CHAPTER 2

Recombinant DNA Methods: Applications to Human Genetics

Helen Donis-Keller, PhD, and
David Botstein, PhD

The past decade has seen a revolution in molecular biology due mainly to the introduction of recombinant DNA methods. These methods have allowed the isolation and characterization of genes from any organism and the determination of the DNA sequence and any encoded protein sequences. It has become possible as well to follow genes through families, and indeed, over evolutionary changes. The ability to isolate and analyze human genes has had, if anything, an even more profound effect on the field of human genetics than on most areas of biology. The new technology has already made possible the understanding of many inherited diseases at the molecular level. Perhaps more important, recombinant DNA methods, for the first time, have permitted direct general application of mendelian ideas to human families, through the use of restriction fragment length polymorphisms (RFLPs).

The purpose of this chapter is to review briefly the technology in a way that makes for better understanding by non-specialists, rather than to introduce actual techniques. For the latter purpose, several extremely helpful technical manuals and other publications[1,2,3,4] are available for those interested in laboratory work; for linkage analysis at the technical level the reader is referred to Ott (1985).[5] In what follows we will describe very briefly principles underlying the use of restriction endonucleases, DNA sequence analysis, gel-transfer hybridization (Southern blotting), gene cloning, and library construction, as well as the principles of genetic mapping using RFLPs. The last-named subject will take us into a description of the statistical ideas and methodologies required to make and use a linkage map of the human genome.

ISOLATION AND CHARACTERIZATION OF GENES AND DNA FRAGMENTS

Restriction Enzymes Cleave Double-Stranded DNA

The era of recombinant DNA methods can be thought of as beginning with the discovery and purification of sequence-specific endonucleases. These are bacterial enzymes that recognize specific nucleotide sequences in DNA from all organisms and cleave the DNA within that sequence or very nearby it. They are called *restriction enzymes* because they apparently evolved in bacteria as defenses against the invasion of foreign DNAs in the form of viruses or plasmids. Each of the enzymes is associated with a modification activity (usually a methylase) that protects the bacterial cell's own DNA against cleavage: a cell carrying a specific restriction/modification system will modify the restriction sites in its own DNA, whereas foreign intruding DNA, being unmodified, will be cleaved upon entry.

Through the pioneering work of Nathans and Smith (1975),[6] it became clear that, whatever role these enzymes played in nature, they could be used as reagents to characterize the sequences of DNAs. Many enzymes with different sequence specificities have been found, some recognizing sequences only 4 base pairs (bp) in length (thus cutting DNA every 256 bp on average, ie, 4^4), others recognizing 6 bp sites (cutting on average every 4,096 bp, ie, 4^6). Recently, enzymes recognizing sites as large as 10 bp have been found. Convenient, up-to-date summaries of available enzymes are the catalogs of the manufacturers (eg, New England Biolabs). Since the enzymes read bits of DNA sequence, the number and arrangement of restriction sites is characteristic of a given DNA sequence: this is called its *restriction map.* One maps DNA molecules for three major purposes: first, as a "signature," since each sequence has a unique map; second, as an aid to manipulations, such as subcloning and RFLP analysis; third, as a prelude to determination of the nucleotide sequence.

The restriction map of a DNA molecule is deduced in practice from the pattern of fragment sizes produced after digestion with a number of different restriction enzymes, separately or in combination. The pattern can conveniently be analyzed by separating the fragments by electrophoresis through an agarose gel; the DNA is stained with ethidium bromide or visualized by hybridization in situ. In most systems of electrophoresis, fragment mobility is inversely proportional to the logarithm of molecular length, allowing easy comparison of lengths. Use of standards of known length allows routine determination of molecular length to a precision of about 5%. A large number of restriction enzymes are commercially available, and the apparatus for agarose gel electrophoresis and visualization of the fragments is simple to use.

Cloning DNA Sequences

The direct determination of the restriction map of a DNA sequence requires its isolation in bulk. For anything beyond small virus genomes, it is necessary to clone the sequences of interest in order to create restriction maps of them directly. There are many systems of cloning, and their basis is beyond the scope of this article. However, it should be borne in mind that DNA clones consist of a *vector*, that is, a DNA segment that contains means of replication and selection in bacteria, and an *insert*, that is, the cloned DNA that has been joined to the vector and replicated in bacteria with it. The means of replication may be simply a site ("origin of replication") recognized and used by the bacterial host DNA synthesis machinery, or else it may involve both a viral origin of replication and additional viral enzymes required for replication. The selection system is usually a gene conferring resistance to an antibiotic such as ampicillin or tetracycline. These elements, when attached to insert DNA, allow the selection for bacteria carrying the composite "clone." The clone will replicate and, by expressing drug resistance, be able to withstand the applied selection. There are three general vector types: (1) plasmid vectors, (2) viral vectors, and (3) cosmid vectors. *Plasmid vectors* are relatively small (2–5 kilobases [kb]) self-replicating circles of double-stranded DNA that accommodate all sizes of insert (0–40 kb) and are frequently used as vectors when large segments of DNA are subdivided or "subcloned." *Viral vectors* may be mammalian viruses such as SV40 or bacteriophages, for example, bacteriophage λ, that have been modified by removing regions of their genomes not essential for replication or packaging, and replacing this DNA with insert DNA (up to 20 kb). The cloning of human genomic DNA into such vectors is schematically represented in Figures 2.1 and 2.2. *Cosmids* are combination vectors with bacteriophage λ and plasmid elements that allow the efficient insertion and replication of very large segments of DNA (up to 40 kb). The advantages of cosmid systems are in the efficiency of cloning (packaging of cloned DNA in viral capsids): the end result is a population of plasmid clones with very large inserts.

The source of inserts can be the DNA of the target organism (in which case the clones are called *genomic* clones), or it can be a messenger RNA copied into DNA by reverse transcription, in which case the clones are called complementary DNA (*cDNA*) clones. Since most human genes are split by intervening sequences that are removed by the cell in the process of making the mRNA, a cDNA clone of very modest length (say, 1000 bp) may contain sequences spanning 200,000 bp in the human genome. A schematic representation of the cDNA cloning procedure is shown in Figure 2.3.

A major difference between cDNA and genomic clones is the presence in the latter of repeated sequences that abound in the genome, but are rarely

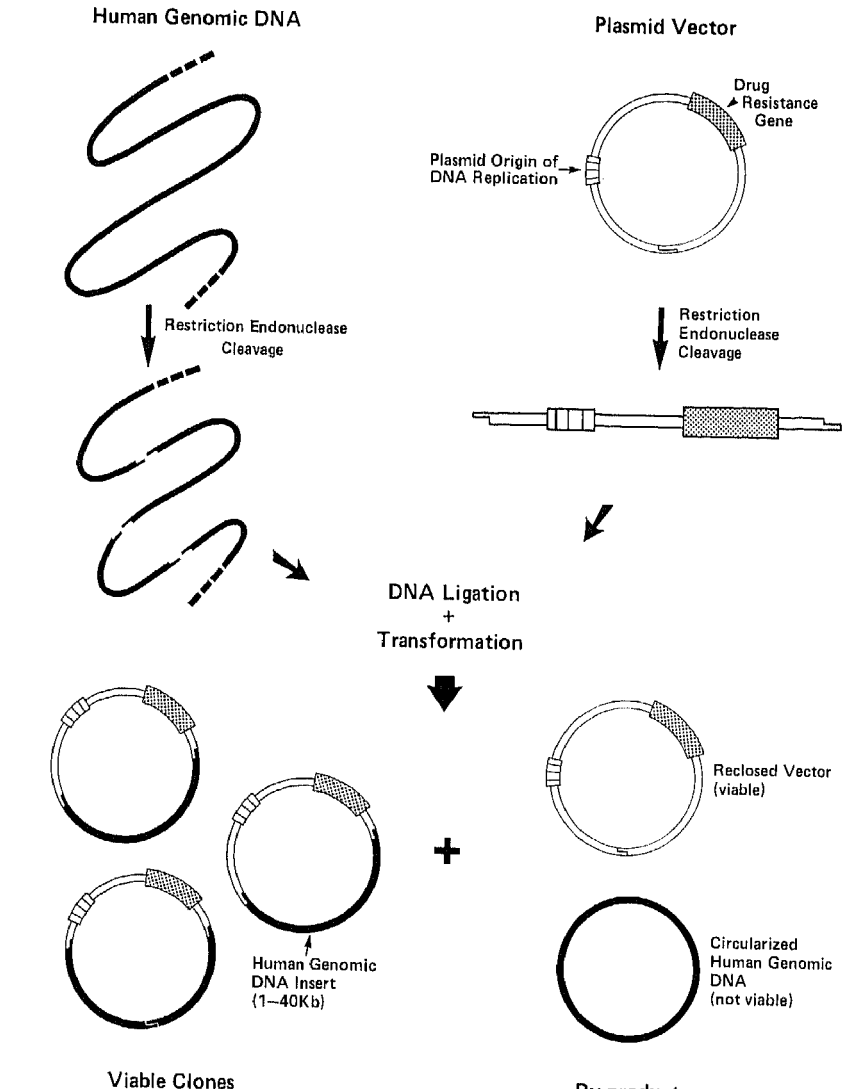


FIGURE 2.1 Cloning human DNA into a plasmid vector. Human genomic DNA is cleaved with a restriction endonuclease and inserted into a standard plasmid vector that contains a means of replication and a selectable marker. Various by-products of the procedure (closed vector and circularized human DNA) are also represented.

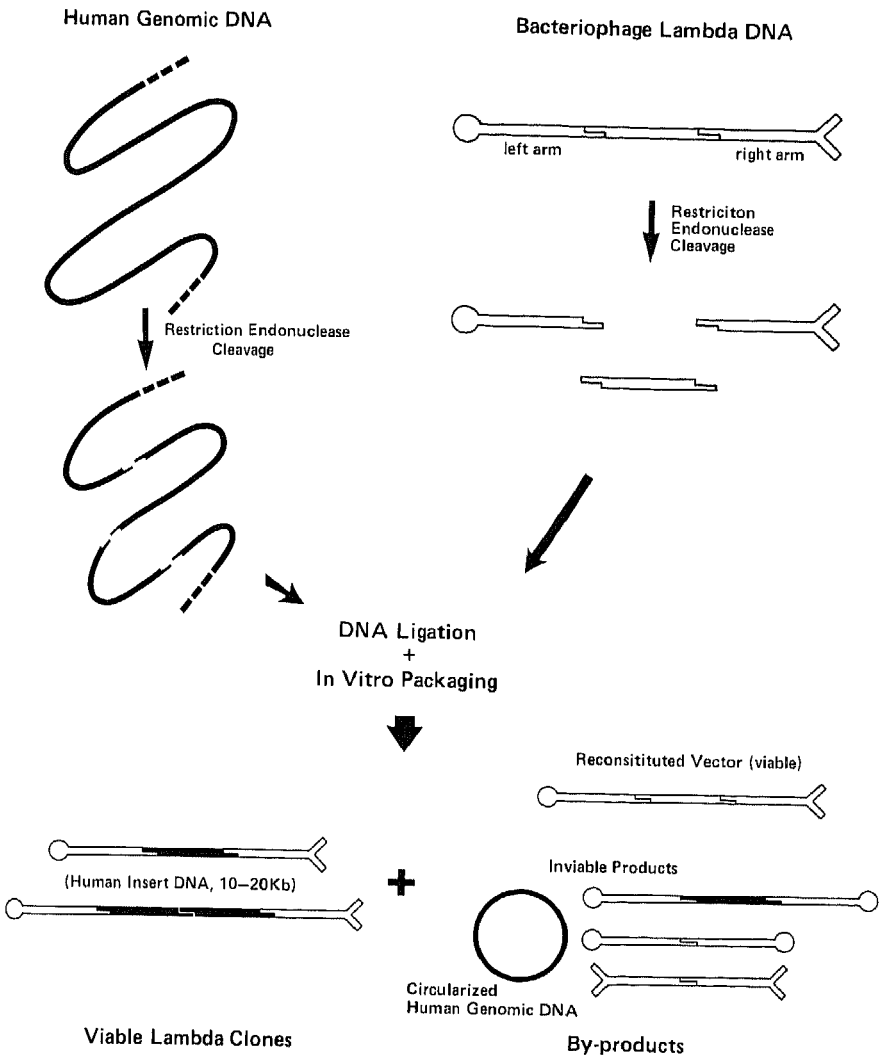


FIGURE 2.2 Cloning human DNA into a bacteriophage λ vector. Human genomic DNA is cleaved with a restriction endonuclease and inserted into a bacteriophage λ vector. The recombinant vector is then packaged into bacteriophage capsid particles in an in vitro reaction. Various by-products of the procedure are also pictured.

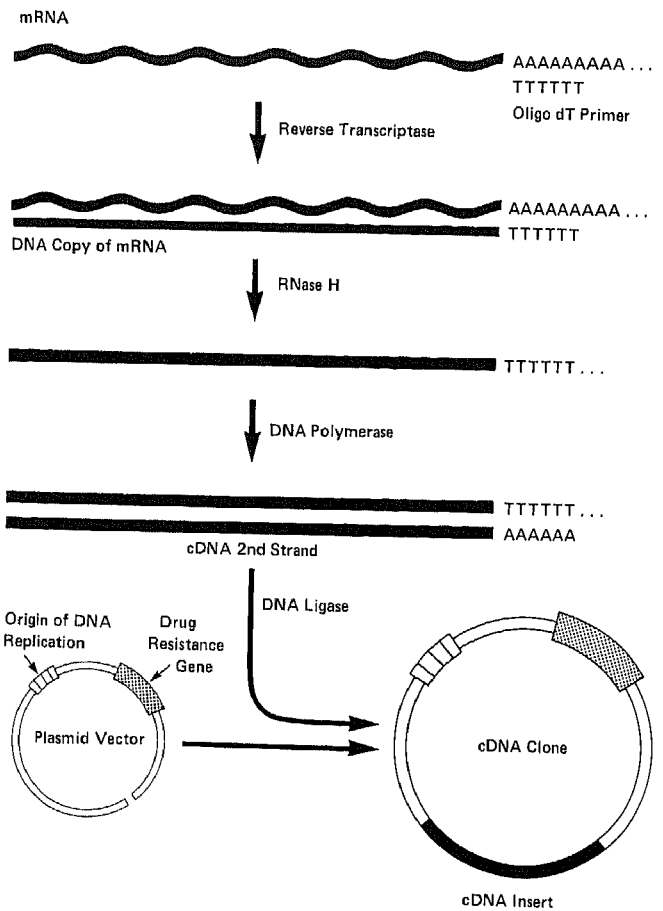


FIGURE 2.3 cDNA cloning. Schematic representation of the in vitro reactions that result in the production of cDNA clones.

transcribed and even more rarely included in the final processed mRNA. A random genomic clone is thus overwhelmingly likely (98%) to contain heavily repeated sequences, whereas a cDNA clone will be repeated, in general, only if the protein specified is a member of a gene family, such as the globins or the immunoglobulins.

In general, particular human genes have been cloned by the cDNA route, in which the protein specified by the gene is known. In such cases the mRNA is isolated from the tissue in which the gene is expressed, copied into DNA using the viral enzyme reverse transcriptase, and ligated into a suitable

vector. Genomic clones are then obtained by using the cDNA clone as a hybridization "probe" in order to retrieve corresponding genomic sequences from "libraries" of genomic sequences.

Constructing DNA Libraries

Since all nucleated cells contain essentially the entire genome (which represents all the human genetic material except mitochrondrial DNA), DNA to be used in the construction of genomic libraries can be easily obtained from lymphocytes isolated from a peripheral blood sample. Such libraries are made by fragmenting randomly the genomic DNA by partial digestion with a restriction enzyme and cloning the fragments into a vector such that enough clones are generated to make it overwhelmingly probable that all sequences are represented several times. Thus, one good genomic library, for example, the Maniatis human genomic library in a λ vector,[7] suffices for virtually all genomic cloning purposes.

A genomic library will contain different types of sequences aside from those that code for genes. Unique sequence DNA is estimated to comprise 70% of the human genome (reviewed in Lewin, 1980),[8] while the remainder consists of moderately repetitive DNA segments like ribosomal genes and others that are highly repeated such as the "Alu" or "Kpn" families.[9,10] Some highly repeated sequences are difficult to clone because they frequently recombine and are lost during standard cloning procedures and therefore may not be represented in genomic libraries.[11,12] It is also possible to construct chromosome-specific libraries by using as a source of insert physically separated chromosomes or hybrid cell lines (originating from fusions of rodent with human cells)[13] that are thought to contain a single human chromosome.[14]

In contrast, cDNA libraries reflect the tissue of origin, being made of copies of mRNAs that vary according to tissue type; different genes are expressed in liver and in brain, for example, and thus different mRNAs are present. In summary, the human geneticist has available two kinds of cloned human DNAs: cDNA, representing particular genes or the expressed genes of particular tissues, and genomic DNA, containing a continuous segment of the human genome. Both kinds of clone have uses in human genetics, many of which involve the examination of human DNAs and their restriction maps indirectly by hybridization.

DNA Sequencing

The ultimate characterization of a DNA molecule is, of course, its complete nucleotide sequence. Since the simultaneous development of the chemical

and enzymatic approaches,[15,16,17] each method has been refined and extended so that it has become routine to determine sequences of genes as they are cloned (See Fig. 2.1).

The chemical method employs the idea of radiolabeling a fragment of DNA just at one end and then chemically cleaving in a base-specific manner. A partial reaction, in that case, will contain a nested set of molecules whose different lengths correspond to the positions of the base at which cleavages were made. The lengths of labeled fragments from four such reactions suffice to indicate the relative positions of each of the bases in the sequence. Figure 2.4 shows a diagrammatic representation of these concepts. A recent extension of the chemical methods allows the determination of sequences without cloning, by using highly radioactive DNA probes directly on genomic DNA cleaved and separated by size instead of direct end labeling.[18]

The enzymatic method[17] is similar to the chemical method except that the DNA to be sequenced is copied in vitro with DNA polymerase and the reaction stopped in a base-specific manner, usually by the addition of chain-termination dideoxynucleotide analogs as substrates. Methods for applying enzymatic methods to DNA from whole genomes by amplification of specific segments have also been published.[19,20]

DNA sequencing is very labor intensive, whether one uses chemical or enzymatic methods. For this reason there is great interest in methods of automation of this process. A prototype automatic system has been announced.[21]

RECOMBINANT DNA METHODS IN GENETIC ANALYSIS

Genetic Markers

The ability to follow the inheritance of marker traits in humans is of particular importance in medical genetics for diagnostic applications, and, in addition, may constitute the initial steps toward the identification and characterization of the genes for which the molecular mechanisms are unknown. Furthermore, even complex disorders in which environmental and genetic factors combine can in principle be studied by following the inheritance of any one or more genetic loci that might be responsible for the phenotype. In addition to the contribution of environmental factors, identification of the genetic loci important in such a disorder may allow the subdivision of the phenotype into different classes, each of which may be effectively treated with a therapy tailored to the subtype. Classical genetic markers, for example, protein polymorphisms such as ABO blood groups and HLA antigens, have found practical benefit for blood and tissue typing, but have limited usefulness in genetic mapping in humans because they are too few in number

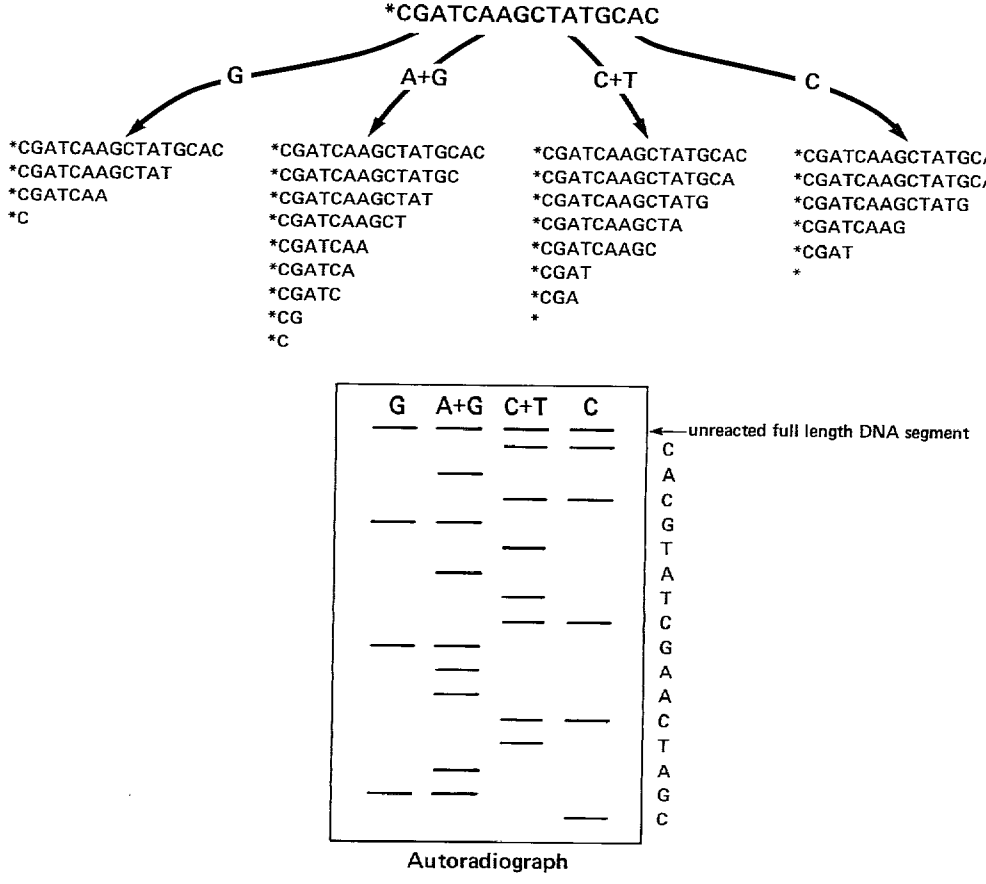


FIGURE 2.4 Maxam Gilbert DNA sequencing. A unique-sequence DNA fragment labeled at the 5′ end with P^{32} is treated with nucleotide base-specific reagents in four separate chemical reactions (dimethyl sulfate is used for purines A and G, and hydrazine is used for pyrimidines T and C). Reaction conditions are controlled so that on average only one nucleotide per DNA fragment reacts with the nucleotide base-specific reagent. The nucleotide chain is then cleaved at the position of the treated base, which produces a nested set of fragments representing cleavage at the site of nucleotide base modification. The reaction products are then separated according to size by electrophoresis through a polyacrylamide gel. The gel is exposed to radiographic film, and the DNA sequence read from the autoradiograph. The first "base" represented in the autoradiograph is P^{32}. See Maxam and Gilbert, 1980, for a complete description of this procedure.

The principle of other sequencing methods is similar except that the differences are in the method of interrupting the chain in a base-specific manner or the method of labeling the 5′ end of the molecule specifically.

and, except for HLA, are insufficiently polymorphic. The advent of recombinant DNA technology opened the way to the development of an entirely new system of genetic markers for humans that enable the mapping of virtually any inherited trait. These markers, RFLPs, exploit the variation in DNA sequence among individuals. The examination of sequence variation at particular regions of the genome allows distinctions to be made among individuals with a resolution much greater than any preexisting method. It has been estimated that sequence differences among individuals occur on average every 50 to 100 nucleotides.[22] Such regions of variation can be sampled by the use of restriction enzymes since they recognize and cleave unique sequences in DNA.*

Detecting RFLPs by Southern Blotting

The methodology most commonly employed to visualize sequence polymorphism is by gel transfer and hybridization of size-fractionated enzyme-cleaved genomic DNA to labeled cloned DNA; this procedure is also known as *Southern blotting*[22] (Southern, 1975). The procedure is carried out as follows (Fig. 2.5):

1. Genomic DNA from a set of unrelated individuals (A–E in Fig. 2.5) is cleaved with a restriction enzyme and fractionated according to size by electrophoresis through an agarose gel.
2. The DNA is denatured to separate the strands and transferred to a solid support such as a nitrocellulose or nylon membrane.
3. A cloned segment of DNA is radioactively labeled and hybridized to the genomic DNA attached to the membrane in the presence of a solution that facilitates the formation of hydrogen bonds. The cloned DNA hybridizes, or anneals with, its complementary sequence present in the genomic DNA. The membrane is then washed to remove any nonspecifically bound labeled DNA and exposed to radiographic film.
4. Any polymorphism is inferred from the differences, if any, between individuals in the pattern of restriction-fragment lengths displayed on the autoradiograph.

*A note on the word *polymorphism* is in order. In classical population genetics a locus is said to be polymorphic when it displays at least one alternative allele present at a frequency of at least 1% in the population. In modern DNA work, one is interested mainly in the likelihood of heterozygosity at a locus, which will be usefully high only when the locus is very polymorphic (minor allele frequencies of 10% or greater, and multiple alleles). The way in which DNA polymorphisms are now employed has resulted in a subtly different usage of the word, although no formal change in definition is required. In RFLP analysis stress is put on the expected degree of *heterozygosity* at a locus, not, as was classically the case, on the absolute frequency in a population of the particular alleles at a locus.

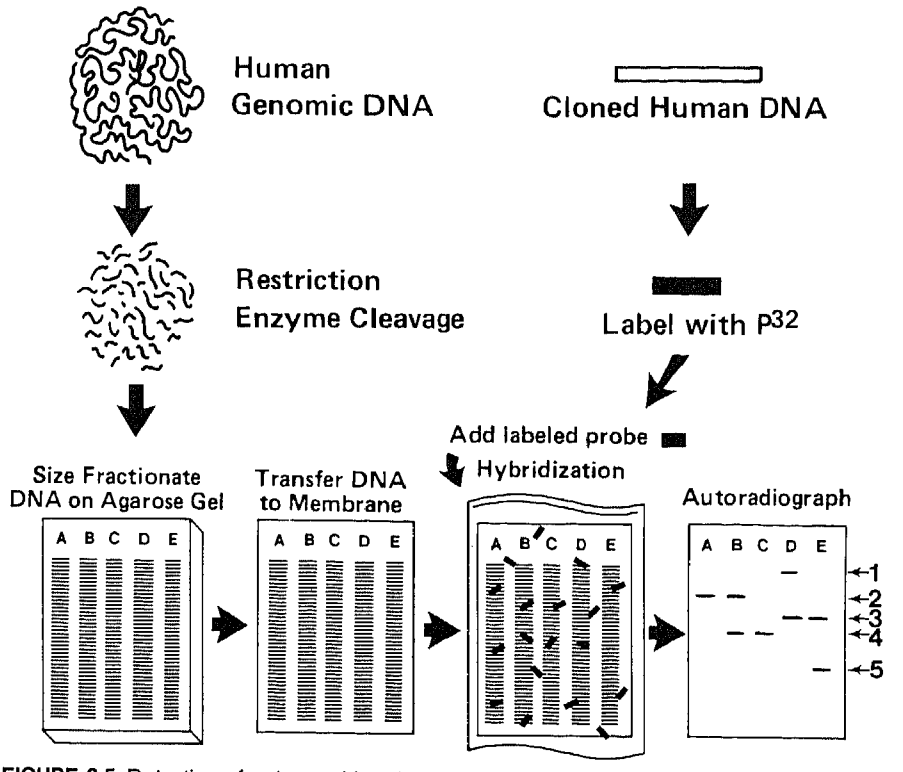


FIGURE 2.5 Detection of polymorphism by gel transfer. A cloned human DNA segment is tested for its ability to reveal polymorphism by hybridization to size-fractionated genomic DNA from five unrelated individuals (A,B,C,D,E) cleaved with a restriction enzyme. Five different fragment lengths (alleles 1–5) are detected with the radioactively labeled human DNA probe. The genotypes of the individuals are A 2,2; B-2,4; C-4,4; D-1,3; E-3,5.

As Figure 2.5 shows, five different patterns are seen among the five unrelated individuals. Since everyone inherits two copies of each chromosome, one from the mother and one from the father, except for the sex chromosomes, one would expect to see two fragment lengths, or alleles, in each individual. Individuals *A* and *C* in Figure 2.5 each have two alleles that are the same length, thus showing only one band, are therefore termed *homozygotes*. Individuals *B, C,* and *E* have a different fragment length for each allele and are termed *heterozygotes*. Therefore, this region of DNA displays polymorphism in DNA fragment length: this is a good RFLP locus. There are five alternative fragment lengths (*alleles*) at this locus. In some cases, more than one fragment length constitutes a single allele at a locus and thus the set of fragments is inherited according to mendelian principles (see Knowlton et al, 1986,[24] and Schumm et al, 1987,[25] for examples). The

geneticist recognizes as a "locus" any length of DNA within which recombination (see below) is so small as to be negligible.

A genomic library with relatively large inserts (10–20 kb) is a good source of cloned DNA probes to test, as above, for polymorphism. Since most of the clones will contain repeated DNA, it is necessary to prehybridize the clones with total genomic DNA so that only unique sequence regions of the clones are available to be tested for polymorphism in the gel transfer experiment[25,26] (Litt et al, 1985; Schumm et al, 1987). Alternatively, about 1% of the human DNA clones in λ vectors contain entirely single-copy DNA and therefore can be tested for polymorphism without the prehybridization step.[27] The cloned DNA sequences are compared as above to genomic DNA from a variety of unrelated individuals.

A set of restriction enzymes that survey different sequences are tested with each genomic clone. It has been observed that the enzymes TaqI and MspI are especially useful in revealing polymorphism in humans because they contain as part of their recognition sequence the dinucleotide CG, which is apparently highly mutable.[25,28] Other restriction enzymes that are particularly useful in revealing polymorphism are RsaI, HindIII, BglII, PstI, EcoR1, and BamH1.[25]

Figure 2.6 shows an examle of a typical screening blot in which the genomic DNA of five unrelated individuals is cleaved with six restriction enzymes and tested for polymorphism with a cloned DNA probe, CRI-RL4-117. All of the individuals apparently share the same sequences at the sites tested with five of the enzymes. However, a polymorphism is visualized with the enzyme MspI. This is a simple two-allele polymorphism that is caused by a single base pair difference at the recognition site of the enzyme. Polymorphism may also occur as a result of DNA rearrangements, that is, by the insertion or deletion of DNA segments. Figure 2.7 shows an example of a rearrangement polymorphism. Polymorphisms at single genetic loci can be much more complicated than the simple ones illustrated here and may consist of clustered site changes and/or multiple DNA rearrangements with a number of fragment lengths that characterize each allele. The most useful polymorphisms have a large number of alleles represented at equal frequencies in the population. With such markers it is likely that most individuals will be heterozygous at the locus and that the alleles contributed by each parent can be identified in the offspring. The usefulness of a polymorphism in an inheritance study is represented by its polymorphism information content (PIC)* value.[22]

*The PIC is a mathematical measure of the usefulness of a polymorphism in linkage studies. It is the probability of identifying in any given offspring the parental chromosome contribution. Some authors use the mean heterozygosity (simply the frequency of heterozygotes) as a measure of polymorphism. The difference between PIC and heterozygosity is that the former estimates

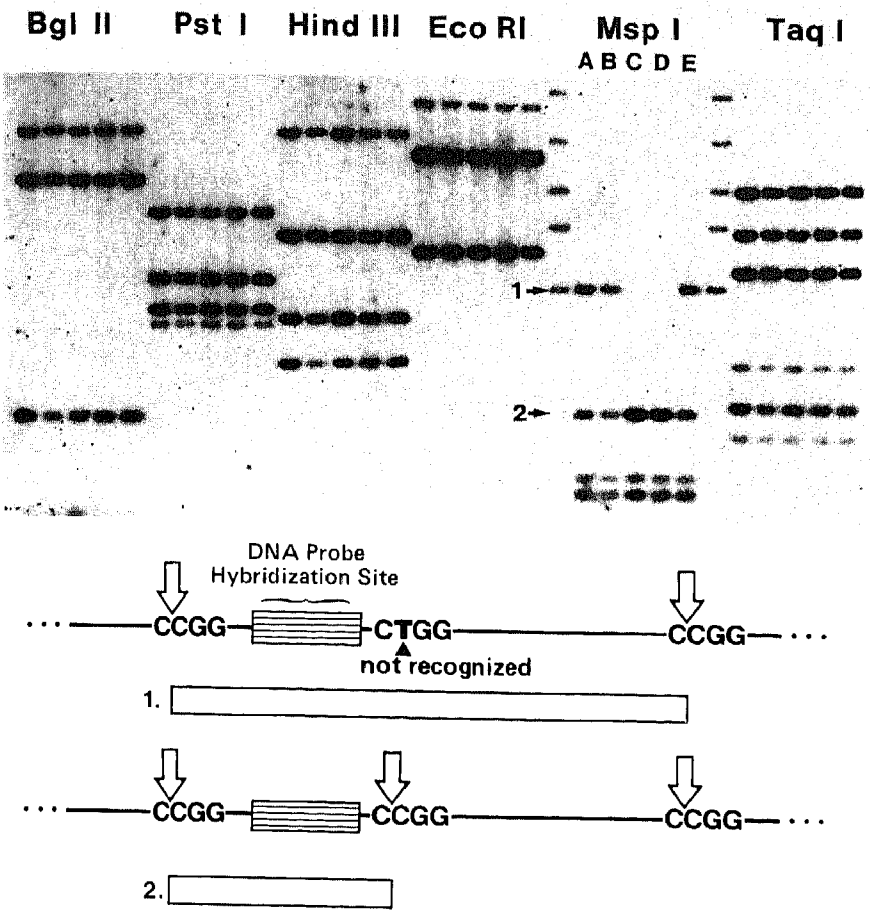


FIGURE 2.6 Screening for polymorphism: base change polymorphism. Southern transfer of genomic DNA from five unrelated individuals (A,B,C,D,E) cleaved with the restriction enzymes indicated and probed with genomic clone CRI-RL4-117. A site change polymorphism is revealed by the enzyme MspI. Individuals A,B, and E are heterozygous at the locus, whereas individuals C and D are homozygous for allele 2 (*arrows*). The recognition site of the enzyme and resulting fragment lengths (alleles) are shown below the autoradiograph. Note: DNA probe hybridization site is not drawn to scale, and only the top strand of the duplex DNA molecule is illustrated with the restriction endonuclease cleavage sites.

---

more exactly the probability that a marker will be informative in practice when used in an inheritance study of a random family. The heterozygosity overestimates the usefulness of a marker at low values; at high values the PIC and heterozygosity converge. A polymorphism with a PIC value of 1 is fully informative for any mating (ie, the parental chromosome contribution can be unambiguously determined in the offspring). Figure 2.8 shows the inheritance of an RFLP fully informative in a three-generation family. Four distinguishable alleles at the locus are present in the parents, and it is easy to trace the inheritance of the alleles in the offspring.

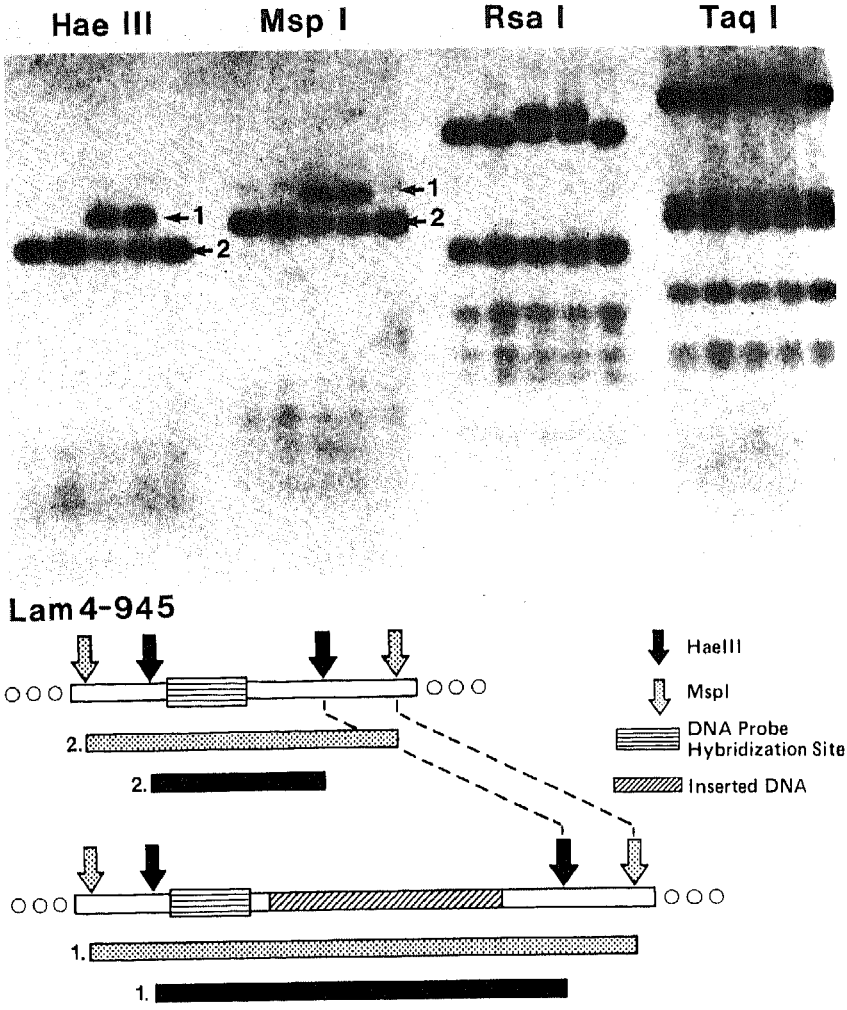


FIGURE 2.7 Screening for polymorphism: DNA rearrangement polymorphism. Southern transfer of genomic DNA from five unrelated individuals cleaved with the restriction enzymes indicated and probed with the genomic clone CRI-Lam4-945. An insertion/deletion polymorphism is indicated from comparison of the similarity of patterns from each of the enzymes; the alleles differ in molecular weight by an absolute amount in the various restriction digests, as indicated by the drawing shown below the autoradiograph.

The examples shown demonstrate polymorphism of DNA fragments that range in length from 1 to 20 kb. It is also possible to resolve smaller and much larger fragment lengths by using different gel systems and electrophoresis conditions. Polymorphisms have been identified with fragments that range in length from ten to several hundred nucleotides,[29] and in principle

polymorphisms should be identifiable with fragments in the hundreds of kilobases resolved by orthogonal field gel electrophoresis (OFAGE)[30] or field inversion electrophoresis systems (FIGE).[31]

Other Methods for Detecting Sequence Variation

Another method of detecting sequence variation that does not rely on restriction enzymes is the use of denaturing gradient gel electrophoresis.[32] This method, which in principle can detect heterozygosity at any sequence (not just restriction sites), has yet to be applied on a large scale to the detection of polymorphisms suitable for use as genetic markers. It has been used, however, to detect many known mutations in the globin genes.

Synthetic nucleotide probes can also be used to identify sequence variation. In this case small oligonucleotide probes are used to test for single base differences. For example, oligonucleotide probes of 17 nucleotides have been used to detect the base change that is responsible for sickle cell anemia.[33] In this case the hybridization conditions were controlled so that even a single mismatch between the sequence of the probe and the genomic DNA resulted in a failure of the probe to hybridize to the corresponding genomic sequence.

Detection Systems

By far the most commonly used method for labeling DNA is nick translation[1,34] using radioactive nucleoside triphosphates in vitro. Random sequence priming and the production of radioactive RNA copies, all in vitro, have also been used to enhance the degree of specificity of radiolabeling.[35] Nonradioactive labeling methods have been suggested. Most notable among these are the ones based on biotinylated nucleosides after nick translation. The biotin can be used to attach avidin coupled to many different detectable materials, including fluorescent compounds and enzymes.[36] Unfortunately, these methods have not yet reached the sensitivity of the radiolabeling method; for the problem of detecting a single copy sequence in the human genome one still needs the high signal-to-noise ratio of the radioactive method.

GENETIC MAPPING

Linkage of Genetic Markers

The indispensable element of any kind of linkage study is heterozygosity. Only if an individual is heterozygous at each of two loci can the linkage relationship between the two loci be established. In the case of humans, the

requirement for heterozygosity means that high degrees of polymorphism are required, since one is dependent upon the natural occurrence of heterozygosity in random matings. As described above, RFLPs have provided an essentially inexhaustible supply of polymorphisms particularly suitable for linkage studies, since both alleles can always be determined in heterozygotes.

Given a set of fully informative RFLPs (ie, ones polymorphic enough that each parent in a randomly chosen family is likely to be heterozygous at every RFLP locus), it is possible to follow each of the parental contributions into the children directly and thus to determine linkage (see Figure 2.8). The principle is illustrated in Figure 2.9, where two of each of the parents chromosomes are shown. On one chromosome is shown the RFLP marker A, and the father has distinguishable RFLP alleles A1 and A2 while the mother has alleles A3 and A4. The other chromosome carries two other RFLP loci, B and C, each with four alleles (B1 and B2 in the father, B3 and B4 in the mother, C1 and C2 in the father, and C3 and C4 in the mother). All the properties of mendelian inheritance can now be demonstrated:

1. Each of the children receives either A1 or A2 from the father and either A3 or A4 from the mother (Mendel's first law of segregation) so that the four possible genotypes of the children are A1/A3, A1/A4, A2/A3, or A2/A4. Likewise for loci B and C.
2. All combinations of A and B alleles deriving from each parent are found in the children. Thus, following only the contribution from the father, we find that the combinations A1,B1; A2,B1; A1,B2; A2,B2 are equally common. This is Mendel's second law, also called the *rule of independent assortment.* Loci A and C also assort independently.
3. In contrast, only a limited number of combinations of B and C alleles deriving from each parent are found in the children: thus, again following the contribution from the father, we find that the combinations B1,C1 and B2,C2 are found in the children, but the combinations B1,C2 and B2,C1 are not. This principle is called *linkage,* and reflects the fact that B and C are physically close together on the chromosome in the example. When linkage is complete, then the alleles B1 and C1 were inherited from the grandparents together and in turn are passed on together to the children.

Unlike the example in Figure 2.9, however, linkage ordinarily is not an all-or-none phenomenon. Some gene pairs indeed lie so close together that they rarely reassort and thus approximate the limiting case of complete linkage, but linked markers usually lie far enough apart that they can become separated in meiosis by a process called *recombination.* Thus, in the example the allele combinations B1,C2 and B2,C1 are recombinant types, whereas the combinations B1,C1 and B2,C2 are *parental.* Linkage is more generally

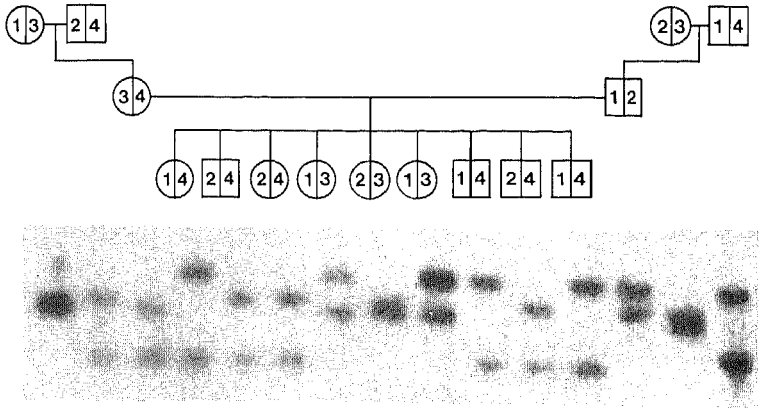


FIGURE 2.8 Inheritance of RFLP alleles. Mendelian inheritance of an RFLP in a three-generation family. A human DNA clone RL4-365 isolated from a human genomic library, labeled with P^{32} by a nick translation reaction and hybridized against the human genomic DNA cleaved with the restriction enzyme BglII detects four alleles (1-4.2kb;2-4.0kb;3-3.8kb;4-3.5kb).

expressed as a deviation from random assortment (ie, equal frequencies [50%] of parental and recombinant types in the progeny, as in loci A and B above) so that the recombinant types appear less frequently than the 50% expected.

It is important to understand that there is nothing in the RFLP pattern or "phenotype" of the parents that indicates whether alleles B1 and C1 lie on the same chromosome (and B2 and C2 on the other) or whether the opposite arrangement (B1 and C2 on one chromosome and B2 and C1 on the other). All we know from examination of their DNA with RFLP probes is that the genotype of the father is B1/B2, C1/C2 and the genotype of the mother is B3/B4, C3/C4. In the example, we know the answer to the question of their chromosomal origin from the way in which the pair was passed on to the children (ie, B1 and C1 together); we could also infer the arrangement (often called the *linkage phase*) from an examination of the DNA of the grandparents.

Yet, as the example shows, when we do not know about linkage in advance, and even when we do not know the linkage phase in advance, it is clear that one can infer both. If there are many children and the RFLPs are informative enough (ie, sufficiently polymorphic to be used in many different families), we can infer linkage by looking for enough data to make the deviation from random assortment statistically significant.

Quantitative assessments of the significance of linkage estimates are made by calculations in which the likelihood that the data observed arose from a particular model (ie, linkage) is compared to the likelihood of observ-

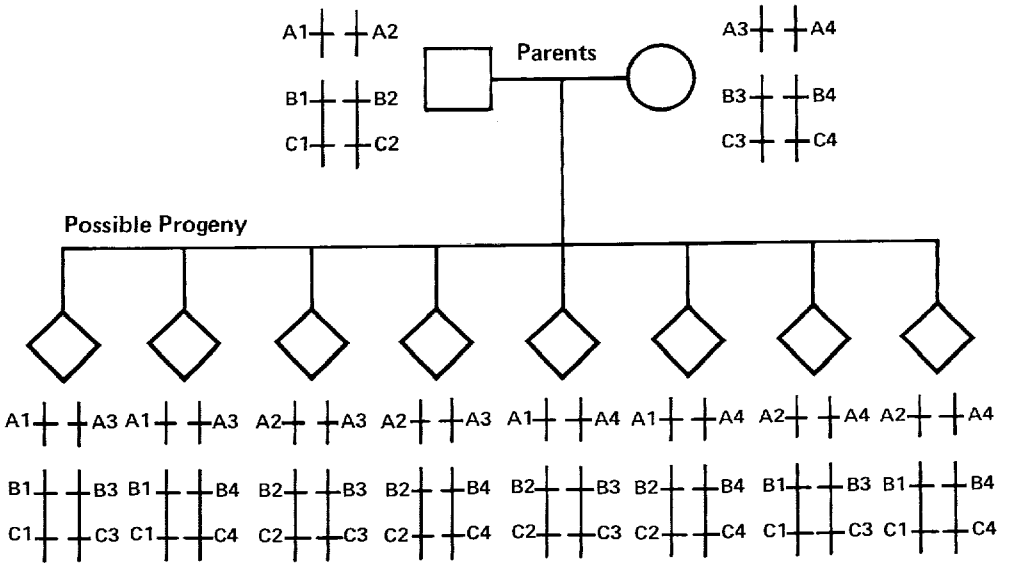


FIGURE 2.9 Using the inheritance of RFLP alleles to determine linkage. Three RFLP loci (A,B,C) on chromosome pairs are shown with possible inheritance patterns in the offspring (see text for details).

ing the same data on the alternative hypothesis (random assortment). By longstanding convention, linkage is regarded as having been demonstrated when the likelihood ratio (often called the odds ratio) reaches 1,000:1 in favor of the linkage hypothesis[37] (see Ott, 1985,[5] for a full discussion of likelihood methods as applied to linkage). The statistic usually calculated is the logarithm of the odds ratio (LOD); LOD scores obtained from different families are additive, and, of course, significance is achieved when the LOD reaches a value of 3. The calculation of a LOD score expected for a simple case is given in Figure 2.10, where it can be seen that a single phase-known meiosis will contribute an expected LOD score of +0.16 if the hypothesis of linkage at a distance of 10% recombination is true, and −0.22 if the alternative hypothesis of nonlinkage is true.

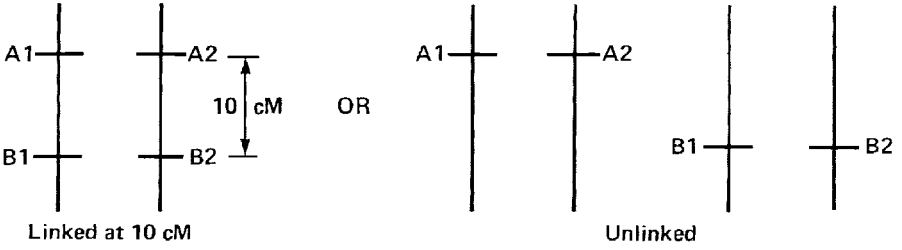
The likelihood method is very flexible, and hypotheses much more complex than the one in Figure 2.10 can be accommodated. Uncertainties of linkage phase, for example, can easily be taken into account by altering the hypothesis and thus the calculation of the likelihood that the data was derived from this hypothesis. Of course, ambiguities (such as lack of knowledge of linkage phase) will reduce the magnitude of a given family's contribution to the overall LOD score; conversely, the presence of multiple affected siblings will raise the contribution. LOD scores can be obtained from several different families, and these scores can be added until the score rises to 3 (or, alternatively, declines to −2, the conventional level for rejection of the linkage hypothesis).

Thus, RFLP markers are used to map disease genes by applying them to members of families in which the disease is segregating and trying, for each RFLP marker, to find evidence for linkage as assessed by the LOD scores. As mentioned, the value of different families with respect to their contribution to achieving an LOD of 3 varies: in general, for simple mendelian diseases, the ideal circumstance is a family in which there are several affected siblings (to provide many opportunities to observe crossovers or the lack of crossovers with the RFLP marker) and, if possible, grandparents (who can reduce ambiguities about linkage phase).

Closing in on a Gene

Genetic linkage to several single gene disorders has been discovered over the past few years. For example, the loci responsible for Duchenne muscular dystrophy,[38] Huntington's disease,[39] adult polycystic kidney disease[40] and cystic fibrosis[41-44] have been identified by testing random DNA markers for linkage in families segregating these disorders.

The major virtue of the linkage approach is that one follows, in family studies, the actual biologic defect in inherited disease because presence and absence of the disease phenotype itself is scored. Before beginning, one has



Possible Genotypes	Frequency of Genotype		Ratio of Odds	Expected Contribution to LOD Score	
	If Linked	If Unlinked	Ratio of Odds (linked/unlinked)	If Linked	If Unlinked
A2B2	0.90	0.50	9/5	0.9 log 9/5	0.5 log 9/5
A2B1	0.10	0.50	1/5	0.1 log 1/5	0.5 log 1/5
				Expected LOD Score =	+0.16 or −0.22

FIGURE 2.10 LOD score from a single-phase known meiosis. The contribution of one meiosis to the LOD (log of the odds) score calculation is shown. For the purpose of illustration, 10 cM is taken as exactly 10% recombination.

no idea of the physical location of the disease gene in the genome unless, of course, it is sex linked. However, upon detection of statistically credible linkage to an RFLP marker, a piece of DNA (the probe) is available that can be located physically by conventional methods. Currently these methods are two: mapping by somatic cell hybrids between rodent and human cells[45] and in situ hybridization to human chromosomes.[46]

Somatic cell hybrid panels are constructed by fusion of rodent cells to human cells and continued propagation of the products of fusion. The rodent complement of chromosomes remains stable during this process, but the human chromosomes are somewhat randomly lost during passage of the cells. Thus, one collects a set of individual cell lines that may contain, for example, a cell line with human chromosomes 6,2,X, and the rodent chromosomes (cell line A); another cell line (B) containing human chromosomes 2,5,8,10, and the rodent chromosomes; and another cell line (C) with human chromosomes 5,6,8,10,X, and the normal complement of rodent chromosomes. By characterization of the cell lines as to their human chromosome constitution a panel of cell lines can be assembled such that an uncharacterized segment of DNA can be assigned to a human chromosome based on its ability to hybridize (or fail to hybridize) to human DNA from the hybrid lines. By a process of scoring the hybridization for concordance or discordance, the chromosomal origin of the DNA segment can be deduced. In the example outlined above, the DNA segment could be unambiguously assigned to chromosome 2 if the DNA segment hybridized to the cell lines containing human chromosome 2 (lines A and B) and not to line C, the cell line containing human chromosomes also present in lines A and B (ie, 5,6,8,10,X).

The other method used to assign DNA segments to chromosomes consist of labeling the DNA segment of interest with tritiated hydrogen and hybridizing the segment directly to stained metaphase chromosomes that have been attached to glass slides. Autoradiographs are made of the fixed chromosomes and the "grains," that is, the radioactive decay positions, are counted and scored as to the chromosomal location (the chromosomes are first identified by their characteristic staining pattern).

These two methods are complementary, and each is capable of localizing the marker (and thus the linked gene) to a physical chromosome with a resolution on the order of a few percent of the human genome. In every case where linkage has been detected to an autosomal mendelian disease locus, this physical mapping has followed immediately.[39,42]

Knowledge of the chromosomal location is helpful in trying directly to clone and identify the disease gene. Many approaches are being tested for this purpose: the most promising include direct microdissection of the relevant region from a number of stained chromosomes;[47] the identification of ex-

pressed sequences in the region of interest from cDNA libraries from appropriate tissues;[48] the use of natural chromosomal rearrangements in the region of interest.[49,50] More speculative approaches include attempts to produce restriction maps of very large chromosomal fragments using OFAGE and FIGE and the development of methods to jump, hop, or skip along chromosomes over distances of hundreds of kilobases using a continually evolving array of new vectors.[51,52]

Virtually all of these physical methods profit from the identification by further linkage studies of additional polymorphic markers in the region of interest. As a rule of thumb, 1% recombination (a minimally detectable distance) corresponds on the average to about 1 million base pairs. The largest segment of DNA that one can clone is about 50 kb. Therefore, simply trying to "walk" along the chromsome by overlapping clones is formidable by current technology. It is eminently worthwhile, having found linkage at about 5 to 10% recombination, to attempt to saturate the region with polymorphism in the expectation of getting down to the minimal 1% (ie, 1 megabase [mb]) before applying the physical methods. In the case of chronic granulomatous disease, in fact, the availability of many markers in a region allowed the direct detection of the mRNA product of the disease gene as well.[48]

Efficient Genetic Mapping Using a Genome RFLP Map

The way in which RFLP markers are used also has an influence on the efficiency with which disease genes can be mapped. If, instead of trying markers randomly one at a time, one first maps the markers with respect to each other, multilocus mapping methods can be applied in which the LOD score contributed by a single meiosis can be twice or more that obtained for the same markers scored singly.[53,54] Maps of RFLP loci with respect to each other are made in the same way as linkage to diseases, using families with many siblings and, where possible, grandparents. Maps of several chromosomes already exist, notably the X chromosome[55] and chromosome 7.[56,57] Although these maps are made, in part, with RFLPs of limited informativeness, their existence clearly indicates the strong likelihood that a complete useful RFLP map of the human genome is at most a few years away.

Applications of the Human RFLP Map

Using a complete RFLP map instead of markers singly has advantages beyond efficiency in mapping simple mendelian traits. It recently has been

found that use of a complete map offers the possiblity of mapping much more complex genetic diseases. Among the situations that clearly can be dealt with are *genetic heterogeneity* (in which a disease is caused by any of several genes); some kinds of multifactorial or polygenic diseases (in which several genes must cooperate to produce the disease); and instances in which the gene(s) involved provide not a certainty of disease, but only a predisposition to it.[54,58]

The informativeness, number, and distribution of RFLPs required to make a useful RFLP map can be calculated as well. In summary, a map of completely informative RFLPs (PIC greater than about 0.8) evenly spaced at intervals of 40 centimorgans (cM) (ie, less than 100 RFLP loci for the whole genome, which is 3300 cM in length) will be nearly optimal for efficient mapping of single-gene diseases in as few as 10 families with two or more affected sibs. Where the RFLPs are less informative, more markers per centimorgan will be required. To achieve even spacing, of course, many more markers will have to be mapped than the number that, in the end, would delimit the even intervals.[59] More complicated genetic problems will require better maps.

It is clear that there are many more diseases in which inheritance plays a major part than just the simple mendelian ones. The complete RFLP map affords an opportunity for studying the inherited components of this larger group of diseases. Theoretic studies[54,58,60] indicate that inherited diseases in which any of as many as five genes can cause a disease (genetic heterogeneity) are susceptible to analysis with a complete map. Such diseases include ataxia telanglectasia and xeroderma pigmentosa and possibly many other conditions. Furthermore, similar calculations suggest that diseases in which more than one gene participates in a given family can also be analyzed successfully.[58] Thus, it seems likely that progress can be made in that very large area in which positive evidence (eg, twin studies or family clustering) for inherited components exists but for which a simple genetic hypothesis does not suffice. This class of diseases is very large and includes bipolar affective disorder (manic depression), many forms of cancer, and diabetes.

We have seen that the use of recombinant DNA methods in clinical medicine has already become a reality, especially in the diagnosis of inherited diseases. The number of inherited diseases[61] makes it inevitable that many more diagnositically useful tests will emerge in the very near future. The extension of linkage methods (especially the completion of the RFLP map) promises to make possible in the not-too-distant future the application of linkage methods to common diseases not simply caused by the inheritance of a single dominant or recessive gene. It is too early to tell exactly how useful genes defining the inherited components of such diseases will be. It seems certain, however, that the continued application of these methods in re-

search will result, at the minimum, in a much better understanding of the relationship between heredity and other factors.

ACKNOWLEDGMENTS

We thank Gita Akots, Valerie Brown, and Cindy Helms for use of the autoradiographs shown in Figures 2.6, 2.7, and 2.8, respectively.

REFERENCES

1. Maniatis T, Fritsh EF, Sambrook J: *Molecular Cloning: A Laboratory Manual.* New York, Cold Spring Harbor Laboratory, Cold Spring Harbor, 1982.
2. Wu R (ed): Recombinant DNA, in *Methods in Enzymology,* New York, Academic Press, 1979, vol 68.
3. Wu R, Grossman L, Moldave K: Recombinant DNA: Part B, in *Methods in Enzymology.* New York, Academic Press, 1982, vol 100.
4. Wu R, Grossman L, Moldave K: Recombinant DNA: Part C, in *Methods in Enzymology,* New York, Academic Press, vol 101.
5. Ott J: *Analysis of Human Genetic Linkage.* Baltimore, Johns Hopkins University Press, 1985.
6. Nathans D, Smith HO: Restriction endonucleases in the analysis and restructuring of DNA molecules. *Ann Rev Biochem* 1975;44:273–290.
7. Lawn R, Fritsch E, Parker R, et al: The isolation and characterization of linked delta and beta globin genes from a cloned library of human DNA. *Cell* 1978;15:1157–1174.
8. Lewin B: Eukaryotic chromosomes, in *Gene Expression,* ed 2. New York, John Wiley & Sons, 1980, vol 2, pp. 503–569, 861–930.
9. Schmid CW, Deininger L: Sequence Organization of the Human Genome. *Cell* 1975;6:345–358.
10. Schmid CW, Jelinek WR: *Science* 1982;216:1065–1070.
11. Wyman AR, Wolfe LB, Botstein D: Propagation of some human DNA sequences in bacteriophage lambda requires mutant *Escherichia coli* hosts. *Proc Nat Acad Sci USA* 1985;82:2880–2884.
12. Wyman AR, Barker D, Wertman KF, et al: Factors which equalize the representation of genome segments in recombinant libraries. To be published.
13. Van Dilla MA, Dearen LL, Albright RL, et al: Human chromosome-specific libraries: Construction and availability. *Biotechnology* 1986;4:537–552.
14. Saxon PJ, Srivatsan ES, Leipzig CV, et al: Selective transfer of individual human chromosomes to recipient cells. *Mol Cell Biol* 1985;5:140–146.
15. Maxam AM, Gilbert W: A new method for sequencing DNA. *Proc Natl Acad Sci USA* 1977;74:560–564.
16. Maxam AM, Gilbert W: Sequencing endlabeled DNA with base-specific chemical cleavages. *Meth Enzymol* 1980;65:499–560.
17. Sanger F, Nicklen S, Coulson AR: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 1977;74:5463–5467.
18. Church GM, Gilbert W: Genomic sequencing. *Proc Natl Acad Sci USA* 1984;81:1991–1995.
19. Saiki RK, Arnheim N, Erlich HA: A novel method for the detection of polymorphic restriction sites by cleavage of oligonucleotide probes: Application to sickle cell anemia. *Biotechnology* 1985;3:1008–1012.

20. Saiki RK, Scharf S, Faloona F, Erlich HA, et al: Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985;230:1350–1354.
21. Smith LM, Sanders JZ, Kaiser RJ, et al: Fluorescence detection in automated DNA sequence analysis. *Nature* 1986; 321:674–679.
22. Botstein D, White R, Skolnick M, et al: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 1980;32:314–331.
23. Southern EM: Detection of specific sequences among DNA fragments separated by electrophoresis. *J Mol Biol* 1975;98:503–517.
24. Knowlton R, Brown V, Braman J, et al: Use of highly polymorphic DNA probes for genotypic analysis following bone marrow transplantation. *Blood* 1986;68:378–385.
25. Schumm JS, Knowlton RG, Barker DF, et al: Identification of more than 500 RFLPs by random screening. *Am J Hum Genet.* In press.
26. Litt M, White R: A highly polymorphic locus in human DNA revealed by cosmid derived probes. *Proc Natl Acad Sci USA* 1985;82:6206–6210.
27. Wyman AR, White R: A highly polymorphic locus in human DNA. *Proc Natl Acad Sci USA* 1980;77:6754–6758.
28. Barker D, Schafer M, White R: Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. *Cell* 1984;36:131–138.
29. Kreitman M, Aguadé M: Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by four cutter hybridization. *Proc Natl Acad Sci USA* 1986;83:3562–3566.
30. Schwartz DC, Cantor CR: Separation of yeast chromosome-sized DNAs by pulse field gradient gel-electrophoresis. *Cell* 1984;67–75.
31. Carle GF, Frank M, Olson MV: Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. *Science* 1986;232:65–68.
32. Myers RM, Lunelsky N, Lerman L, et al: Detection of single base substitutions in total genomic DNA. *Nature* 1985;313:495–498.
33. Studencki AB, Wallace RB: Allelespecific hybridization using oligonucleotide probes of very high specific activity: Discrimination of the human betaA and beta-Sglobin genes. *DNA* 1984;3:715.
34. Rigby PWJ, Dieckmann M, Rhodes C, et al: Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. *J Mol Biol* 1977;113:237–254.
35. Melton DA, Krieg PA, Rebagliati, et al: Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucl Acids Res* 1984;12:7035–7056.
36. Leary JJ, Brigati DJ, Ward DC: Rapid and sensitive colorimetric method for visualizing biotin-labeled DNA probes hybridized to DNA or RNA immobilized on nitrocellulose: Blo-blots. *Proc Natl Acad Sci USA* 1983;80:4045–4049.
37. Morton NE: Sequential tests for the detection of linkage. *Am J Hum Genet* 1955;7:277–318.
38. Davies KE, Pearson PL, Harper PS, et al: Linkage analysis of two cloned DNA sequences flanking the Duchenne muscular dystrophy laws on the short arm of the human x chromosome. *Nucl Acids Res* 1983;11:2303–2312.
39. Gusella JF, Wexler NS, Conneally PM, et al: A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 1983;306:234–238.
40. Reeders SP, Breuning MH, Davies KE, et al: A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature* 1985;317:542–544.
41. Tsui LC, Buchwald M, Barker DF, et al: Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* 1985;230:1054–1057.
42. Knowlton R, Cohen-Haguenauer O, Van Cong N, et al: A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature* 1985;318:380–381.

43. Wainwright B, Scambler P, Schmidtke J, et al: Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature* 1985;318:384–385.
44. White R, Woodward S, Leppert M, et al: A closely linked genetic marker for cystic fibrosis. *Nature* 1985;318:382–384.
45. Kuhn LC, McClelland A, Ruddle FH: Gene transfer, expression, and molecular cloning of the human transferrin receptor gene. *Cell* 1984;37:95–103.
46. Harper ME, Saunders GF: Localization of single-copy DNA sequences on G-banded human chromosomes by *in situ* hybridization. *Chromosoma* 1981;83:431–439.
47. Rohme D, Fox H, Herrmann B, et al: Molecular clones the mouse t complex derived from microdissected metaphase chromosomes. *Cell* 1984;36:783–788.
48. Royer-Pokora B, Kunkel LM, Monaco AP, et al: Cloning the gene for an inherited human disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature* 1986;322:32–38.
49. Ray PN, Belfall B, Duff C, et al: Cloning of the breakpoint of an X;21 translocation associated with Duchenne muscular dystrophy. *Nature* 1985;318:672–675.
50. Kunkel LM, Monaco AP, Middlesworth W, et al: Specific cloning of DNA fragments absent from the DNA of a male patient with an X-chromosome deletion. *Proc Natl Acad Sci USA* 1985;82:4778–4782.
51. Collins FS, Weissman SM: Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proc Natl Acad Sci USA* 1984;81:6812–6816.
52. Poustka A, Pohl TM, Barlow DP, et al: Construction and use of human chromosome jumping libraries from NotI-digested DNA. *Nature* 1987; 325:353–357.
53. Lathrop GM, Lalouel JM, Julier C, et al: Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am J Hum Genet* 1985;37:482–498.
54. Lander E, Botstein D: Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. *Proc Natl Acad Sci USA,* to be published.
55. Drayna D, White R: The genetic linkage map of the human X chromosome. *Science* 1985;230:753–758.
56. Donis-Keller H, Barker DF, Knowlton RG, et al: Highly polymorphic RFLP probes as diagnostic tools, in *Cold Spring Harbor Symposia Quantitative Biology: Molecular Biology of Homo Sapiens.* New York, Spring Harbor Press, 1987;51:317–324.
57. Barker DF, Green P, Knowlton RG, et al: A genetic linkage map of 63 chromosome 7 DNA markers. *Proc Natl Acad Sci USA* 1987. In press.
58. Lander E, Botstein D: Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map, in *Cold Spring Harbor Symposia: Molecular Biology of Homo Sapiens,* New York, Cold Spring Harbor Laboratory Press, 51:49–62.
59. Lange K, Boehnke M: How many polymorphic genes will it take to span the human genome? *Am J Hum Genet* 1982;34:842–845.
60. Cavaill-Sforza LL, King M-C: Detecting linkage for genetically heterogeneous diseases and detecting heterogeneity with linkage data. *Am J Hum Genet* 1986;38:599–616.
61. McKusick VA: *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes* ed 6. Baltimore, Johns Hopkins University Press.