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?

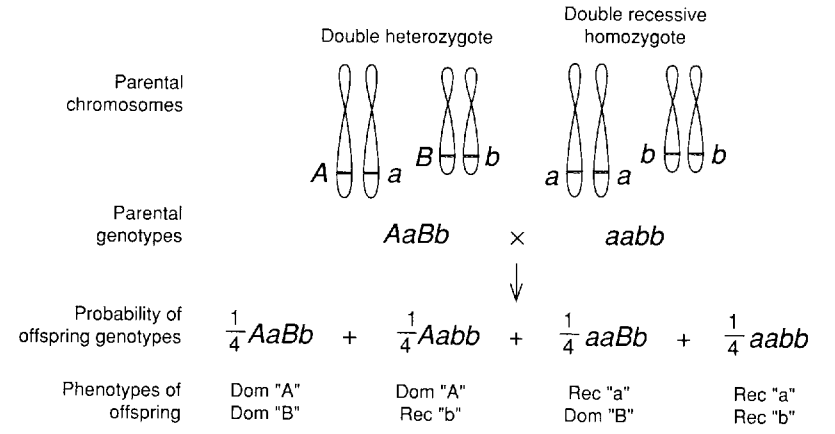
Answer: Only matings involving an individual who is heterozygous for both traits (genotype *AaBb*) reveal deviations from independent assortment and thus reveal linkage. Moreover, the most obvious deviations occur in the test cross, a mating between a double heterozygote and a doubly recessive homozygote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

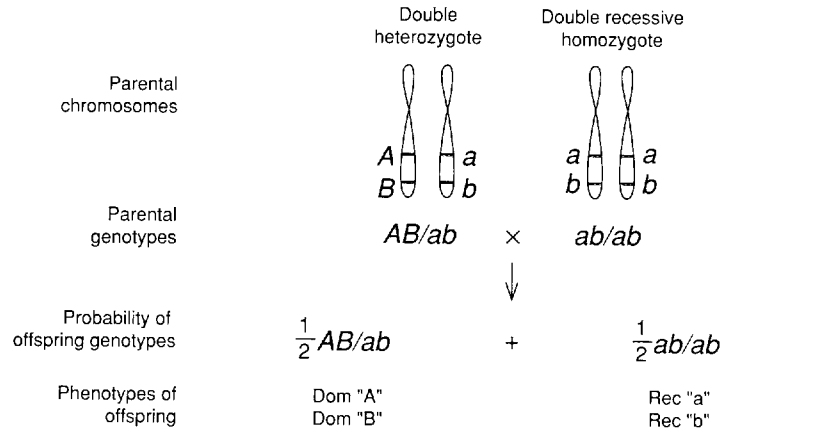
The reader should note the difference in the arrangement of alleles in cases 2a and 2b and how each arrangement, or *linkage phase*, in the heterozygous parent leads to different two-trait phenotypes among the offspring. In case 2a, *A* and *B* are on one chromosome and *a* and *b* are on the other (a 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)

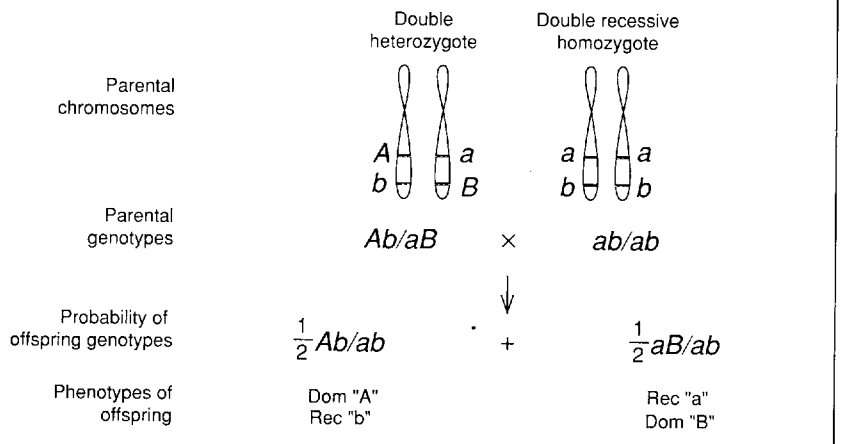
Case 1: Independent Assortment



Case 2a: Linkage (with Linkage Phase I)

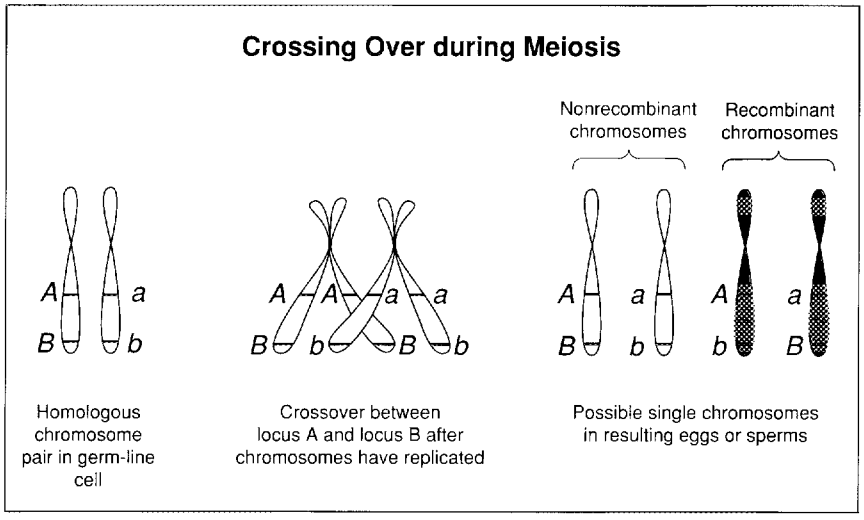


Case 2b: Linkage (with Linkage Phase II)



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?

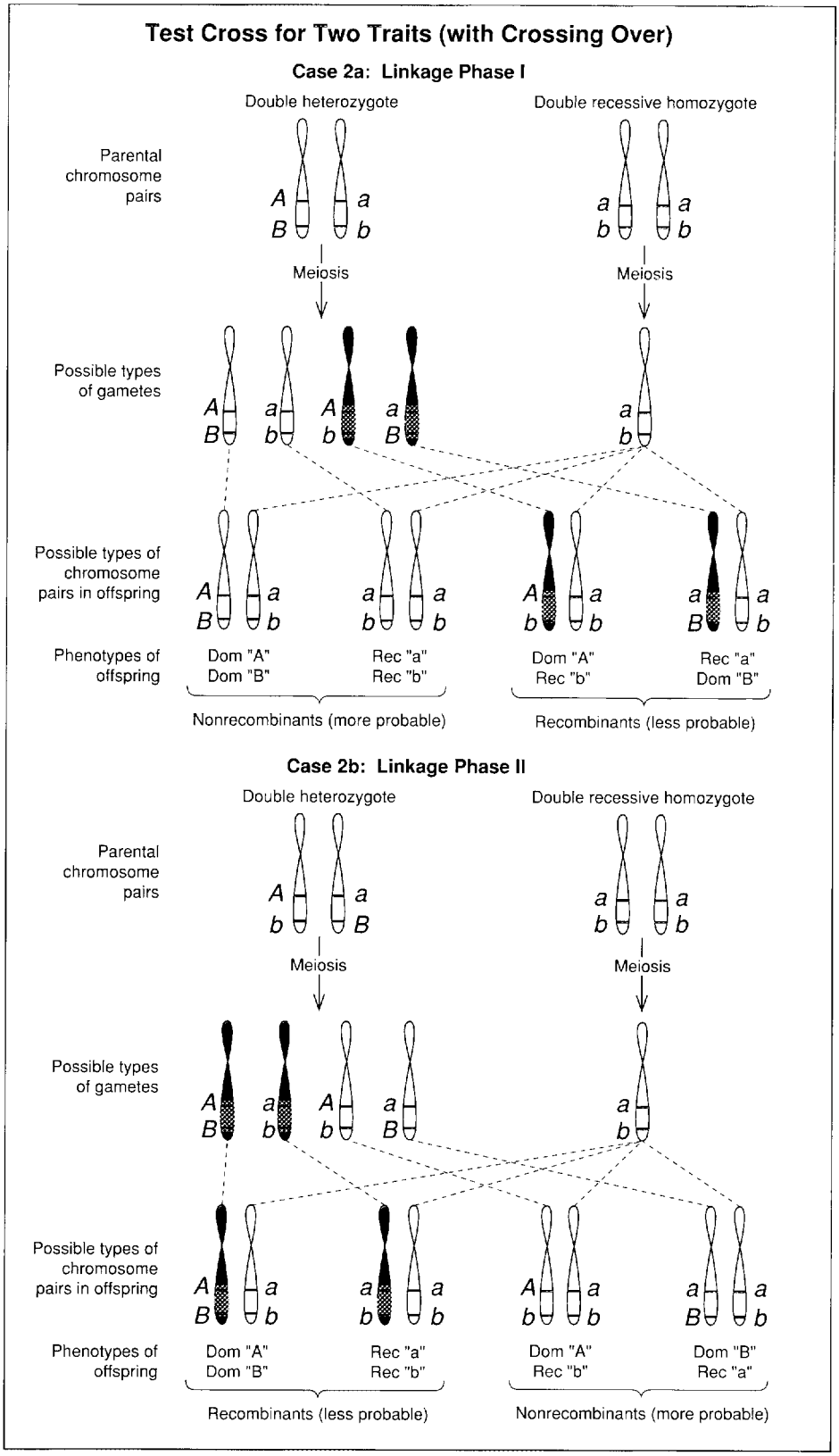
Answer: When we include the possibility of recombinant offspring in cases 2a and 2b (above), the distinction between case 1 (independent assortment) and cases 2a and 2b (linkage) becomes less obvious.

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.

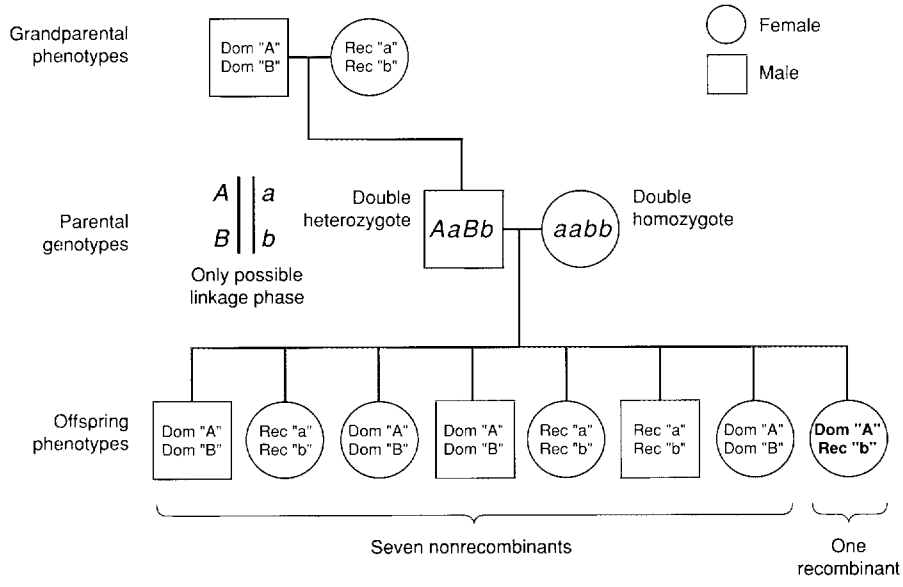


When we try to determine linkage among human traits, the problems we encounter are that human matings are not controlled (and therefore test-cross matings are rare), the data needed to infer the possible linkage phase in the heterozygous parent may not be available, and the number of offspring produced 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.

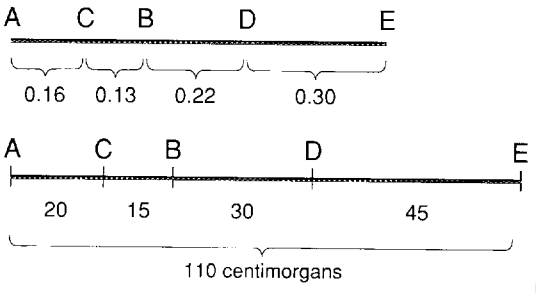
Question: Once we have determined the recombination fractions for many pairs of genes, how do we construct linkage maps of the chromosomes?

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"). ■