Chapter 2

Technologies for Mapping DNA

CONTENTS

Figures

Tables

Chapter 2

Technologies for Mapping DNA

ORGANIZATION AND FUNCTION OF GENETIC INFORMATION

what Is a Genome?

The fundamental physical and functional unit of heredity is the gene. Genetics is the study of the patterns of inheritance of specific traits. The chemical bearer of genetic information is *deoxyribonucleic acid (DNA).* The DNA of multicellular organisms such as insects, animals, and human beings is associated with protein in highly condensed microscopic bodies called *chromosomes.* A single set, or *haploid* number, of chromosomes is present in the egg and sperm cells of animals and in the pollen cells of plants. All body cells, or *somatic cells,* carry a double set, or *diploid* number, of chromosomes, one originating from each parental set, **The entire complement of genetic material in the set of chromosomes of a particular organism is defined as its *genome.***

How Are Genomes organized?

Long before genetic material was identified as DNA, maps of genes on chromosomes were constructed, and many of the details of transmission of genes from generation to generation were elucidated [Judson, see app, A]. The gene for colorblindness, for example, was assigned to the human X chromosome in 1911 (80), about 40 years before the discovery of the structure of DNA. In fact, it has been known for nearly a century that the genetic material:

- has a structure that is maintained in stable form,
- is able to serve as a model for replicas of itself,
- has an information code that can be expressed, and
- is capable of change or variation.

Each of these features can be described in molecular terms based on the structure and function of DNA.

To know how DNA controls cell function, and ultimately the structure and function of an entire organism, it is necessary to understand its structure. In multicellular organisms, DNA is generally found as two linear strands wrapped around each other in the form of a double helix. A DNA strand is a polymeric chain made of *nucleotides,* each consisting of a nitrogenous base, a deoxyribose sugar, and a phosphate molecule (figure z-1). The arrangement of nucleotides along the DNA backbone is called the *DNA sequence.* There are four nucleotides used in DNA sequences: adenosine (A), guanosine (G), cytidine (C), and thymidine (T). The two strands of DNA in the helix are held together by weak bonds between *base pairs* of nucleotides. In nature, base pairs form only between A's and T's and between G's and C's. **The size of a genome is generally given as its total number of base pairs.**

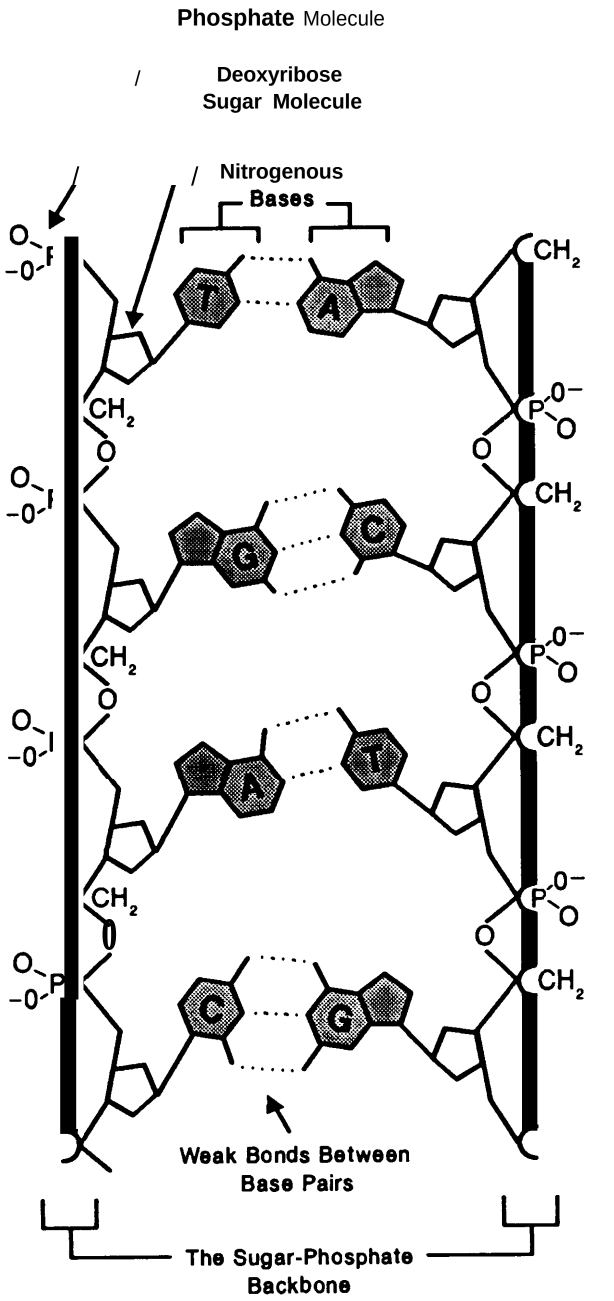
A full genome of DNA is regenerated each time a cell undergoes division to yield two daughter cells. During cell division, the DNA double helix unwinds, the weak bonds between base pairs break, and the DNA strands separate. Free nucleotides are then matched up with their complementary bases on each of the separated chains, and two new complementary chains are made (figure 2-2). In human and other higher organisms, DNA replication occurs in the nucleus of the cell. This DNA replication process was first proposed in 1953 by Francis H.C. Crick and James D. Watson (19,73,74).

What Is the Genetic Code?

Most genes carry an information code that specifies how to build *proteins.* Proteins are an essential class of large molecules that function in the formation and repair of an organism's cells and tissues. Proteins can be components of essential structures within cells, or they can carry out more active roles in the overall function of a particular cell type. Included in this important class of molecules are hormones such as insulin, antibodies to fight cellular infections, and receptors on the cell's surface for modulating interactions between a particular tissue and its surroundings (68). Enzymes are a specialized group of proteins that increase the rate of the biochemical processes that take place in metabolism.

Proteins are long chains of smaller molecules, called *amino acids,* that fold into the unique structures necessary for protein function. The information for generating proteins of specific amino

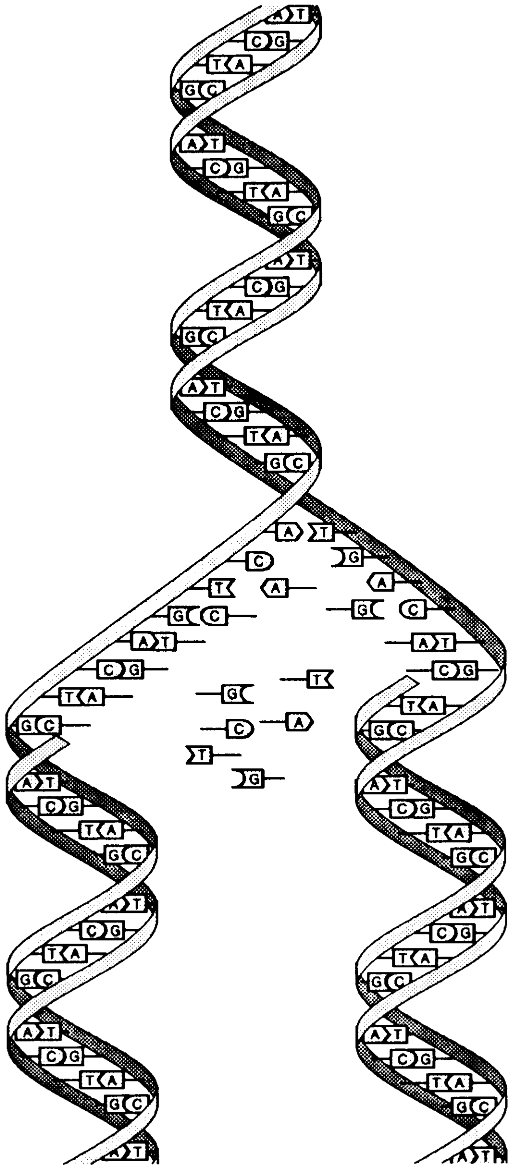
Figure 2-1.—The Structure of DNA


Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
T
A
G
C
CH2
O
-O
P
O-
Weak Bonds Between Base Pairs
The Sugar-Phosphate Backbone

The four nitrogenous bases, adenine (A), guanine (G), cytosine (C), and thymine (T), form the four letters in the alphabet of the genetic code. The pairing of the four bases is A with T and G with C. The sequence of the bases along the sugar-phosphate backbone encodes the genetic information.

SOURCE: Office of Technology Assessment, 1988.

Figure 2-2.—Replication of DNA



When DNA replicates, the original strands unwind and sewe as templates for the building of new, complementary strands. The daughter molecules are exact copies of the parent, each daughter having one of the parent strands.

SOURCE: Office of Technology Assessment, 1988.

acid sequences is found in the genetic code-a code based on sequences of nucleotides that are “read” in groups of three (table 2-I). Genetic information is transmitted from DNA sequences to protein via another large molecule called *messenger ribonucleic acid (mRNA).* The structure of ribonucleic *acid (RNA) is* very similar to that of DNA. Figure 2-3 illustrates the major steps in gene expressi’on, namely:

Table 2.1.—The Genetic Code

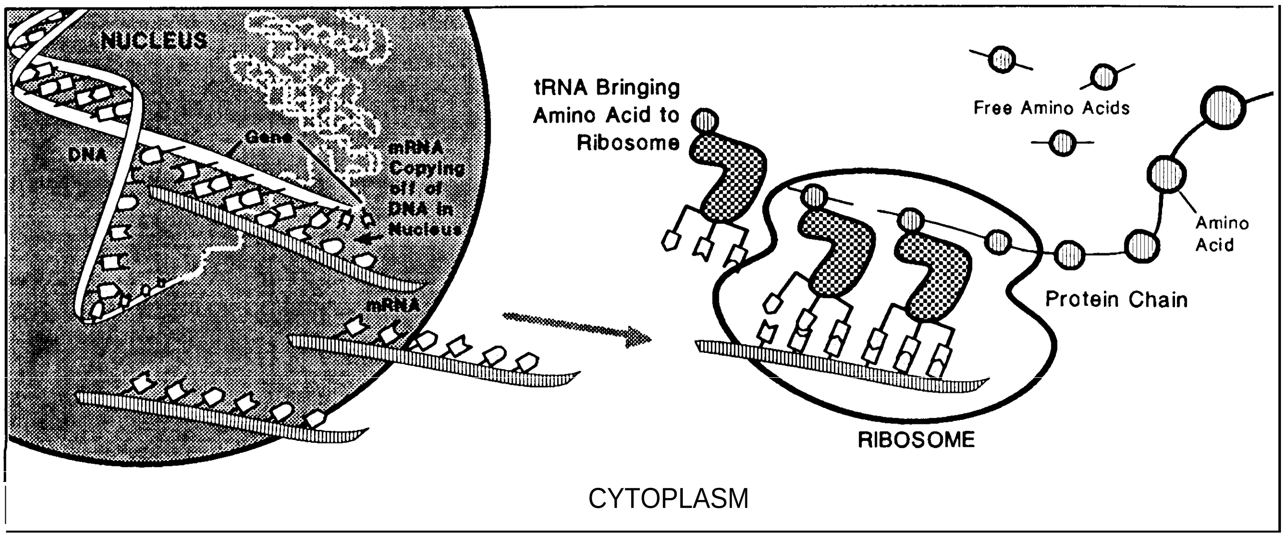
Codon	Amino Acid	I Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid
Uuu	Phenylalanine	Ucu	Serine	UAU	Tyrosine	UGU	Cysteine
Uuc	Phenylalanme	Ucc	Serine	UAC	Tyrosine	UGC	Cysteine
UUA	Leucine	UCA	Serine	UAA	stop	UGA	stop
UUG	Leucme	UCG	Serine	UAG	stop	UGG	Tryptophan
Cuu	Leucme	Ccu	Proline	CAU	Histidine	CGU	Arginine
Cuc	Leucine	ccc	Proline	CAC	Histidine	CGC	Arginine
CUA	Leucine	CCA	Proline	CM	Glutamine	CGA	Arginine
CUG	Leucme	CCG	Proline	CAG	Glutamine	CGG	Ar~inine
AUU	Isoleucine	ACU	Threonine	MU	Asparagine	AGU	Serine
AUC	Isoleucine	ACC	Threonine	MC	Asparagine	AGC	Serine
AUA	Isoleucine	ACA	Threonme	AAA	Lysine	AGA	Arginine
AUG	Methionine (start)	ACG	Threonine	AAG	Lysine	AGG	Arginine
GUU	Valine	GCU	Valine	GAU	Aspartic acid	GGU	Glycine
GUC	Valine	GCC	Alanine	GAC	Aspartic acid	GGC	Glycine
GUA	Valine	GCA	Alanine	GAA	Glutamic acid	GGA	Glycine
GUG	Valine	GCG	Alanine	GAG	Glutamic acid	GGG	Glycine

Each codon, or triplet of nucleotides in RNA, codes for an amino acid. Twenty different amino acids are produced from a total of 64 different RNA codons, but some amino acids are specified by **more** than one codon (e.g., phenylalanine is specified by UUU and by UUC). In addition, one codon (AUG) specifies the start of a protein, and three **codons** (UAA, UAG, and UGA) specify the termination of a protein. Mutations in the nucleotide sequence can change the resulting protein structure if the mutation alters the amino acid **specified by a codon or if it alters the reading frame by deleting or adding a nucleotide.**

U=uracil (thymine) **A =adenine**
C=cytosine **G =guanine**

SOURCES: Office of Technology Assessment and National Institute of General Medica Sciences, 198S.

Figure 2“3.—Gene Expression



In the first step of gene expression, messenger RNA (mRNA) is synthesized, or transcribed, from genes by a process somewhat similar to DNA replication. In higher organisms, this process takes place in the nucleus of a cell. In response to certain signals (e.g., association with a particular protein), sequences of DNA adjacent to, or sometimes within, genes control the synthesis of mRNA. Protein synthesis, or translation, is the second major step in gene expression. Messenger RNA molecules are known as such because t hey carry messages specif ic to each of t he 20 different amino acids that make up proteins. Once synthesized, mRNAs leave the nucleus of the cell and go to another cellular compartment, the cytoplasm, where their messages are t translated into the chains of amino acids that make up proteins. A single amino acid is coded by a sequence of three *nucleotides* in the mRNA, called a codon. The main component of the translation machinery is the ribosome—a structure composed of proteins and another class of RNAs, ribosomal RNAs. The ribosome reads the genetic code of the mRNA, while a third kind of RNA molecule, transfer RNA (tRNA), mediates protein synthesis by bringing amino acids to the ribosome for attachment to the growing amino acid chain. Transfer RNAs have three nucleotide bases that are complementary to the codons in the mRNA (see table 2-l).

SOURCE: Office of Technology Assessment, 1988.

- *transcription* of DNA into mRNA, and
- *translation* of mRNA into protein.

By these processes, the genetic code directs amino acids to be joined together in the order specified by the sequence of nucleotides in the messenger RNA, which was in turn determined by the sequence of nucleotides in the DNA.

In molecular terms, a gene is a region of a chromosome whose DNA sequence can be transcribed to produce a biologically active RNA molecule. Messenger RNAs constitute the major class of biologically active RNAs. Other RNAs may act as lattices to stabilize certain cell structures or may participate directly in important cellular processes such as protein synthesis.

How Big Is the Human Genome?

The diploid human genome consists of 46 chromosomes—22 pairs of *autosomes* and 1 pair of sex *chromosomes* (two X chromosomes for females and an X and a Y chromosome for males). A single egg cell has 22 different autosomes and a single X chromosome, whereas sperm cells carry 22 different autosomes and either an X or a Y chromosome.

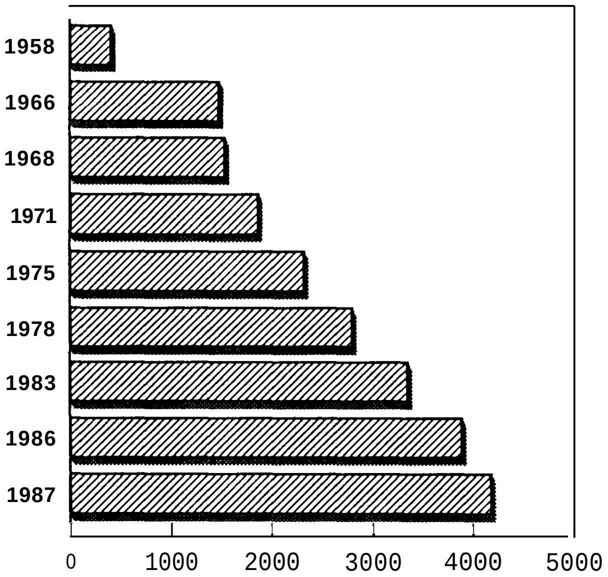
Scientists estimate the total number of human genes per haploid genome at 50,000 to 100,000. The characterization of the structure of human genes on chromosomes was made possible recently through *recombinant DNA technology* (the use of molecular biology tools to combine DNA from one organism with that of another). It is now known that human genes can vary in size from fewer than 10)000 base pairs to more than 2 million. The entire haploid genome is approximately 3 billion base pairs, So far, researchers are far from having determined where each human gene is located on the 24 chromosomes. Victor McKusick of The Johns Hopkins University maintains Mendelian *Inheritance in Man,* an encyclopedia of expressed genes [see app. D]. According to the October 1, 1987, count, 4,257 genes were represented in the encyclopedia; of those, at least 1,200 had been mapped to specific chromosomes or regions of chromosomes (51). Figure 2-4 illustrates the years of effort invested thus far in identifying even this small fraction of the total number of human genes.

How Does the Human Genome Compare to Other Genomes?

Before much was known about the DNA sequences that make up genomes, it was thought that the amount of DNA per haploid genome would increase in proportion to the biological complexity of the organism. Since chromosomes can vary in size, the total amount of DNA in a haploid cell is a better indicator of actual genome size than the number of chromosomes. Table 2-2 shows that higher plants and animals do have much more DNA than lower organisms, There are some notable exceptions, however, to the correlation between overall genome size and complexity of the organism, A good example is the salamander, which has a haploid DNA content more than 30 times greater than that of humans, even though it is obviously a smaller, less complex organism. Similarly, the cells of some species of plants have a greater DNA content than human cells (72).

This inconsistency between DNA content and the apparent complexity of an organism is known to geneticists as the *C-va.lueparadox* (C-value refers to the haploid genome size). A great deal of research has been devoted to determining the scientific basis for the C-value paradox. Variations

Figure 2-4.–Number of Human Gene Loci Identified From 1958 to 1987



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

Table 2-2.—Haploid Amounts of DNA in Various Organisms

Organism	Number of base pairs (millions)
Bacterium	4.7
Yeast. ... ,	15
Nematode	80
Fruit fly	155
Chicken	1,000
Human	2,800
Mouse	3,000
Corn	15,000
Salamander	90,000
Lilv	90,000

SOURCES:
B. Alberts, D.Bray, J. Lewis, et al., Mo/ecu/arBio/ogy of fhe Ce//(NewYork, NY: Garland Publishing, 1983)
C. Burks,GenBank", LosAlamosNational Laboratory, Los Alamos, NM,personal communication, March 1988.
T. Cavalier-Smith (cd), *The Evo/utionof* Genorne Size (New York, NY: Wiley& Sons, 1985)
J. Darnell, H. Lodish, and D. Baltimore, *Molecular Cc// Biology (New* York, NY: Scientific American, 1988),

in genome size usually arise from increases in the amount of DNA per chromosome, not from increases in numbers of chromosomes. The genomes of all higher organisms contain sequences of DNA that occur as large numbers of repeated units, either clustered in one chromosomal region or in regions dispersed throughout the entire genome. These repeated sequences contribute to wide variations in total DNA content among what are often closely related species.

In large genomes such as the human genome, *intron* sequences also contribute to size. Introns are DNA sequences occurring within the coding region of a gene. They are transcribed into mRNA, but are cut (spliced) out of the message before it is translated into protein. Introns can increase the number of base pairs in a gene by more than tenfold. Many genes also have long regions at their ends that are transcribed into mRNA but are not translated into protein. In addition, some protein-coding genes have given rise to gene *families* that make several closely related protein products. Other gene families consist of hundreds or thousands of closely related genes (72).

The untranslated sequences within or at the ends of genes, gene families, and moderately or frequently repeated DNA sequences between genes still do not account for all of the DNA in the genomes of higher organisms, nor for the variations in genome size among these organisms. Many scientists interpret these facts to mean that some fraction of DNA in the human genome is expendable; although there is little agreement on the size of this fraction, some believe it to be more than 90 percent of the genome (27,54). The implication of the C-value paradox, that much of human DNA is expendable, is one reason that some esteemed scientists do not favor a major effort to obtain a complete nucleotide sequence of the human genome. They believe time would be better spent identifying and understanding the function of gene products that contribute to the cellular processes leading to the development of an organism as complex as man (1). On the other hand, some scientists consider the C-value paradox to be one of the many mysteries that might be unraveled once entire genomes have been analyzed in greater detail.

Why Does Hereditary Information Change?

Hereditary variation is the result of changes occurring by *mutation* —a change in the sequence or number of nucleotides—which occurs during DNA replication. Mutations formed in sex cells are inherited by offspring, whereas those that occur in somatic cells remain only in the affected organism. Some diseases, such as certain human cancers, arise from factors in both of these categories. Mutations are also acquired by artificial means, such as exposure to chemicals or certain forms of radiation.[1] Such factors can cause a change in a single DNA base pair that may modify or inactivate a protein, if one is encoded in that region of the chromosome.

More extreme mutations, involving changes in the structure of a single chromosome or changes in chromosome number, can occur; for example:

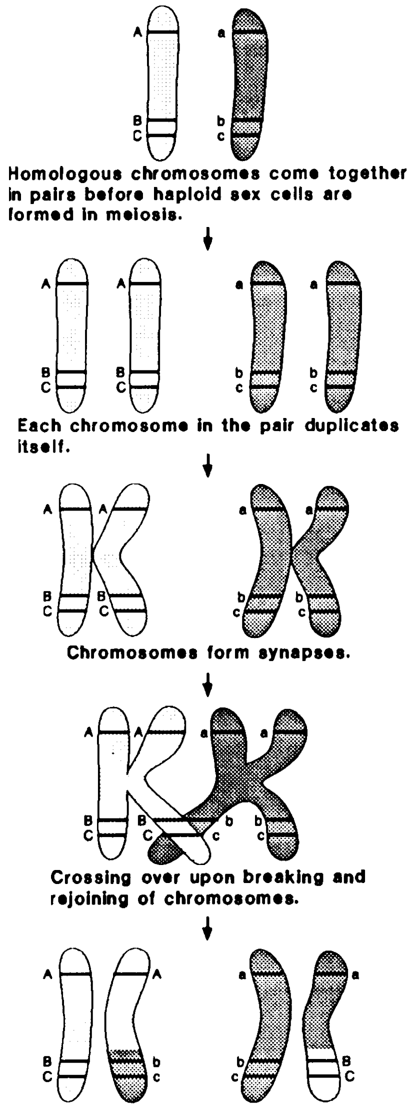
- deletion of a chromosome,
- *duplication* of a chromosome or a piece of a chromosome,

[1] A 1986 OTA report, ***Technologies for Detecting Heritable Mutations in Human** Beings,* addresses the kinds and effects of mutations in human beings and new technologies for detecting mutations and measuring mutation rates.

- *translocation,* or insertion of a chromosome fragment from one chromosome pair into an unmatched member of a different pair, and
- *inversion,* or the breakage of a chromosome fragment followed by its rejoining in the opposite orientation.

In diploid cells, there is a tendency for each DNA molecule to undergo some form of modification or rearrangement with each cell division. The progenitors of sex cells area special class of diploid cells that undergo two rounds of cell duplication in a process called meiosis. Meiosis results in four instead of two daughter cells, each with a haploid set of chromosomes. Before the first meiotic cell division, each member of a chromosome pair is replicated, forming two sets of chromosome pairs. At this stage, the cell has two identical copies of chromosomes of maternal origin and two identical copies of chromosomes of paternal origin. Also at this time, the chromosome pair of maternal origin is in close association with that of paternal origin, and an event called *crossing over* can occur; that is, one maternal and one paternal chromosome can break, exchange corresponding sections of DNA, and then rejoin (figure 2-5). (This process is also referred to as *recombination.)* In this way, two of the four resulting sex cells have chromosomes with new combinations of genes, while the other two cells carry the parental (original) combinations of genes. Since chromosomes originating maternally or paternally can carry different forms of any given gene, new combinations of traits are created by such crossovers.

Figure 2-5.—Separation of Linked Genes by Crossing Over of Chromosomes During Meiosis



Homologous chromosomes come together in pairs before haploid sex cells are formed in meiosis.

Each chromosome in the pair duplicates itself.

Chromosomes form synapses.

Crossing over upon breaking and rejoining of chromosomes.

Chromosomes with new gene combinations after crossing over.

SOURCE: Office of Technology Assessment, 1988.

GENETIC LINKAGE MAPS

Because of recombination during meiosis, certain groups of traits originating on one chromosome are not always inherited together (figure 2-5). The closer, or more linked, genes are on a particular chromosome, the smaller the probability that they will be separated during meiosis. Each chromosome is inherited independently of all others, so only genes on the same chromosome can be linked.

Gene mapping, broadly defined, is the assignment of genes to chromosomes. A genetic linkage map permits investigators to ascertain one genetic locus relative to another on the ba-

sis of how often they are inherited together. Strictly speaking, a genetic locus is an identifiable region, or *marker,* on a chromosome. The marker can bean expressed region of DNA (a gene) or some segment of DNA that has no known coding function but whose pattern of inheritance can be determined, Variation at genetic loci is essential to genetic linkage mapping, **The markers that serve to identify chromosome locations must vary in order to be useful for linkage studies in families, because only when the parents have different forms at the marker locus can linkage to a gene be followed in their children.** *Alleles* are the alternative forms of a particular genetic locus. For example, at the locus for eye color, there are blue and brown alleles. During meiosis, all of the genetic loci on a chromosome remain together unless they are separated by crossing over between chromosome pairs.

Distance on genetic maps is measured by how often a particular genetic locus is inherited separately from some marker. This measure of genetic distance is called *recombination frequency*. The amount of recombination is expressed in units called *centimorgans.* One centimorgan is equal to a 1 percent chance of a genetic locus being separated from a marker due to recombination in a single generation.

During the generation of sex cells in human beings, if a gene and a DNA marker are separated by recombination in 1 percent of the cases studied, then they are, on average, separated by 1 million base pairs. The relationship between genetic map distance (recombination frequencies) and physical map distance (measured in DNA base pairs) can vary, however, by five-or even tenfold. Recombination can vary from near zero, if genetic loci are very close, to 50 percent, between genetic loci that are far apart on the chromosome or on different chromosomes. Some chromosome regions are highly prone to recombination and exhibit high recombination frequencies, while other chromosome regions appear to be resistant to recombination. Interestingly, the rate of recombination in the same region of a particular chromosome typically varies among males and females, and it is often greater in females. The reasons for this have not been established. Double or multiple crossover events can also occur between two loci that are widely separated. Each of these variations in recombination frequencies complicates the relationship between genetic and physical maps. Nevertheless, if a genetic linkage map were constructed with a set of markers separated by an average of 1 centimorgan, then most genes could be located within a range of 100)000 to 10 million base pairs.

Genetic linkage between two or more observable traits can be established with greater certainty in large populations. For this reason, large families are preferred for mapping studies. If two genetic loci are closely linked, then their separation by recombination during meiosis is unlikely and a large family must be studied to determine how close they are on the genetic map. As more loci are placed on the genetic map, it becomes possible to determine the location of a new trait on the basis of its inheritance pattern compared to two or three others already on the map. The frequency with which multiple traits are inherited together generally must be calculated for many individuals over many generations before genetic mapping results are statistically significant.

The X chromosome is particularly amenable to linkage analysis because male traits directly reflect the genes on the single X chromosome present. For this reason, the genetic linkage map of the X chromosome is the most nearly complete of all chromosome maps.

Mapping of genetic loci on autosomes, on the other hand, is not as easy, unless the gene is found to be linked to a marker that has already been mapped through the study of family inheritance patterns. The first assignment of a gene to a specific autosomal chromosome came in 1968, when researchers showed that the Duffy blood group, which can be identified in families by biochemical methods, is linked to a variation in chromosome 1 (23). About the same time, the feasibility of correlating specific genes with particular chromosomes or chromosome regions by a technology called somatic cell hybridization was demonstrated (75). This and other experimental methods developed in the 1970s radically advanced the study of human genetics, allowing investigators to locate autosomal genes on human genetic and physical maps [Judson, see app. Al (50,58).

LINKAGE MAPS OF RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

The advent of recombinant DNA technology in the 1970s brought about a tremendously useful new way to create genetic linkage maps. Examination of DNA from any two individuals reveals that variations in DNA sequence occur at random about once in every 300 to 500 base pairs (37). These variations occur both within and outside of genes, and most do not lead to functional changes in the protein products of genes. Kan and Dozy (40) were the first to demonstrate this phenomenon experimentally, by showing that one particular DNA sequence, recognized by the restriction enzyme HpaI, was lost in certain individuals. *(Restriction enzymes* are proteins that recognize specific, short nucleotide sequences and cut the DNA at those sites.) This alteration in the DNA correlated with the inheritance of sickle cell disease.

This important discovery led researchers to propose that natural differences in DNA sequence *(pol'orphisms)* might replace other chemical and morphological markers as a way to track chromosomes through a family (5,64). In addition to polymorphisms in *restriction enzyme cutting sites,* it is possible to detect differences among individuals in the number of copies of short DNA sequences repeated in tandem. Polymorphic sequences can occur within a restriction enzyme cutting site or between sites. In either case, the lengths of DNA fragments generated upon cutting the DNA with restriction enzymes will vary among individuals having different alleles at such locations. These polymorphic sequences are thus commonly referred to as *restriction fragment length polymorphism (RFLP)* markers.

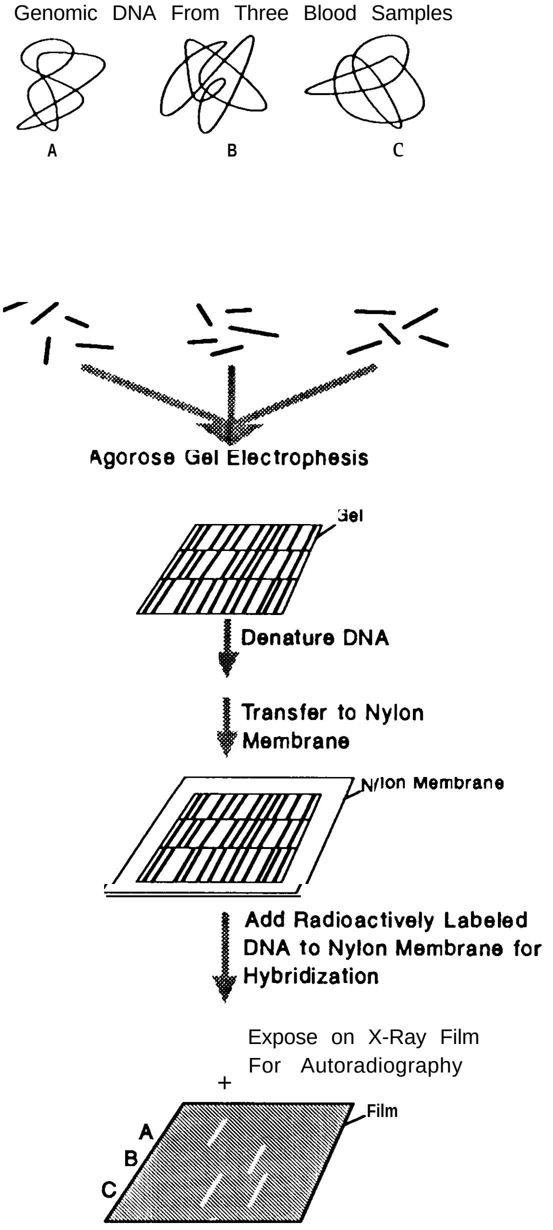
In 1983, genetic linkage between a RFLP marker on chromosome 4 and Huntington's disease (a neurological disease that usually strikes its victims by the age of 35) was discovered (31), paving the way for the general use of RFLPs as markers for genes responsible for inherited disorders. The more frequently a RFLP marker is inherited with the gene, the more likely it is to be physically close to the gene, and hence the more useful it is as a gene marker. The major limitations to the usefulness of RFLP markers are how polymorphic they are (how much they vary among individuals), how many other markers exist in the same region, and the extent to which DNA samples of large families are available for analysis (43,44).

RFLP Mapping

A RFLP map is a type of genetic linkage map, consisting of markers distributed throughout the genome. Construction of the map involves determining the linkages between RFLP markers, their arrangement along the chromosomes, and the genetic distances between them. RFLP markers are identified and mapped by comparing the sizes and numbers of restriction enzyme fragments generated from different individuals. Just as genetic loci representing expressed DNA segments have alternate, or allelic, forms, so may RFLPs. The value of any marker depends mostly on how many variants it displays. The more often the marker varies in a population, the more likely it is that an individual will inherit two different alleles at the marker location (one on each member of a matched pair of chromosomes), making it possible to detect recombination between markers in that individual's offspring (76).

In RFLP mapping, DNA obtained from white blood cells (lymphocytes) or other tissues of several different individuals are first cut into fragments using restriction enzymes (figure 2-6). The fragments are then separated by size. This is accomplished by a procedure called *electrophoresis,* in which a mixture of DNA fragments of various sizes is placed in a polymeric gel (e.g., agarose) and then exposed to an electric field. Because the chemical makeup of DNA gives it a net negative charge, the DNA fragments will travel in an electric field toward a positive electrode. Large DNA fragments will move more slowly than small ones, thus the mixture is separated, or resolved, according to size. With very large pieces of DNA, the use of restriction enzymes yields numerous fragments along the entire length of the gel, making it necessary to identify RFLPs using radioactively labeled, single-strand segments of DNA called *DNA probes (65).* RFLP markers are identified by virtue of their ability to form base pairs (hybridize) with DNA probes that have complementary sequences of nucleotides. Some useful probes for RFLP mapping are fragments of genes; others are randomly isolated DNA segments that identify polymorphisms; still others are complementary to sequences with variable numbers of tandemly repeated, shorter sequences that occur frequently within the genome. A technique called *autoradiography* is used to show the image of a band on an X-ray film wherever the agarose gel held a restriction enzyme fragment that hybridized to the DNA probe. Where polymorphisms occur, different patterns will be observed among samples taken from different individuals [Myers, see app. A] (figure 2-6).

Figure 2-6.—Detection of Restriction Fragment Length Polymorphisms Using Radioactively Labeled DNA Probes


Genomic DNA From Three Blood Samples
A
B
C
Agorose Gel Electrophesis
Gel
Denature DNA
Transfer to Nylon Membrane
Nylon Membrane
Add Radioactively Labeled DNA to Nylon Membrane for Hybridization
Expose on X-Ray Film For Autoradiography
+
Film
A
B
C

Variations in DNA sequences at particular marker sites are observed as differences in numbers and sizes of DNA fragments among samples taken from different individuals (shown here as samples A, B, and C).

SOURCE: Off Ice of Technology Assessment, 1988,

When Is a Map of RFLP Markers Complete?

Botstein and co-workers (5) predicted in 1980 that only 150 different markers would be needed to link all human genes to chromosomal regions containing RFLPs. In practice, however, it has been estimated that many more markers may have to be studied and evaluated in order to find the minimum number which would be randomly distributed over the genome (45). It now appears that hundreds of DNA probes for highly polymorphic sequences, scattered widely over the genome, will be required for a complete human linkage map (77).

With a l0-centimorgan map, for example, there is a greater than 90 percent chance of being able to determine the rough chromosomal location of any gene associated with an inherited disease. Raymond White and colleagues at the Howard Hughes Medical Institute at the University of Utah have taken advantage of one such tandemly repeated sequence, known as VNTR, to create a set of probes useful for making a complete RFLP map of the human genome (76). White's RFLP map, with continuously linked landmarks separated on average by 10 centimorgans (about 10 to 20 million base pairs), is nearly complete. At the ninth international Human Gene Mapping Workshop, held in September 1987, he reported 475 markers covering 17 human chromosomes, based on the DNA from 59 different three-generation families. White's group and other geneticists believe that a l-centimorgan RFLP marker map, determined from normal families and consisting of thousands of markers spaced an average of 1 million base pairs apart, would be the ideal research tool (see ch. 3 for further discussion) (17,52).

Another group, led by Helen Donis-Keller at Collaborative Research, Inc. (Waltham, MA), reported its own RFLP linkage map, consisting of 403 markers an average of 9 centimorgans apart. A new gene or marker on their map can be located relative to the existing markers 95 percent of the time (24).

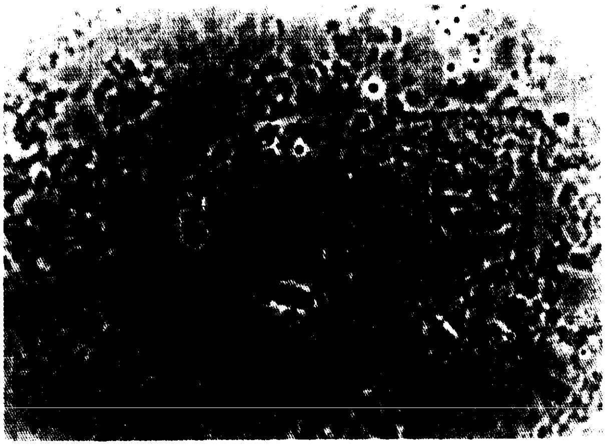
As physical markers that can be followed genetically, RFLPs are the key to linking the genetic and physical maps of the human genome. RFLP linkage maps, as well as linkage maps of expressed genes, can be correlated with banding patterns and other identifiable regions of chromosomes by somatic cell hybridization and in *situ* hybridization. These and other relatively low resolution physical mapping technologies are described in the following sections.

LOW-RESOLUTION PHYSICAL MAPPING TECHNOLOGIES

A physical map is a representation of the locations of identifiable landmarks on DNA. For the human genome, the physical map of lowest resolution is found in the banding patterns on the 22 autosomes and the X and Y chromosomes observable under the light microscope. This map has at most 1,000 landmarks (i.e., visible bands) (57).

Another type of relatively low resolution physical map illustrates the positions of expressed segments of DNA relative to certain regions of the chromosome or to specific chromosome bands. Expressed genes include those that are transcribed into mRNA and then translated into protein, and another class of essential genes that are transcribed into RNA but not translated into proteins. Included in the latter class are transfer and ribo - somal RNAs involved in protein synthesis, RNAs involved in the removal of intron sequences from mRNAs, and an RNA associated with the cellular protein secretion machinery. Procedures are available to make DNA copies, or *complementary DNAs* (cDAVIS), of RNA transcripts. These cDNAs can in turn be mapped to genomic DNA sequences by somatic cell hybridization, *in situ* hybridization, and other low-resolution physical mapping methods. A physical map illustrating the location of expressed genes is often referred to as a cDNA map. As noted earlier, only 1,200 of the 50,000 to 100,000 human genes have been physically mapped to chromosomes.

A high-resolution physical map can be made by cutting up the entire human genome with restriction enzymes and ordering the resultant DNA segments as they were originally oriented on the chromosomes. This third type of physical map, a contig map, can be related to the maps of chromosome bands and expressed genes. **The physical map of highest possible resolution or greatest molecular detail, is the complete nucleotide sequence of the human genome.** Thus there is a continuum of mapping techniques that ranges from low to high resolution (see table 2-2). These techniques are discussed in this section, on low-resolution physical mapping, and in the following one, on high-resolution physical mapping methods.



M

B *D* *m* *g* ry
g m m
g m

Somatic Cell Hybridization

The somatic cell hybridization technique for gene mapping typically employs human fibroblast and rodent tumor cells grown in culture. The hu -

man and mouse cells are fused (hybridized) together using certain chemicals, Sendai virus, or an electric field, as illustrated in figure 2-7 (58). The chromosomes of each of the fused cells become mixed, and many of the chromosomes are lost from the hybrid cell. Human chromosomes are preferentially lost over rodent chromosomes, but there is generally no preference for which human chromosomes are lost. The individual hybrid cells are then propagated in culture and maintained as cell lines. In practice, the hybrid cell lines resulting from cell fusions contain different subsets of between 8 to 12 human chromosomes in addition to rodent chromosomes (58).

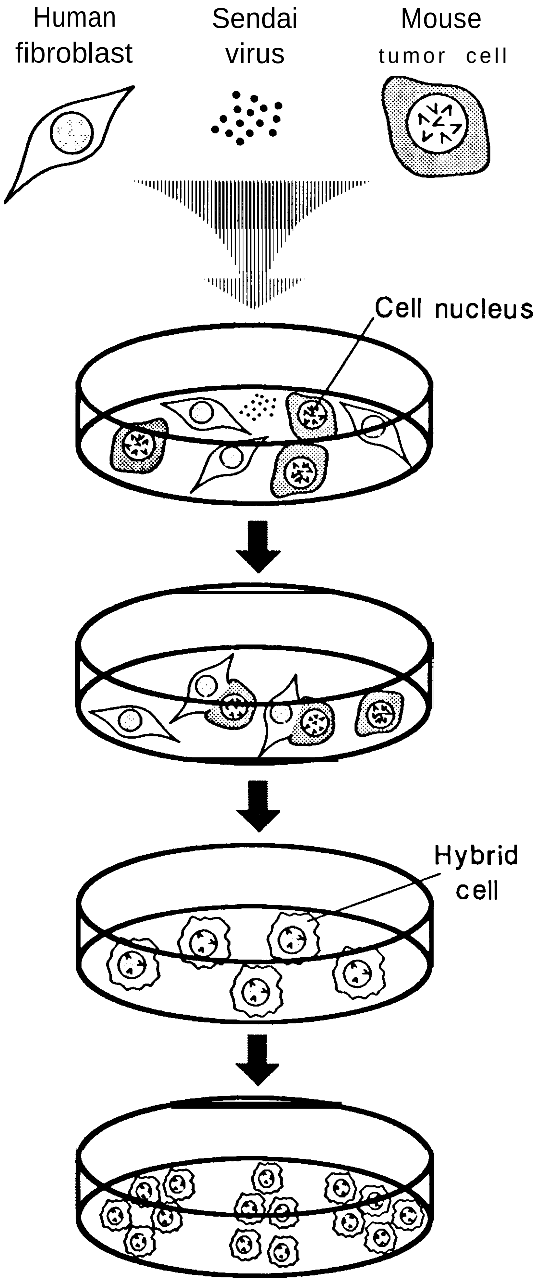
Using a large set (panel) of somatic cell hybrids containing different chromosome combinations, it is possible to correlate the presence or absence of a particular chromosome with a particular gene. Assignment of a gene to a chromosome is made by detecting a protein produced by a hybrid cell line and associating it with the chromosome unique to that cell line. Alternatively, if the gene to be mapped has already been isolated by DNA cloning procedures, then the gene can be used directly to identify complementary nucleotide sequences in the DNA extracted from somatic cell hybrids.

Modifications of the somatic cell hybridization method have been devised to generate single chromosome hybrids; to date, hybrid cells containing single copies of human chromosomes 7, 16, 17, 19, X, and Y are available (58). Somatic cell hybrid lines carrying chromosomes with deletion or translocation mutations are also useful low-resolution mapping tools because they make it possible to infer the location of a particular gene.'

Chromosome Sorting

Chromosome sorting offers an alternative to the screening of somatic cell hybrid panels for low-resolution gene mapping. In this approach, DNA hybridization is used to map genes to chromosomes that have been differentiated by flow

Figure 2-7.—Somatic Cell Hybridization



Somatic cell hybrids are generated by the process of cell fusion, an event that can be enhanced by adding Sendai virus. Initially, the hybrid cell contains complete sets of chromosomes from both parental cells, but hybrids of human and mouse cells are unstable and chromosomes from the human cells are preferentially lost. After a few generations in culture, a line of hybrid cells is established that contains both mouse and human chromosomes.

SOURCE: Office of Technology Assessment, 1988.

[7] The Institute for Medical Research, in Camden, New Jersey, estab lished a repository for SCHs with chromosome rearrangements called the Human Mutant Cell Library. The availability of this centralized storage facility has accelerated the rate of mapping human genes to specific chromosomal locations.

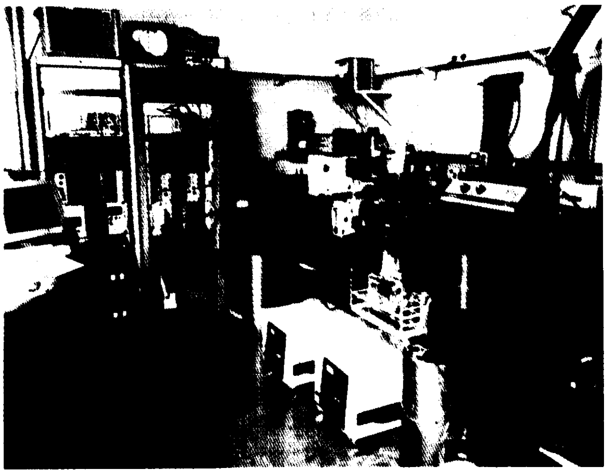


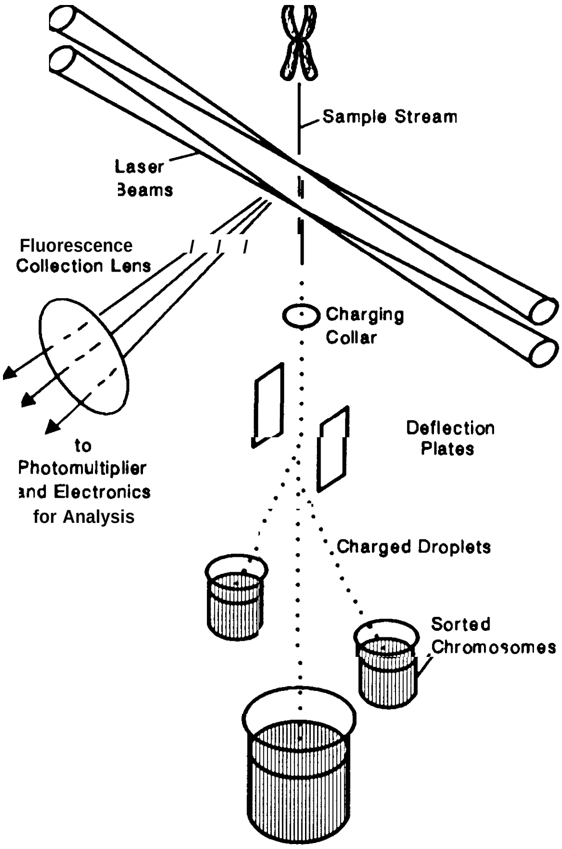
Photo credit: Larry Deaven, Los Alamos National Laboratory, Los Alamos, NM

Flow cytometry facility for chromosome sorting at Los Alamos National Laboratory.

cytometry and purified by flow sorting. Fluorescent markers that bind to chromosomes are used in flow cytometry as the basis of separating chromosomes from one another in a flow sorter (figure 2-8) (21,29,30,46). Because human chromosomes differ in the degree to which they bind the fluorescent markers, it is possible to use this approach to physically separate some chromosomes from others. The dual-laser chromosome sorter has been used successfully to separate all the human chromosomes except chromosomes 10 and 11. In addition, chromosomes from cell lines with translocations and deletions can be used to narrow the location of the gene to a certain chromosomal region (46).

To determine on which chromosome a gene lies, chromosomes are sorted onto different paper filters made of nitrocellulose. There the DNA is denatured and hybridized with a radioactively labeled DNA probe complementary to the gene to be mapped. (In general, the cDNA is available for use as a probe for the gene.) On whichever chromosome the gene appears, the two sequences will hybridize, and the hybridization can be observed using autoradiography.

Figure 2-8.—Chromosome Purification by the Flow Sorter



Chromosomes stained with a fluorescent dye are passed through a laser beam. Each time, the amount of fluorescence is measured and the chromosome deflected accordingly. The chromosomes are then collected as droplets.

SOURCE: Courtesy of Los Alamos National Laboratory, Los Alamos, NM

Karyotyping

At a stage of cell division when chromosomes have duplicated but not yet separated from one another, they condense to form structures with features that can be observed under a light microscope. The structure of human chromosomes can be studied by chemically fixing white blood cells at the appropriate stage of cell division and then photographing the chromosome spreads as they appear on slides under the microscope. Individual chromosomes are identified in the photograph, cut out, and, in the case of autosomes, matched with their morphologically identical chromosome partner to generate a *karyotype*. Karyotyping has been most useful for correlating gross chromosomal abnormalities with the characteristics of specific diseases (e.g., Down's syndrome and Turner's syndrome).

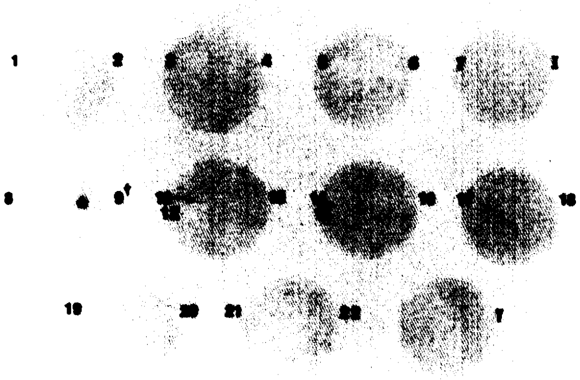


Photo credit: Roger Lebo, the University of California, San Francisco Medical Center. Reprinted with permission from Alan R. Liss, Inc.

Assignment of a gene and genes with related sequences to specific human chromosomes. Samples of the 21 different human chromosomes were sorted onto 11 circular filters and then hybridized to a radioactively labeled DNA probe from the aldolase gene (aldolase is an enzyme involved in the metabolism of sugars). Most of the radioactive signal in the autoradiograph appears on the filter with chromosome 9, indicating that the complementary DNA sequence is in that chromosome. The autoradiograph also shows some hybridization to chromosomes 17 and 10, indicating that aldolase genes with different, but similar, sequences are found on these chromosomes.

Chromosome Banding

Using fluorescent dyes as chromosome-specific stains, Caspersson and others (10-I2) developed optical methods for observing the banding patterns on human chromosomes. These methods reveal more details of chromosome morphology than does simple karyotyping. The bands are chromosomal regions that appear as stripes on chromosome spreads when viewed under the light microscope. Each of the 24 different human chromosomes has a unique banding pattern, thus the bands can be used to identify individual chromosomes. Genes can be mapped to specific bands by identifying differences between the banding patterns on chromosomes from normal individuals and those on the chromosomes from an individual with a significant chromosomal alteration.

Nearly 1,000 distinct bands have been detected on the 24 human chromosomes by staining and light microscopy, and an average of 100 genes is represented in a single band (50). Chromosome banding is a useful procedure for finding the general location of a gene, but it does not offer sufficient resolution to identify the exact position of a gene relative to other genes mapped in the same region (58).

In Situ *Hybridization*

Family linkage and somatic cell hybridization are not direct mapping methods; they are based on the correlation between traits and the frequency of transmission of those traits in families. Karyotyping and analysis of chromosome banding allow a specific trait to be correlated with a particular chromosome or a large region of a chromosome. Advances in molecular biology have overcome the limitations of those techniques by providing means for more precise mapping of genetic markers. One such method is *in situ* hybridization of isolated genes or gene fragments to chromosomal DNA.

The *in situ* hybridization technique was originally developed by Mary Lou Pardue and Joseph Gall for detection of genes encoding ribosomal RNAs in chromosomes from *Drosophila* salivary glands (56). In the typical in *situ* hybridization experiment, the DNA corresponding to a particular gene or gene fragment is used to probe for complementary sequences in chromosomes (28). The chromosomes to be analyzed are fixed on a microscope slide, where the DNA strands are chemically treated and separated. Next, the radioactively labeled DNA probe is mixed with the chromosomes on the slide. Under proper conditions, the DNA probe hybridizes with the gene sequence wherever it is located on the prepared chromosomes.

Results of *in situ* hybridization can be seen by exposing the slides to a photographic emulsion for a long period, then analyzing the photographs under a microscope. Wherever the radioactively labeled DNA strands have paired with complementary chromosomal regions, tiny silver grains appear. The location of a specific gene can be found by counting the number of grains in each region and using computer methods to analyze the data (58). Although *in situ* hybridization has been a principal method for the mapping of human genes to autosomes, higher-resolution methods are nec -

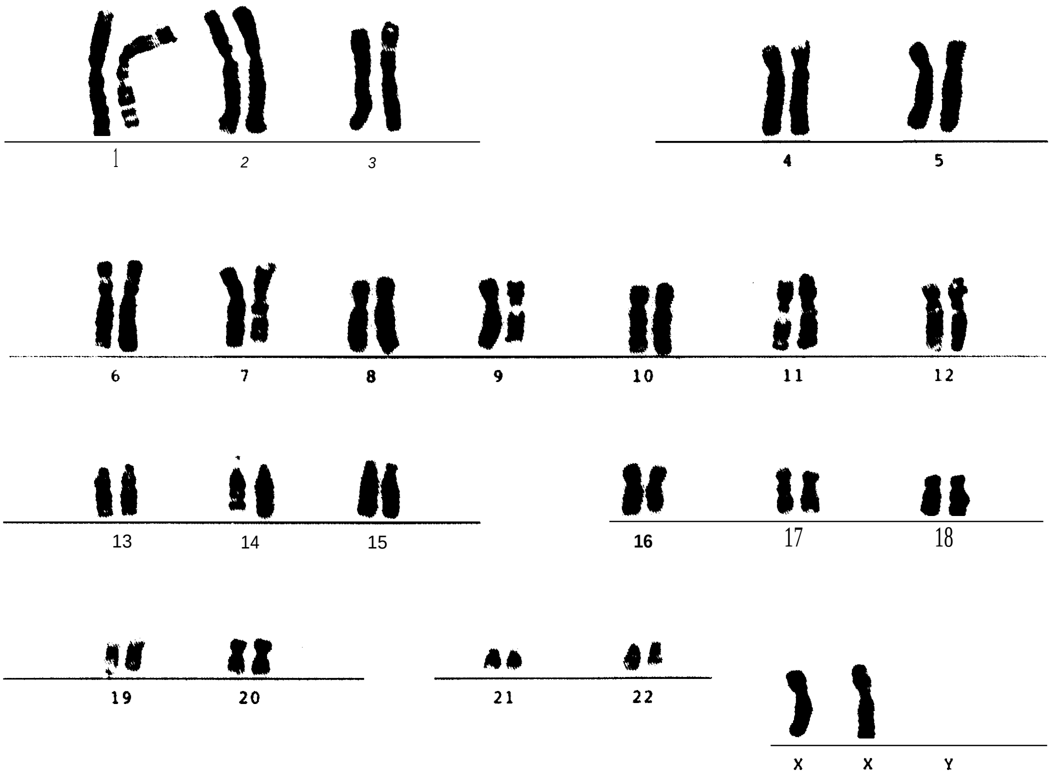


Photo credit: The Genetics and IVF Institute, Fairfax, VA

Human karyotypes of a normal female.

essary. The procedure is limited to a resolution of about 10 million base pairs, a substantial portion of the total length of most chromosomes. Since many genes could fit into such a region, the exact location of the gene of interest must still be determined precisely (58).

Other Methods for Mapping Genes

Several other techniques for mapping human genes are available, including gene dosage mapping and comparative mapping of species. In gene dosage mapping, a correlation is made between the amount of gene product and the presence of extra genes or the absence of a gene or chromosome fragment, Biochemical analysis of cellular contents isolated from an individual with a particular genetic disease, or from somatic cell hybrid lines derived from that individual's cells, is performed to measure amounts of gene products. The structure of the altered chromosome (or chromosomes) is then characterized by one or more of the methods already described.

Comparative mapping of species can provide useful human gene mapping information. This is particularly true among mammals, where it has been demonstrated that different species have similar patterns of gene organization on certain chromosomes [Computer Horizons, Inc., see app. A]. For example, tabulations show that all of the human autosomes except chromosome 13 have at least two linked genes which are also linked in the mouse (35).

Comparison of the banding patterns of chromosomes from different species have also proved useful in matching chromosomes between species, even though differences in total numbers of chromosomes exist. There is, for example, a striking resemblance between chimpanzee and human

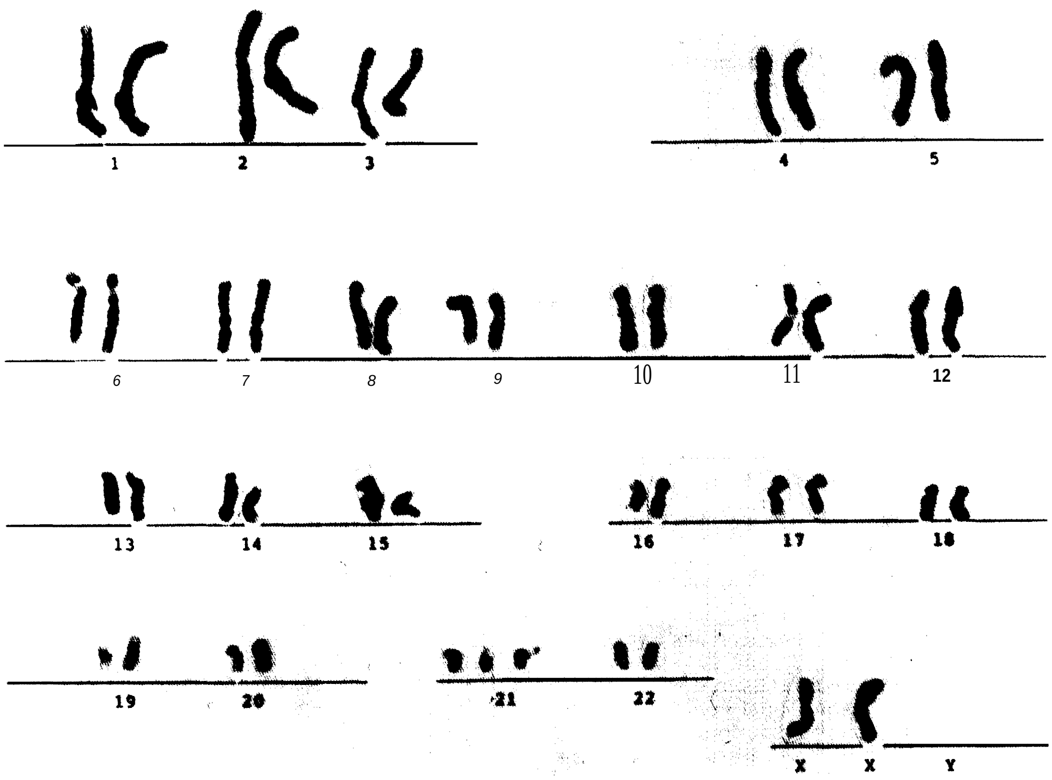


Photo credit: The Genetics and IVF Institute, Fairfax, VA

A female with the extra chromosome 21 associated with Down's syndrome.

chromosomes (81). In recent years, DNA sequence comparisons between widely differing organisms have been used to isolate or confirm the identity of specific human genes (see ch. 3).

HIGH-RESOLUTION PHYSICAL MAPPING TECHNOLOGIES

Construction of high-resolution physical maps of whole genomes involves cutting the component DNA with restriction enzymes, analyzing the chemical characteristics of each fragment, and then reconstructing the original order of the fragments in the genome. Generally, the DNA fragments to be ordered are isolated from chromosomes; united with carrier, or vector, DNA molecules originating from viruses, bacteria, or the cells of higher organisms; and introduced into suitable host cells, where the isolated DNA can be reproduced in large quantities, A fragment of DNA is said to be cloned when it is stably maintained as part of a DNA vector in a single line of cells. A set of clones representing overlapping segments of DNA encompassing an entire genome is called a genomic *library*. In order to make a physical map, the clones in the genomic library must be ordered in relation to one another's position on the chromosome. The following sections describe in more detail the methods currently available for creating high-resolution physical maps and their application to the genomes of specific organisms.

Cloning Vectors as Mapping Tools

Any genome mapping project first requires the isolation, usually by cloning technologies, of fragments of chromosomal DNA. Several different

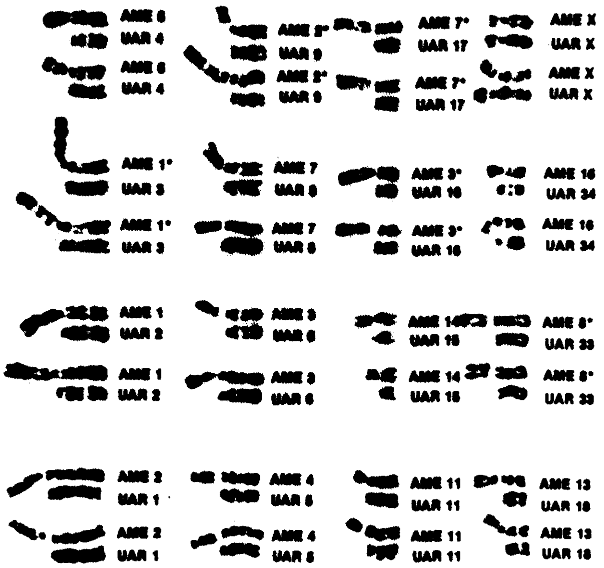


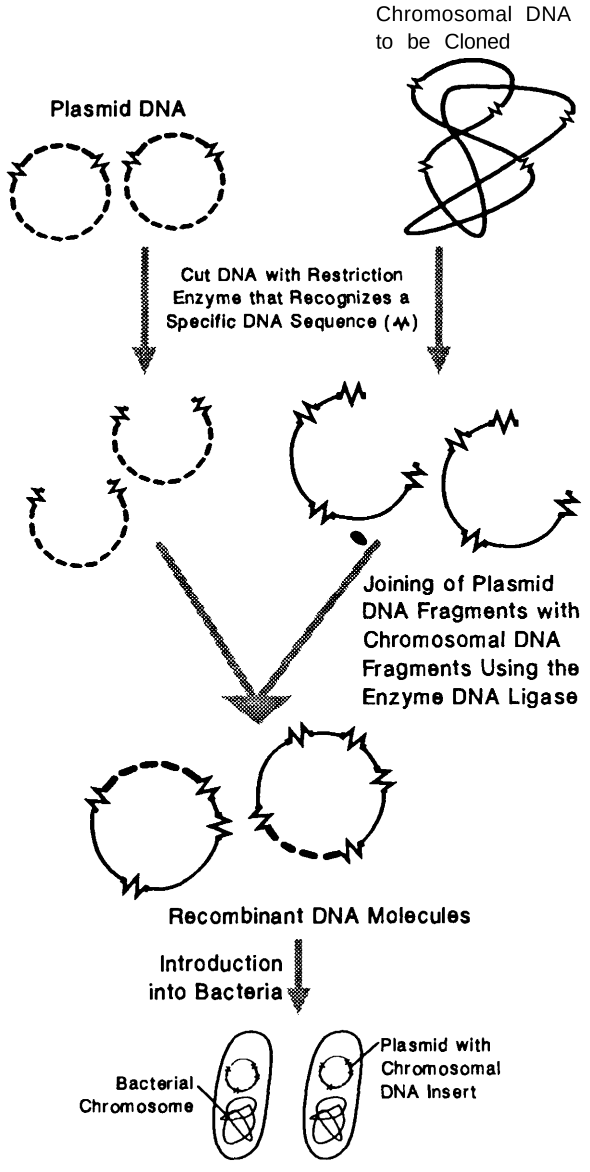
Photo credit: Stephen O'Brien, The National Cancer institute, Frederick, MD. Reprinted with permission from **Nature 317:14G144, 1985.**

A comparative alignment of chromosomes from the giant panda (AME) and the brown bear (UAR). The putative matches between the whole chromosomes, or chromosome segments, of each animal were based on the thickness of stained bands on the chromosomes and on the spacing and intensity of the bands. This type of molecular information has been used to establish the phylogeny of these bears and to demonstrate some of the problems in using the appearance of animals, instead of their chromosome structure, in studies of evolution.

types of cloning vectors have been developed using recombinant DNA technology:

- Plasmid *vectors* are circular DNA molecules of 1,000 to 10,000 base pairs that can carry additional DNA sequences in fragment inserts up to 12,000 base pairs (2,4). Plasmids exist as minichromosomes in bacterial cells (usually between 10 to 100 copies per cell) and are separate from the main bacterial chromosome.
- *Phage lambda chromosomes* are about 50,000 base pairs and can accept foreign DNA inserts up to about 23,000 base pairs (33,79). Just as viruses infect human cells, phage infect bacterial cells and generate hundreds of descendants.
- *Cosmid vectors* are plasmids that also contain specific sequences from the bacterial phage lambda. Cosmids are about 5,000 base pairs, but because they contain phage lambda sequences, they can carry DNA inserts up to about 45)000 base pairs (figure 2-9) (25,34,36).
- Yeast *artificial chromosomes* are dasrnids containing portions of yeast chrohosomal DNA that function in replication. These artificial chromosomes can accommodate foreign DNA fragment inserts nearly 1 million base pairs long (6).

Figure 2-9.—DNA Cloning in Plasmids


Chromosomal DNA to be Cloned
Plasmid DNA
Cut DNA with Restriction Enzyme that Recognizes a Specific DNA Sequence (⋀⋁)
Joining of Plasmid DNA Fragments with Chromosomal DNA Fragments Using the Enzyme DNA Ligase
Recombinant DNA Molecules
Introduction into Bacteria
Bacterial Chromosome
Plasmid with Chromosomal DNA Insert

SOURCE: Office of Technology Assessment, 1988.

Most of the physical mapping work carried out to date has employed bacteriophage and cosmid cloning vectors because the yeast artificial chromosome vectors have only recently been developed [Myers, see app. A].

physical Mapping of Restriction Enzyme Sites

With the exception of DNA sequencing, restriction enzyme mapping is the method that gives the highest-resolution picture of DNA as it is organized in a chromosome. Several basic steps are involved in the construction of this type of physical map for part or all of a genome:

- purifying chromosomal DNA,
- fragmenting DNA by restriction enzymes,
- inserting all the resulting DNA fragments into DNA vectors to establish a collection (library) of cloned fragments, and
- ordering the clones to reflect the original order of the DNA fragments on the chromosome.

Variations in any of these steps can affect the resolution of the physical map.

Purification of Chromosomal DNA

Whole chromosomes are the best source of DNA for genomic libraries. Mixtures of chromosomes can be extracted directly from cells, but for organisms with complex genomes, such as human beings, it might be desirable to first separate the different chromosomes and then create sets of clones from the individually purified chromosomes.

Mixtures of whole chromosomes extracted from human cells can be sorted by flow cytometry. Somatic cell hybrid lines carrying one or a few human chromosomes can also be used as a highly enriched source of particular chromosomes. **The refinement of existing methods and the development of new technologies for obtaining large amounts of purified human chromosomes will be crucial in the early stages of human genome mapping projects.**

Fragmentation of DNA

The availability of chromosome fragments of decreasing size allows mapping at higher resolution. A technology called pulsed field gel electrophoresis (PFGE) allows separation of DNA molecules ranging in size from 20,000 to 10 million or more base pairs (8,9)13)61) [Myers, see app. A].

During PFGE, large DNA fragments are subjected to an electric field that is switched back and forth across opposite directions for short pulses of time. This alteration in the direction of the electric field allows very large DNA molecules (up to tens of millions of base pairs) to migrate into the agarose gel and separate from one another, even though the normal size limit for electrophoretic separation of DNA molecules during conventional agarose gel electrophoresis is about 50,000 base pairs. This method is so powerful that it has been used successfully to separate all 14 of the yeast chromosomes from each other (figure 2-10) [Myers, see app. A]. Since intact human chromosomes have an average size of approximately 100 million base pairs, the PFGE technique is only useful for separating large fragments made from individual, purified human chromosomes.

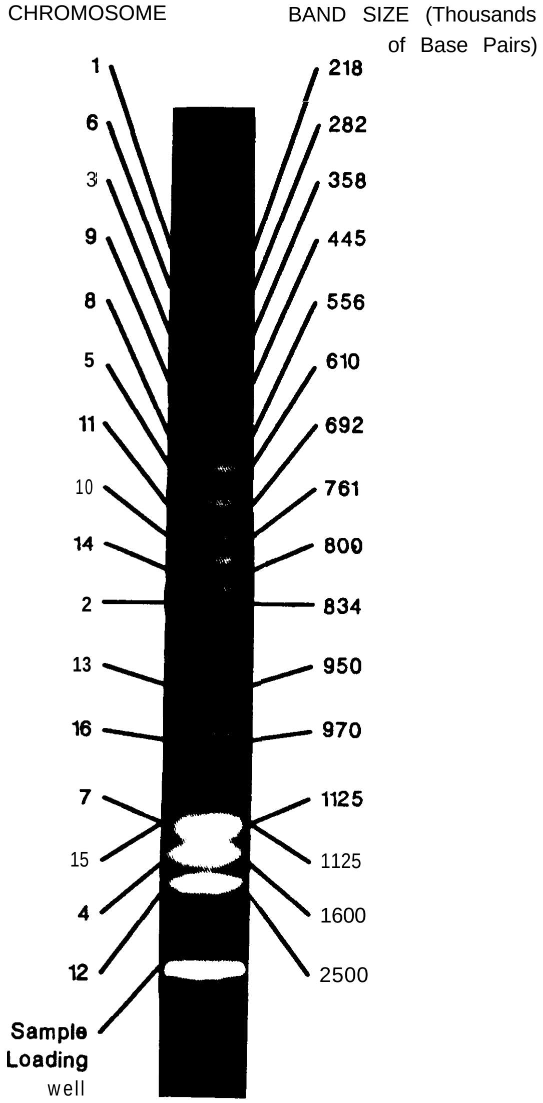
The level of detail possible on a physical map depends on the restriction enzyme or enzymes used. There are a few restriction enzymes that cut DNA very infrequently, generating small numbers of large fragments (ranging from several thousand to a million base pairs). Most restriction enzymes cut DNA more frequently, generating large numbers of small fragments (ranging from fewer than a hundred to greater than a thousand base pairs). The relative order of a small set of large fragments is easier to determine than the order of a large set of short fragments, but it gives a lower-resolution physical map. The choice of enzyme thus depends on the purpose of the physical map. If the aim is to have fragments of a size amenable to DNA sequencing, then a mapped restriction site at least every 500 base pairs would be ideal, but a mapped site every 2,000 to 3,000 base pairs would also be practical. **Given the technology currently available, sequencing the DNA of the 3-billion-basepair haploid human genome might require the prior mapping**

of as many as 6 million restriction enzyme cutting sites (69) [Myers, see app. Al.

Construction of Libraries of DNA Fragments

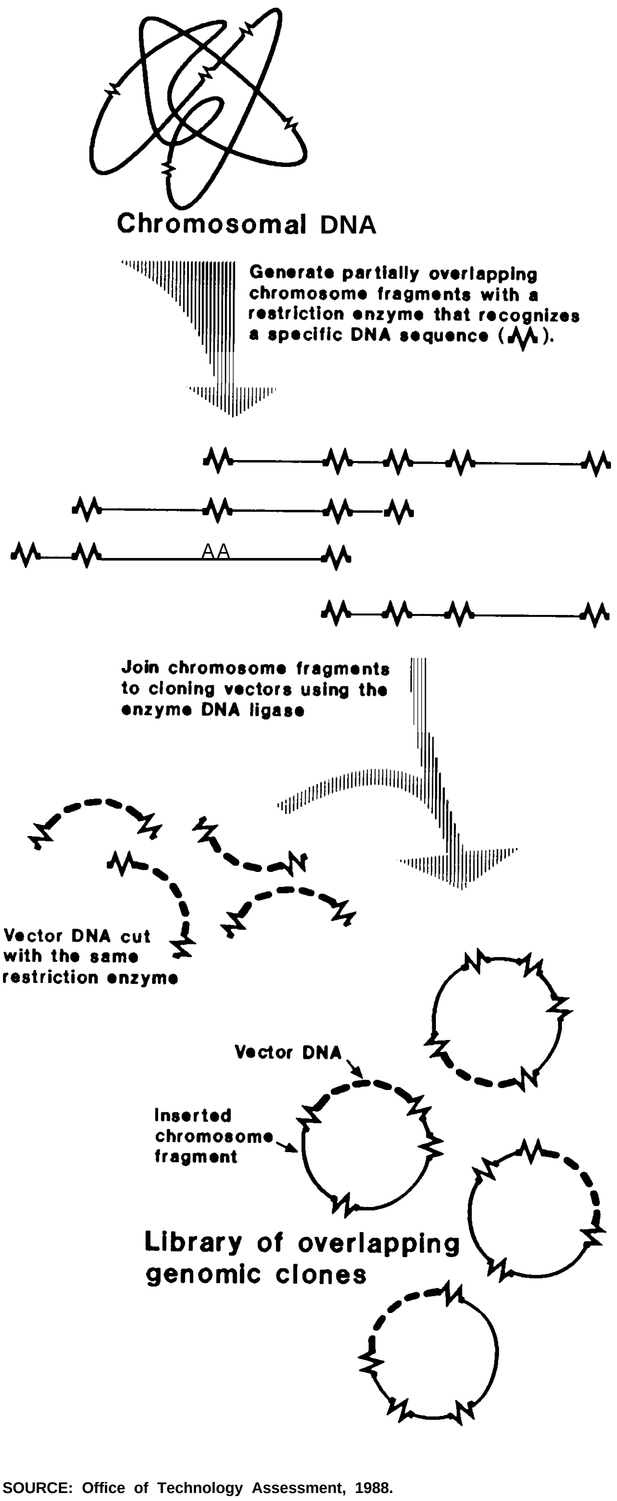
For physical mapping projects, it is important to have as much DNA as needed. The use of cloned DNA fragments offers this advantage. Fragments of DNA from whole chromosomes are generally cloned into vectors such as plasmids, cosmids,

Figure 2-10.—Separation of Intact Yeast Chromosomes by Pulsed Field Gel Electrophoresis



SOURCES: Chris Traver and Ronald Davis, California Institute of Technology, Pasadena, CA.

Figure 2.11.—Constructing a Librarv of Clones Containing Overlapping Ch;omosomal Fragments


Chromosomal DNA
Generate partially overlapping chromosome fragments with a restriction enzyme that recognizes a specific DNA sequence ( ).
AA
Join chromosome fragments to cloning vectors using the enzyme DNA ligase
Vector DNA cut with the same restriction enzyme
Vector DNA
Inserted chromosome fragment
Library of overlapping genomic clones

SOURCE: Office of Technology Assessment, 1988.

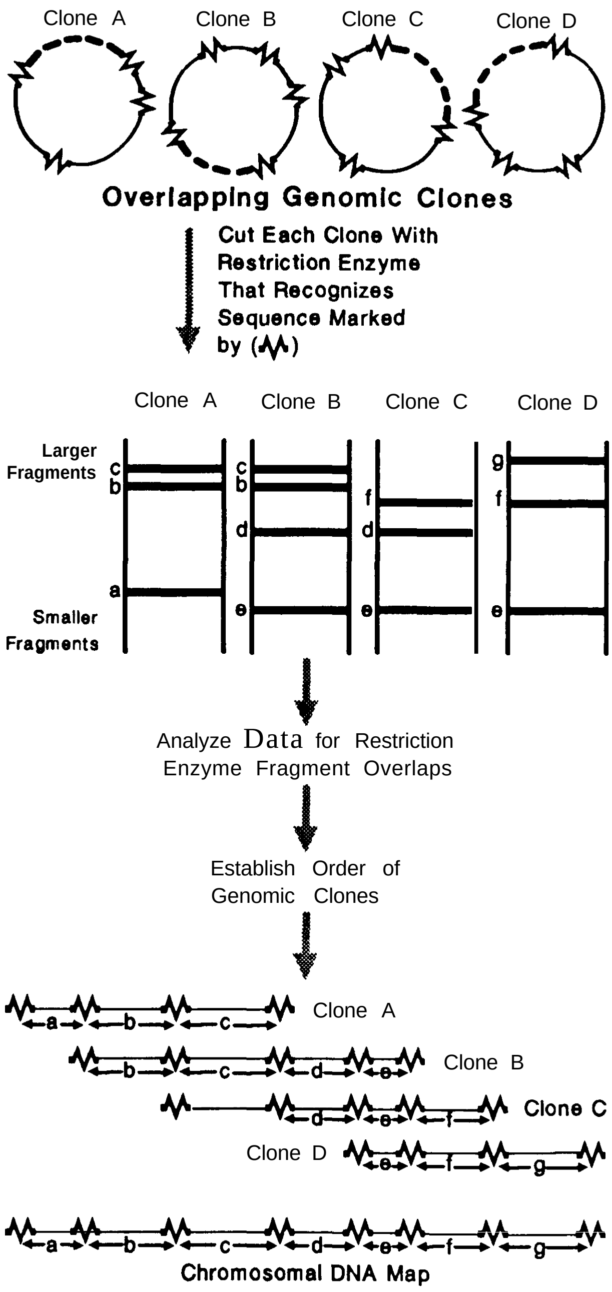
phage chromosomes, and artificial yeast chromosomes. These vectors can be stably maintained in host cells (bacteria or yeast) that multiply rapidly to provide the amounts of DNA necessary for restriction enzyme mapping and DNA sequencing. DNA fragments are usually cloned by cutting the vector of choice with a restriction enzyme and then connecting the newly generated ends of the vector to the ends of the DNA fragments with the enzyme DNA ligase. The resuhing collection of clones is called a *library.* There is no obvious order to the library, and the relationship between the components can only be established by physical mapping.

In order to establish that any two clones represent chromosomal segments that normally occur next to one another in the genome, it is necessary to have collections of clones representing partially overlapping regions of chromosomal DNA (figure 2-11). To create libraries of overlapping clones, the chromosomal DNA is treated with a frequent-cutting restriction enzyme, one that cuts every 500 base pairs or so, but conditions are controlled so that the enzyme is not allowed to cut the DNA at all the possible restriction enzyme sites. Instead, by lowering the amount of restriction enzyme used, only partial cutting is allowed. The experimental conditions for partial cutting are adjusted so that DNA fragments are generated with an average size equal to the vector's capacity (usually 20,000 to 50,000 base pairs). In theory, no one of the cutting sites will be recognized by the restriction enzyme more frequently than another, so a population of overlapping segments representing all possible cutting sites in the original DNA sample should be generated. These fragments are then cloned in the appropriate vector.

Determination of the Order of Clones

The clones in a library are ordered by subdividing the chromosomal DNA inserts into even smaller fragments and identifying which clones have some common subfragments. Figure 2-12 illustrates how this is done. A particular DNA clone (vector plus the chromosomal DNA insert) is cleaved with one or more restriction enzymes (other than that used to make the clones) under conditions in which all sites are recognized and cut. The resulting fragments are then run on a gel made of agarose. After electrophoresis, a pattern of fragments is observed along the length of the gel. If the DNA fragments are present in sufficient amounts, they can be seen under ultraviolet light after staining the gel with the dye

Figure 2-12.—Making a Contig Map


Clone A
Clone B
Clone C
Clone D
Overlapping Genomic Clones
Cut Each Clone With Restriction Enzyme That Recognizes Sequence Marked by (-ʌʌ-)
Clone A
Clone B
Clone C
Clone D
Larger Fragments
Smaller Fragments
c
b
a
c
b
d
e
f
d
e
g
f
e
Analyze Data for Restriction Enzyme Fragment Overlaps
Establish Order of Genomic Clones
a
b
c
Clone A
b
c
d
e
Clone B
d
e
f
Clone C
Clone D
e
f
g
a
b
c
d
e
f
g
Chromosomal DNA Map

SOURCE: Office of Technology Assessment, 1988.

ethidium bromide; otherwise, the phosphates at the ends of the DNA fragments are labeled with a radioactive isotope and viewed after autoradiography. A unique pattern of bands appears (corresponding to DNA fragments) for any given clone because of the unique arrangement of restriction enzyme sites in the region of the chromosome from which that clone was derived. If two clones contain overlapping segments of DNA, then a portion of the banding pattern for each will be identical. For example, if the clone order is A-B-C-D, then the restriction enzyme fragments from clone A will partially overlap with those from clone B, clone B fragments with clone C, and so on (figure 2-12).

Groupings of clones representing overlapping, or contiguous, regions of the genome are known as *contigs (18)66).* on an incomplete physical map, contigs are separated by gaps where not enough clones have been mapped to allow the connection of neighboring contigs. Of all the steps in physical mapping, the connection of all the contigs is the one that faces the greatest number of technical problems. Therefore, **the time required to achieve a complete physical map of any ge-**

GENOMIC MAP OF BACTERIOPHAGE T4

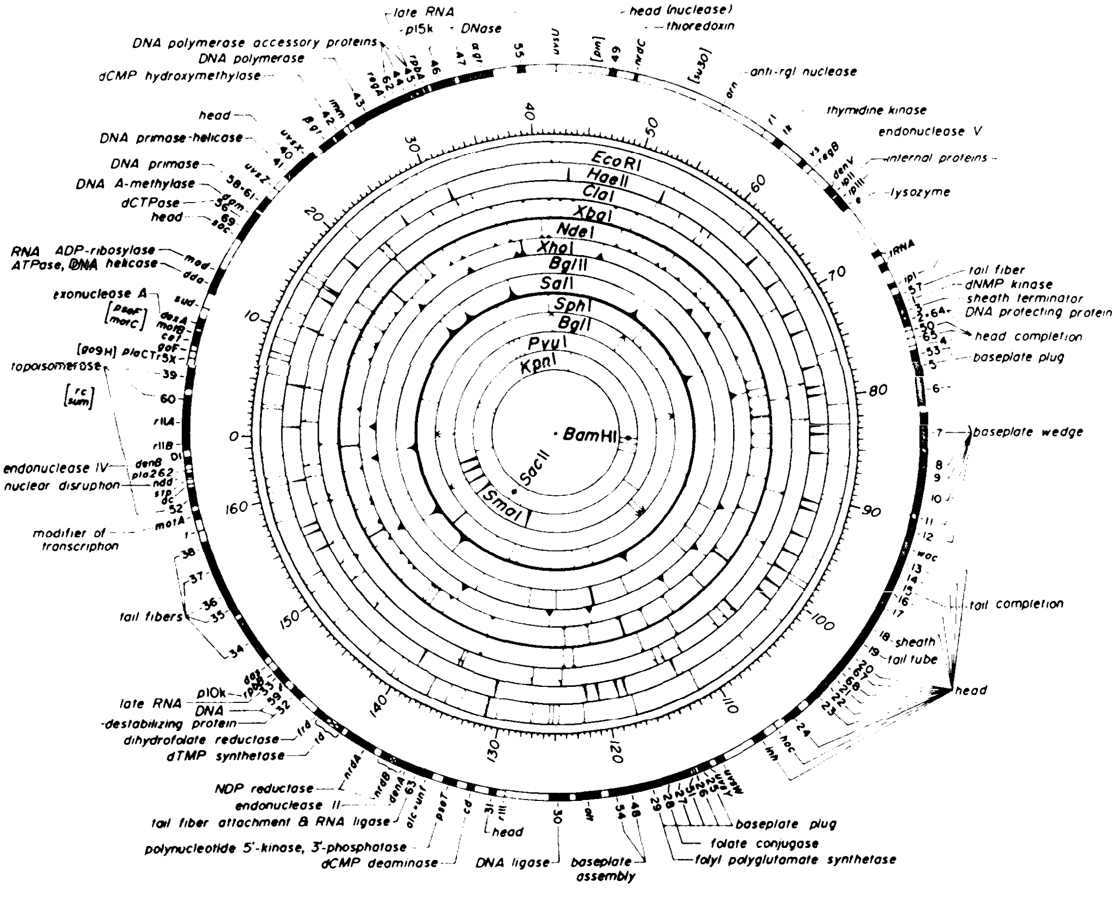


Photo credit: Elizabeth Kutter and Burton Guttman, Evergreen State College, Olympia, WA. Reprinted with permission from Stephen J. O'Brien (ed.), Genetic Maps 1987, *vol. 4, (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1987).*

Genomic map of bacteriophage T4. The genome of bacteriophage T4 contains about 166,000 base pairs. Shown are maps illustrating the names of mapped genes (outer circle), genetic map distances (second largest circle), which are measured in *mimdes* in bacteria and their phage, and posit ions of DNA cutting sites for a variety of restriction enzymes (inner circles).

nome is a function of the time required to connect neighboring contigs.

Physical Mapping of Nonhuman Genomes

So far, high-resolution mapping of entire genomes has focused on nonhuman organisms. Most of the technologies applicable to human genetic and physical mapping, therefore, have been developed from work on other organisms. Mapping of complete genomes is well underway for several species of bacteria and yeast, for the nematode, and is beginning for the fruit fly. These organisms have long served as excellent model systems for what are sometimes found to be universal genetic and biochemical mechanisms governing cell physiology. The technologies employed in these high-resolution genome mapping projects range from making contig maps to fine mapping by DNA sequencing.

The Bacterial Genome

For many years, bacteria (mainly *Escherichia coli)* and phage (viruses that infect bacteria) have been principal research subjects of molecular biologists, molecular geneticists, and biochemists. Because of the relative ease of studying gene function in E. coli, it is the organism whose genetics and biochemistry are closest to being completely understood. The DNA of this bacterium is contained in a single circular chromosome of 4.7 million base pairs (63). The genetic map of E. coli is quite extensive, with about 1,200 of the 5,000 or so known genes already cloned (3). In addition, the nucleotide sequence of over 20 percent of this bacterial genome is known (26).

Progress on the physical map of the E. coli genome is good. Cassandra Smith and co-workers at Columbia University made a complete physical map of this genome using a restriction enzyme called Not I, which cuts DNA only infrequently (63). Not I recognizes a sequence eight nucleotides long that is expected to occur by chance once every 34,000 base pairs. Only 22 Not I sites were found in the *E. coli* genome (63).

A higher-resolution physical map of *E. coli* was generated by Kohara and colleagues (42) at Nagoya University in Japan. These researchers devised an innovative, rapid mass-analysis mapping approach involving eight different restriction enzymes. In a period of time equivalent to only one-half of a person-year, this group generated a high-resolution physical map covering 99 percent of the *E. coli* genome, leaving only seven gaps. An independent effort to generate a high-resolution map of the cutting sites for three different, frequent-cutting restriction enzymes is also near completion at the University of Wisconsin, Madison (20).

The work of the researchers in Wisconsin and Japan is important because it generates an ordered set of clones. A map indicating the order of a library of genomic clones is immediately useful to anyone wishing to examine DNA corresponding to a gene whose position on the map is known. The physical map is correlated with the genetic map at many sites in E. *coli,* primarily as a result of including in the analysis clones containing known genes. Kohara and co-workers demonstrated that the use of large fragments for connecting groups of clones is not necessary for *E. coli.* Because of the computational limitations on connecting great numbers of small fragments, however, large-fragment maps, analogous to the Not I map of E. coli, will no doubt play a significant role in mapping large genomes, such as the human genome.

The Yeast Genome

An ongoing project to map the 15 million base pairs in the *Saccharomyces cerevisiae* (baker's yeast) genome has been described by olson and colleagues (55) at Washington University. These researchers initiated the mapping project to facilitate the organization of the vast amount of information already available on this organism. As Olson writes:

> Just as conventional cartography provides an indispensable framework for organizing data in fields as diverse as demography and geophysics, it is reasonable to suppose that "DNA cartography" will prove equally useful in organizing the vast quantities of molecular genetic data that may be expected to accumulate in the coming decades (55).

A large fraction of the S. cerevisiae genome (about 95 percent) is available in clones that have been joined together in over 400 contiguously mapped stretches. These contigs are being correlated with a complete large-fragment restriction map for the yeast genome. These combined maps make it possible to construct or identify a mapped region 30,000 to 100,000 base pairs in length around virtually any starting point, typically a cloned gene [Mount, see app. A].

The Nematode Genome

The nematode *Caenorhabditis elegans* is a popular organism among developmental biologists because the origin and function of all 958 cells in the adult animal are known, offering researchers the opportunity to study the basis of organis - mal development. With its 3-day generation time, C. *elegans* is also suited to genetic studies, Molecular biologists, interested in the molecular basis of development, would find an ordered set of clones from the nematode genome particularly useful for their work [Mount, see app. A].

Coulson and Sulston at the Medical Research Council in England initiated a C. *elegans* mapping project to provide such tools and to establish communications among the laboratories working on this organism, Like the S. *cerevisiae* genome mapping project, this resulted in a set of clones that covers most of the genome (18). One difference is that the C. *elegans* clones are put into order by the fingerprinting method: Distances from each cleavage site for one enzyme to the nearest site for a second enzyme were measured, and clones sharing a number of such distances (measured as lengths of restriction fragments observed on polyacrylamide gels) were considered to overlap. This process makes identification of overlapping regions somewhat easier (because the information is denser), at the expense of more precise physical map information. A second difference is that cosmid clones were used in the nematode project, while phage clones were used in the yeast project. Cosmid clones can accommodate larger DNA inserts than phage clones, but they can also be less stable, with portions of the inserts becoming deleted more often (17). At present, over 700 contigs, ranging from 35,000 to 350,000 base pairs in length and representing 90 percent of the C. *elegans* genome, have been characterized (71).

The Fruit Fly Genome

The genetics of the common fruit fly, *Drosophila melanogaster,* are the best characterized of any multicellular organism. One reason for studying fruit flies is that it is possible to carry out a saturating screen to detect mutations of a particular type. In a saturating screen, every gene that could mutate to produce the defect being studied is identified. (This accomplishment is crucial to a complete understanding of many cellular processes.) The saturating screen technique allows for a comprehensive genetic analysis because the entire genome can be examined for the presence of genes that are involved in a particular process. The most celebrated example is an exhaustive study of mutations that are lethal to the fly in its larval stage (39,53,78) [Mount, see app. A].

Until recently, the physical mapping of the 165 million base pairs in the *D. melanogaster* genome had not been undertaken by any one laboratory. Roughly 500 to 1,000 genomic clones have been obtained in various laboratories in various vectors; all of these clones have been localized to a chromosomal map position by *in situ* hybridization to polytene chromosomes (a multicopy set of *D. melanogaster* chromosomes unique to its salivary gland). A listing of these clones is maintained by John Merriam and colleagues at the University of California, Los Angeles, and the clones are made available to all researchers [Mount, see app. A].

Work by Michael Ashburner and co-investigators at Cambridge University on a comprehensive map of overlapping cosmid clones of the *D. melanogaster* genome was approved for funding by the European Economic Community in late 1987. This project is expected to follow the fingerprinting strategy of the nematode project, with the important difference that cytological maps (maps of banding patterns derived from microscopic analysis of stained chromosomes) of *D. melanogaster* chromosomes will be exploited. First, the technique of microdissection cloning (whereby DNA is excised from precise regions of the salivary gland polytene chromosomes and cloned) will be used to generate region-specific genomic clones. These microdissection clones are not of sufficient quality to be used directly, but they can be used to correlate cosmids in a standard genomic library with specific chromosomal regions. This step makes it easier to assemble the contiguous clones into groups. Finally, the position of all contigs with respect to the cytological map will be confirmed by in situ hybridization, whereby cosmid clones from the various contigs would be hybridized to salivary gland chromosomes [Mount, see app. A].

Strategies for Physical Mapping of the Human Genome

It is likely that making contig maps of large genomes, such as the human genome, will require a combination of bottom-up mapping and top-down mapping (55). Bottom-up mapping starts by making genomic clones, then fragmenting these clones further to decipher the overlaps necessary for connecting clones into contigs. Top-down mapping (e.g., Smith's E. coli map) is of lower resolution because it is derived from minimal fragmentation of source DNA. The critical distinction between the two methods is the size of the genomic DNA fragments used. Bottom-up mapping starts with relatively small genomic clones, while top down mapping starts with large genomic clones. The advantage of top-down mapping is that it offers more continuity (fewer gaps), while the bottom-up method has higher resolution (more detail). In formulating strategies for mapping the human genome, it will be necessary to decide what level of molecular detail is necessary to begin a human genome mapping project. Will information-rich strategies like those used to develop high-resolution E. coli restriction enzyme maps (20,42) or the DNA signposts offered by a RFLP map be the best first-generation human genome maps?

Contig Mapping

Scientists in the fields of molecular biology and human genetics who reviewed an OTA contract report on possible strategies for making contig maps of the human genome [Myers, see app. A] favored the following strategy: to map the genome one chromosome at a time, dividing and subdividing each chromosome into smaller and smaller segments before beginning restriction enzyme mapping and ordering of clones. After subdivision, restriction maps of these smaller segments would be determined and the information linked together to form continuous maps of whole chromosomes. In principle, this strategy could be broken down into five consecutive steps:

1. isolation of each human chromosome,
2. division of each chromosome into a collection of overlapping DNA fragments 0,5 to 5 million base pairs in length,
3. subdivision and isolation of each of these chromosomal fragments into overlapping DNA fragments about 40,000 base pairs in length,
4 determination of the order of the 40,000-base-pair DNA fragments as they appear in the chromosomes and determination of the positions of cutting sites for a restriction enzyme within each of these fragments, and
5. use of the mapping information gained in step 4 to link together each of the overlapping o.5 to 5-million-base-pair fragments isolated in step 2 [Myers, see app. A].

The substantial progress made so far on contig maps of nonhuman genomes implies that technologies already exist to begin construction of a global physical map of the human genome. The haploid human genome (approximately 3 billion base pairs) is at least 30 times larger than that of the nematode, the largest genome for which comprehensive physical mapping has been attempted. Sulston predicted that the mapping work he and his co-workers have done over the past 4 years could be repeated within 2 person-years, because much of their time was spent devising computer methods for data analysis (17). If the size of a genome were linearly related to the time required to physically map it, then the human ge - nome could be mapped to the same degree of completion as the nematode genome (90percent) in about 60 person-years. Such calculations are simplistic, however, because features of the human genome other than its size make it potentially more difficult to map. For example, some DNA sequences are repeated frequently throughout the human genome, in contrast to the nematode genome, and these are likely to interfere with the physical mapping process.

Techniques for isolating large chromosomal fragments should offer solutions to some of the physical mapping problems expected to arise from the occurrence of repetitive sequences in the human genome. The two most promising methods developed to date are the PFGE technology (8)9)13,61) and the yeast artificial chromosome cloning technology (6).

A National Research Council advisory panel on mapping and sequencing the human genome recommended improvements in technologies for the following to facilitate the construction of physical maps of large genomes:

- separating intact human chromosomes;
- separating and immortalizing identified fragments of human chromosomes;
- cloning the cDNAs representing expressed genes, especially those that represent rare cell-, tissue-, and development-specific mRNAs;
- cloning very large DNA fragments;
- purifying very large DNA fragments, including higher-resolution methods for separating such fragments;
- ordering the adjacent DNA fragments in a DNA clone collection; and
- automating the various steps in DNA mapping, including DNA purification and hybridization analysis, and developing novel methods that allow simultaneous handling of many DNA samples (52).

DNA Sequencing

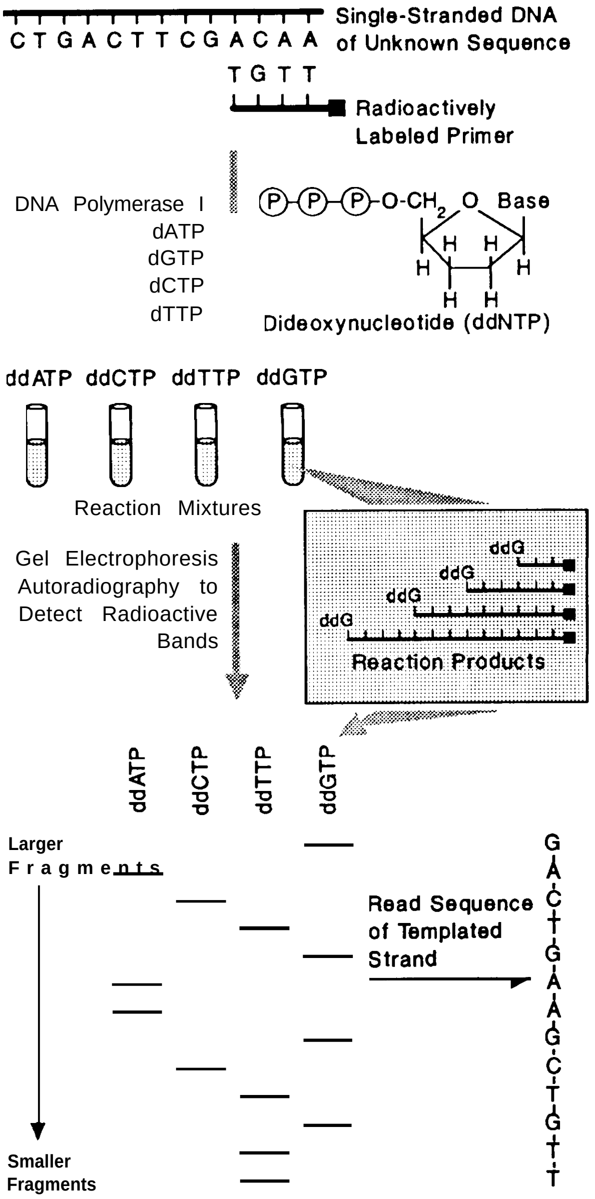
Strategies for sequencing the entire human genome are much more controversial than those for generating contig maps. Some scientists favor sequencing only expressed genes, identified with a cDNA map (17). Others propose that sequencing should continue to be targeted at specific regions of interest, as is currently done. Still others hold the view that the whole genome should be sequenced because it could reveal sequences with important functions that would otherwise go unidentified (see ch. 3). The National Research Council panel proposed first that pilot programs be conducted with a goal of sequencing approximately 1 millon continuous nucleotides (which is about five times as large as the largest continuous stretch of DNA sequenced to date) (52). Second, improvements in existing DNA sequencing technologies would be vigorously encouraged. Finally, extensive sequencing of other genomes, including the mouse, fruit fly, nematode, yeast, and bacterial genomes, was recommended for purposes of comparison (52).

The potential uses of human genome maps and sequences will likely dominate strategic decisions on which of the possible methods should be used to construct them (ch. 3). The strategy currently favored–preparing physical maps of individual chromosomes—requires that decisions be made on which chromosomes should be mapped first. Mapping smaller chromosomes first in pilot projects (e.g., chromosomes 21 and 22) would be the logical strategy from a technical perspective. Alternatively, selecting chromosomes linked to the largest numbers of markers for human genetic diseases (e.g., chromosome 7 and the X chromosome) might make the impact of genome mapping on clinical medicine more immediate. Efforts are already underway in a number of U.S. and foreign laboratories (ch. 8) to physically map (at various levels of resolution) human chromosomes known to be of general clinical significance or to carry genes of specific interest to the researchers involved. Scientists at Los Alamos and Livermore National Laboratories have begun mapping chromosomes 16 and 19, respectively. These chromosomes were chosen for their relatively small sizes and number of clinically relevant genetic markers. Researchers at Columbia University have begun work on a physical map of chromosome 21 for similar reasons.

DNA Sequencing Technologies

Two methods for sequencing DNA are standard in laboratories today. One technique, developed by Fred Sanger and Alan Coulson at the Medical Research Council in England (60), uses enzymes (figure 2-13), while the other, developed by Alan Maxam and Walter Gilbert at Harvard University, involves chemicals that degrade DNA (figure 2-14) (48,49). The two methods differ in the means by which the DNA fragments are produced; they are similar in that sets of radioactively labeled DNA fragments, all with a common origin but terminating in a different nucleotide, are produced in the DNA sequencing reactions.

George Church at Harvard Biological Laboratories has adapted the Maxam and Gilbert DNA sequencing method in an innovative technology, Figure 2.1 3.— DNA Sequencing by the Sanger Method


Single-Stranded DNA of Unknown Sequence
C T G A C T T C G A C A A
T G T T
Radioactively Labeled Primer
DNA Polymerase I
dATP
dGTP
dCTP
dTTP
Ⓟ-Ⓟ-Ⓟ-O-CH2 O Base
H H
H H
H H
Dideoxynucleotide (ddNTP)
ddATP ddCTP ddTTP ddGTP
Reaction Mixtures
Gel Electrophoresis Autoradiography to Detect Radioactive Bands
ddG
ddG
ddG
ddG
Reaction Products
ddATP
ddCTP
ddTTP
ddGTP
Larger Fragments
Smaller Fragments
Read Sequence of Templated Strand
G A C T G A A G C T G T T

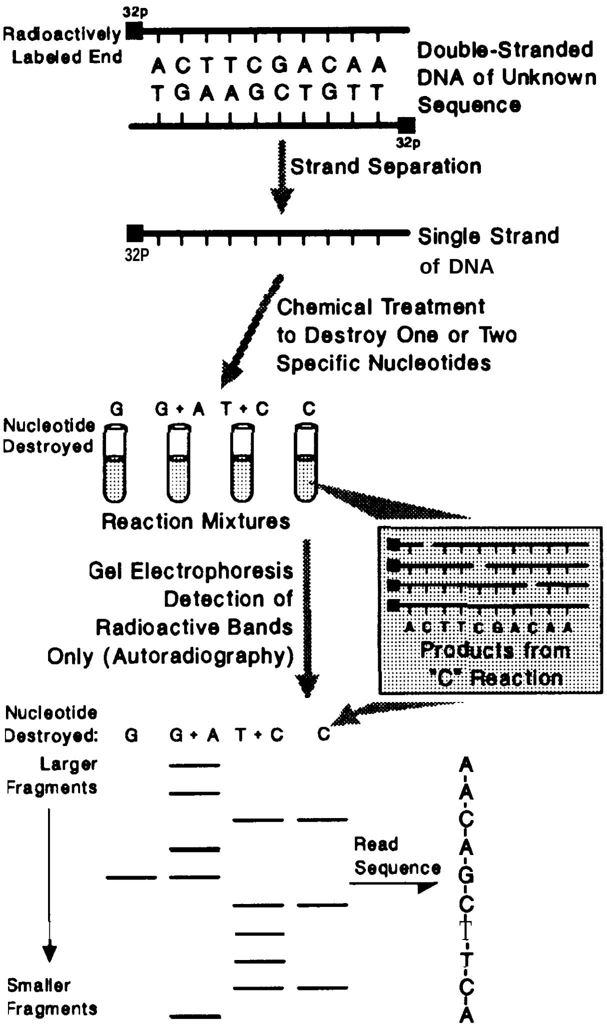
In the Sanger method, a cloned DNA fragment is mixed with a short piece of synthetic DNA complementary to only one end (the origin) of the cloned fragment. An enzyme called DNA polymerase is then used to catalyze the synthesis of a complementary strand. During the polymerization reaction, a modified nucleotide, a dideoxynucleotide, is included with a mixture of the four naturally occurring nucleotides (A, G, T, and C), one of which is labeled with a radioactive phosphorous or sulfur atom, causing growth of the DNA chain to stop whenever the modified nucieotide is inserted. Four separate reactions, each containing all four normal nucleotides but a different dideoxynucleotide, can be carried out. A series of radioactively labeled DNA strands will be made, the lengths of which depend on the distance from the origin to the nucleotide position where the chain was terminated. For example, if a short DNA template has four G's, conditions are set up such that some molecules will be made with no G dideoxynucleotide analogs, some will terminate at the fourth G position, some at the third G position, and so on. Similarly, the other three dideoxynucleotides will insert infrequently and randomly at the appropriate positions in the other three nucleotide-specific reactions. The series of labeled DNA strands is subsequently analyzed by polyacrylamide gel electrophoresis. Radioactively labeled DNA is electrophoresed through a vertical slab of polyacrylamide gel (polyacrylamide is a polymeric resin in which DNA molecules from 1 to 400 bases long can be separated from one another), an X-ray film is then placed over the gel and exposed, and the resulting autoradiograph shows a ladderlike pattern of bands. The sequencing reactions corresponding to each of the four different bases are run as four adjacent lanes on the polyacrylamide gel, and the resulting ladders of bands are read alternately to give the sequence of the DNA.

SOURCE: Office of Technology Assessment, 1988.

called multiplex sequencing, that enables a researcher to analyze a large set of cloned DNA fragments as a mixture throughout most of the DNA sequencing steps. Mixtures of clones are operated on in the same way as a single sample in traditional sequencing. This is accomplished by tagging each DNA clone in the mixture with short, unique sequences of DNA in the first step and then deciphering the nucleotide sequence of each cloned fragment in the final step. Multiplex sequencing allows the simultaneous analysis of about 40 clones on a single DNA sequencing gel, increasing the efficiency of the standard procedure by more than a factor of 10 (14). Church and coworkers have been applying the multiplex sequencing strategy to determine the complete nucleotide sequences of two species of bacteria, *E. coli* and *Salmonella* typhimurium (14).

The major problem with current DNA sequencing technology is the large number of DNA sequences that remains to be determined. Multiplex is only one of several new sequencing protocols that could be of great value to large genome sequencing projects. Church and Gilbert devised a method related to multiplex sequencing that aI1ows sequencing directly from genomic DNA (15). Another method, developed by researchers at Cetus (Emeryville, CA), involves the selective amplification of specific DNA sequences without prior

Figure 2-14.–DNA Sequencing by the Maxam and Gilbert Method


32p
Radioactively Labeled End
A C T T C G A C A A
T G A A G C T G T T
Double-Stranded DNA of Unknown Sequence
32p
Strand Separation
32P
Single Strand of DNA
Chemical Treatment to Destroy One or Two Specific Nucleotides
Nucleotide Destroyed
G
G + A
T + C
C
Reaction Mixtures
Gel Electrophoresis Detection of Radioactive Bands Only (Autoradiography)
A C T T C G A C A A
Products from "C" Reaction
Nucleotide Destroyed: G G + A T + C C
Larger Fragments
Smaller Fragments
Read Sequence
A-A-C-A-G-C-T-T-C-A

In the Maxam and Gilbert procedure, chemical reactions specific to each of the four bases are used to modify DNA fragments at carefully controlled frequencies. One end of one strand in a double-stranded DNA fragment is radioactively labeled, and the labeled DNA is used in each of four separate react ions and treated with a chemical that specifically nicks one or two of the four bases in the DNA. When these DNA molecules are treated with another chemical, the DNA fragments are broken where the base was nicked and are destroyed. Just as in the Sanger sequencing method, the products of the Maxam and Gilbert sequencing procedure are fragments of varying lengths, each ending at the G, C, T, or A where the chemical reaction took place. By limiting the amount of chemicals used in each of the base-specific reactions so they will react only a few times per molecule, it is possible to obtain all possible double-stranded DNA fragments equal in length to the distance from the radioactively labeled origin to each of the bases. For any given DNA fragment sequenced, each of the four reactions is eiectrophoresed separately, as described in figure 2-13, and the sequencing patterns determined from the autoradiograph.

SOURCE: Office of Technology Assessment, 1988.

cloning (59). Each of these methods could potentially eliminate the steps of cloning and DNA preparation in sequence analysis (41).

Finally, DNA sequencing methods that do not involve either gel electrophoresis or chemical or enzymatic reactions have also been proposed. At the Los Alamos National Laboratory, researchers are investigating ways to use enhanced fluorescence detection methods in flow cytometry as an alternative to gel techniques for DNA sequencing. others have suggested scanning tunneling electron microscopes to read bases directly on a strand of DNA (57,62).

AUTOMATION AND ROBOTICS IN MAPPING AND SEQUENCING

The longest single stretch of DNA sequence determined to date, the genome of the Epstein-Barr virus, contains fewer than 200,000 base pairs. The total number of nucleotides sequenced to date using both chemical and enzymatic sequencing technologies is about 16 million base pairs [Computer Horizons, Inc., see app. A]. This is the current size of GenBank®, the U.S. repository of DNA sequence data [app. D]. Since GenBank” includes only reported data, 16 million base pairs represents a low estimate of the total number of base pairs sequenced. Reported DNA sequences range from those of small viruses to those of animals and plants (table 2-3). So far, less than one-tenth of 1 percent (1.9 million base pairs) of the nearly 3 billion base pairs in the haploid human genome has been sequenced and reported (7). The current DNA sequencing rate is estimated to generate only about 2 million base pairs per year of sequence information (7), a powerful incentive for

Table 2-3.—Amount of Genome Sequenced in Several Well-Studied Organisms

Organism	Genome size (base pairs)	Percent sequenced
Escherichia co/i (bacterium) . .	4.7 million	16
Saccharomyces cerevisiae (yeast)	15 million	4
Caenorhabditis elegan (nematode)	80 million	.06
Drosophila melanogaster (fruit fly)	155 million	.26
A4us musculs (mouse)	3 billion	.04
Homo sapiens (human)	2.8 billion	.08

SOURCES:
C. Burks, GenBank®, Los Alamos National Laboratory, Los Alamos, NM, personal communication, March 1988.
GenBank" Release No. 54, December 1987.

devising methods of automating the procedures involved in preparing for and carrying out DNA sequencing. Some recent reviews (16,38,41,47,57) provide detailed accounts of the robotic and automated systems currently available and describe the kinds of systems being developed or planned for genome mapping and sequencing.

Any degree of automation will help lower the overall costs of genome projects, both in time and in dollars. The primary objective in the use of automation is standardization, driven by the need for repetitive, highly accurate determinations (41). Some of the existing automated devices are designed for repetitive DNA cloning steps, such as the preparation and restriction enzyme cutting of cloned DNA samples. Similarly, efforts are being made to automate the pouring, loading, and running of gels for separating DNA and for sequencing DNA. Many of the steps in physical mapping could be adapted to automation. Cloning procedures, DNA probe synthesis, and DNA hybridizations are only a few of those being explored for application to genome projects. A system that automates some steps in growing DNA clones, to be used, for example, as gene probes or for DNA sequencing, was recently introduced by Perkin-Elmer Cetus Instruments (Norwalk, CT) (67).

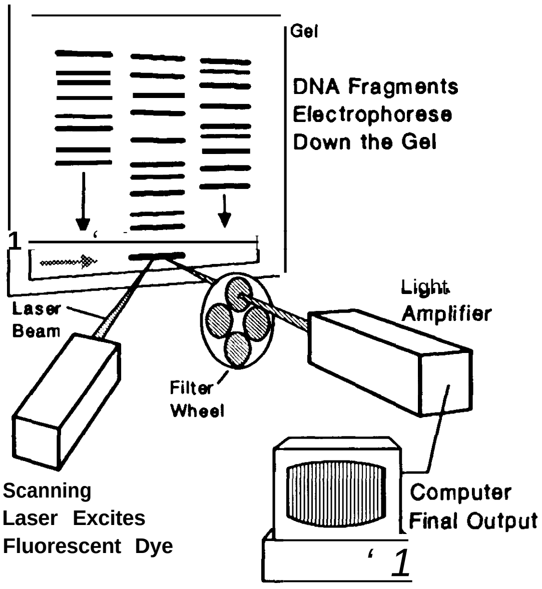
The area of automation that has received the most attention is DNA sequencing. An international workshop on automation of DNA sequencing technologies was held in 1987 in okayama, Japan, and the proceedings give an extensive account of the state of the art from an international perspective (32). There are five steps in the task of DNA sequence analysis:

1. cloning or otherwise isolating the DNA,
2. preparing the DNA for sequence analysis,
3. performing the chemical (Maxam and Gilbert) or enzymatic (Sanger) sequencing reactions,
4. running the sequencing gels, and
5. reading the DNA sequence from the gel.

Steps 3 through 5 are the functions most often performed by the instruments developed as of early **1988 (16,57, 70)**; however, none of the companies involved has yet commercialized an integrated system that performs all of the functions.

In 1986, Applied Biosystems, Inc. (Foster City, CA) introduced the first commercial, automated DNA sequencer (16). This instrument was made using technology developed by Leroy Hood and co-workers at the California Institute of Technology. This and similar machines perform steps 4 and 5. The Applied Biosystems, Inc. system is based on the Sanger sequencing reaction, with modifications to use different fluorescent dyes instead of radioactive chemicals to label the primers. Because the sequencing reaction primers are individually labeled with different dyes, each of the four enzymatic reactions can be run together in a single lane on the polyacrylamide gel. A laser activates the dyes, and fluorescence detectors read the DNA sequence at the bottom of the gel as each fragment appears. The sequence is determined directly by a computer (figure 2-15). E.I. du Pent de Nemours & Co. (Wilmington, DE) introduced in 1987 an automated system that slightly modifies the technology used by Hood and Applied Biosystems, Inc.; this system can potentially reduce the number of artifacts read by the fluorescence detectors. Hitachi, Ltd. (Tokyo, Japan) is also expected to market an instrument that automates steps 4 and 5, but it too is based on the fluorescence technology developed by Hood and colleagues. In early 1988, another U.S. company, EE&G Bimolecular (Wellesley, MA), began marketing a machine that automates the same DNA sequencing and gel-reading methods used manually in most laboratories. Bio-Rad Laboratories (Richmond, CA) marketed an instrument that

Figure 2“15.—Automated DNA Sequencing Using Fluorescently Labeled DNA


Gel
DNA Fragments Electrophorese Down the Gel
1
Laser Beam
Light Amplifier
Filter Wheel
Scanning Laser Excites Fluorescent Dye
Computer Final Output
‘ 1

SOURCE: Leroy Hood, California Institute of Technology, Pasadena, CA.

scans autoradiographs of DNA sequencing gels and analyzes the data.

Most of these DNA sequencing systems are based on manual enzymatic sequencing reactions, while the gel running and reading are automated. Only one commercial enterprise, Seiko of Japan, has reported automating the chemical or enzymatic steps (step 3) in the DNA sequencing protocol (7o). In addition, the University of Manchester Institute of Science (Manchester, England) has built an automatic reagent manipulating system to carry out the Sanger sequencing reactions (47).

Robotics are used to give automation flexibility, to extend its capabilities to complex operations typically performed by highly skilled laboratory workers. Conceivably, laboratory robots would allow programmable devices to do physical work as well as to process data (41). Several robotic devices have been designed and used successfully by companies involved in the commercialization of recombinant DNA products or processes. Genetics Institute’s (Cambridge, MA) Autoprep” Plas - mid Isolation System provides small quantities of plasmid DNA and vector DNA for DNA sequencing (22). Researchers at the same company also developed a robot to purify and isolate synthetic oligonucleotides for use as probes in cloning and DNA sequencing (38).

Technical advances are occurring rapidly and simultaneously in biology, robotics, and computer science, so it is difficult to predict what the future will bring in the development of automated technology. Some yet-to-bedeveloped technology could make many of the current physical mapping procedures obsolete.

CHAPTER *2* REFERENCES

1. Ayala, F.J., “Two Frontiers of Human Biology: What the Sequence Won’t Tell Us,” Issues *in Science and Techno]gy,* (Spring) :51-56, 1987.
2. Bernard, H. U., and Helinski, D. R., “Bacterial Plasmid Cloning Vehicles, ” in *Genetic Engineering, VOI. 2,* J,K. Setlow and A. Hollaender, eds., (New York: Plenum Press, 1980).
3.Blattner, F.,University of Wisconsin, Madison, personal communication, October 1987.
4. Bolivar, F., and Backman, K., “Plasmids of *E. Coli* as Cloning Vectors, ” *Methods in Enzymolog 68:245-267, 1979.*
5. Botstein, D., White, R. L., Skolnick, M., et al., “Construction of a Gen,etic Linkage Map in Man Using Restriction Fragment Length Polymorphisms,” *American Journal of Human Genetics 32:314-331,* 1980.
6. Burke, D.T., Carle, G. F., and Olson, M. V., “Cloning of Large Segments of Exogenous DNA Into Yeast Using Artificial-Chromosome Vectors)” *Science 236:806-812, 1987.*
7. Burks, C., Los Alamos National Laboratory, Los Alamos, NM, personal communication, February 1988.
8. Carle, G. F., Frank, F., and Olson, M. V., “Electrophoretic Separations of Large DNA Molecules by Periodic Inversion of the Electric Field,” *Science* 232:65-68, 1986.
9 Carle, G. F., and Olson, M. V., “An Electrophoretic Karyotype for Yeast ,“ *Proceedings of the National Academy of Sciences USA 82:3756-3760, 1985.*
10 Caspersson, T., Lomakka, C., and Zech, L., “Fluores-

cent Banding," *Hereditas 67:89-102, 1971.*

11 Caspersson, T., Zech, L., and Johansson, C., "Differential Banding of Alkylating Fluorochromes in Human Chromosomes, " *Experimental Cell Research 60:315-319, 1970.*

12. Caspersson, T., Zech, L., Johansson, C., et al., "Quinocrine Mustard Fluorescent Banding," *Chromosome 30:215-227, 1970.*

13. Chu, G., Vollrath, D., and Davis, R. W., "Separation of Large DNA Molecules by Contour-Clamped Homogeneous Electric Fields," Science 234: 1582-1585, 1986.

14. Church, G. M., "Genome Sequence Comparisons)" grant proposal, Feb. 10, 1987.

15. Church, G. M., and Gilbert, W., "Genomic Sequencing, " *Proceedings of the National Academy of Sciences USA 81:1991-1995, 1984.*

16. Connell, C., Fung, C., Heiner, J., et al., "Automated DNA Sequence Analysis, "BioTechniques 5:342-348, 1987.

17. Costs of Human Genome Projects, OTA, workshop, Aug. 7, 1987.

18. Coulson, A., Sulston, J., Brenner, S., et al., "Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans)" Proceedings of the National Academuv of Sciences USA 83:7821-7825, 1986.*

19. Crick, F. H. C., and Watson, J. D., "The Complementary Structure of Deoxyribonucleic Acid," *Proceedings of the Royal Society(A) 223:80-96, 1954.*

20. Daniels, D. L., and Blattner, F. R., "Mapping Using Gene Encyclopedias, " *Nature 325:831-832, 1987.*

21. Deaven, L. L., Van Dilla, M. A., Bartholdi, M. F., et al., "Construction of Human Chromosome-Specific DNA Libraries Fran. Flow. Sorted Chromosomes," *Cold Spring Harbor Symposia on Quantitative Biology* 51:159-167, 1986.

22. DeBonville, D.A., and Riedle, G. E., "A Robotic Workstation for the Isolation of Recombinant DNA, " in *Advances in Laboratory Automation-Robotics, vol. 3,* p. 353.

23. Donahue, R. P., Bias, W.B., Renwick, J. H., et al., "Probable Assignment of the Duffy Blood Group Locus to Chromosome 1 in Man," *Proceedings of the National Academy of Sciences USA 61:949-955, 1968.*

24. Donis-Keller, H., Green, P., Helms, C., et al., "A Genetic Linkage Map of the Human Genome, " Cell 51:319-337, 1987.

25. Evans, G. A., and Wahl, G. M., "Cosmid Vectors for Genomic Walking and Restriction Mapping," in *Methods in Enzymology: A Guide to Molecular Cloning,* vol. 152, (in press).

26. Foley, B., Nelson, D., Smith, M.T., et al., "Cross-Sections of the Genbank Database, " *Trends in Genetics 2:233-236, 1986.*

27. Gall, J.G., letter to the editor, *Science* **233:1367-1368, 1986.**

28. Gerhard, D-S., Kawasaki, E. S., Bancroft, F.C., et al., "Localization of a Unique Gene by Direct Hybridization," *Proceedings of the National Academy of Sciences USA 78:3755-3759, 1981.*

29, Gray, J. W., Dean, P. N., Fuscoe, J. C., et al., "High-Speed Chromosome Sorting, "Science 238:323-329, 1987,

30. Gray, J. W., Langlois, R. G., Carrano, A. V., et al., "High Resolution Chromosome Analysis: One and Two Parameter Flow Cytometry, " *Chromosome 73:9-27, 1979.*

31. Gusella, J. F., Wexler, N. S., Conneally, P. M., et al., "A Polymorphic DNA Marker Genetically Linked to Huntington's *Disease, "Nature 306:234-238, 1983.*

32. Hayashibara International Workshop on Automatic and High Speed DNA-Base Sequencing, Hayashibara Biochemical Laboratory, Okayama, Japan, July 7-9, 1987.

33, Hendrix, R. W., Roberts, J, W., Stahl, F. W., et al., *Lambda H* (Cold Spring Harbor, NY: Cold Spring Harbor Press, 1982).

34. Hohn, B., and Collins, J., "A Small Cosmid for Efficient Cloning of Large DNA Fragments, " *Gene* 11:291-298, 1980.

35. "Human Gene Mapping 8," *Cytogenetics and Cell Genetics 40:1-4, 1985.*

36. Ish-Horowicz, D., and Burke, J. F., "Rapid and Efficient Cosmid Cloning, " *Nucleic Acids Research 9:2989-2998, 1981.*

37. Jeffries, A.J., "DNA Sequence Variants in the @-&Globin Genes of Man)" Cell 1:1-10, 1979.

38. Jones, S. S., Brown, J. E., Vanstone, D. A., et al., "Automating the Purification and Isolation of Synthetic DNA)" *Biotechnology 5:67-70, 1987.*

39. Jikgens, G. E., Wieschaus, C., Nikslein-Volhard, C., et al., "Mutations Affecting the Pattern of the Larval Cuticle in *Drosophila melanogaster* II: Zygotic Loci on the Third Chromosomes, " *Rouxk Archive of Developmental Biolo~* 193d:283-295.

40. Kan, Y.W., and Dozy, A.M., "Polymorphism of DNA Sequence Adjacent to Human Beta-Globin Structural Gene: Relationship of Sickle Mutation, " *Proceedings of the National AcademuV of Sciences USA 75:5631-5635, 1978.*

41, Knobeloch, D. W., Hildebrand, C. E., Moyzis, R. K., et al., "Robotics in the Human Genome Project, " *Biotechnology 5:1284-1287, 1987.*

42. Kohara, Y., Akiyama, K., and Isono, K., "The Physical Map of the Whole *E. Coli* Chromosome: Application of a New Strategy for Rapid Analysis and

Sorting of a Large Genomic Library,” Cell 50:495-508, 1987.

43. Lander, E. S., and Botstein, D., “Mapping Complex Genetic Traits in Humans: New Strategies Using a Complete RFLP Linkage Map, ” Cold *Spring Harbor Symposia on Quantitative Biology 51:49-62, 1986.*

44. Lander, E. S., and Botstein, D., ‘(Strategies for Studying Heterogeneous Genetic Traits in Humans by Using a Linkage Map of Restriction Fragment Length Polymorphisms,” *Proceedings of the National Academy of Sciences USA 83:7353-7357, 1986.*

45. Lange, K., and Boehnke, M., “How Many Polymorphic Genes Will It Take To Span the Human Genome?” *American Journal of Human Genetics 34:842-845, 1982.*

46. Lebo, R. V., Anderson, L. A., Lau, Y.-F. C., et al., “Flow-Sorting Analysis of Normal and Abnormal Human Genomes,” *Cold Spring Harbor Symposia on Quantitative Biology 51:169-176, 1986.*

47. Martin, W.J., and Davies, W .R., “Automated DNA Sequencing: Progress and Prospects, ''BioTechnol0~ 4:890-895, 1986.

48. Maxam, A. M., and Gilbert, *W.*, ‘(A New Method for Sequencing DNA,” *Proceedings of the National Academy of Sciences USA 74:560-564, 1977.*

49. Maxam, A.M., and Gilbert, W., “Sequencing End-Labeled DNA with Base-Specific Chemical Cleavage,” *Methods in Enzymo]oo 65:499-560, 1980.*

50. McKusick, V. A., “The Morbid Anatomy of the Human Genome: A Review of Gene Mapping in Clinical Medicine,” *Medicine 65:1-33, 1986.*

51. McKusick, V. A., and Ruddle, F. H., “Toward a Complete Map of the Human Genome,” *Genomics 1:103-106, 1987.*

52. National Research Council, *Mapping and Sequencing the Human Genome,* (Washington, DC: National Academy Press, 1988.)

53. Nusslein-Volhard, C., Wieschaus, E., and Kluding, H., “Mutations Affecting the Pattern of Larval Cuticle in *Drosophila melanogaster* I: Zygotic Loci on the Second Chromosome,” *Roux’s Archives of Developmental Biology 193:267-282, 1984.*

54. Ohno, S., “An Argument for the Genetic Simplicity of Man and Other Mammals, ” *Journal of Human Evolution 1:651-662, 1972.*

55. Olson, M. V., Dutchik, J. E., and Graham, M.Y., “Random-Clone Strategy for Genomic Restriction Mapping in Yeast,” *Proceedings of the National Academy of Sciences USA 83:7826-7830, 1986.*

56. Pardue, M. L., and Gall, J. G., “Chromosomal Location of Mouse Satellite DNA,” *Science* **168:1356-1358, 1970.**

57, Rotman, D., “Sequencing the Entire Human Genome,” *Industrial Chemist* (December) :18-21, 1987.

58. Ruddle, F., Bentley, K. L., and Ferguson-Smith, A., “Physical Mapping Review,” contract report to the Office of Technology Assessment, 1987.

59, Saiki, R. K., Sharf, S., Faloona, F., et al., *Science* 230: 1350-1354, 1985.

60, Sanger, F., Nilken, S., and Coulson, A. R., “DNA Sequencing With Chain-Terminating Inhibitors,” *Proceedings of the National Academy of Sciences USA 74:5463-5468, 1980.*

61. Schwartz, D. C., and Cantor, C. R., “Separation of Yeast Chromosome-Sized DNAs by Pulsed Field Gel Electrophoresis,” Cell 37:67-75, 1984.

62. Shera, B., Lawrence Livermore National Laboratory, Livermore, CA, personal communication, February 1988.

63. Smith, C.L., Econome, J.G., Schutt, S., et al., “A Physical Map of the *Escherichia coli* Genome, ” *Science 236:1448-1453, 1987.*

64. Solomon, E., and Bodmer, W. F., “Evolution of Sickle Variant Gene,” *The Lancet* April 28, 1979, p.923.

65. Southern, E. M., “Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis,''*Journalof Molecular Biology 98:503-517, 1975.*

66. Staden, R., “A New Method for Storage and Manipulation of DNA Gel Reading Data,” *Nucleic Acids Research 8:3673-3694, 1980.*

67. Stinson, S., “System Automates DNA Amplification,” *Chemical and Engineering News,* Dec. 21, 1987, p. 24.

68. U.S. Congress, Office of Technology Assessment, New *Developments in Biotechnology, 4: U.S. Investment in Biotechnology* (Washington, DC: U .S. Government Printing office, in press),

69. Vogel, F., and Motulsky, A. G., *Human Genetics: Problems and Approaches (New* York: Springer-Verlag, 1986), pp. 369-370.

70. Wada, A., “Automated High-Speed DNA Sequencing, ” *Nature 325:771-772, 1987,*

71. Waterson, R., Medical Research Council, Cambridge, England, personal communication, October 1987.

72. Watson, J. D., Hopkins, N. H,, and Roberts, J. W., *Molecular Biology of the Gene* [Menlo Park: The Benjamin/Cummings Publishing Co., 1987).

73, Watson, J.D., and Crick, F. H. C., “Genetic Implications of the Structure of Deoxyribonucleic Acid, ” *Nature 171:964-967, 1953.*

74. Watson, J.D., and Crick, F. H. C., “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature 171:737-738, 1953.*

75. Weiss, M., and Green,, H., “Human-Mouse Hvbrid

Cell Lines Containing Partial Complements of Human Chromosomes and Functioning Human Genes," *Proceedings of the National Academy of Sciences USA 58:1104-1111, 1967.*

76. White, R., and Lalouel, J.-M., "Chromosome Mapping With DNA Markers," *Scientific American 258:40-48, 1988.*

77. White, R., Lippert, M., Bishop, D. T., et al., "Construction of Linkage Maps with DNA Markers for Human Chromosomes, ''Nature 313:101-105, 1985.

78. Wieschaus, E. C., Niisslein-Volhard, C., and Jikgens, G., "Mutations Affecting the Pattern of the Larval Cuticle in *Drosophila Melanogaster* III: Zygotic Loci on the X-Chromosome and Fourth Chromosome, " *Roux's Archive of Developmental Biology 193:296-307, 1984.*

79. Williams, B. G., and Blattner, F. R., "Bacteriophage Lambda Vectors for DNA Cloning," in *Genetic Engineering,* vol. 2, J.K. Setlow and A. Hollaender (eds.), (New York, NY: Plenum Press, 1980).

80. Wilson, E. B., "The Sex Chromosomes, " *Arch. Mikrosk. Anat. Entw"cklungsmech. 77:249, 1911.*

81. Yunis, J.J., Sawyer, J. R., and Dunham, K., '(The Striking Resemblance of High-Resolution G-Banded Chromosomes of Man and Chimpanzee, " ***Science*** 208:1125-1148, 1980.