AN INTRODUCTION TO RECOMBINATION AND LINKAGE ANALYSIS

MARY SARA McPEEK*

Abstract. With a garden as his laboratory, Mendel (1866) was able to discern basic probabilistic laws of heredity. Although it first appeared as a baffling exception to one of Mendel's principles, the phenomenon of variable linkage between characters was soon recognized to be a powerful tool in the process of chromosome mapping and location of genes of interest. In this introduction, we first describe Mendel's work and the subsequent discovery of linkage. Next we describe the apparent cause of variable linkage, namely recombination, and we introduce linkage analysis.

Key words. genetic mapping, linkage, recombination, Mendel.

1. Mendel. Mendel's (1866) idea of enumerating the offspring types of a hybrid cross and his model for the result provided the basis for profound insight into the mechanisms of heredity. Carried out over a period of eight years, his artificial fertilization experiments involved the study of seven characters associated with the garden pea (various species of genus *Pisum*), with each character having two **phenotypes**, or observed states. The characters included the color of the petals, with purple and white phenotypes, the form of the ripe seeds, with round and wrinkled phenotypes, and the color of the seed albumen, i.e. endosperm, with yellow and green phenotypes.

Mendel first considered the characters separately. For each character, he grew two true-breeding parental lines, or strains, of pea, one for each phenotype. For instance, in one parental line, all of the plants had purple petals, and furthermore, over a period of several years, the offspring from all self-fertilizations within that line also had purple petals. Similarly, he grew a true-breeding parental line of white-flowered peas. When he crossed one line with the other by artificial fertilization, all the resulting offspring, called the **first filial** or F_1 **generation**, had purple petals. Therefore, the purple petal phenotype was called **dominant** and the white petal phenotype **recessive**. After self-fertilization within the F_1 generation, among the offspring, known as the **second filial** or F_2 **generation**, 705 plants had purple and 224 plants had white petals out of a total of 929 F_2 plants. This approximate 3:1 ratio (p-value .88) of the dominant phenotype to the recessive held for the other six characters as well.

Mendel found that when F_2 plants with the recessive phenotype were self-fertilized, the resulting offspring were all of the recessive type. However, when the F_2 plants with the dominant phenotype were self-fertilized, 1/3 of them bred true, while the other 2/3 produced offspring of both phenotypes, in a dominant to recessive ratio of approximately 3:1. For instance, among

* Department of Statistics, University of Chicago, Chicago, Il 60637.

100 F_2 plants with purple petals, 36 bred true, while 64 had both purple and white-flowered offspring (the numbers of these were not reported). Mendel concluded that among the plants with the dominant phenotype, there were actually two types, one type which bred true and another hybrid type which bred in a 3:1 ratio of dominant to recessive.

Mendel's explanation for these observations is that each plant has two units of heredity, now known as **genes**, for a given character, and each of these may be one of two (or more) types now known as **alleles**. Furthermore, in reproduction, each parent plant forms a reproductive seed or **gamete** containing, for each character, one of its two alleles, each with equal chance, which is passed on to a given offspring. For instance, in the case of petal color, the alleles may be represented by **P** for purple and **p** for white. (In this nomenclature, the dominant allele determines the letter of the alphabet to be used, and the dominant allele is uppercase while the recessive allele is lowercase.) Each plant would have one of the following three **genotypes**: **pp**, **pP** or **PP**, where types **pp** and **PP** are known as **homozygous** and type **pP** is known as **heterozygous**. Plants with genotype **pp** would have white petals, while those with genotype **pP** or **PP** would have purple petals. The two parental lines would be of genotypes **pp** and **PP**, respectively, and would pass on gametes of type **p** and **P**, respectively. The F_1 generation, each having one **pp** parent and one **PP** parent, would then all be of genotype **pP**. A given F_1 plant would pass on a gamete of type **p** or of type **P** to a given offspring, each with chance 1/2, independent from offspring to offspring. Then assuming that maternal and paternal gametes are passed on independently, each plant in the F_2 generation would have chance 1/4 to be of genotype **pp**, 1/2 to be of genotype **pP**, and 1/4 to be of genotype **PP**, independently from plant to plant. In a large sample of plants, this multinomial model would result in an approximate 3:1 ratio of purple to white plants with all of the white plants and approximately 1/3 of the purple plants breeding true and the other approximately 2/3 of the purple plants breeding as in the F_1 generation. Mendel's (1866) observations are consistent with this multinomial hypothesis. Mendel's model for the inheritance of a single character, in which the particles of inheritance from different gametes come together in an organism and then are passed on unchanged in future gametes has become known as **Mendel's First Law**.

Mendel (1866) also considered the characters two at a time. For instance, he considered the form of the ripe seeds, with round (**R**) and wrinkled (**r**) alleles, and the color of the seed albumen, with yellow (**Y**) and green (**y**) alleles. Mendel crossed a true-breeding parental line in which the form of the ripe seeds was round and the color of the seed albumen was green (genotype **RRyy**) with another true-breeding parental line in which the form of the ripe seeds was wrinkled and the color of the seed albumen was yellow (genotype **rrYY**). When these characters were considered singly, round seeds were dominant to wrinkled and yellow albumen

TABLE 1.1
The sixteen equally-likely genotypes among the F_2 generation (top margin represents gamete contributed by father, left margin represents gamete contributed by mother).

	RY	Ry	rY	ry
RY	RRYY	RRYy	RrYY	RrYy
Ry	RRYy	RRyy	RrYy	Rryy
rY	RrYY	RrYy	rrYY	rrYy
ry	RrYy	Rryy	rrYy	rryy

was dominant to green. All of the F_1 offspring had the yellow and round phenotypes, with genotype **RrYy**. In the F_2 generation, according to the results of the previous experiments, 1/4 of the plants would have the green phenotype and the other 3/4 the yellow phenotype, and 1/4 would have the wrinkled phenotype and the other 3/4 the round phenotype. Thus, if these characters were assumed to segregate independently, we would expect to see 1/16 green and wrinkled, 3/16 yellow and wrinkled, 3/16 green and round, and 9/16 yellow and round, i.e. these phenotypes would occur in a ratio of 1:3:3:9. The experimental numbers corresponding to these categories were 32, 101, 108, and 315, respectively, which is consistent with the 1:3:3:9 ratio (p-value .93). Mendel further experimented with these F_2 plants to verify that each possible combination of gametes from the F_1 generation was, in fact, equally likely (see Table 1.1). ¿From these and other similar experiments in which characters were considered two or three at a time, Mendel concluded that the characters did segregate independently. The hypothesis of independent segregation has become known as **Mendel's Second Law**.

The above example provides an opportunity to introduce the concept of recombination. When two characters are considered, a gamete is said to be **parental**, or **nonrecombinant**, if the genes it contains for the two characters were both inherited from the same parent. It is said to be **recombinant** if the genes it contains for the two characters were inherited from different parents. For instance, in the previous example, an F_1 individual may pass on to an offspring one of the four gametes, **RY**, **Ry**, **rY**, or **ry**. **Ry** and **rY** are the parental gametes, because they are each directly descended from parental lines. **RY** and **ry** are recombinant gametes because they represent a mixing of genetic material which had been inherited separately. Mendel's Second Law specifies that a given gamete has chance 1/2 to be a recombinant.

Fisher (1936) provides an interesting statistical footnote to Mendel's work. His analysis of Mendel's data shows that the observed numbers of plants in different classes actually fit too well to the expected num-

bers, given that the plant genotypes are supposed to follow a multinomial model (overall p-value .99993). That Mendel's data fit the theoretical ratios too well suggests some selection or adjustment of the data by Mendel. Of course, this in no way detracts from the brilliance and importance of Mendel's discovery.

2. Linkage and recombination. Mendel's work appeared in 1866, but languished in obscurity until it was rediscovered by Correns (1900), Tschermak (1900) and de Vries (1900). These three had independently conducted experiments similar to Mendel's, verifying his results. This began a flurry of research activity. Correns (1900) drew attention to the phenomenon of complete gametic coupling or **complete linkage**, in which alleles of two or more different characters appeared to be always inherited together rather than independently, i.e. no recombination was observed between them. Although this seems to violate Mendel's Second Law, an obvious extension of his theory would be to assume that the genes for these characters are physically attached. Sutton (1903) formulated the chromosome theory of heredity, a major development. He pointed out the similarities between experimental observations on chromosomes and the properties which must be obeyed by the hereditary material under Mendel's Laws. In various organisms, chromosomes appeared to occur in homologous pairs, each pair sharing very similar physical characteristics, with one member of each pair inherited from the mother and the other from the father. Furthermore, during **meiosis**, i.e. the creation of gametes, the two chromosomes within each homologous pair line up next to each other, with apparently random orientation, and then are pulled apart into separate cells in the first meiotic division, so that each cell receives one chromosome at random from each homologous pair. In fact, the chromosomes each duplicate when they are lined up before the first meiotic division, so after that division, each cell actually contains two copies of each of the selected chromosomes. During the second meiotic division, these cells divide again, forming gametes, with each resulting gamete getting one copy of each chromosome from the cell. Still, the net result is that each gamete inherits from its parent one chromosome at random from each homologous pair. The chromosome theory of heredity provided a physical mechanism for Mendel's Laws if it were assumed that the independent Mendelian characters lay on different chromosomes, and that those which were completely linked lay on the same chromosome.

An interesting complication to this simple story was first reported by Bateson, Saunders and Punnett (1905; 1906). In experiments on the sweet pea (*Lathyrus odoratus*), they studied two characters: flower color, with purple (dominant) and red (recessive) phenotypes, and form of pollen, with long (dominant) and round (recessive) phenotypes. They found that the two characters did not segregate independently, nor were they completely linked (see Table 2.1). When crosses were performed between a

TABLE 2.1
The counts of observed and expected genotypes in Bateson, Saunders and Punnett's (1906) data. In each of the three subtables, the top margin represents form of pollen, and the left margin represents flower color.

	expected no linkage		observed data		expected complete linkage	
	L	l	L	l	L	l
P	1199.25	399.75	1528	106	1599	0
p	399.75	133.25	117	381	0	533

true-breeding parental line with purple flowers and long pollen (genotype **PPLL**) and one with red flowers and round pollen (genotype **ppll**), in the F_2 generation, there were long and round pollen types and purple and red flowers, both in ratios of 3 to 1 of dominant to recessive types, following Mendel's First Law. However, among the purple flowered plants, there was a preponderance of long-type pollen over round in a ratio of 12 to 1, whereas among the red flowered plants, the round-type pollen was favored, with a ratio of long to round type pollen of 1 to 3. The authors were baffled as to the explanation for this phenomenon which is now known as **linkage** or partial coupling, of which complete linkage or complete coupling is a special case.

It was Thomas Hunt Morgan who was able to provide an explanation for Bateson, Saunders and Punnett's observations of linkage and similar observations of his own on *Drosophila melanogaster*. Morgan (1911), building on a suggestion of de Vries (1903), postulated that exchanges of material, called **crossovers**, occurred between homologous chromosomes when they were paired during meiosis (see Figure 2.1). In the example of Bateson, Saunders, and Punnett (1905; 1906), if a parental line with purple flowers and long pollen were crossed with another having red flowers and round pollen, then the members of the F_1 generation would each have, among their pairs of homologous chromosomes, a pair in which one of the chromosomes had genes for purple flowers and long pollen (**PL**) and the other had genes for red flowers and round pollen (**pl**). During meiosis, when these homologous chromosomes paired, if no crossovers occurred between the chromosomes in the interval between the genes for flower color and pollen form, then the resulting gamete would be of parental type, i.e. **PL** or **pl**. If crossing-over occurred between the chromosomes in the interval between the genes, the resulting gamete could instead be recombinant, **Pl** or **pL** (see Figure 2.1). Without the crossover process, genes on the same chromosome would be completely linked with no recombination allowed, but they typically exhibit an amount of recombination somewhere

FIG. 2.1. *(a) During meiosis, each chromosome duplicates to form a pair of sister chromatids that are attached to one another at the centromere. The sister chromatids from one chromosome are positioned near those from the homologous chromosome, and those four chromatid strands become aligned so that homologous regions are near to one another. (b) At this stage, crossovers may occur, with each crossover involving a nonsister pair of chromatids. (c) At the first meiotic division, the chromatids are separated again into two pairs that are each joined by a centromere. (d) The resulting chromatids will be mixtures of the original two chromosome types due to crossovers. (e) In the second meiotic division, each product of meiosis receives one of the four chromatids. (f) depicts the same stage of meiosis represented by (b), but here only a portion of the length of the four chromatids is shown. Suppose that the interval depicted is flanked by two genetic loci. Consider the chromatid whose lower end is leftmost. That chromatid was involved in one crossover in the interval, thus its lower portion is dark and its upper portion is light, showing that it is a recombinant for the flanking loci. On the other hand, consider the chromatid whose lower edge is second from the left. That chromatid was involved in two crossovers in the interval, thus its lowermost and uppermost portions are both dark, showing that it is non-recombinant for loci at the ends of the depicted interval. In general, a resulting chromatid will be recombinant for an interval if it was involved in an odd number of crossovers in that interval.*

between perfect linkage (0% recombination) and independence (50% recombination). That the chance of recombination between genes on the same chromosome should be between 0 and 1/2 is a mathematical consequence of a rather general assumption about the crossover process, no chromatid interference, described later.

Although we now know that crossing-over takes place among four chromosome strands, rather than just two, the essence of Morgan's hypothesis is correct. In diploid eukaryotes, during the pachytene phase of meiosis, the two chromosomes in each homologous pair have lined up next to each other in a very precise way, so that homologous regions are adjacent. Both chromosomes in each pair duplicate, and the four resulting chromosome strands, called **chromatids** are lined up together forming a very tight bundle. The two copies of one chromosome are called **sister chromatids**. Crossing-over occurs among the four chromatids during this phase, with each crossover involving a non-sister pair of chromatids. After crossing-over has occurred, the four resulting chromatids are mixtures of the original parental types. Following the two meiotic divisions, each gamete receives one chromatid. For genes on the same chromosome, a recombination occurs whenever the chromatid which is passed on to the gamete and which contains the two genes was involved in an odd number of crossovers between the genes (see Figure 2.1).

3. Linkage Analysis. A consequence of the crossover process, Morgan (1911) suggested, would be that characters whose genes lay closer together on a chromosome would be less likely to recombine because there would be a smaller chance of crossovers occurring between them. This is the key to **linkage analysis**: the smaller the amount of recombination observed between genes, i.e. the more tightly linked they are, the closer we could infer that they lie on a chromosome. This provides a way of locating genes relative to one another by observing the pattern of inheritance of the traits which they cause. It is remarkable that a comparison of various traits among family members may yield information on the microscopic structure of chromosomes. Despite many important advances in molecular biology since since Morgan's suggestion in 1911, linkage analysis is still a very powerful tool for localizing a gene of interest to a chromosome region, particularly because it may be used in cases where one has no idea where the gene is or how it acts on a biochemical level.

Modern linkage analysis uses not only genes that code for proteins that produce observable traits, but also neutral **markers**. These are regions of DNA that are **polymorphic**, that is, they tend to differ from individual to individual, but unlike genes, the differences between alleles of neutral markers may have no known effect on the individual, although they can be detected by biologists. While these markers may not be of interest themselves, they can be mapped relative to one another on chromosomes and used as signposts against which to map genes of interest. Genes and

markers are both referred to as **genetic loci**.

As an undergraduate student of Thomas Hunt Morgan, Sturtevant (1913) applied the principle of linkage to make the first genetic map. This consisted of a linear ordering of six genes on the X-chromosome of *Drosophila*, along with genetic distances between them, where he defined the **genetic distance** between two loci to be the expected number of crossovers per meiosis between the two loci on a single chromatid strand. He called this unit of distance one **Morgan**, with one one-hundredth of a Morgan, called a **centiMorgan** (**cM**), being the unit actually used in practice. Sturtevant (1913) remarked that genetic distance need not have any particular correspondence with physical distance, since as we now know, the crossover process varies in intensity along a chromosome. The crossover process generally cannot be observed directly, but only through recombination between the loci. For nearby loci, Sturtevant (1913) took the genetic distance to be approximately equal to the recombination fraction, i.e. proportion of recombinants, between them. Once he had a set of pairwise distances between the loci, he could order them. Of course, it is possible to have a set of pairwise distances which are compatible with no ordering, but in practice, with the large amount of recombination data typically obtained in *Drosophila* experiments, this does not occur. Sturtevant realized that the recombination fraction would underestimate the genetic distance between more distant loci, because of the occurrence of multiple crossovers. There are several obvious ways in which Sturtevant's (1913) method could be improved. First, the recombination fraction is not the best estimate of genetic distance, even for relatively close loci. Second, it is desirable to have some idea of the variability in the maps. Also, depending on what is known or assumed about the crossover process, it may be more informative to consider recombination events among several loci simultaneously. In order to address these issues properly it is necessary to have a statistical model relating observed recombinations to the unobserved underlying crossovers. We proceed to outline some of the issues involved.

Haldane (1919) addressed the relationship between recombination and crossing-over through the notion of a **map function**, that is, a function M connecting a recombination probability r across an interval with the interval's genetic length d by the relation $r = M(d)$. Haldane's best-known contribution is the map function he introduced, and which is now known by his name, $M(d) = [1 - exp(-2d)]/2$. The Haldane map function arises under some very simple assumptions about the crossover process. Recall that crossing-over occurs among four chromatid strands, and that each gamete receives only one of the four resulting strands. We refer to the occurrence of crossovers along the bundle of four chromatid strands as the **chiasma process**. Each crossover involves exactly two of the four chromatids, so any given chromatid will be involved in some subset of the crossovers of the full chiasma process. The occurrence of crossovers along a given chromatid will be referred to as the **crossover process**. To obtain the Haldane map func-

tion, assume first that the chiasma process is a (possibly inhomogeneous) Poisson process. Violation of the assumption is known as **chiasma interference** or **crossover position interference**. Second, assume that each pair of non-sister chromatids is equally likely to be involved in a crossover, independent of which were involved in other crossovers. This assumption is equivalent to specifying that the crossover process is obtained from the chiasma process by independently thinning (deleting) each point with chance 1/2. Violation of this assumption is known as **chromatid interference**, and the assumption itself is referred to as no chromatid interference (**NCI**). This pair of assumptions specifies a model for the occurrence of crossovers which is known as the **No-Interference (NI) model**. Deviation from this model is known as **interference**, which encompasses both chiasma interference and chromatid interference.

Since genetic distance is the expected number of crossovers d in an interval on a single chromatid strand, the assumption of NCI implies that the expected number of crossovers of the full chiasma process in the interval is $2d$. Under the assumption of no chiasma interference, the chiasma process is then a Poisson process with intensity 2 per unit of genetic distance. To obtain the Haldane mapping function, we apply **Mather's Formula** (1935), which says that under the assumption of NCI, $r = [1-P(N=0)]/2$, where r is the recombination probability across an interval, and N is the random variable corresponding to the number of crossovers in the chiasma process in that interval. Under the NI model, $P(N=0) = exp(-2d)$, giving the Haldane map function.

Following is a well-known derivation of Mather's Formula (see e.g. Karlin and Liberman 1983): If we assume NCI, then each crossover has chance 1/2 to involve a given chromatid, independent of which chromatids are involved in other crossovers. In that case, if there are N crossovers in the chiasma process on an interval, with $N > 0$, then the chance of having i crossovers in the crossover process on a given chromatid is

$$\left(\begin{array}{c} N \\ i \end{array} \right) \times \frac{1}{2^i} \times \frac{1}{2^{N-i}}$$

for $0 \leq i \leq N$. On a given chromatid, a recombination will occur in the interval if the chromatid is involved in an odd number of crossovers in the interval. Thus, the chance of a recombination given that $N > 0$ crossovers have occurred in the chiasma process is

$$\frac{1}{2^N} \times \sum_{i=0}^{[[\frac{N-1}{2}]]} \left(\begin{array}{c} N \\ 2i+1 \end{array} \right) = \frac{1}{2},$$

and the chance is 0 if $N = 0$, so the chance of a recombination is $Pr(N > 0)/2$. One consequence of Mather's Formula is that under NCI, the chance of recombination across an interval increases, or, at least, does not decrease,

as the interval is widened. Another is that the chance of recombination across any interval has upper bound 1/2 under NCI. These two observations appear to be compatible with virtually all published experimental results.

Haldane's map function provides a better estimate of genetic distance than the recombination fraction used by Sturtevant (1913). Instead of estimating d by the observed value of r, one could instead plug the observed value of r into the formula $d = -1/2\ln(1-2r)$. One could perform separate experiments for the different pairs of loci to estimate the genetic distances and hence obtain a map. Standard deviations could easily be attached to the estimates, since the number of recombinants in each experiment is binomial.

One could also look at a number of loci simultaneously in a single experiment. Assuming that the experiment was set up so that all recombination among the loci could be observed, the data would be in the form of 2^m counts, where m is the number of loci considered. This is because for each locus, it would be recorded whether the given chromosome contained the maternal or paternal allele at that locus. If we number the loci arbitrarily and assume that, for instance, the probability of maternal alleles at loci 1,3,4 and 5 and paternal alleles at loci 2 and 6 is equal to the probability of paternal alleles at loci 1,3,4 and 5 and maternal alleles at loci 2 and 6, then we could combine all such dual events and summarize the data in 2^{m-1} counts. We index these counts by i, where $i = (i_1, i_2, ..., i_{m-1}) \in \{0,1\}^{m-1}$ and $i_j = 0$ implies that both loci i_j and i_{j+1} are from the same parent, i.e. there is no recombination between them, while $i_j = 1$ implies that loci i_j and i_{j+1} are from different parents, i.e. they have recombined. Fisher (1922) proposed using the method of maximum likelihood for linkage analysis, and this is the method largely used today. We now describe the application, to the type of data described above, of the method of maximum likelihood using Haldane's NI model. This is the simplest form of what is known as **multilocus linkage analysis**.

In a given meiosis, the NI probability of the event indexed by i is simply

$$p_i = \prod_{j=1}^{m-1} \theta_j^{i_j}(1-\theta_j)^{1-i_j} = 1/2 \prod_{j=1}^{m-1} (1-e^{-2d_j})^{i_j}(1+e^{-2d_j})^{1-i_j},$$

where θ_j is the probability of recombination between loci i_j and i_{j+1} and d_j is the genetic distance between them. The formula reflects the fact that under NI, recombination in disjoint intervals is independent. Note that the formulation depends crucially on the presumptive order of the markers. The same recombination event will have a different index i if the order of the markers is changed, and a different set of recombination probabilities or genetic distances will be involved in the above formula. For a given order,

one can write down the likelihood of the data as

$$L(\theta, n) \propto \prod_i p_i^{n_i} = \prod_{j=1}^{m-1} \theta_j^{\sum_{i:i_j=1} n_i} (1-\theta_j)^{\sum_{i':i'_j=0} n_i},$$

where n_i is the number of observations of type i. The likelihood is maximized by

$$\hat{\theta}_j = \sum_{i:i_j=1} n_i \div \sum_{i'} n_{i'}$$

for all j, that is, just the observed proportion of recombinants between loci i_j and i_{j+1}. Since the assumption of NCI implies $\theta_j \leq 1/2$, one usually takes the constrained maximum likelihood estimate, $\hat{\theta}_j = \min(\sum_{i:i_j=1} n_i \div \sum_{i'} n_{i'}, 1/2)$. All other recombination fractions between non-adjacent pairs of loci can be estimated by using the fact that under NI, if loci A, B, and C are in order ABC, then the chance of recombination between A and C, θ_{AC}, is related to the chance of recombination between A and B, θ_{AB}, and that between B and C, θ_{BC}, by the formula $\theta_{AC} = \theta_{AB}(1-\theta_{BC}) + (1-\theta_{AB})\theta_{BC}$. The variance in the estimate $\hat{\theta}_j$ is $\theta_j(1-\theta_j)/n$, and $\hat{\theta}_j$ and $\hat{\theta}_k$ are independent for $j \neq k$. Thus, under the assumption of NI, the multilocus linkage analysis reduces to a pairwise analysis of recombination between adjacent markers when the data are in the form given above.

To estimate order, one may consider several candidate orders and maximize the appropriate likelihood under each of them. The maximum likelihood estimate of order is that order whose maximized likelihood is highest. When one wants to map a new locus onto a previously existing map, one can follow this procedure, considering as candidate orders those orders in which the previously mapped loci are in their mapped positions and the new locus is moved to different positions between them.

Outside of the world of experimental organisms, the reality of multilocus linkage analysis is quite different from what has been portrayed so far. Humans cannot be experimentally crossed, and therefore human linkage data does not fit neatly into 2^{m-1} observed counts. In some individuals, maternal and paternal alleles may be identical at some loci, so that recombination involving those loci cannot be observed in their offspring. Ancestors may not be available for the analysis, so it may not be possible to definitively determine whether particular alleles are maternally or paternally inherited. When some information is missing, the information that is available may be in the form of complicated pedigrees representing interrelationships among individuals. In these cases, multilocus linkage analysis under NI does not reduce to a pairwise analysis. Maximization of the NI likelihood is an extremely complex undertaking and is the subject of considerable current research. For an introduction to linkage analysis in humans, see Ott (1991).

Most linkage analyses, whether in humans or in experimental organisms, are today still performed using the NI model. In fact, the phenomenon of interference is well-documented in a wide range of organisms. In their experiments on *Drosophila*, Sturtevant (1915) and Muller (1916) noticed that crossovers did not seem to occur independently, but rather the presence of one seemed to inhibit the formation of another nearby. ¿From recombination data, it may be impossible to distinguish whether observed interference is due to chromatid interference, chiasma interference, or both, because of a lack of identifiability. If the chiasma and crossover processes themselves could be observed, this would eliminate the difficulty. In certain fungi such as *Saccharomyces cerevisiae*, *Neurospora crassa*, and *Aspergillus nidulans*, the problem is made less acute for two reasons. First of all, these genomes are very well mapped, with many closely spaced loci, and for certain very near loci, the observation of a recombination or not between them is nearly equivalent to the observation of a crossover or not between them. Secondly, in these organisms, all four of the products of meiosis can be recovered together and tested for recombination. This type of data is known as **tetrad data**, as opposed to **single spore** data in which only one of the products of meiosis is recovered. As a result of these features, some tetrad data give approximate discretized versions of the chiasma and crossover processes. From this sort of data, it is clear that chiasma or position interference is present, and that the occurrence of one crossover inhibits the formation of another nearby (Mortimer and Fogel 1974). The existence and nature of chromatid interference has proved more difficult to detect than position interference. Statistical tests of chromatid interference based on generalizations of Mather's formula demonstrate some degree of chromatid interference, but the results are not consistent from experiment to experiment (Zhao, McPeek, Speed, 1995).

Various crossover models that allow for interference of one or both types have been put forward and examined. These include Fisher, Lyon and Owen (1947), Owen (1949, 1950), Carter and Robertson (1952), Karlin and Liberman (1979), Risch and Lange (1979), Goldgar and Fain (1988), King and Mortimer (1990) Foss, Lande, Stahl, and Steinberg (1993), McPeek and Speed (1995), Zhao, Speed, and McPeek (1995). The model used overwhelmingly today in linkage analysis is still the no interference model, due to its mathematical tractability. However, the chi-square model of Foss, Lande, Stahl, and Steinberg (1993), McPeek and Speed (1995), and Zhao, Speed, and McPeek (1995) may now be a viable contender.

4. Conclusion. Mendel showed that through careful quantitative observation of related individuals, the mechanism of heredity of traits could be studied. Linkage analysis, proposed by Morgan in 1911 and still used today, is equally startling in that it is based on the principle that careful quantitative observation of related individuals can actually illuminate the positions of genes on chromosomes. While the phenomenon of linkage

between traits allows one to infer that their genes are on the same chromosome, it is the phenomenon of recombination, that has the effect of varying the degree of linkage, which allows these traits to be mapped relative to one another on the chromosome. One of the most useful characteristics of linkage analysis is the fact that it can be used to map genes that are identified only through their phenotypes, and about which one may have no other information.

5. Recommended reading. Whitehouse (1973) gives a thorough historical introduction to genetics. Bailey (1961) is a detailed mathematical treatment of genetic recombination and linkage analysis, while Ott (1991) is an introductory reference for genetic linkage analysis in humans.

Acknowledgements. I am greatly indebted to Terry Speed for much of the material in this manuscript. This work was supported in part by NSF Grant DMS 90-05833 and NIH Grant R01-HG01093-01.

REFERENCES

Bailey, N. T. J. (1961) Introduction to the Mathematical Theory of Genetic Linkage, Oxford University Press, London.

Bateson, W., Saunders, E. R., and Punnett, R. C. (1905) Experimental studies in the physiology of heredity, *Rep. Evol. Comm. R. Soc.*, **2**: 1-55, 80-99.

Bateson, W., Saunders, E. R., and Punnett, R. C. (1906) Experimental studies in the physiology of heredity, *Rep. Evol. Comm. R. Soc.*, **3:** 2-11.

Carter, T. C., and Robertson, A. (1952) A mathematical treatment of genetical recombination using a four-strand model, *Proc. Roy. Soc. B*, **139**: 410-426.

Correns, C. (1900) G. Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbartarde, *Ber. dt. bot. Ges.*, **18:** 158-168. (Reprinted in 1950 as "G. Mendel's law concerning the behavior of progeny of varietal hybrids" in *Genetics, Princeton*, **35:** suppl. pp. 33-41).

de Vries, H. (1900) Das Spaltungsgesetz der Bastarde, *Ber. dt. bot. Gesell.*, **18:** 83-90. (Reprinted in 1901 as "The law of separation of characters in crosses", *Jl. R. Hort. Soc.*, **25:** 243-248.

de Vries, H. (1903) Befruchtung and Bastardierung, Leipzig. (Reprinted as "Fertilization and hybridization" in C. S. Gager (1910) Intracellular pangenesis including a paper on fertilization and hybridization, Open Court Publ. Co., Chicago, pp. 217-263).

Fisher, R. A. (1922) The systematic location of genes by means of crossover observations, *American Naturalist*, **56:** 406-411.

Fisher, R. A. (1936) Has Mendel's work been rediscovered? *Ann. Sci.*, **1:** 115-137.

Fisher, R. A., Lyon, M. F., and Owen, A. R. G. (1947) The sex chromosome in the house mouse, *Heredity*, **1:** 335-365.

Foss, E., Lande, R., Stahl, F. W., Steinberg, C. M. (1993) Chiasma interference as a function of genetic distance, *Genetics*, **133:** 681-691.

Goldgar, D. E., Fain, P. R. (1988) Models of multilocus recombination: nonrandomness in chiasma number and crossover positions, *Am. J. Hum. Genet.*, **43:** 38-45.

Haldane, J. B. S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors, *J. Genetics*, **8:** 299-309.

Karlin, S. and Liberman, U. (1979) A natural class of multilocus recombination processes and related measures of crossover interference, *Adv. Appl. Prob.*, **11:** 479-501.

Karlin, S. and Liberman, U. (1983) Measuring interference in the chiasma renewal formation process, *Adv. Appl. Prob.*, **15:** 471-487.

King, J. S., Mortimer, R. K. (1990) A polymerization model of chiasma interference and corresponding computer simulation, *Genetics*, **126:** 1127-1138.

Mather, K. (1935) Reduction and equational separation of the chromosomes in bivalents and multivalents, *J. Genet.*, **30:** 53-78.

McPeek, M. S., Speed, T. P. (1995) Modeling interference in genetic recombination, *Genetics*, **139:** 1031-1044.

Mendel, G. (1866) Versuche über Pflanzenhybriden, *Verh. naturf. Ver. Bruenn*, **4:** 3-44. (Reprinted as "Experiments in plant-hybridisation" in Bateson, W. (1909) Mendel's principles of heredity, Cambridge Univ. Press, pp. 317-361.)

Morgan, T. H. (1911) Random segregation versus coupling in Mendelian inheritance, *Science,* **34:** 384.

Mortimer, R. K. and Fogel, S. (1974) Genetical interference and gene conversion, in R. F. Grell, ed., Mechanisms in Recombination, Plenum Publishing Corp., New York, pp. 263-275.

Muller, H. J. (1916) The mechanism of crossing-over, *The American Naturalist,* **50:** 193-221, 284-305, 350-366, 421-434.

Ott, Jurg (1991) Analysis of human genetic linkage, rev. ed., The Johns Hopkins University Press, Baltimore.

Owen, A. R. G. (1949) The theory of genetical recombination, I. Long-chromosome arms. *Proc. Roy. Soc. B*, **136:** 67-94.

Owen, A. R. G. (1950) The theory of genetical recombination, *Adv. Genet.*, **3** : 117-157.

Risch, N. and Lange, K. (1979) An alternative model of recombination and interference, *Ann. Hum. Genet. Lond.*, **43:** 61-70.

Sturtevant, A. H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association, *Jour. Exp. Zool.,* **14:** 43-59.

Sturtevant, A. H. (1915) The behavior of the chromosomes as studied through linkage, *Zeit. f. ind. Abst. u. Vererb.,* **13:** 234-287.

Sutton, W. S. (1903) The chromosomes in heredity, *Biol. Bull. mar. biol. Lab., Woods Hole,* **4:** 231-248.

Tschermak, E. von (1900) Über künstliche Kreuzung bei *Pisum sativum*, *Ber. dt. bot. Ges.,* **18:** 232-239. (Reprinted in 1950 as "Concerning artificial crossing in *Pisum sativum*" in *Genetics, Princeton*, **26:** 125-135).

Whitehouse, H. L. K. (1973) Towards an understanding of the mechanism of heredity, St. Martin's Press, New York.

Zhao, H., McPeek, M. S., Speed, T. P. (1995) A statistical analysis of chromatid interference, *Genetics*, **139:** 1057-1065.

Zhao, H., Speed, T. P., McPeek, M. S. (1995) A statistical analysis of crossover interference using the chi-square model, *Genetics*, **139:** 1045-1056.