SCIENTIFIC AMERICAN February 1988 Volume 258 Number 2

Chromosome Mapping with DNA Markers

Variable sequences in the DNA of human chromosomes act as genetic landmarks. Individual markers serve for tracing defective genes; collectively the markers provide the elements of a chromosome map

by Ray White and Jean-Marc Lalouel

Say that a disease is known to run in families, following a classic Mendelian pattern of inheritance. Somewhere among the 100,000 genes on the 23 pairs of human chromosomes a single gene is defective. The symptoms and progress of the disease have been described in meticulous detail, but its biochemistry is an enigma, and even predicting who will actually get the disease is guesswork. Such has been the case not just for a handful of rare afflictions but for most of the 3,000 known genetic diseases, including such familiar scourges as Huntington's disease and cystic fibrosis. Where does one begin the search for a causative mechanism, a diagnostic test and, ultimately, a treatment?

It is now possible to start by closing in on the defective gene itself. The territory to be surveyed is vast: the human chromosomes consist of linear molecules of double-strand DNA with a total length of about three billion base pairs (the chemical subunits that encode information along DNA). A typical gene, a complete unit of genetic information, is minuscule by contrast, encompassing perhaps 10,000 base pairs. And yet by correlating the inheritance of a distinctive segment of DNA—a "marker"—with the inheritance of a disease, one can now localize the mutant gene to within one or two million base pairs, or less than a thousandth of the human genome (the total complement of DNA). That kind of precision puts the gene within reach of molecular tools for cloning DNA and testing its activity. The identification of a genetic marker that is closely linked with a disease also means the gene's inheritance can be followed. It opens the way to simple tests for diagnosing carriers and future disease victims.

The basic strategy, known as linkage analysis, is a venerable tool of classical genetics. In our laboratory at the University of Utah and in many others, however, it has gained new power from the techniques of molecular biology, which make available a greatly expanded set of markers: molecular variations known as RFLP (for restriction-fragment length polymorphism) markers. Linkage analysis has now revealed RFLP markers for a number of disease genes, and many more diseases will soon yield to the strategy. It is also serving a more general purpose. By following the inheritance of many RFLP markers simultaneously in healthy families, we and other workers have begun to plot their positions in relation to one another and map them onto the physical framework of the chromosomes. The goal is a complete map of markers: an array of reference points that spans the genome and makes it possible to pinpoint disease genes far more efficiently than can be done with isolated markers.

The linkage strategy exploits the way genes are inherited. An ordinary human cell contains 23 pairs of homologous, or matching, chromosomes, one chromosome per pair inherited from the mother and the other from the father. In meiosis, the series of cell divisions that gives rise to germ cells (sperm or eggs), the homologous chromosomes in a progenitor cell are duplicated and then distributed among four germ cells, each of which receives 23 single chromosomes. The parental chromosomes are not transmitted intact, however. In the course of meiosis homologous chromosomes repeatedly recombine: they "cross over" and exchange segments of equal length [*see illustration on page 43*]. As a result each chromosome that is transmitted in a germ cell is generally a patchwork of segments from the two parental chromosomes. Recombination is the phenomenon that enables one to find linkage between a marker and a disease.

What makes it possible to detect recombination and employ it in linkage analysis are the many differences between homologous chromosomes. They often carry two different alleles, or versions, of many of their matching genes and also of many apparently meaningless DNA sequences within and between genes. The recombinant chromosomes that are parceled out to the germ cells at meiosis represent new combinations of these features. An allele from a locus on one chromosome and an allele from a different locus on the other, homologous

chromosome can be combined and passed on together; at the same time the alleles at two loci on a single chromosome can be separated, so that only one of them is inherited.

The closer together two loci lie on the same parental chromosome, the less often their alleles are separated as DNA is exchanged between homologous chromosomes during meiosis. Hence one can gain a measure of the distance between a gene of particular interest—one that has a disease-causing mutant allele, for example—and a marker by correlating the inheritance pattern of their alleles. If the individuals in an afflicted family who develop the disease almost always inherit the same version of the marker, the mutant gene and the marker must lie very close together on the same chromosome. The marker and the disease gene are said to be linked.

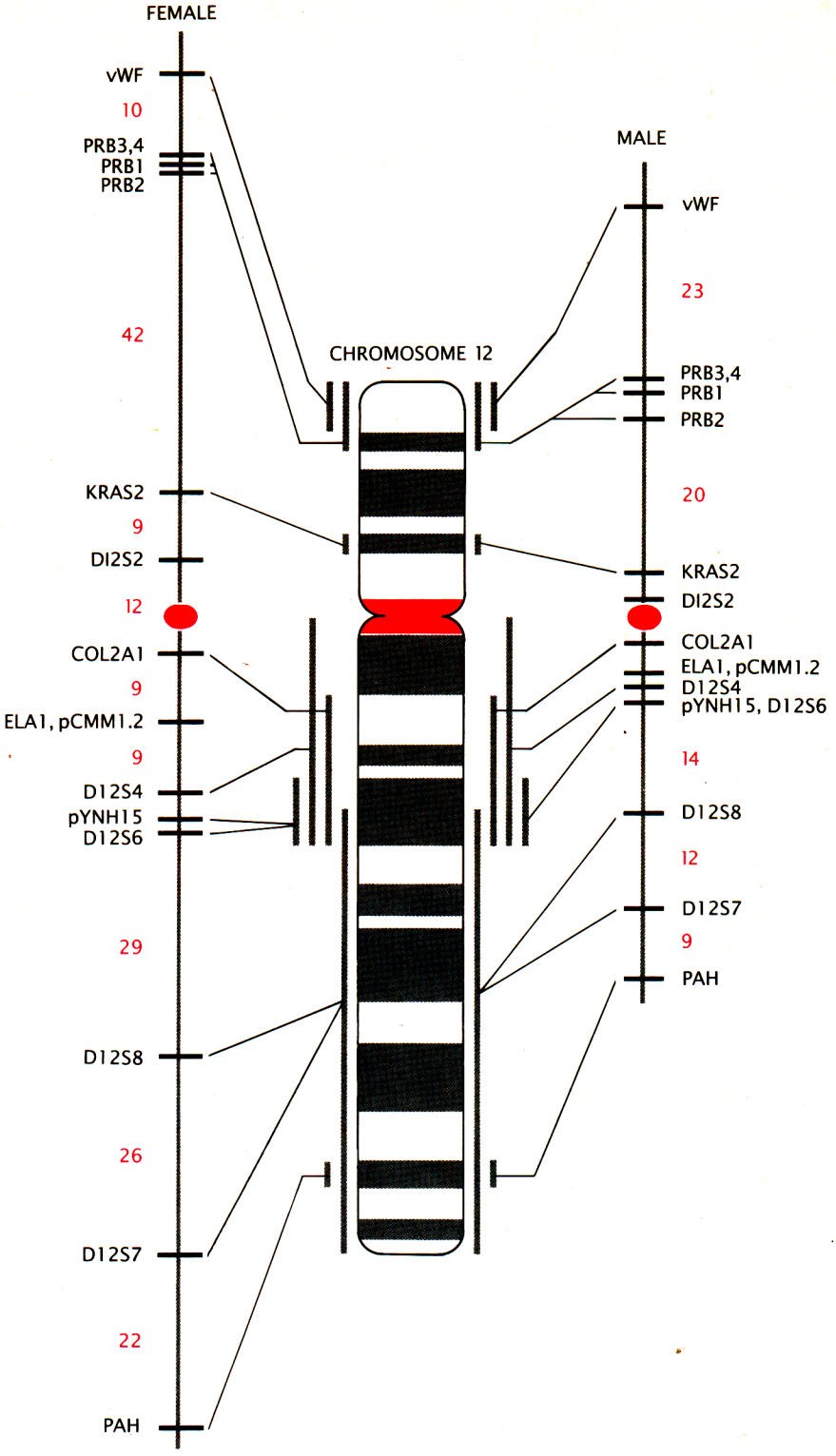
Other markers lying farther from the disease gene will recombine with the gene more frequently, so that the disease will be less likely to be inherited together with any given marker allele. In the extreme case, for a marker and a disease lying well apart on a chromosome, the recombination frequency reaches 50 percent. The marker and the gene are then unlinked: a given marker allele has only an even chance of being passed on with the disease. The same pattern of 50 percent coinheritance emerges when a marker and a mutant allele are borne on entirely different chromosomes.

Correlating the inheritance of a marker and a disease requires two things. The marker must be readily detectable, and it must be found in a number of distinguishable variants throughout the population. Linkage can be detected only if a person carrying mutant and normal alleles of a disease gene also carries two different versions of the marker; if the two marker alleles are indistinguishable, crossovers between the disease and the marker will be undetectable in the offspring. There will be no way to tell a linked marker from an unlinked one.

Until a few years ago only a limited set of markers met both criteria. The genes coding for certain enzymes, blood-group antigens (which determine blood type) and other proteins have multiple alleles, which manifest themselves by giving rise to protein polymorphisms: detectably different versions of the protein each gene codes for. Only 25 to 30 such marker systems of any value were known, however, covering only small sections of a few chromosomes. For want of markers most of the human genome remained inaccessible to the linkage approach.

With the advent of recombinant-DNA technology in the mid-1970's linkage mapping could be transformed into a practical and powerful tool for human genetics. The transformation can be dated to a genetics retreat sponsored by the University of Utah in April, 1978. There David Botstein of the Massachusetts Institute of Technology, Ronald W. Davis of Stanford University and Mark H. Skolnick of Utah proposed that the DNA sequence itself might yield numerous and readily detectable markers. Recognizing the potential power of the new approach, one of us (White) soon decided to test the hypothesis by committing his laboratory to the development of a set of DNA-based markers that would make it possible to detect linkage anywhere in the human genome. Botstein, White, Skolnick and Davis published the first paper detailing the approach in 1980. In the meantime

EXTENSIVE FAMILIES with living grandparents—modern counterparts to this turn-of-the-century family—are the ideal setting for studies of genetic linkage. In linkage studies the relative positions of sites in the chromosomes are inferred from the frequency with which genetic variations at those sites are passed on together from parents to children. By examining the inheritance of a genetic disease and arbitrary genetic markers in afflicted families one can assign a chromosomal location to the disease gene; by correlating inheritance of many markers in large, healthy families one can make maps of chromosomes.



MAP of chromosome 12 was made by tracing the inheritance of DNA markers: sites where the two copies of a chromosome often carry detectably different DNA sequences. The markers are arrayed in their statistically likeliest order and are separated by distances reflecting their recombination frequency, or the percent of the time marker versions carried on the same parental chromosome are separated by a recombination event during the formation of sperm or eggs. The recombination frequency between two markers rises with increasing physical separation, but the precise relation between recombination frequency and distance can vary depending on several factors, including sex. On chromosome 12, for example, the overall rate of recombination seen when the chromosome is passed on by a woman is higher than when it is passed on by a man, and so its genetic map is represented as being longer in women. An approximate chromosomal position has been determined for some of the DNA markers (*center*).

many other workers were beginning to find markers in human DNA and to speculate about their uses, and it was clear that this approach was an idea whose time had come.

The new linkage strategy gains its power from the very high level of normal polymorphism that can be found in the sequence of base pairs making up DNA. Between homologous chromosomes there is a difference in sequence, on the average, every 200 to 500 base pairs. Identifying these allelic variants would provide a practically limitless supply of markers scattered throughout the human chromosomes.

Molecular tools known as restriction enzymes provide a means of detection. Each restriction enzyme, made by a particular species of bacteria, binds to DNA wherever it finds a specific short sequence of base pairs and cleaves the molecule at a specific site within that sequence. A variation in DNA sequence that creates or eliminates a restriction site will alter the length of the resulting DNA fragment or fragments. The variation creates a restriction-fragment length polymorphism—an RFLP.

The RFLP defines a potential marker. A single restriction enzyme finds millions of cutting sites in the total human DNA, however. How can one or two variant fragments be detected among millions? The fragments are first sorted by electrophoresis: an electric field draws them through a gel, in which their mobility is inversely proportional to their length. A powerful and sensitive technique called Southern blotting after Edward M. Southern, who developed it at the University of Edinburgh, serves for picking out the fragments of interest.

Southern blotting relies on the unique character of the DNA molecule. The bases along two strands of DNA can pair only according to set rules, and so the sequence on one strand constitutes a unique match for the sequence on the other. A length of single-strand DNA can therefore act as a probe, detecting and binding to the complementary sequence in a sample of ordinary DNA that has been "denatured": heated or exposed to high *p*H in order to separate its strands. In Southern blotting the DNA fragments on an electrophoresis gel are denatured and blotted onto a membrane, where they are exposed to probe DNA labeled with a radioactive isotope. The probe hybridizes, or binds, only to the fragment or fragments that bear the complementary sequence of bases. The

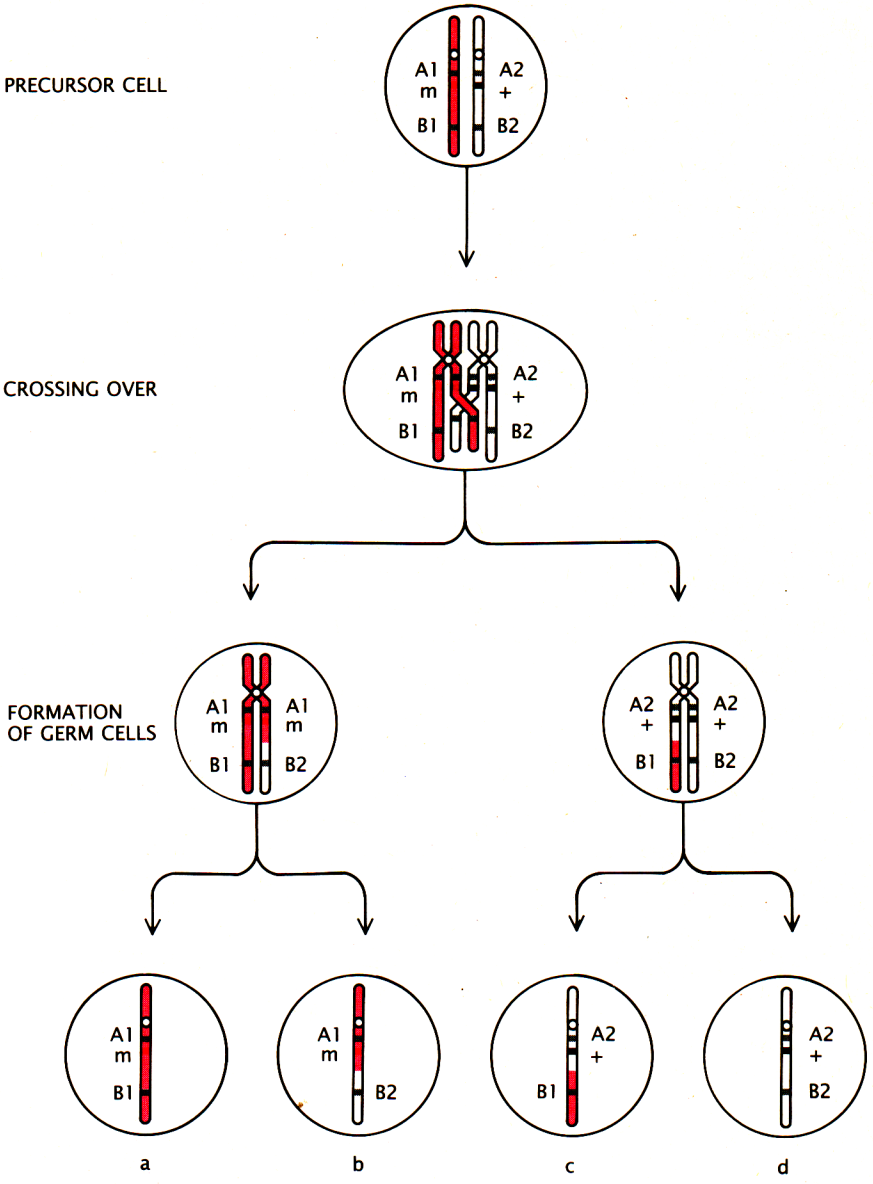
radioactive label makes it possible to detect the position of the fragments, which reveals their size.

To detect an RFLP, then, one needs to find a probe that is complementary to DNA near the restriction-enzyme cutting site. A segment of DNA is chosen, often at random, from a collection (a "library") of cloned DNA fragments representing the full human genome. It is denatured, made radioactive and applied to Southern blots of DNA samples that have been digested with a restriction enzyme. If the radioactive bands appear at different places on blots of DNA from different individuals, the cloned DNA has detected the variable cutting pattern that results from a DNA polymorphism. The probe and the RFLP it detects constitute a unique genetic marker system. With it one gains a point of reference in the genome: the short stretch of polymorphic DNA, whose inheritance pattern can now be traced.

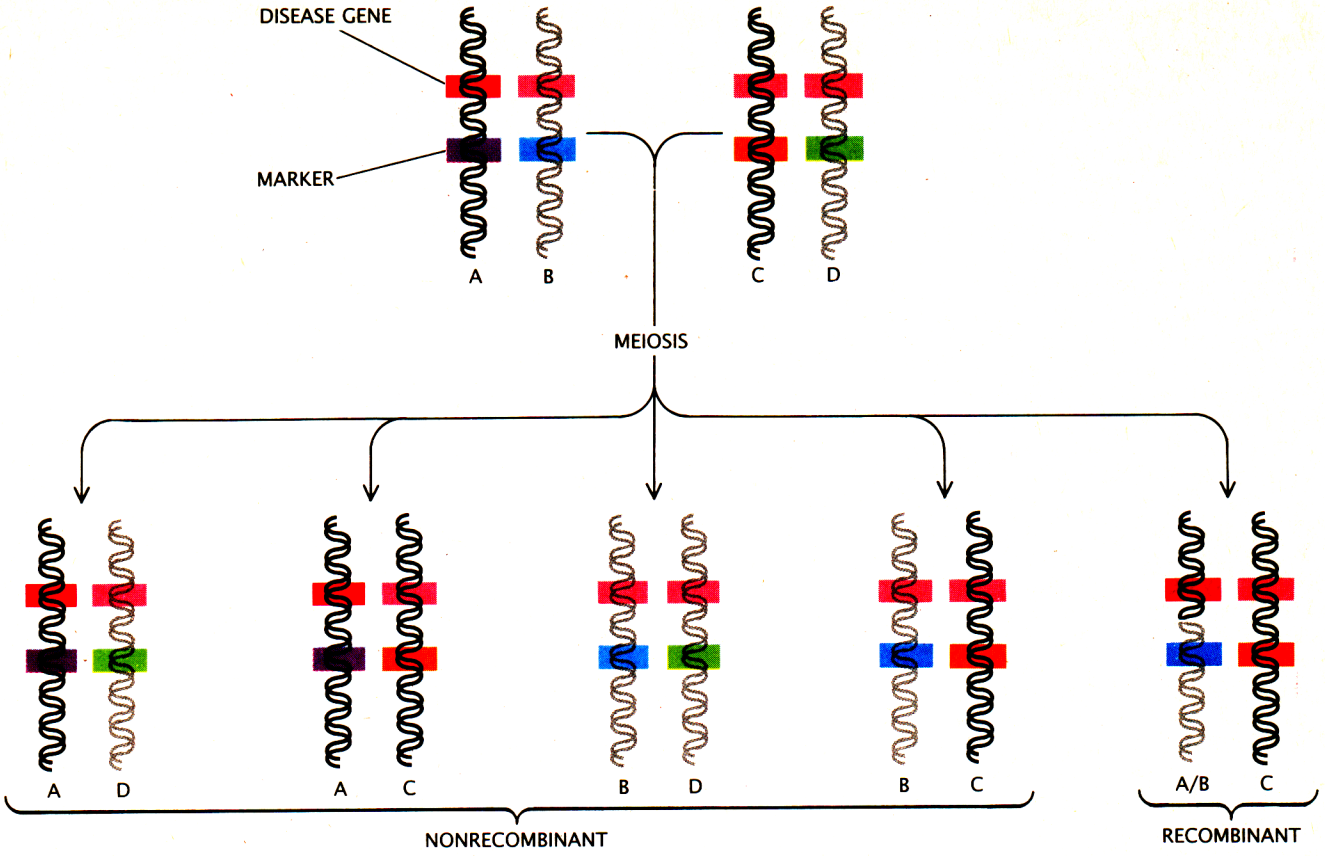
This DNA marker, defined by the RFLP, is found in one form or another in every individual, healthy or diseased. But if a genetic disease is passed down a pedigree together with a particular allele of the RFLP, the mutant gene can be assumed to lie in the same chromosomal region as the marker. In a second afflicted family the same marker will also show linkage, although the specific form of the marker that accompanies the disease may differ. Linkage to an arbitrary DNA marker reveals nothing about the physical position of the gene itself, of course, and for many purposes (such as diagnostic tests) physical location is immaterial. Nevertheless, the probe can also pick out the chromosome carrying the marker and the disease gene. If the probe is applied to a full set of human chromosomes, for example, it will hybridize to the chromosome bearing the marker site.

The value of any marker depends in large part on how many variants it displays throughout the population: the more versions of the marker there are, the more likely it is that an individual harboring a disease gene will carry two different alleles at the marker locus, making it feasible to detect recombination between the disease and the marker in offspring. Many RFLP's result from a change in a single base pair or the addition or deletion of a few base pairs at the restriction-enzyme cutting site. Such variation has a simple effect: the restriction site is either present or absent. The RFLP exists in only two forms, and so at least half of all individuals will probably be homozygous at the marker locus: they will carry the same variant on both homologous chromosomes. (Occasionally two restriction sites occur sufficiently close together to be detected by a single probe, yielding in effect a single marker with four alleles.)

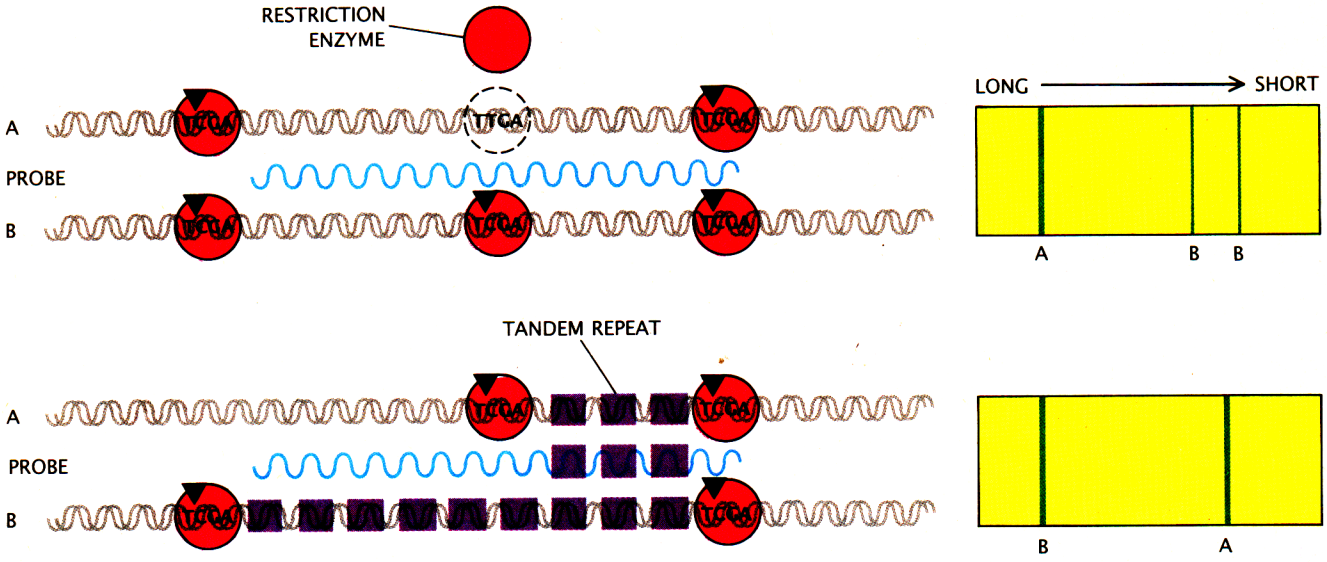
Another kind of DNA polymorphism creates many different versions of an RFLP. At many sites on the human DNA a single sequence that does not code for any protein is repeated many times. The origin and significance of these "tandem repeats" is a mystery, but for linkage mapping they offer a practical advantage in that the number of repeats at a given



RECOMBINATION makes it possible to detect genetic linkage. The diagram follows one idealized pair of homologous, or matching, chromosomes through meiosis, the process of cell division that produces germ cells (sperm or eggs). The chromosomes carry different alleles of two markers (*A, B*); one chromosome also bears a mutant, disease-causing allele of a gene (*m*) and the other chromosome bears the normal allele (+). In the precursor cell the disease is associated with allele *1* of both marker *A* and marker *B*. In the first phase of meiosis the chromosomes are replicated. The homologous chromosomes then "cross over," exchanging segments of equal length. Here crossing over takes place between loci *A* and *B*. The result is two germ cells (*a, d*) that carry the parental combinations of alleles and two (*b, c*) that contain recombinant chromosomes. In cell *b* the mutant gene is still found with allele *1* at locus *A* but is now joined by allele *2* at locus *B*. A low frequency of crossovers between the disease gene and marker *A* in many meioses would indicate that the disease and the genetic marker are closely linked.



LINKAGE between a disease gene and a marker is evident in the family history of the disease. Genetic features of a hypothetical couple and their children are shown. One parent suffers from a genetic disease caused by a single mutant allele (*red*); the other is healthy and hence carries only normal versions of the gene (*pink*). Children who inherit the disease usually also inherit a particular marker allele (*purple*) from the diseased parent, since the disease gene and the linked marker tend not to recombine.



DNA MARKERS—sites at which homologous chromosomes often differ in DNA sequence—are detected as RFLP's (restriction-fragment length polymorphisms). The DNA is digested with a restriction enzyme, which cuts wherever it finds a specific short sequence of nucleotides (in this case the base sequence *TCGA*). In one kind of marker (*top left*) a sequence difference causes a restriction site that is present on one homologous chromosome to be absent from the other. As a result the restriction fragments from each chromosome will differ in length. A DNA probe whose base sequence is complementary to that of DNA at the marker site reveals the fragments after they are sorted by electrophoresis (*top right*). Another kind of marker (*bottom left*) is characterized by a VNTR—a variation in the number of tandem repeats (short, repeated DNA sequences). The span between cutting sites differs between matching chromosomes, again resulting in distinctive fragments detected after electrophoresis (*bottom right*).

locus can vary from a few to hundreds of copies. Restriction fragments generated by cutting near these tandem repeats vary in length correspondingly [*see bottom illustration on opposite page*]; hence the RFLP comes in not just two forms but many. Given this variability in the population as a whole, the odds are good that a given individual will carry different versions of the RFLP on homologous chromosomes. A Southern blot will reveal two distinct fragments of different lengths, one from each homologous chromosome.

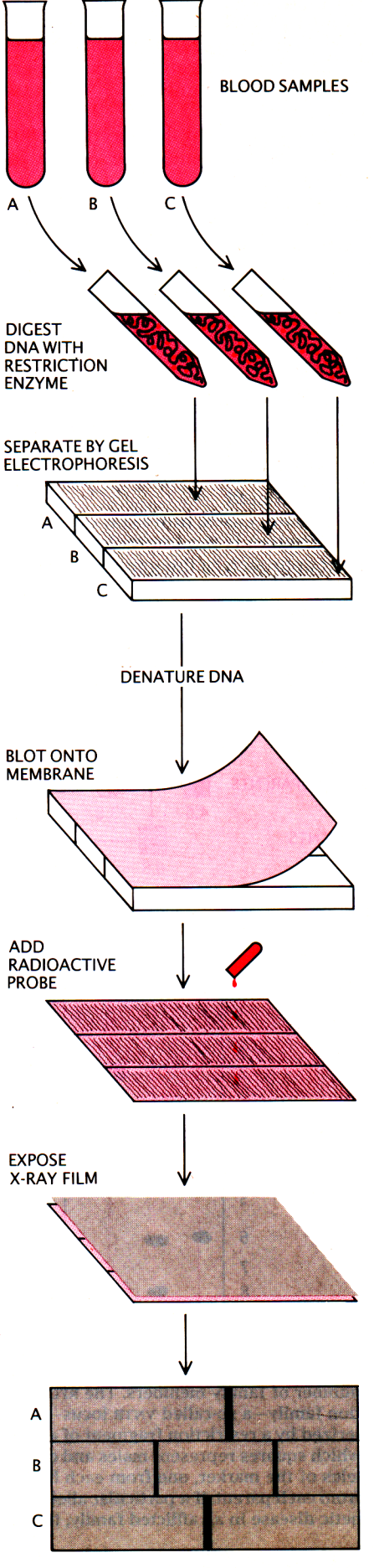
Probes for markers based on variations in the number of tandem repeats (VNTR's) can be developed more systematically than probes for ordinary markers. Alec J. Jeffreys of the University of Leicester recently recognized that the repeated sequences at many VNTR loci in different parts of the genome show similarities. The evolutionary explanation is again obscure, but the partial sequence homology means that under certain conditions a probe complementary to one VNTR locus can serve to pick out probes specific for other loci from a library of cloned DNA. Of the nearly 600 DNA markers developed so far in our laboratory, about 300 are VNTR's.

Such markers can serve as elements in an overall linkage map of the genome, or they can be developed for the more immediate purpose of tracking down a specific disease gene. Finding a marker whose inheritance is correlated with the appearance of a disease can be a staggering task on unmapped chromosomes. Since one often begins by knowing nothing about the chromosomal location of the disease gene or of any marker whose inheritance pattern is traced in an afflicted family, one can unwittingly search for linkage to tens of markers lying in a chromosomal region that is actually remote from the disease gene while leaving another, linked region unexamined. Nevertheless, the linkage strategy has already scored some remarkable successes.

If one knows which chromosome to search, the number of markers that must be tested can be reduced from several hundred, on the average, to as few as half a dozen. A genetic disease that almost always affects males but is inherited through the mother, for example, can be assumed to result from a recessive gene on the sex-determining *X* chromosome. (A mother carrying the disease has a second *X* chromosome bearing a normal copy of the gene, which masks the recessive disease gene; a son who inherits the mutation has only one *X* chromosome and therefore develops symptoms.) To find the gene, one need only test markers known to be carried on the *X* chromosome.

Genes for *X*-linked diseases were among the earliest to be traced through RFLP analysis; the first was the gene that causes Duchenne muscular dystrophy and probably also Becker muscular dystrophy (mapped by Kay Davies of the University of Oxford and Robert Williamson of St. Mary's Hospital in London). Yet an increasing number of diseases stemming from defects on the autosomes (the 22 pairs of nonsex chromosomes) have also yielded to the linkage strategy.

Huntington's disease became the first autosomal disease to be linked with a DNA marker when James F. Gusella of the Harvard Medical School and his colleagues studied afflicted families living in this country and near Lake Maracaibo in Venezuela. The group was fortunate in having to trace only eight markers through the families before finding one that was linked to the disease. Since then our laboratory and others have discovered markers for the genes causing disorders including cystic fibrosis, peripheral neurofibromatosis, or von Recklinghausen's disease (a disorder characterized by "café au lait" spots on the skin and a tendency to develop tumors and other disorders of the bone and nervous system), and familial polyposis coli (whose victims develop many colon polyps and run a very high risk of colon cancer).



RFLP ANALYSIS begins with a blood sample. DNA is extracted from the nuclei of white blood cells and digested with a restriction enzyme. The resulting DNA fragments are separated by gel electrophoresis, which sorts them in order of size. The RFLP is then detected by Southern blotting. First the DNA in the gel is heated to denature it, or separate its two strands, and is blotted onto a nylon membrane. A probe—a radioactively labeled segment of single-strand DNA that is complementary to the RFLP locus—is applied to the membrane. The probe hybridizes with the fragments from the locus; a sheet of X-ray film placed over the membrane detects the radioactively tagged fragments and thereby reveals which versions of the RFLP are present. In RFLP analysis of families, DNA samples from several individuals are often analyzed at the same time.

Tantalizingly, evidence of linkage has even been seen for forms of Alzheimer's disease and manic depression that run in families.

A "hit" can open the way to identifying the gene itself, which in turn provides a starting point for investigating the molecular mechanisms of the disease. By cloning the gene and determining its base-pair sequence one can deduce the composition of the protein it codes for and perhaps identify a specific defect. The protein can be synthesized and antibodies to the protein can be generated in experimental animals. Properly labeled, the antibodies can reveal the distribution of the protein in tissues affected by the disease. Such knowledge might hold the key to a treatment.

In many cases, however, the initial localization is too imprecise for a direct approach to the gene by current DNA technologies. The Huntington's disease gene, for example, recombines with its first identified marker at a frequency of about 5 percent, which implies that the marker lies as many as five million base pairs away from the gene. For identifying and cloning a gene the suspect stretch of DNA must be reduced to about a million base pairs, which means finding markers that recombine with the gene at a frequency of only about 1 percent. Ideally the markers will also flank the gene, bracketing the stretch of DNA to be tested.

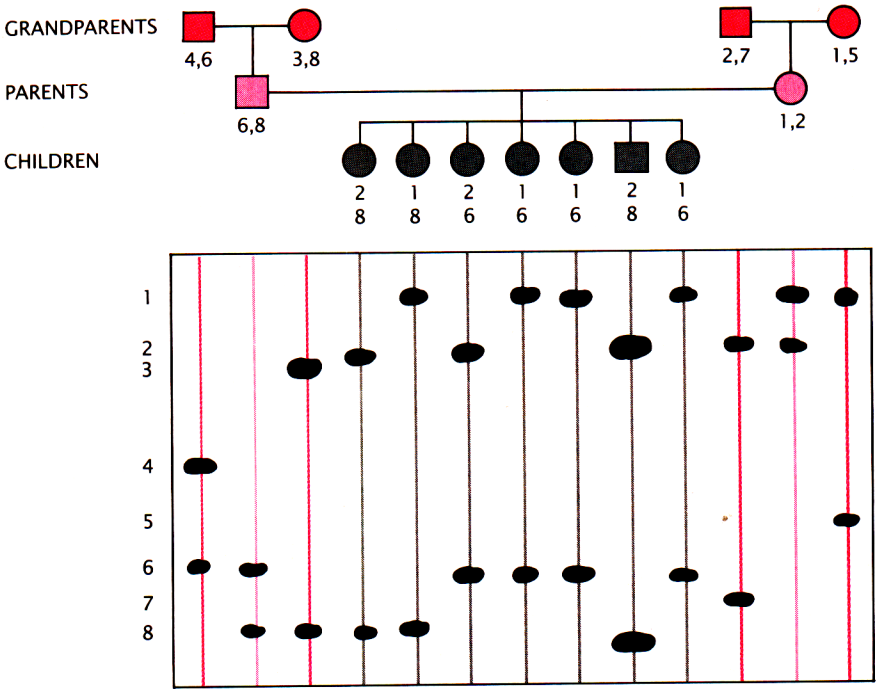
Tightly linked, flanking markers for cystic fibrosis, peripheral neurofibromatosis and familial polyposis are in hand, and a new, tightly linked marker has been identified for Huntington's disease. The search for the causative gene of each disease is in high gear. The approaches vary, but a common tactic is to comb a library of cloned chromosomal segments for one that is recognized by probes for both flanking markers. The segment—which presumably includes both markers and the gene they flank—can then be broken down further and each of the fragments cloned and tested for biological activity. Typically, a fragment can serve as a probe for messenger RNA (the sign that a gene is being expressed) in tissue affected by the disease. If it detects a messenger RNA that is unique to affected tissue, the probe itself may include all or part of the disease gene.

A different strategy has already culminated in the identification of the genetic defect in Duchenne muscular dystrophy. The region of the *X* chromosome that Davies and Williamson had linked with the disease shows missing segments in many patients; hence the disease may sometimes result from the absence of part or all of a normal gene. By identifying a region that is deleted in common among disease victims, Louis M. Kunkel of the Harvard Medical School and his colleagues were able to isolate and clone the gene.

Even before a disease gene is identified, linkage can sometimes point to possible causative mechanisms. The linked marker may fall near a gene of known function, which may then become a candidate for causing the disease. To take one instance, the marker for peripheral neurofibromatosis occurs on chromosome 17, which also carries the gene encoding the cellular receptor for nerve-growth factor (a substance that is vital for the survival and growth of nerve cells). That gene came under suspicion as a possible site of the mutation responsible for neurofibromatosis, but it was later found to lie some distance from the locus of the disease. Other genes on chromosome 17 may now become candidates for involvement in the disorder.

Reasonably tight linkage between a marker and a disease also makes it possible to devise tests for carriers and unborn victims—tests that are urgently needed given the frequency and insidious character of many genetic diseases. In populations of northern European descent, for example, one individual in 20 carries the cystic fibrosis gene. Because the gene is recessive, the carrier shows no symptoms, but if two carriers marry, their children stand a one-in-four chance of inheriting two defective genes and developing the disease. Huntington's disease is caused by a dominant gene (manifested even if the matching gene is normal), but its symptoms generally do not appear until middle age—after the unwitting victim has transmitted the disease to half of his or her children.

Before the presence of a disease gene can be established in an individual at risk, DNA from several other family members, both diseased and healthy, must be analyzed to determine which marker allele (or alleles,



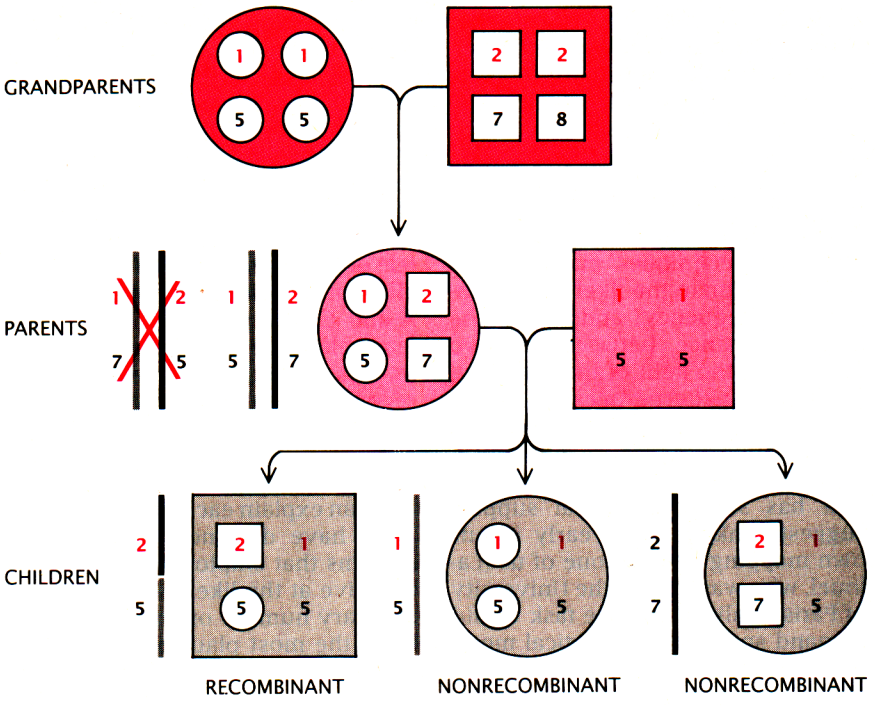
INHERITANCE OF AN RFLP can be traced by comparing restriction fragments from a number of family members. The RFLP marker that was analyzed in this three-generation family—a so-called VNTR locus—has many different alleles, each of them characterized by a restriction fragment of a specific size. Each individual in this pedigree (in which squares represent males and circles represent females) carries two different alleles of the marker, one from each homologous chromosome; children get one allele from each parent. If a particular allele of the RFLP is consistently associated with a genetic disease in an afflicted family, the marker and the defective gene may be linked.

in the case of a recessive disease that takes two copies of a gene to show itself) is inherited with the disease in this particular family. Finding a telltale allele in DNA from a prospective parent then indicates that he or she risks passing on the disease. Because DNA samples can be taken from a fetus soon after conception, the disease can also be diagnosed prenatally, enabling parents to make an informed decision about continuing the pregnancy. It is worth noting that in families at risk for some genetic diseases, fetal testing is actually increasing the number of births, simply because many couples would not conceive at all if they could not be sure the child was healthy before bringing it to term.

The construction of linkage maps, showing both arbitrary linkage markers and characterized genes arrayed along the chromosomes, has gone forward in parallel with the search for specific disease linkages. Linkage mapping represents a more deliberate and systematic approach to tracing mutant genes. From a complete linkage map workers trying to locate a disease gene will be able to choose and test a set of markers spaced at equal, large intervals along the chromosomes. Then, having discovered a linkage that restricts the gene's location to a specific chromosomal segment, they might test markers from a fine-scale map of the region in search of the tight linkage needed for further molecular studies.

The capacity to scan the genome for linkage not only will allow single-gene defects to be pinpointed more efficiently but also will hasten the search for the genetic bases of diseases caused by multiple aberrant genes. In addition, linkage maps will ultimately make it possible to check many points along the chromosomes simultaneously for a pattern of inheritance matching the family history of a disease, such as diabetes, coronary heart disease and certain cancers, to which susceptibility seems to be inherited. It might then be possible to close in on genes that confer predisposition to these illnesses.

Producing such a map extends the linkage strategy: now one is searching for linkage not between a DNA marker and a disease but between arbitrary DNA markers. Finding that alleles of different markers tend to be passed on together suggests the markers reside on the same chromosome, and the particular frequency



DATA FROM THREE GENERATIONS can solve genetic mapping's "phase" problem, posed by two markers on the same chromosome. Unless one knows the phase of two markers (*color and black*) in a parent—how their alleles (*numbers*) are distributed between the homologous chromosomes—one cannot unambiguously detect recombination in the children. Analyzing DNA from grandparents (the mother's parents in this case) can reveal which two alleles each grandparent contributed. Since the mother must have received alleles *1* and *5* on the chromosome she inherited from her mother, alleles *2* and *7* can be assigned to the matching chromosome from her father. The other configuration of alleles is thus ruled out, and a recombination event that has taken place in the mother's chromosome can be identified unambiguously in the first child.

with which the markers recombine reflects their "genetic distance."

Although linkage mapping is simple in concept, it presents an enormous bookkeeping and analytical challenge. A large-scale linkage map of the genome, sufficient to locate any disease gene within a span of between 10 and 20 million base pairs, would include between 100 and 200 evenly spaced markers. To have markers at even intervals, however, one must assemble a much larger set of random markers on the map. DNA must be collected from hundreds of individuals in dozens of large families and tested for the RFLP's characterizing each marker locus.

The analysis of these vast data sets is complicated by the fact that perhaps two-thirds of the markers in any individual are uninformative. They carry two identical alleles, with the result that linkage between the marker and any other locus cannot be detected in offspring. For two markers that may be linked, moreover, there is often no way to determine their "phase": how their alleles are distributed between the two homologous chromosomes. Without knowledge of which alleles are on the same chromosome in a parent, one cannot unambiguously detect recombination between the markers in the child.

These limitations are minimized when the data are gathered from extensive pedigrees. We have been fortunate in being able to draw on excellent family resources for our own mapping effort. For one thing, more than 50 Utah families with eight children or more have generously volunteered to give blood samples, from which we take DNA for analysis and establish permanent cell lines. The presence of many children means that the parents' chromosomes can be followed through a large number of meioses, giving more accurate estimates of recombination frequencies than could be had from families with few children. In addition almost all the Utah families we sampled have four living grandparents, whose DNA can often indicate the phase of markers in the parents. If one knows, for example, that allele 1

of marker *A* and allele 3 of marker *B* both originated in a grandfather, then his child—one parent—must carry both alleles on the same chromosome if the markers are linked.

Even so, the inevitable limitations in the data mean that the map must be founded on probabilities. Statistical techniques make it possible to estimate the likeliest recombination frequency, and hence the genetic distance, between any two markers in the light of the observed inheritance pattern. An estimated recombination frequency of 50 percent suggests two markers are unlinked; a smaller frequency—say 10 percent—that has strong statistical support suggests linkage. Very early in our own mapping venture one of us (Lalouel, who was then at the University of Paris) realized that the task would demand specialized statistical methodology and computer programs. He and his colleague Mark Lathrop designed algorithms and programs capable of both maintaining the huge data base and performing joint analysis of the inheritance patterns of many markers.

Having identified a set of linked markers, one still needs to determine their order along the chromosome. In principle one could calculate the probability of each possible order's giving rise to the observed inheritance pattern and choose the likeliest arrangement. As few as 15 linked marker loci, however, can be arranged in 6.5 times 10^{11} different orders, an impossibly large number. In practice one can quickly eliminate entire families of improbable orders by considering loci in subsets of, say, three at a time.

For a flavor of the reasoning, suppose that in a large family specific alleles of linked markers *A*, *B* and *C* are usually passed on as a group: a child inherits all or none of them. In one child, however, the original alleles of *A* and *C* are inherited with another allele of *B*; in a second child the original allele of *B* is joined by other alleles of *A* and *C*. The sequence *A-B-C* is the least plausible sequence because it implies that double recombination—recombination both between *A* and *B* and between *B* and *C*—took place in both cases. (Under either alternative order, *A-C-B* or *B-A-C*, one recombination can explain each observation.)

We have designed computerized systems that employ such strategies to arrive at the likeliest order for an arbitrary number of linked markers. Once the most plausible order for a cluster of linked markers has been established, they can be assigned to a specific chromosome, for example by hybridizing one of the marker probes to a set of intact chromosomes. Linkage clusters are thereby assembled into a chromosome map.

The genetic distances on a chromosome's linkage map are related to physical distances (numbers of base pairs), but the relation is by no means direct. We have found, for instance, that the recombination frequency of a given pair of markers often differs significantly between the sexes. That is, the probability that two markers on a chromosome inherited from the mother will have recombined during her meiosis may be quite different from the probability of recombination between the markers on the same chromosome inherited from the father. On chromosome 13, for example, recombination frequencies are several times higher in females. On chromosome 11 the opposite is true in one interval, and in an adjacent interval the two sexes show similar recombination frequencies. The molecular basis for these intriguing variations is mysterious, but as a practical matter we have been preparing two maps of each chromosome, one map for each sex, showing identical marker orders but different genetic distances.

We have completed preliminary maps for most of the human chromosomes. Another group has recently published a similar collection of preliminary maps, based on a smaller number of reference families. Yet linkage mapping is an inherently collaborative enterprise: every group is looking for landmarks on the same terrain. Markers developed and studied in one laboratory often complement markers from another laboratory, in some cases bridging gaps between linked clusters.

A framework for cooperation has been set up by Jean Dausset at the Center for the Study of Human Polymorphism (CEPH) in Paris. The CEPH has undertaken to collect, store and distribute DNA from 40 families. The collection draws mostly on our Utah families but also includes DNA from families studied by other workers. Investigators from around the world (including our own group) get complete sets of DNA from the collection; in return workers report their markers and inheritance patterns to the CEPH, which makes the data available to all investigators and so lays the groundwork for a single genetic map.

The completion, probably within the next few years, of a high-resolution map will consummate the transformation of the human genome from uncharted territory to well-surveyed ground. Such a map can be expected to yield precise locations for most of the remaining well-characterized genetic diseases. A complete linkage map will also prove invaluable for guiding another large-scale investigation of the genome, which is still in the planning stage: an effort to determine the complete base-pair sequence of human DNA. Small islands of DNA along the chromosomes will most likely be sequenced first. The linkage markers within each island will serve to locate it in the larger landscape of the genome.

DISEASE	CHROMOSOME	DATE
HUNTINGTON'S DISEASE	4	1983
DUCHENNE MUSCULAR DYSTROPHY	*X* (GENE CLONED)	1983
POLYCYSTIC KIDNEY DISEASE	16	1985
CYSTIC FIBROSIS	7	1985
CHRONIC GRANULOMATOUS DISEASE	*X* (GENE CLONED)	1985
PERIPHERAL NEUROFIBROMATOSIS	17	1987
CENTRAL NEUROFIBROMATOSIS	22	1987
FAMILIAL POLYPOSIS COLI	5	1987
MULTIPLE ENDOCRINE NEOPLASIA IIa	10	1987

TABLE OF DISORDERS gives a small sample of the genetic diseases for which the chromosomal location of the defective gene has been determined with the help of linkage studies. The table also indicates which chromosome carries the gene and the linked marker and gives the year linkage was first reported. Reasonably tight linkage can make the marker useful for diagnosing the disease in members of an afflicted family; very tight linkage can open the way to identification and cloning of the disease gene.