I: What Is the Genome Project?

Bob Moyzis: This discussion is meant to address scientists, particularly physical scientists, who know very little about the Human Genome Project and may have many misconceptions about it. Let's share our perceptions of how this project got started. Why are we doing it, and why did the idea of taking on the entire human genome gain support in the scientific community?

David Botstein: The answers are complicated because the human genome is largely unexplored territory. It's tremendously information-rich, and different people have had different ideas about the best way to go about finding out what's there. The initial proponents of the Genome Project, especially Charles DeLisi in the Department of Energy [DOE], said, "The human genome is the blueprint for the development of a single fertilized egg, into a complex organism of more than 10^{13} cells. The blueprint is written in a coded message given by the sequence of nucleotide bases—the As, Cs, Gs, and Ts that are strung along the DNA molecules in the genome. So let's read the entire sequence from one end to the other, put the whole thing in a computer, and give it to the theoreticians and computer analysts to decode the instructions." And what instructions does the human genome contain? Everyone who has taken high-school biology knows that DNA contains genes, that genes are the coded messages for making proteins, and that proteins carry out all of the functions of an organism. So why not begin by reading the sequence? Now many of us, including me, thought the straight sequencing approach was crazy because it ignores biology. Yes, we can read the sequence, pick out a gene, and use the genetic code to translate the coding regions of the gene into the sequence of amino acids that composes the protein. But then we run into a big problem: How do we know what the protein does? At present we have no way to determine the function of a protein from its amino-acid sequence alone. Wally Gilbert likes to say that if we had a catalog of all the protein amino-acid sequences, we would be able to deduce protein functions. Some day we may get there, but right now that's science fiction, not science.

Bob Moyzis: Interpreting protein function is a problem. But the straight sequencing approach, as initially proposed, presented other serious difficulties. First and foremost, the technology to sequence the whole genome was just not available. That was the conclusion of the human genome workshop sponsored by the DOE in 1986 in Santa Fe, New Mexico, and it is true today. We're not too bad at reading stretches of DNA 10,000 bases long—the average length of a gene—but present technologies are still too labor-intensive and too expensive to think of sequencing the 6 billion bases in the human genome. However, the technology is changing rapidly, a point we'll return to later. We're also not certain how to pick out the genes from all the other DNA sequences in the genome or how to separate the gene sequences into protein-coding regions, or exons, and noncoding regions, or introns. We're making progress, but the problems are still unsolved. On the other hand, most participants at the 1986 meeting agreed that a major effort in genetic and physical mapping was appropriate. That conclusion was confirmed by the report, published in 1987, of the DOE's Health and Environment Research Advisory Committee. Many individuals with a physical-science background do not understand that a DNA sequence *without* a genetic map is nearly useless.

David Botstein: Most of us were unaware of the DOE workshop and report, but the idea of understanding the human genome stirred up so much interest that the National Research Council organized its own committee to assess the feasibility of the Project. Some members of that committee are here—Maynard Olson, Lee Hood, and I. We independently recommended that the Human Genome Project go ahead—but, as Bob pointed out, in an entirely different manner than originally proposed. We said, "Let's postpone sequencing the genome until we develop better sequencing technology and focus on developing the tools, the genetic and physical maps, needed to interpret the sequence once we have it. Let's build some biology into this effort."

Bob Moyzis: But we still have a problem of perception in the scientific community. The conclusion of every meeting and report on the Human Genome Project has been that the goal is *not* to immediately sequence the entire human genome. That idea died an early death. But every negative report about the Project says that we are going to be doing this mindless sequencing.

Maynard Olson: Critics often do not take the time to understand what they are criticizing.

Norton Zinder

Until recently people tried to guess which protein from among the tens of thousands of human proteins was produced by the mutant gene . . . the new approach is to avoid playing around with lots of proteins and instead to find the responsible gene in the DNA.

Norton Zinder: I'd like to go back to an earlier point, that different people are interested in different aspects of the genome. The most ambitious interest is a very long-term goal—to understand the whole blueprint. But there's a large group of people, and maybe they're in the majority, who are more practical. They are interested in understanding human disease, and they support the Genome Project because the maps that will be developed are just the tools needed to find the genes responsible for inherited diseases. Victor McKusick has compiled a catalog of over 4000 such diseases and many of them are Mendelian, which means that they are each caused by a single mutant gene. People are very excited about the prospect of finding those genes.

Bob Moyzis: It's ironic that the genetic-mapping community had little to do, I feel, with initiating the Human Genome Project, recent books documenting the history of this project notwithstanding. Once the Project gained momentum, however, it was clear that the human genetic-mapping community would be a primary user of the maps, particularly in the search for the genes causing the Mendelian diseases that Norton just mentioned. Our audience may be surprised to learn that the method used to infer that a single gene is the cause of an inherited disease goes all the way back to Mendel. Despite all the advances we've made in molecular genetics, Mendel's laws and his indirect methods of inference still provide the basic methods for much of what is done in genetics.

David Botstein: Mendel identified the basic unit, the quantum, of heredity, which is the gene. Mendel's laws are the quantum mechanics of genetics. They provide a quantitative link between physical traits, the traits we see, and genetic traits, which are the unseen messages in the genetic material. In the case of humans, looking for Mendelian patterns of inheritance is often the only method we have for connecting phenotype with genotype.

Bob Moyzis: Mendel's laws apply only to discrete variable traits—for example having or not having unusually short fingers, a trait called brachydactyly. Because those traits [normal or short digits] are inherited according to the ratios predicted by Mendel, geneticists can infer a number of things. First, that digit length is determined by a single pair of genes, one inherited from each parent, and second that the brachydactyly gene has two versions, or alleles, say A and a, where A is the rare dominant allele that causes the anomalous digit length.

Most variable traits are not Mendelian. They result from the complex interaction of many genes. On the other hand, many inherited diseases *are* the result of a single mutant gene. How do we determine that? We can't do controlled-breeding experiments and analyze thousands of offspring as Mendel did. But if we trace the disease through the generations of families affected by the disease, we can use statistical analysis to infer from a relatively small sample whether a single gene-pair is involved, and if so, whether the mutant gene, the allele that causes the disease, is dominant or recessive. [For a discussion of Mendel's laws, see "Understanding Inheritance."]

Norton Zinder: Yes, but how do we go further toward understanding the disease? Until recently people tried to guess which protein from among the tens of thousands of human proteins was produced by the mutant gene. They would use various biochemical and cytological methods to compare normal and disease-affected tissues, but often

the disease gives no clue as to what proteins might be involved. The new approach is to avoid playing around with lots of proteins and instead to find the responsible gene in the DNA, sequence the gene, determine its protein product, and then try to determine what the protein does.

How do we find the gene responsible for a Mendelian trait? Until 1980 we had no practical method. Then David Botstein came up with a brilliant idea that's been used successfully to locate several of the more common disease genes and given great impetus to the Genome Project. The idea is based on a very old method for inferring the order of and relative distances between genes that lie along a single chromosome, what we call genetic-linkage mapping.

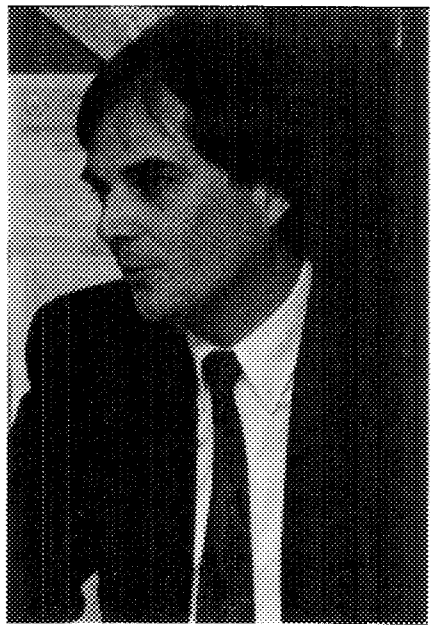
David Cox: Methods for constructing classical linkage maps are basic to what we are doing in the Genome Project, and again, they are an extension of Mendelian inference. Suppose we focus on two different Mendelian, or single gene, traits and trace the pattern of their co-inheritance from one generation to the next just as Mendel did. We may find that the phenotypes of two traits don't follow Mendel's law of independent assortment, but rather, that specific forms of those traits are almost always co-inherited. Statistically, that means the gene pairs for the two traits are linked and therefore lie on the same chromosome pair.

If we had a blackboard, we could show the particular type of mating, called the test cross, that reveals linkage between two different gene pairs. The gist of it is that if one parent is heterozygous for both traits—has the genotype *AaBb*—and the other parent is homozygous recessive for both traits—has the genotype *aabb*—then the combinations of the two traits in the offspring tell us whether or not the two gene pairs are on the same chromosome pair. [See "Classical Linkage Mapping."]

The interesting thing is that some fraction of the time the alleles—particular forms of the two genes—on a given chromosome are *not* co-inherited. How do they break apart? During the formation of either eggs or sperms, a pair of homologous, or matching, chromosomes can exchange corresponding chunks of DNA in a process called crossing over and thereby produce chromosomes containing new combinations of alleles. The recombinant chromosomes can then be inherited by an offspring.

How often do two alleles get separated by crossing over? It depends on how far apart they are. And that's the key to estimating the distance separating the two alleles. That distance is called the genetic distance. People have done many such linkage studies and constructed linkage maps giving the order of and genetic distances between genes that specify Mendelian traits and that lie on the same chromosome. The problem is that linkage analysis provides no way of locating genes on the chromosome itself.

Norton Zinder: The breakthrough in finding human genes was Botstein's idea to apply the methods of linkage not to variable physical traits that we see with our eyes but to variations in the base sequence of the DNA, that is, to variations in the spelling of the DNA. Variations in spelling are called polymorphisms, and they may occur anywhere along the genome—not only in the genes. The important point is that if the variations at some locus, some region, of a chromosome can be detected by a DNA probe, the region



Bob Moyzis

We are asked frequently whether the isolation of a disease gene immediately leads to a cure. Of course it does not, but without isolation of the gene, finding a cure is almost impossible.

becomes a DNA marker, that is, a variable DNA trait that can be traced through families in the same way we trace variable physical traits. [See "Modern Linkage Mapping with Polymorphic DNA Markers—A Tool for Finding Genes."]

In fact, we can construct a linkage map of DNA markers spaced throughout the genome provided we can find the appropriate probes. The search for DNA probes that detect variable loci is done at random and is very time-consuming. Once a probe for a DNA marker is found, however, not only can the marker be used in linkage analysis but also the probe can be used to find the physical location of the marker on the genome. And then we have a way of locating disease genes on the genome. Because if a disease is co-inherited most of the time with some marker, then the disease gene must be physically close to the site of the marker.

Bob Moyzis: There's a tremendous amount of effort involved in this approach, but it works. It's been used to find a number of disease genes, including the genes for cystic fibrosis and neurofibromatosis. That's why the first priority of the Genome Project, as outlined in the joint DOE/NIH five-year plan, is to construct linkage maps of polymorphic markers and furthermore to include enough markers on the linkage maps so that no two are very far apart. At the same time we will build physical maps consisting of cloned DNA fragments that cover the genome in a more or less continuous way, so we can locate the markers from the linkage maps on the DNA itself.

And once we integrate the physical maps and the linkage maps, we'll be able to find the genes related to virtually all inherited diseases, including multigenic diseases such as cancer and neurological disorders. That's the plan, and it's what we're doing right now. We're also developing more efficient technology for sequencing and applying that technology to the sequencing of million-base stretches of DNA.

Norton Zinder: Most people don't see this project the way we do. That's why there are so many misconceptions about it. This Project is creating an infrastructure for doing science; it's not the doing of the science per se. It will provide the biological community with the basic materials for doing research on human biology.

This Project is creating an infrastructure for doing science; it's not the doing of science per se. It will provide the biological community with the basic materials for doing their research on human biology. And the whole endeavor is technology-driven because getting 6 billion of anything is a hard job. At every level it is a bootstrapping operation.

The whole endeavor is technology-driven because getting 6 billion of anything is a hard job. At every level it is a bootstrapping operation. First, we have to improve the technology to do mapping and sequencing on a large scale, and then we have to do the mapping and sequencing.

Bob Moyzis: Norton, why don't you expand on what you mean by *creating an infrastructure for doing science.*

Norton Zinder: There are two kinds of biological science. The one most of us like to talk about—synthetic science—concerns topics like physiology, biochemistry, and biological function. The second is analytical science, which many of us take for granted. Analytical science answers questions such as: What is hemoglobin made of? How many disulfide bridges are in that protein? Does it have two amino-acid chains or just one? And answering such questions generates the technical means for doing synthetic science.

Now the Genome Project is analytical science. It will determine the structure of the genome down to the order of the nucleotide bases along the DNA molecule in each chromosome. Some biologists complain that not every base is important and that we are doing analysis for the sake of doing analysis. But careful analysis often leads to surprises.

Let me give you one beautiful example. No one knew that many proteins are initially made with a sequence of amino acids, called the signal sequence, that allows those proteins to be transported from the membrane where they are made—the endoplasmic reticulum—to other locations in the cell. The signal sequence is usually removed after the protein reaches its destination, so its existence was not detected. But when the RNA template for the protein hemoglobin was sequenced, we discovered that it coded for this extra sequence of amino acids not found in mature

hemoglobin. This one fact led to the whole theory of protein translocation, and it is the kind of discovery that will almost certainly come from sequencing the human genome.

Maynard Olson: Wally Gilbert is among those who say that the Genome Project isn't science because it's about improving the technology for doing things we already know how to do rather than about new ideas. But that's a rather naive view of what science is. As Sydney Brenner once said, "In molecular biology there are technical advances, discoveries, and ideas, and they usually occur in that order." Was von Leeuwenhoek doing science when he developed the microscope and realized how to use it for biology?

For more than a hundred years advances in biology correlated more closely with advances in optics than with anything else that was happening. As biologists could see better, they made discoveries about organisms, cells, and subcellular structures, and from these came more powerful ideas. We know science doesn't always work that way. Darwinism and Mendelism are counterexamples, where abstract ideas really led the way. But most of the time biology is driven forward by new technology.

Norton Zinder: I'm known to be overly cautious about predicting new technological developments, and at the moment we need new technology to meet the goals of the Genome Project. But during my forty years in molecular biology, I've learned to have great faith that when people start thinking about doing something, they're going to come through with a means of doing it and that means invariably opens up a whole world of new possibilities. Back in 1969 Gunther Stent wrote a book saying that we were at the end of the great discoveries in molecular biology. At that point we knew the genetic code and we knew that DNA was the genetic material. The next step was to learn how to manipulate DNA so we could study just how it really works, but there seemed to be no way of doing that because DNA molecules are so chemically monotonous—they are just long strings of four different nucleotides. Then came the discovery of restriction enzymes, enzymes that recognize specific nucleotide sequences and cut DNA at just those sites. And that changed everything because we had a way to break up DNA molecules in a reproducible way. Questions we couldn't conceive of even asking suddenly became accessible to study.

Bob Moyzis: The discovery of restriction enzymes started the recombinant-DNA revolution in the 1970s. I was a graduate student at Johns Hopkins University when pioneers like Hamilton Smith isolated the first restriction enzymes. Smith later received the Nobel Prize for his work, and this was an incredibly exciting time at Hopkins.

Using restriction enzymes, it became possible to cut pieces of DNA from, say, mouse, and combine them with a piece of bacterial DNA. One could then propagate that recombinant DNA molecule in a host organism, usually the bacterium *E. coli*, and then either harvest the recombinant clones for further analysis or study the expression of the foreign DNA insert in the host organism. So restriction enzymes turned out to be a tremendous breakthrough.

Norton Zinder: I had the good fortune to experience the impact of a technological breakthrough firsthand because it was a breakthrough in which I actually participated. It was 1948, and I was a graduate student working on the genetics

Maynard Olson

For more than a hundred years advances in biology correlated more closely with advances in optics than with anything else that was happening . . . We know science doesn't always work that way. Darwinism and Mendelism are counterexamples, where abstract ideas really led the way. But most of the time biology is driven forward by new technology.

Norton Zinder

During my forty years in molecular biology, I've learned to have great faith that when people start thinking about doing something, they're going to come through with a means of doing it and that means invariably opens up a whole world of new possibilities.

of *E. coli*. At that time it was almost impossible to make new bacterial mutants, and without new mutants, geneticists can't work. The standard practice was to irradiate the bacteria and test them, one at a time, for some new trait. The type of trait we were looking for was a biochemical defect that would affect their ability to grow in the absence of some growth factor. Unfortunately, almost all the bacteria would die, and in a month's work, you would find maybe one mutant. Well, the day after Joshua Lederberg and I thought of using penicillin as a negative selection factor for mutants, we had more mutants than we could ever analyze in our lifetimes.

Maynard Olson: Let me fill in Norton's story. The idea was to deprive the bacteria of a growth factor, say a certain amino acid. Since normal, or wild-type, bacteria manufacture all the amino acids, they would continue to grow. But penicillin was known to kill only growing cells. So when you apply penicillin to the culture, it kills the wild-type bacteria, whereas the mutants that stopped growing because they didn't manufacture the amino acid would sit there in a latent state, unaffected by the penicillin. Then you washed the penicillin away and isolated the new mutants.

Norton Zinder: From that moment on all of the intermediary metabolisms of *E. coli*, that is, all the biochemical steps needed to synthesize important chemical compounds, became accessible to study, and bacterial genetics moved forward in ways that led us to understand a great deal about how genes really work. It led, for example, to my discovery of bacterial transduction, which is the introduction of genes from one bacterial mutant into another by a bacterial virus. Bacterial transduction is a natural progenitor of recombinant-DNA technology.

Maynard Olson: We need to remind ourselves that when Norton was doing those experiments, molecular biology was barely a field. Only a few people like Norton, with eclectic interests in microbiology, biochemistry, physiology, and so on, were thinking about biological processes in a new way and trying to understand their origins in the genetic material. But recombinant-DNA technology has had a huge impact on the way biologists work because it enables almost anyone to study DNA. The field of molecular biology is now defined by a certain experimental paradigm, and people interested in population genetics, developmental biology, protein chemistry, or whatever are all, in a sense, molecular biologists. They all search for answers at the level of the DNA. And they all use more or less the same experimental techniques. You take DNA out of cells, find out something about it, change it, put it back into cells, and then you see how the cells work differently. That's the basic paradigm.

Norton Zinder: Molecular biology is a powerful approach because all of biology starts from genes. I'm not saying genes are everything, but without them you don't get very far. That's why our colleagues, whether they are molecular biologists, neurobiologists, or students of African killer bees are all trying to locate and clone the genes relevant to their interests. When the Genome Project delivers these global maps of the human genome, the search for human genes at least will be a lot easier.

David Botstein: It's worth expanding that point. Our recent success in isolating human disease genes has made everybody optimistic about the usefulness of the Human Genome Project. But those genes were found one at a time. Once we have the linkage maps of highly polymorphic markers and the

physical maps of ordered, cloned DNA fragments, the search for disease genes will become routine.

The first step in isolating a disease gene will be to trace the markers one at a time through several generations of a family or families affected by the disease. The markers that are inherited most often with the disease are physically closest to the causative gene. After identifying markers that flank the region containing the gene, you find the markers on the physical map, pick out the DNA between the markers, find the gene in the DNA, read the sequence, and use the genetic code to translate the base sequence of the gene into the amino-acid sequence of the protein.

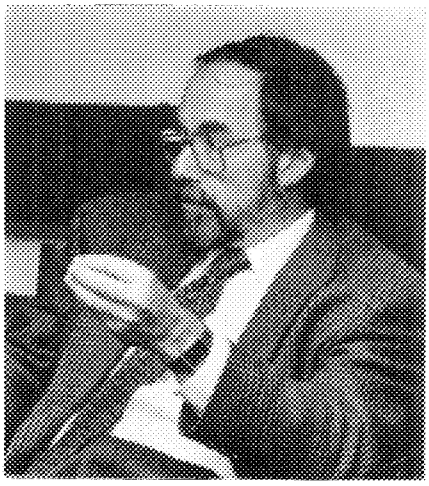
Now I said earlier that we have no way of deducing the function of a protein from its amino-acid sequence. But sometimes there is an empirical way. The sequence may be similar to the sequence of another protein whose function is known, and almost without exception that other protein is in a simpler model system—either yeast, or *Drosophila*, or something else that you can study in the laboratory. That is the reason mapping and sequencing the genomes of nonhuman organisms are part of the Human Genome Project.

We can figure out the function of a human gene by analogy with the function of a similar, or homologous, gene in an experimental organism. For example, we found that the gene responsible for muscular dystrophy codes for a protein that is similar to certain cytoskeletal proteins that have been well studied in a number of organisms. The gene for cystic fibrosis is similar to the multidrug-resistance gene, which had been studied to death in some systems and could be recognized immediately. The gene for neurofibromatosis codes for a gap protein that had been studied even more than the preceding two and whose mechanism of action is quite well understood.

Bob Moyzis: Before those genes were found, little was known about the causes of the diseases at the molecular or biochemical level. But after isolating a disease gene, finding another gene of known function, and identifying the mutation in the DNA responsible for the disease, one can then begin to identify the molecular mechanism of the disease and begin to design a therapy to counteract the defect caused by the mutant gene.

We are asked frequently whether the isolation of a disease gene immediately leads to a cure for the disease. Of course it does not, but *without* isolation of the gene, finding a cure is almost impossible. For example, our chances of combating the AIDS virus would be very slim if its genome had not been isolated and sequenced. With that information in hand, rational drug treatments to inhibit viral replication can be devised and tested.

Another informative example is muscular dystrophy. For over twenty years various drug treatments were tested on what was considered an animal model system for muscular dystrophy, namely, mutant chickens that exhibited similar muscle degeneration. Once the muscular-dystrophy gene was identified, it was discovered that the physical defect in the chickens was completely unrelated to the physical defect in humans. Hence, all those years of drug research were of little value. A mouse mutant with the mouse homolog of the muscular-dystrophy gene, however, has now been identified. Ironically, that mutant had been known for years, but it was unrecognized as a muscular-dystrophy



David Baltimore

The only way to study the genetics of the higher perceptual and integrative human functions is by studying human beings. We can't study the genetics of human beings in the way biologists like because you can't mate them in a controlled way. So we have to get the information we need out of natural matings. The linkage and physical maps will help us do that.

David Botstein

There just isn't enough information in noncontrolled crosses between humans to pinpoint the genes involved in very complex traits. For that you need model systems. And that's precisely why mapping and sequencing the genomes of model organisms is an integral part of the Human Genome Project.

mutant until the human gene was isolated. Now, because the underlying molecular defect is known, rational drug regimes can be tested on the new animal model system.

David Baltimore: I'd like to point out that investigators were searching for disease genes and finding them long before the Genome Project existed. We were looking at homologies between DNA from humans and model organisms. No one needed a new Project to continue doing what we were doing before.

But the Genome Project is something quite different because it will allow us to examine human variability, for example, variations in mathematical ability or in what we call intelligence. Those variations are caused by the interaction of many genes. And certainly the best way that biologists have to unravel which genes are involved in complex traits is to find a set of markers that are linked to the disease and then find the genes associated with those markers. In other words, we need the linkage maps and the physical maps that will be generated by the Human Genome Project. Those maps will allow us to do new kinds of science.

I am particularly *uninterested* in the sequence of the entire human genome because I believe that level of detail is not very useful. But I'm very interested in studying the genome at a level where we can get at multigenic traits and at subtle aspects of human genetics. That is why we are mapping the human genome rather than the mouse genome, and the rationale for doing so should not be to find human disease genes, because we're doing moderately well at finding them right now.

But the only way to study the genetics of the higher perceptual and integrative human functions is by studying human beings. We can't study the genetics of human beings in the way biologists like because you can't mate them in a controlled way. So we have to get the information we need out of natural matings. The linkage and physical maps will help us do that. So I believe that the Human Genome Project will open up an entirely new level of human biology. To my mind that is the only reasonable rationale for the whole program.

David Botstein: With some claim to proprietorship of the method you are describing for studying multigenic traits, let me say that without some organized effort like the Genome Project, we can't even find the genes for single-gene diseases in an efficient way. But because the Human Genome Project exists and the maps are being made, people are having the courage to set up relatively simply experiments on multigenic traits.

One experiment, proposed by Jasper Rine of the Berkeley Genome Center, involves selecting dogs with different behavioral characteristics, treating those characteristics as multigenic traits, and figuring out by experimental matings what genes are involved. Human genes similar to those genes will be identified and studied to see whether they are involved in determining similar behavioral characteristics in humans. We can't do that without the experimental work on model organisms. There just isn't enough information in noncontrolled crosses between humans to pinpoint the genes involved in very complicated traits. For that you need the model systems. And that's precisely why mapping and sequencing the genomes of model organisms is an integral part of the Human Genome Project.

David Baltimore: I'm not arguing against model systems. My point is that

the Genome Project will allow us to study complex traits that are specific to human beings, something we couldn't do before.

David Galas: Yes, the Genome Project will allow us to examine human variability and complex human traits, but that's only one of the reasons for doing this project. Although human disease genes are only a small fraction of the information in the human genome, they are very important to society, and the time has now come when it doesn't make sense to continue chasing individual genes. Just look at the funding history of cystic fibrosis. It cost over $100 million to find that one gene and took eight years of prodigious effort.

David Cox: The others we've found have been just as time-consuming and expensive. Each one has cost many, many millions of dollars. So to say we're doing moderately well with disease genes misses the point.

David Galas: We would spend much more money trying to find disease genes one at a time than we are going to spend on the entire Genome Project.

Bob Moyzis: I agree. Having participated in both the cloning of single genes and the mapping of entire chromosomes, I would estimate that the Human Genome Project is a hundred times more efficient. Further, the Genome Project will result in the identification of very rare disease genes. Such orphan genes, like orphan drugs, will never receive the funding needed for their isolation. But a *complete* map will make it possible to isolate all disease genes efficiently, including orphan genes.

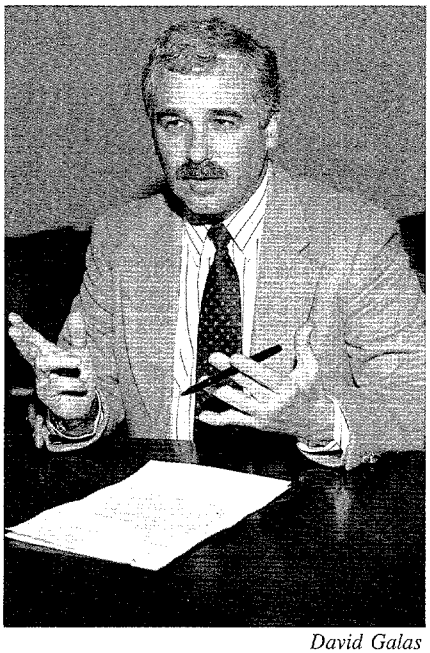
David Galas: We're going from targeted hunts for individual genes to a search for all the genes, which can then be studied one by one. It's a change in the paradigm for gathering information about genes, and it's much more efficient. If you're a guy who wants to study a particular gene, you won't have to first map the region, find the gene and sequence it. Instead, all that information will already be available.

Bob Moyzis: It's a paradigm shift, however, that's threatening to some investigators. They do not like the *perceived* loss of control. They should realize, however, that the tools that will come out of the Genome Project will serve to liberate their research.

David Cox: Of course! Then people will be able to spend their time studying the biology, not isolating the genes. The Genome Project will provide the maps and the sequences, and those raw materials will be used not only to understand human diseases, but also to study much more global biological questions about complex disorders involving many genes and about the interaction of genes with their environment. We'll be able to study how different genes are turned on and off in different tissues and at different times, and we'll study the developmental processes that turn a fertilized egg into a mature organism. But first we have to get the raw materials.

David Botstein: Everybody agrees that the physical maps and the linkage maps will revolutionize a certain kind of genetics, and the major emphasis of the Genome Project during its first five years is to make those maps. But if we get only that far and don't go down to the level of the DNA sequence, we will have missed a great fraction of the possible benefit of the Project.

We need to know the sequences of many, many genes if we are ever to be able to predict the function of a protein from



David Galas

Although human disease genes are only a small fraction of the information in the human genome, they are very important to society, and the time has come when it doesn't make sense to continue chasing individual genes. Just look at the funding history of cystic fibrosis. It cost over $100 million to find that one gene and took eight years of prodigious effort.

its DNA sequence or to understand the bigger picture of how genes are organized and regulated.

My favorite analogy with physics is spectroscopy. We're now cataloguing genes just like Fraunhofer catalogued atomic spectra. He had no idea what the lines meant in physical terms, but he knew they were important. And people made their living measuring fine structure, and hyperfine structure, and *superhyperfine* structure—not that such a thing exists—for different elements in the periodic table. But none of what all that information meant got worked out until a theory of the atom was developed, until Bohr and Schrödinger and those guys developed quantum theory. All of a sudden everybody said, "Aha, I can explain those lines because the atom has such and such a structure."

In much the same way, we're collecting the spectra, the sequences, of different genes, but the long-term goal of biology is to determine the functions of those sequences, that is, to understand as much as we can about the information encoded in the genome of the fertilized egg.

David Baltimore: A significant part of the biology community does not believe that sequencing the entire genome is the way to reach such an understanding. That's one of the reasons why the Genome Project is so controversial.

David Botstein: Perhaps I should explain why sequencing the entire genome is a controversial issue. As far as we know now, the informative part of the genome—the part that codes for proteins—is a small fraction of the total genome. Much of the DNA is junk, or of unknown and maybe unimportant function. The arguments that a large fraction of the DNA is relatively unimportant exist and are pretty convincing. Most reasonable people estimate that the protein-coding regions compose on the order of 10 percent of the genome. The 10 percent I'm referring to are the bits of information in the information-theory sense—the exons. You can strip a human gene of its introns and insert only the exons into a bacterial cell, and the stripped gene functions, that is, makes a protein. That's been the result for all the human genes tried so far.

My favorite analogy with physics is spectroscopy. We're now cataloguing genes just like Fraunhofer catalogued atomic spectra. He had no idea what the lines meant in physical terms, but he knew they were important.

Probably the great majority of biologists would initially say, "It makes obvious sense to sequence the informative bits first because sequencing with current technology is very expensive, laborious, and boring." But before the informative bits can be sequenced, they must be found. So the choice about the approach to sequencing the human genome is really not obvious. It depends on the answer to a technical question: Is it more expensive to figure out which are the informative bits and then sequence them, which is our current approach, or to sequence the entire genome and then find the informative bits? The first five-year plan of the Genome Project is agnostic on this issue. It says, "We want to develop the technology for faster and cheaper sequencing as quickly as we can, and we are supporting pilot sequencing projects that lead in both directions." The compromise between the "let's go out and get every nucleotide" gang and the guys who thought that the idea was nuts was to say, "We're going to postpone most large-scale sequencing, and depending on how far we get in improving technology, we'll decide what approach to take on the human genome." Sequencing is the area that really needs some breakthroughs. If sequencing were about a hundred times cheaper or a hundred times faster, then it wouldn't make any sense not to sequence the whole genome.

Bob Moyzis: We'll return to the prospects for getting that hundredfold improvement in sequencing a bit later, but now I'd like to counter the notion that most of the genome is junk. Even if exons make up only 10 percent of the genome, that doesn't mean the other 90 percent of the genome is totally superfluous, that you can get rid of it without any effect. Remember that a few hundred years ago a lot of physiologists said the brain was useless because they had no idea what it did. The history of science is full of such statements.

I've spent several years identifying and cloning the human telomere, and we're now attempting similar work on human centromeres. Those regions don't code for proteins, but they're not junk. The telomeres ensure the stability of the chromosomes during DNA replication, and the centromeres are involved in the proper parceling out of the chromosomes during cell division. Unequal parceling out, or aneuploidy, is the major cause of both embryonic abnormalities and metastatic cancer. All other genetic defects added together do not add up to the human suffering caused by aneuploidy. Similarly, the regulatory

regions necessary for controlling gene expression are not junk. They compose a significant fraction of the DNA and are often far removed from the genes they regulate.

I think the non-protein-coding regions are the most interesting regions of the genome because they are the regions that make it all work. There are many DNA codes other than the protein code, and determining the other codes is probably the most basic scientific justification for the Human Genome Project. It seems to me that when people say that 90 percent of the genome is junk, they really mean that those regions are uninteresting to their area of research. If you are interested in how proteins fold or how ions pass through cellular membranes, then the primary amino-acid sequences of the proteins encoded in DNA are probably the only aspect of the Genome Project that will interest you. Those are important and exciting areas of research, and the functioning of chromosomes is likely to shed *little* light on the answers. However, I believe that no molecular biologist interested in understanding how the genome works—how genes are differentially expressed in different tissues, for example, or how deletion of information causes genetic diseases—thinks the answers are only in the protein-coding regions. To quote Mary Lou Pardue, "One person's junk is another person's collector's item."

David Botstein: Okay, Bob, your point is well taken, but I think everybody is in agreement that no one's going to sequence from one end of the human genome to the other given current technology and the uncertainty about the function of most of the genome. The technology just isn't there to do it.

Right now, the Genome Project is funding a few large-scale sequencing projects, that is, projects to sequence continuous stretches of DNA from one million to several million bases in length. Sequencing such long stretches has never before been attempted. But we are not sequencing any old stretch of DNA but rather are focusing on model-system DNA, which can be interpreted fairly easily, or on stretches that encompass well-studied families of genes such as the HLA complex, or on cDNAs.

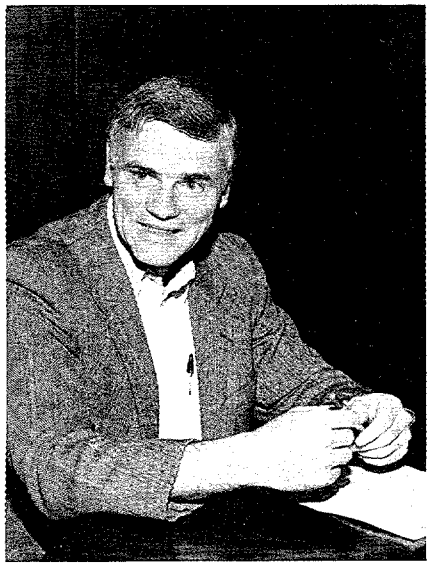
Lee Hood: It's also necessary to support some biology along with the mapping and sequencing. Some of us at Caltech applied to both NIH and DOE for a grant for large-scale sequencing, and they both argued that we shouldn't do any biology as part of the Project. Well, the fact is that you're not going to get any good people to do the sequencing if you're not going to let them do any biology on the sequences they generate. It's insane to think that good laboratories are only going to sequence and not do anything else. They may take the money for sequencing, but they will end up spreading it around doing other kinds of things.

At Caltech we are sequencing the regions in the human and mouse genomes that code for the proteins of the immune system that recognize foreign antigens. Those proteins make up the receptors on the surfaces of T-cells. The T-cell receptor genes of the mouse and humans combined encompass between 6 million and 7 million base pairs of DNA. We've already sequenced close to 500,000 base pairs of that DNA.

We plan to set up a group whose primary purpose will be to push hard on sequencing as much DNA as possible. It will be a core of technicians managed by a senior postdoctoral fellow interacting with a group of more junior postdoctoral fellows interested both in sequencing

David Botstein

The approach to sequencing the human genome . . . depends on the answer to a technical question: Is it more expensive to figure out which are the informative bits [the protein-coding regions] and then sequence them . . . or to sequence the entire genome and then find the informative bits?



Lee Hood

The most widespread criticism is that the Project is taking away from other aspects of biological science and especially away from individual investigators . . . On the other hand, people don't seem to remember that the Genome Project is less than 1 percent of the total NIH research budget.

and biology. Also, as we do the large-scale sequencing itself, we will learn what new technologies need to be developed to get the job done efficiently. So the biology and the development of efficient sequencing technology will go hand in hand with the large-scale DNA sequencing.

David Cox: We have many different strategies for mapping and sequencing, and what the Genome Project is about right now is determining the most effective way to use them. The biological community has long been familiar with cloning DNA, making maps of restriction sites, and sequencing DNA, and those technologies are steadily being improved. So ultimately the entire human genome is going to be mapped and a large fraction of it sequenced. The issue is efficiency.

The money spent cloning and sequencing is a significant fraction of every laboratory's budget. If the maps and the cloned DNA were available, biologists could spend their time studying how the gene relates to the biology, and the science would move along much more efficiently and rapidly. So the rationale of the Genome Project is to put a lot of money up front into getting the maps of the human genome and thereby free up the rest of the scientific community to do biology. From a business point of view the Genome Project makes a lot of sense.

Norton Zinder: And the only way we're going to accomplish the goals in a reasonable time is through a targeted program. The goals are to develop the technology for mapping and sequencing the human genome and then to do the mapping and sequencing. It's as simple as that. It just takes work and money. The question is: How much work do we want to put in and how much money?

Bob Moyzis: Most reports, including that of the National Research Council's recommendation to Congress, indicated that $3 billion spread out over 15 years, which amounts to $200 million per year, was appropriate. If we reach that level of funding, it will be enough to generate the maps, but I question whether the necessary technology developments as well as the transfer of technology to industry can be accomplished within that budget.

The information from the Genome Project needs to be used for individualized medical diagnosis, and so we need to develop rapid, efficient ways to screen millions of people for hundreds of genes. Yet I see little current support for accomplishing that goal. Lee Hood is one of the few individuals thinking about and working on this problem. But still, by the standards of the biological community, the Project's current funding—$57 million from the DOE and $105 million from the NIH—makes it seem very much like big science, and as such it's been a target for criticism.

Lee Hood: The most widespread criticism is that the Project is taking money away from other aspects of biological science and especially away from individual investigators. That concern has not softened too much because the NIH isn't funding grants at very high levels and people feel the pinch. On the other hand, people don't seem to remember that the Genome Project is less than 1 percent of the total NIH research budget.

David Cox: From a psychological point of view the Project has led to a terrified scientific community. Researchers are saying, "Wait a minute. What am I going to do while you're making that map if I'm not getting any money to do my research?"

Bob Moyzis: There's also the fear that the Human Genome Project will stamp out the creativity of the individual researcher, that because it is a large project it will destroy the sociology that has produced so many dramatic advances in molecular biology over the last fifteen years. The Project requires a lot more coordination than biologists are accustomed to.

David Botstein: The goal is too big for standard cottage-industry science. We need to be able to think about the whole genome at once, and that requires more organization than we usually have. As Norton said, we need a targeted effort. The nice thing is that this large effort doesn't have to be on one piece of real estate. It can be, but it does not need to be.

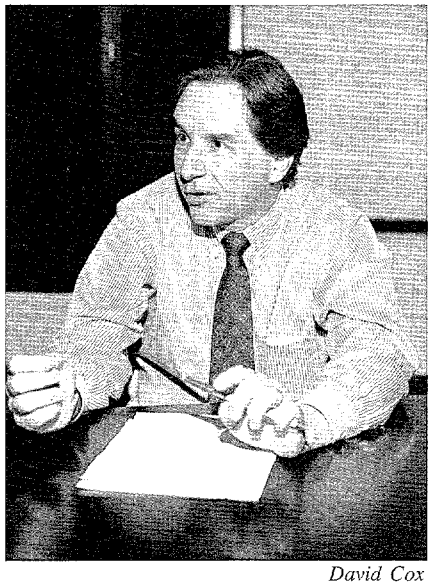
David Galas: And in fact the effort is rather dispersed. The NIH probably will very soon have about ten genome centers located at universities, and the DOE currently has three centers at Los Alamos, Livermore, and Berkeley national labs. But we also have a lot of smaller projects at other national labs and a large number of individual research grants at universities. So, in a sense this project is certainly nothing like big science in any way it's ever been described before. The Genome Project is different from projects at any of the discipline-oriented NIH institutes in that it tends to be a bit more focused and a bit more integrated because the maps we're aiming for can't be made by just a couple of people. And all the people working on the Project have to coordinate their efforts. Ultimately, compiling, collating, and checking all the data will be the real problem.

Bob Moyzis: The size of this project is not totally outside the scale of what has been happening elsewhere in biology. Individual lab efforts much larger than the physical-mapping effort at Los Alamos are not unusual. The Genome Project just makes more visible the movement toward larger, more coordinated research projects. The handwriting is on the wall, but many are reluctant to see it happen. As I mentioned earlier, there is a fear of losing control.

Lee Hood: Another concern of our critics is that this project won't produce anything useful for biology, that it is a misconceived project, and that it's boring science.

Bob Moyzis: Boring science is somebody taking for the 500th time yet another gene and sequencing the 200 nucleotides at the end to try to figure out whether there's another regulatory sequence out there that's going to somehow explain how the gene is turned on or off. That's molecular biology as it is currently done. My perception is that this project will revolutionize how people think about biology.

David Galas: Your comment reminds me of a poster, a satire on the state of molecular-biological research, that was displayed at a meeting on the Molecular Biology of Mammalian Gene Expression not too long ago. It was a generic poster outlining the formula for studying gene expression. This is what you do: You get a cDNA, you find the gene by hybridization, you look at expression in various tissues, you pull out the gene, you get the genomic clone, you sequence upstream, you sequence downstream, you do some gel-shift experiments, you do footprints, then you do direct mutagenesis, and then you show that this is the factor that binds this and that. Just plug in your favorite gene and it works! People learn something from that approach, but is it any less mindless than doing maps?



David Cox

From a psychological point of view the Project has let to a terrified scientific community. Researchers are saying, "Wait a minute. What am I going to do while you're making that map if I'm not getting any money to do my research?"

Nancy Wexler

The public thinks they have to wait fifteen years and then the human genome will be delivered on a platter, like the Hubble telescope, flaws and all. But as the genes spill out and the diseases are understood, the Project yields immediate benefits.

Nancy Wexler: To me the beauty of this project is that any new piece of information is immediately relevant. As soon as you obtain a sequence for a human gene, you can look at model organisms to find genes with similar sequences and perhaps identify the function of the gene. The public thinks that they have to wait fifteen years and then the human genome will be delivered on a platter, like the Hubble telescope, flaws and all. But as the genes spill out and the diseases are understood, the Project yields immediate benefits.

Bob Moyzis: That's an important difference between this so-called big science project and other projects, especially in the physical sciences. The infrastructure we are constructing—that's Norton's term—is useful long before it is finished. We should not, however, confuse this immediate usefulness with the ultimate goals. Multigenic traits, for example, will not be accessible until the linkage maps are complete. It's then that most of the fun begins.

David Cox: But I've heard many scientists ask, "How can I be sure that you will give me the tools from the Genome Project that I need to get on with my research?" Those not directly involved with the Genome Project feel they are being pushed out. A lot of thought currently taking place in the Genome Project is about how to get useful information out to the scientific community because that is the purpose of this project, and it has to start happening sooner than fifteen years, sooner than five years, and in fact sooner than two years.

The Genome Project must constantly assess what new tools can be made available to the scientific community and, at the same time, not jeopardize the whole reason for doing the project, which is to generate the maps in a cost-efficient and timely manner. Those two competing concerns must constantly be juggled.

There is a tool that the Genome Project will make available in the next year or so, a kit of 150 polymorphic DNA markers spaced evenly along the genome. That sparse version of the linkage maps we'll ultimately make will be the first product we give out to the community.

David Galas: As Nancy and Bob pointed out, the Genome Project is constantly generating not only new technologies and new data but also different ways of doing things in the molecular biology lab. As we go along, there's going to be a major increase in the usefulness of the Genome Project to the rest of biology with no decrease in the rate of the mapping.

Bob Moyzis: All the technology developed in the course of reaching the goals of the Genome Project becomes immediately useful for smaller projects. Even the large-scale physical-mapping projects have valuable spin-offs. Previously, students would spend their entire graduate career isolating, at best, one gene. Then they would pass it on to somebody else to do all the fun stuff of finding out what the gene does. Now that the physical-mapping projects make it possible to access large amounts of DNA quickly, a student can do some very interesting biology and do it a lot faster than he or she was able to do before.

David Botstein: This is the third or fourth field that I've watched grow. And what you see in a field that's really taking off is an exponential growth in the number of young people attending meetings. And that is what we're seeing in the genome business.

Like it or not, the Genome Project is going to transform the science of biology in a major way . . . The people who criticize the Genome Project on its scientific merit, who say it's boring, are largely lacking the vision to understand where this thing is going.

David Galas: Like it or not, the Genome Project is going to transform the science of biology in a major way. We will learn about so many things at a greater level of detail than ever before, and that detail will reveal principles that could not be approached up to now. The people who criticize the Genome Project on its scientific merit, who say it's boring, are largely lacking the vision to understand where this thing is going.

Lee Hood: The sound and fury from our critics has lessened slightly, but I suspect the volume will get turned up again as people go to Congress to try and squelch the genome initiative during the next budgetary hearings. Now that the Project is ongoing and the money is committed, I don't think the criticism will succeed in squelching it overtly. But, if our critics succeed in intimidating the NIH from spending money in ways that are consistent with the mission of the Genome Project, then they will have succeeded in squelching it by the back-door route. If most of the money gets spent on small projects that don't have much to do with the Genome Project itself, then the Project will flounder.

Right now the NIH is spending $8 billion a year on research, and the Genome Project is $105 million this year. So making the Genome Project into a more directed effort rather than spreading the money around is not going to change the character of American biological science in a fundamental way. That worry is unfounded.

The Genome Project is at the very beginning, and the NRC recommendation of $200 million per year is quite a bit more than we're now getting. So, quite apart from how well we're doing in managing the Project, if we've got a lot less money, the task will take longer. Frankly, the $200 million per year that the NRC suggested was really a guess. If anything, it'll cost more. So, we have to temper the suggested time line with the reality of the resources that we have available.

Classical Linkage Mapping

Classical linkage analysis is used to determine the arrangement of genes on the chromosomes of an organism. By tracing how often different forms of two variable traits are co-inherited, we can infer whether the genes for the traits are on the same chromosome (such genes are said to be linked), and if so, we can calculate the genetic distance separating the loci of the linked genes. The order of and pairwise distances between the loci of three or more linked genes are displayed as a genetic-linkage map.

For simplicity, we will consider traits of the type that Mendel studied, namely, traits exhibiting two forms, or phenotypes, one dominant and one recessive. Each such Mendelian trait is determined by a single pair of genes, either *AA*, *Aa*, or *aa*, where *A* is the dominant allele (form) of the gene and *a* is the recessive allele. Many inherited human diseases fall into this category. The two phenotypes are the presence or absence of the disease, and they are determined by a single gene pair, either *DD*, *DN*, or *NN*, where *D* is the defective allele that causes disease and *N* is the normal allele. If *D* is dominant, as in Huntington's disease and retinoblastoma, a person who inherits only one copy of *D*, and therefore has the genotype *DN*, can manifest the disease. Alternatively, if *D* is recessive, as in neurofibromatosis, cystic fibrosis, and most other inheritable human diseases, a person must inherit a copy of *D* from each parent (genotype *DD*) to manifest the disease phenotype. The two members of a gene pair are located at corresponding positions on a pair of homologous chromosomes. The chromosomal position of the gene pair for trait "A" will be called locus A. In the figures the dominant phenotype will be referred to as dom "A" and the recessive phenotype as rec "a."

First let's consider the inheritance of two unlinked traits, "A" and "B." Here, unlinked means that the gene pairs for the two traits are on different chromosome pairs. Since the chromosomes on which the genes reside are inherited independently, the genes are also inherited independently. In other words each offspring of a parent with the genotype *AaBb* has an equal chance of inheriting *AB*, *Ab*, *aB*, or *ab* from that parent. The latter statement is the law of independent assortment discovered by Mendel. (See the discussion of Mendelian genetics in "Understanding Inheritance.")

Now let's suppose instead that traits "A" and "B" are linked and that a parent carries the dominant alleles *A* and *B* on one chromosome of a homologous pair and the alleles *a* and *b* on the other chromosome. The offspring usually co-inherit either *A* with *B* or *a* with *b*, and, in this case, the law of independent assortment is not valid. Thus to test for linkage between the genes for two traits, we examine certain types of matings and observe whether or not the pattern of the combinations of traits exhibited by the offspring follows the law of independent assortment. If not, the gene pairs for those traits must be linked, that is they must be on the same chromosome pair.

Question: What types of matings can reveal that the genes for two traits are linked?

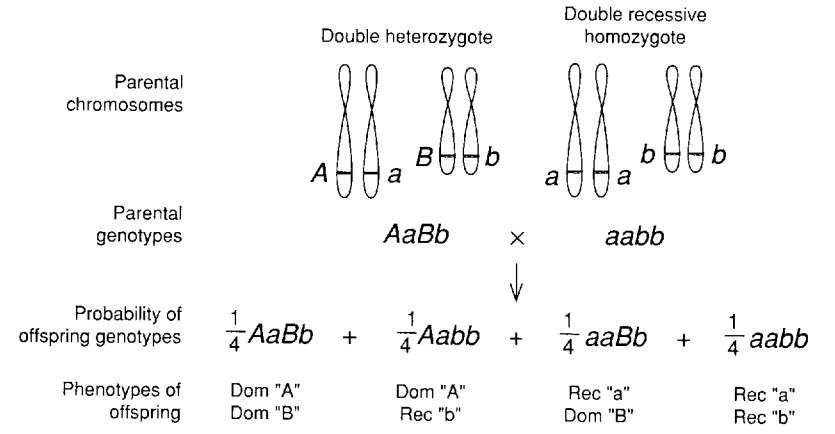
Answer: Only matings involving an individual who is heterozygous for both traits (genotype *AaBb*) reveal deviations from independent assortment and thus reveal linkage. Moreover, the most obvious deviations occur in the test cross, a mating between a double heterozygote and a doubly recessive homozygote (genotype *aabb*). Recall that individuals with the genotype *AaBb* manifest both dominant phenotypes; those with the genotype *aabb* manifest both recessive phenotypes.

A Simplified Example: Consider a test cross between a double heterozygote (*AaBb*) and a double recessive homozygote (*aabb*). Without additional information, all we know is that the genes of the heterozygous parent could be arranged in any one of the three configurations shown in cases 1, 2a, or 2b. Recall, however, that a parent transmits only one member of each chromosome pair to each of its offspring, so each of the possible arrangements would yield a different result. In case 1, where the gene pairs for traits "A" and "B" are on different chromosome pairs, the offspring can exhibit all four possible two-trait phenotypes, each with a probability of 1/4, in agreement with the law of independent assortment. In cases 2a and 2b, where the gene pairs are linked (and we ignore the effects of crossing over, a phenomenon described below), the offspring exhibit only two of the four composite phenotypes, each with a probability of 1/2. Thus if the genes for traits "A" and "B" are linked, it would appear that the results of the test cross would depart significantly from predictions based on independent assortment.

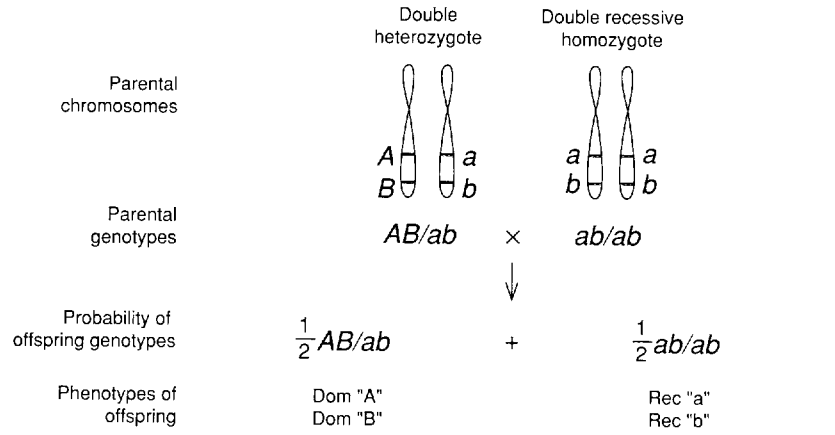
The reader should note the difference in the arrangement of alleles in cases 2a and 2b and how each arrangement, or *linkage phase*, in the heterozygous parent leads to different two-trait phenotypes among the offspring. In case 2a, *A* and *B* are on one chromosome and *a* and *b* are on the other (a genotype denoted by *AB/ab*, where the slash separates the alleles on different chromosomes). Consequently, the offspring from this test cross exhibit either both dominant or both recessive phenotypes, each with a probability of 1/2. In case 2b, *A* and *b* are on one chromosome and *a* and *B* are on *different* members of the homologous pair (genotype *Ab/aB*), and so the offspring exhibit the other two composite phenotypes, each a combination of a dominant and a recessive trait and, again, each with a probability of 1/2. In this simplified example, it appears quite easy to distinguish linkage from independent assortment, provided the test cross results in a large number of progeny. However, in simplifying the example we have made a significant omission.

Test Cross for Two Traits (without Crossing Over)

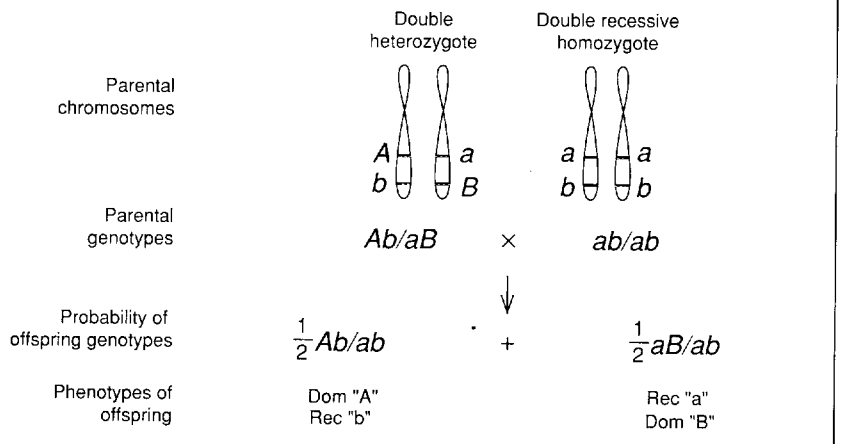
Case 1: Independent Assortment



Case 2a: Linkage (with Linkage Phase I)

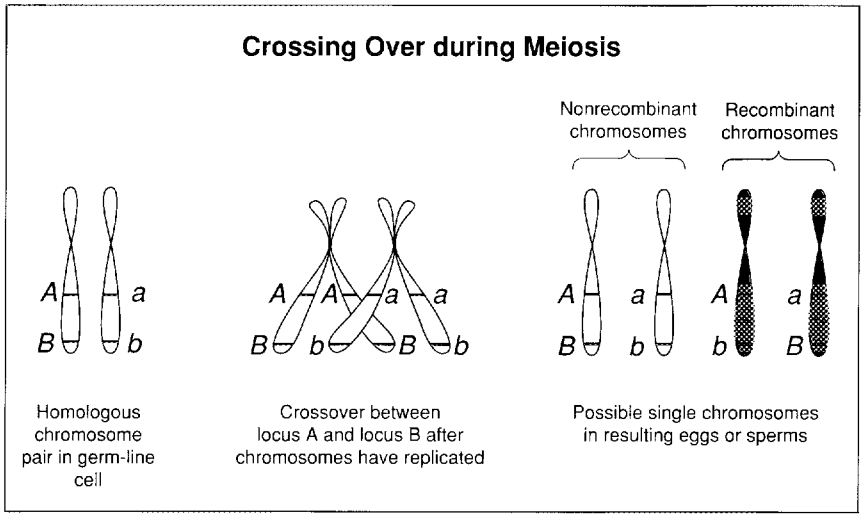


Case 2b: Linkage (with Linkage Phase II)



Question: Are two alleles on the same chromosome* always *inherited together?

Answer: No. During meiosis (the formation of eggs or sperms), two homologous chromosomes may exchange corresponding segments of DNA in a process called crossing over. Crossing over leads to formation of gametes that possess chromosomes containing new combinations of alleles, or recombinant chromosomes. Crossing over is not a rare phenomenon. In fact, each human chromosome pair within a germ-line cell undergoes, on average, about 1.5 crossovers during meiosis.



Example: Consider again a doubly heterozygous parent with the genotype *AB/ab*. That is, *A* and *B* are on one member of the homologous chromosome pair and *a* and *b* are on the other. During meiosis each chromosome is replicated and the resulting four chromosomes are parceled out so that only one enters each gamete. If crossing over does not occur between locus A and locus B (as assumed in case 2a above), each egg or sperm produced by the parent receives a chromosome containing either *A* and *B* or *a* and *b*. Those chromosomes are said to be nonrecombinant for traits "A" and "B." On the other hand, if crossing over happens to occur between locus A and locus B, as shown in the figure at left, then some gametes will receive a chromosome containing a new combination of alleles, either *A* and *b* or *a* and *B*. Those chromosomes (shaded red) are said to be recombinant for traits "A" and "B." (Note that only individuals who are doubly heterozygous for two traits can produce gametes containing chromosomes that are recombinant for those traits.) The appearance of a recombinant, an offspring containing a recombinant chromosome, is called a recombination event.

Question: How do recombination events complicate the determination of linkage between the genes for two traits?

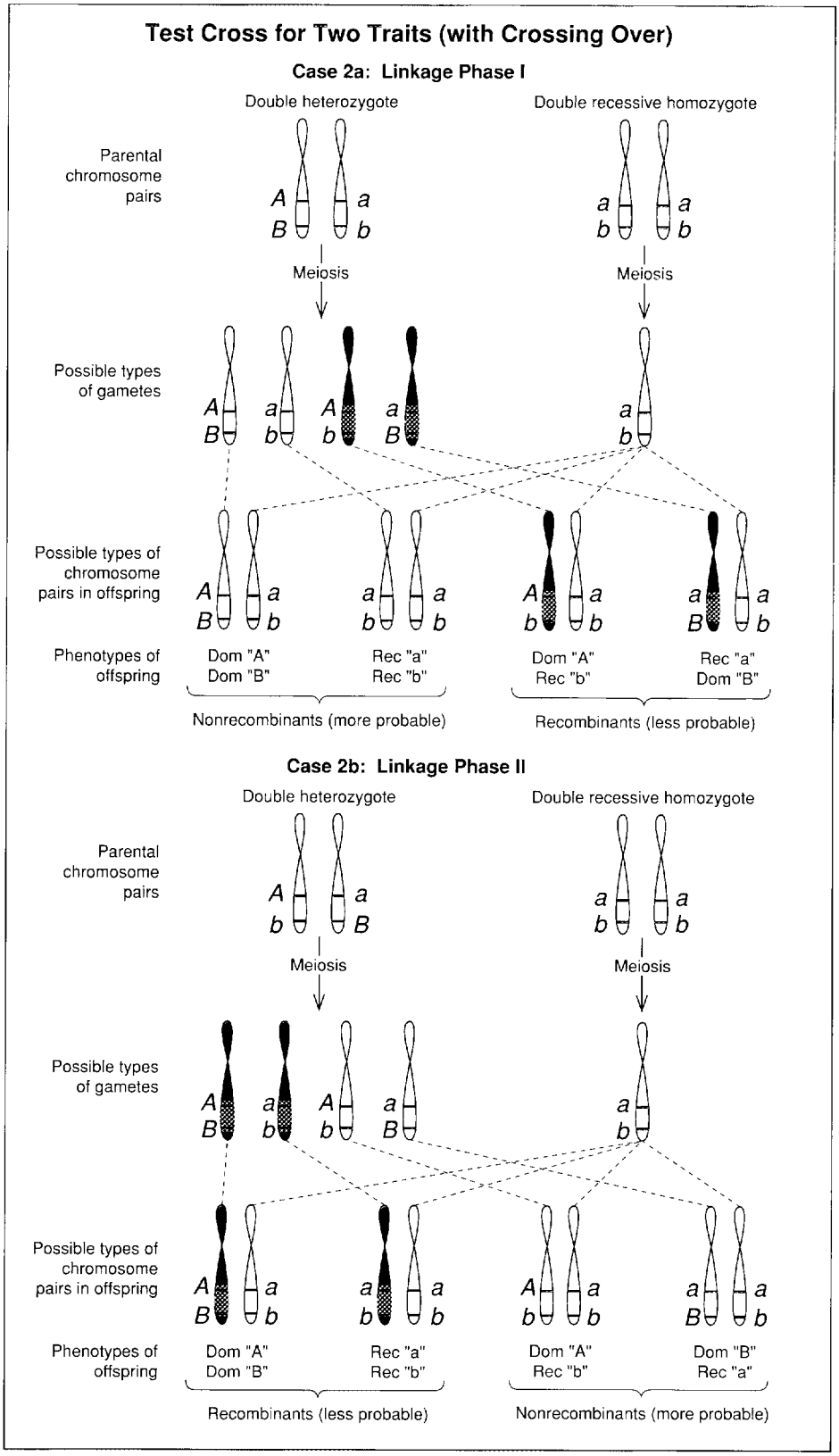
Answer: When we include the possibility of recombinant offspring in cases 2a and 2b (above), the distinction between case 1 (independent assortment) and cases 2a and 2b (linkage) becomes less obvious.

A More Realistic Example: The figure on the page opposite shows the test crosses for cases 2a and 2b, this time including the possibility of recombinants among the offspring. The doubly heterozygous parent may produce recombinant chromosomes (shown in red), which can then be inherited to produce recombinant offspring. In each case the recombinants have the composite phenotypes that were absent when the possibility of crossing over was not included (see cases 2a and 2b above). In other words, both cases 2a and 2b can produce all four composite phenotypes, just as does case 1 (independent assortment). However, whereas in case 1 the probabilities of producing the phenotypes were equal, in case 2 the probability of

producing recombinants is usually less than the probability of producing nonrecombinants. Thus linkage will be apparent from the results of a test cross provided three criteria are met: (1) the loci of the linked genes must be relatively close together; (2) a large number of progeny must be available to obtain good statistics (therefore we may have to examine a large number of matings); and (3) the test cross must involve only one possible linkage phase; that is, we must be able to infer which linkage phase is present in the heterozygous parent if indeed the genes are linked.

If these criteria are met, then we know which offspring are recombinants. Further, by comparing the number of recombinant offspring with the total number of offspring, we can arrive at an estimate of the probability of producing a recombinant. That probability is called the *recombination fraction* and, as we will see below, is related to the distance separating the loci of the linked genes.

We will also see that as the loci of two linked gene pairs get farther and farther apart, the recombination fraction for the two gene pairs approaches 0.5, so that the two recombinant phenotypes are produced with the same probability as the two nonrecombinant phenotypes. In other words, when the recombination fraction is 0.5, all four composite phenotypes are produced with equal probability, just as they are in case 1, and we infer that the gene pairs are unlinked even though they are on the same chromosome pair.

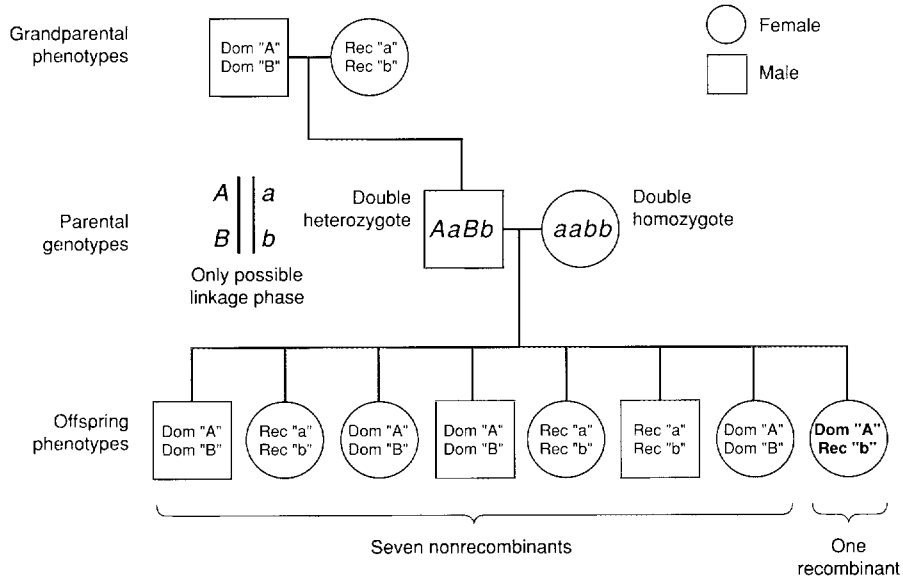


When we try to determine linkage among human traits, the problems we encounter are that human matings are not controlled (and therefore test-cross matings are rare), the data needed to infer the possible linkage phase in the heterozygous parent may not be available, and the number of offspring produced by two parents is typically much smaller than that produced by a pair of experimental organisms.

Question: How do we estimate, from the offspring of a single family, the likelihood that two gene pairs are linked?

Answer: For simplicity, we consider a three-generation family for which we have enough information to infer the linkage phase in the heterozygous parent, if indeed the gene pairs for the two traits under study are linked. We can then identify which offspring are recombinants for the two traits, again under the hypothesis of linkage, and divide the number of recombinant offspring by the total number of offspring to obtain an estimate of the recombination fraction. Finally, we evaluate the likelihood of obtaining the data we have under two opposing hypotheses: that the gene pairs are linked, and that the gene pairs are unlinked. The ratio of the two likelihoods is a measure of how reliably the data distinguish linkage from independent assortment.

Test Cross for Linkage in a Three-Generation Family



$$\text{Likelihood ratio} = \frac{L(\text{data} \mid \theta)}{L(\text{data} \mid \frac{1}{2})} = \frac{\theta^1(1-\theta)^7}{(\frac{1}{2})^8}$$

$$\text{"Most likely" recombination fraction} = \theta_{max} = \frac{1}{8}$$

$$\text{Lod score} = \log_{10} \frac{L(\text{data} \mid \theta_{max})}{L(\text{data} \mid \frac{1}{2})} \approx 1.08$$

Data from this family indicate that the odds are about $10^{1.08}$, or 12.6 to 1 in favor of linkage between traits "A" and "B".

Example: Consider a test cross between a male double heterozygote (*AaBb*) and a female double recessive homozygote (*aabb*). The doubly heterozygous father inherited both dominant alleles from his father, and therefore, if the gene pairs for traits "A" and "B" are linked, the father must carry alleles *A* and *B* on the same chromosome. Thus, under the hypothesis of linkage, we know the linkage phase in the father, and therefore, we know that an offspring exhibiting one dominant and one recessive trait is a recombinant. Among the offspring shown here, one is a possible recombinant and seven are possible nonrecombinants. Thus the genes for traits "A" and "B" appear to be linked, with a recombination fraction of 1/8.

We need a method to evaluate the statistical significance of our results. The conventional approach is to apply maximum-likelihood analysis, which estimates the "most likely" value of the recombination fraction θ as well as the odds in favor of linkage versus nonlinkage. We begin with the conditional probability $L(\text{data} \mid \theta)$, which is the likelihood of obtaining the data if the genes are linked and have a recombination fraction of θ. In particular, the likelihood of obtaining one recombinant and seven nonrecombinants when the recombination fraction is θ is proportional to $\theta^1(1-\theta)^7$, since θ is, by definition, the probability of obtaining a recombinant and $(1-\theta)$ is the probability of obtaining a nonrecombinant.

We then determine θ_{max}, the value of θ at which L has its maximum value, or equivalently, at which $dL/d\theta = 0$. In this simple case, where we have only one linkage phase to consider, θ_{max} is identically equal to 1/8, the value we obtained by direct inspection of the data. (If both linkage phases are possible, both must be taken into account in the likelihood function.)

Next we compute the ratio of likelihoods $L(\text{data}\,|\,\theta = \theta_{max})/L(\text{data}\,|\,\theta = 1/2)$, where $L(\text{data}\,|\,\theta = 1/2)$ is the likelihood of obtaining the data when $\theta = 1/2$, or equivalently, when the gene pairs are unlinked. This ratio gives the odds in favor of linkage with a recombination fraction of θ_{max} versus nonlinkage. For this family we find that the odds are about 12.6 to 1 in favor of linkage with a recombination fraction of 1/8 versus independent assortment, or nonlinkage.

Geneticists usually report the results of linkage analysis in terms of a lod score, which is the logarithm (to the base 10) of $L(\text{data}\,|\,\theta = \theta_{max})/L(\text{data}\,|\,\theta = 1/2)$. For this family the lod score is about 1.1. A lod score of 3, which corresponds roughly to 1000-to-1 odds that two gene pairs are linked, is considered definitive evidence for linkage. The analysis of many families with large numbers of siblings is usually required to achieve lod scores of 3 or more.

Question: Why is the recombination fraction for linked gene pairs related to the distance separating the gene pairs?

Answer: If we assume that crossing over occurs with equal probability along the lengths of the participating chromosomes (an assumption first made by Thomas Hunt Morgan around 1910), then the distance between the loci of two gene pairs determines the probability that recombinant chromosomes will be formed during meiosis, which, by definition, is the recombination fraction. In particular, if two loci are far apart, a greater number of crossovers between the two will occur and recombinant chromosomes will be formed during a greater number of meioses than if the loci are close together. In other words, the value of the recombination fraction increases with the distance between the gene pairs, and thus it provides a measure of the physical distance separating the two pairs. Additionally, pairwise comparison of recombination fractions for several gene pairs on the same chromosome pair establishes the order of the loci along the chromosome pair.

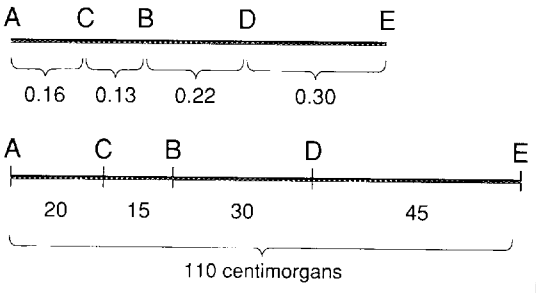
Question: Once we have determined the recombination fractions for many pairs of genes, how do we construct linkage maps of the chromosomes?

Answer: First, we use the recombination fractions to separate the gene pairs into linkage groups. A linkage group is a set of gene pairs each of which has been linked to at least one other member in the set and all of which, therefore, must be on the same chromosome pair. Then, because the recombination fraction increases with the distance separating the loci of two gene pairs, we can use them to order the loci of the gene pairs. The ordering is carried out much as one would order a set of points on a line, given the lengths of the line segments joining the various pairs of points. Next each recombination fraction is converted to a genetic distance, a quantity defined below. Finally, the loci are plotted on a line in a manner such that the plotted distance between any two loci is proportional to the genetic distance between the two loci.

Construction of a Linkage Map

Linkage data

Interval between gene loci	Recombination fraction	Genetic distance (centimorgans)
AB	0.25	35
AC	0.16	20
AD	0.36	65
AE	0.44	110
BC	0.13	15
BD	0.22	30
BE	0.39	75
CD	0.30	45
CE	0.42	90
DE	0.30	45



Example: The table shows the recombination fractions for a linkage group of five gene pairs, *Aa, Bb, Cc, Dd,* and *Ee*. The loci of these gene pairs are A, B, C, D, and E, respectively, and AB, for example, denotes the interval between locus A and locus B. The recombination fractions corresponding to the intervals AB, BC, and AC are 0.25, 0.13, and 0.16, respectively. Consequently, locus C is inferred to lie between locus A and locus B, as shown in the linkage map. All five loci can be ordered by this type of inference, as shown in the figure.

The next step is to convert the recombination fractions into genetic distances. The genetic distance between locus A and locus B is defined as the average number of crossovers occurring in the interval AB. When the interval is so small that the probability of multiple crossovers in the interval is negligible, the recombination fraction is about equal to the average number of crossovers, or to the genetic distance. However, as two loci get farther apart, the probability of multiple crossovers in the interval between them increases. Further, an even number of crossovers between two loci returns the alleles at those loci to their original positions and therefore does not result in the production of recombinant chromosomes. Consequently, the recombination fraction underestimates the average number of crossovers in the interval, or the genetic distance between the two loci. We therefore use what is called a mapping function to translate recombination fractions into genetic distances.

In 1919 the British geneticist J. B. S. Haldane proposed such a mapping function (see below). The table lists the genetic distance, according to Haldane's function, that corresponds to each recombination fraction, and those distances are displayed as a linkage map.

Question: What is Haldane's mapping function?

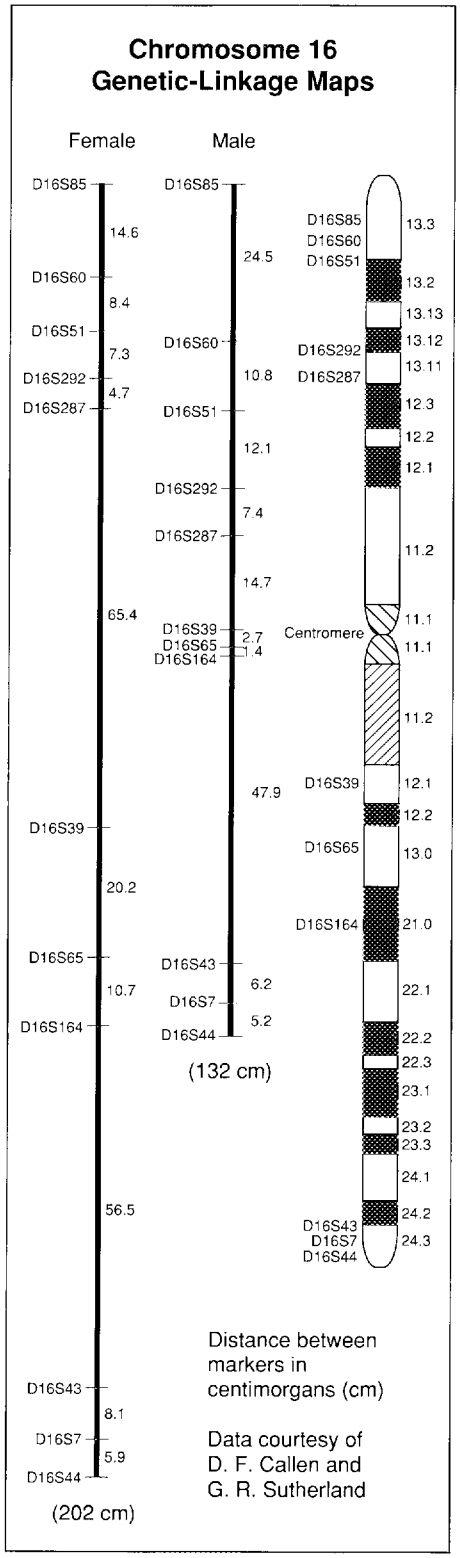
Answer: Haldane defined the genetic distance, x, between two loci as the average number of crossovers per meiosis in the interval between the two loci. He then assumed that crossovers occurred at random along the chromosome and that the probability of a crossover at one position along the chromosome was independent of the probability of a crossover at another position. (It follows from those assumptions that the distribution of crossovers is a Poisson distribution.) Using those assumptions, he derived the following relationship between θ, the recombination fraction and x, the genetic distance (in morgans): $\theta = \frac{1}{2}\left(1 - e^{-2x}\right)$, or, equivalently, $x = -\frac{1}{2}\ln(1 - 2\theta)$. Note that as the genetic distance between two loci increases, the recombination fraction approaches a limiting value of 0.5. Also, when the recombination fraction is small, x and θ are approximately equal. In practice geneticists treat them as equal for recombination fractions of 0.1 or less. As indicated, the unit of genetic distance is the morgan, or, more often used, the centimorgan, a distance between two loci such that on average 0.01 crossovers occur in that interval. Cytological observations of meiosis indicate that the average number of crossovers undergone by the chromosome pairs of a germ-line cell during meiosis is 33. Therefore, the average genetic length of a human chromosome is about 1.4 morgans, or about 140 centimorgans.

Question: How can we estimate the physical distance between the two gene loci from the genetic distance between them?

Answer: Since the average genetic length of a human chromosome is about 140 centimorgans and the average physical length of the DNA molecule in a human chromosome is about 130 million base pairs, 1 centimorgan corresponds to approximately 1 million base pairs of DNA. However, this correspondence is very rough because it is based on the assumption that the probability of crossing over is constant along the lengths of the chromosomes. In reality, however, the probability of crossing over varies dramatically from point to point, and a genetic distance of 1 centimorgan may correspond to a physical distance as large as 10,000,000 base pairs or as small as 100,000 base pairs. Also, because the probability of crossing over is higher in female humans than in male humans, genetic distances are greater in females than in males.

Example: Shown here are two genetic-linkage maps for chromosome 16, one derived from data for males and the other from data for females. The female linkage map is 70 centimorgans longer than the male linkage map. But we know from other data that the physical length of the DNA molecule in either a male or female chromosome 16 is the same (about 100 million base pairs). Note that the loci listed on the linkage map are those not of genes but rather of DNA markers (see "Modern Linkage Mapping").

CAVEAT: Classical linkage analysis can be applied only to genes for variable traits, and, most efficiently, to genes for single-gene variable traits such as many inherited human diseases. It can tell us whether the gene pairs for two or more variable traits are on the same homologous chromosome pair, but alone it cannot tell us on which chromosome pair the gene pairs reside. Furthermore, it can tell us the order of the gene pairs in a linkage group, but alone it cannot tell us where any one of the gene pairs is physically located. Finally, classical linkage analysis provides a genetic distance between two linked gene pairs, but that distance is not always proportional to the length of the DNA segment separating the gene pairs. Thus, classical linkage analysis alone does not help us to isolate the particular segment of DNA that contains a particular gene. However, when linkage analysis is applied to inherited variations in DNA itself, it does serve that function (see "Modern Linkage Mapping"). ■



Modern Linkage Mapping

with polymorphic DNA markers—a tool for finding genes

Problem: In "Classical Linkage Mapping" we showed how to construct maps that give the order of and genetic distances between gene pairs for variable, single-gene traits that are linked (lie on the same homologous chromosome pair). Prominent among the variable, single-gene traits of humans are inherited diseases. Several thousand such genetic disorders have been identified, and many of the genes for those disorders were mapped through classical linkage analysis. However, the maps included no reference to the physical reality of DNA, and therefore they did not provide the information necessary to isolate a segment of DNA containing a disease-causing gene. Then, in 1980, David Botstein, Raymond L. White, Mark Skolnick, and Ronald W. Davis transformed linkage mapping into a tool for finding genes.

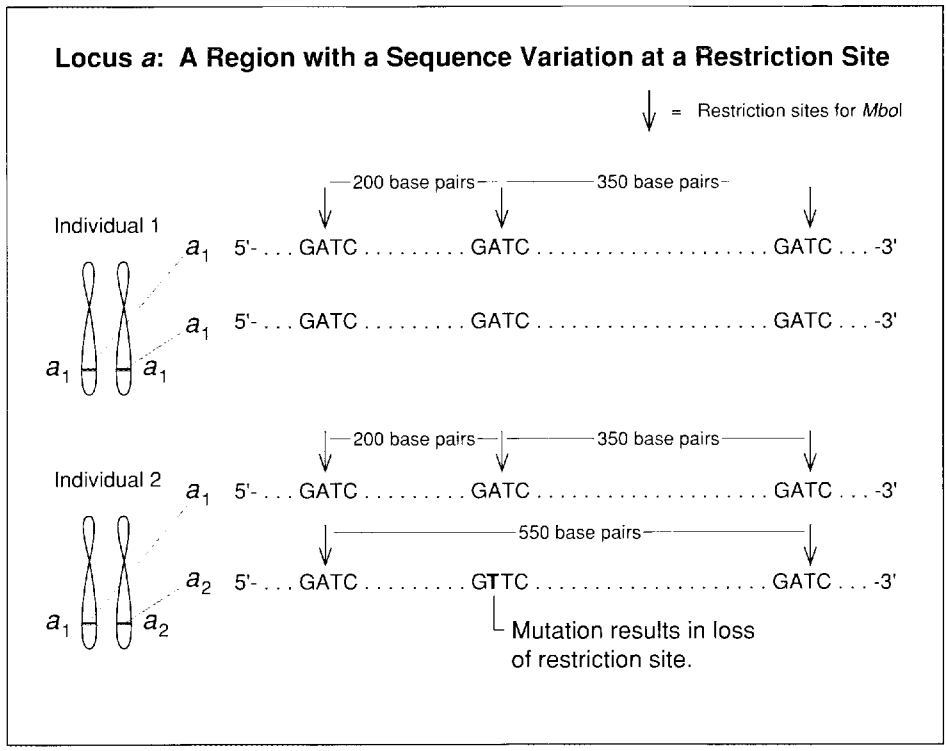
The Botstein Idea: If we could compare the base sequences of corresponding regions of the DNA from several individuals, we would find many regions with identical sequences—but we would also find many regions where the base sequence varies slightly from one individual to another. Those variable regions are called DNA polymorphisms. Now suppose we have available DNA probes that can not only reveal the presence of variable regions but also distinguish one sequence variation from another. Suppose further that some of the variable regions are fairly stable, so that a given sequence within such a region is transmitted from one generation to the next. In other words, each variable region exhibits only a limited number of sequence variations among the population. Such a variable region, together with the DNA probe that detects the sequence variations within that region, is called a polymorphic DNA marker.

Polymorphic DNA markers are very useful for several reasons. First, because they are variable, we can construct a linkage map of DNA markers just as we construct a linkage map of the genes that determine variable phenotypic traits. That is, we trace the co-inheritance of pairs of DNA markers to determine the genetic distances between them. Second, we can trace the co-inheritance of a marker and a variable phenotypic trait to determine the genetic distance between the marker and the gene responsible for the variable phenotypic trait. Finally, we can use the DNA probe for a marker to find the physical location of the marker on a chromosome. The physical loci of the polymorphic DNA markers can then serve as landmarks in the search for a specific gene. For example, if we know from the linkage map that a gene for a particular phenotypic trait lies between two particular DNA markers, then the gene of interest can be found in the stretch of DNA connecting the physical loci of the two markers. In summary, DNA markers provide a way to connect loci on linkage maps with physical loci in the human genome, which in turn, provides a way to find genes of interest.

Question: What is an example of a base-sequence variation within a region that can turn the region into a DNA marker?

Answer: The base-sequence variation within a region must be easily detectable to make the region a candidate for a DNA marker. One type of detectable variation is a single base change that results in the creation or loss of a restriction-enzyme cutting site. Such sites are short sequences, four to eight base pairs in length, at which a restriction enzyme cuts a DNA molecule. For example, each cutting site for the restriction enzyme *Mbo*I has the base sequence 5′-GATC.

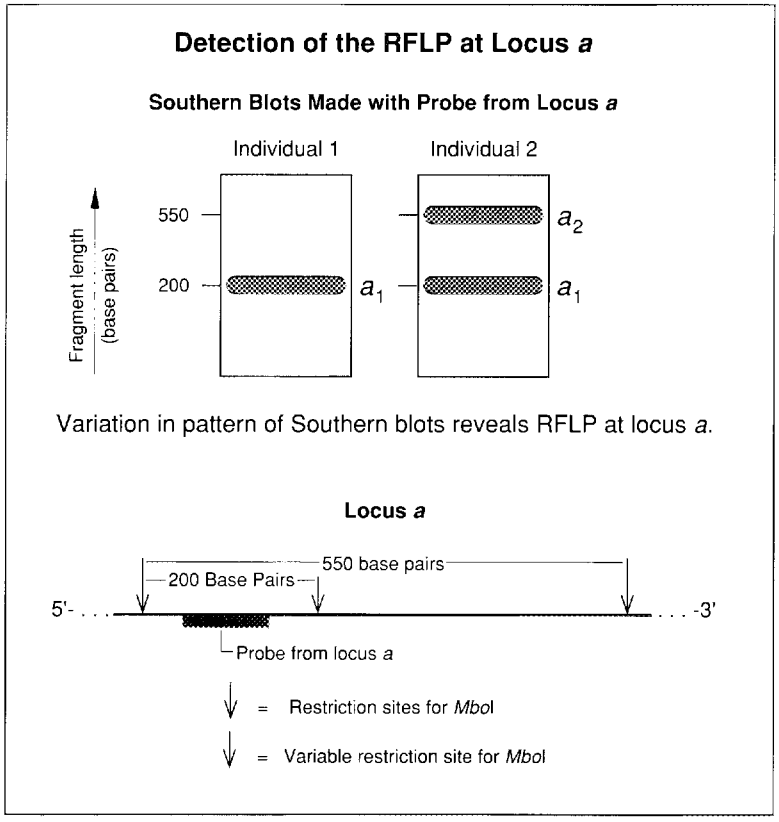
Example: Consider locus *a*, a variable region on a particular pair of of homologous chromosomes. The figure shows the DNA segments that compose locus *a* in the homologous chromosome pairs of two individuals. Also shown are the positions of the cutting, or restriction, sites for the restriction enzyme *Mbo*I within locus *a* and the distance between successive sites. Individual 1 carries two copies of a_1, a version, or allele, of locus *a* that has three restriction sites for *Mbo*I. Individual 2 carries one copy of a_1 and also a copy of another allele, a_2. Note that a_2 is missing the middle restriction site present in a_1. The absence of that restriction site is due to a change in a single base pair (shown in red). If *Mbo*I is allowed to cut the DNA from these two individuals, a_1 will be cut into two fragments of lengths 200 base pairs and 350 base pairs, whereas a_2 will be cut into one fragment of length 550 base pairs.



Question: How do we detect which alleles of locus* a *are present in the DNA molecules of two individuals?

Answer: We measure the lengths of the fragments from locus *a* produced by cutting the DNA with *Mbo*I and note the differences between the lengths of the fragments from the two individuals. We do so by making a Southern blot (see "Hybridization" in "Understanding Inheritance"). We begin by extracting many copies of the DNA from the blood cells of each individual. We then chop up, or digest, the DNA in each sample with the restriction enzyme *Mbo*I. The next step is to separate the resulting fragments (called restriction fragments) according to length by gel electrophoresis (see "Gel Electrophoresis" in "Understanding Inheritance"). Because shorter fragments travel farther through the gel than longer fragments, the lengths of the fragments can be determined from their final positions on the gel. We then transfer (blot) the fragments onto a filter paper in a manner that preserves their final gel positions.

Next, we allow a radioactively labeled DNA probe from locus *a* to hybridize, or bind by complementary base pairing, to the restriction fragments. The probe hybridizes only to fragments from locus *a* and thereby reveals their positions and therefore their lengths. Finally, we make an autoradiogram of the filter paper in which the positions of the fragments that have hybridized to the probe are imaged as dark bands.



Example: The figure shows Southern blots for the DNA of individuals 1 and 2 made with the enzyme *Mbo*I and a probe for locus *a*. The position of the probe is shown in the diagram of locus *a*. That particular probe binds to the restriction fragments of length 200 base pairs from allele a_1 and to the restriction fragments of length 550 base pairs from allele a_2. Since individual 1 carries allele a_1 only, the Southern blot of individual 1 shows one band at a position corresponding to a length of 200 base pairs. Individual 2 carries alleles a_1 and a_2 and therefore has a Southern blot showing two bands, one at 200 base pairs and one at 550 base pairs. The variation within locus *a* that causes this difference between the two Southern blots (the presence or absence of a restriction site) is called a restriction fragment length polymorphism, or RFLP, which is one type of polymorphic DNA marker. (Another type of polymorphic DNA marker is described in "The Polymerase Chain Reaction and Sequence-tagged Sites.")

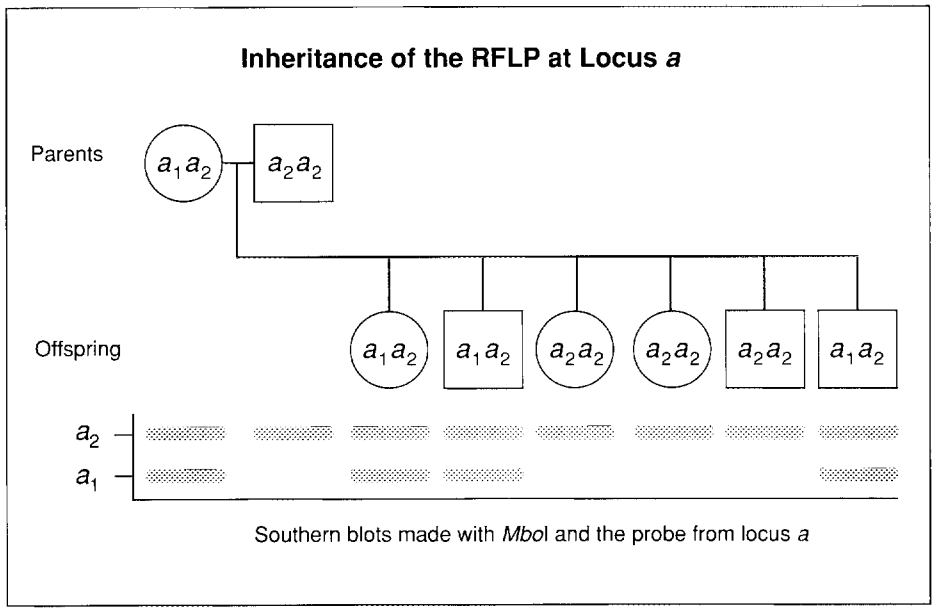
Question: How do we find polymorphic DNA markers?

Answer: Originally, this was done by a process involving patience and preferably luck. We randomly choose one clone from a collection of human DNA clones, use it as a probe in the making of Southern blots of the DNA of many individuals, and see whether the Southern blots vary from one individual to the next. A variation implies that the probe is part of a variable region of the genome and therefore defines that region as a polymorphic DNA marker. If the clone chosen does not reveal a difference, we continue choosing clones until a difference does show up. More recently, with the wide application of the polymerase chain reaction (PCR) and the discovery that there are a large number of highly variable, short di-, tri-, and tetranucleotide repeat sequences flanked by unique DNA sequences, it has become possible to select such regions of DNA and then develop them into highly polymorphic markers.

Question: How are polymorphic DNA markers used in linkage analysis?

Answer: In linkage analysis a polymorphic DNA marker is analogous to a gene that has two or more alleles. Each parent carries a pair of alleles of the marker, one on each member of a chromosome pair, so each parent may be either homozygous or heterozygous for the marker. Also, each parent transmits only one allele of the marker to each offspring.

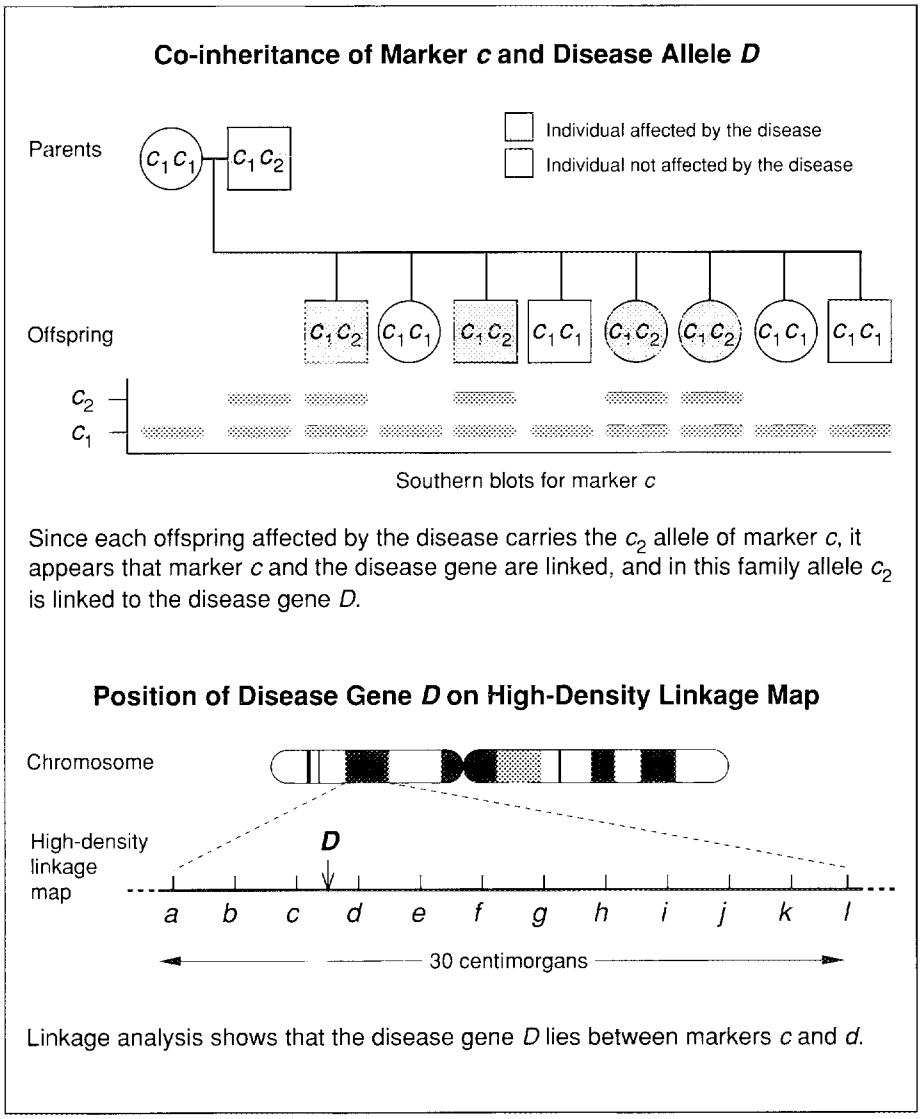
Example: The figure at right shows an example of the inheritance of the RFLP at locus *a*. Beneath each parent and each of their six children is shown the Southern blot for the marker. The father is heterozygous for the marker, carrying alleles a_1 and a_2. Among the offspring three are heterozygous and three are homozygous for a_2. The heterozygous offspring have inherited the allele a_1 from their father. Note that the alleles of a polymorphic DNA marker are inherently easier to trace than the alleles of a gene because the alleles of a polymorphic DNA marker are codominant. That is, none of them are recessive and each is directly observable.



We can also trace the inheritance of two markers, find out whether they are linked (on the same chromosome), and determine the recombination fraction for the two markers and thus the genetic distance between their loci. The linkage analysis exactly parallels that described for phenotypic traits in "Classical Linkage Mapping." In particular, an informative mating, one that reveals linkage between a pair of markers, must involve a parent who is heterozygous for both markers.

Question: Why does the Genome Project have as one of its top priorities the construction of a high-density linkage map of polymorphic DNA markers?

Answer: By 1996 the Genome Project hopes to have produced a set of linkage maps, each containing polymorphic DNA markers spaced along each human chromosome at intervals of 2 to 5 centimorgans, genetic distances that roughly correspond to physical distances of 2 to 5 million base pairs of DNA. Such a set of maps will enable researchers to find any gene of interest relative to the loci of approximately 1500 markers. In other words, the markers will form a set of reference points along the genome.

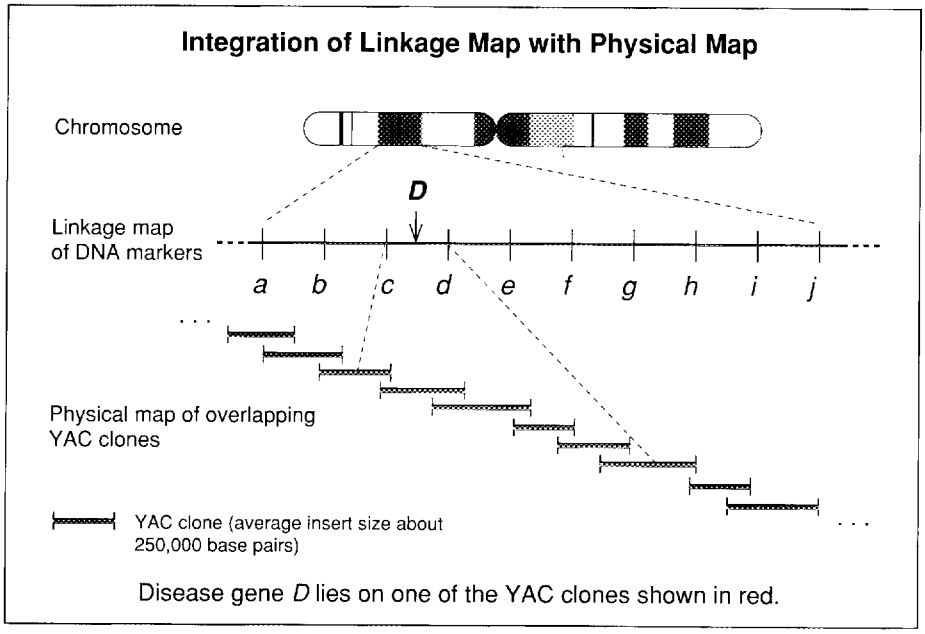


Example: Suppose we are interested in locating a mutant gene D that causes an inherited disease. We can find families affected by the disease and trace the co-inheritance of the disease with the reference markers on a linkage map. If we have a 2-centimorgan linkage map of highly informative markers (see "Informativeness and Polymorphic DNA Markers"), we can find markers flanking the gene that are less than 2 centimorgans away on either side. The pedigree in the figure shows the type of data needed to establish that the marker c and the disease gene D are tightly linked, that is, c and D are so close together that recombination events between them are rarely observed. Similar data between marker d and D would allow us to infer that D lies between c and d, as indicated in the lower part of the figure. This example shows the characteristic pattern of inheritance of an autosomal dominant disorder identified by allele c_2 of marker c.

Question: Once we have found DNA markers flanking a disease gene, how do we localize the disease gene on the DNA itself?

Answer: In addition to creating a linkage map of polymorphic DNA markers, the Genome Project is creating a physical map for each human chromosome. A physical map consists of an ordered set of overlapping cloned fragments that spans the entire length of the DNA molecule in the chromosome. As the physical maps and the linkage maps are constructed, the linkage map for each chromosome is being integrated with the physical map for that chromosome. That is, each locus on the linkage map will be associated with a locus on the physical map. Thus, if we find two markers that flank a disease gene, we will be able to ascertain how many base pairs of DNA separate the markers, and we will also have all that DNA available as cloned fragments. We therefore know that the disease gene is in one of those cloned fragments, and we can employ various methods to find the DNA segment that contains the gene. (Those methods are not necessarily straightforward, as explained on pages 111 and 142 of "Mapping the Genome.")

Example: The figure at right shows a schematic representation of a human metaphase chromosome (dark bands indicate A-T rich regions), a portion of a linkage map of polymorphic DNA markers, the position of a disease gene *D* on that map (as determined by linkage analysis), and the corresponding physical map of cloned fragments. Dotted lines connect the loci on the linkage map with the corresponding loci on the physical map and on the metaphase chromosome. Highlighted in red are the clones that must be searched to find the disease gene.



CAVEAT: In practice we need flanking markers that are within 1 centimorgan of the gene on either side so that the search for the disease gene will involve no more than about 2 million base pairs of DNA. Consequently, the long-term goal of the Genome Project is to find enough highly polymorphic DNA markers so that they are spaced at intervals of 1 centimorgan on the linkage maps, or a total of about 3300 markers. If they are found by a random search, we will have to find about ten times that number to achieve the 1-centimorgan map. The search for markers has been accelerated in several ways. For example, new types of markers are being systematically sought (see pages 133–134 in "The Polymerase Chain Reaction and Sequence-tagged Sites"), and automated techniques are being developed to detect DNA markers in large numbers of individuals. ■

Informativeness of Polymorphic DNA Markers

Carl E. Hildebrand, David C. Torney, and Robert P. Wagner

As mentioned in "Modern Linkage Mapping," one of the five-year goals of the Human Genome Project is to find highly informative polymorphic DNA markers spaced at 2- to 5-centimorgan intervals along the genetic linkage map of each human chromosome. In this context, informative means useful for establishing through linkage analysis that the marker is near a gene or another marker of interest. Recall that linkage between two variable loci can only be determined from matings in which one parent is heterozygous (carries two different alleles) for the marker or gene at each locus (see "Classical Linkage Mapping"). Thus a marker is highly informative for linkage studies if any individual chosen at random is likely to be heterozygous for that marker. As shown below, markers with many alleles, or highly polymorphic markers, tend to be highly informative.

Informativeness can be quantitatively measured by a statistic called the polymorphism information content, or PIC. This statistic is defined relative to a particular type of pedigree: one parent is affected by a rare dominant disease and is heterozygous at the disease-gene locus (genotype DN, where D is the dominant, disease-causing allele of the gene and N is the normal allele of the gene). The other parent is unaffected by the disease (genotype NN). The polymorphic DNA marker in question has several alleles, a_i, which are codominant, that is, each one can be detected so that the genotype at the marker locus ($a_i a_j$) can always be determined for any individual. Moreover, the marker locus is linked to (on the same chromosome pair as) the disease-gene locus. The important property of this type of pedigree is that the genotypes of the parents and the offspring at both the marker locus and the disease-gene locus can always be inferred. In this context, an offspring is said to be *informative* if we can infer from his or her genotype which marker allele is linked to (on the same chromosome as) the disease allele and would therefore be co-inherited with the disease allele in subsequent generations.

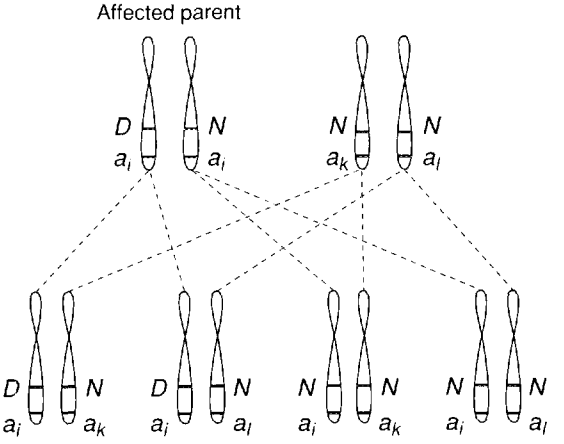
The PIC value of the marker is defined as the expected fraction of informative offspring from this type of pedigree. The figure divides the possible matings from such a pedigree into three categories depending on the genotypes of the parents at the marker locus. Each category has a different fraction of informative offspring. Note that the marker locus is assumed to be near the gene locus, so recombination between the two is a rare event and is not taken into account. In (a) the disease-affected parent is homozygous at the marker locus (genotype $a_i a_i$) and therefore none of the offspring are informative. In (b) both parents have the same heterozygous genotype at the marker locus ($a_i a_j$). Then, if each possible type of offspring is produced with equal probability, half of the offspring are informative. For all other combinations of marker alleles in the parents, all offspring are informative. The fully informative matings are summarized in (c).

PIC is the expected fraction of informative offspring from the type of pedigree shown in the figure. Under the assumption of Hardy-Weinberg equilibrium (that in the general population the frequencies of the alleles at the marker locus are independent of the frequencies of the alleles at the disease

Mating Categories for Evaluation of PIC

PIC is the expected fraction of informative offspring from a mating between an affected individual carrying a single copy of a dominant disease allele *D*, and an unaffected individual. This mating is divided into three categories depending on which alleles a_i (i = 1, 2, ...) are present at the locus of a polymorphic marker with *n* alleles. Each category produces a different fraction of informative offspring. Recall that the genotypes of each offspring are known, but the arrangement of alleles on the chromosomes is not known. Thus an offspring is informative if his or her genotype allows us to infer that *D* and a_i are linked in the affected parent and will therefore be coinherited. Informative offspring are shown in red.

(a) *k* and *l* can take on any values

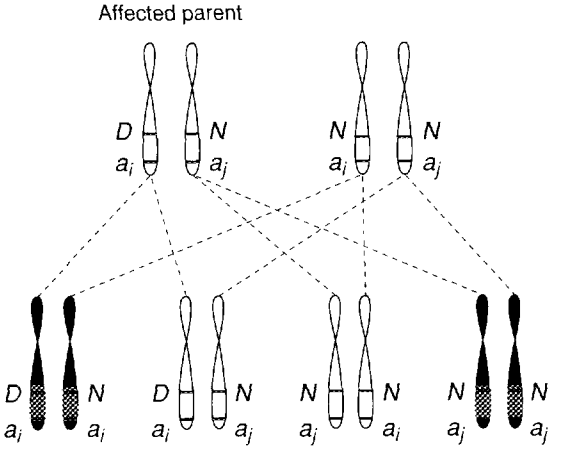


D = disease allele at disease locus
N = normal allele at disease locus
a_i = marker allele at marker locus
p_i = frequency of marker allele a_i

The affected parent is homozygous at the marker locus. Therefore, all offspring inherit a_i from the affected parent, and the inheritance of a_i cannot be used to predict the coinheritance of *D*.

Frequency of mating = p_i^2
Fraction of informative offspring = 0.

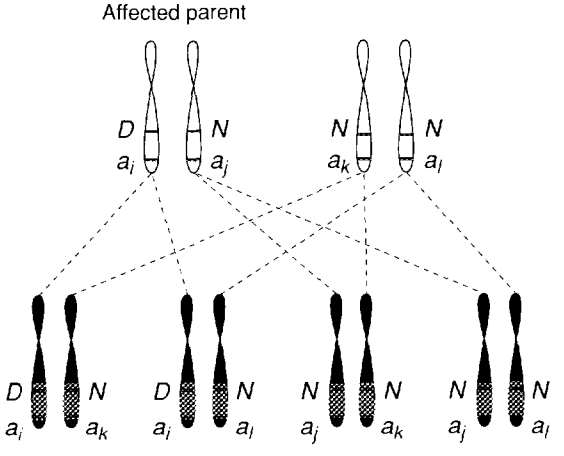
(b) $i \neq j$



Both parents are heterozygous at the marker locus (genotype a_ia_j). In the absence of crossing over two types of offspring are informative (red), that is, we can deduce from the genotypes of those offspring that *D* and a_i are linked (or on the same chromosome) in the affected parent. Specifically, the offspring genotype DNa_ia_i tells us directly that *D* and a_i were coinherited from the affected parent and therefore must be on the same chromosome. The offspring genotype DNa_ja_j tells us that *N* and a_j were coinherited from the affected parent and by the process of elimination the *D* and a_i must be on the same chromosome in that parent.

Frequency of mating = $2p_ip_j$ ($2p_ip_j$)
Fraction of informative offspring = 0.5

(c) $i \neq j$ and *k*, *l* can be any combination except *i*, *j* and *j*, *i*



The affected parent is heterozygous at the marker locus, and the unaffected parent carries a different combination of marker alleles than that in the affected parent. Thus the genotypes of all offspring allow one to deduce that *D* and a_i are linked in the affected parent.

Frequency of mating = $2p_ip_j$ ($1-2p_ip_j$)
Fraction of informative offspring = 1.0

locus) and the further assumption that a pair of alleles occurs with a frequency equal to the product of the two frequencies, we can determine the frequency of each mating category from the frequencies p_i of each marker allele a_i. Then (following Botstein et al., 1980 or Roychoudhury and Nei, 1988), to calculate PIC we multiply the frequency of each mating type by the expected fractions of informative offspring from that mating type and add the products:

$$\mathrm{PIC} = 1 - \sum_{i=1}^{n} p_i^2 - 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} p_i^2 p_j^2 = 1 - \sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} p_i^2\right)^2 + \sum_{i=1}^{n} p_i^4 \quad ,$$

where p_i = frequency of the marker allele, a_i and n = number of different alleles. Thus to evaluate the PIC value of a marker, we must determine the frequencies of each marker allele. We present an example (from Weber et al., 1990) in which the polymorphic marker is on human chromosome 16 and has four alleles each containing the dinucleotide repeat $(GT)_n$, where n takes on the values 170, 168, 166, and 154. A population of 120 chromosomes indicated that the frequencies of those four alleles are 0.01, 0.12, 0.2, and 0.67, respectively. Using the equation for PIC, we find that the PIC value for this marker equals 0.44. Thus 44 percent of the offspring should be informative in the type of pedigree illustrated in the figure. Theoretically, PIC values can range from 0 to 1. At a PIC of 0, the marker has only one allele. At a PIC of 1, the marker would have an infinite number of alleles. A PIC value of greater than 0.7 is considered to be highly informative, whereas a value of 0.44 is considered to be moderately informative. A gene or marker with only two alleles has a maximum PIC of 0.375. Clearly markers with greater numbers of alleles tend to have higher PIC values and thus are more informative.

An alternative measure of the degree of polymorphism of a marker is the heterozygosity, the probability that any randomly chosen individual is heterozygous for any two alles at a marker locus having allele frequencies p_i. Thus, heterozygosity $= 1 - \sum_{n=1}^{n} p_i^2$, where $\sum_{n=1}^{n} p_i^2$ is the homozygosity. PIC, therefore, will always be lower than the heterozygosity and can be considered to be the heterozygosity corrected for partially informative matings. Polymorphic loci containing many tandem repeats of a short sequence two to six bases long tend to have many alleles and are thus good candidates for highly informative markers. Those markers can be detected using PCR (see "The Polymerase Chain Reaction and Sequence-tagged Sites"). ■

Further Reading

David Botstein, Raymond L. White, Mark Skolnick, and Ronald W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32:314-331.

James L. Weber, Anne E. Kwitek, and Paula E. May. 1990. Dinucleotide repeat polymorphisms at the D16S260, D16S261, D16S265, D16S266, and D16S267 loci. *Nucleic Acids Research* 18:4034.

Jurg Ott. 1991. *Analysis of Human Genetic Linkage*,, revised edition. Baltimore: The Johns Hopkins University Press.

Arun K. Roychoudhury and Masatoshi Nei. 1988. *Human Polymorphic Genes*. New York/Oxford: Oxford University Press.