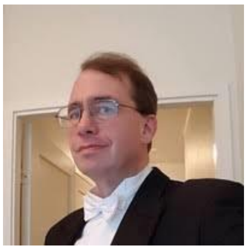
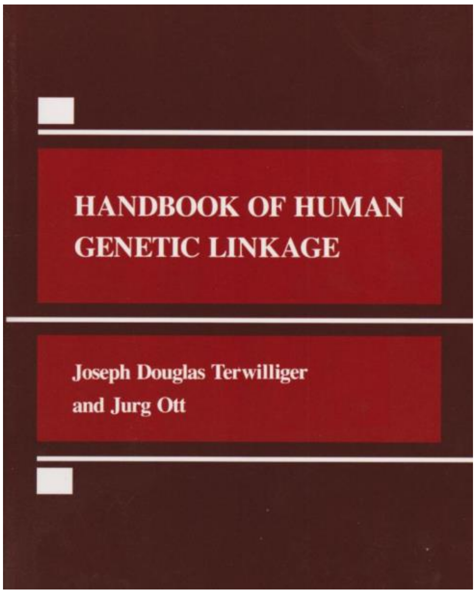
Handbook of Human Genetic Linkage



Joseph Douglas Terwilliger[1]
Jurg Ott[2]

[1]Columbia University, Department of Genetics and Development, New York;
[2]Rockefeller University, New York



ACKNOWLEDGMENTS
The authors would like to acknowledge the assistance of Marcy Speer, Patrick J. Gambino, Yin Yao, and Pentti Tienari for their critical comments, and assistance in preparation of the document.

NOTES TO UPDATED VERSION

We kept the original references in author-year form. New references are consecutively numbered, [1], [2], and so on. The text is largely based on Joe's original manuscript and may deviate occasionally from the printed version.

Most of this book was written by Joe Terwilliger in the early 1990's, at a time without next-generation sequencing (NGS), but the gist of the text is still relevant for people who want to learn how to use linkage analysis. I have been adding more modern material on an ongoing basis, where I thought it would be helpful (example: Phenocopies at the beginning of Chapter 9; appendix B.7 on **sequence data in families**).

With the newest versions of the Linkage programs, handling loops is now automatic – no more need to break loops manually. See chapter 7!

Support through an NSFC grant, project number 31470070, from the Chinese Government is gratefully acknowledged.

Jurg Ott, 8 June 2021

1 Background material

1.1 The purpose of this book

This is a practical guide to human linkage analysis, with emphasis on the use of various computer programs. Little theoretical background will be provided. For theoretical and methodological background and references, we refer the reader to Ott (1999) [1], to which this book is sort of a companion. Since this is predominantly a technical how-to book, we do not want to reference sources already referenced in the companion book except for major issues and new sources.

Text and exercises (files are available online) of the current **online version** are essentially the same as in our original *Handbook* [2] except that software and computers have developed greatly and we updated some of the corresponding references.

Much of this book consists of detailed instructions on how to carry out linkage analyses such that also novices will be able to successfully complete them. The book is intended to be used for self-study or as a manual in linkage courses. Most exercises employ the LINKAGE programs version 5.1, though brief discussions of other programs will be given (version 5.2 is very similar to version 5.1). All exercises in this book are written for execution on Windows or Linux PCs. Some ILINK results may differ slightly between machines. For example, in this book, ILINK analyses were done on a VAX, to illustrate some minor differences between implementations.

To carry out on a PC the exercises described in the subsequent chapters, you need to enter data into a computer file with the aid of a text editor or word processing program. You may use WORD or WordPerfect for that purpose but must make sure that the files you create are in ASCII or "text" file format. To test whether a file is in ASCII or in a word processor's own format, type it out on the screen, that is, issue the DOS command TYPE FNAME, where FNAME is the name of your data file. If it appears normal on the screen, then it is in ASCII format; otherwise you will see strange symbols or the lines will not start at the left margin. Useful text editors are NOTEPAD, NOTEPAD++, or the Crimson Editor; in Linux, I like *gedit* and *Leafpad*.

To obtain copies of the LINKAGE programs, please consult my webpage and the Appendix material.

1.2 Notation and definitions

We sometimes use mathematical notation, which may not be familiar to all readers. Here is a list of symbols and their meaning.

Intervals are indicated with parentheses or brackets, depending on whether the endpoints do or do not belong to the interval. [a,b] defines an interval between the values *a* and *b*, both values inclusive (closed interval), whereas (a,b) denotes an interval excluding the endpoints *a* and *b* (open interval).

That x belongs to a certain set of values or to an interval is indicated by $\in$, which stands for “member of” or “element of”. For example, $x \in [a, b)$ is equivalent to $a \le x < b$.

Binomial coefficients are written in the usual manner as $\binom{n}{k}$ (pronounced *n* choose *k*), where $\binom{n}{k}$ = $n!/[k!(n-k)!]$ with $n!$ (*n* factorial) being defined as $n \times (n-1) \times (n-2) \times ... \times 2 \times 1$.

A vertical bar, $|$, can have more than one meaning. In probability statements, it indicates a condition. For example, $P(X|Y)$ is short for "the conditional probability that X occurs given that Y occurs or is true." For sets of parameter values, the same symbol indicates a restriction on the range of values considered. For example, $(\theta \mid 0.02 < \theta < 0.15)$ reads “values of θ such that they are larger than 0.02 and smaller than 0.15.”

1.3 Review of linkage analysis principles

In this section, a brief summary of human genetic linkage analysis will be given, what it is, and how it works. We will see that there are two different aspects a linkage analysis, testing and estimation. These two ideas will be compared and contrasted, and we will see when, and how to use each of these aspects. Various reviews of linkage analysis and pedigree analysis have been published [3, 4].

1.3.1 Basic introduction to human genetics

As we all know, human beings are sexually reproducing organisms. A man's sperm cell infiltrates and fertilizes his wife's egg cell, with the resulting zygote containing a complete and unique set of genetic information defining many of the biological characteristics of the newly developing human being. This genetic information is stored in coded form within molecules of deoxyribonucleic acid (DNA). DNA is composed of a linear arrangement of smaller molecules, known as nucleotides, whose sequence forms a code which contains information defining the structure of various protein molecules to be synthesized by the cell, the regulation of the production of such molecules, and a great many other functions, the sum total of which define much of who a person will become. While the major function of DNA is the encoding of the structure of various protein molecules, only a small fraction of the total DNA in any cell is actually involved in this process. Those sections of the DNA that are responsible for coding protein structures are called *genes*. They are inherited according to the mendelian laws. In general, any piece of DNA (or of a chromosome such as a secondary restriction) inherited in a mendelian manner is called a *locus*; thus genes are a particular type of loci.

In humans, this linear string of DNA, containing both genes and non-coding sequences, is divided into 23 segments called *chromosomes*. Further, each child receives one copy of each chromosome from the mother, and one from the father, for a total of 46 chromosomes, comprised of 22 pairs of so-called *autosomes*, and one pair of *sex chromosomes*. The sex chromosomes come in variants X and Y, and are involved in the determination of the sex of an individual, with all men having one X chromosome and one Y chromosome (XY), and all women having two X chromosomes (XX).

As was stated above, each offspring has two copies of each chromosome (except that the sex chromosomes in males are X and Y), one derived from each of the parents, and transmits one copy of each chromosome to his or her offspring. Each chromosome segregates independently, meaning, for example, if a mother transmits her maternally derived copy of chromosome 1 to an offspring, she still has an equal chance to transmit her paternally or maternally derived copy of chromosome 2, etc. This division of the DNA into independently segregating chromosomes allows for an increase in diversity of the population, as opposed to forcing each individual to receive half of his DNA from each of two of his four grandparents (i.e. one entire genome as opposed to 23 separate chromosomal entities). However, nature has a way of even further increasing the diversity of the species, through a process known as recombination.

1.3.2 Recombination

Every human being produces germ cells (sperm or egg), containing one copy of each chromosome (*haploid* chromosome set). When a sperm fertilizes an egg, the two haploid chromosome sets are combined making a new zygote containing two copies of each chromosome (*diploid* chromosome set). This diploid cell then develops into a new human being, whose every cell contains an identical full diploid set of chromosomes. However, since each cell contains two identical copies of each chromosome, an obvious question is where do these haploid germ cells come from, and how is it determined which copy they will receive of each chromosome? The answer is that they go through a complicated process called *meiosis*. For a full detailed description of meiosis, the reader is referred to any genetics textbook, like Ayala and Kiger (1984), for example. To sum up the relevant points, when a germ cell is formed, the two *homologous* copies of each chromosome pair up, and each member of a pair goes into one daughter cell and the other member into the other daughter cell (which eventually become gametes). In this distribution of homologous chromosomes, it is random whether the paternally or maternally derived chromosome of each pair goes into a specific daughter cell. Consequently, a gamete will contain some chromosomes from the father and some chromosomes from the mother of the individual producing the gamete.

There is, however, one additional source of variation, which is the main focus of our study. When the pairs of homologous chromosomes line up side by side, they undergo a process called *crossing over* [5], which results in a so-called *recombination* [5]. In recombination, portions of the maternal homolog recombine with the paternal homolog to form a hybrid chromosome in the place of the original ones. Let us assume we have chromosomes MMMMMMMMMM and PPPPPPPPPP lined up beside each other, where M stands for a maternally derived gene and P for a paternal gene. They could recombine in such a way that a crossover takes place between genes 2 and 3 and another one occurs between genes 6 and 7. The two resulting chromosomes would then be represented by MMPPPPPMMM and PPMMMMMPPP. For a more detailed description of recombination involving four chromosome strands see Ott (1991).

Recombination is a frequent process, and it appears that at least one such crossover event must occur on each chromosomal arm (or chromosome) in each meiosis (Sturt, 1976). Small acrocentric chromosomes typically show only one crossover, with larger chromosomes experiencing two or three crossovers. The entire basis of linkage analysis is that recombination events occur between two genetic loci (genes, DNA markers, chromosomal aberrations, etc.) in a rate related to the distance between them on the same chromosome. In other words, loci that are physically very close to each other tend to be inherited together more often than not. The goal of linkage analysis is to determine whether two loci tend to co-segregate more often than they should if they were not physically close together on the same chromosome.

1.3.3 Linkage analysis

We have seen that each individual carries two copies of each chromosome, one derived from each parent. Each of the two chromosomes may carry different variations of the DNA sequence at a given locus. These variations are then referred to as *alleles* (some people use the term gene for allele). Traditionally, the term allele has referred to different variant forms of a protein-encoding gene, that typically had different phenotypic expressions, like the *A*, *B*, and *O* alleles at the ABO blood group. However, with the advent of DNA polymorphisms that are inherited in a mendelian form, the term allele has been expanded to include any mendelianly inherited variation in the DNA sequence at a given locus.

The two homologous chromosomes segregate independently. Therefore, an allele at one locus on one chromosome segregates together with a given allele at another locus on another chromosome with 50% probability. Alleles at loci on the same chromosome should co-segregate at a rate that is somehow related to the distance between them on the chromosome. This rate is the probability of a recombination event occurring between the two loci, or *recombination fraction* (hereafter denoted by θ). Multiple recombination events can occur on the same chromosome. If in a gamete two loci have experienced two crossovers between them, then the final result shows a *non-recombination* between the two loci (for example, the first and last loci on chromosome MMPPPPPMMM considered above).

The recombination fraction ranges from θ = 0 for loci right next to each other through θ = ½ for loci far apart (or on different chromosomes), so that it can be taken as a measure of the *genetic distance* or *map distance* between gene loci. This measure works fine for small distances. The unit of measurement is 1 map unit = 1 centimorgan (cM), corresponding approximately to a recombination fraction of 1%. However, because of the occurrence of multiple crossovers, the recombination fraction is not an additive distance measure. Therefore, it must be transformed by a *map function* into the map distance. For example, the Haldane map function turns θ = 0.27 (27%) into 39 cM, and the Kosambi map function translates θ = 0.27 into 0.30 Morgans (30 cM, centimorgans; see Ott (1991), or Liberman and Karlin (1984), for more information).

Two loci are said to be genetically *linked* when θ < ½, and the phenomenon that this occurs is termed *genetic linkage*. The object of *linkage analysis* is to estimate θ and to test if θ is less than ½, that is, whether an observed deviation from 50% recombination is statistically significant. The estimate of the recombination fraction, usually denoted by $\hat{\theta}$, is in simple cases the proportion of recombinants (proportion of children carrying a recombinant gamete) out of all opportunities for recombination and ranges in principle between 0 and 1. Because maximum likelihood estimates are defined on the set of admissible parameter values and the recombination fraction cannot exceed ½ (unless there is so-called chromatid interference), its estimate is usually also restricted to [0, ½].

Notice that the term linkage refers to *loci*, not to specific alleles at these loci. For example, it is wrong to say that, in a given pedigree, the disease gene is linked with the *A* allele at the marker locus. In a child, alleles at different loci are said to be *in coupling* (as opposed to being in repulsion) when they originated from the same parent. Further, two loci residing on the same chromosome are said to be *syntenic*; they may or may not be linked.

A rudimentary test of linkage between two loci could be set up by comparing, in a chi-square test, an observed number *k* of recombinations and *n* – *k* of non-recombinations with their expected numbers of *n*/2 each under no linkage. The main problem with this, however, is that in most human pedigree data, it is not possible to count recombinants and non-recombinants. For this reason, people typically use likelihood-based methods for testing linkage. Using sophisticated analysis programs, like LINKAGE, it is possible to evaluate the likelihood of a given pedigree under different assumptions about the recombination fraction between two loci. Further, since the Neymann-Pearson lemma tells us that if there is a best test of a given hypothesis, it is of the form of a likelihood ratio test, we have good theoretical basis for using the likelihood ratio test as our test of choice. In linkage analysis, our likelihood ratio is formed as L(θ)/L(θ = 0.5), with the denominator corresponding to the likelihood of our data under the assumption of no linkage. Likelihoods will be discussed in more detail in the exercises to follow.

In linkage analysis, the test is typically formulated in terms of the common (base 10) logarithm of this ratio, or *lod score*. The formula for the lod score is $Z(\theta) = \log_{10}[L(\theta)/L(0.5)]$, or equivalently, $Z(\theta) = \log_{10}[L(\theta)] - \log_{10}[L(\theta = 0.5)]$. These log likelihoods are then calculated via one of the many available linkage analysis programs. The main emphasis of this book will be on how to utilize the LINKAGE program package to compute these lod scores in practical situations. In some simple cases, the likelihoods can be evaluated by hand, as you will see later, but in the majority of family data, this is either impractical or impossible. This is true especially when we have diseases with complicated modes of inheritance as one of our loci.

The most common application of linkage analysis is to try and find the location, in the genome, of a gene responsible for a certain mendelianly inherited disease. In these situations, we often have complicated modes of inheritance, in which we are not certain which individual has which alleles at the disease locus. Consider, for example, the situation where we have a dominant disease. Assuming that *D* represents the dominant disease-causing allele, and + represents the normal or "wild type" allele, we must define the genotype-phenotype relationships or *penetrances*. In this case, we know that unaffected individuals have genotype +/+, and that affected individuals have either genotype D/+ or D/D, however, we cannot discern one from the other phenotypically. The computer programs will compute accurate likelihoods, allowing for both possibilities for each affected individual. As you will see in the course of this book, likelihoods can be computed for these, as well as much more complicated penetrance models. The important thing for our purposes now, is to understand how these likelihoods can then be converted to lod scores, and what these lod scores tell us.

Consider the situation that among children of a phase-known marriage we can count recombinants and non-recombinants. The total number of meioses (children) is *n*, of which *k* are recombinants and *n* – *k* are non-recombinants. Let *r* be the true recombination fraction (the probability that a recombination occurs) while θ is a formal parameter, that is, the recombination fraction assumed in likelihood calculations. As outlined in my book [1], the probability that *k* of *n* children are recombinants is equal to

$$P(k)=\binom{k}{n} r^k (1-r)^{n-k}$$

and the corresponding lod score is given by $Z(\theta) = \log_{10}[P(k;\theta)/P(k;0.5)] = \log_{10}[2^n \theta^k (1-\theta)^{n-k}]$.

1.3.4 Testing

Basically there are two aspects of linkage analysis that can be performed by the use of such lod scores. The first is a test of linkage. In other words, do our data provide us with sufficient information to say that we have found linkage between our two genes? Since we usually have marker loci with known genetic location, and a disease for which we want to find the genetic cause, we can rephrase this test as "Is there evidence for linkage of our disease gene to our marker locus?". We have already defined a test statistic, the lod score, typically denoted as Z(θ), but we have not yet decided how to apply this test. In other words, for what critical region is our test statistic determined to provide sufficient evidence to say that we have found linkage to our disease gene? By convention in linkage analysis, people accept a critical value for this statistic of 3 as significant evidence for linkage. A theoretical examination of this cutoff point (see Ott, 1991) assures that only 1/20 times will a lod score of 3 be spurious for mendelian disorders, and so this is taken as the minimum acceptable level for a significant test with a simple autosomal disease. Similarly, for diseases known from segregation analysis to be on the X-chromosome, a lod score of 2 is considered to represent a significant linkage finding, since the prior probability of linkage is much higher than for an autosomal disease. For complex diseases, however, this lod score of 3 threshold may be too low, but we will defer further discussion of this until chapter 25.

The lod score test is usually performed by maximizing the lod score over all values of θ on the interval [0, 0.5]. If the maximum of this lod score curve exceeds 3, the test of linkage is significant, and the location of our disease has likely been found. But, if we are doing a genomic screen, what should an investigator do when he finds a lod score that is not significant, yet is still quite large, since typically not all potentially available families are analyzed in the early stages of a linkage analysis? In general, if a lod score of around 2 is found, it may be advisable to type further families for that marker, or to look for other markers in the vicinity of that one. If upon typing further individuals, the lod score drops, then it was most likely spurious, but it may also rise, to exceed the threshold of 3, in which case you have found a significant linkage.

1.3.5 Estimation

After significant evidence for linkage of the disease to a given marker has been found, the next step is to determine the exact location of this gene, to make it easier to isolate, and eventually study the gene itself. We have already explained that there is a monotonic relationship between recombination fraction and physical distance on the chromosome, so if we can determine the recombination fraction between the disease and marker, we will have some idea where to look for the gene itself. If you remember, the lod score test was based on the maximum of the lod score, maximized over θ. We know that the maximum of the likelihood function occurs at the same point as the maximum of the log of the likelihood function, so we can just find the value of θ at which our lod score is maximized, and use this as our estimate of the recombination fraction between disease and marker. This is referred to as the *maximum likelihood estimate* (MLE) of θ, and is denoted by $\hat{\theta}$.

Standard likelihood theory tells us that we can obtain a consistent estimate of any parameter given a set of data (AND the correct model), by maximizing the likelihood of the data with respect to that parameter, that is, in the limit of a large number of observations, the MLE is unbiased with a variance tending to become zero. For finite data sets, however, MLEs in human genetics are generally biased. In the presence of modelling or diagnostic or marker typing errors [6], MLEs may be inconsistent (asymptotically biased).

In addition to point estimates discussed above, one may also obtain interval estimates. Two types of intervals are discussed below, *confidence intervals* and *support intervals*.

Confidence intervals for a parameter such as the recombination fraction or a gene frequency are intimately connected with statistical tests about the parameter in question. On the basis of a set of observations, one may test the null hypothesis, H_0:$p = p_0$, whether the parameter estimate is significantly different from an assumed parameter value, p_0. The test may be carried out for a multitude of parameter values, p_0. The set of all those parameter values p_0 for which the test is not significant constitutes the confidence interval for *p*. Therefore, a significant test result for some value p_t implies that p_t is outside the confidence interval for *p*, and vice versa.

A support interval is, in principle, quite a different construct. It is based on the *support* (a synonym for $\log_e$ likelihood) for a parameter provided by a set of observations. The *m*-unit support interval (Edwards, 1992) consists of all those parameter points with associated $\log_e$ likelihood within *m* units of the maximum $\log_e$ likelihood; the parameter values inside a support interval are considered "plausible" because their support is only *m* units lower than that of the best supported parameter value. Examples of how to compute confidence and support intervals are given in Appendix A (BINOM program). In human genetics, support is also used to mean $\log_{10}$ likelihood. For example, one speaks of 1-lod-unit or 3-lod-unit support intervals.

There is a connection between confidence and support intervals. Consider a regular test situation in which you want to test the null hypothesis that a parameter *p* has a certain value, p_0. You do this by obtaining a maximum likelihood estimate, , and want to contrast versus p_0. Under the null hypothesis, the test statistic, $X^2 = 2ln[L(\hat{p})/L(p_0)]$, follows an asymptotic chi-square distribution with 1 degree of freedom (df), where L(*p*) denotes the likelihood at the value *p*. In this situation, a 2-unit support interval may be interpreted as an approximate 95% confidence interval for *p*, and a 3.32-unit support interval as an approximate 99% confidence interval.

The above discussion is relevant to the test of the null hypothesis of no linkage (H_0: $\theta = 0.5$) versus linkage (H_1: $\theta < 0.5$). As is well-known, the test is declared significant when $Z_{max} \geq 3$. True to the intimate relation between statistical tests and confidence intervals, a confidence interval should contain those values, θ_0, for which the test of H_0: $\theta = \theta_0$ is not significant. In linkage analysis, we construct support intervals rather than confidence intervals but also expect a meaningful relationship between support interval and test result. Consequently, the support interval associated with the test criterion $Z_{max} \geq 3$ must be a 3-lod-unit support interval. The earlier recommendation (Conneally et al., 1985) of testing with $Z_{max} \geq 3$ but constructing 1-lod-unit support intervals leads to an inconsistency between statistical test and support interval when $1 < Z_{max} < 3$. Therefore, that recommendation also stated that no support interval should be constructed when $Z_{max} < 3$. We feel that this solution to the problem is unsatisfactory and that the test for linkage and its associated support interval should be consistent. Therefore, we recommend the use of 3-lod-unit support intervals. The exercises in this book adhere to this rule.

We realize that support interval and confidence interval are constructs belonging to different schools of thought (likelihood approach versus statistical testing approach); each has merits in its own right. We hope that our way of intertwining these two constructs is not offensive to representatives of either school.

1.4 Installing the LINKAGE programs

We assume here that you obtained the PC version of the LINKAGE programs from our website, which provides detailed instructions for installation of these programs. A more efficient and sophisticated program version is FastLINK, notably in Linux. Users are encouraged to install FastLINK rather than the regular LINKAGE package. This is especially important for pedigrees with loops (see section 7).

Part I: Two-point Linkage Analysis

2 File system used by LINKAGE

In this introductory chapter, we will go over the basic file structure of the LINKAGE programs (version 5.1). Topics covered will include the basics of how to enter pedigree data in preparation for doing a linkage analysis. A schematic view of how files and programs interact is shown in Figure 2-1.

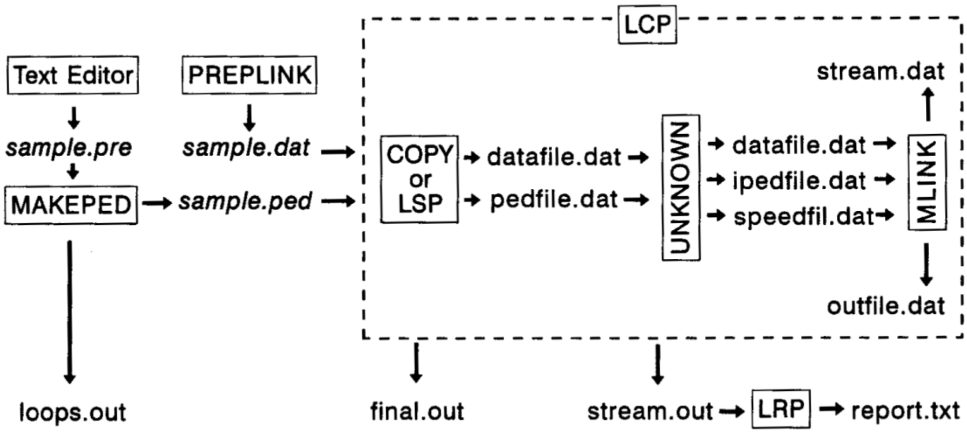


Figure 2–1. Schematic view of how files and programs interact

2.1 PEDIGREE DRAWINGS

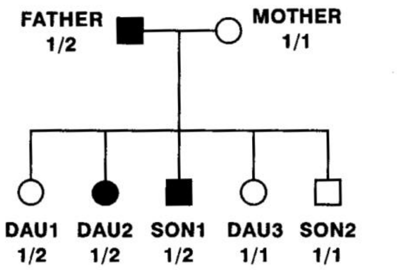


Figure 2–2. Pedigree drawing for EX1.PRE

First, let us assume the pedigree structure and data shown in Figure 2-2. For those of you not too familiar with this type of diagram, circles refer to females, and squares refer to males, individuals who are filled in are affected with the disease in question, while those who are white are unaffected. Further, under each individual, his marker data are given. In this example there is one marker locus indicated.

2.2 PEDIGREE FILES

The first thing you must do is to create a pedigree file with your word processor, in which you describe the pedigree to be analyzed. In such a file, you must enter one line per individual, containing the following information:

Column 1 : Pedigree identifier	The identifier can be a number or a character string
Column 2 : Individual's ID	The identifier can be a number or a character string
Column 3 : The individual's father	If the person is a founder, just put a 0 in each column
Column 4 : The individual's mother	If the person is a founder, just put a 0 in each column
Column 5 : Sex (gender)	(1 = Male, 2 = Female)
Column 6+: Genetic data	(Disease and Marker Phenotypes)

In this case, our first genetic locus is the disease, which will be coded as an affection status type of locus. (In the LINKAGE programs, there are 4 different ways of entering the phenotypic data, called locus types. Their usage will be explained in more detail later.) You should then enter the phenotypic data by entering a *2* if the individual is affected, a *1* if unaffected, and a *0* if the person's affection status is unknown. In this case, the second genetic locus is our marker, which we will code as an allele numbers type of locus. This is the most straightforward way of entering codominant marker information. To do this, you must enter the allele number corresponding to each of the two alleles, separated by at least one space. The most important thing to remember about this type of locus is that you must number your alleles with integers starting from 1. For example if you had a two allele locus at which you had alleles 3.6, and 5.2, you would have to renumber them as 1 and 2 to enter them in your pedigree file as an allele numbers type of locus. If an individual has not been typed, and the marker phenotype is unknown, you must enter 0 for each allele (eg. Phenotype = *0 0*). Note that an individual cannot have one allele known, and the other unknown in this type of locus. They must be either both known, or both unknown! A clever way of evading this problem in simple situations will be dealt with in chapter 10.

For example, let us enter the pedigree data for the pedigree shown in Figure 2-2, in a file called EX1.PRE. We will assign the pedigree the name *ex1*, and give the pedigree members the names *father*, *mother*, *dau1*, *dau2*, *son1*, *dau3*, and *son2*, where *dau* is short for daughter. Most work will be done in a command (cmd) window (also called DOS box). If you use NOTEPAD as your text editor, you may type *notepad ex1.pre* to which the editor responds by saying that no such file can be found and whether you want to create it.

To enter the father, we would first enter the pedigree name *ex1*, followed by a space, and in the next column, we'd type *father*, the individual's name (or ID number). Since he is a founder, we'd enter zeroes in the next two columns, since his parents are unknown (NOTE: Either both parents must be unknown, or neither parent. If one parent is known, but the other is unknown, you would have to add a “dummy” parent with unknown phenotypes, and no parents). Next, you would enter a 1 for his sex (1 = male, 2 = female). The father is affected, so you would type a 2 in the affection status column, and then at the marker he has alleles 1 and 2, so you would enter 1 and 2 in the next two columns, followed by <Enter>. Now, the line in your pedigree file for this individual should look like this:

```
ex1  father 0 0  1  2  1 2
```

The number of spaces between fields doesn't matter, but making the spacings of different lengths will help you recognize alleles belonging to the same locus. Now, enter the rest of the pedigree on your own. *CAUTION: Be certain that your word processing program produces an ASCII file, with the file ending immediately after the last line. NO blank lines are permitted at the end, nor blank spaces on the line following the last individual*. To ensure that your input file does not contain any empty trailing lines, press Ctrl-End, which will position the cursor at the end of the file. If this position is not in column 1 of the line immediately following the last pedigree data line, press the Backspace key repeatedly until the cursor is all the way to the right on the last pedigree data line, then press Enter.

```
    ex1     father      0           0           1     2     1     2
    ex1     mother      0           0           2     1     1     1
    ex1     dau1        father      mother      2     1     1     2
    ex1     dau2        father      mother      2     2     1     2
    ex1     son1        father      mother      1     2     1     2
    ex1     dau3        father      mother      2     1     1     1
    ex1     son2        father      mother      1     1     1     1
```

2.3 MAKEPED

Now that you've entered these data into a pedigree file, save it as *EX1.PRE*. Next, you will need to process this pedigree file with the MAKEPED program, which will add several pointers required by the LINKAGE programs. To do this, type the following at the DOS prompt.

```
MAKEPED infile outfile n
```

where *infile* is the name of the pedigree file without pointers (*EX1.PRE* in this case) and *outfile* is the name of the file to be created by MAKEPED (*EX1.PED* in this case). It is a good general convention to use the extension .PRE to refer to a pedigree file before it is processed by MAKEPED, and to use the extension .PED afterwards. The letter *n* on the command line is optional and tells the program that no loops are present and all probands are to be selected automatically (see below); with *n* as the third parameter on the command line, MAKEPED will run without querying the user as outlined below. If no *n* is given, the program will then ask

```
      Does your pedigree file contain any loops? (y/n) →
```

to which you will respond *n*, since there were neither consanguinity nor marriage loops in this simple pedigree. We will discuss how to handle loops in a subsequent exercise. Next, it will ask

```
      Do you want probands selected automatically? (y/n) →
```

In this case, you would enter *y*, since we will not be calculating genetic risks. The MAKEPED program will now run for a few seconds (actually, the program will further run a separate program called LOOPS, which checks for undeclared loops, and will be discussed later), and produce a pedigree file EX1.PED that is readable by the LINKAGE programs. If you look at the file in your word processor, you will notice that all ID numbers are now integers, and not characters any more, and there are several extra columns. The meanings of each column are as follows:

Column 1: Pedigree Number
Column 2: Individual ID number
Column 3: ID of father
Column 4: ID of mother
Column 5: First Offspring ID
Column 6: Next Paternal Sibling ID
Column 7: Next Maternal Sibling ID
Column 8: Sex (*1* = Male, *2* = Female. Unknown sex not permitted)
Column 9: Proband Status (*1* = proband, higher numbers indicate doubled individuals formed in breaking loops. All other individuals have a *0* in this field.)
Cols 10+: Disease and Marker Phenotypes (as in the original Pedigree File.)

Also, at the end of each line, the program will give the original pedigree names, and individual names from your pedigree file, so you can still identify which individual in the processed file corresponds to which person in the *.PRE file. However, to be safe, it is always better to make any future modifications to the pedigree data in the *.PRE file, and then rerun MAKEPED.

2.4 PARAMETER FILES (PREPLINK)

Now that you've specified the pedigree to be analyzed, it is necessary to generate a parameter file, in which you define the model parameters for each locus in the pedigree file, and other parameters required for the analysis. To create this file, use the PREPLINK program. Just type PREPLINK at the DOS prompt to begin running this program.

The screen should appear similar to the following:

```
****************PRESENT STATUS******************
(a) Number of Loci                   : 2
(b) Sexlinked                        : N
(c) Calculate Risk                   : N
(d) Mutation                         : N
(e) Haplotype Frequencies            : N
(f) Locus Order                      : 1 2
(g) Interference                     : N
(h) Recombination Sex Difference     : N
(i) Program Used                     : MLINK
(j) Recombination Values             :
       0.100
**********OTHER OPTIONS*************************
(k)   See or Modify Loci Description
(l)   See or Modify Recombination to Vary
(m)   Read Datafile
(n)   Write Datafile
(o)   Exit
***************************************************
Enter letter to see or modify values
```

Now, we will have to make the specifications in this program match our desired analysis. To begin with, we should check the first line *(a) Number of Loci*, which is currently given a value of 2 by default. Since this is correct for our analysis, we needn't change this. The second line, *(b) Sexlinked*, would be used to tell the program whether we will be using autosomal markers or X-linked ones. Currently, the default value is *N*, meaning the disease and markers are autosomal, so we needn't alter this. Next, we come to the option *(c) Calculate Risk*. If one wanted to compute genetic risks, one could select this option to specify the risk locus and allele. Similarly, option *(d) Mutation* would allow one to specify one locus at which mutations can occur. However, there is the additional restriction that mutation can only occur from one specific allele to another. Hence it is primarily useful for disease loci, with normal alleles being mutated into disease alleles with a specific frequency. For our purposes, however, we will assume the absence of mutation, so we can ignore this option, which by default is set to *N*o. Option *(e) Haplotype frequencies* is also set to *N*o by default, since by default, the programs assume linkage equilibrium, and compute haplotype frequencies from gene frequencies at each locus. If one wanted to incorporate linkage disequilibrium data in the analysis, it would be imperative to specify haplotype frequencies for all possible haplotypes with this option. We assume linkage equilibrium, so we can leave this set at the default as well. The locus order can be input here as well, but since we have only two loci, it is immaterial, so we should leave it at the default setting, *1 2*. *(g) Interference* has only been incorporated in a very rudimentary fashion, as you will see in later chapters. For now, just ignore this option, and leave it set to *n*o, since it is unavailable for general use. Option *(h) Recombination Sex Difference* can be very important, since in general there are different rates of recombination in male and female meioses. This is discussed in detail in Part II, and for now, we will assume that there is *N*one. Option *(i)* allows you to choose the program with which to perform the analysis. We will stick with the default program, MLINK, and defer further discussion of this until later. Similarly, the *(j) Recombination values* option allows you to set the recombination fraction at which to compute lod scores. In general, it is not necessary to specify program used, recombination values, recombination sex difference, or locus order in PREPLINK, since you can override these choices interactively when running the LCP program, as you will see in upcoming chapters.

Next, we must specify the genetic parameters which define the loci to be analyzed. To do this, you must now choose option *(k) See or modify loci description*. When you type *k*, followed by pressing the <Enter> key, you will see a screen like the following:

```
****************************************************
(1) Allele Numbers GENE FREQS : 0.500000 0.500000
(2) Allele Numbers GENE FREQS : 0.500000 0.500000
****************************************************
(a) SEE OR MODIFY A LOCUS
(b) DELETE LOCUS
(c) ADD LOCUS
(d) CHANGE ORDER TO CORRESPOND TO PEDIGREE FILE (NOT CHROMOSOME ORDER)
(e) CHANGE LOCUS TYPE
(f) RETURN TO MAIN MENU
****************************************************
enter letter to modify values
```

We will now make Locus 1 correspond to the first locus in our pedigree file, which was the disease locus. To change locus 1 from allele numbers to affection status, choose option *(e)*, to which you will be prompted

```
ENTER LOCUS TO CHANGE
```

to which you should respond *1*. Next you will be given a menu of options as follows:

```
ENTER NEW LOCUS TYPE:
(a) BINARY FACTORS
(b) QUANTITATIVE TRAIT
(c) AFFECTION STATUS
(d) ALLELE NUMBERS
```

You should choose *(c) AFFECTION STATUS*, after which you will see a menu like the one above, with allele numbers changed to affection status in the description of locus 1. We still must modify the other parameters, like gene frequency and penetrances at locus 1, so choose option *(a) SEE OR MODIFY A LOCUS*, and specify locus 1. You will then see a current default description of locus 1, as follows:

```
******************************************
LOCUS NUMBER:            1
******************************************
(a) Number of Alleles         : 2
(b) Number of Liability Classes     : 1
(c) Penetrances:
GENOTYPE    1 1   0.000000
GENOTYPE    1 2   0.000000
GENOTYPE    2 2   1.000000
(d) Gene Frequencies :
 0.500000 0.500000
(e) EXIT
******************************************
enter letter to modify values
```

Since we have assumed two alleles (one normal, one disease), the *Number of Alleles* is correct. Likewise, line *(b) Number of Liability Classes : 1* is correct. We'll get into the meaning and application of liability classes later on. Let us just assume for the moment that we have a fully penetrant dominant disease (for more information about penetrance, see chapter 5), and that allele *2* is the disease allele. If this is the case, our *(c) Penetrances* must be modified to reflect this. The penetrances given as default correspond to a recessive disease with full penetrance. (Can you see this?) Now enter *c* to modify the penetrances to correspond to a fully penetrant dominant disease as follows (the program presents the old penetrance value and prompts you with ?, at which you must enter the new value, which may or may not coincide with the old value.) :

```
ENTER NEW PENETRANCES
GENOTYPE 1 1 OLD PEN 0.000000
?
0
GENOTYPE 1 2 OLD PEN 0.000000
?
1
GENOTYPE 2 2 OLD PEN 0.000000
?
1
```

Once you've responded as above, your locus description will again be shown with these modified values. Note that you must enter a new penetrance, followed by <Enter> whenever you are prompted, even if it will remain the same as the default value. Next, you must modify the gene frequencies, since the disease allele is certainly not at such a high frequency in the population. Let us assume the disease allele has a population frequency of 0.00001, giving the normal allele a frequency of 0.99999. So, choose option *(d)*, and respond as follows:

```
ENTER 2 NEW GENE FREQUENCIES
0.99999 0.00001
```

The order in which you enter them is important, since you defined the penetrances above in such a way that allele 2 is the disease causing allele (can you see this?), so you must be certain that allele *2* receives the correct gene frequency of 0.00001. Now, this locus is properly specified, so we can choose option *(e) Exit* to take us back to the menu screen where each locus is specified. The top should now look like this:

```
*******************************************
(1) affection status    GENE FREQS : 0.999990 0.000010
(2) allele numbers      GENE FREQS : 0.500000 0.500000
*******************************************
```

Now, we should look at locus 2, by choosing *(a) SEE OR MODIFY A LOCUS*, and specifying locus 2. We will then see a screen like this:

```
*********************
Locus Number : 2
*********************
(a) Number of Alleles: 2
(b) Gene Frequencies:
 0.500000 0.500000
(c) EXIT
*********************
enter letter to modify values
```

Since there are only two alleles at this codominant marker locus, and we are assuming equal gene frequencies, we needn't change anything here. In general, one ***must*** have reliable estimates of the gene frequencies for all alleles at each locus, as it has been repeatedly demonstrated (Ott, 1992, for example) that assuming equal gene frequencies for a given marker locus can lead to increased false positive evidence for linkage when the true gene frequencies deviate from equality (almost always, this is the case). At this time, you should *(c) EXIT*, followed by *(f) RETURN TO MAIN MENU*, and *(n) WRITE DATAFILE*. You will then be asked to supply the name of the file to be saved, which should be *EX1.DAT*. The extension **.DAT* is used by convention to refer to the Parameter File for the analysis. You may next choose *(o) EXIT*, as you have finished specifying the parameters for the analysis. If you wish, you may now look at this parameter file in your word processor. It should look like the following:

```
2 0 0 5 << NO. OF LOCI, RISK LOCUS, RISK ALLELE, SEXLINKED (IF 1) PROGRAM
0 0.0 0.0 0 << MUT LOCUS, MUT RATE, HAP FREQUENCIES (IF 1)
 1 2
1 2 << AFFECTION, NO. OF ALLELES
 0.999990 0.000010 << GENE FREQUENCIES
 1 << NO. OF LIABILITY CLASSES
 0.0000 1.0000 1.0000 <<PENETRANCES
3 2 << ALLELE NUMBERS, NO. OF ALLELES
 0.500000 0.500000 << GENE FREQUENCIES
 0 0 <<SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.1000 << RECOMBINATION VALUES
 1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

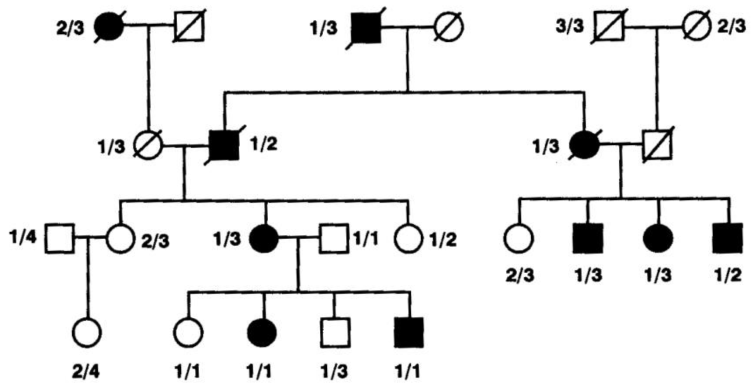


Figure 2–3. Pedigree drawing for USEREX2.*

These are just the parameters you selected in PREPLINK, presented in a format readable by the LINKAGE programs. The bottom three lines are specific for the program to be used for the analysis, but for our purposes at this time, we will ignore this section of the file.

At this point, we have learned how to enter pedigree data in a form readable by the LINKAGE programs. We have also learned how to use the MAKEPED and PREPLINK programs to help in making these files. Throughout your career in linkage analysis, you will keep coming back to these programs, so it is important to understand them, and become fluent in the usage of these programs. Before going on to the next chapter, where we'll actually begin to do our own linkage analyses, we will do a practice example with entering another set of data in the required formats.

EXERCISE 2

For the pedigree shown in Figure 2-3, please create pedigree and parameter files (USEREX2.*), in the way we learned in this chapter. In this drawing, the number directly under each individual is his marker phenotype, if known. Use PREPLINK to generate a parameter file, specifying the first locus to be a fully penetrant dominant disease (Affection Status locus type) with disease allele frequency of 0.00001, and the second locus to be a codominant locus (Allele Numbers locus type) with three equally frequent alleles (gene frequencies = 0.33333). Remember that unknown individuals are coded as *0* at an affection status locus, and *0 0* at an allele numbers locus. Watch out for an intentional “typo”!

3 Running the LINKAGE programs MLINK and ILINK

In this chapter, you will be performing your first real 2-point linkage analyses, using the pedigree files you created in the previous chapter. We will be using the MLINK, and ILINK programs to perform these analyses. After this chapter, you will be able to do your own basic analyses. In practice, one typically does not use the analysis programs directly, but sets up an analysis using the LCP program, which is discussed in chapter 4. For some types of application, however, it is important that one knows how to handle the analysis programs directly.

3.1 THEORETICAL ANALYSIS

Let us begin this first linkage analysis with a theoretical analysis of the pedigree from Figure 2-2 in the previous chapter, for which you have already made the pedigree file (EX1.PED) and the parameter file (EX1.DAT). Let us examine this family more closely. The only way a meiosis can provide information about linkage is when the parent in which the meiosis occurred is heterozygous at both loci. Otherwise, no information can be obtained about linkage from this parent (i.e. if a parent is *1/1* at a marker locus, there is no way to tell which *1* allele was transmitted to any given offspring, so no linkage information is available). Since *mother* is homozygous at the marker locus, and also at the disease locus (since it is a fully penetrant dominant disease, she must be homozygous normal to be unaffected), she is uninformative for linkage. So, we need only look at the fate of the paternally derived alleles in this family. We know that *father* is heterozygous at the marker locus (*1/2*), and also at the disease locus (since he is affected, yet he has unaffected children he must carry one disease allele, and one normal allele). Hence, he is informative for linkage. However, we do not know in what phase these alleles exist in *father*, but we know there are only two choices, either he has phase D 1 / N 2 or he has phase N 1 / D 2. This type of nuclear family is called a phase unknown pedigree, since only genotype information is available, and not haplotype information on the doubly heterozygous parent. Each of these phases then has an equal 50% chance of being correct a priori (since we assumed absence of linkage disequilibrium). So, we can then examine the offspring to count recombinants and non-recombinants under each phase. We can disregard what each child got from the mother, and consider the pedigree to be reduced to what is shown in Figure 3-1, containing only the alleles derived from *father*.

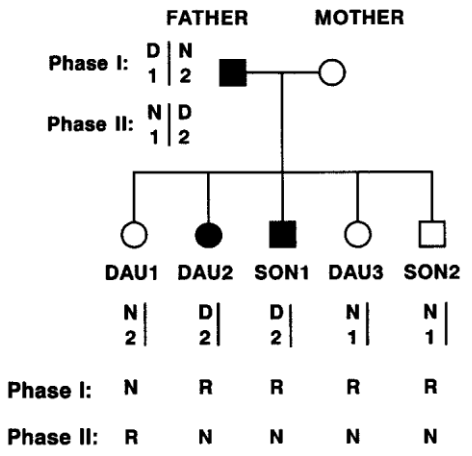


Figure 3–1. Linkage information contained in EX1.*

Note that there are 3 different haplotypes observed in the children, (a) D 2, (b) N 1, and (c) N 2. If phase 1 were correct, haplotypes (a) and (b) would both be recombinants (i.e. non-parental types), and haplotype (c) would be non-recombinant (i.e. parental type). Similarly, under phase 2, the opposite situation would pertain, and haplotypes (a) and (b) would be non-recombinant, and haplotype (c) would be recombinant. Thus, under phase 1, we would have 4 recombinants and 1 non-recombinant in this family, and under phase 2, we would have 4 non-recombinants and 1 recombinant. Since the probability of a recombination is equal to θ (the so-called *recombination fraction*), and the probability of a non-recombination is therefore equal to (1 – θ), we can calculate the likelihood of this pedigree as a function of θ. Under phase 1, the probability of observing the data would be equal to P(data) = $K\theta^4(1-\theta)$, and under phase 2, the probability of the data would be $K\theta(1-\theta)^4$, where K is a constant coefficient in each case. Since each phase has a prior probability of 0.5, we can compute the probability of the phase unknown family observed as follows by the law of total probability: P(data) = P(Phase 1)P(data | Phase 1) + P(Phase 2)P(data | Phase 2); which in this case is equal to: P(data) = $(0.5)[K\theta^4(1-\theta)] + (0.5)[K\theta(1-\theta)^4]$.

Since the likelihood is defined as P(data), we can set up a likelihood ratio test for linkage as [1]

$$\Lambda = \frac{P(data|\theta)}{P(data|\theta=0.5)} = \frac{L(\theta)}{L(\theta=0.5)} = \frac{K[\frac{1}{2}\theta^4(1-\theta)+\frac{1}{2}\theta(1-\theta)^4]}{K[\frac{1}{2}(0.5)^4(0.5)+\frac{1}{2}(0.5)(0.5)^4]}.$$

Since *K* can be factored out of both numerator and denominator, it is irrelevant for the likelihood ratio, and thus is typically ignored in all linkage analyses. Thus, in our family, the likelihood ratio reduces to just

$$\Lambda = \frac{\frac{1}{2}\theta^4(1-\theta)+\frac{1}{2}\theta(1-\theta)^4}{\frac{1}{2}(0.5)^4(0.5)+\frac{1}{2}(0.5)(0.5)^4}.$$

The so-called lod score is then just the common logarithm of the likelihood ratio, and is equal to

$$Z(\theta)=\log_{10}(\Lambda)=\log_{10}[\tfrac{1}{2}(\theta^4(1-\theta))+\tfrac{1}{2}(\theta(1-\theta)^4)]-5\log_{10}(0.5).$$

The maximum of this lod score occurs at the same point as the maximum of the likelihood, so by maximizing the lod score over θ, we will find the maximum likelihood estimate of the recombination fraction θ. In this case, the maximum occurs at approximately θ = 0.21, with a corresponding lod score of

$$Z(0.21)=\log_{10}(\Lambda)=\log_{10}[\tfrac{1}{2}(0.21)^4(0.79)+\tfrac{1}{2}(0.21)(0.79)^4]-5\log_{10}(0.5),$$

which equals Z(θ = 0.21) = 0.124929.

3.2 MLINK

Now that we have analytically derived the correct answer, let us confirm our results by performing the same analysis with the LINKAGE programs. Let us first analyze our data using the MLINK program, which computes lod scores at a user-defined set of recombination fractions. In this example, let us compute the lod scores starting at θ = 0, 0.1, 0.2, 0.3, 0.4, and 0.5. In other words, we will start at θ = 0, and calculate lod scores in steps of 0.1 until we get to θ = 0.5. To do this we will go back into PREPLINK, and set up the parameter file to specify this analysis with the MLINK program. Note: The starting value may be 0.0001 instead of exactly 0, which will avoid a lod score of negative infinity.

First, enter *PREPLINK* on the command line, to activate the program. Next, choose option *(m) Read Datafile*, and specify that the name of the datafile to be read is *EX1.DAT* (the file we created in the last chapter). Now choose option *(i) Program Used*. This is where we can specify the analysis to be run by LINKAGE. You should see a menu of choices like the following:

```
***********************************
(a) MLINK : Y
(b) ILINK : N
(c) LINKMAP : N
(d) RETURN TO MAIN MENU
********************************
Use ILINK for CILINK or LODSCORE
Use LINKMAP for CMAP
enter letter to modify values
```

At this point, you should choose *(a)* to select the MLINK program, followed by *(d) RETURN TO MAIN MENU.*

The next thing we will need to adjust is the starting recombination value, by selecting choice *(j) Recombination values* from the main menu. You should then be prompted with

```
      ENTER 1 NEW THETAS
```

to which you should respond *0*, as we wish to start calculating lod scores from Θ = 0. After entering this, you will be automatically returned to the main menu of PREPLINK.

There is still one more thing we need to tell the program, which is that we wish to vary the recombination fraction in steps of 0.1 up to θ = 0.5. To do this select option *(l) See or modify recombination to vary*, from the main menu. You should see a screen like the following:

```
****************************************
(a) RECOMBINATION TO VARY : 1
(b) STARTING VALUE : 0.0000
(c) INCREMENT : 0.0100
```

```
(d) FINISHING VALUE : 0.5000
(E) RETURN TO MAIN MENU
****************************************
enter letter to modify values
```

We now must set all of these values to specify our desired analysis. The first line, "*(a) RECOMBINATION TO VARY*" may sound confusing. Here we are dealing with only a two-point analysis, so there is only one recombination fraction involved. However, if we were doing a multi-point analysis, there would be multiple inter-locus recombination fractions, and MLINK can only vary one of them, hence you need to specify which one here. For our purposes, leave it at *1*. Now, you must set the *(b) STARTING VALUE* to *0.0000*, the *(c) INCREMENT* to *0.1000*, and the *(d) FINISHING VALUE* to *0.5000*. This should be clear, since we want to increment θ by steps of 0.1, stopping at θ = 0.5. After making these adjustments, please enter *(e) RETURN TO MAIN MENU*, and *(n) Write Datafile*, calling it *EX1.DAT* (By the way, *YES*, you do want to overwrite the EX1.DAT file that exists, as you have just modified it, and no longer need the old one).

Now that we have fully prepared ourselves for the analysis, let us begin. The first thing you must do is to copy your pedigree file to a new name, PEDFILE.DAT, by typing

```
COPY EX1.PED PEDFILE.DAT
```

at the DOS prompt. Similarly, you will need to copy the parameter file to DATAFILE.DAT, by typing

```
COPY EX1.DAT DATAFILE.DAT
```

This is required when the LINKAGE programs are run directly, though we'll be learning how to avoid all of this tedium in the next chapter.

We are now finally ready to do the linkage analysis. First, we must run the UNKNOWN program. This program is very important in eliminating impossible genotypes from consideration by the analysis programs. If you have a pedigree with a large number of individuals with unknown marker or disease genotypes, this program will save massive amounts of time. Also, it checks for inconsistencies in your data. If you have entered the data in such a way that a non-Mendelian situation arises, the UNKNOWN program will inform you of this by saying "Incompatibility detected in this family at Locus 1", or something like that. In any event, the LINKAGE programs are set-up, on most computers, so that one MUST run the UNKNOWN program before the analysis programs, since it runs quickly, detects inconsistencies, and saves much time from the analysis programs in most situations. To run this program, type *UNKNOWN* at the DOS prompt. When the program has completed, type *DIR*, and you will see that it has produced 2 new files, SPEEDFIL.DAT, and IPEDFILE.DAT. In this case, since everybody is typed, and all genotypes are uniquely determined, this file should be empty (i.e. the size should be 0 bytes). On the other hand, for the same reason, the IPEDFILE.DAT should be essentially identical in substance to the PEDFILE.DAT, with different spacings, and without the comments at the end of each line.

Now, you are finally ready to perform your first linkage analysis with the MLINK program. To do this, simply enter MLINK at the DOS prompt, and the program will begin to calculate lod scores for you. When the program is finished, there should be new files produced called OUTFILE.DAT, and STREAM.DAT. For our purposes at this time, we will ignore the STREAM.DAT file, though you'll see in later chapters why it is important. Now, look at the OUTFILE.DAT file in your word processor. It should resemble the following:

```
LINKAGE (V5.1) WITH 2-POINT AUTOSOMAL DATA
 ORDER OF LOCI: 1 2
-----------------------------------
-----------------------------------
THETAS 0.500
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.830722 -8.612355
-----------------------------------
TOTALS -19.830722 -8.612355
```

```
-2 LN(LIKE) = 3.966144326368E+001 LOD SCORE = 0.000000
-----------------------------------
-----------------------------------
THETAS 0.000
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -100000002004087734272.000000 -43429358520716623872.000000
-----------------------------------
TOTALS -100000002004087734272.000000 -43429358520716623872.000000
-2 LN(LIKE) = 2.000000040082E+020 LOD SCORE = -43429358520716623872.000000
-----------------------------------
-----------------------------------
THETAS 0.100
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.780789 -8.590670
-----------------------------------
TOTALS -19.780789 -8.590670
-2 LN(LIKE) = 3.956157852618E+001 LOD SCORE = 0.021685
-----------------------------------
-----------------------------------
THETAS 0.200
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.544641 -8.488112
-----------------------------------
TOTALS -19.544641 -8.488112
-2 LN(LIKE) = 3.908928168151E+001 LOD SCORE = 0.124243
-----------------------------------
-----------------------------------
THETAS 0.300
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.613033 -8.517814
-----------------------------------
TOTALS -19.613033 -8.517814
-2 LN(LIKE) = 3.922606586242E+001 LOD SCORE = 0.094541
-----------------------------------
-----------------------------------
THETAS 0.400
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.758215 -8.580866
-----------------------------------
TOTALS -19.758215 -8.580866
-2 LN(LIKE) = 3.951642988211E+001 LOD SCORE = 0.031489
-----------------------------------
-----------------------------------
THETAS 0.500
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.830722 -8.612355
-----------------------------------
TOTALS -19.830722 -8.612355
-2 LN(LIKE) = 3.966144326368E+001 LOD SCORE = 0.000000
```

You should summarize these data, by extracting the important information, and writing it in a little table. The most important pieces of information are the log$_{10}$(Likelihood), and the Lod Score at each theta. From this output file, you can extract the information shown in table 3-1.

θ	Log_{10}(Likelihood)	Lod Score
0	-infinity	-infinity
0.1	-8.590670	0.021685
0.2	-8.488112	0.124243
0.3	-8.517814	0.094541
0.4	-8.580866	0.031489
0.5	-8.612355	0.000000

Table 3-1: Analysis results of EX1.PED; EX1.DAT

Essentially, the lod scores can be calculated as shown above, by subtracting the $\log_{10}$(Likelihood) at θ = 0.5 from each of the other likelihoods. You should try this by hand to verify that this is how the likelihoods are calculated. Further, you may wish to verify these lod scores by using the analytical formula we derived above, and substituting in the appropriate values of θ to compute lod scores. They should be identical. You should see that from MLINK we get a good idea, not only of where the maximum lod score occurs, but we get to see the whole lod score curve, which can provide information about how accurate your maximum likelihood estimate is. We'll get into this in subsequent chapters. Notice that at θ = 0.20, the lod score is 0.124243, very close to our theoretical maximum, for θ = 0.21. As a test to verify our theoretical calculation, let's use the MLINK program to calculate the lod score at θ = 0.21 to verify that we did it correctly by hand. Just read your EX1.DAT file back in to PREPLINK, and modify option *(l) See or modify recombination to vary*. This time set *(b) STARTING VALUE* to *0.21*, *(c) INCREMENT* to *0.1*, and *(d) FINISHING VALUE* to *0.22*. Then *(e) RETURN TO MAIN MENU*, write the new EX1.DAT file, and exit PREPLINK. Note that in the manner we set up our recombination to vary, it will start at θ = 0.21, and move in steps of 0.1 until θ > 0.22, the finishing value. So, in this case, it will only calculate the lod score at θ = 0.21, since 0.21 + 0.1 = 0.31 > 0.22. Now, copy the EX1.DAT file to DATAFILE.DAT, and we can begin the analysis (EX1.PED is unchanged, and still the same as our PEDFILE.DAT). Let's check it out by running the UNKNOWN program again, followed by the MLINK program, as outlined above. Your new OUTFILE.DAT file should look like this:

```
LINKAGE (V5.1) WITH 2-POINT AUTOSOMAL DATA
 ORDER OF LOCI: 1 2
-----------------------------------
-----------------------------------
THETAS 0.500
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.830722 -8.612355
-----------------------------------
TOTALS -19.830722 -8.612355
-2 LN(LIKE) = 3.966144326368E+001 LOD SCORE = 0.000000
-----------------------------------
THETAS 0.210
-----------------------------------
PEDIGREE | LN LIKE | LOG 10 LIKE
-----------------------------------
 1 -19.543061 -8.487426
-----------------------------------
TOTALS -19.543061 -8.487426
-2 LN(LIKE) = 3.908612143922E+001 LOD SCORE = 0.124929
```

Look at this file, and you can see that the lod score at θ = 0.21 is 0.124929, both according to the MLINK program, and our theoretical analysis. You might have noticed that in this OUTFILE.DAT, the likelihoods and lod scores are given not only for θ = 0.21, but also for θ = 0.5. The reason for this is that in order to compute a lod score, you need the likelihood at θ = 0.5 as the denominator of the likelihood ratio. For this reason, no matter what θ's you want lod scores computed for, the MLINK program will always compute the likelihood at θ = 0.5 first (Note that the lod score at this point is ALWAYS 0, since L(0.5)/L(0.5) = 1, and $\log_{10}(1) = 0$).

3.3 ILINK

We will also frequently use the ILINK program for two-point analyses. This program doesn't give you the lod scores at predefined points, but rather attempts to numerically maximize the likelihood, and only returns the likelihoods at the maximum likelihood estimate of the iterated parameter (in this case, the recombination fraction). So, let's try to use this program to find the maximum likelihood estimate of θ. We know from our theoretical evaluation that the maximum is at approximately θ = 0.21. Let's see if ILINK returns a similar result.

Go back to PREPLINK, and set up the parameter file to do an ILINK analysis. First call up PREPLINK, and read in the EX1.DAT file. Then go back to option *(i) Program Used*, and select the *ILINK* program. Return to the main menu, and choose option *(l) See or modify iterated parameters*. Note that this is different from what option *(l)* said when we had specified the MLINK program. Anyway, you should now see a menu like the following:

```
****************************************
(a) RECOMBINATION VALUES TO BE ITERATED (1) OR FIXED (0) :
 0
(b) LOCUS FOR WHICH VALUES MAY BE ITERATED : 1
(c) RETURN TO MAIN MENU
****************************************
NB : IF YOU WISH TO ITERATE OTHER PARAMETERS THE DATAFILE
 MUST BE MODIFIED AFTER EXITING FROM PREPLINK
enter letter to modify values
```

The first option *(a) RECOMBINATION VALUES TO BE ITERATED (1) OR FIXED (0)* is straightforward. In this case, there is only one recombination value, since we only have two loci. We merely need to tell the program whether it should try to maximize the likelihood with respect to this parameter, or whether it should remain fixed at the specified value. To start out, let's use ILINK to find the maximum likelihood estimate of θ, so we should select option *(a)*, and then enter a *1* to iterate the recombination value. Now, we are confronted by the option *(b) LOCUS FOR WHICH VALUES MAY BE ITERATED*. The ILINK program can do more than just maximize the likelihood over recombination fractions. It can also estimate gene frequencies, penetrances, disequilibrium, sex difference in recombination, etc. For now, though, we are only going to be using it to estimate recombination fractions (which are not locus-specific values). Still, the program allocates memory as if it were going to maximize the likelihood over all of these parameters. For this reason it is a good idea to have one of the marker loci be the locus for which values may be iterated, since there are fewer parameters at such a locus (no penetrances!). So, choose option *(b)*, and choose locus *2* as the new locus for iterated values. Now return to the main menu. There is one thing left that we MUST change, and that is *(j) recombination values*. Since we are maximizing the likelihood over the recombination fraction, the value we enter here should be thought of as just a starting value for the maximization procedure. The closer it is to the maximum, the quicker the program will converge to the estimate. Also, in some cases, different starting values may yield different estimates, as the algorithm could get trapped in a local maximum, or the program could stop for reasons other than that the maximum was found. Also, you CANNOT start the iteration at θ = 0, as this is a boundary value for θ, and the maximization algorithm used in ILINK will not work properly when started at a boundary. For this reason, it is a good idea to start at 0.1. In real life, you may want to try multiple starting values, just to be sure they all end up the same place roughly. Anyhow, once this is changed, you can save the new EX1.DAT file, and leave PREPLINK. Copy this file to DATAFILE.DAT, and run the UNKNOWN program again. Now, instead of typing MLINK at the DOS prompt, type *ILINK* to run the ILINK program.

The ILINK program produces three output files, OUTFILE.DAT, FINAL.DAT, and STREAM.DAT. Again, we will defer discussion of STREAM.DAT to a later chapter. The most important output file from ILINK is the FINAL.DAT file, which should resemble the following:

```
CHROMOSOME ORDER OF LOCI :
 1 2
****************** FINAL VALUES **********************
PROVIDED FOR LOCUS 2 (CHROMOSOME ORDER)
*******************************************************
GENE FREQUENCIES :
 0.500000 0.500000
```

```
******************************************************
THETAS:
 0.212
******************************************************
-2 LN(LIKE) = 3.908608568137E+001
LOD SCORE = 1.249370510945E-001
NUMBER OF ITERATIONS = 4
NUMBER OF FUNCTION EVALUATIONS = 12
PTG = -1.660797216309E-005
******************************************************
******************************************************
```

This gives us the final estimate of θ = 0.212 (pretty close to our approximate solution), with an associated lod score of 0.124937 (slightly larger than our value, as the program estimated θ to 3 decimal places, and we only did it to 2). The gene frequencies given are for locus 2, which we specified as the locus at which parameters were to be estimated. They were not estimated, but since the program allows for this estimation, it output their final values. The –2ln(like) is given here, since it has a nice statistical interpretation, being in units of chi-squared asymptotically (We'll be using this later). The other values are not so important, but just specify how long it took the program to converge, and information about the gradient of the likelihood surface. Let us now look at the other output file, OUTFILE.DAT (remember that this was the name of the important file produced by MLINK). This should look like the following:

```
DIFFER INTER = 3.452668897808E-004 TRUNC UPPER = 1.858135866348E-002

ITERATION 1 T = 0.100 NFE = 2 F = 3.956157852618E+001
X= 1.000000000000E-001
G= -1.119885140454E+001
P= 1.119885140454E+001
TBND = 8.036538458179E-002 RESET T = 4.015494478452E-002

ITERATION 2 T = 0.020 NFE = 5 F = 3.929637287819E+001
X= 3.248446298996E-001
G= 2.951007149097E+000
P= -4.689220798602E-002
FSMF = 2.652056479930E-001 PTG =-1.383792410037E-001 TMIN= 6.436095502122E-002
INITIAL T = 1.000000000000E+000
TBND = 6.927475669229E+000 RESET T = 1.000000000000E+000

ITERATION 3 T = 2.000 NFE = 8 F = 3.909481766725E+001
X= 2.310602139275E-001
G= 8.576558660115E-001
P= -3.842391630530E-002
FSMF = 2.015552109388E-001 PTG =-3.295449721438E-002 TMIN= 5.586901493504E-002
INITIAL T = 1.000000000000E+000
TBND = 6.013447772779E+000 RESET T = 1.000000000000E+000

ITERATION 4 T = 0.500 NFE = 11 F = 3.908608568137E+001
X= 2.118482557749E-001
G= 2.680003275200E-002
P= -6.196996965180E-004
FSMF = 8.731985879677E-003 PTG = -1.660797216309E-005 TMIN = 5.586901493504E-002
EXIT CONDITION 5
Specified tolerance on normalized gradient met
```

Most of the stuff in this file isn’t that important to the user, with the exception of the final line in the file, where it indicates the exit condition. In this case, it says "Specified tolerance on normalized gradient met". This means the program converged to the maximum, within a predefined tolerance level. Sometimes, however, the program exits for other reasons, like "Excessive cancellation in gradient", meaning that the final results you've obtained in your FINAL.DAT file are not really the maximum likelihood estimates, and perhaps you should restart your ILINK analysis using the end points from FINAL.DAT as starting values for a new ILINK analysis.

Let us confirm that we have achieved a maximum, by starting ILINK using θ = 0.213 as the starting value, and letting it go from there. To do this, read the EX1.DAT file back into PREPLINK, and change the recombination value to 0.213, save the file EX1.DAT, copy it to DATAFILE.DAT, and rerun UNKNOWN and ILINK. The FINAL.DAT file should indicate that the new estimate of θ is 0.211, with a lod score of 0.124938, which is roughly the same as before, and only better in the 6th decimal place. Since this kind of accuracy is unimportant, we can stop here, satisfied that the best estimate of θ is approximately 0.21.

Again, we can also use ILINK to evaluate the lod score at the specific point θ = 0.21, to verify that ILINK computes the same lod score as MLINK, and as we did when we calculated it by hand. To do this, read the EX1.DAT file into PREPLINK, and set the recombination fraction to 0.21. Then, go to *(l) See or modify iterated parameters*, and set *(a) RECOMBINATION VALUES TO BE ITERATED (1) OR FIXED (0)* to *0*, as we now want to fix the recombination fraction and just calculate the lod score. Then write the new EX1.DAT file, and exit PREPLINK. Now copy the EX1.DAT to DATAFILE.DAT as before, and rerun UNKNOWN and ILINK, and this time, the FINAL.DAT file should indicate the lod score as 0.124929, to 6 decimal places.

In this chapter, we introduced the principles of basic 2-point linkage analysis. We further calculated, theoretically, 2-point lod scores for our sample pedigree, and then introduced the UNKNOWN, MLINK, and ILINK programs of the LINKAGE package, and we used them to perform 2-point analyses with our example pedigree. In the next chapter, we will see that there is an easier way...

EXERCISE 3

Analyze the example pedigree from exercise 2. Try to analyze this pedigree analytically. When you see how amazingly complicated and time-consuming it can be, give up, and use the MLINK and ILINK programs to compute 2-point lod scores as described in this chapter. Please compute 2-point lod scores for θ = 0, 0.1, 0.2, 0.3, and 0.4 with MLINK (If you are running into problems, consult the answers given in chapter 12 at the end of Part I). Then, do the same with ILINK (can you see how to do this? Hint: It takes 5 separate runs of ILINK...). Finally, find the best estimate of θ with ILINK, starting with a value of θ = 0.1, and then refine the estimate, by starting ILINK again from the value of θ given in the first FINAL.DAT file. Are you satisfied with this precision?

4 Setting up a Linkage Analysis Using LCP

In the last chapter, we learned how to use the LINKAGE programs to do a linkage analysis. Specifically, we learned how to manipulate the datafile for each different analysis we wanted to perform. There is a program, the Linkage Control Program (LCP) that was written to make the whole process a lot simpler. It allows the user to specify different analyses without modifying the parameter file each time by hand. This program allows the user to specify any number of different analyses, and then let the computer do the analyses non-interactively, freeing up the researcher to do other things. In most situations, one will use the LCP program whenever performing a linkage analysis, and we will be using it from this chapter on whenever we wish to use the LINKAGE programs (unless specifically specified). The main feature of LCP is that it allows the user to set up datafile and pedfile with many loci and then specify subsets of loci for analysis.

4.1 LCP

We will now learn how to use LCP to set up an analysis of the data in EX1.PED and EX1.DAT. The LCP program is used to write a so-called *batch file* which contains a series of commands that will define all of the steps involved in setting up and performing any set of linkage analyses a person might desire.

First, start by typing *LCP* at the DOS prompt to activate the program. You will be presented with a screen like the following:

```
COMMAND file name [PEDIN.BAT] : PEDIN.BAT
LOG file name [FINAL.OUT] : FINAL.OUT
STREAM file name [STREAM.OUT] : STREAM.OUT
PEDIGREE file name [PEDIN.DAT] : PEDIN.DAT
PARAMETER file name [DATAIN.DAT] : DATAIN.DAT
Secondary PEDIGREE file name [] :
Secondary PARAMETER file name [] :
```

The so-called COMMAND file name is the name of the batch file which LCP will produce. The default name is PEDIN.BAT, and there is usually no reason to change it, unless you will want to repeat the same analysis later for some reason. The LOG file is the file, typically called FINAL.OUT, containing all of the results of all the analyses performed. Basically, this consists of a collection of the OUTFILE.DAT and FINAL.DAT files you were introduced to in the last chapter. The STREAM file, similarly, is a collection of STREAM.DAT files, which we will defer discussion of until later on. Finally, we have to indicate the PEDIGREE and PARAMETER files in which the data to be analyzed is stored. There is a very useful help screen available in LCP, which can be accessed at any time by hitting <Ctrl-H>. You may wish to look at this now, to see a summary of the "control characters", which may be useful to you later.

At this point, we should adjust this first screen to correspond to our desired analysis. The only names we will need to modify at this time are the PEDIGREE and PARAMETER files. Please go to these lines, using the cursor keys, and delete the file names currently shown (Use <CTRL-U> to delete any entire line in LCP). Now, replace these names with your file names, *EX1.PED* for the PEDIGREE file, and *EX1.DAT* for the PARAMETER file. Everything else on this screen is set up correctly, so you should now advance to the next screen by hitting the <Page Down> key (Linux: ctrl-N advances to the next screen). The next screen should look like this:

```
General pedigrees : <-
Three-generation pedigrees :
Experimental cross pedigrees :
```

The general version of the LINKAGE programs can be accessed through the *General pedigrees* option on this page. The *Three-generation pedigrees* option allows the user to choose one of the specialized versions of the LINKAGE programs designed to analyze codominant markers only in CEPH-type families, which are very specific types of three-generation pedigrees (discussed in detail in Part II). The *Experimental cross pedigrees* option allows the user to use specific versions of the LINKAGE programs designed to analyze animal crosses, and will not be used in this book, since they are not designed for human data. For our purposes, we will now choose the *General pedigrees* option from this menu, and then hit the <Page Down> key again to move to the next page. Now, you will have a list of the programs to choose from. We want to do the same analysis as in the last chapter, so first choose the *MLINK* program from this menu, and hit <Page Down>. You should now see the following menu of choices:

```
Specific evaluation : <-
Lod score table :
Multiple pairwise lod table :
```

The *Specific evaluation* option is the one we used in the last chapter, and the one we will use at this time. However, in many situations, the *Lod score table* option may be more useful, as it allows you to pick a set of recombination fractions at which to perform the lod score calculations, while under the *Specific evaluation* option, you must increment the steps by a constant factor, as we saw in the previous chapter. For now, choose the *Specific evaluation* option, and hit <Page Down>. The next menu has only one choice (a limitation of LCP, not the analysis programs), for *No sex difference*, so just hit <Page Down> again, and finally, you will see a menu of choices that should remind you of what we changed in PREPLINK in the last chapter:

```
                 Locus Order:
       Recombination Fractions: .1
       Recombination varied   : 1
            Increment Value   : .1
            Stop Value        : .5
```

Now, you should modify the *Locus order* to be *1 2* (or *2 1*, it makes no difference), with starting *Recombination fraction* of *0*. *Recombination varied*, as discussed in the previous chapter, remains at *1*. *Increment value* of *0.1*, and *Stop value* of *0.5* should be set. This, as you remember will calculate the lod scores at θ = 0, 0.1, 0.2, 0.3, 0.4, and 0.5. Now **BE SURE TO HIT <PAGE DOWN>** in order to write this analysis to the PEDIN.BAT file! However, there is no need to exit from the program at this point. In chapter 3, we also analyzed the pedigree in question with the ILINK program. So, let us hit <Page Up> (Linux: ctrl-P) three times, to bring us back to the menu:

```
LODSCORE :
ILINK :
LINKMAP :
MLINK : ←
```

Now, just move the arrow up to the *ILINK* program, and hit <Page Down> to select it. Now, you will see a screen that is basically meaningless in the context of 2-point analysis:

```
Specific order : ←
All orders :
Inversions of adjacent loci :
```

These options are very important when you are trying to do a multipoint analysis, so we will defer discussion of these until Part II. For now, just select specific order, and hit <Page Down>. You should now see a selection of options regarding sex difference in recombination fractions. In this analysis, we assumed there was no sex-difference, so just go to the *No sex difference* line, and hit <Page Down> again. Finally, you will be presented with the screen on which you will select the analysis parameters:

```
Locus order [] :
Recombination fractions :
```

Here, you should enter the *Locus order* of *1 2* (or equivalently, *2 1*, as above), and the starting value for the *Recombination fraction* should be set to *0.1*, as in the last chapter. Now hit <Page Down> to enter this analysis, and we will then be ready to exit the program, and LCP will write the PEDIN.BAT file. Do this by entering <Ctrl-Z>.

If you look at the file directory (by typing *DIR* at the DOS prompt), you should see a new file, PEDIN.BAT, which you just created with the LCP program. Feel free to look at it in your word processor. It is quite long, and can be confusing to interpret, but basically, it is instructing the computer on how to do all of the cumbersome manipulations you had to do by hand in the previous chapter. Anyway, return to DOS, and type *PEDIN* to call this file. It will then perform the analyses you have requested. When it is finished, look at the FINAL.OUT file in your word processor. It should contain a brief summary of each analysis you requested, followed by the OUTFILE.DAT file (for the MLINK run), and then the FINAL.DAT file (for the ILINK run). These should contain the same lod scores as the analysis in the last chapter. See how easy it is when you have LCP to do all of the hard work for you.

We would like now to briefly explain how this batch file works. Basically, the batch file uses another program, LSP (you'll see this again later...), to modify the parameter files, replacing what you did with PREPLINK before, producing PEDFILE.DAT and DATAFILE.DAT files. LSP also writes a summary of the analysis which is added to the FINAL.OUT file. Next, it calls the UNKNOWN program and the desired analysis program (MLINK or ILINK). It then takes the appropriate output file, and appends it to the FINAL.OUT file, and starts the process over for the next analysis, deleting all intermediate files in the process. It is really quite handy, and efficient, and makes your life a lot easier.

In this chapter, you learned how to use the linkage control program, LCP, to handle most of the drudgery of the linkage analysis process. You used this program to efficiently direct the computer through all of the required steps involved in using the LINKAGE programs in an automated manner.

EXERCISE 4

Please use LCP to perform the analyses you did in exercise 3, and check that the results are, in fact, identical.

5 Elementary usage of the Affection status locus type

In this chapter, you will be learning how to manipulate the affection status locus type to allow for autosomal dominant and recessive diseases. Further, the concept of reduced penetrance will be introduced on an elementary level. You will also have some more example pedigrees to enter, giving you a chance to practice all of the skills you've learned thus far.

5.1 AUTOSOMAL DOMINANT DISEASE

If you remember, the example pedigree we've been working on so far has been a phase unknown pedigree with a dominant disease segregating. Now, let us suppose that we have collected data on the parents of *father*, giving us the pedigree shown in Figure 5-1, with *fgrandpa* being unaffected, and *2/2* at the marker locus, while *fgrandma* is affected, and *1/1* at the marker locus.

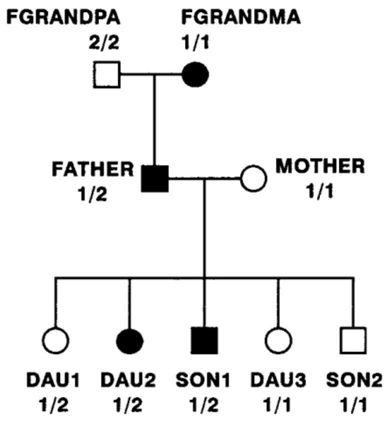


Figure 5–1. Phase-known Pedigree EX2.*

This information tells us that *father* has to have received the disease and the *1* allele together from *fgrandma*, and the normal allele together with the *2* allele from *fgrandpa*. Thus, we know *father*'s phase with certainty to be + 2/D 1. In this case, we can easily determine which children are recombinant and which are non-recombinant. In this family, *dau1* is non-recombinant, having received the + 2 haplotype from *father*, while the others are all recombinant, having received the non-parental haplotypes D 2 (*dau2* and *son1*) or + 1 (*dau3* and *son2*). The associated likelihood is then just $K\theta^4(1 - \theta)$, K being a constant, giving us a lod score of $\log_{10}[\theta^4(1 - \theta)/(0.5)^5]$, which equals $5\log_{10}2 + 4\log_{10}\theta + \log_{10}(1 - \theta)$. This lod score function can be easily maximized by just taking the first derivative with respect to θ as $dZ(\theta)/d\theta = \log(e)\,[4/\theta - 1/(1 - \theta)]$. Setting this equal to zero, and solving for θ yields a maximum likelihood estimate of $\hat{\theta} = 0.8$. Calculating the lod score at that point yields $Z(\theta = 0.8) = \log_{10}([0.8]^4[0.2]/[0.5]^5) = 0.4185$. However, in human genetics, recombination fractions larger than 50% are meaningless, since when two genes are completely unlinked, the maximum possible recombination fraction is only 0.5. For this reason, the estimate is typically truncated to 0.5, given that the larger θ is meaningless. (An estimate of > 0.5 may also indicate that there is some data error, if the corresponding lod score is large.) In this case, therefore, our truncated MLE (maximum likelihood estimate) of θ would be 0.5, with associated lod score of 0.

Please make pedigree (EX2.PED) and parameter (EX2.DAT) files for this new pedigree. Make the disease autosomal dominant, as you did for the original pedigree (EX1.*). If you are unclear as to how to do that, refer back to chapter 2. Give the disease locus gene frequencies of 0.99999 for the normal (wild type) allele, and 0.00001 for the disease allele. Then define an allele numbers locus with 2 equally frequent alleles. Now, use LCP to set up an analysis of this pedigree with the MLINK program, starting from θ = 0, in steps of 0.1, up to θ = 0.5, and with the ILINK program, using a starting value of θ = 0.1. After getting the first estimate of θ, restart the ILINK program to refine the estimate, by setting up a further analysis with LCP. When all is said and done, you should interpret the output from FINAL.OUT, and come up with the results given in table 5-1.

θ	$\log_{10}$(Likelihood)	Lod Score
0.0	-infinity	-infinity
0.1	-12.357079	-2.540602
0.2	-11.204114	-1.387637
0.3	-10.557742	-0.741265
0.4	-10.124935	-0.308458
0.5	-9.816477	0.000000

ILINK: $\hat{\theta}$ = 0.798,   Z($\hat{\theta}$) = 0.4185

Table 5-1: Analysis results of EX2.PED; EX2.DAT

5.2 RECESSIVE DISEASE

Let us continue now, by looking at an example of a *recessive* disease. Consider the family in Figure 5-2, with grandparents, parents, and offspring.

Please make a pedigree file corresponding to this pedigree, and name it EX3A.PRE. Process this file with MAKEPED as above (to make EX3A.PED), and create an appropriate parameter file (EX3.DAT) in PREPLINK as before, only this time specify the *PENETRANCES* for a recessive disease, which are as follows:

```
GENOTYPE 1 1 OLD PEN 0.000000
?
0
GENOTYPE 1 2 OLD PEN 0.000000
?
0
GENOTYPE 2 2 OLD PEN 1.000000
?
1
```

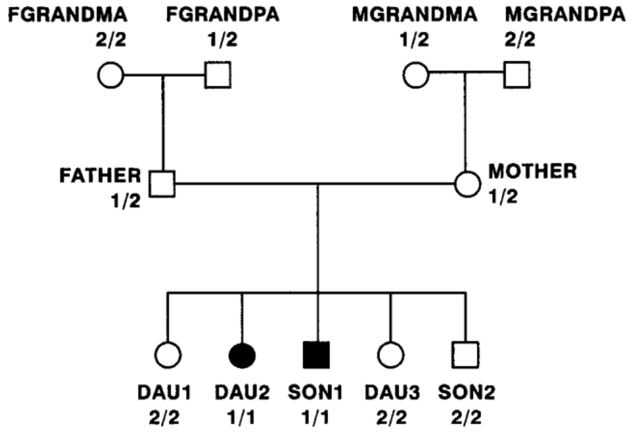


Figure 5–2. Recessive Pedigree EX3A.*

Again modify the gene frequencies to be 0.99999 and 0.00001, as in the previous dominant disease example. The marker should be a two allele *Allele Numbers* type locus with equal gene frequency for the two alleles. Save this file as *EX3.DAT*.

Now analyze this family with MLINK and ILINK as you did above, and examine your output. It should correspond to the results given in Table 5-2.

θ	Log_{10}(Likelihood)	Lod Score
0	-15.418525	0.976874
0.1	-15.606399	0.789000
0.2	-15.824343	0.571056
0.3	-16.065391	0.330008
0.4	-16.291110	0.104290
0.5	-16.395399	0.000000

ILINK: $\hat{\theta}$ = 0.000, Z($\hat{\theta}$) = 0.976874.

Table 5-2: Analysis results of EX3A.PED; EX3.DAT

Looking at this family, we can see that no information about linkage is obtained from the grandparents, since we have no way of knowing which grandparent was carrying the disease allele in each case. If you do not think this is correct, reanalyze this family, omitting *fgrandma*, *fgrandpa*, *mgrandpa*, and *mgrandma*. When you do this, the $\log_{10}$(Likelihood)'s will change, since there are fewer people in the family, but the lod scores should remain identical. Make this new pedigree file, *EX3B.PED* as above, and then use LCP to set up an analysis of this new family, using the same parameter file, *EX3.DAT*. The resulting output should be as shown in Table 5-3.

θ	Log_{10}(Likelihood)	Lod Score
0	-13.612340	0.976874
0.1	-13.800214	0.789000
0.2	-14.018158	0.571056
0.3	-14.259206	0.330008
0.4	-14.484925	0.104290
0.5	-14.589214	0.000000

ILINK: $\hat{\theta}$ = 0.000, Z($\hat{\theta}$) = 0.976874.

Table 5-3: Analysis results of EX3B.PED; EX3.DAT

Herein, you can see one fundamental difference between recessive and dominant disease analysis. In the dominant case we saw above, adding the grandparents helped to establish the phase in the parents, adding much additional information about linkage. In this recessive case, we added the grandparents, yet got absolutely no change in the lod scores. The grandparents contribute no phase information whatsoever, since we have no way of telling which unaffected grandparent was carrying the disease. In general, you can see that typing grandparents can be very useful in dominant diseases, while in recessive diseases it may be a waste of effort. Can you see how this might affect the experimental approach to mapping a disease, based on its mode of transmission?

This pedigree is more difficult to analyze by hand, as there are four possible phase combinations for the parents. We know that each parent is doubly heterozygous, giving two phase probabilities for each parent. Since the parents are independent, we have four combinations of phase at the two parents jointly. Fortunately, all of the children are homozygous, so we can tell with certainty what each parent transmitted to each child. The formal equations involved are therefore quite complex, so we will not provide them here, as they are not very illustrative.

Throughout this book we assume a single mutant site (base-pair) to be responsible for dominant or recessive disease. Exome and whole-genome sequencing analysis have shown, however, that more than one nucleotide in a gene may be mutated and lead to disease. An important situation is the occurrence of compound heterozygous sequence variants, for which two parents are each heterozygous but for variants at different DNA sites in the same gene. A child may inherit a mutant allele from each parent and is then affected in a recessive manner. But at each site, considered by itself, the mutation appears to be dominant. If parents are non-consanguineous, the most likely explanation for a recessive disease is compound heterozygosity for two different pathogenic mutations, and specific filtering rules in genome sequencing have been derived for finding such cases [7]. On the other hand, when two related (unaffected) parent have an affected offspring, the child is likely a classical recessive. Among many examples, we quote one with Charcot-Marie-Tooth disease [8] and another with compound heterozygous mutations in the GARS gene leading to mitochondrial disease [9].

5.3 EQUIVALENT NUMBERS OF RECOMBINANTS AND NONRECOMBINANTS

For most family data, the likelihood is a complicated function of θ and other parameters such as allele frequencies. Sometimes it is useful to approximate this function by the LOD score resulting from known numbers of recombinants and nonreombinants [1, 10]. This will allow, for example, to gauge the numbers of recombination events in a dataset or, if this approximation is satisfactory, to interpolate LOD scores at θ values not present in published data.

Based on the LOD score, $Z(\theta) = n\log_{10}[2(1 - \theta)] + k\log_{10}[\theta/(1 - \theta)]$, with k recombinants in n meioses, we may obtain "estimates" of k and n in one of two ways, either (1) working with the maximum LOD score and the θ value at which it occurs, or (2) working with two LOD score values and the corresponding two θ values at which the two LOD scores were obtained. Derivation of the formulas required for obtaining k and n is straightforward although a little tedious and is not given here. A spreadsheet implementing these formulas is available on my webpage as http://www.jurgott.org/linkage/EquivalentLods.xlsx.

5.4 INCOMPLETE PENETRANCE

We have spent most of this chapter learning how to model different modes of transmission of a disease using the affection status locus type. One other common complication in such modelling is that there is not always complete penetrance. For many diseases, even if one has the disease predisposing genotype, the individual will not necessarily become affected. In fact often, there are very complicated probability models for this phenomenon, as you will see in later chapters. For now, let us just consider the simplest case, in which a random individual with the disease predisposing genotype has only a 50% chance of becoming affected. Let us go back to the phase unknown example from files EX1.PED and EX1.DAT, and modify EX1.DAT to allow for incomplete penetrance of this dominant disease. Just read the file EX1.DAT back into PREPLINK, then choose *(k) See or modify locus parameters*, and *(a) SEE OR MODIFY A LOCUS*, specifying locus *1*, the disease. Now, modify the penetrances as follows:

```
GENOTYPE 1 1 OLD PEN 0.000000E+00
?
0
GENOTYPE 1 2 OLD PEN 1.000000E+00
?
0.5
GENOTYPE 2 2 OLD PEN 1.000000E+00
?
0.5
```

Then save the new parameter file as *INC.DAT*, and reanalyze this pedigree with MLINK and ILINK, in the same way as before, and examine the new lod scores. The results are given in Table 5-4.

θ	Log_{10}(Likelihood)	Lod Score	Z(θ)from EX1
0.0	-8.612351	0.374811	-infinity
0.1	-8.703932	0.283230	0.021685
0.2	-8.800774	0.186388	0.124243
0.3	-8.892292	0.094870	0.094541
0.4	-8.961068	0.026094	0.031489
0.5	-8.987162	0.000000	0.000000

ILINK : $\hat{\theta} = 0$ $Z(\hat{\theta}) = 0.374811$

Table 5-4: Analysis results of EX1.PED; INC.DAT

The reason the estimate of θ is now 0 is because the one likely recombinant is most likely assumed to be a case of non-penetrance (i.e. the individual has the disease predisposing genotype, but did not express the disease for some reason). Let us take a look at this situation in a theoretical manner.

As we know from looking at this family in chapter 3, there are two equally likely (a priori) phases for *father*. He can be either (1) + 1/D 2 or he can be (2) D 1/+ 2. Let us see what effect this reduced penetrance has on our analysis, shall we. First of all, let us assume, since the disease gene frequency is so small, that *mother* is +/+ at the disease locus. Then, we know with certainty that *dau2* and *son1* have disease locus genotype D/+. However, *dau1*, *dau3*, and *son2* could each be either +/+ or D/+. Let us examine what happens under phase (1). In this case, let us first consider that *dau1* could have received + 2 from *father*. In this case, the probability of her being unaffected would be 1, since her disease locus genotype is +/+. The probability of receiving the + 2 haplotype would be (1 – θ). If she received D 2 from *father*, then she would be unaffected with probability 0.5, since her disease locus genotype is D/+. Also, the probability of receiving this haplotype is θ, since it is recombinant. Thus, the overall probability of observing *dau1*, and unaffected offspring with marker genotype 1/2 would be (1 – θ) + 0.5(θ) = 0.5(2 – θ) . Similarly, *dau3* and *son2*, who are identical phenotypically, can be shown to have probability of θ + 0.5(1 – θ) = 0.5(1 + θ). Thus, the total probability of this family, assuming Phase I would be $[0.5(2 – \theta)][0.5(1 + \theta)]^2\theta^2$. Under Phase II, it can be shown to be $[0.5(2 – \theta)]^2[0.5(1 + \theta)][1 – \theta]^2$. Thus, the total likelihood of this family would be

$$0.5([0.5(2 – \theta)][0.5(1 + \theta)]^2\theta^2 + [0.5(2 – \theta)]^2[0.5(1 + \theta)][1 – \theta]^2)$$

which can be reduced to the following:

$$K((2 – \theta)(1 + \theta)[(1 + \theta) \theta^2 + (2 – \theta)(1 – \theta)^2]).$$

The lod score can then be represented as

$$Z(\theta) = \log_{10}\left\{\frac{(2-\theta)(1+\theta)[(1+\theta)\theta^2 + (2-\theta)(1-\theta)^2]}{(\tfrac{3}{2})(\tfrac{3}{2})[(\tfrac{3}{2})(\tfrac{1}{2})^2 + (\tfrac{3}{2})(\tfrac{1}{2})^2]}\right\}.$$

One can then compute lod scores as above. For example, if you plug in θ = 0, you will get Z(θ = 0) = $\log_{10}([(2)(1)[0 + 2]/[(3/2)^3(1/2)]) = \log_{10}(4/(27/16)) = \log_{10}(64/27) = 0.37481$. As you can see, the addition of incomplete penetrance makes it extremely difficult to compute lod scores analytically, making the computer programs much more important and useful. When you get into complicated pedigrees, with complicated disease models, it becomes even more untractable. Still, you see that we can verify the results obtained with the LINKAGE programs by hand when we question the results.

In this chapter, we learned how to use the affection status locus type to specify various different disease models, including simple applications of reduced penetrance.

EXERCISE 5

Reconsider the pedigree from exercise 2, and analyze the pedigree assuming a penetrance of 75% for the dominant disease (in file USEREX5.DAT). How does this affect your results? Does it make sense to you? Try parametrizing the disease as an autosomal recessive disease with 70% penetrance. Does this result seem to make sense? What happens here, if you reduce the penetrance to 30% in the dominant and recessive cases separately? Given that this disease is really autosomal dominant with full penetrance, do the results of this analysis make sense? What would you expect if the disease were really recessive, and was analyzed as if it were dominant? *Hint*: Use LCP!

6 Sex-Linked Recessive diseases

In this chapter, we will introduce the concept of linkage analysis with sex-linked recessive diseases, and how to analyze them in the LINKAGE programs. You will learn how to modify your parameter and pedigree files to enter such data, and how to use the LINKAGE programs to analyze it. Since practically all sex-linked diseases are recessive, one usually refers to them simply as sex-linked.

6.1 SEX-LINKED DISEASES

In humans, there are twenty-two pairs of homologous autosomal chromosomes. Each person receives one copy of each chromosome from their mother, and one from their father, thus explaining most observed patterns of inheritance. Up to now, we have only considered such autosomal loci. However, humans also have so-called *sex chromosomes*, which are responsible, in part, for determining the sex of an individual. In humans, these chromosomes are designated *X* and *Y*, with *XX* individuals being female, and *XY* individuals being male. Clearly, every person must receive one X chromosome from their mother, while fathers give their X chromosome to their daughters, and their Y chromosome to their sons. On the sex chromosomes, recombination can only occur in females, since they have two homologous X-chromosomes, while men have the non-homologous X-Y pair. Actually, there is a small region of homology with recombination occurring between X and Y, known as the *pseudoautosomal region*, in which loci behave as if they were autosomal [11]. In general, though, it can be assumed that no recombination occurs on the X chromosome in males. Also, every male has only one allele at each X-chromosomal locus, in the so-called *hemizygous* state. This property makes it such that if everyone in a pedigree is typed, we will always know with certainty the phase in each individual, with the exception of female founders. That is to say, we will know in each non-founder female, which allele came from which parent (Males of course only receive X-chromosomal alleles from their mothers). This makes sex-linkage very easy to analyze.

A large number of diseases in humans, like hemophilia, and some forms of retinitis pigmentosa, are known to be inherited as X-linked recessive traits (McKusick, 1990). In fact, most X-chromosomal diseases that are known are fully penetrant recessives (sometimes with delayed age of onset), and are quite often lethal, like X-linked Agammaglobulinemia (Kwan et al, 1990). These diseases will be discussed again in part III. In recessive X-linked diseases, the only affected people are females who are homozygous for the disease allele, and males who are hemizygous with the disease allele. The only way a female could be homozygous, however, is if she received a disease allele from her affected father. Since many of these diseases are either lethal, or cause an affected man to be unlikely to sire children, usually only males are affected. If the disease were not lethal, and the population gene frequency of the disease allele were, say, 0.01, the probability of a random male in the population to be affected would be 0.01, while the probability of a random female being affected would be $(0.01)^2 = 0.0001$. In any case, the majority of affecteds will be male (In this example, they are 100 times more likely to be affected. In part III, you will see that for lethals, it is virtually impossible for women to become affected). Similarly, if there were a disease gene on the Y chromosome, *ONLY* males could be affected. In general, sex-linked diseases are characterized by this preponderance of male affecteds, and by the absence of any male-to-male transmission of the disease, since males can only get X chromosomes from their mothers! Let us now see how to analyze a sample pedigree segregating an X-linked recessive disease.

6.2 PREPARATION OF PEDIGREE AND PARAMETER FILES

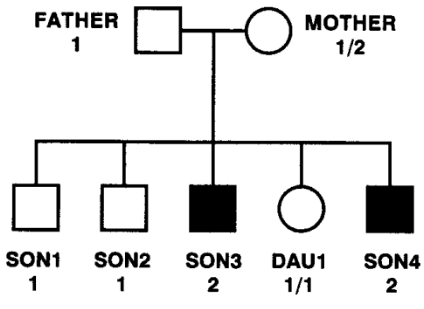


Figure 6–1. Sex-linked recessive pedigree EX4.*

Let us consider the pedigree shown in Figure 6-1. Enter this pedigree (EX4.PRE) in the standard LINKAGE format, as you have done in the previous autosomal examples. The only difference is that now, males will have only one allele at each locus while females will have two. The way we are required to code this in allele numbers format is to enter the allele number twice in males, as if they were homozygous. That is to say, if a male has allele 2 at a locus, you would enter his marker phenotype as *2 2*. This makes sure there are the same number of columns for males and females in the pedigree file at each locus. If you mistakenly enter a heterozygous genotype for any male in the pedigree file, the UNKNOWN program will detect it as an inconsistency, since males can only have one allele per X-linked locus. The EX4.PRE file should be as follows:

```
ex4   father      0           0      1    1     1 1
ex4   mother      0           0      2    1     1 2
ex4   son1        father      mother 1    1     1 1
ex4   son2        father      mother 1    1     1 1
ex4   son3        father      mother 1    2     2 2
ex4   dau1        father      mother 2    1     1 1
ex4   son4        father      mother 1    2     2 2
```

Process this file with MAKEPED in exactly the same way as for autosomal traits, and save it as EX4.PED.

Now, you will need to create a parameter file with the PREPLINK program. In this case, we will have a fully penetrant sexlinked recessive disease with gene frequency of 0.01 for the disease allele, and an allele numbers type of marker locus with 3 equally frequent alleles. To make your EX4.DAT parameter file, call up the PREPLINK program. From the main menu, you must first select the *(b) Sexlinked* option, to tell the program you will be looking at a sexlinked disease and markers. This will toggle the *N* in the row following *(b) Sexlinked* to a *Y*, meaning that from now on the program will operate under the assumption that all loci are sex-linked. Next, you must choose option *(k) See or modify locus description*. Change the first locus to affection status, as you have done previously, and change the second locus to allele numbers. Then choose option *(a) SEE OR MODIFY A LOCUS*, specifying locus *1*. You should see a screen like the following:

```
*************************************
LOCUS NUMBER : 1
*************************************
(a) NUMBER OF ALLELES : 2
(b) NUMBER OF LIABILITY CLASSES : 1
(c) PENETRANCES :
 MALES:
ALLELE 1 0.00000E+00
ALLELE 2 1.00000E+00
 FEMALES:
GENOTYPE 1 1 0.00000E+00
GENOTYPE 1 2 0.00000E+00
GENOTYPE 2 2 1.00000E+00
(d) GENE FREQUENCIES :
 0.500000 0.500000
(e) EXIT
*************************************
enter letter to modify values
```

This is slightly different from what we saw in the autosomal case, as we now have to specify separate penetrances for males and females. Notice that for males, there are two penetrances, corresponding to the two possible hemizygous genotypes (alleles), while females have three, corresponding to the three possible 2-allele genotypes. Looking at the indicated penetrance values closely, we can see that the default is for an X-linked recessive disease, with the disease causing allele being allele 2. As an exercise in entering the penetrances for a sex-linked disease, let us redefine the penetrances so that we have a fully penetrant X-linked recessive disease, with the disease causing allele being allele 1. The results will be identical! (If you don't believe this, try it both ways as an exercise). Let us now enter the appropriate penetrances for our fully penetrant X-linked recessive disease, with option *(c) PENETRANCES*. You should adjust the values as follows:

```
ENTER NEW PENETRANCES
 MALES:
ALLELE 1 OLD PEN 0.00000E+00
?
1
ALLELE 2 OLD PEN 1.00000E+00
?
0
 FEMALES:
```

```
GENOTYPE 1 1 OLD PEN 0.00000E+00
?
1
GENOTYPE 1 2 OLD PEN 0.00000E+00
?
0
GENOTYPE 2 2 OLD PEN 1.00000E+00
?
0
```

In this case, we specified that all males hemizygous for allele 1, and all females homozygous for allele 1 are affected with probability 1, while all other individuals cannot be affected with the disease. In this example, we have defined the penetrances such that allele 1 is the disease allele. To make this specification complete, you must modify the *(d) GENE FREQUENCIES* accordingly. Please enter new gene frequencies as follows:

```
ENTER 2 NEW GENE FREQUENCIES
0.01 0.99
```

Then, *(e) EXIT*, and *(a) SEE OR MODIFY A LOCUS*, specifying locus *2*. At this allele numbers locus, you can adjust the values just as you would for an autosomal locus, by setting the number of alleles to *3*, and the gene frequencies to *0.33333*, *0.33333*, and *0.33334* (so they sum to 1 !). You have already indicated that the analysis is to be done on sexlinked data, so the programs will know that males must be hemizygous. Now, go back to the main menu, and save the file as *EX4.DAT*.

6.3 PERFORMING THE LINKAGE ANALYSIS

You have already specified in the parameter file, EX4.DAT, that you will be analyzing sex-linked data, so you needn't do anything different in LCP. Just use LCP the same way as you have been throughout the course of this book, and specify that the program should do the analysis with MLINK, starting at θ = 0, in steps of 0.05 this time, and stopping at θ = 0.45 (We already know the lod score at θ = 0.5 MUST be 0 by definition, and the likelihood will be calculated as the first iteration in any case). Also specify an analysis with the ILINK program, with starting value of θ = *0.2* (for variety's sake). Now, exit LCP, and do the linkage analysis. The results from the FINAL.OUT file are shown in table 6-1.

θ	Log_{10}(Likelihood)	Lod Score
0.0	-4.644217	0.903088
0.05	-4.733319	0.813986
0.1	-4.827180	0.720125
0.15	-4.926119	0.621185
0.2	-5.030163	0.517142
0.25	-5.138642	0.408663
0.3	-5.249414	0.297891
0.35	-5.357506	0.189799
0.4	-5.453323	0.093982
0.45	-5.521958	0.025347
0.5	-5.547305	0.000000

ILINK : $\hat{\theta}$= 0, Z($\hat{\theta}$) = 0.903088

Table 6-1: Analysis results of EX4.PED; EX4.DAT

Let us take a look at this example analytically, to try and understand where this lod score came from. In this case, we know for a fact that the disease allele had to come from *mother*, since *father* was unaffected, making him hemizygous normal. Further, *mother* has to be heterozygous, since she is unaffected herself, but has affected sons. We then know that *son1* and *son2* got the normal allele from *mother* together with the 1 allele at the marker locus. Similarly, we know that *son3* and *son4* received the disease allele together with the 2 allele from *mother*. However, *dau1* could have received either the normal or disease allele from *mother*, since they are phenotypically indistinguishable. Hence, even though we know she received the 1 allele at the marker locus, we do not have any information about her disease locus genotype, so she is uninformative for linkage, and contributes nothing to the linkage analysis. To verify this, you should try analyzing the pedigree, leaving her out. The lod scores (but not the likelihoods) at every point will be identical.

What we are left with is essentially a phase unknown situation, in which all the information comes from transmission from *mother* to *son*s. The analysis is then straightforward, as in the first example in the book. *Mother* is equally likely to have phase (1) d 1/+ 2 or phase (2) + 1/d 2. If she has phase (1), then all four offspring are obligate recombinants, with probability of θ^4. If she has phase (2), they are all non-recombinants, with probability of $(1 - \theta)^4$. Hence the overall likelihood of this pedigree as a function of θ is just $K[0.5\theta^4 + 0.5(1 - \theta)^4] = K[\theta^4 + (1 - \theta)^4]$. The lod score is then just $\log_{10}([\theta^4 + (1 - \theta)^4]/[(0.5)^4 + (0.5)^4])$ $= \log_{10}(2^3[\theta^4 + (1 - \theta)^4])$. If you plug in different values of θ, for example, $\theta = 0.05$, you should get a lod score of $\log_{10}(2^3[(0.05)^4 + (0.95)^4]) = 0.81399$, which is the same as you calculated with the LINKAGE programs above.

Females are not always uninformative for linkage. Let us assume that in the pedigree we analyzed above, *dau1* married *husband*, who is unaffected, and has marker locus allele 3 as shown below, and had a child, *gson*, who was affected with marker locus allele 1 (of course, since *dau1* is homozygous for the 1 allele...), as shown in Figure 6-2.

In this case, we cannot tell whether a recombination occurred from *dau1* to *gson*, since *dau1* is homozygous at the marker locus. However, the fact that *gson* is affected forces *dau1* to be heterozygous at the disease locus, telling us that she received the 1 allele together with the disease allele from her mother. This will have a significant effect on the linkage analysis. Please add in this information to the pedigree in file EX4.PRE (saving it again as *EX4.PRE*), process the file with MAKEPED (saving it as *EX4.PED*), and reanalyze this pedigree, using the same command file (*PEDIN.BAT*) as before (It is not necessary to rerun LCP, since we are going to do the same analyses involving files with the same names as before). Just type *PEDIN* to analyze this new pedigree. The results should be as shown in table 6-2.

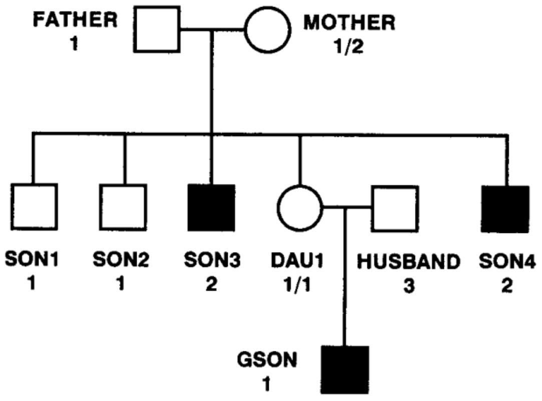


Figure 6–2. Extension of pedigree EX4.*

θ	Log₁₀(Likelihood)	Lod Score
0.0	-infinity	-infinity
0.05	-6.816792	-0.185952
0.1	-6.609155	0.021685
0.15	-6.530573	0.100267
0.2	-6.506597	0.124243
0.25	-6.512741	0.118099
0.3	-6.536299	0.094951
0.35	-6.567995	0.062845
0.4	-6.599351	0.031489
0.45	-6.622368	0.008472
0.5	-6.630840	0.000000

ILINK : $\hat{\theta} = 0.211$; $Z(\hat{\theta}) = 0.124938$

Table 6-2: Analysis results of modified EX4.PED; EX4.DAT

Do you notice anything interesting about these lod scores? If you look at the lod scores obtained in the first exercise in the book, the phase unknown family with five offspring, you will notice that the lod scores are identical. To understand why, let us again look at the analytical computation of the lod score. Using the notation above, if *mother* has phase (1), the four *son*s are recombinants, but *dau1*, who received a d 1 haplotype from *mother,* is a non-recombinant, with likelihood of $K\theta^4(1 - \theta)$. Similarly under phase (2), the four boys represent non-recombinants, and *dau1* had to have received a recombinant haplotype from mother, so the likelihood would be $K\theta(1 - \theta)^4$. The overall likelihood of the pedigree would be

$$K[\theta^4(1 - \theta) + \theta(1 - \theta)^4],$$

for a lod score of

$$\log_{10}([\theta^4(1 - \theta) + \theta(1 - \theta)^4]/[(0.5)^4(0.5) + (0.5)(0.5)^4]) = \log_{10}(2^4[\theta^4(1 - \theta) + \theta(1 - \theta)^4]),$$

which is identical to the lod score computed for files EX1.*, with the phase unknown family.

In this chapter, we learned how to analyze a sex-linked disease in the LINKAGE programs, and saw how to do it analytically for some simple examples. Different properties of sex-linked traits were introduced, especially the concept of hemizygosity, and its effects on linkage analysis. Further, we saw that if we change our pedigree file (or parameter file), without changing its name, we can rerun an analysis without going through LCP again, by just calling the same PEDIN.BAT file as was used in the previous analysis. Note that to do this, all file names must be unchanged, and the exact same analysis must be performed (i.e. MLINK starting from θ = 0, through θ = 0.45, in steps of 0.5, ILINK with starting recombination value of 0.2 in our example).

EXERCISE 6

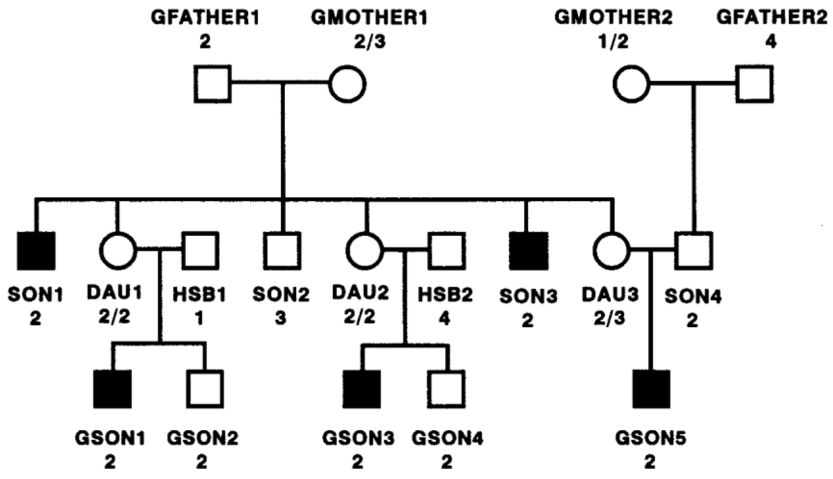


Figure 6–3. Sex-linked recessive pedigree USEREX6.*

Analyze the pedigree in Figure 6-3, assuming an X-linked fully penetrant recessive disease, with gene frequency of 0.01, and a marker with four equally frequent alleles (in files USEREX6.*). Try to analyze this family both analytically, and using the LINKAGE programs MLINK and ILINK, at the same points as the analysis in this chapter above. Try to see exactly what is going on at the genotype level, and which meioses are informative for linkage. Then find the equation for the lod score in this family. Compare the results with what was obtained from the LINKAGE programs. They should be the same...

7 Loops and Fastlink

In this chapter you will learn what a *loop* is, and how to deal with them in the LINKAGE programs. We will introduce the concept of *consanguinity* loops, and *marriage* loops, and their ramifications in a linkage analysis. Further, you will be learning how to deal with these in the LINKAGE programs, and finally, we will briefly introduce a utility program called *LOOPS*, which is designed to detect the presence of unbroken loops in a LINKAGE pedigree file, and is automatically run every time MAKEPED is used to process a pedigree file. For the first time, we will be analyzing more than one pedigree in one pedigree file. In most real linkage studies, you will have multiple pedigrees, and this is your first introduction to this situation.

With the introduction of the latest version of FastLINK, loops can be handled automatically. Thus, **everything in this book about breaking loops is theoretically correct but obsolete in practice**. As described in one of the FastLINK documents, loops can be handled automatically by the *unknown* program in FastLINK: You first feed your pedigree data into MAKEPED claiming that no loops are present even though the data contain loops. Then, based on the resulting *pedfile.dat* file, invoke “`unknown –l`” (‘ell’ for loops), which will produce a file called *lpedfile.dat*. This file is best copied over *pedfile.dat* (delete *pedfile.dat* and rename *lpedfile.dat* to read *pedfile.dat*) after which you proceed normally, that is, by running *unknown* again (but without the –l flag) and then the appropriate analysis program.

7.1 INTRODUCTION TO LOOPS

In the few very simple examples we've looked at of analytical calculation of likelihoods, you have seen just how complicated it can be, even in small nuclear families. In more complex pedigrees, it quickly becomes intractable to do such computations by hand. It also quickly becomes apparent that it is not straightforward how to design a computer program to handle any general pedigree structure in a theoretically justifiable manner. The *Elston-Stewart algorithm* [12] provides a way of calculating the likelihood in a recursive manner, allowing for the possibility of computer-based linkage analyses in general pedigrees. One of the main features of this algorithm is its dependence on *clipping* (or peeling). In clipping, small nuclear families, within a larger pedigree, are analyzed, and all of the information is collapsed on to one of the parents (or other relatives), whose own sibship is analyzed next, and so-on, until all of the information is collapsed onto one final person, the proband (For a detailed description, please consult Ott, 1991, pp.169-172). This is all pretty straightforward unless there is a loop in the pedigree. A loop is present in a pedigree when it is possible to start at any individual in a pedigree drawing, and draw a connected sequence of lines ending up back at the original individual, without retracing your steps. If a loop exists in a pedigree, the collapsing would circle around and around in the connected series of individuals, causing the algorithm to get caught in an infinite loop.

To circumvent this problem, the LINKAGE programs require that in each loop, one individual who is both an offspring and a parent must be “*doubled*”. This effectively “*breaks*” the loop, as the program will consider these two doubled individuals to be genotypically identical including phase, but still separate individuals. This allows the breaking of the infinite loop described above, and permits the likelihood to be computed, albeit slowly. The most efficient persons to double in any given pedigree are those members of the loop with the greatest amount of known genotypic information, including phase. This greatly reduces the computational time, as explained in Ott (1991).

There are two primary types of loops, consanguinity loops, and marriage loops. In a consanguinity loop, inbreeding is required. In other words, the parents of a given individual must be related. In a pure marriage loop, no inbreeding occurs, yet a completed pedigree circuit is created, nevertheless. An example would be two brothers who are married to two sisters, as you will see below. The distinction is immaterial for the LINKAGE programs, and is only important in that it points out that it is possible to have loops in a pedigree without inbreeding!

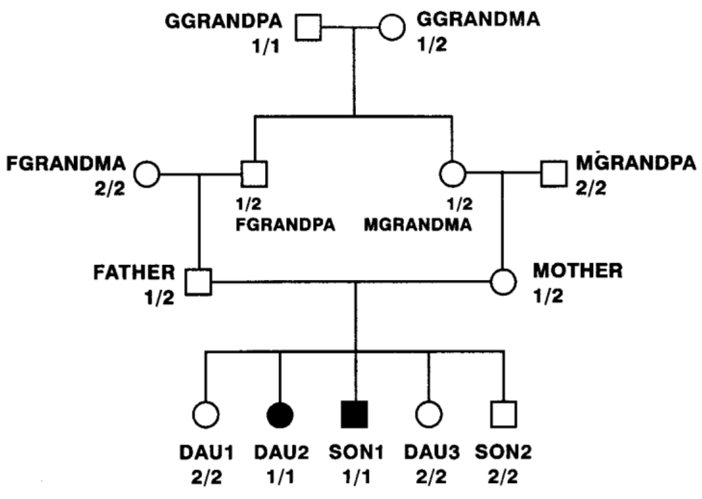


Figure 7–1. Pedigree from EX3A with added loop—EX5.*

7.2 CONSANGUINITY LOOPS

Let us consider again the recessive disease pedigree from EX3A.PED and EX3.DAT, which you analyzed in chapter 5. Suppose that we went out and collected more data on that extended family, only to learn that there was a consanguinity loop, as indicated in Figure 7-1.

Clearly this relatedness will have a major effect on the linkage information in the pedigree. Given the very small gene frequency of the disease allele, it is most likely that the disease allele was present in only one of the founders (persons with no parents in the pedigree). Hence, that makes *fgrandpa* and *mgrandma* the most likely carriers of the disease allele. Also, it adds a lot of linkage information, as you will see. So, let us analyze this new pedigree with the LINKAGE programs, as before. You can use the same parameter file, EX3.DAT, as nothing needs to be changed there. The only difference is in the pedigree file, so please enter the pedigree in the figure into a new file, EX5.PRE, in the same format as before (with pedigree identifier *ex5*), and then call up the MAKEPED program to process the pedigree file. Proceed in the MAKEPED program exactly as before, only when the program prompts you with the following:

```
Does your pedigree file contain any loops? (y/n)? ->
```

Respond *Y*, after which you will be asked if you have a file of loop assignments. The answer is *N*o. A file of loop assignments would be another file which indicates which individual in which pedigree to break the loop(s) at. You do not have one, and generally will not. Next, you will be asked the following:

```
Enter identifiers for each pedigree and person...
enter pedigree 0 when finished.
 Pedigree -> ex5
 Person -> father
```

In this case, you could break the loop at either the father, the mother, the paternal grandfather, or the maternal grandmother. You may wonder how to decide where to break it. It makes no difference in the results where you break the loop, but it may affect the computing time greatly. You should always break the loop at the individual with the least ambiguity in genotype. In this case, *father* and *mother* have got to be heterozygous at the disease locus, so you know their genotypes more exactly than the grandparents, since the disease allele doesn't have to come from a specific grandparent. So in this case, lets break the loop at *father*. Then type *0* when prompted with "Pedigree ->" to continue. You will next be asked:

```
Do you want these selections saved for later use? (y/n) ->
```

Respond *n*, since "saving" them will only create a file of loop assignments like the one you were asked for above, and will not affect the final *.PED file in any way. It is generally of little use to you to make a file of loop assignments. Next, have the program select all probands automatically, as before, to complete the processing of your file EX5.PED. Now analyze this pedigree with MLINK and ILINK as before, using the same parameter file (EX3.DAT), through the LCP program, and compare your results to those in Table 7-1.

θ	Log_{10}(Likelihood)	Lod Score
0	-11.622653	1.879919
0.1	-11.901587	1.600985
0.2	-12.221869	1.280703
0.3	-12.591054	0.911518
0.4	-13.015882	0.486690
0.5	-13.502572	0.000000

ILINK: = 0.001, Z() = 1.878120

Table 7-1: Analysis results of EX5.PED; EX3.DAT

Please compare these values with what was obtained in the prior analysis, in which your maximum lod score was only about half as large. In the absence of recombination, every phase-known meiosis contributes roughly 0.3 units of lod score. In this case, adding this consanguinity (two additional individuals typed), the additional information is roughly equivalent to three new phase known meioses. Why is this the case? Well, the fact that both affecteds are homozygous for the *1* allele, suggests that the disease is probably segregating together with the *1* allele throughout the pedigree, from a common ancestor. The concept of homozygosity mapping (Smith, 1953; Lander and Botstein, 1987) is based on this idea, that homozygous affecteds, whose parents are related, have most likely received a common haplotype without recombination from a common founder, allowing us to gain linkage information from all presumed non-recombinant meioses from the original founder haplotype to the affected kids. In recessive disease pedigrees, there is often a good deal of inbreeding. One of the reasons why marriages between closely related individuals is illegal is because of this increased propensity for recessive genetic disease among their offspring. However, for linkage analysis, these families are a godsend. Consider a family with the same structure as above, only with one affected child, with marker genotype 1 1. Consider everyone else in the pedigree to be untyped at the marker locus, and unaffected. Then analyze this pedigree with the LINKAGE programs assuming that the gene frequency of the 1 allele is 0.01, and there are 2 alleles at the marker locus (the disease is as defined for the previous example). Analyze this family (shown in Figure 7-2) with LINKAGE as you have learned to do. From only one typed individual who is affected and homozygous for this rare allele, you get a maximum lod score, at θ = 0, of 1.14.

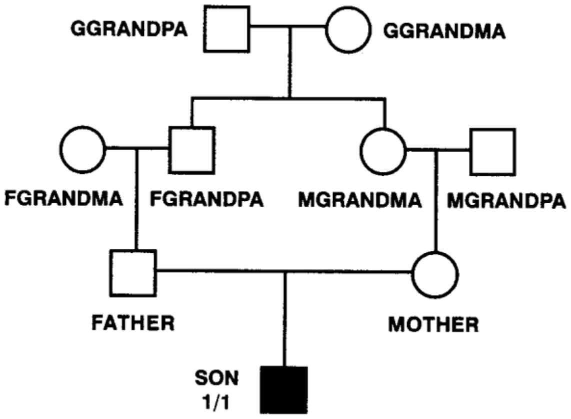


Figure 7–2. Homozygosity mapping pedigree.

This is because both the disease allele and the marker are so rare that they most likely entered the pedigree only once. Further, if they are both known to be present in the affected child, then he must have received them together without recombination from a common ancestor, assuming each allele entered the pedigree only once. The fact that we do not know which ancestor carried the disease allele, nor do we know which one carried the 1 allele adds noise to our analysis, but when one has a recessive disease, this shows that you can gain more linkage information per individual typed with inbred families in many situations, especially when you have a marker locus with a great many rare alleles. It is important to consider, however, that while inbred families make a linkage analysis more cost effective, in terms of the number of people you need to type, there is a large "cost" in computer time. In fact, the time it takes to complete a likelihood calculation in a pedigree increases exponentially with the number of loops in the pedigree.

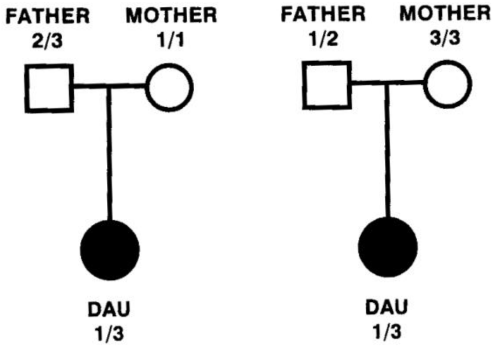


Figure 7–3. Two nuclear pedigrees—no loop—EX6A.*

7.3 MARRIAGE LOOPS

Let us consider the case where you have two pedigrees to analyze at the same time. This can be done quite simply by just including both pedigrees in the same pedigree file with different pedigree names. At present, you will be analyzing two nuclear families with one affected offspring each, as shown in Figure 7-3. Please make one pedigree file containing both pedigrees, and name the pedigrees *ex61*, and *ex62*.

Name these pedigree files *EX6A.PRE*, and then *EX6A.PED*. Make a parameter file as before with the PREPLINK program, specifying the disease to be recessive, with gene frequency for the

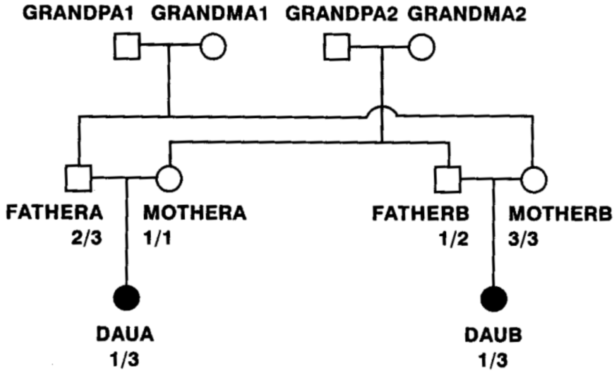


Figure 7–4. Same two nuclear pedigrees connected via marriage loop—EX6B.*

disease allele of 0.00001, as before. At the marker locus, however, allow for three alleles, with gene frequencies of 0.25, 0.40, and 0.35. Call this file *EX6.DAT*. Now analyze these families with MLINK and ILINK as before. As expected, there is no information in these pedigrees, since they are phase unknown matings with only one offspring, and no linkage disequilibrium. However, for interest's sake, take note of the magnitude of the log likelihoods. You should find lod scores of 0 everywhere.

But wait, you just discovered a relationship between these two pedigrees, that they are actually two siblings marrying two other siblings, as shown in Figure 7-4. This marriage creates a loop in the pedigree, because you can start from *fathera*, and connect him back to himself by going through ***fathera-motherb-daub-fatherb-mothera-daua-fathera***. This is a loop which must be broken before you can analyze this pedigree in the LINKAGE programs, as outlined above. Please enter this combined pedigree as one large pedigree and name this file *EX6B.PRE*. Call MAKEPED, and specify that there is indeed a loop in this pedigree, a marriage loop, which can be broken at either of the four parents. You must break a loop at an individual in the loop who is an offspring and a parent. The best strategy is to select one such individual about whom the most complete genotypic information, including phase, is known. In this example, we know that each of the parents are carriers, but the parents who are heterozygous at the disease locus carry somewhat less phase information. Clearly the homozygous parents are uninformative for linkage, and we know that in *motherb*, the disease and normal alleles are each on the same haplotype with a 3 allele. So, in this example, it may be slightly more economical to break the loop at either *mothera* or *motherb*, as they are homozygous at the marker locus, while the *father*s are heterozygous, and phase unknown. Proceed as in the previous example, and analyze the family with MLINK and ILINK. Output is given in table 7-2.

```
  θ      Log10(Likelihood)    Lod Score
_________________________________________
  0         -14.718037      0.374809
 0.1        -14.901117      0.191729
 0.2        -15.023110      0.069736
 0.3        -15.082021      0.010825
 0.4        -15.095476     -0.002630
 0.5        -15.092846      0.000000

ILINK: = 0.001, Z() = 0.373373
```

Table 7-2: Analysis results of EX6B.PED; EX6.DAT

Again, the addition of the information on the relatedness of these two pedigrees has taken two completely uninformative nuclear families, and given us one larger informative family with positive lod scores. Can you see why? How about the log likelihoods? Are they different from the combined family and the two smaller ones? Can you explain the magnitude of the difference?

The answer to all of these questions centers around the fact that in the case where you have two separate nuclear families, there are four independent disease alleles segregating. These alleles each have frequency of 0.00001, so the probability of observing four separate such alleles is $(0.00001)^4$. Clearly, once we add the relationships in, we only need two such disease alleles to explain the pedigree as observed. If one member of each grandparental pair is a carrier, then all of the parents can be carriers. The chance to observe two independent realizations of the disease allele is much larger, at $(0.00001)^2$, which is 10^{10} times more likely. Hence the likelihoods are much larger in the second pedigree, than the other two combined. As for the linkage information, again we obtain the additional information from the same phenomenon that characterized homozygosity mapping. Here, we have two first cousins who are both affected, and who have the same marker locus genotype. Further, we can tell that the daughters both got a *1* allele and a disease allele from either *grandma2* or *grandpa2*. Further, the *3* allele came from the other grandparental pair together with an additional disease allele. Since both grandparental pairs contributed an identical haplotype to their grandchildren, there is some linkage information present. Still, there is not much, since each child has one homozygous parent, further diluting the information present in this family. As an analogous case to the homozygosity mapping example above, consider this same pedigree, only make everyone unknown at the marker locus except the two daughters. Then, alter the parameter file so that the gene frequencies of the three marker locus alleles are 0.01, 0.98, and 0.01. In this way, we force the *1* and *3* alleles to have occurred only once each in this pedigree (most likely). Then rerun the analysis with the MLINK and ILINK programs as you have been doing. The maximum lod score from only having typed these two affected cousins is now

1.18, with all of the linkage information coming from the loop, and the fact that the observed alleles were so rare as to have been most likely occurred only once in the pedigree.

7.4 LOOPS PROGRAM

A computer program was written (Xie and Ott, 1992) to detect the presence of unbroken loops in pedigree files after they've been processed by the MAKEPED program. This program uses concepts of graph theory to detect connected graphs in the pedigree. Essentially, each individual is a node, and each marriage point is a node. If any unbroken path can be traced between nodes of the graph, from one individual back to himself without retracing, then the program informs you that an unbroken loop still exists, and it gives you a list of the nodes involved in the connected graph. Please re-process the files *EX5.PRE* and *EX6B.PRE* with MAKEPED, specifying that there are no loops. When MAKEPED calls the LOOPS program, it will tell you that a loop has been detected, and it will give you a listing of the nodes of the connected graph, in the file *LOOPS.OUT*. Then you can go back, and reprocess the file in MAKEPED, breaking the indicated loop. After breaking the loops, the LOOPS program should tell you that there are no loops detected in the pedigree. This program is now automatically run after MAKEPED, to detect any undeclared loops you may have overlooked. Also, if there are no loops, but you made data entry errors resulting in a loop, this will help you catch those. It is possible to use MAKEPED without using the LOOPS program (by typing MAKEPED1 instead of MAKEPED), but as a safeguard you should always use the LOOPS program.

In this chapter, you learned about loops, and why they are useful in a linkage analysis. You also saw how to analyze pedigrees with loops in the LINKAGE programs. Further, you were introduced to how to analyze multiple pedigrees at the same time, and were shown how to use the LOOPS program to detect any loops which may be present in a pedigree. It should be emphasized that although analyzing pedigrees with loops may be more cost efficient for the molecular biologist, they can be very slow to analyze. You will see just how slow when you try to solve exercise 7.

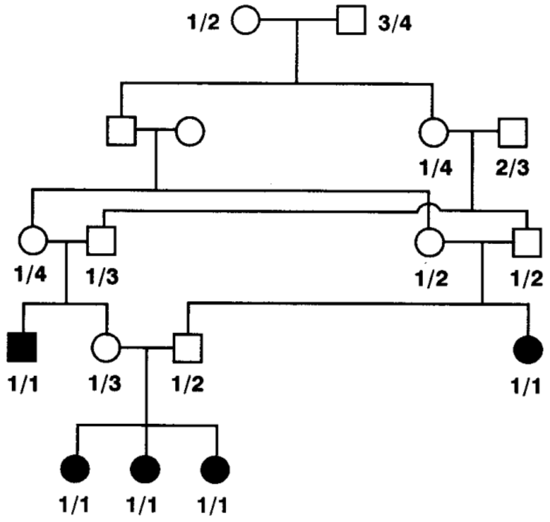


Figure 7–5. Complicated pedigree with multiple loops—USEREX7.*

EXERCISE 7

This problem involves the identification and breaking of consanguinity and/or marriage loops in pedigree data. Please look at the pedigree drawn in Figure 7-5. Make a pre-MAKEPED pedigree file and a parameter file for this pedigree (USEREX7.*). In this case, we will be analyzing a fully penetrant recessive disease with gene frequency of 0.001, and a four allele codominant allele numbers type of marker with equal gene frequencies for each allele. Remember that one should always break loops at an individual with the least genotypic ambiguity. How many loops are there in this pedigree? Where is the best place to break them? Be sure to check your processed file with the LOOPS program to make sure you've broken them all!

Try to solve this on your own. Answers are given in chapter 12, but since this is an advanced problem, we want you to try to think this through yourselves. If you are unsure how to break loops, reread the chapter about consanguinity and marriage loops.

Then, analyze this family with the MLINK and ILINK programs in the same manner as in the first problem. If you get the wrong answers, check to verify that you've correctly broken each and every loop present in the family. Also, be sure you haven't broken too many loops either...

8 Locus Types I : Allele numbers and binary factors

In this chapter, you will be introduced to the *binary factors* locus type, which is perhaps the most basic way of inputting data into the LINKAGE programs. You will also see the relationship between binary factors and the allele numbers locus type. We will further see a simple way to use this locus type to code the ABO blood group, and, for the first time, how to use LCP to analyze a subset of the total number of loci present in a given dataset.

In practice, people nowadays work mostly with two-allelic markers, so the binary factors locus type is rarely used. The one important thing in this chapter is the LCP program.

8.1 CODOMINANCE

In a binary factors type of locus, a phenotype consists of the presence or absence of a number of so-called *factors*. At a given locus, each allele is presumed to cause the presence of a different subset of the total number of factors codable by the sum of all alleles at that locus. It is important to point out that every allele has its own binary factors notation, and the "*or*" of any two alleles occurring in a genotype yields the notation for the corresponding phenotype. As a simple example, let us consider a locus with two codominant alleles (1 = 2, read 1 "is codominant with" 2). The first allele is assumed to cause the presence of one factor, and the second allele is assumed to cause the presence of another. In this case, the "factor" may be the presence of a given band on a Southern blot, corresponding to the allele in question, or the factor may simply refer to the observation at the phenotypic level of the presence of the same allele. It needn't have any real phenotypically meaningful interpretation – let us just assume that our locus can cause either of two factors to be observed. Alleles are coded by a sequence of 0's and 1's. For each factor at a given locus, we must enter one column in our datafile, consisting of either a *0* (if the factor is not present) or a *1* (if the factor is present). In this case, our allele 1 would be *1 0*, because the first factor (presence of allele 1) is present, and the second factor is absent (allele 2 is, of course, not present in allele 1). Analogously, allele 2 would be *0 1* at this locus. However, what we observe in an individual are phenotypes, which are composed of two alleles. Clearly in our codominant locus, we can have three genotypes, *1 1*, *1 2*, and *2 2* (in allele numbers format), each of which corresponds to a unique phenotype, since the locus is codominant. How do we combine alleles in this locus type? Well, it is basically a logical OR operation. If any factor is present in **either** the maternally or paternally derived allele, then it is present in the combined phenotype, much like the ***OR*** operation on a computer. So, if we have genotype 1/1, in binary factors, that is *1 0/1 0*. If we perform this logical *OR* operation we get *1 0 OR 1 0 = 1 0*. The first factor is present in either the first or the second allele (in this case, in both), and the second factor is absent from BOTH alleles. Hence, the phenotype would be *1 0*. For genotype 1/2, we would now have have binary factors genotype *1 0/0 1*, which would correspond to a phenotype of *1 0* OR *0 1 = 1 1*, since factor 1 is present in the first allele, and factor 2 is present in the second allele. Similarly, genotype 2/2 would correspond to a binary factors phenotype of *0 1*. It is also important to note that in the LINKAGE programs binary factor notation, a phenotype of *0 0* does **NOT** mean absence of both markers (which would not make sense here), but rather is the code for unknown (i.e. untyped) phenotypes. The entire situation can be summarized in table 8-1.

Real		Binary	Factor Status		Binary
		Factors			Factors
Geno.	Pheno.	Genotype	Factor 1	Factor 2	Phenotype
1/1	1 1	1 0/1 0	Present	Absent	1 0
1/2	1 2	1 0/0 1	Present	Present	1 1
2/2	2 2	0 1/0 1	Absent	Present	0 1
Unknown	0 0	Unknown	Unknown	Unknown	0 0

Table 8-1: Binary factors representation of codominant locus with 2 alleles, 2 factors.

This table shows both how to code such a codominant system as a binary factors type of locus with two alleles and two factors. This simplest application of the binary factors locus type has been implemented in the LINKAGE programs in a more user-friendly way, through the allele-numbers locus type, which you have been using throughout this book. The simple correspondence between allele numbers and binary factors is illustrated in table 8-1. As an exercise, go back to the original phase-unknown pedigree from chapter 2, which you entered in the files EX1.DAT and EX1.PED. At this point, add a third locus to the pedigree file (EX1.PRE), in binary factors format, which should be the binary factors equivalent of the allele numbers marker locus already entered in this file. Put these new phenotypes directly after the allele numbers phenotypes on each line, in the format from table 8-1. Your file should eventually look like the following:

```
ex1   father      0           0           1      2      1 2   1 1
ex1   mother      0           0           2      1      1 1   1 0
ex1   dau1        father      mother      2      1      1 2   1 1
ex1   dau2        father      mother      2      2      1 2   1 1
ex1   son1        father      mother      1      2      1 2   1 1
ex1   dau3        father      mother      2      1      1 1   1 0
ex1   son2        father      mother      1      1      1 1   1 0
```

Save this file, as *EX7.PRE*, and process it with MAKEPED to produce a pedigree file called *EX7.PED*. Next, we must modify our parameter file to indicate the addition of a new locus. Read the *EX1.DAT* file into the PREPLINK program by typing the following at the DOS prompt:

```
PREPLINK EX1.DAT
```

Now, from the main menu, select option *(k) See or modify loci description,* and from the subsequent menu, choose the *(c) ADD LOCUS* option. You will be returned to the same menu again, only with an additional locus indicated at the top of the screen. Now we will first need to *(e) CHANGE LOCUS TYPE*, specifying locus *3*, and specifying that it be changed to *(a) BINARY FACTORS*. Next, you should *(a) SEE OR MODIFY A LOCUS*, specifying locus *3*, and you will see a screen like the following:

```
******************************************
LOCUS NUMBER: 3
******************************************
(a) NUMBER OF ALLELES : 2
(b) NUMBER OF FACTORS : 2
(c) FACTORS PRESENT (1) OR ABSENT (0) FOR EACH ALLELE :
1 0
0 1
(d) GENE FREQUENCIES :
 0.500000 0.500000
(e) EXIT
******************************************
enter letter to modify values
```

Let us disregard what is already there, and set up this locus as discussed above. First let us set the number of alleles by choosing option *(a) NUMBER OF ALLELES*, and specifying that there are *2* alleles. Then, specify the *(b) NUMBER OF FACTORS* to be *2* as well, as explained above. Now, choose option *(c) FACTORS PRESENT...* You should respond to the questions as follows:

```
ENTER NEW FACTORS. 1=PRESENT 0=ABSENT
LEAVE A SPACE BETWEEN FACTORS
ALLELE 1
1 0
ALLELE 2
0 1
```

This is as we discussed above as well. Now, the gene frequencies should be set to be equal for the two alleles, as in the first exercise. Now, go back to the main menu, and save this new file as *EX7.DAT*. We are now ready to analyze this family with the LINKAGE programs.

Call up the **LCP program**, specifying pedigree file to be *EX7.PED*, and parameter file to be *EX7.DAT*. Proceed as in previous examples, by selecting the MLINK program. This time you want to do the same analyses as before, starting from θ = 0, in steps of 0.1, stopping at θ = 0.4 (remember that θ = 0.5 always gives a lod score of 0 by definition). However, this time, after specifying that the analysis be performed between loci 1 and 2 (the disease and the allele numbers locus), specify a further analysis with Locus Order = *1 3* (the disease and the new binary factors locus). To do this, just hit <Page Down> after entering the analysis parameters for 1 vs 2, and then returning to the same screen, and only modifying the first line to read *Locus Order 1 3*, followed by <Page Down>. Then go back, as before to specify the same analyses with ILINK, again specifying first 1 vs. 2, hitting <Page Down>, and repeating the analysis with 1 vs. 3. Then exit LCP with <Ctrl-Z>, and perform the analysis by typing *PEDIN* at the DOS prompt. Examine the FINAL.OUT file in your word processor, when the analysis is finished, and compare the likelihoods and lod scores for the two analyses, 1 vs. 2 , and 1 vs. 3. You should notice that everything is equal, as they are exactly the same locus, entered in a different format. If they are not always equal, then recheck your pedigree and parameter files for possible errors!

8.2 MULTIPLE FACTORS - TWO ALLELES (THE CEPH DATABASE)

There is a large databank for human genetics at the CEPH (Centre d'Étude du Polymorphisme Humain) in Paris (Dausset et al, 1990). In this database, much of the phenotypic data has been entered in this binary factors format, since historically, this was the first data entry format available. However, there are frequently situations in which there are two allele codominant systems with multiple factors. Many of these systems contain additional constituently present factors, for example, constant bands on a Southern blot. These add no additional genetic information, but many investigators thought it was more meaningful to include full banding patterns in the input files, for future reference. As an example of this, let us consider a two-allele codominant situation with five factors, of which factors 1, 3, and 4 are constituently present. Now, allele 1 would be coded as *1 1 1 1 0* (factors 1-4 present, factor 5 absent) and allele 2 would be coded as *1 0 1 1 1* (factors 1,3,4,5 present, factor 2 absent). In this case, the same logical *OR* operation can determine our phenotypic correspondence. In this case, we would have the situation summarized in Table 8-2.

Real		Binary Factors	
Geno.	Pheno.	Genotype	Phenotype
1/1	1 1	1 1 1 1 0 / 1 1 1 1 0	1 1 1 1 0
1/2	1 2	1 1 1 1 0 / 1 0 1 1 1	1 1 1 1 1
2/2	2 2	1 0 1 1 1 / 1 0 1 1 1	1 0 1 1 1
Unknown	0 0	Unknown	0 0 0 0 0

Table 8-2 : Alternative binary factors representation of codominant locus with 2 alleles, 5 factors.

As you can see, this is essentially the same as the two factor situation we utilized above, for if one eliminates the constituently present factors, 1, 3, and 4, we are left with exactly the same binary phenotypes as we had before. Still, for practice in using this form of the Binary Factors locus type, which is predominant in the CEPH database, please go back to the last example, files *EX7.PED* and *EX7.DAT*, and add a new locus, of the Binary Factors type, with two alleles and five factors as described in table 8-2. Make this locus identical in genetic information to loci 2 and 3 in this file. Save the new pedigree (*EX8.PED*) and parameter files (*EX8.DAT*). Then, analyze the disease versus locus 4, to be sure the likelihoods and lod scores are all identical with the ones you obtained from analyzing loci 1 vs 2, or 1 vs 3.

8.3 DOMINANCE AND RECESSIVITY

For the next extension, let us consider how to use the binary factors locus type in situations where the allele numbers locus type cannot be used. One such case occurs when one wishes to analyze a fully penetrant dominant locus. Let us assume that we have a two allele locus, which determines a dominant trait, and allele 1 is dominant over allele 2 (*1 > 2*). This can be visualized by assuming that allele one produces some protein, while allele 2 produces nothing, and the protein is produced equally effectively with one or two copies of allele 1. An intuitive way of coding this would be to specify one factor, the protein in question, and have allele 1 be *1*, for factor 1 present (the protein is produced), and allele 2 being *0*, for factor 1 absent (no protein produced). If this were the case, we would have a genotype : phenotype relationship as outlined in Table 8-3.

	Binary Factors	
Genotype	Genotype	Phenotype
1/1	1/1	1
1/2	1/0	1
2/2	0/0	0
Unknown	Unknown	0

Table 8-3 : Incorrect attempt at characterizing a dominant trait in binary factors notation with one factor.

This would be fine and dandy, except for the fact that unknown individuals as well as 2/2 individuals would be coded as 0. This is unacceptable, so we need to add an additional factor that is constituently present, for example, a factor "typed" with an intuitive meaning that if the individual was typed at the locus, this factor would be present, and if the individual was not typed, the factor would be absent. This factor would behave much like the functionless factors added in EX8.PRE, only now, having the additional function of discriminating between 2/2 individuals and unknown individuals, as shown in table 8-4.

	Binary Factors	
Genotype	Genotype	Phenotype
1/1	1 1/1 1	1 1
1/2	1 1/0 1	1 1
2/2	0 1/0 1	0 1
Unknown	Unknown	0 0

Table 8-4 : Correct representation of a dominant trait in binary factors notation with 2 factors.

Please go back to this same example, and recode the fully penetrant disease as if it were a binary factors locus type, keeping the gene frequencies the same. Then check it by comparing an analysis of locus 1 vs 2 with that of locus 5 (the new binary factors representation of the disease) vs locus 2. They MUST be identical as well.

A recessive condition is also immediately specified as well, since whenever we have a dominance relationship there is a built-in recessive relationship. In this case we had *1 > 2* (Allele 1 is dominant over Allele 2), which means that *2 < 1* (Allele 2 is recessive to Allele 1), since it is required for there to be 2 copies of the non-protein producing allele, in this example, for the "absence of protein" phenotype to be observed. This is the definition of recessivity. Thus, we can use the same coding scheme as above to code a recessive system, only being sure to specify allele 2 as the recessive disease predisposing allele!

8.4 SYSTEMS WITH DOMINANCE AND CODOMINANCE, ABO BLOOD GROUP

More complicated dominance/codominance relationships can be specified as well with the binary factors notational system. Consider the ABO blood group. To this point, we have not learned any way of coding this in a LINKAGE input file, even though it is among the most basic loci in humans. In fact, many early linkage studies were done with this very locus, and we do not yet know how to use this kind of data. Well, these relationships can be easily coded for in the LINKAGE programs as a binary factors locus. Think about this for a while on your own. If you cannot figure it out, then go on to read the following chapter.

In the ABO blood group, it is well known that there are three alleles, A, B, and O. The A allele causes, among other things, the production of a certain cell-surface antigen "A". Similarly, the B allele produces cell-surface antigen "B". The O allele, on the other hand produces NO cell-surface antigens. The actions of these three alleles, relative to the production of cell-surface antigens, are independent, so it can be clearly seen that the dominance relationship can be summarized as follows: *(A = B) > O* (A and B are codominant, and both are dominant over the O allele). We know how to code the codominance, and we know how to code the dominance, but how can we combine the two phenomena in the same locus? Well, let us first begin by considering that the cell-surface antigens A and B could be considered as binary factors. If this were the case, allele A would be *1 0*, allele B would be *0 1*, and allele O would be *0 0*. This locus could then be summarized as in Table 8-5.

ABO		Binary Factors	
Phenotype	Genotype	Genotype	Phenotype
A	A/A	1 0 / 1 0	1 0
	A/O	1 0 / 0 0	1 0
B	B/B	0 1 / 0 1	0 1
	B/O	0 1 / 0 0	0 1
AB	A/B	1 0 / 0 1	1 1
O	O/O	0 0 / 0 0	0 0
Unknown	Unknown	Unknown	0 0

Table 8-5 : Incorrect representation of ABO blood group in binary factors format with 2 factors.

This is an intuitively satisfying approach, except that now O blood type individuals are indistinguishable from unknown individuals in terms of their binary factors phenotypes. Disregarding the O allele completely, however, we see a straightforward codominance relationship between A and B alleles. Further, disregarding the A allele, we can see a straight dominance relationship specified with *B* > *O*. The same holds for *A* > *O*. Thus, our dominance relationships are correct, and we merely need to allow the program to distinguish O phenotypes from unknown individuals. This is the same situation we encountered in the straight dominance example above. To remedy this situation add an additional constituently present factor (“genotyped”), to allow for this discrimination, as shown in Table 8-6.

ABO		Binary Factors	
Phenotype	Genotype	Genotype	Phenotype
A	A/A	1 0 1 / 1 0 1	1 0 1
	A/O	1 0 1 / 0 0 1	1 0 1
B	B/B	0 1 1 / 0 1 1	0 1 1
	B/O	0 1 1 / 0 0 1	0 1 1
AB	A/B	1 0 1 / 0 1 1	1 1 1
O	O/O	0 0 1 / 0 0 1	0 0 1
Unknown	Unknown	Unknown	0 0 0

Table 8-6 : Correct representation of ABO blood group in binary factors notation - 3 alleles.

In this chapter, you learned about the binary factors locus type, and how to use it to enter phenotypic data under codominance, dominance, recessivity, and any combination thereof. We also studied the relationship between Binary Factors and Allele Numbers locus types, and learned that the allele numbers locus type is just a shorthand form for entering codominant binary factors data. Additionally, for the first time, you used LCP to extract subsets of loci from a pedigree and parameter file to perform analyses on various subsets of the loci in these files.

EXERCISE 8

Go back and re-enter the data from all previous user exercises, adding binary factors representations of all allele numbers locus types, and fully penetrant diseases. Compare the results to make sure they are compatible.

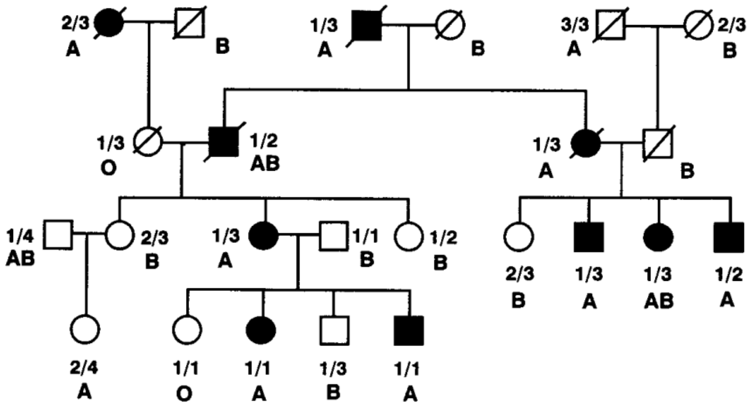


Figure 8–1. ABO blood group added to pedigree from Figure 2-3—USEREX8.*

Add the ABO blood group data to the pedigree in exercise 2, as shown in Figure 8-1 (Make new files USEREX8.*, containing the disease and both loci). Use gene frequencies of 0.28 for the *A* allele, 0.06 for the *B* allele, and 0.66 for the *O* allele (Cavalli-Sforza and Bodmer, 1971). Analyze this data in 2-point analysis, disease vs. ABO, and marker 1 vs. ABO. Are these results consistent? Why or why not?

9 Advanced applications of Affection Status I : Incomplete Penetrance Revisited

In this chapter we will be using the *affection status* locus type to allow for various penetrance models which are sometimes quite complicated. We will review the basics of incomplete penetrance, and learn how to allow for age-dependent penetrance through the use of so-called *liability classes*, also allowing for phenocopies in a possibly age-dependent manner.

We learned in section 5.4 that a disease-predisposing genotype may not always lead to disease. For example, consider a simple dominant disease locus with two alleles, *d* and +, so that genotypes *dd* and *d*+ cause disease while individuals with genotype ++ are unaffected. With incomplete penetrace, the conditional probability of being affected is g = P(affected|*dd*) = P(affected|*d*+). Often, one also encounters patients with the ++ genotype; these are called **phenocopies** with respect to the disease locus under consideration (Goldschmidt coined the term “Phänokopie” in 1935 [13]). Thus, the penetrance for a phenocopy is the conditional probability, f = P(affected|++), with $f < g$. There are various potential reasons for phenocopies to occur. One possibility is that phenocopies are due to a second disease locus at a different genomic location (digenic inheritance [14-16]). A simple strategy to find such a second locus is as follows [17]: Mask all non-phenocopy affecteds by making their phenotype “unknown” and then search for linkage of markers to a locus compatible with the phenocopies.

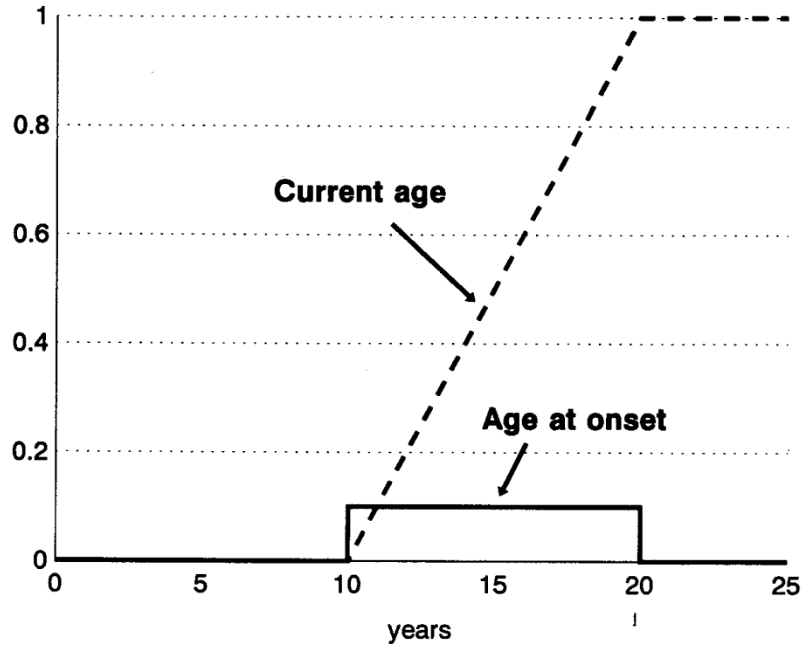


Figure 9–1. Graphical representation of straight line age-of-onset functions

9.1 AGE-DEPENDENT PENETRANCE

One potential problem in a linkage analysis is the presence of incomplete penetrance. In chapter 5, we saw how to allow for constant reduced penetrance. However, in most cases, the penetrance reduction is not constant, but is rather dependent on age. Let us return to the pedigree EX1.PED, from chapter 2, and reanalyze it under the assumption of age dependent penetrance. In this case, we will assume that the penetrance is 0 for persons up to 10 years of age, and then age of onset is uniformly distributed, with a maximum penetrance of 1 at age 20. This age of onset function basically means that given the susceptible genotype, everyone will become affected at some point between age 10 and 20, with each age of onset being equally likely. Graphically, the density function and corresponding distribution function are shown in Figure 9-1.

To incorporate such an age of onset function in the LINKAGE programs, you will have to use what are known as liability classes in conjunction with the affection status locus type. In this case, we can have separate penetrance definitions for persons of each age from age ≤ 10 (penetrance = 0), through age ≥ 20 (penetrance = 1) as follows: Read the file INC.DAT into the PREPLINK program. Then *(k) See or modify loci description*, and *(a) SEE OR MODIFY A LOCUS*, for Locus *1*, the disease. Now, we want to define the age of onset penetrances by using liability classes, but how many liability classes do we need? Well, to figure this out, we will need to know the ages of the persons in the pedigree (current age, or age last seen). The ages are as follows: *father* = 50, *mother* = 45, *dau1* = 8, *dau2* = 13, *son1* = 16, *dau3* = 17, *son2* = 22. The penetrances for each of them are easily computed, given our assumptions about age dependent penetrance. For simplicity, and since we have only information about current age, and not age of onset, we will use the distribution function of the age of onset as our penetrance function, i.e. penetrance = P(becoming affected at or before the current age │ Genotype). For persons aged 10 and under, the penetrance is 0; for persons 20 and over, the penetrance is 1; and for persons between 10 and 20, the penetrance is just 0.10 × (age - 10), which is the equation of the distribution function, i.e. the line between coordinates (age = 10, penetrance = 0) and (age = 20, penetrance = 1). Therefore, the penetrances for each person in the pedigree are as follows: *father* = 1, *mother* = 1, *dau1* = 0, *dau2* = 0.3, *son1* = 0.6, *dau3* = 0.7, *son2* = 1. Since there are 5 different penetrance classes needed to specify the penetrances for all the persons in this pedigree, we will need to use PREPLINK now, to change *(b) NUMBER OF LIABILITY CLASSES* to

5. Then, you must modify the penetrances for each of the five liability classes as follows: Select *(c) PENETRANCES:*, and modify the penetrances as follows: Let liability class 1 be for persons aged less than 10 years (*dau1*, for example). The new penetrances should be entered as follows:

```
LIABILITY CLASS: 1
GENOTYPE 1 1 OLD PEN 0.000000
?
0
GENOTYPE 1 2 OLD PEN 1.000000
?
0
GENOTYPE 2 2 OLD PEN 1.000000
?
0
```

This is because the probability of being affected is 0 for all genotypes in this age class. Next you will be asked to provide penetrances for liability class 2, which, for convenience sake, should be the next largest penetrance value, which would be for the 13 year old *dau2*, who has penetrance of 0.3. Since this disease is still assumed to be autosomal dominant with reduced penetrance, the value 0.3 should be given for both susceptible genotypes, *1 2*, and *2 2* (*2* being the disease allele). Continue as follows:

```
LIABILITY CLASS : 2
GENOTYPE 1 1 OLD PEN 0.000000
?
0
GENOTYPE 1 2 OLD PEN 0.000000
?
0.3
GENOTYPE 1 2 OLD PEN 0.000000
?
0.3 ...
```

Continue to enter penetrances for the remaining liability classes as indicated in table 9-1.

Liability	Penetrances for Genotypes				Relevant
Class	Age	1 1	1 2	2 2	Individuals
1	<10	0	0	0	dau1
2	13	0	0.3	0.3	dau2
3	16	0	0.6	0.6	son1
4	17	0	0.7	0.7	dau3
5	>20	0	1	1	father, mother, son2

Table 9-1 : Penetrance class definitions for INC2.PED;INC2.DAT

After you have entered these penetrances, return to the main menu, and save this file as INC2.DAT. Next, copy the file EX1.PRE to INC2.PRE, and read it into your word processor. You must now modify the phenotypes given for the affection status locus, by adding an extra column, indicating the liability class for each individual. In this case, you can determine which liability class goes with each individual by consulting table 9-1. Please add an extra column after the affection status (1 = unaffected, 2 = affected), with the appropriate liability class number. One important thing to note, that is different from all other locus types, is that when an individual is unknown at an affection status locus with multiple liability classes, you must specify the individual as being 0 (Unknown), in any liability class. Which liability class you use is immaterial, but this second column **MUST** be non-zero. A safe way of coding this information would be to put the individual in the appropriate liability class, based on his age, in case you go back to add new phenotypic information later, although using **0 1** as the code for unknowns will give appropriate results for ALL unknowns as well! The final pedigree file (INC2.PRE) should look like the following:

```
ex1   father      0           0           1      2 5   1 2
ex1   mother      0           0           2      1 5   1 1
ex1   dau1        father      mother      2      1 1   1 2
ex1   dau2        father      mother      2      2 2   1 2
ex1   son1        father      mother      1      2 3   1 2
ex1   dau3        father      mother      2      1 4   1 1
ex1   son2        father      mother      1      1 5   1 1
```

Now, process the file with MAKEPED, to produce pedigree file INC2.PED, and analyze the pedigree with MLINK and ILINK, through the LCP shell program, as you have been doing throughout the book. The output is presented in table 9-2.

θ	Log_{10} Likelihood	Lod Score
0.0	-8.152963	0.789145
0.1	-8.321516	0.620593
0.2	-8.505736	0.436372
0.3	-8.698514	0.243595
0.4	-8.868216	0.073893
0.5	-8.942108	0.000000

ILINK: $\hat{\theta}$= 0.000, Z($\hat{\theta}$) = 0.789145.

Table 9-2 : Analysis results of INC2.PED; INC2.DAT

Can you understand why these results are so different from those we obtained when we allowed for complete penetrance in chapter 3, and from the simple incomplete penetrance model used in chapter 5? The individual *dau1* was the individual who was a likely recombinant if penetrance was complete, but when one allows for the fact that at her young age, she had no possibility to be affected, regardless of her genotype, all the evidence for a recombination disappears.

If there is full penetrance in reality, though, and you model the disease as having reduced penetrance, you will tend to lose information, as you will see in the next example. Please go back to the known fully penetrant recessive disease example we analyzed in chapter 5 (files EX3.*). Let us model this disease as if it were autosomal recessive with constant reduced penetrance of 50%. Modify the penetrances for the affection status locus in the parameter file, EX3.DAT, by the same method as in the dominant disease above. This time, however, change the penetrances to fit an autosomal recessive trait with 50% penetrance.

Then, save this file as INCREC.DAT, and reanalyze the pedigree in EX3A.PED. The lod scores obtained (with constant penetrance of 50%) are shown in table 9-3. With age-dependent penetrance classes as in the dominant case above, the following six lod scores should be obtained: 0.699815, 0.523004, 0.344353, 0.177339, 0.049606, and 0.

θ	Log_{10} Likelihood	Lod Score
0.0	-16.020575	0.776033
0.1	-16.200742	0.595866
0.2	-16.391282	0.405326
0.3	-16.579914	0.216694
0.4	-16.733805	0.062803
0.5	-16.796608	0.000000

ILINK: $\hat{\theta}$= 0.000, Z($\hat{\theta}$) = 0.776033.
Table 9-3 : Analysis results of EX3A.PED; INCREC.DAT

If you go back to the example in chapter 5 when we analyzed this pedigree under the model of complete penetrance, you will see that our maximum lod score was 0.976874. So, information was lost by specifying the reduced penetrance model. Whereas all unaffected persons were previously known not to be homozygous for the disease, under this model, it is possible for any of the affecteds to have the disease susceptibility genotype. Hence, it is not clear whether or not they represent recombinants. Here, there is a loss of information from allowing for reduced penetrance, while in our previous example, the lod score rose substantially from such allowances. There are situations where a wrong model may give a higher lod score than the true model, but under the true genetic model, your results will be better, in general.

Instead of working with penetrance classes, the LIPED program (chapter 11) allow you to specify a straight-line onset curve such as the one shown in Figure 9-1 and enter current ages of individuals, where for an unaffected individual the age is given as a negative number. Working with the same data as those leading to the results shown in Table 9-2, we assume that the N/N genotype is not susceptible to disease. The input file to LIPED then looks as follows:

```
1 0000 0.0          Straight-line age of onset
           0   m   0   0
 2 2   <- number of alleles
 1 3   <- number of phenotypes
 3 0   <- locus type
 3     <- output option
 dis   D   N
    0010    9990
   D   D  10  20   1  10  20   1
   D   N  10  20   1  10  20   1
   N   N   0   0   0   0   0   0
 SNP   1   2 1/1 1/2 2/2
    3000    7000
   1   1   1   0   0
   1   2   0   1   0
   2   2   0   0   0
   7   0Family EX1.PED, Handbook Figure 3-1 and section 9.1
  FA   0   0   m  50 1/2
  MO   0   0   f -45 1/1
  D1  FA  MO   f  -8 1/2
  D2  FA  MO   f  13 1/2
  S1  FA  MO   m  16 1/2
  D3  FA  MO   f -17 1/1
  S2  FA  MO   m -22 1/1
9000
```

and furnishes the following results:

```
  R MALE  R FEM.   LOG10[L(R)]   LOD-SCORE
  0.5000 0.5000      -7.46284      0.000
  0.4500 0.4500      -7.44322      0.020
  0.4000 0.4000      -7.38894      0.074
  0.3500 0.3500      -7.31063      0.152
  0.3000 0.3000      -7.21924      0.244
  0.2500 0.2500      -7.12289      0.340
  0.2000 0.2000      -7.02646      0.436
  0.1500 0.1500      -6.93253      0.530
  0.1000 0.1000      -6.84224      0.621
  0.0500 0.0500      -6.75597      0.707
  0.0010 0.0010      -6.67530      0.788
  0.0001 0.0001      -6.67385      0.789
```

Clearly, there is excellent agreement between the results from LINKAGE and LIPED.

9.2 DISTRIBUTION FUNCTIONS VS. DENSITY FUNCTIONS

It is of importance to delineate exactly what age it is that you are reporting as part of a phenotype. If the individual concerned is affected with the disease, then there are two different types of age observation possible. If it is known at what age the person became affected with the disease, then one should report this *age of onset* for the disease. Sometimes, however, it is not known at what age an individual became affected, and is merely known that they are currently affected with the disease, in which case the *current age* (or *age of last examination*) should be reported in the phenotype. If the age of onset is only approximately known, it may still be useful to incorporate such information in the analysis. These two situations require very different penetrance definitions. If one is dealing with current age, then all that is known is that at some point before the current age, the person became affected. Hence, the probability we are interested in is P(affected before current age), which is a cumulative distribution function. This would be analogous to the age of onset distribution applied in the first example in this chapter. If the data is available on age of onset, however, it is imperative to consider something a little bit different, P(affected at age of onset), which can be taken more accurately from the probability density function, not the distribution function. This difference is very subtle, but can be of major importance. For unaffected individuals, of course, there is no such thing as an age of onset, but merely the age of last exam (or current age). In this situation, we are interested in P(not affected before age of last exam) = 1 – P(affected before age of last exam). Therefore, this is also based on the distribution function described above.

Let us consider the simple case of a dominant disease. For any given individual, the penetrances, *f*, for the three genotypes are as follows: P(aff│DD) = P(aff│Dd) = *f*(age); P(aff│dd) = 0. For any affected individual, therefore, genotype dd would be impossible, and the other two genotypes would have equal penetrance, so the only discriminatory power comes from the elimination of genotype dd. In this case, you can see that the likelihood of any affected individual may be written as *f*(age)P(DD) + *f*(age)P(Dd) + 0 P(dd) = *f*(age)[P(DD)+P(Dd)]. If P(DD) is a function of θ, then the likelihood ratio would be

$$\frac{f(age)[P(Dd;\theta) + P(DD;\theta)]}{f(age)[P(Dd;\theta = 0.5) + P(DD;\theta = 0.5]}.$$

The numerical value of *f*(age) is unimportant, therefore, since it will only act as a constant multiplier of the likelihood in both numerator and denominator of the likelihood ratio, and will factor out of the lod score equation. So, in the absence of phenocopies, the numerical value of the penetrance is immaterial for affected individuals. However, for unaffected individuals, the corresponding penetrances are as follows: P(Not Aff.│DD) = P(Not Aff.│Dd) = 1 – *f*(age); P(Not Aff.│dd) = 1. In this case, the numerical value of *f*(age) plays an important role in discriminating between genotypes, since there is no longer this equal weighting factor. The likelihood of any unaffected individual is thus [1 – *f*(age)] × [P(Dd│θ) + P(DD│θ)] + P(dd│θ). There is no longer a common factor to cancel in numerator and denominator, and thus the numerical value of the penetrance is very crucial to the analysis.

9.3 PHENOCOPIES

The situation is much more complicated, however, when the presence of *phenocopies* is allowed for. If there are separate age distributions for phenocopies and genetic cases, then there can be rather complicated situations arising. For unaffecteds, the distribution function must be used for both penetrances, for the same reasons outlined above. However, for affecteds, the distinction between density and distribution function can be crucial. Consider that if *f*(age) is the penetrance for genetic cases, and f_p(age) is the penetrance for phenocopies, then the likelihood of any affected individual is just *f*(age)[P(DD) + P(Dd)] + f_p(age)P(dd). No longer is there a common factor which can factor out of numerator and denominator in the likelihood ratio. Therefore, one must pay special attention to the numerical value of the penetrances. Let us consider the extreme case of a genetic disease with age dependent penetrance according to a straight line age of onset distribution function, on the range of 10 years to 20 years, with the penetrance for individuals at age 20 or above being full. Similarly, let the phenocopies follow a similar straight line distribution function, starting from 0 up to age 15 and rising to 10% at age 30. It is significant to point out that in general a uniform distribution is far from ideal to use, since its density function is either 0 or a fixed larger value. In general, it is wiser to use a lognormal or normal age of onset density function, as it will allow age to be a greater discriminatory factor in terms of interpreting the genotypic background given phenotypes. The distribution functions and densities corresponding to our Uniform distributions are outlined in table 9-4.

	Phenocopies		Genetic Cases	
Age	Distribution	Density	Distribution	Density
0-10	0	0	0	0
10-15	0	0	0.1×(age-10)	0.1
15-20	0.0067×(age-15)	0.0067	0.1×(age-10)	0.1
20-30	0.0067×(age-15)	0.0067	1	0
30-	0.1	0	1	0

Table 9-4 : Uniform distribution and density functions shown for phenocopies and genetic cases.

If we think of the likelihood of an individual, f(age)[P(DD) + P(Dd)] + f_p(age)P(dd) as being a function of the ratio of phenocopies to genetic cases, then we can parametrize it as f(age)[P(DD) + P(Dd)] + kf(age)[P(dd) = f(age) [P(DD) + P(Dd) + kP(dd)]; where $k = f_p(\text{age})/f(\text{age})$. In this representation, f(age) is now equal in numerator and denominator, and will be factored out of the likelihood ratio. Therefore, the entire amount of information available from the penetrances comes from the ratio f_p(age)/f(age). Values of f_p(age)/f(age) are shown in table 9-5 for all age ranges.

Age	Distribution	Density
0-10	UNDEFINED	UNDEFINED
10-15	0	0
15-20	0.067×(age-15)/(age-10)	0.067
20-30	0.0067×(age-15)	∞
30-	0.1	UNDEFINED

Table 9-5 : Values of f_p(age)/f(age) for each age group.

These ratios are extremely different between the groups, as you can see. Consider an individual with age of onset at age 25. If we were to use the density functions, we would see that it was impossible for him to have been a genetic case, so he would definitely have been interpreted as a phenocopy. However, based on the distribution function, the ratio would be only 0.067. The interpretations are highly different, since the latter would not give much discriminatory power, and might favor the genetic causes, while the former would imply that it was definitely a phenocopy. These ratios are the most important factor in any penetrance model. Clearly, when the ratio is 1:1, there is no phenotypic basis for discriminating between the possible genotypes, and the situation is analogous to that in which the true phenotype is unknown. The more deviant from 1 this ratio is, the greater the power to discriminate between genotypes based on phenotype.

In selecting a model, it is important to select a phenocopy probability that makes sense on a population level. If the population prevalence of a disease, φ, is known, then it is imperative that the disease gene frequency, p, and penetrances, f, satisfy $\varphi = f_{DD}p^2 + 2f_{Dd}p(1-p) + f_{dd}(1-p)^2$. Assuming a dominant disease, in the extreme case, for $p = 0$, clearly $\varphi = f_{dd}$, since all cases are non-genetic, so it is clear that $f_{dd} \leq \varphi$ in all cases. Analogously, for a sex-linked recessive, the situation is that only males can be affected, so that effectively $\varphi = pf_D + (1-p)f_{+}$., where φ is the prevalence in males. In this situation, again, $f_{+} \leq p$, and all penetrances must be selected such that the prevalence equations are satisfied.

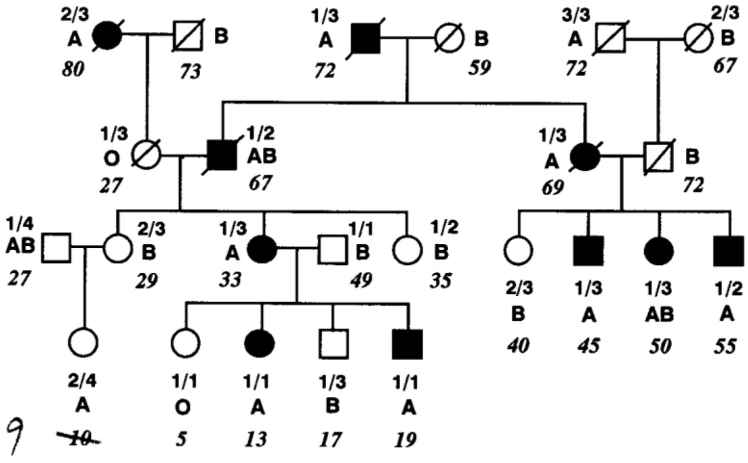


Figure 9–2. Ages added to pedigree from Figure 2-3—USEREX9.*

EXERCISE 9

Again, using the same family as in exercise 8 (files USEREX8.*), assume that there is now incomplete penetrance following a straight line age of onset curve starting from 0.1 at ages less than 10 up to 0.9 at ages 60 and above, where the current ages are given in Figure 9-2 (for more information, consult Ott (1991), p. 160). Please divide your pedigree members into appropriate age classes, and define liability classes containing the penetrances for individuals in each age class. For simplicity's sake, we recommend using age classes as follows (0-9), (10-19), (20-29),... Please use the midpoint of each age class as the basis of determining the penetrances for people in that class ,e.g. use age 15 to calculate penetrances for people in the age class (10-19). Please analyze the pedigree data starting from θ = 0 up through θ = 0.5 in steps of 0.1 with MLINK, and also analyze the data with ILINK, as you have done throughout this chapter. Please consider both disease vs. marker 1, and disease vs. ABO.

10 Advanced Applications of the Affection Status Locus Type II: It's not just for diseases any more...

In this chapter, you will be learning how to use the affection status locus type to allow for nonstandard situations. We will consider complicated dominance relationships, including allowing the identification of obligate carriers. Then, we see how to include data about loci at which there is only partial information available. Along the same lines as that, we will discuss elementary approaches to modelling errors in marker typings, and diagnostic uncertainty.

10.1 GENERALIZED DEFINITION OF THE *2* PHENOTYPE

The fact that a locus is coded as an "affection status" does not in any way imply that the locus must be a disease locus. It merely defines the way the data will be entered in the pedigree and parameter files for the analysis. Further, it is an unfortunate convention that people think of the phenotype *2* as affected, and of the phenotype *1* as unaffected at this type of locus. While this may typically be the case in any given analysis with a disease, in reality the phenotype *2* really means presence of a given phenotype defined by the penetrances given in the parameter file. Further, the phenotype *1* really means the absence of the phenotype defined by the penetrances given in the parameter file. This may sound confusing, but it really is only an unfortunate historical consequence. You see, this locus type was originally formulated with the idea that it would be used only for diseases. Its full potential was apparently not realized at that time. In the case of a disease, the *2* phenotype means presence of the phenotype "affected with the disease" defined by the penetrances in the parameter file, and the *1* phenotype means absence of that phenotype defined by the same penetrances. However, in no way does the program assume anything regarding the biological meaning of the *2* phenotype. As an extreme case, in a recessive disease, the trait "unaffected" is dominant over the "affected" trait. One could use an affection status locus to code the trait normal, by defining penetrances corresponding to a fully penetrant dominant trait, with gene frequency of the trait allele being very high. Then, the phenotype *2* could correspond to presence of the "unaffected" phenotype, and *1* would mean absence of the "unaffected" phenotype (or, simply affected with the disease). As an additional example, reconsider the fully penetrant dominant disease data from files EX8.*. Please add a sixth locus to this pedigree. First read the parameter file, EX8.DAT into PREPLINK, and add an additional locus to the file, as you have already learned to do. At this additional locus, our goal is to set up the penetrances such that the *2* phenotype will correspond to the absence of disease. In other words, our phenotype in this case is "absence of disease". We have already seen that recessivity and dominance are the same thing, in that when (1 > 2), clearly (2 < 1). So in our fully penetrant dominant disease, (disease > non-disease), so (non-disease < disease). Thus, we should set up this new locus to represent a fully penetrant recessive condition.

Change the newly added locus in the parameter file to an affection status locus type, and set up the penetrances to correspond to a fully penetrant recessive condition, with penetrances as follows:

```
ENTER NEW PENETRANCES
GENOTYPE 1 1 OLD PEN 0.00000E+00
?
0
GENOTYPE 1 2 OLD PEN 0.00000E+00
?
0
GENOTYPE 2 2 OLD PEN 1.00000E+00
?
1
```

The only thing we need to be careful about now is our gene frequency representation. Clearly, the recessive allele is the non-disease allele, so it must get the corresponding gene frequency of 0.99999. In this representation, the non-disease allele is allele number 2, so you must modify the gene frequencies, such that allele 1 has frequency of 0.00001, and allele 2 has frequency of 0.99999. Finally, save this new file as EX9.DAT.

Now, go back to the pedigree file, EX9.PRE, and add the new affection status phenotypes at the end of each line (after the 2 binary factors loci). Be sure to code all diseased individuals as *1*, and all normal individuals as *2*. This should be exactly the opposite of the first locus, in the sense that all individuals who were coded as *2* at locus 1, should now be coded as *1* at locus 6, and vice versa. Please analyze locus 6 vs 2, by setting up the MLINK and ILINK analyses in LCP in the same manner as you have been doing throughout the book. The results should be identical to the lod scores and likelihoods obtained in the original analysis of this pedigree in chapter 3. Now, you should be able to clearly see that the *2* phenotype does not necessarily mean affected, but has a biological meaning only in accordance with how you defined the penetrances in the parameter file.

10.2 CODOMINANT MARKER LOCI

We continue this chapter by pointing out that you can code markers as affection status loci as well, if you wish to. Let us reconsider the same pedigree from files EX9.*, with the fully penetrant autosomal disease. We shall now add a sixth locus to this pedigree, in which we will recode the codominant allele numbers locus as an affection status locus type. For our example, we merely want to define a simple codominant marker as an affection status locus to show the potential equivalence of locus types, so for the marker we've been looking at in this example, the penetrances are presented in table 10-1.

Liab. Class	Allele Numbers	Penetrances for Genotypes			Affection Status Coded as
		1 1	1 2	2 2	
1	1 1	1	0	0	2 1
2	1 2	0	1	0	2 2
3	2 2	0	0	1	2 3

Table 10-1: Affection status representation of 2-allele codominant system.

Now you see the penetrance relationship between genotypes and phenotypes, where penetrance is defined as P(phenotype | genotype) So, in row 1, you see the probability of phenotype *1 1* given genotype *1 1* is 1, and the probability of phenotype *1 1* given genotype *1 2* or *2 2* is 0. This is just a straight codominant locus. The last column is the affection status notation by which you would define each phenotype. The first *2* means presence of some phenotype for which penetrances are defined, while the second number indicates which liability class the appropriate penetrances are given in. If one has phenotype *2* in liability class *1*, it just means that at the locus in question, the penetrances in liability class one are the probabilities of the observed phenotype (whatever that may be) given each of the possible genotypes. So you would know that the person had probability 1 of having this phenotype if he had marker type *1 1*, and probability 0 of having this phenotype if he had genotype *1 2* or *2 2*. Thus, you know this person has genotype *1 1*.

However, you could also assign the phenotype *1* in liability class *1*, in which case the penetrances for the individual's phenotype would be one minus the penetrances given in the liability class 1. In other words a person with phenotype *1* in liability class 1 would have the given phenotype with probability 0 for genotype *1 1* , and with probability 1 for genotype *1 2* or *2 2*. So all you would know about this individual is that he is definitely not a *1 1*, but could be either *1 2* or *2 2*. This is one way of coding a dominance relationship. But for our purposes, we just want to code a codominance relationship, so we will give everyone phenotype *2*, with the appropriate liability class assigned to each phenotype. Now, following the scheme in the last column of table 10-1, please add a sixth column to EX9.PRE, corresponding to the marker type of each individual in affection status notation. Then save it as EX10.PRE, and process it with MAKEPED to make a file EX10.PED. Now, call up the PREPLINK program, read in EX9.DAT, and add a sixth locus, this time of affection status type. Then, make the screen, under see or modify locus 6, look like the following:

```
*********************
(a) NUMBER OF ALLELES          : 2
(b) NUMBER OF LIABILITY CLASSES     : 3
(c) PENETRANCES:
LIABILITY CLASS : 1
GENOTYPE 1 1 1.00000E+00
GENOTYPE 1 2 0.00000E+00
GENOTYPE 2 2 0.00000E+00
LIABILITY CLASS : 2
GENOTYPE 1 1 0.00000E+00
GENOTYPE 1 2 1.00000E+00
GENOTYPE 2 2 0.00000E+00
LIABILITY CLASS : 3
GENOTYPE 1 1 0.00000E+00
GENOTYPE 1 2 0.00000E+00
GENOTYPE 2 2 1.00000E+00
(d) GENE FREQUENCIES :
0.500000 0.500000
(e) EXIT
***************************
```

Then, save the new file as EX10.DAT. Invoke LCP to use MLINK to analyze loci 1 and 6, and compare this with an analysis of loci 5 and 2, in the manner described above. Then compare the results again. Of course, they should again be completely identical, or else you should recheck your input files, and reconsider the logic of your model.

10.3 CARRIER STATUS

There are frequently recessive diseases for which some carriers of the disease allele (heterozygotes) do show some mild phenotypic effect that distinguishes them from homozygous normal individuals. In such cases, it may be useful to incorporate this information in the linkage analysis. However, it is very important to point out that if the only reason for calling an individual an obligate carrier is based on genetic reasons, (i.e. because he has affected children), it is NOT in principle a good idea to tell the program this individual is a carrier. If this person is an obligate carrier, the programs will determine that themselves, although at some minor expense in computing time. Further, when you add in this kind of information, it makes errors more likely, and occasionally can be based on unwarranted assumptions. So, this is only a valuable thing to do if one has a PHENOTYPIC reason to call an individual a carrier. If all carriers are recognizable, then the disease is essentially codominant, with each genotype corresponding to a unique phenotype (+/+ = Unaffected, +/d = Carrier, d/d = Affected). Then, one can code the disease in the same manner as the codominant marker system in the previous example. So, let us return to the recessive disease pedigree we analyzed in chapter 5 (files EX3.*, with EX3A.ped). This time, we will assume that we know that the following unaffected people are carriers: *fgrandpa*, *mgrandma*, *father*, *mother*. Similarly, *fgrandma*, *mgrandpa*, *dau1*, *dau3*, and *son2* are phenotypically determined to be homozygous unaffecteds. Now, enter this new third locus at the end of each line of the datafile. You can use the same penetrance scheme as explained for codominant marker loci, assigning the *2* allele, for example, to be the disease causing allele. Please make the appropriate modifications to the pedigree and parameter files, and save them as EX11.*. Then analyze them as usual with MLINK and ILINK. The solutions are given in table 10-2.

θ	Log(Likelihood)	Lod Score
0.0	-15.418534	3.010294
0.1	-15.876108	2.552720
0.2	-16.387632	2.041196
0.3	-16.967550	1.461277
0.4	-17.637017	0.791811
0.5	-18.428828	0.000000

ILINK: $\hat{\theta}$ = 0.000   Z($\hat{\theta}$) = 3.010294

Table 10-2: Analysis results for EX11.PED; EX11.DAT

In this example, our lod score tripled from 0.98 to 3.01, due to the phenotypic information we used to distinguish carriers from homozygous normal individuals. Of course, if there were no linkage, one would expect that adding the additional information would likely make the lod scores significantly smaller. This example just shows how much more information one can get from a codominant locus versus a recessive locus in the same family. Now, every meiosis is phase known and informative, from *mother* and *father* to their children. These ten phase known non-recombinants give us a lod score of $\log_{10}[(1 - \theta)^{10}/(0.5)^{10}]$, which is maximized at $\theta = 0$, to give us $\log_{10}[2^{10}] = 10 \times \log_{10}[2] = 3.010$. When the disease is recessive, with no phenotypic means to discriminate carriers from homozygous normal individuals, there is much less phase and genotype information available, causing the observed drastic reduction in information.

10.4 DIAGNOSTIC UNCERTAINTY

Now, let us consider a situation in which you do not know whether or not someone is truly affected, but a clinician can assign a probability with which he believes the individual to be affected (BASED ON NON-GENETIC REASONS!). How can we use the LINKAGE programs to allow for such diagnostic uncertainty? Well, it is possible to model such diagnostic uncertainty in the affection status locus type by using liability classes to define penetrances for each such degree of uncertainty. Ott (1991) described a method whereby one can generate the penetrances for individuals who have a given uncertainty of diagnosis by forming a weighted average of the penetrances for unaffected and affected phenotypes. Essentially if one has probability ***p*** of being affected with a disease, then one could compute the penetrance given genotype 1/1 as *p*[P(affected | 1/1)] + (1 – *p*)[P(unaffected | 1/1)], and so on for the other genotypes. The justification for such an ad hoc approach is discussed in detail in Ott (1991).

Let us assume, for example, that you have a disease that you wish to analyze that is inherited as an autosomal dominant disorder. The problem is that you have a certain degree of uncertainty in the diagnosis. Let us assume that you have four different observed phenotypes at the disease locus as follows: 1) Definitely affected; 2) Definitely unaffected; 3) Affected with 80% certainty; 4) Unaffected with 80% certainty. Assuming the disease to be fully penetrant, and using the affection status locus type, devise a method of using all of this information in your analysis. Please enter the data from Figure 10-1, in which the values under each person in the drawing correspond to the age of the individual, his diagnostic class (as defined above), and his ABO blood type.

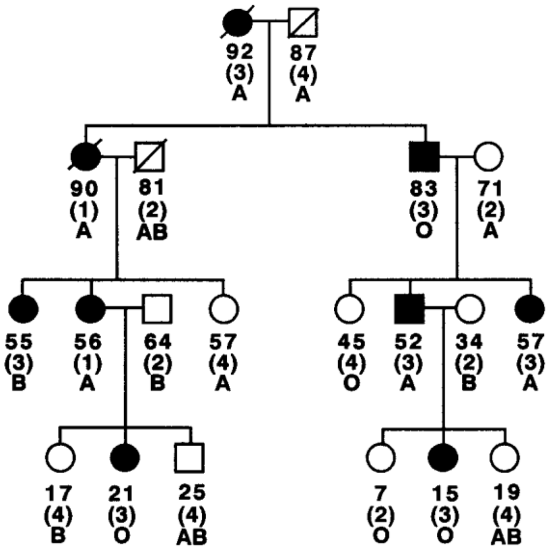


Figure 10–1. Diagnostic uncertainty pedigree—EX12.*

Please enter all the necessary information about this pedigree in LINKAGE format (files EX12.PED; EX12.DAT), with gene frequency for the disease allele of 0.1, and at ABO, use binary factors notation, and gene frequencies of 0.26 for the A allele, 0.06 for the B allele, and 0.68 for the O allele (these allele frequencies are somewhat different from those used in exercise 8). Then run MLINK and ILINK on this family in the manner you have been doing throughout the book. (You will further develop this problem in exercise 10.)

For individuals in diagnostic class 3, the penetrance is, for AA or Aa individuals, (0.8) × (1) + (0.2) × (0) = 0.8, and that for people with aa genotype is (0.8) × (0) + (0.2) × (1) = 0.2. In essence, we can code our four phenotypes with the following penetrance classes:

	AA	Aa	aa
1)	1	1	0
2)	0	0	1
3)	0.8	0.8	0.2
4)	0.2	0.2	0.8

In this simple case, the penetrances for phenotype 2 are equal to 1 minus the penetrance for phenotype 1, and the same relationship applies between phenotypes 3 and 4. Therefore, we need only two liability classes, corresponding to phenotypes 1 and 3. Then individuals in class 2 would be coded as *1* in the liability class for phenotype 1, and class 4 individuals would be coded as *1* in the liability class for phenotype 3. In essence we need the following two liability classes:

CLASS 1)	1	1	0
CLASS 2)	0.8	0.8	0.2

And, the codes for the four phenotypic classes are as follows: Class 1) = *2 1*, Class 2) = *1 1*, Class 3) = *2 2*, Class 4) = *1 2*, Unknown) = *0 1* or *0 2*, with no difference.

MLINK results are given in table 10-3.

θ	Log(Likelihood)	Lod Score
0.0	-17.663268	-0.208789
0.1	-17.357955	0.096524
0.2	-17.317604	0.136875
0.3	-17.347853	0.106626
0.4	-17.397023	0.057456
0.5	-17.454479	0.000000

ILINK: $\hat{\theta}$= 0.186 Z($\hat{\theta}$) = 0.137380

Table 10-3: Analysis results for EX12.PED; EX12.DAT

10.5 ABO BLOOD GROUP REVISITED

In chapter 8, we learned how to code the ABO blood group as a Binary Factors type of locus. Now that you have a basic idea about Binary Factors loci, we will discuss another important application of Affection Status loci. This is sometimes useful if there are complicated dominance relationships at your marker locus, as there are at the ABO blood group. All this means is that you must define the penetrances for each phenotype in separate "liability classes". The penetrances for the ABO blood group phenotypes are shown in table 10-4.

		Penetrances for Genotypes						Affection
Liab.	Pheno-							Status
Class	type	AA	AB	AO	BB	BO	OO	Coded As
1	A	1	0	1	0	0	0	2 1
2	B	0	0	0	1	1	0	2 2
3	AB	0	1	0	0	0	0	2 3
4	O	0	0	0	0	0	1	2 4

Table 10-4: Affection status representation of ABO blood group

In this example, you can see how the affection status locus type can be used to specify complicated dominance relationships.

In summary, in this chapter, we saw that the affection status locus type can be used to characterize any type of mendelian inheritance, and is not limited to disease traits. In fact, a more appropriate name for this locus type would be "dichotomous locus type". We earlier saw that the Allele Numbers locus type can handle a subset of all possible binary factors in a simpler notation. Similarly, in this chapter, we have seen that the binary factors locus type can code for a small subset of all possible modes of inheritance codable under an affection status locus type. Any phenotype:genotype equivalence that can be described with the allele numbers or binary factors notation can also be handled in an affection status. Plus, the affection status has the additional advantage of allowing for penetrance values other than 0 or 1, making it even more flexible than the other locus types. If one wanted to, one could do all linkage analyses with the LINKAGE programs without ever using any allele numbers or binary factors loci, but as we have seen, they provide convenient shorthand notations for simple fully penetrant genetic loci.

EXERCISE 10

Go back to the second part of exercise 8 (files USEREX8.*), and recode the ABO blood group as an affection status locus, in the manner described above. Then, perform the same analyses as you did in exercise 8 with ABO, replacing the binary factors representation of it with the new affection status representation. The results of the analysis must be identical, if you have made no errors. Also, recode the disease from that exercise as a recessive trait (In this case, the *2* phenotype represents “normal”!), that is, think of the "normal" trait as the trait of interest, which should be coded as “2” in the pedigree file. Finally, recode the codominant marker locus from this example as an affection status locus. Reanalyze the pedigree with these new affection status loci, and compare the results. They should be identical to the results obtained from the earlier allele numbers and binary factors representations.

Reconsider the example from this chapter's text (EX12.*), with the 80% diagnostic certainty. Please assume that that disease had constant 70% penetrance, instead of the full penetrance you had previously assumed. Analyze the pedigree with MLINK and ILINK.

Next, go back, and insert the 0.5% phenocopy rate in the analysis, along with the diagnostic uncertainty and reduced penetrance models. Reanalyze the pedigree with MLINK and ILINK.

Next, let us assume that the disease is also inherited with reduced penetrance following a straight line age-of-onset curve. In this case the penetrances for the susceptible genotypes range from 10% for those under 10 up to 90% for those over 50, while the penetrance curve for those with non-susceptible genotypes (phenocopy rate) also follows a straight line age-of-onset curve with penetrances ranging from 0.2% under age 20, rising to 1% for those over age 60. Please now define a liability class notation for this locus by dividing the population into six age classes as follows [0,9], [10,19], [20,34], [35,49], [50,59], [60,100], calculating penetrance values for the median age of each group (e.g [0,9] : use age 5; [20,34] : use age 27). Note that you will now need many liability classes. Since not all of them are used in this specific pedigree, you can simply recode the locus to just include the liability classes you will need for this analysis, based on the diagnostic class/age combinations appearing in this pedigree.

Let us assume now, that it is possible for some individuals to be either 1) type A or AB, and others can be 2) type B or AB, others can be 3) A or O, and still others can be either 4) B or O. In other words, the blood types are not uniquely determined, but some information was obtained from a partial test. In this pedigree, let us now assume that the 56 year old mother of three in the third generation has type 1) above, the probably affected 15 year old girl in the fourth generation has type 2) above, the 57 year old probably affected female on the extreme right of the drawing in the third generation has type 3) above, and the 17 year old probably unaffected female in the fourth generation on the extreme left of the drawing has type 4) above. Now recode your pedigree file, and reanalyze the family.

11 THE LIPED PROGRAM

LIPED (for LIkelihoods in PEDigrees) was the first generally available program for linkage analysis (Ott, 1974) [18]. It has changed little since it was extended to handle general pedigrees (Ott, 1976) except that age-of-onset functions were later incorporated. It contained one error relating to the likelihood of qualitative traits; fortunately, that bug was caught by Dr. Robert Elston soon after the program was distributed. The fact that is has been bug free ever since is probably the main reason why it was used for 20 years [19] and was by many people considered the gold standard against which other programs are compared. Below we will give you a general description of the LIPED program and provide an example of how to use it.

11.1 Characteristics of LIPED

LIPED is written in Fortran 77 and runs almost unmodified on most computers. Many researchers have adapted it to their computers and made various modifications to it. We support LIPED for Windows and Linux, compiled with GNU gfortran. It is distributed with various example input files and an extensive documentation.

Only two loci can be handled by LIPED at any one time, but the program is set up to carry out various two-point analyses in a single run. All preprocessing steps such as the ones carried out by the MAKEPED program are incorporated in a single program. Below, the main similarities and differences between LIPED and LINKAGE are pointed out.

As in the LINKAGE programs, the likelihood is calculated recursively by the use of the Elston-Stewart algorithm, but no iterative parameter estimation is possible as in the ILINK program. In both, LIPED and LINKAGE, program constants are set for maximum numbers of alleles, loops, etc., and new values of these constants require recompilation of the programs.

LIPED calculates lod scores for a sequence of recombination fractions or a sequence of points of male (θ_m) and female (θ_f) recombination fractions, which are displayed in a rectangular coordinate system with axes of θ_m and θ_f.

Five locus types are distinguished in LIPED:

1) Qualitative phenotypes with penetrances of 0 or 1

2) Qualitative phenotypes with any penetrances

3) Quantitative phenotypes following a normal distribution

4) Disease phenotypes (affected/unaffected) with age of onset following a lognormal distribution

5) Disease phenotypes with age of onset following a straight-line “penetrance curve.” This and the last locus type make it very easy to handle age dependent penetrance because the different cases of age of onset known or unknown are allowed for through their appropriate density or distribution function. In the LINKAGE programs, properly allowing for age dependent penetrance can require the setting up of many liability classes.

LIPED cannot directly handle the situation that parents in a pedigree both have parents in the same pedigree. Such a situation must be accommodated by doubling one of the parents much in the same way as one breaks a loop (see example below). This requirement is a consequence of the fact that only Elston-Stewart type peeling (going up through pedigrees) is incorporated in LIPED but not more general pedigree traversing algorithms.

For the analysis of a given two-locus problem, LIPED is generally somewhat slower than MLINK, particularly, of course, when upwards branching is present requiring the doubling of individuals in LIPED but not in MLINK.

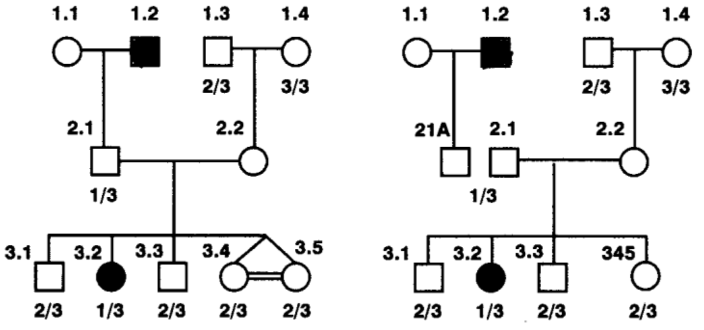


Figure 11–1. Pedigree with monozygotic twins—LIPED

11.2 An example: monozygotic twins

Several examples are provided in the LIPED documentation such that we do not give extensive details on how to use the program. Only one example is outlined here, which also demonstrates how monozygotic twins should be handled by LIPED. This example could be calculated in pretty much the same way by MLINK and is thus not specific to LIPED.

Figure 11-1 shows a small three-generation pedigree in which a dominant disease is segregating together with a marker locus with three alleles.

Penetrance is incomplete (90%) but no phenocopies occur. Individuals *3.4* and *3.5* are monozygotic twins, that is, they represent two phenotypic expressions of the same genotype. Therefore, for linkage analysis purposes, these two individuals must be represented as a single individual (here denoted by *345*). In addition, the penetrances for that single individual must be the squares of the penetrances appropriate for one of the two twins. For example, given the genotype D/D, the penetrance for affected is 0.90 and that for unaffected (phenotype of our two twins) is 0.10. Thus, the penetrance for the single individual representing the two monozygotic twins must be $(0.10)^2 = 0.01$. To allow for this different set of penetrances, a separate phenotype (liability class) is introduced, here denoted by the symbol *MT*. Incidentally, no such extra liability class is required in the MENDEL program [20] because it directly provides for the presence of monozygotic twins.

Whereas the left side of Figure 11-1 shows the original pedigree, the pedigree manipulated for input to LIPED is displayed on the right side (note that one of the parents must be doubled). The input file appropriate for this problem is shown below. Comments to the right of <= are optional and are given here only for better clarity. Details on how to set up an input file may be found in the online LIPED manual.

```
1 0000 0.0          MZ twin analysis
           0   m   0 0/0      <= unknown parent (here 0), male (here m), unknown phenotypes
 2 3                    <= number of alleles at loci 1, 2, ...
 3 6                    <= number of phenotypes
-1 0                    <= locus types
 2                      <= output option
 Dis   D   +  AF  NA  MT      <= locus name, allele symbols, phenot. symbols
   0.001   0.999              <= allele frequencies (D = recessive disease allele)
   D   D  .9  .1 .01          <= genotype and penetrances
   D   +  .9  .1 .01
   +   +   0   1   1
 S61   1   2   3 1/1 1/2 1/3 2/2 2/3 3/3  <= allele and phenotype symbols
      .2      .5      .3
   1   1   1
   1   2   0   1
   1   3   0   0   1
   2   2   0   0   0   1
   2   3   0   0   0   0   1
   3   3   0   0   0   0   0   1
  11   1Pedigree with monozygotic twins
 1.1   0   0   f  NA 0/0
 1.2   0   0   m  AF 0/0
 21A 1.1 1.2   m  NA 1/3
 1.3   0   0   m  NA 2/3
 1.4   0   0   f  NA 3/3
 2.2 1.3 1.4   f  NA 0/0
 2.1   0   0   m  NA 1/3
 3.1 2.1 2.2   m  NA 2/3
 3.2 2.1 2.2   f  AF 1/3
 3.3 2.1 2.2   m  NA 2/3
 345 2.1 2.2   f  MT 2/3
 2.1 21A                <= This line identifies the doubled individual
9000
```

The five 1's and 0's on the first input line indicate (1) the presence of a single marker locus, (2) the LIPED should print to screen and an output file called *liped.out*, (3) no special measures to prevent underflow, (4) autosomal inheritance, and (5) allele frequencies (as opposed to haplotype frequencies); 0.0 indicates absence of mutation. On the second input line, the four items define the symbols for “no parent” (here 0), male sex (here “m”), and unknown phenotypes at the two loci (here 0 and 0/0). Subsequent lines indicate numbers of alleles, numbers of phenotypes, and locus type for the different loci; and an output option (*2* specifies equal male and female recombination fractions with values as shown in the output below).

Next, a group of input lines is reserved for each locus. In each group, the first line specifies the locus symbol, allele symbols for use in the few lines immediately below, and phenotype symbols used in the pedigree data. The next line furnishes the allele frequencies in eight spaces each. After that, as many lines as there are genotypes are expected, each line defining a genotype and the associated penetrances for each phenotype (blank instead of a number is interpreted by FORTRAN as zero). Be aware that here the penetrances are the conditional probabilities with respect to the phenotypes listed on the first line in each group whereas in LINKAGE, penetrances refer to the “2” phenotype. Since the monozygotic twins are unaffected (phenotype MT), the penetrances for genotypes D/D and D/+ are small.

To beak a loop, or one of two inheritance lines ascending from two parents, a suitable individual must be doubled. The principle is the same as in the LINKAGE programs except that here it is the user who must carry out the doubling: One of the two doubles must be coded as an offspring and the other as a mate without parents. The two are later identified as representing the same individual.

On line 14 from the bottom, the first number (here *m* = 11, right justified in columns 1-4) specifies the number of family members to follow, and the second number (in columns 5-8) denotes the number of pairs of doubled individuals present. The comment (starting in column 9) following the two numbers is optional and will appear on the output. The next *m* lines describe the pedigree data: in each line one must have a unique individual id, parents' id's, sex, and phenotypes. After the m-th line, for each pair of doubled individuals, the id's of the two doubles are listed on one line. Finally, a code consisting of four digits directs the LIPED program to take further actions as indicated by input lines following that code. Here, a stop code (9000) is provided.

To run this input file, one invokes the program by typing LIPED. The output file, *liped.out*, looks as follows:

```
      Program LIPED  Version for PC   June 1995/Jan 2015   J. Ott
       Copyright (c) Jurg Ott 1988-2015.
 -----------------------------------------------------------------
  Program started on Sun Jan 25 20:46:28 2015

 PROBLEM  1   MZ twin analysis
 **********
 (autosomal linkage)

 Pedigree    1   Pedigree with monozygotic twins
 -------------
               11 individuals

 LOCUS  0    Dis     VS.     LOCUS  1    S61
 -------------------------------------------

 GENE FREQUENCIES FOR  0    Dis  0.0010  0.9990
 GENE FREQUENCIES FOR  1    S61  0.2000  0.5000  0.3000

  R MALE  R FEM.   LOG10[L(R)]   LOD-SCORE
  0.5000 0.5000       -10.40960       0.000
  0.4000 0.4000       -10.33163       0.078
  0.3000 0.3000       -10.15452       0.255
  0.2000 0.2000        -9.95545       0.454
  0.1000 0.1000        -9.76631       0.643
  0.0500 0.0500        -9.67790       0.732
  0.0010 0.0010        -9.59526       0.814
  0.0001 0.0001        -9.59378       0.816
```

As these results show, the maximum lod score (at $\hat{\theta} = 0$) is equal to 0.816. If the two twins are replaced by a single person with unmodified penetrance (here of 0.10), the resulting lod score is 0.779, which is, in this case, smaller than the correct lod score of 0.816. On the other hand, if the two twins are handled like regular twins (each with penetrance of 0.10), a maximum lod score of 1.039 is obtained, which is clearly higher than the correct lod score. Generally, treating monozygotic twins like fraternal twins tends to have the same effect as duplicating an offspring, that is, “inventing” data for one additional individual. In the long run, this does not introduce a bias in the recombination fraction estimate but it consistently inflates the lod score although, in any specific case, the lod score may also be decreased (when the twins represent recombinants).

EXERCISE 11

Modify the input to LIPED shown above to verify that falsely replacing the twins by a single individual with penetrance as for any unaffected phenotype results in a maximum lod score of 0.779. Why is the lod score now smaller?

Similarly, prepare the input such that the twins appear as fraternal sibs, leading to a lod score of 1.039. You may have to consult the LIPED manual if you experience problems. Interpret the result.

12 Solutions to Part I Exercises

EXERCISE 2

The pre-MAKEPED file, USEREX2.PRE should be as follows:

```
userex2 1 0 0 2 2 2 3
userex2 2 0 0 1 1 0 0
userex2 3 0 0 1 2 1 3
userex2 4 0 0 2 1 0 0
userex2 5 0 0 1 1 3 3
userex2 6 0 0 2 1 2 3
userex2 7 2 1 2 1 1 3
userex2 8 3 4 1 2 1 2
userex2 9 3 4 2 2 1 3
userex2 10 5 6 1 1 0 0
userex2 11 8 7 2 1 2 3
userex2 12 0 0 1 1 1 4
userex2 13 8 7 2 2 1 3
userex2 14 0 0 1 1 1 1
userex2 15 8 7 2 1 1 2
userex2 16 10 9 2 1 2 3
userex2 17 10 9 1 2 1 3
userex2 18 10 9 2 2 1 3
userex2 19 10 9 1 2 1 2
userex2 20 12 11 2 1 2 4
userex2 21 14 13 2 1 1 1
userex2 22 14 13 2 2 1 1
userex2 23 14 13 1 1 1 3
userex2 24 14 13 1 2 1 1
```

This is not the only individual identifying scheme that is acceptable. One could just as easily have named individuals in the pedigree with character ID's, but in this case, we decided to identify each individual by a number. Upon processing this file with MAKEPED, the following USEREX2.PED file should result:

```
1 1 0 0 7 0 0 2 0 2 2 3 Ped: userex2 Per: 1
1 2 0 0 7 0 0 1 1 1 0 0 Ped: userex2 Per: 2
1 3 0 0 8 0 0 1 0 2 1 3 Ped: userex2 Per: 3
1 4 0 0 8 0 0 2 0 1 0 0 Ped: userex2 Per: 4
1 5 0 0 10 0 0 1 0 1 3 3 Ped: userex2 Per: 5
1 6 0 0 10 0 0 2 0 1 2 3 Ped: userex2 Per: 6
1 7 2 1 11 0 0 2 0 1 1 3 Ped: userex2 Per: 7
1 8 3 4 11 9 9 1 0 2 1 2 Ped: userex2 Per: 8
1 9 3 4 16 0 0 2 0 2 1 3 Ped: userex2 Per: 9
1 10 5 6 16 0 0 1 0 1 0 0 Ped: userex2 Per: 10
1 11 8 7 20 13 13 2 0 1 2 3 Ped: userex2 Per: 11
1 12 0 0 20 0 0 1 0 1 1 4 Ped: userex2 Per: 12
1 13 8 7 21 15 15 2 0 2 1 3 Ped: userex2 Per: 13
1 14 0 0 21 0 0 1 0 1 1 1 Ped: userex2 Per: 14
1 15 8 7 0 0 0 2 0 1 1 2 Ped: userex2 Per: 15
1 16 10 9 0 17 17 2 0 1 2 3 Ped: userex2 Per: 16
1 17 10 9 0 18 18 1 0 2 1 3 Ped: userex2 Per: 17
1 18 10 9 0 19 19 2 0 2 1 3 Ped: userex2 Per: 18
1 19 10 9 0 0 0 1 0 2 1 2 Ped: userex2 Per: 19
1 20 12 11 0 0 0 2 0 1 2 4 Ped: userex2 Per: 20
1 21 14 13 0 22 22 2 0 1 1 1 Ped: userex2 Per: 21
1 22 14 13 0 23 23 2 0 2 1 1 Ped: userex2 Per: 22
1 23 14 13 0 24 24 1 0 1 1 3 Ped: userex2 Per: 23
1 24 14 13 0 0 0 1 0 2 1 1 Ped: userex2 Per: 24
```

In this file, the additional pointers were added, as explained in chapter 2. The parameter file, USEREX2.DAT should look like the following (after PREPLINK):

```
 2 0 0 5 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0 0.0 0.0 0 << MUT LOCUS, MUT RATE, HAPLOTYPE FREQUENCIES (IF 1)
 1 2
1 2 << AFFECTION, NO. OF ALLELES
 0.999990 0.000010 << GENE FREQUENCIES
 1 << NO. OF LIABILITY CLASSES
 0.0000 1.0000 1.0000 << PENETRANCES
3 3 << ALLELE NUMBERS, NO. OF ALLELES
 0.333330 0.333330 0.333330 << GENE FREQUENCIES
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.10000 << RECOMBINATION VALUES
 1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

Depending on the version of PREPLINK you are using, there may be some differences in the number of decimal places in the output files, but the format should be the same as that indicated above, for this example.

EXERCISE 3

The analytic solution in this pedigree is extremely complicated, and involves a large number of complicated formulas. For this reason, most pedigrees need to be analyzed with computer programs like LINKAGE. In this case, however, we can get some idea of what the recombination fraction should be from examining the most likely situation in this pedigree. The disease appears to be segregating with the *1* allele in this pedigree. If the disease actually were segregating with the *1* allele, there would be one obligate recombinant, in the unaffected female in the bottom generation with marker genotype *1 1*. Otherwise, there are 12 meioses informative for the disease in which there are no obligate recombination events. So, we would guess that the recombination fraction estimate should be somewhere around 1/13 = 0.077. Of course, the actual estimate will not be exactly equal to this, since the pedigree is phase unknown in the upper branches, and since there are a few untyped individuals.

When you call up the UNKNOWN program, it will give you the following message:

```
ERROR: Incompatibility detected in this family for locus 2
```

When a message like this comes out of the UNKNOWN program, it typically implies there is some error in the input files. First, you should look over the pedigree to make sure that there are no Mendelian inconsistencies (i.e. a *1/1* father and a *2/2* mother having a *2/2* child would be inconsistent with Mendelian inheritance). However, unless you made a typing error when entering the data, this should not be the case. The next thing to look at would be the description of the loci in the parameter file. It would seem that the parameter file created in exercise 2 is compatible with the description of the loci given in that exercise. However, upon closer scrutiny of the pedigree, it is clear that while the marker locus was entered as a three allele system, in the third generation, a man married into the pedigree with genotype *1/4*, which would be impossible at a three-allele locus. Hence, we need to recode this locus in PREPLINK, to allow for four alleles instead of three. For now, let us assume the four alleles are equally frequent, with gene frequency 0.25 for each allele. Later, in Part III, we will learn more appropriate ways of dealing with gene frequency estimation, but we will defer further discussion of it until then. Now, when you run the analysis you should get the results shown in Table 12-1. Notice how close the estimated recombination fraction of 0.079 was to our approximation of 1/13 = 0.077. Perhaps we should try and modify our DATAFILE.DAT file, such that we use a starting value of 0.079 for the recombination fraction (in the ILINK analysis), and see if we can further refine this estimate of θ. When you do this, your new estimate should be = 0.077, which is exactly 1/13, with $Z(\hat{\theta})$ = 1.783275. Running the ILINK analysis at each recombination fraction (as can be easily done in MLINK) would be accomplished by modifying the bottom 3 lines of your datafile. Initially they should look like this:

```
0.10000 << RECOMBINATION VALUES
0 << THIS LOCUS MAY HAVE ITERATED PARS
1
```

This causes the ILINK program to start at θ = 0.10000, and then iterate (since the last line contains a *1*) the recombination fraction until the MLE is found. In order to compute the likelihood at θ = 0.10000, you would simply need to modify the bottom line of the parameter file, by replacing the *1* (iterate recombination fraction) to a *0* (fix the recombination fraction). Then, you would simply specify the desired recombination fraction on the third line from the bottom of the file. Every time you wish to compute the lod score for a different θ, simply alter the third line from the bottom, and run the ILINK program again. The results will be exactly the same as those shown in Table 12-1).

θ	Log(Likelihood)	Lod Score
0.0	-infinity	-infinity
0.1	-26.896276	1.767576
0.2	-27.187291	1.476561
0.3	-27.650514	1.013338
0.4	-28.180207	0.483646
0.5	-28.663852	0.000000

ILINK: $\hat{\theta}$= 0.079. Z($\hat{\theta}$) = 1.783267

Table 12-1: Analysis results from USEREX2.* with MLINK and ILINK

EXERCISE 4

Using LCP, you should specify on the first screen the values as follows:

```
COMMAND file name [PEDIN.BAT] : PEDIN.BAT
LOG file name [FINAL.OUT] : FINAL.OUT
STREAM file name [STREAM.OUT] : STREAM.OUT
PEDIGREE file name [PEDIN.DAT] : USEREX2.PED
PARAMETER file name [DATAIN.DAT] : USEREX2.DAT
Secondary PEDIGREE file name [] :
Secondary PARAMETER file name [] :
```

Then, after selecting the *MLINK* program, with *Specific Evaluations*, and *No sex difference* options, you should complete the *MLINK - Lod Score Specification* screen as follows:

```
 Locus order [] : 1 2
 Recombination fractions [.1] : 0
 Recombination varied [1] : 1
 Increment varied [.1] : 0.1
 Stop value [.5] : 0.4
```

Then, go back and select the *ILINK* program, with *Specific order*, and *No sex difference* options, and set up the *ILINK - Locus Order Specifications* screen as follows:

```
 Locus order [] : 1 2
 Recombination fractions [.1] : 0.1
```

Then, hit <Page Down> to save this analysis, and hit <Ctrl-Z> to exit. Type *PEDIN* at the DOS prompt, and then examine the FINAL.OUT file. The results contained within this file should be identical to those in table 12-1.

EXERCISE 5

The results of the analysis of the USEREX2.PED pedigree under 75% penetrance (autosomal dominant) are presented in table 12-2. Note that in this analysis, the estimated recombination fraction was *0*. This is because it was more likely that the "obligate recombinant" individual in the pedigree was more likely to be non-penetrant than to be a recombinant, when both options were allowed for. The lod score also jumped by about 0.4 units. This is largely due to the decreased estimate of θ. With the same amount of data, the smaller the estimated recombination fraction is, the stronger the evidence for linkage (this makes sense, since the fewer recombinants observed in a fixed sample, the greater the likelihood that the two loci are linked).

θ	Log(Likelihood)	Lod Score
0.0	-27.197236	2.209401
0.1	-27.583813	1.822823
0.2	-28.029907	1.376729
0.3	-28.519204	0.887432
0.4	-29.006410	0.400227
0.5	-29.406636	0.000000

ILINK: $\hat{\theta}$= 0.001. Z($\hat{\theta}$) = 2.206964

Table 12-2: Analysis results from USEREX2.PED with autosomal dominant model with 75% penetrance

The analysis of this pedigree under the autosomal recessive model with 70% penetrance is presented in table 12-3. Note that now the lod scores are almost all uniformly negative. This makes sense, because now, all of the affected individuals are forced to be homozygous for the disease allele, and thus uninformative for linkage. The unaffecteds (who did not carry the disease allele under the dominant model) are now forced to be carriers of the disease allele in at least one copy, if they have affected children (who must also be homozygous for the disease allele). Thus, all of the information is coming from the individuals who were uninformative in the dominant analysis. If the disease truly were dominant, you would expect these (actual homozygous normal) individuals to transmit either allele to their affected children with equal probability, and therefore, the expected estimate of θ should be 0.5 in these situations.

θ	Log(Likelihood)	Lod Score
0.0	-64.367120	-10.326114
0.1	-54.633665	-0.592695
0.2	-54.195452	-0.154446
0.3	-54.031058	0.009948
0.4	-53.993094	0.047912
0.5	-54.041006	0.000000

ILINK: $\hat{\theta}$= 0.389; Z($\hat{\theta}$) = 0.0482982

Table 12-3: Analysis results from USEREX2.PED with autosomal recessive model with 70% penetrance

The results of the same analyses when only 30% penetrance is allowed for are presented in table 12-4 for the dominant case, and table 12-5 for the recessive case. From these results, you can see the general effect of lowering the penetrance is to flatten the lod score curve. The magnitudes (positive and negative) are reduced significantly when the penetrance is reduced, since the ability to discriminate genotypes among the unaffected individuals is severely limited. When there is 70% penetrance, the unaffecteds have penetrances 1 for +/+ genotypes, and 0.3 for +/D or D/D genotypes (assuming a dominant disease. The ratio of penetrances is thus 10/3, or about 3.33:1. However, when the penetrance is reduced to 30%, the ratio of penetrances is reduced to 10/7, or about 1.4:1. When you consider that a penetrance ratio of 1:1 implies an inability to distinguish between genotypes based on a given phenotype, then you can see that a ratio of 1.4:1 isn't much different from calling the unaffected individuals unknown in phenotype, and so you are losing information. Of course, when penetrance is complete, the penetrances for unaffecteds are 1 for +/+ genotypes, and 0 for *+/D* or *D/D* genotypes, for a penetrance ratio of 1:0 or infinity, since all unaffecteds must have had genotype +/+. This is the reason reducing the penetrance has such a major effect on the analysis, and it points out that the effect is due to increased uncertainty about the genotypes of unaffected individuals. Of course, for affected individuals, the penetrance ratio is always 0/f = 0, whatever the value of f (the penetrance) is. Hence, it does not alter the ability to discriminate the genotypes of affected individuals, unless you are allowing for phenocopies as you will see in chapter 9.

```
 θ          Log(Likelihood)          Lod Score
------------------------------------------------
0.0         -30.608103               1.805107
0.1         -30.984986               1.428225
0.2         -31.384362               1.028849
0.3         -31.790104               0.623107
0.4         -32.155375               0.257836
0.5         -32.413210               0.000000
```

ILINK: $\hat{\theta}$ = 0.001; Z($\hat{\theta}$) = 1.802592

Table 12-4: Analysis results from USEREX2.PED with autosomal dominant model with 30% penetrance

If the disease were actually recessive, and you analyzed it as a dominant disease, you would expect quite a different result. First of all, you would probably have to assume reduced penetrance for the disease, since parents (in a recessive disease) are typically unaffected. Since in a recessive disease, both parents contribute one disease allele to affected children, if you are looking at a linked marker, it would have to show linkage in both parental meioses. If you were to analyze the disease as if it were dominant, then you would only be considering segregation of the disease from one of the parents, with the other parent typically considered to be homozygous normal. Thus, you would be throwing away half of your truly informative meioses for the disease and marker. Still, you are retaining roughly half of the meioses, in which the marker is, of course, still cosegregating with the linked disease allele. Therefore, you would in general expect to still find positive lod scores, and reasonable estimates of the recombination fraction, although your power should be roughly chopped in half.

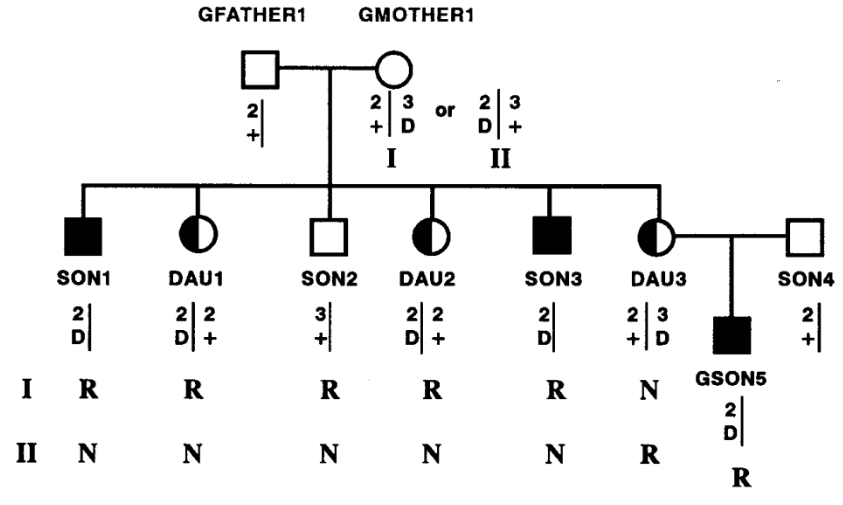


Figure 12–1. Linkage information from pedigree in Figure 6–3

EXERCISE 6

If one looks closely at this pedigree, it is clear that the information for linkage can be collapsed into what is shown in Figure 12-1. Clearly, *GMOTHER2* and *GFATHER2* contribute no information about linkage, since they have only one son, and did not transmit the disease allele. Hence, they can be left out. Similarly, *GSON1* and *GSON2* merely serve to identify *DAU1* as a carrier of the disease allele, but since she is a homozygote at the marker locus, they provide no other information about linkage. The same holds for *GSON3* and *GSON4*, who only tell us that *DAU2* is a carrier of the disease allele. On the other hand, *GSON5* does provide evidence for linkage. Clearly, the fact that he is affected tells us that his mother, *DAU3* is a carrier of the disease allele. Further, she is a heterozygote, who had to have receive both the disease and the *3* allele from *GMOTHER1*. Since she then transmitted the *2* allele with the disease allele to *GSON5*, there was an obligate recombination event. We also know the phase in each of the 6 children in this collapsed nuclear family, as indicated in Figure 12-1. However, we do not know the phase in *GMOTHER1*. Since each of the two phases have equal probability, we can compute the likelihood of this nuclear family from counting recombinants and non-recombinants under each phase as just $\frac{1}{2}[\theta^5(1-\theta)] + \frac{1}{2}[(1-\theta)^5\theta]$. Since there is one additional obligate recombinant from *DAU3* to *GSON5*, this whole likelihood should be multiplied by θ to get the likelihood for the entire pedigree. The likelihood ratio, therefore, is just $(\theta[\theta^5(1-\theta) + (1-\theta)^5\theta])/(0.5)^6$, and the lod score, accordingly is just $6 \times \log_{10}(2) + \log_{10}(\theta) + \log_{10}[\theta^5(1-\theta) + (1-\theta)^5\theta]$.

The results of the analysis of this example with MLINK and ILINK are shown in table 12-6. To confirm that your analytical result is compatible with the computer analysis, let us compute the numerical value of the lod score from the formula above at θ = 0.10. From the formula above, $Z(0.10) = 6 \times \log_{10}(2) + \log_{10}(0.10) + \log_{10}[(0.10)^5(0.90) + (0.90)^5(0.10)] = 1.80618 - 1 - 1.22872 = -0.42254$, which is exactly what was found with the MLINK program. To further confirm that the reduction of the data into the genotypic information contained in Figure 12-1, please enter the pedigree shown in that figure in LINKAGE pedigree and parameter files, coding the disease locus as an allele numbers locus with the + allele being coded as *1*, and the disease allele coded as *2*. Assume gene frequency of 0.01 for the *2* allele at this locus. Leave the second locus as a four allele allele numbers locus, and analyze this reduced pedigree (remember, the reason for using an allele numbers locus is to allow carriers [*1 2*] to be distinguished from homozygous normal individuals [*1 1*]). When you analyze this new pedigree, the lod scores should be identical to those in table 12-6, although the likelihoods will be much larger, since the pedigree is much smaller.

EXERCISE 7

If you were to run the MAKEPED program without declaring any loops, the LOOPS program would report that "*Loop(s) present in Family 1*". The LOOPS.OUT file should look like the following:

```
Program LOOPS version 1.17

Programmed by Xiaoli Xie July 1992
Design: Jurg Ott and Xiaoli Xie

Loop(s) in family 1!
Individuals in parentheses are married
The individuals and/or marriages involved are:
Loop 1: 3-(3,4)-7-8-(6,5)-5-(1,2)-3
Loop 2: 7-8-(5,6)-10-9-(3,4)-7
Loop 3: 8-(5,6)-10-(10,9)-13-12-(7,8)-8
Note: ID numbers are as assigned by MAKEPED
```

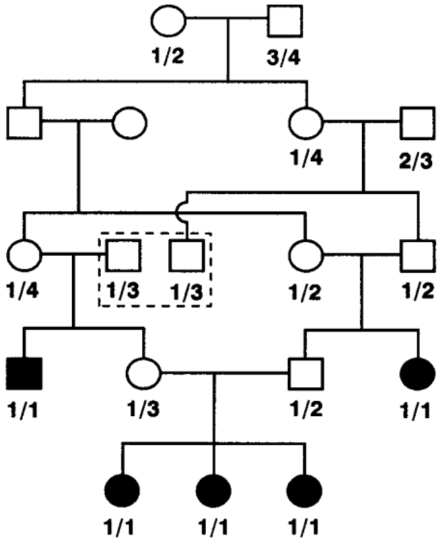


Figure 12–2. Pedigree from Figure 7–5 with one loop broken

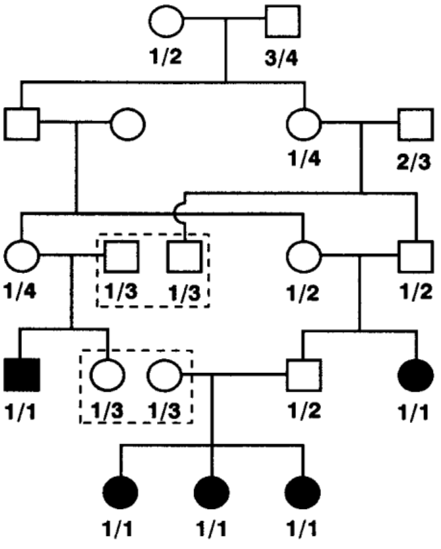


Figure 12–3. Pedigree from Figure 7–5 with two loops broken

Thus, the program reports having found three loops in this pedigree, the first being a consanguinity loop connecting the first set of first cousins who married (the first two people in the third generation); the second being a marriage loop with the two brothers in the third generation marrying two sisters; and the third being a consanguinity loop between the two cousins who married in the fourth generation of the pedigree, with the three affected sons. Of course, it is possible to list other loops, like the consanguinity between the second couple who married in generation 2, etc. But the program is saying that if you were to break the three loops it outlined, then there would be no further loops in the pedigree. To see this, let us go through the pedigree, and break one loop at a time. It is important to remember that you need to break the loops by doubling the individuals about whom there is the least amount of genotypic ambiguity (including phase).

Let us begin by breaking the first loop discovered by the LOOPS program by doubling the *1/3* unaffected male in the third generation. The resulting pedigree would look like that shown in Figure 12-2. Of course, there are still loops left in this pedigree, so we should next decide to break another loop in this pedigree. You could run the MAKEPED program, and break this one loop, and have the LOOPS program identify the remaining loops. In this case, after running the LOOPS program, the first loop detected would involve the two cousins who married in the fourth generation. By looking at the pedigree you can see that this loop is still present, even after breaking the first loop. Let us then proceed to break this loop at the unaffected mother of the three affected girls in the fourth generation (the *12*th individual in the pedigree), as shown in Figure 12-3. By looking at this figure, however, you can see that there is still at least one loop remaining, since there is still a first cousin marriage in the second generation. If you run MAKEPED at break the two loops as we did, the LOOPS program will again detect a loop, the loop being indicated in the LOOPS.OUT file as the first cousin marriage in the third generation. If we break this loop by doubling the unaffected male with *1/2* genotype in the third generation (he has the most genotypic information, since his parents are typed, and he has to be a carrier of the disease), we will be left with the pedigree in Figure 12-4. From examination of this pedigree, there don't seem to be any loops remaining, as can be more clearly seen if we redraw the pedigree, separating the doubled individuals in such a way that the absence of loops is more apparent, as shown in Figure 12-5. Clearly when the pedigree is redrawn like this, the absence of remaining loops is apparent. If you now run the MAKEPED program, and break the loops in the places we did, the LOOPS program should report that "*No loop detected in Family 1*", confirming what we have seen in figure 12-5.

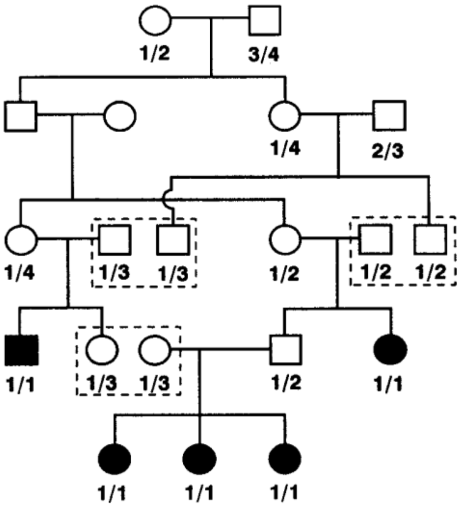


Figure 12–4. Pedigree from Figure 7–5 with three loops broken

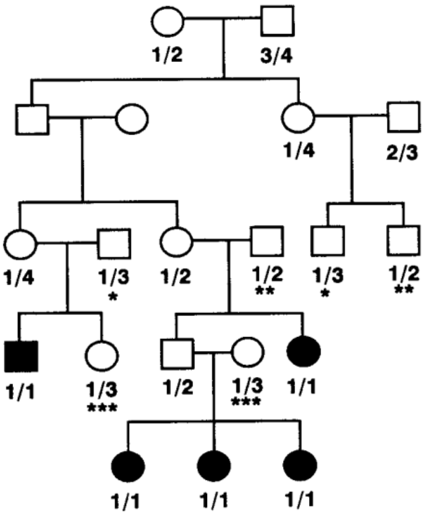


Figure 12–5. Summary of broken loops from pedigree from Figure 7–5

θ	Log(Likelihood)	Lod Score
0.0	-66.341955	-9.203737
0.1	-57.732587	-0.594369
0.2	-57.299879	-0.161660
0.3	-57.133608	0.004611
0.4	-57.091982	0.046237
0.5	-57.138219	0.000000

ILINK: $\hat{\theta}$= 0.394; Z($\hat{\theta}$) = 0.046381

Table 12-5 : Analysis results from USEREX2.PED with autosomal recessive model with 30% penetrance.

θ	Log(Likelihood)	Lod Score
0.0	-infinity	-infinity
0.05	-12.566381	-0.907257
0.10	-12.081665	-0.422540
0.15	-11.853246	-0.194121
0.20	-11.733741	-0.074617
0.25	-11.676429	-0.017305
0.30	-11.658802	0.000322
0.35	-11.665186	-0.006062
0.40	-11.679781	-0.020657
0.45	-11.683900	-0.024775
0.50	-11.659124	0.000000

ILINK: $\hat{\theta}$= 0.857; Z($\hat{\theta}$) = 0.559736

Table 12-6 : Analysis results from USEREX6.*

Of course, we did not select the only possible individuals at which to break the loops. They could just as easily have been broken elsewhere, but it is important to remember that the best strategy (computing time-wise) is to break all loops at the individuals with the minimum amount of genotypic ambiguity (including phase), as the computing time increases exponentially with the number of possible genotypes for the doubled individual.

θ	Log(Likelihood)	Lod Score
0.00	-13.838789	4.210090
0.05	-14.282988	3.765890
0.10	-14.748045	3.300833
0.15	-15.235013	2.813866
0.20	-15.743824	2.305054
0.25	-16.270839	1.778040
0.30	-16.802341	1.246537
0.35	-17.301193	0.747686
0.40	-17.699404	0.349474
0.45	-17.943024	0.105854
0.50	-18.048878	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 3.125729

Table 12-7: Analysis results from USEREX7.*

The results of the linkage analysis with this pedigree are shown in table 12-7. In this one small pedigree, because of the intense level of consanguinity, there is enough evidence for linkage to have a significant test result, with Z($\hat{\theta}$) > 3. Since we have a significant test result for linkage in this case, for the first time, it is meaningful to construct a 3-unit-of-lod-score support interval around this maximum, to give us some idea of the accuracy of our estimate of $\hat{\theta}$= 0. In this case, our support interval would cover all values of θ with Z(θ) ∊ [1.21, 4.21], which in this example would mean our support interval for θ would extend throughout the interval [0, 0.30). So, while our test of linkage is significant, we have little power in this small dataset to accurately estimate the recombination fraction. Obviously, in a real life situation, the solution would be to collect more families, or to type more markers, and do a multipoint analysis, which will be discussed in part II.

EXERCISE 8

The results from the binary factors representations of all markers in all previous user exercises should be identical to the results obtained originally when the allele numbers representations of those loci were used, so they will not be repeated again here.

θ	Log(Likelihood)	Lod Score
0.00	-25.714063	3.499118
0.10	-26.276072	2.937108
0.20	-26.890589	2.322592
0.30	-27.569724	1.643457
0.40	-28.332891	0.880289
0.50	-29.213181	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 3.495397

Table 12-8: Analysis of Disease vs. ABO in USEREX8.*

For the pedigree USEREX8.PED, with the four-allele marker and the ABO blood group markers, you were supposed to do linkage analyses between the two markers, and between ABO and the disease. If you remember, the analysis of disease vs. marker 1 was done in exercise 3, with the results shown in table 12-1, with $\hat{\theta}$= 0.077 (Z($\hat{\theta}$) = 1.78), after further refinement of the estimate in the table of 0.079. The results of disease vs. ABO are presented in table 12-8, and the results of the analysis of ABO vs marker 1 are given in table 12-9. The recombination fraction between the disease and marker 1 was estimated at 0.077, while the recombination fraction between the disease and ABO was estimated to be 0.

θ	Log(Likelihood)	Lod Score
0.00	-infinity	-infinity
0.10	-30.341911	0.310354
0.20	-29.883658	0.768607
0.30	-29.974499	0.677766
0.40	-30.289189	0.363077
0.50	-30.652265	0.000000

ILINK: $\hat{\theta}$= 0.220; Z($\hat{\theta}$) = 0.778700

Table 12-9 : Analysis of Marker 1 vs. ABO in USEREX8.*

However, the recombination fraction between marker 1 and ABO was estimated to be 0.220. How can this be consistent, you ask? Well, if you examine the pedigree closely, it is clear that there are a large number of meioses in which the disease is not informative, but the ABO and marker 1 are informative. In these meioses, there appear to be a large number of obligate recombinants between ABO and the marker. If you remember, when the penetrance of the disease was lowered to 75%, the recombination fraction estimate between marker 1 and the disease was 0. Please analyze the disease versus ABO, assuming 75% penetrance for the disease. The results of this analysis are presented in table 12-10. Now, we have a situation in which both markers show 0% recombination with the disease in this pedigree, and yet the two markers show about 22% recombination between themselves. This causes even more confusion. Consider the fate of the unaffected child in the last generation with marker type *1 1*, and ABO type *O*. When the analysis is done with marker 1, the most likely scenario has this child carrying the disease allele, but being non-penetrant.. Yet, when the analysis is done with the ABO blood group, this child is most likely not carrying the disease allele, since there is estimated to be 0% recombination, and the disease allele is segregating with the *A* allele, yet this child received the *O* allele from his mother. This apparent discrepancy can only be dealt with by doing a multipoint analysis, which will be discussed in Part II.

Θ	Log(Likelihood)	Lod Score
0.00	-26.944701	3.011264
0.10	-27.446887	2.509077
0.20	-27.990525	1.965439
0.30	-28.583064	1.372901
0.40	-29.234883	0.721082
0.50	-29.955965	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 3.007926

Table 12-10: Analysis of Disease vs. ABO in USEREX8.* with 75% penetrance for the autosomal dominant disease

EXERCISE 9

The formula for the penetrance for each age class would be obtained as follows: Clearly for individuals under age 10, the penetrance would be simply 0.1, and for individuals over age 60, the penetrance would be 0.9. For individuals in the middle, we would need to determine the formula for the line connecting the points (10,0.1) and (60,0.9), which is simply (age-10)[0.8/50] + 0.10. If we then divide our set into age classes, the appropriate penetrances are given in table 12-11. After making the appropriate modifications to the pedigree and parameter files, you should get the results shown in table 12-12 from the analysis of disease vs marker 1, and those in table 12-13 from the analysis of disease vs ABO.

		Penetrances		
Liability Class	Age range	+/+	+/D	D/D
1	< 10	0	0.10	0.10
2	10-19	0	0.18	0.18
3	20-29	0	0.34	0.34
4	30-39	0	0.50	0.50
5	40-49	0	0.66	0.66

6	50-59	0	0.82	0.82
7	≥ 60	0	0.90	0.90

Table 12-11 : Penetrances for individuals in different age classes according to the age of onset function defined in exercise 9.

θ	Log(Likelihood)	Lod Score
0.0	-27.591976	1.982445
0.1	-27.995795	1.578626
0.2	-28.426063	1.148357
0.3	-28.869965	0.704456
0.4	-29.280893	0.293528
0.5	-29.574421	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 1.979753

Table 12-12 : Results of analysis of disease vs. marker 1 in files USEREX8.*

θ	Log(Likelihood)	Lod Score
0.0	-27.912368	2.211381
0.1	-28.308307	1.815443
0.2	-28.727879	1.395871
0.3	-29.171507	0.952243
0.4	-29.638251	0.485498
0.5	-30.123749	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 2.208726

Table 12-13 : Result of analysis of disease vs. ABO in files USEREX8.*

EXERCISE 10

When you have a disease with 80% diagnostic certainty and 70% penetrance, to determine the penetrance values, you simply apply the formula given in chapter 10, p × P(affected │ genotype) + (1 – p) × P(unaffected│genotype), where p is the probability with which you believe the individual to be affected (the diagnostic certainty). In this case, then, for people who are affected with 80% certainty, the penetrances are for people with genotype AA or Aa (notation from chapter 10), (0.8)(0.7) + (0.2)(0.3) = 0.62, and for people with genotype aa, (0.8)(0) + (0.2)(1) = 0.2. Thus the penetrance classes with the disease being 70% penetrant are as follows:

	AA	Aa	aa
1)	0.7	0.7	0
2)	0.62	0.62	0.20

Thus, in this case, there is very little genotype discrimination possible for the affected with 80% certainty liability class, since the penetrance ratio is only 0.62/0.20 = 3.1:1, which is much smaller than 0.7/0 = ∞. The results of the analysis of the pedigree under this model are given in table 12-14.

θ	Log(Likelihood)	Lod Score
0.0	-17.296418	0.437915
0.1	-17.399164	0.335169
0.2	-17.505815	0.228518
0.3	-17.600109	0.134224
0.4	-17.674374	0.059959
0.5	-17.734333	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 0.437261

Table 12-14: Result of analysis of pedigree with diagnostic certainty of 80%, and penetrance of 70%

Inserting the phenocopy penetrance of 0.005 is relatively straightforward. We simply go back to the formula above, and plug in the new values, so the new penetrance for AA and Aa individuals with 80% diagnostic uncertainty is now 0.8 × 0.7 + 0.2 × 0.3 = 0.62 (the same as above), and the penetrance for aa individuals is now 0.8 × 0.005 + 0.2 × 0.995 = 0.203. Using this information now defines our liability classes as follows:

	AA	Aa	aa
1)	0.7	0.7	0.005
2)	0.62	0.62	0.203

The results of the analysis under this model are given in table 12-15.

θ	Log(Likelihood)	Lod Score
0.0	-17.306317	0.431483
0.1	-17.408201	0.329599
0.2	-17.513479	0.224322
0.3	-17.606208	0.131592
0.4	-17.679049	0.058751
0.5	-17.737800	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 0.430832

Table 12-15: Result of analysis of pedigree with diagnostic uncertainty of 80%, penetrance of 70%, and penetrance for homozygous normal individuals of 0.5%

When we allow for the age-dependant penetrance, we can compute the penetrances as (age – 10)[0.8/40] + 0.10 for genotypes AA and Aa and for individuals between age 10 and 50. For genotype aa, the penetrance can be computed as (age – 20)[0.008/40] + 0.002 for individuals between age 20 and 60. Since we are only provided with information about current age, we are only able to use the distribution functions for the age-dependent penetrance calculations. The penetrances are given in table 12-16 for the age classes specified in the exercise. In this pedigree, however, three of these twelve possible age/diagnosis combinations are never used, 100% diagnostic certainty in classes (10 – 19) and (35 – 49), and 80% diagnostic certainty in class (< 10). In order to make the program run more efficiently, only allow for the necessary nine liability classes in your analysis. The results of your analysis should match up with those in table 12-17.

	Penetrances					
	Diagnostic Class 1			Diagnostic Class 3		
Age Class	AA	Aa	aa	AA	Aa	aa
< 10	0.1	0.1	0.002	0.26	0.26	0.201
10-19	0.2	0.2	0.002	0.32	0.32	0.201
20-34	0.44	0.44	0.0034	0.464	0.464	0.202
35-49	0.74	0.74	0.0064	0.644	0.644	0.204
50-59	0.90	0.90	0.009	0.74	0.74	0.205
≥ 60	0.90	0.90	0.01	0.74	0.74	0.206

Table 12-16: Age-dependent penetrance distribution with and without diagnostic uncertainty for exercise 10

θ	Log(Likelihood)	Lod Score
0.0	-16.576215	0.656951
0.1	-16.758859	0.474307
0.2	-16.926905	0.306261
0.3	-17.065199	0.167967

0.4	-17.165189	0.067976
0.5	-17.233166	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 0.655672

Table 12-17: Results of analysis using the age dependent scheme outlined in table 12-16

To allow for uncertainty of phenotype at the ABO blood group locus, we will first have to recode the ABO blood group phenotypes as an affection status locus. Then, we can allow for the ambiguity in genotype as shown in table 12-18. As you can see in this table, the phenotype *A or AB* is exactly complementary to the phenotype *B or O*, in terms of penetrance, so you could get away with one additional liability class to code for both of these options as shown in table 12-18. The same relation applies to *B or AB* and *A or O*. To make things most efficient, please use the minimum of six liability classes, instead of eight, as outlined in the table. The results of your analysis should match those in table 12-19.

	Penetrances						
Phenotype	AA	AB	AO	BB	BO	OO	Coded as:
A	1	0	1	0	0	0	*2 1*
B	0	0	0	1	1	0	*2 2*
AB	0	1	0	0	0	0	*2 3*
O	0	0	0	0	0	1	*2 4*
A or AB	1	1	1	0	0	0	*2 5*
B or AB	0	1	0	1	1	0	*2 6*
A or O	1	0	1	0	0	1	*1 6* or *2 7*
B or O	0	0	0	1	1	1	*1 5* or *2 8*

Table 12-18 : Penetrances for ABO Blood Group, including uncertain phenotype allocations.

θ	Log(Likelihood)	Lod Score
0.0	-15.742181	0.587896
0.1	-15.910975	0.419103
0.2	-16.065113	0.264965
0.3	-16.189603	0.140475
0.4	-16.275887	0.054190
0.5	-16.330077	0.000000

ILINK: $\hat{\theta}$= 0.001; Z($\hat{\theta}$) = 0.586712

Table 12-19 : Results of analysis with age dependent penetrance scheme outlined in table 12-16 and ABO Blood group uncertainty of phenotype scheme outlined in table 12-18.

EXERCISE 11

For the handling of monozygotic twins in a linkage analysis, a recommendation often given is that the two twins should be replaced by a single individual since the two individuals represent two identical copies of the same genetic material. This works fine with full penetrance because $1^2 = 1$. With incomplete penetrance, however, treating the two twins as a single individual with penetrance as for any individual with the same phenotype represents an error possibly resulting in a change of lod score. The direction of the lod score change depends on the phenotype and whether that phenotype is indicative of a recombinant or nonrecombinant. To represent the twin pair as a single individual with penetrance as for any unaffected sibling, in the input line for individual 345 (third line from the bottom), we simply replace MT by NA. This changes the penetrance from 0.01 to 0.10. Consequently, the penetrance ratio for genetic versus nongenetic cases changes from 0.01/1 = 0.01 to 0.10, that is, it is now (falsely) closer to 1 such that the phenotype of the single twin has less weight than the phenotypes of the two twins jointly. An unaffected individual with marker type 2/3 in the given sibship appears to be a nonrecombinant. Because nonrecombinants increase the lod score and recombinants decrease it, in this case bringing the penetrance ratio closer to 1 (giving the phenotype less weight) decreases the lod score.

To make the monozygotic twins appear as fraternal sibs, we replace the single line for the 345 individual by two lines corresponding to two offspring, 3.4 and 3.5. Each of these two individuals has phenotype NA. As outlined above, this deviation from the correct analysis amounts to adding nonexistent linkage information. Because the added information is in the direction of a nonrecombination, the resulting lod score turns out to be too high.

Part II Multipoint Linkage Analysis with the LINKAGE package

13 Gene Mapping in CEPH families

In this chapter, you will be introduced to the use of a specialized set of linkage analysis programs designed for use in 3-generational CEPH-type pedigrees. When trying to construct a genetic map of a given chromosomal region, one should always type his new markers in the CEPH pedigrees, and do the analysis with these programs, for reasons to be outlined in this chapter.

Since this book was originally written, marker mapping has progressed greatly and detailed maps have been published [21], so the material in chapter 13 is rather outdated.

13.1 WHAT IS CEPH?

CEPH (Centre d'Etude du Polymorphisme Humain) is a center for genetic studies in Paris. Researchers at CEPH and at the University of Utah in Salt Lake City have assembled a large homogeneous panel of families for use in making maps of genetic markers. These families all consist of nuclear pedigrees with many offspring, and in most cases, the grandparents. This basic three-generational pedigree structure has come to be known as "CEPH-type" pedigrees, and is illustrated in Figure 13-1. The standard family set comprises forty families, with an extended family set of 64 families (Dausset et al, 1990).

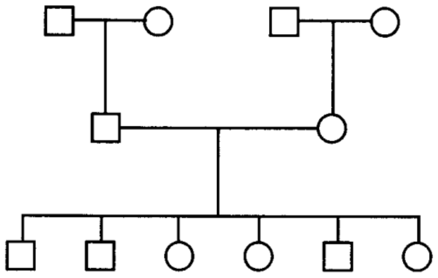


Figure 13-1. Example of CEPH-pedigree structure

Blood from each member of this panel is stored at CEPH, and can be made available to researchers around the world when they wish to type a new marker against the panel. In this way, each time a new marker is generated, it will be typed throughout this panel, adding to the CEPH database. This database, consisting of all the markers typed in the panel, can then be accessed by linkage analysts, so that they can try to generate good genetic maps, and determine the map locations of all newly generated markers based on the information on other markers in the database. Whenever someone generates a new marker for use in a genetic analysis, he should always try and map its location using the CEPH panel, not the disease pedigrees he is working with.

13.2 WHY USE THE CEPH PANEL?

Investigators often ask why they should take the time, expense, and trouble to type a new marker throughout this dataset. Well, to generate fine marker maps requires enormous datasets, and the CEPH panel has enough markers typed through it, such that not only is it a large dataset, but there are usually lots of markers nearby that have already been well mapped. This makes the process of mapping a new marker much more efficient, simpler, and more accurate. Furthermore, it is not clear how the presence of various disease-inducing mutations may affect recombination rates in the immediate region surrounding them. For example, if the mutation causing the disease of interest is a chromosomal inversion, the frequency of recombination may be reduced, due to lack of homology. Similarly, if a large deletion exists, then it would appear to make more distantly spaced markers appear closer together. Other mutations could induce hot-spots of recombination, etc. So, for these regions, one would like to do all marker mapping in the same genetically-healthy pedigree set. One should, however, be cautioned that there have been rather large error rates in the marker typings seen in the CEPH pedigrees in the past (Brzustowicz et al, 1993), so if you want to generate a very high resolution map, it may be advisable to check the marker typings at neighboring loci, especially if you observe double recombinants over a very short region.

13.3 CLINKAGE

There is a set of specialized linkage analysis programs (CMLINK, CILINK, and so on; these special programs may not be included in your standard LINKAGE package) written especially to analyze the CEPH pedigrees. These programs make use of very efficient computational algorithms, specific to the analysis of pedigrees with the "CEPH structure". As was mentioned above, CEPH-type pedigrees consist of a nuclear family with up to four grandparents.

In the normal LINKAGE programs, there are very strict limitations about the number of loci, and number of alleles per locus that can be analyzed, due to memory constraints. However, the CLINKAGE programs have many fewer of these constraints, and as such can analyze many loci at a time, with many alleles at each locus. Further, these analyses are incredibly rapid relative to very small analyses with the more general LINKAGE programs. There are severe restrictions, however, in the applicability of these programs. First of all, they can only be used to analyze codominant marker loci of the allele numbers or binary factors type, so no trait loci or quantitative variables can be used. Secondly, the only parameters one can vary are the recombination fractions, and female to male map distance ratios; one cannot use these programs to estimate linkage disequilibrium, gene frequencies, or interference. So, while these programs are incredibly fast, and efficient for generating genetic maps of codominant loci, and ordering a set of loci relative to one another, they are limited to these applications. Of course, the pedigree structure limitations as well make these programs useful primarily just for analyzing large sets of loci typed against the CEPH panel. We will be investigating each of these programs, and doing some sample exercises with them in the next few chapters.

13.4 General map construction strategies

To order a number *n* of marker loci and estimate the lengths of the *n-1* intervals between adjacent markers, one would ideally apply the maximum likelihood strategy. This would amount to computing the likelihood for each possible order, and that order with the highest associated likelihood would then be the best estimated order of loci. For numbers of loci as small as six or seven, depending on the amount of data, this approach is often feasible. However, to order a large number of loci, the sheer number of possible orders (n!/2) and the length of time it takes to calculate pedigree likelihoods prohibit considering all orders, and alternative methods of map construction must be employed.

One approach might be to evaluate all orders but to calculate a simple, approximate measure for the plausibility of the data under each order. Several such measures have been proposed, for example, the sum of adjacent recombination fractions (SARF, the smaller the better), which is obtained by adding the recombination fraction estimates in each interval. Another measure is the sum of two-point lod scores for all pairs of adjacent loci (SAL, large values are good). An overview of these methods may be found in Weeks (1991). In our limited experience with these methods, they do not work very well with real data, and as a matter of fact, hardly anyone in the gene mapping field uses them.

The most popular strategy in current use is to calculate exact (or almost exact) likelihoods for only a limited number of markers at a time. Therefore, rules are required by which one selects an initial set of markers for building a "trial map" and builds upon it by adding new markers. The MAPMAKER and CRI-MAP programs have built-in rules for iterative map construction. Of course, the final map resulting from such a procedure is not guaranteed to be the overall best map (by the rigorous likelihood criterion), so one needs various ways of corroborating the proposed best map. Also, MAPMAKER and CRI-MAP make some approximations in the likelihood calculations although that does not seem to be crucial. For references to these programs, see Appendix B. Below, we sketch some basic steps in map construction. For more detailed information, you may want to consult some of the recent chromosome map reports (e.g., Mills et al., 1992; Petrukhin et al., 1993).

If one does not want to rely on automatic map building by one of these programs, another strategy is to use only the most informative markers to construct a "skeletal" map whose order can be established unequivocally. Informativeness of a marker in a particular set of data my be assessed by computing the lod score at θ = 0 of this marker against itself; those markers with the highest lod scores are most informative (Mills et al., 1992). Additional markers are then "dropped" into each interval of this map, where each time all map distances are recalculated. If likelihood calculation with all markers is not possible, than only a subset of markers in the vicinity of the position of the new markers are used.

The final five to ten best map orders should then be scrutinized carefully. A common strategy to corroborate a given map is to invert pairs of adjacent loci one at a time (or evaluate all orders for any triple of adjacent loci), each time recalculating the location score. Also, map building is usually carried out assuming equal male and female recombination rates. This restriction must be relaxed and map distances re-estimated allowing for sex dependent recombination rates. Because the programs used in the map building phase generally calculate approximate likelihoods, it is recommended that the location scores for the best maps be recalculated by the CILINK program as it carries out exact likelihood calculations. For each of the best maps, three assumptions on the ratio of female-to-male map distance ratio, $R = x_f/x_m$, should be evaluated: 1) R = 1 throughout the map, 2) R ≠ 1 but constant in each interval, and 3) R ≠ 1 and possibly different in each interval. That hypothesis with a significant maximum location score is retained, where significance is assessed by chi-square tests as outlined in chapter 18.

An important component of map construction is error detection. Apart from retyping everyone and applying specialized statistical methods to pinpoint pedigree errors (e.g., Ott 1993; Brzustowicz et al. 1993; Haines, 1992), marker errors are typically recognized by the occurrence of double crossovers over a short map distance. Since this strategy requires the assumption of a locus order, the following two steps are generally repeated as often as necessary: 1) Build a map using one of the techniques incorporated in the MAPMAKER or CRI-MAP programs, and 2) follow up on individuals in whom multiple crossovers occur within, say, 30 cM (the CHROMPICS option of CRI-MAP is most useful at this stage). Each occurrence of such multiple crossovers is followed up, perhaps by retyping individuals in the lab, to possibly correct any errors.

14 The locus ordering problem: CILINK

In this chapter, you will learn how to use CILINK of the CLINKAGE package in locus ordering problems. Locus ordering problems are of critical importance in any multipoint linkage analysis. One requires an accurate map, for a multipoint linkage analysis of a disease to be meaningful. For this reason, the locus ordering problem is one of the most significant and troublesome topics in all of linkage analysis.

14.1 HOW DOES ONE ORDER A SET OF LOCI?

One of the most crucial problems in linkage analysis is to order sets of closely linked markers. Once someone achieves a positive test of linkage (i.e. Z > 3), the next step is to try and find the location of this linked gene. For example, we may have a known map of markers, and find that a new one is linked to this set of markers. Then, we would need to find out exactly where along this map the new marker falls. Another possible situation would be that you know that a set of three genes are linked to each other, but you have no idea of their relative orientation. One would then need to order the three loci in question. So, the question remains, how can we figure out the order of a series of loci from pedigree analysis?

Let us consider the case of three loci known to be linked to one another. Let us further assume that for each meiosis in our sample, we can determine whether alleles at any pair of the two loci were cosegregating or not. In other words, we could determine for any meiosis whether or not a recombination event occurred between any pair of the loci. Consider that we have loci A, B, and C. There are then four possible observed meioses, as indicated in table 14-1.

	Interval		
Meiosis	AB	AC	BC
Type I	R	R	N
Type II	R	N	R
Type III	N	R	R
Type IV	N	N	N

Table 14-1: Four possible meiotic three-locus recombination events, irrespective of locus order.

Other combinations are not possible, since what happens between markers AB, and AC uniquely determines what happened between B and C, irrespective of true locus order. Now, we can reformulate the four meiosis types indicated above into a 2 × 2 table, for any particular locus order, as in Table 14-2.

	Interval 1	
Interval 2	R	NR
R	W	X
NR	Y	Z

Table 14-2: 2x2 table representation of the four possible three-locus recombination events.

If one were to consider the specific locus orders possible for this experiment, you would have ABC, ACB, and BAC. Then, the corresponding 2 × 2 tables would be as shown in table 14-3 (where roman numerals refer to meiosis type).

	Interval 1					
	Order ABC		Order ACB		Order BAC	
Interval 2	R	NR	R	NR	R	NR
R	II	III	III	II	I	III
NR	I	IV	I	IV	II	IV

Table 14-3: 2x2 table representations under each possible locus order.

Clearly, one way to choose which order is best, is to pick the order which would require the fewest number of double recombinants, since these are necessarily rare events, with frequency $\theta_1\theta_2$, or less if there is interference. In this case, if we had collected 100 informative meioses, and found 90 Type IV meioses, 6 Type III meioses, 3 Type II meioses, and 1 Type I meiosis, then our tables are shown in table 14-4.

	Interval 1					
	Order ABC		Order ACB		Order BAC	
Interval 2	R	NR	R	NR	R	NR
R	3	6	6	3	1	6
NR	1	90	1	90	3	90

Table 14-4: Sample data placed in the three possible 2x2 tables

By inspection locus order BAC seems to be the best order, as it has the minimum number of double recombinants. Under the assumption of no interference, which is routinely made by the LINKAGE programs, the recombination fractions should be independent, so the rows and columns of the 2 × 2 tables in table 14-3 should be as close as possible to independent. One measure of the deviation from such independence is a simple chi-square test of independence on each of these tables. If you use the Linkage Utility Program, CONTING, you can calculate these chi-square values. Further information about the CONTING program will be given in Part III. For the data in table 14-4, the corresponding chi-square values are: ABC – 22.16, ACB – 54.09, BAC – 2.07 (Using a Yates correction, the corresponding values of χ^2 would be 14.56, 44.48, and 0.19 respectively). This seems to indicate that order BAC gives a substantially better fit to our model than any other locus order, however, the chi-square approximation is not really valid because the expected number of double recombinants in the upper left cell is less than 1.

There is a very good book chapter about different methods for ordering loci by Dan Weeks (Weeks, 1991), which provides a comprehensive overview of the theory and practice of a number of approaches. The user is referred to this source for a detailed analysis of this very important problem in human genetics. For now, we will deal with what can be done with the LINKAGE programs as far as locus ordering is concerned. Typically, the analyst will be unable to perform the analysis described above, since in real pedigree data, one never has all meioses informative and phase known for all pairs of markers. To get around this problem, the LINKAGE programs can be used. The ILINK program can be used to maximize the likelihood under each possible order, by estimating each intermarker recombination fraction jointly. Thus, all the pedigree information is taken advantage of, including the partially informative meioses. As was mentioned in the previous chapter, however, the general LINKAGE programs are very slow, and quickly run out of memory as the number of loci, and number of alleles per locus increases. As mentioned in chapter 13, there is a special version of LINKAGE, called CLINKAGE that is optimized for likelihood calculations in CEPH type pedigrees. Using these programs, one can rapidly compute likelihoods for any given marker order, and any reasonable number of markers.

14.2 CILINK

For the remainder of this chapter we will concentrate on the practical usage of the CILINK program, which is the CEPH-pedigree-specific version of the ILINK program. Let us consider the case where we have five

CEPH-type pedigrees, each with the same, identical pedigree structure, as shown in Figure 14-1, with marker locus phenotypes as given in table 14-5:

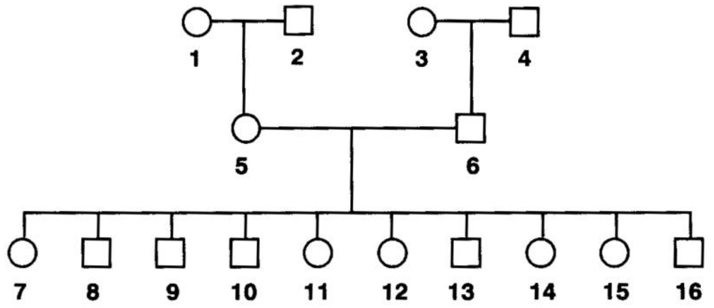


Figure 14–1. Pedigree structure for each of the five CEPH pedigrees in CEPH1.*

Pedigree	Individual	Marker 1	Marker 2	Marker 3	Marker 4
1	1	1 1	1 1	1 1	1 2
1	2	2 2	2 2	2 2	3 4
1	3	3 3	3 3	3 3	5 6
1	4	4 4	4 4	4 4	7 8
1	5	1 2	1 2	1 2	2 3
1	6	3 4	3 4	3 4	6 7
1	7	1 3	1 4	1 3	2 6
1	8	1 4	1 4	1 4	2 7
1	9	1 3	1 3	1 3	2 6
1	10	1 3	1 4	1 3	2 7
1	11	1 3	1 3	1 3	2 6
1	12	2 4	2 4	2 4	3 7
1	13	2 3	2 3	2 3	3 6
1	14	2 4	2 3	2 4	3 7
1	15	2 4	2 4	2 4	3 7
1	16	2 4	2 4	2 4	3 7
2	1	1 1	1 1	1 1	1 2
2	2	2 2	2 2	2 2	3 4
2	3	3 3	3 3	3 3	5 6
2	4	4 4	4 4	4 4	7 8
2	5	1 2	1 2	1 2	2 3
2	6	3 4	3 4	3 4	6 7
2	7	1 3	1 3	1 3	2 6
2	8	2 3	2 3	2 3	3 6
2	9	1 3	1 3	1 3	2 6
2	10	2 3	1 3	1 3	2 6
2	11	1 3	1 3	1 3	2 6
2	12	2 4	2 4	2 4	3 7
2	13	1 4	1 4	1 4	2 7
2	14	2 4	2 4	2 4	3 7
2	15	1 4	1 4	1 4	2 7
2	16	2 4	2 4	2 4	3 7
3	1	1 1	1 1	1 1	1 2
3	2	2 2	2 2	2 2	3 4
3	3	3 3	3 3	3 3	5 6
3	4	4 4	4 4	4 4	7 8
3	5	1 2	1 2	1 2	2 3
3	6	3 4	3 4	3 4	6 7
3	7	1 3	1 3	2 3	2 6
3	8	1 3	1 3	1 3	2 6
3	9	2 4	2 4	2 4	3 7
3	10	1 4	1 4	1 4	2 7
3	11	2 3	2 3	2 3	3 6
3	12	2 4	2 4	1 4	3 7
3	13	1 3	1 3	1 3	2 6
3	14	1 4	1 4	2 4	2 7
3	15	2 3	2 3	2 3	3 6
3	16	2 4	2 4	2 4	3 7
4	1	1 1	1 1	1 1	1 2
4	2	2 2	2 2	2 2	3 4
4	3	3 3	3 3	3 3	5 6

4	4	4 4	4 4	4 4	7 8
4	5	1 2	1 2	1 2	2 3
4	6	3 4	3 4	3 4	6 7
4	7	1 3	1 3	1 3	2 6
4	8	2 3	2 3	2 4	3 6
4	9	1 4	1 4	1 4	2 7
4	10	2 4	2 4	2 4	3 7
4	11	2 3	2 3	2 3	3 6
4	12	1 4	1 4	1 4	2 7
4	13	1 3	1 3	1 3	2 6
4	14	1 4	1 4	1 3	2 7
4	15	1 3	1 3	1 3	2 6
4	16	1 4	1 4	1 4	2 7
5	1	1 1	1 1	1 1	1 2
5	2	2 2	2 2	2 2	3 4
5	3	3 3	3 3	3 3	5 6
5	4	4 4	4 4	4 4	7 8
5	5	1 2	1 2	1 2	2 3
5	6	3 4	3 4	3 4	6 7
5	7	1 3	1 3	1 3	2 6
5	8	1 3	1 3	1 3	2 6
5	9	1 4	1 4	1 4	2 7
5	10	1 4	1 4	1 3	2 7
5	11	1 4	1 4	1 4	2 7
5	12	1 3	1 3	1 3	2 6
5	13	2 3	2 3	2 3	3 6
5	14	1 3	1 3	1 3	2 6
5	15	1 4	1 4	1 4	2 7
5	16	2 3	2 3	2 3	3 6

Table 14-5 : Genotypes for each individual at four marker loci in five CEPH-type pedigrees of the structure shown in figure 14-1.

Note that in this case, all meioses are fully informative, and occur in the same proportions as in table 14-4. Please make a standard LINKAGE format parameter file for this set of pedigrees, with the four allele numbers marker phenotypes as indicated below each individual, in file CEPH1.PRE. Then, process this file with MAKEPED to produce CEPH1.PED. Then, you should call up PREPLINK to create a parameter file (called CEPH1.DAT). You must specify that there are *4* loci, three of them with 4 equally frequent alleles, and one with 8 equally frequent alleles, with each locus being of the allele numbers type. Do not worry about the other options, like locus order, or any of the program specific parameters, as we will be using LCP shortly to set this up.

When you have finished, call up the LCP program, selecting the three-generational pedigrees option (these are the CLINKAGE programs). Choose the *CILINK* program, *All orders* option, with *No sex difference* in recombination fraction. Then you should set up the parameter screen as follows:

```
                  Locus Set [] : 1 2 3
 Recombination Fractions [.1] : .1 .1
```

Note that there are two recombination fractions to indicate here. The program will then reorder these three loci in all possible orders, and maximize the likelihood (under the assumption of no interference) for each given order respectively. Please hit <Page Down> to save this analysis setup, then hit <Ctrl-Z> to exit, and type *PEDIN* to run the analysis. You will note that instead of calling the UNKNOWN program, the LCP now calls a program called CFACTOR, which does the CEPH-type pedigree factorization, making these programs so efficient.

When the programs have finished running, use the Linkage Report Program (LRP) to examine the output. To do this, type *LRP* at the DOS prompt. The first screen should specify the STREAM file name as STREAM.OUT. (Remember this stream file was produced by each of the LINKAGE programs, and now you will see how to use it). This name is correct, so just hit <Page Down> to continue. At the next screen, you should select *Three-generation pedigree reports*, and then the *Multi-Point order report (CILINK)* option. Select the *table format*, and request that the report be *output to the screen*. You should then see the following information on the screen:

```
Order -2LN Like Odds
-------------------------------------------
 .070 .040
3----1----2        -1.9294E+02 1.00E+00 <==

 .040 .090
1----2----3        -1.8316E+02 1.33E+02

 .070 .090
1----3----2        -1.6602E+02 7.00E+05
```

Just as in our 2 × 2 analysis of this same dataset above, we note that locus order 3-1-2 (C-A-B = B-A-C) is the best order. Note that the recombination fraction estimates for each locus order are the same as they would be if estimated from the marginals of the 2 × 2 tables in table 14-4. The main question here is how to interpret these results. Typically, for a locus order to be conclusive, it is required that one have 1000:1 odds supporting the best order over the second best order, or a $\log_{10}$(likelihood) difference of at least 3 between the top two orders (analogous to the lod score of 3 requirement in a linkage test). In this case, we only have 133:1 odds, as shown in the third column above. To compute the "lod scale" equivalent, we can look at the values of –2ln(Like) shown above. Subtracting the –2ln(Like) of the best order from each of the others would give us the results shown in table 14-6 (Lod scale equivalents are computed by dividing the differences in –2ln(like) by 2ln(10) ≈ 4.6; odds (likelihood ratio) computed as 10^Z, where Z is the $\log_{10}$(Likelihood) difference):

Order	Δ2ln(Like)	Lod Scale	Odds
3--1--2	-0-	-0-	1
1--2--3	9.78	2.13	133
1--3--2	26.92	5.85	700,000

Table 14-6: CILINK analysis of loci 1, 2, and 3 in each possible order.

In this example, then, we can eliminate order 1-3-2, but cannot distinguish between orders 3-1-2 and 1-2-3, at the level of 1000:1 odds. The final interpretation is that you need to analyze more data to make a *framework map* (a map with best order supported by a likelihood ratio of at least 1000:1 over the next best order), but it seems much more likely that the true order is 3-1-2, though on this small sample size, it cannot be said to have been established that this is the case. The only real answer is to collect more families, or type the remainder of the CEPH families for these markers. One thing important to realize is that this analysis was based on only five CEPH-type families. In reality there are many more families than this, so you will certainly have better power for locus ordering when the whole panel is typed.

As an exercise to show how much efficiency is gained by using the CILINK program over the ILINK program, please repeat the above analysis using the ILINK program instead of the CILINK program. When you look at the results in table 14-7, you will notice that the values of –2ln(Like) are very different from those found with CILINK, but that the differences and odds are the same as indicated below, showing that they are equally valid, but that you cannot compare likelihoods generated with these two programs separately. The differences are due to the factoring algorithms used by CILINK.

Order	-2ln(Like)	Δ2ln(Like)	Lod Scale	Odds
3--1--2	555.66	-0-	-0-	1
1--2--3	565.44	9.78	2.13	133
1--3--2	582.58	26.92	5.85	700,000

Table 14-7: ILINK analysis of loci 1, 2, and 3 in each possible order.

EXERCISE 14

Please order the set of five loci given in the pre-MAKEPED file shown below. Also, try and draw out the pedigree structures. Are they all compatible with the CEPH pedigree structural requirements? Please enter this file as USEREX14.PRE, and process it with MAKEPED to produce USEREX14.PED.

```
44 1 12 13 1 1 0 0 1 0 1 0 1 1 1 0 0 1 1
44 2 14 15 2 1 0 1 1 1 0 1 0 1 1 0 0 1 0
44 3 1 2 2 0 0 0 1 1 0 1 1 1 1 0 0 1 0
44 4 1 2 2 1 0 0 0 1 1 1 1 1 1 0 0 1 1
44 5 1 2 2 1 0 1 0 1 1 1 1 1 1 0 0 1 1
44 6 1 2 1 1 0 0 0 1 1 1 1 1 1 0 0 1 1
44 7 1 2 1 1 0 1 1 0 1 1 1 1 1 0 0 1 0
44 8 1 2 1 1 0 1 1 0 0 1 1 1 1 0 0 0 0
44 9 1 2 2 1 0 0 0 1 1 1 1 1 1 0 0 1 1
44 10 1 2 2 1 0 1 1 1 0 1 1 1 1 0 0 1 1
44 11 1 2 1 0 0 0 1 0 0 1 1 1 1 0 0 1 0
44 12 0 0 1 1 1 0 0 1 1 0 1 1 1 0 0 0 1
44 13 0 0 2 1 1 0 1 1 0 0 1 1 1 0 0 1 1
44 14 0 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1
44 15 0 0 2 1 0 0 0 1 1 1 0 1 1 0 0 1 1
45 1 10 11 1 1 0 0 1 1 0 1 1 1 0 1 1 0 1
45 2 12 13 2 0 1 0 0 1 0 1 0 1 1 0 1 1 1
45 3 1 2 2 1 1 0 1 1 0 1 0 1 0 1 1 1 1
45 4 1 2 1 1 1 0 1 1 0 1 1 1 0 0 1 1 1
45 5 1 2 1 1 1 0 1 1 0 1 0 1 1 1 0 0 1
45 6 1 2 1 1 1 0 0 1 0 1 1 1 0 0 0 1 1
45 7 1 2 2 1 1 0 0 0 0 1 0 1 1 1 0 1 1
45 8 1 2 1 1 1 0 0 1 0 1 1 1 1 0 1 0 1
45 9 1 2 1 1 1 0 1 1 0 1 0 1 1 1 0 0 1
45 10 0 0 1 1 0 0 0 1 0 0 1 1 1 0 1 0 1
45 11 0 0 2 1 0 0 1 1 0 1 1 1 1 1 0 0 1
45 12 0 0 1 0 1 0 0 1 0 1 1 1 1 0 1 1 1
45 13 0 0 2 0 1 0 0 1 0 1 0 1 1 0 0 0 1
46 1 11 12 1 1 1 0 0 0 0 1 1 1 0 1 1 1 1
46 2 13 14 2 1 1 0 1 0 0 1 0 1 1 1 0 0 1
46 3 1 2 1 1 1 0 0 0 0 1 1 1 0 1 1 1 1
46 4 1 2 1 1 1 0 1 1 0 1 1 1 0 1 1 1 1
46 5 1 2 1 1 1 0 0 0 0 1 1 1 0 1 1 1 1
46 6 1 2 1 1 0 0 1 1 0 1 1 1 1 0 1 1 1
46 7 1 2 1 1 1 0 0 0 0 1 1 1 1 0 1 1 1
46 8 1 2 2 0 1 0 0 0 0 1 0 1 1 1 0 0 1
46 9 1 2 2 1 1 0 0 0 0 1 0 1 1 1 0 0 1
46 10 1 2 2 0 1 0 1 0 0 1 0 1 1 1 0 0 1
46 11 0 0 1 0 0 0 0 0 0 1 1 1 0 0 1 1 1
46 12 0 0 2 0 0 0 0 0 0 1 0 1 0 0 0 0 1
46 13 0 0 1 1 1 0 1 0 0 1 1 1 0 1 0 1 1
46 14 0 0 2 1 0 0 1 0 0 1 1 1 0 0 0 0 1
47 1 12 13 1 1 0 0 0 1 0 1 1 1 0 1 0 0 1
47 2 14 15 2 1 0 0 1 1 0 0 1 1 1 1 0 1 1
47 3 1 2 2 1 0 0 0 1 0 0 1 1 1 1 0 0 1
47 4 1 2 1 1 0 0 1 1 0 0 1 1 0 1 0 1 1
47 5 1 2 1 1 0 0 0 1 0 0 1 1 1 1 0 0 1
47 6 1 2 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1
47 7 1 2 2 1 0 0 1 1 0 0 1 1 1 1 0 1 1
47 8 1 2 2 1 0 0 1 1 0 0 1 1 1 1 0 0 1
47 9 1 2 1 1 0 0 0 1 0 1 1 1 0 1 0 1 1
47 10 1 2 1 1 0 0 0 1 0 1 1 1 1 1 0 0 1
47 11 1 2 1 1 0 0 1 1 0 0 1 1 1 1 0 0 1
47 12 0 0 1 1 0 0 0 1 0 1 1 1 0 1 0 1 1
47 13 0 0 2 1 1 0 1 1 0 1 0 1 1 1 0 0 1
47 14 0 0 1 1 0 0 0 1 0 1 1 1 1 0 0 0 1
47 15 0 0 2 1 1 0 1 0 0 1 1 1 0 1 0 1 0
49 1 11 12 1 0 1 0 1 0 0 1 1 1 1 0 0 0 1
49 2 13 14 2 1 1 0 1 1 0 1 1 1 1 1 0 1 1
49 3 1 2 2 1 1 0 1 1 0 0 1 1 1 0 0 1 1
49 4 1 2 1 0 1 0 1 1 0 1 1 1 1 1 0 0 1
49 5 1 2 1 1 1 0 1 1 0 1 1 1 1 0 0 1 1
49 6 1 2 2 1 1 0 1 0 0 1 1 1 1 0 0 1 1
49 7 1 2 2 1 1 0 1 0 0 0 1 1 1 0 0 0 1
49 8 1 2 2 0 1 0 0 0 0 0 1 1 1 0 0 0 1
49 9 1 2 2 0 1 0 0 0 0 1 1 1 1 0 0 0 1
49 10 1 2 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1
```

```
49 11 0 0 1 0 1 0 1 1 0 1 1 1 1 1 0 0 1
49 12 0 0 2 1 1 0 1 0 0 1 0 1 1 0 0 0 1
49 13 0 0 1 1 1 0 0 1 0 1 1 1 0 1 0 0 1
49 14 0 0 2 1 0 0 1 0 0 1 1 1 1 1 0 1 1
50 1 10 11 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1
50 2 12 13 2 1 1 0 1 1 0 1 0 1 0 1 1 0 1
50 3 1 2 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1
50 4 1 2 1 1 0 0 0 1 1 1 0 1 0 0 1 0 1
50 5 1 2 2 1 0 0 0 1 1 1 0 1 0 1 0 0 1
50 6 1 2 2 1 1 0 1 1 0 1 0 1 0 1 1 0 1
50 7 1 2 2 1 1 0 1 0 1 1 0 1 0 1 1 0 1
50 8 1 2 2 1 0 0 0 1 1 1 0 1 0 1 1 0 1
50 9 1 2 2 0 0 0 1 0 1 1 0 1 0 1 1 0 1
50 10 0 0 1 1 0 0 1 1 0 1 1 1 0 0 1 0 1
50 11 0 0 2 1 0 0 0 0 0 1 1 1 1 1 0 1 1
50 12 0 0 1 1 0 0 0 1 0 1 0 1 0 1 0 0 1
50 13 0 0 2 0 1 0 1 0 0 1 1 1 0 0 1 0 1
53 1 11 12 1 1 0 1 0 1 0 0 1 1 0 0 0 1 1
53 2 13 14 2 1 0 0 0 0 0 0 1 1 0 1 1 0 1
53 3 1 2 2 1 0 1 0 1 0 0 1 1 0 0 0 0 0
53 4 1 2 2 1 0 1 0 1 0 0 1 1 0 0 0 1 1
53 5 1 2 2 0 0 0 0 1 1 0 1 1 0 0 0 0 1
53 6 1 2 2 1 0 0 0 1 1 0 1 1 1 0 1 0 0
53 7 1 2 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1
53 8 1 2 2 1 0 1 0 1 0 0 1 1 0 0 0 1 1
53 9 1 2 1 1 0 1 0 1 0 0 1 1 1 1 0 1 1
53 10 1 2 2 1 0 1 0 0 0 0 1 1 1 0 1 0 0
53 15 1 2 1 1 0 0 0 0 0 0 1 1 1 0 1 0 1
53 11 0 0 1 0 1 1 0 1 0 1 1 1 0 0 0 0 0
53 12 0 0 2 1 0 0 0 1 1 1 1 1 0 0 0 0 1
53 13 0 0 1 1 0 0 0 0 0 1 1 1 0 1 0 0 1
53 14 0 0 2 1 1 0 0 1 1 0 1 1 0 1 1 1 1
54 1 17 14 1 1 1 0 0 1 1 1 1 1 1 1 0 1 0
54 2 15 16 2 1 0 0 1 0 0 1 1 1 0 1 1 0 1
54 3 1 2 2 1 0 0 1 0 1 1 1 1 0 1 0 1 1
54 4 1 2 2 1 1 0 1 1 0 0 1 1 0 1 1 1 1
54 5 1 2 2 1 1 0 1 1 0 1 1 1 1 0 1 1 1
54 6 1 2 2 1 1 0 1 0 1 1 0 1 1 1 0 1 1
54 7 1 2 2 1 0 0 1 0 1 0 1 1 0 1 0 1 1
54 8 1 2 2 1 0 0 1 0 1 0 1 1 0 1 1 1 1
54 9 1 2 1 1 0 0 0 0 0 1 0 1 1 1 0 1 1
54 10 1 2 1 1 1 0 1 1 0 1 0 1 1 1 0 1 1
54 11 1 2 2 1 1 0 1 1 0 1 1 1 1 1 0 1 1
54 12 1 2 1 1 0 0 1 0 1 1 1 1 0 1 0 1 1
54 13 1 2 1 1 1 0 1 1 0 0 0 0 1 0 1 1 1
54 17 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
54 14 0 0 2 1 1 0 1 0 1 1 1 1 1 1 0 1 1
54 15 0 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 1
54 16 0 0 2 1 1 0 1 1 0 1 0 1 0 1 0 0 1
55 1 15 16 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1
55 2 0 0 2 1 0 0 0 1 0 1 0 1 0 1 1 0 0
55 3 1 2 2 1 0 0 0 1 1 1 0 1 0 1 1 1 1
55 4 1 2 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1
55 5 1 2 1 1 0 0 0 1 1 1 0 1 0 1 1 1 1
55 6 1 2 1 1 0 0 0 1 0 1 0 1 0 1 1 0 1
55 7 1 2 2 1 0 0 0 0 0 1 0 1 0 1 1 0 1
55 8 1 2 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1
55 9 1 2 1 1 0 0 0 1 0 1 0 1 0 1 0 1 1
55 10 1 2 1 1 0 0 0 1 0 1 0 1 0 1 0 0 1
55 11 1 2 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1
55 12 1 2 2 1 0 0 0 1 1 1 0 1 0 1 1 1 1
55 13 1 2 2 1 0 0 0 1 1 1 0 1 0 1 0 1 1
55 14 1 2 1 1 0 0 0 1 1 1 0 1 0 1 0 1 1
55 17 1 2 2 1 0 0 0 1 0 1 0 1 0 1 0 0 1
55 15 0 0 1 1 0 0 0 1 1 1 1 1 1 1 0 0 1
55 16 0 0 2 1 0 0 0 1 0 1 0 1 1 0 1 0 1
56 1 18 13 1 1 0 0 1 0 0 1 0 1 1 1 0 0 1
```

```
56 2 12 14 2 1 1 0 1 1 0 1 0 1 1 1 0 1 0
56 3 1 2 1 1 0 0 1 1 0 1 0 1 1 0 0 1 1
56 4 1 2 1 1 1 0 1 1 0 1 0 1 0 1 0 1 1
56 5 1 2 2 1 1 0 1 1 0 1 0 1 0 1 0 1 1
56 6 1 2 2 1 0 0 0 0 0 1 0 1 0 1 0 1 1
56 7 1 2 2 1 1 0 1 1 0 1 0 1 1 1 0 1 1
56 8 1 2 2 1 1 0 0 0 0 1 0 1 0 1 0 1 1
56 9 1 2 1 1 1 0 0 0 0 1 0 1 1 1 0 1 1
56 10 1 2 1 1 1 0 0 0 0 1 0 1 1 1 0 1 1
56 11 1 2 2 1 0 0 1 0 0 1 0 1 1 0 0 1 1
56 15 1 2 2 1 1 0 1 1 0 1 0 1 1 1 0 1 1
56 16 1 2 2 1 1 0 1 1 0 1 0 1 0 1 0 1 1
56 17 1 2 2 1 0 0 1 0 0 1 0 1 1 1 0 1 1
56 18 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
56 13 0 0 2 1 0 0 1 0 0 1 1 1 0 1 0 1 1
56 12 0 0 1 0 1 0 1 1 0 1 0 1 1 1 0 1 1
56 14 0 0 2 1 1 0 1 1 0 1 0 1 1 0 0 1 0
57 1 18 15 1 1 1 0 1 1 0 1 0 1 0 1 1 1 0
57 2 19 16 2 0 1 0 1 1 0 1 1 1 0 1 0 0 1
57 3 1 2 2 0 1 0 1 1 0 0 0 0 0 1 1 1 1
57 4 1 2 1 0 1 0 1 0 0 0 0 0 0 1 0 1 1
57 5 1 2 1 0 1 0 1 0 0 0 0 0 0 1 0 1 1
57 6 1 2 2 0 1 0 1 0 0 1 0 1 0 1 1 1 1
57 7 1 2 1 0 1 0 0 1 0 1 0 1 0 1 1 1 1
57 8 1 2 2 0 1 0 1 1 0 1 0 1 0 1 1 1 1
57 9 1 2 1 1 1 0 1 1 0 1 1 1 0 1 0 1 1
57 10 1 2 2 0 0 0 1 1 0 1 0 1 0 1 0 1 1
57 11 1 2 2 0 1 0 1 0 0 1 0 1 0 1 1 1 1
57 12 1 2 1 0 1 0 1 1 0 1 1 1 0 1 0 1 1
57 13 1 2 2 1 1 0 1 1 0 1 1 1 0 1 0 1 1
57 14 1 2 1 0 1 0 1 1 0 1 1 1 0 1 1 1 1
57 17 1 2 2 1 1 0 1 1 0 1 1 1 0 1 0 1 1
57 18 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
57 15 0 0 2 1 1 0 1 1 0 1 0 1 1 0 1 1 0
57 19 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
57 16 0 0 2 1 1 0 0 1 0 1 0 1 0 1 0 0 1
58 1 11 12 1 0 1 0 1 1 0 0 1 1 0 1 1 0 1
58 2 13 14 2 1 1 0 1 1 0 1 1 1 0 1 0 0 1
58 3 1 2 1 1 1 0 1 1 0 0 1 1 0 1 1 0 1
58 4 1 2 1 0 1 0 1 1 0 0 1 1 0 1 1 0 1
58 5 1 2 2 1 1 0 1 1 0 0 1 1 0 1 0 0 1
58 6 1 2 1 0 1 0 1 0 0 0 1 1 0 1 1 0 1
58 7 1 2 1 1 1 0 1 1 0 1 1 1 0 1 1 0 1
58 8 1 2 2 0 1 0 1 0 0 0 1 1 0 1 0 0 1
58 9 1 2 2 0 1 0 1 0 0 1 1 1 0 1 1 0 1
58 10 1 2 1 0 1 0 1 0 0 0 1 1 0 1 0 0 1
58 15 1 2 1 1 1 0 1 1 0 1 1 1 0 1 0 0 1
58 11 0 0 1 1 1 0 0 0 0 0 1 1 0 0 0 1 1
58 12 0 0 2 1 1 0 1 0 0 1 1 1 0 0 0 0 1
58 13 0 0 1 0 0 0 1 1 0 1 1 1 0 1 1 0 1
58 14 0 0 2 1 0 0 0 1 0 1 1 1 0 1 1 1 1
59 1 0 0 1 1 1 0 1 1 0 1 1 1 0 1 0 0 0
59 2 0 0 2 1 1 0 0 1 1 1 0 1 1 1 0 1 1
59 3 1 2 1 0 1 0 0 1 1 1 1 1 0 1 0 1 0
59 4 1 2 2 0 1 0 0 1 1 1 1 1 1 1 0 1 0
59 5 1 2 2 0 0 0 1 0 1 1 1 1 0 1 0 0 1
59 6 1 2 2 1 1 0 0 1 0 1 0 1 1 1 0 1 1
59 7 1 2 2 0 1 0 0 1 0 1 1 1 1 1 0 1 0
59 8 1 2 1 1 1 0 1 1 0 1 0 1 1 1 0 0 0
59 9 1 2 1 1 1 0 1 1 0 1 0 1 0 0 0 1 1
59 10 1 2 1 0 1 0 0 1 0 1 1 1 0 1 0 1 0
59 11 1 2 1 1 1 0 0 1 1 1 1 1 0 1 0 1 0
59 12 1 2 1 1 0 0 1 1 0 1 0 1 0 1 0 0 0
59 13 1 2 1 1 0 0 1 1 0 1 0 1 0 1 0 0 1
59 14 1 2 1 1 0 0 1 1 0 1 0 1 0 1 0 0 1
59 15 1 2 1 1 1 0 0 1 1 1 1 1 0 0 0 1 1
59 16 1 2 1 1 1 0 1 1 0 1 0 1 1 1 0 1 1
```

```
 59 17 1 2 1 1 1 0 0 1 0 1 0 1 0 1 0 1 1
 59 18 1 2 2 1 1 0 0 1 1 1 0 1 1 1 0 1 1
```

The corresponding parameter file is shown below, and should be entered as USEREX14.DAT:

```
5 0 0 3
0 0.00000000 0.00000000 0
 1 2 3 4 5

2 3
 0.53571430 0.41785710 0.04642857
3
 1 0 0
 0 1 0
 0 0 1

2 3
 0.34931510 0.54452060 0.10616440
3
 1 0 0
 0 1 0
 0 0 1

2 2
 0.74083770 0.25916230
3
 1 0 1
 0 1 1

2 3
 0.24662160 0.62500000 0.12837840
3
 1 0 0
 0 1 0
 0 0 1

2 2
 0.33532940 0.66467060
2
 1 0
 0 1

0 0
 0.10000000 0.10000000 0.10000000 0.10000000
0
 1 1 1 1
```

Please use CILINK to order these five loci, as was described above. Can you order these five loci conclusively? If not, how many orders could not be excluded? What conclusions would you draw from this analysis? Can you run this analysis with ILINK?

15 CMAP and adding a new locus

15.1 MULTIPOINT LOD SCORES - THEORETICAL INTRODUCTION

Often, one has some idea of the map of a certain region, and desires to add a specific additional marker to this already established map. In such circumstances, it may be beneficial to fix the map of known markers, and compute lod scores for each possible location of the new marker. In other words, you simply move the new marker across the map, computing the likelihood if this new marker was actually located at each possible point along the map. This may be done in one of two ways: 1) the interval lengths of the current map are kept fixed, and only the new locus is moved across the known map (using CMAP), or 2) the new locus is successively placed in each possible interval with reanalysis of each locus order (re-estimation of all recombination fractions, using CILINK). The second method is in general preferable, and was described in chapter 14. Here, we describe the first method.

Let us assume that we have a known map of markers with fixed intermarker recombination fractions as shown in table 15-1.

```
Theta:              0.10 0.10
Locus:       ------1----2----3----
Interval:       0    1    2    3
```

Table 15-1: Known map of markers 1, 2, and 3.

Now, we have an additional marker to map against this known map. We would then compute the likelihood for each possible position of the new marker in each interval along the map. The CMAP (and LINKMAP) programs work by placing the new marker in each possible intermarker interval, and subdividing each interval into a certain number of equal segments, computing the likelihood at each step. If we divided each interval into two segments, we'd have likelihoods computed for the positions for the new locus, N, as shown in table 15-2.

Position	MAP (thetas)
A	N-(0.500)-1-(0.100)-2-(0.100)-3
B	N-(0.250)-1-(0.100)-2-(0.100)-3
C	N-(0.000)-1-(0.100)-2-(0.100)-3
D	1-(0.050)-N-(0.056)-2-(0.100)-3
E	1-(0.100)-N-(0.000)-2-(0.100)-3
F	1-(0.100)-2-(0.050)-N-(0.056)-3
G	1-(0.100)-2-(0.100)-N-(0.000)-3
H	1-(0.100)-2-(0.100)-3-(0.250)-N
I	1-(0.100)-2-(0.100)-3-(0.500)-N

Table 15-2: List of map positions for the new locus (N) to be analyzed with CMAP against the fixed map of markers 1,2, and 3.

The program equally divides each interval into two (in this case) segments, by dividing the recombination fraction by 2 for the entire interval (i.e. $\theta_{1,2} = 0.100$, so when the new locus is placed in this interval, this recombination fraction is divided by 2 to get $\theta_{1,N} = 0.05$. You will note that $\theta_{N,2}$ is not equal, but is computed according to the Haldane mapping function to make the map distance from locus 1 to locus 2 remain constant in Haldane centiMorgans.

Now that you have the $\log_{10}$ likelihoods computed for each of these points, how can you convert them into lod scores that can be interpreted easily? In this case, our null hypothesis is that the set of markers 1, 2, and 3 are linked to each other at fixed recombination fractions, but that marker N is unlinked to the entire map of markers. Thus, we would compute multipoint lod scores for map position "D" as $\log_{10}[L(D)/L(A)]$, where A and D are defined in table 15-2. Position A corresponds to the new marker unlinked to the fixed map of 1-(0.1)-2-(0.1)-3, and is thus characterized by our null hypothesis likelihood. Another measure that is often used is the so-called location score, which is based on 2 times the natural log likelihood of the same likelihood ratio, $2\ln[L(D)/ L(A)]$. This is often preferred, as it asymptotically follows a chi-square distribution with 1 degree of freedom. However, it has become traditional to express your results in terms of multipoint lod score for ease of comparison with the two-point situation.

A significant test result is again signified by the magic number of 3, so a multipoint lod score of greater than 3 is considered "proof" of linkage to this set of markers. Estimation can be carried out as well by this approach, analogously to the two-point case, with the most likely location of the new marker being at the point of its maximum likelihood. One must be certain to use a 3-unit-of-lod-score support interval for the location of the new marker. While there have been recommendations (Conneally et al, 1985) that a 1-unit-of-lod-score support interval may be acceptable in estimation of a two-point recombination fraction, in the multipoint case, one has the additional problem of needing to select one locus order over all other orders at 1000:1 odds as explained in the previous chapter. For this reason, one can only develop a meaningful support interval based on 3-units-of-lod-score (10^3 = 1000:1 odds). Even in the 2-point case, the 3-unit-of-lod-score support interval may be more appropriate and meaningful, since one only excludes the hypothesis of $\theta = 0.5$ when $Z(\hat{\theta}) - Z(0.5) \geq 3$, which is the same as saying the value $\theta = 0.5$ is outside a 3-unit support interval around the maximum lod score. For these reasons, we advocate the use of the 3-unit-of-lod-score support interval in all situations to avoid logical inconsistencies of the type described above.

15.2 COMPUTING MULTIPOINT LOD SCORES WITH CMAP

Let us try and map this new marker with the CMAP program in this family. For now, let us assume that we had actually found a significant result for ordering the first three markers as 3-(0.07)-1-(0.04)-2. Then, we would like to use marker 4 as our test locus, and compute the multipoint lod scores for this marker at a series of points across the fixed map. To do this, please call up the LCP program, entering the appropriate pedigree and parameter file names (CEPH1.*), and specifying the *Three-generational pedigrees* option, and then the *CMAP* program. Choose the *All map intervals* option, and *No sex difference* in recombination fractions. The next screen should be completed as follows:

```
                    Test loci [] : 4
       Order of fixed loci [] : 3 1 2
 Recombination fractions [.1] : 0.07 0.04
Number of evaluations in interval [5] : 5
```

The test locus is the new locus (4) for which likelihoods are to be calculated at a series of points across the fixed map of loci (3 1 2), separated by fixed recombination fractions (0.07 0.04). The last line, for number of evaluations per interval determines at how fine of a grid of points you wish to compute the lod scores. In this case, each intermarker interval will be divided into five equal subintervals, as described above. Please enter this problem, and hit <Page Down> to create the batch file. After you have run the analysis, please call up the LRP program to examine the results. This time, be sure to select the *Location score report (CMAP)* option, instead of the CILINK option you used in the locus ordering chapter. You should see the results given in table 15-3.

```
     Order          Loc. Score      -2LN Like       Odds          Lod     x
──────────────────────────────────────────────────────────────────────────────
4====3----1----2
 .500 .070 .040    +0.0000E+00   -1.9294E+02   7.13E+28        0.00   -∞
 .400 .070 .040    +3.0788E+01   -2.2373E+02   1.47E+22        6.69  -0.8047
 .300 .070 .040    +5.5433E+01   -2.4838E+02   6.54E+16       12.05  -0.4581
 .200 .070 .040    +7.4593E+01   -2.6753E+02   4.52E+12       16.22  -0.2554
 .100 .070 .040    +8.6796E+01   -2.7974E+02   1.01E+10       18.87  -0.1116
 .000 .070 .040    -1.1368E+02   -7.9266E+01   3.45E+53      -20.61    0

3====4====1----2
 .000 .070 .040                   infinity                            0
 .014 .058 .040    +1.0393E+02   -2.9687E+02   1.93E+06       22.59   0.0142
 .028 .044 .040    +1.1265E+02   -3.0559E+02   2.46E+04       24.49   0.0288
 .042 .031 .040    +1.1696E+02   -3.0990E+02   2.85E+03       25.43   0.0439
 .056 .016 .040    +1.1857E+02   -3.1151E+02   1.27E+03       25.78   0.0594
3----1====4====2
 .070 .000 .040                   infinity                           0.0754
 .070 .008 .033    +1.3115E+02   -3.2409E+02   2.37E+00       28.51   0.0835
 .070 .016 .025    +1.3281E+02   -3.2575E+02   1.03E+00       28.87   0.0917
 .070 .024 .017    +1.3287E+02   -3.2581E+02   1.00E+00 <==   28.88   0.1000
 .070 .032 .009    +1.3134E+02   -3.2429E+02   2.15E+00       28.55   0.1085
 .070 .040 .000                   infinity                           0.1171
```

```
3----1----2====4
 .070 .040 .000                   infinity                              0.1171
 .070 .040 .100    +1.0877E+02   -3.0171E+02   1.71E+05       23.65   0.2287
 .070 .040 .200    +8.8456E+01   -2.8140E+02   4.41E+09       19.23   0.3725
 .070 .040 .300    +6.3906E+01   -2.5685E+02   9.46E+14       13.86   0.5752
 .070 .040 .400    +3.4843E+01   -2.2778E+02   1.94E+21        7.57   0.9218
 .070 .040 .500    +0.0000E+00   -1.9294E+02   7.13E+28        0.00     ∞
```

Table 15-3: Results of CMAP analysis of the new locus (4) against the fixed map of loci 1-2-3.

You will note that the first column here is the location score. To compute multipoint lod scores from these location scores, you should divide each of these location scores by $2\ln(10) \approx 4.6$. These multipoint lod scores have been added to table 15-3, with their corresponding map locations. In this case, our 3-unit support interval is completely contained within the interval (1,2), so this locus has been uniquely placed into one interval, and we have significant evidence for this locus order (assuming that the fixed map was correct).

Note that the lod score for map N-(0.000)-3-(0.070)-1-(0.040)-2 was –20.61, while the lod score for map 3-(0.000)-N-(0.070)-1-(0.040)-2 was –∞. This happened because sometimes the recombination fractions shown are accurate to only three decimal places. The program divides the recombination fraction into equal segments, and computes lod scores for each subdivision of the intervals. However, sometimes the division isn't completely accurate, but is only to the level of machine precision. When dealing with recombination fractions of 0, the slightest deviation from 0 can have a quite major effect on the lod score, since only at 0 itself is the lod score equal to –∞. Therefore, whenever you have two different values for the lod score at $\theta = 0$ from a given marker, and one of them = –∞, this is the correct value. In general, when the left-side recombination fraction is 0, this is the more accurate value. Similarly, it is not always the case that the likelihood when the new locus is unlinked to the set of markers on the left side equals that computed with the disease unlinked on the right side. Again, this is due to rounding error, and in every case, it is more accurate to use the left-side value. We therefore recommend that one always normalize the lod scores to the likelihood with the new marker unlinked to the set of known loci on the left hand side of the fixed map.

15.3 MAP DISTANCE

It is always useful to be able to represent the multipoint lod scores in graphical format. To do this, you would need to express each putative location of the new marker in terms of its map location in Morgans. For arguments' sake, we typically take the location of the leftmost marker to be at map position 0. All other map distances can be computed by the Haldane mapping function. The Haldane mapping function must, in principle, be used here, as the likelihood computations are all performed assuming absence of interference (although some researchers often erroneously use the Kosambi mapping function to give the illusion of a shorter map). For example, in interval 0 (to the left of marker 3), the map distances of each point from marker 3 can be computed by converting the recombination fractions $\theta_{N,3}$ to map distance by the relationship $x_{HALD}(\theta) = -\frac{1}{2}\ln(1-2\theta)$, where x_{HALD} is the distance in Haldane Morgans corresponding to recombination fraction θ. In this case, since these recombination fractions are to the left of locus 3, we should prefix them with a minus sign, since we are standardizing locus 3 to be map position 0, without loss of generality. Map positions between loci 3 and 1 can be computed similarly, by converting $\theta_{3,N}$ to map distance by the same formula. However, after you get to the right of locus 1, things get more complicated. You no longer are given recombination fractions $\theta_{3,N}$, but instead have $\theta_{3,1}$, and $\theta_{1,N}$. At this point you should convert $\theta_{3,1}$ to map distance, and $\theta_{1,N}$ to map distance, and add them together. Similarly, for points to the right of marker 2, you would have $\theta_{3,1}$, $\theta_{1,2}$, and $\theta_{2,N}$. Convert each of these to map distance using the above formula, and sum the map distances together. For example, for map position 3-(0.07)-1-(0.04)-2-(0.20)-N, the corresponding map position (distance from locus 3) would be $X_{HALD}(0.07) + X_{HALD}(0.04) + X_{HALD}(0.20) = 0.0754 + 0.0417 + 0.2554 = 0.3725$. These $\theta \rightarrow x$ conversions can be performed with the Linkage Utility Program MAPFUN. To do this, simply call up the MAPFUN program. The program will prompt you with the following:

```
Calculate Map distance from given theta [M, or MS for summing ts]
 or Theta from given map distance [T]? (-1 exits)
```

In this situation we want to compute map distances from given θ's, so you should select option M. You will then be asked:

```
Enter theta [+ mapping parameters] (-1 exits)
```

Just enter the appropriate value of θ, for example 0.07 (= $\theta_{3,1}$). Do not worry about other mapping parameters. These have to do with other mapping functions, which we do not need to deal with for now. The program should give you results as follows:

```
 Theta   xHALD   xKOS xCaFal   xRAO  xFELS xSTURT xBINOM
                    param => 0.350  0.000  2.000  4.000
0.0700 0.0754 0.0705 0.0700 0.0701 0.0705 0.0751 0.0740
```

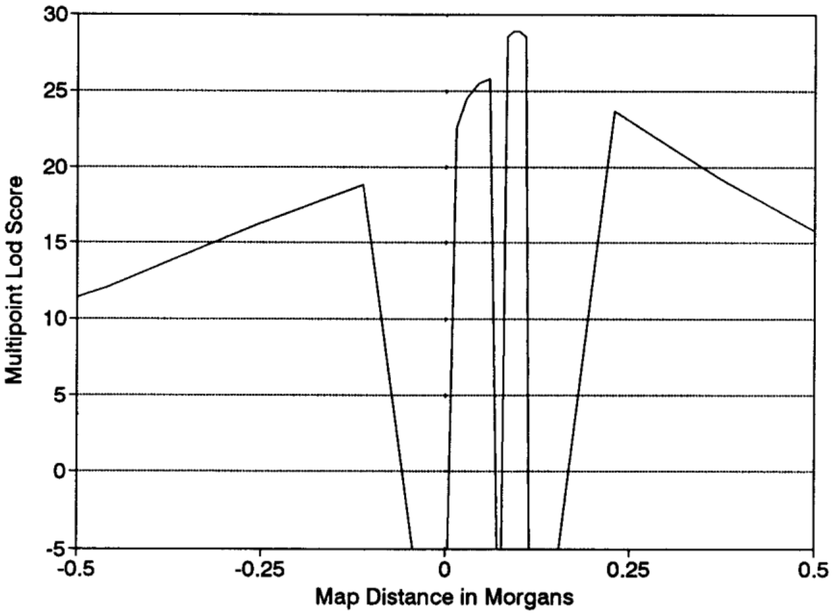


Figure 15-1. Graph of multipoint lod scores from Table 15-3

The only map distance you need is the Haldane map distance. The other mapping functions are obtained from various models of interference, which will be discussed in a later chapter. In general, if you have computed multipoint lod scores with the LINKAGE or CLINKAGE programs, the programs are basing the likelihood calculations on the assumption of no interference, so the only valid and meaningful conversion is from θ to x_{HALD}. Two-point θ's can validly be converted by any map function you choose to assume, since no multilocus gamete probabilities are involved therein. So, for our purposes, all the other output can safely be ignored. Our result is $x_{HALD} = 0.0754$. Now enter –1 twice to exit the program. Note that the reverse transformation (x → θ) can be performed by this program for a variety of mapping functions as well. The map distances for the example with CMAP are given in table 15-3. A graphical display of the multipoint lod score curve is then possible, by plotting Z(x) vs. x, as shown in Figure 15-1.

EXERCISE 15

As an exercise, try using CMAP to map locus *4* against the second-best locus order determined with CILINK, 1 – (0.04) – 2 – (0.09) – 3, for the data in files CEPH1.*. Can the locus *4* be uniquely assigned to one of these map intervals by the 3-unit support interval criterion? Compute map distances and multipoint lod scores across this other map.

Let us also go back and try running CILINK on all four loci together. Note that this would be impossible with regular ILINK, just as this whole chapter's exercises would've been impossible with regular LINKMAP on a PC. Is one order preferred significantly above all others? Is there anything interesting about the two best orders? Does the addition of this new marker locus help us at all in ordering loci 1, 2, and 3? Why or why not?

Go back to the dataset from exercise 14, and use CILINK to make a map of loci 1, 2, and 3. Then, fixing this map, go back and try to add locus 4. Check and see if these results are compatible with 4-point CILINK results. Then try to add locus 5 to the same map of loci 1 2 and 3. Compute all multipoint lod scores and their corresponding map locations as in this chapter. Are the results compatible with the CILINK results from the last chapter? What differences, if any, are there in the results, and the significance levels of our test for locus order and why?

16 Mapping a disease locus against a fixed map of markers

16.1 MULTIPOINT TESTING AND ESTIMATING LINKAGE WITH DISEASES

There are two phases of any genetic linkage analysis, the testing phase, and the estimation phase. Clearly, one must have a significant test result before proceeding to the estimation step, in order to prevent logical inconsistencies. In the testing phase, there is no great advantage to using multipoint analysis, when you are analyzing highly polymorphic markers. You see, the maximum multipoint lod score possible would be equal to the maximum possible two-point lod score when all meioses were informative in both cases. The only time a benefit is obtained in the testing phase is when multipoint analysis allows us to increase the percentage of meioses informative for at least one marker and the disease. In general, however, when there are questions about the model parameters, or diagnostic criteria, as is common in complex disorders, it is advisable to rely on some kind of two-point analysis for the testing phase of any linkage analysis. Not only is two-point analysis more robust in testing situations, but it also can be done at a much lower cost in computer time. Because the mode of inheritance of complex traits is typically unknown and the analysis model presumably is different from the true model, estimates of θ tend to be inflated (Risch, 1990). In multipoint analysis, this inflation tends to "drive" the trait outside of a map even though the trait locus may be in the middle of the map of markers (Risch, 1990). Complex traits will be discussed in more detail in Part III.

The greatest utility of multilocus analysis comes in estimating the location of your linked disease gene, provided that the mode of inheritance is well known. As we have seen in the last few chapters, multipoint analysis can be a very powerful tool for ordering loci along a chromosome. The simplest way of seeing how to do this is by selecting the location for a new gene which would minimize the number of double (and to a lesser extent, single) recombinants. We saw how the CILINK and CMAP programs can be used to localize a new marker relative to other nearby markers. However, when the new locus to be mapped is a disease gene, the situation becomes somewhat different. No longer does one have the advantage of a 1:1 correspondence between phenotype and genotype. One now is dealing with this additional variable of penetrance. Since more complicated penetrance models are required for disease mapping than for marker mapping, one can no longer take advantage of the rapid and efficient algorithm used by the CLINKAGE programs. Instead, one must rely on the general LINKAGE programs, which are much more restrictive in terms of the number of loci that can be used, and the maximum number of permissible alleles at each locus. The fundamental concepts and usage of ILINK and LINKMAP are the same as those you have already seen for CILINK and CMAP, but just computing efficiency is lower, and the generality higher.

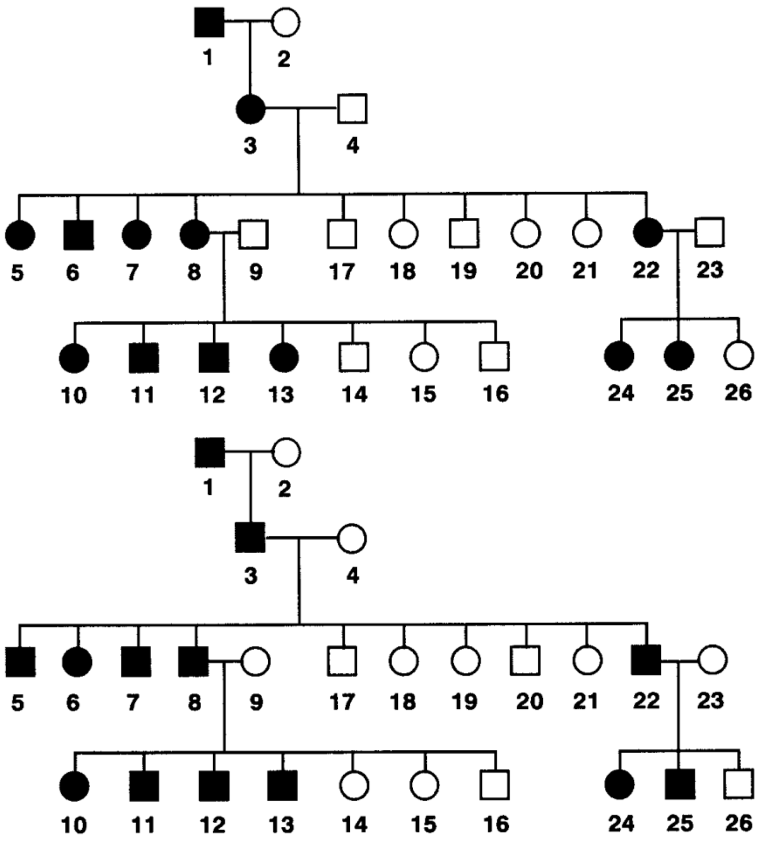


Figure 16–1. Pedigree structures for pedigrees in MULTDIS1.*

16.2 DISEASE GENE MAPPING

Please create pedigree and parameter files (MULTDIS1.*) for the two pedigrees shown in Figure 16-1 (with marker locus phenotypes indicated in table 16-1).

Pedigree	Individual	M1	M2	M3	M4	M5	M6	M7	M8
1	1	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
1	2	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	3	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	4	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	5	1 2	1 2	1 1	1 2	1 2	1 1	1 2	1 1
1	6	1 1	1 2	1 2	1 1	1 2	1 2	1 1	1 2
1	7	1 2	1 2	1 2	1 1	1 2	1 2	1 2	1 2
1	8	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	9	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	10	1 2	2 2	1 2	1 2	2 2	1 2	2 2	2 2
1	11	1 1	1 2	1 2	2 1	1 2	1 2	1 2	2 2
1	12	1 1	2 1	1 2	1 1	2 2	1 2	1 1	2 2
1	13	1 2	1 2	1 1	2 2	1 2	1 2	1 2	1 1
1	14	1 1	2 1	1 1	1 1	2 1	1 2	1 1	1 1
1	15	1 2	1 1	1 1	2 2	1 1	1 1	2 2	1 1
1	16	1 2	2 1	1 1	1 2	2 1	1 2	1 2	1 1
1	17	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
1	18	1 1	1 1	1 1	1 1	1 1	1 2	1 1	1 2
1	19	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	20	1 2	1 1	1 1	1 2	1 1	1 2	1 2	1 1
1	21	1 2	1 1	1 1	1 1	1 1	1 1	1 2	1 1
1	22	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	23	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	24	1 1	2 2	2 2	1 1	2 2	2 2	2 2	2 2
1	25	1 2	1 2	1 2	1 2	1 2	1 1	1 2	1 1
1	26	1 2	1 2	1 1	1 2	1 1	1 1	1 2	1 1
2	1	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
2	2	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	3	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	4	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	5	1 2	1 2	1 1	1 2	1 2	1 1	1 2	1 1
2	6	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	7	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	8	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	9	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	10	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	11	1 1	1 2	1 2	1 1	1 2	1 2	1 2	1 2
2	12	1 1	1 1	1 2	1 1	1 2	1 2	1 1	1 2
2	13	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	14	1 1	1 1	1 1	1 1	1 1	1 2	1 1	1 1
2	15	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
2	16	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	17	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	18	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	19	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	20	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
2	21	1 2	1 1	1 1	1 1	1 1	1 1	1 2	1 1
2	22	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	23	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	24	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
2	25	1 2	1 2	1 2	1 2	1 2	1 1	1 2	1 1
2	26	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1

Table 16-1 : Marker locus phenotypes for pedigrees in figure 16-1.

The disease is assumed to be a fully penetrant dominant disorder, with gene frequency of 0.0001, and each of the eight markers is assumed to have two equally frequent codominant alleles.

Now that we have the pedigree and parameter files prepared, we must decide on a course of action for this analysis. Before going any further with a multipoint type of analysis, it is wise to perform two-point analysis with the disease versus each of the markers to see if there is any two-point evidence to support a linkage to this region. So what we would like you to do is use the LCP program to prepare a batch file which will perform two-point analyses with the disease vs. each of the markers. To do this, invoke LCP, specify the appropriate parameter file and pedigree file. Then, select the *MLINK* program, with *specific evaluations* option, and *No sex difference*. Then enter the appropriate analyses to be done as follows: First, to analyze locus 1 (Disease) vs. locus 2, enter the following options:

```
Locus Order [] : 1 2
Recombination Fractions [.1]: 0
Recombination varied [1]: 1
Increment value [.1]: .1
Stop value [.5]: .5
```

followed by <Page Down>. To analyze the disease vs. Locus 3, repeat the process, except instead of entering Locus Order [] : *1 2*, enter *1 3*. Normally, one would continue doing this until you have selected the disease to be analyzed vs.each of the loci, but in the interest of time, only go up to disease vs. 4. At this point, press <Ctrl-Z> to exit from LCP, and write the batch file. To perform these analyses, you would type *PEDIN* to invoke your newly created batch file, PEDIN.BAT. Another way to get similar results would be with the ILINK program which will iteratively find the maximum likelihood estimate of the recombination fraction. Please try this analysis again with ILINK to verify that your results are consistent - note that ILINK results will give you an estimate, while MLINK gives you a sense of the shape of the likelihood curve. To do this, again invoke the LCP program, select parameter and pedigree files as before, and then select the *ILINK* program, with *Specific orders* option, and *No sex difference*. Then enter your problems. The following example is as above for the comparison of markers 1 and 2:

```
Locus Order [] : 1 2
Recombination Fractions [.1] : 0.1
```

Then hit <Page Down>, and enter the next comparison, 1 vs. 3. Normally, one would continue in this manner until all possible 2 point comparisons involving the disease locus (i.e. 1 vs. 2,..., 1 vs. 9) had been specified, but again, for time reasons, only analyze the disease versus the first three loci (2, 3, and 4). Note that sometimes ILINK gives an estimate of theta, > 0.50. This is an artifact of the method the program uses to maximize the likelihood. All such estimates can be thought of as being = 0.5, as that is the maximum meaningful value for a recombination fraction. Lod scores at θ > 0.5 may serve as a check of data consistency; if Z(> ½) is large (> 2, say), the data may contain errors, and should be checked. Table 16-2 is a table of two-point intermarker recombination fractions which should coincide with those estimates you obtained, though there may be some minor differences.

	2	3	4	5	6	7	8	9
$\hat{\theta}$:	0.363	0.031	0.085	0.324	0.000	0.206	0.295	0.150
Z($\hat{\theta}$):	0.540	7.700	6.090	0.940	10.225	2.727	1.290	4.698

Table 16-2: ILINK estimates of 2-point θ, and Z() for each marker locus vs. disease.

The most tightly linked marker appears to be locus 6. For this locus, please perform an MLINK analysis in steps of 0.01, in order to find the upper bound for the 3-unit of lod score support interval. In this case, the support interval should cover the interval [0,0.19), since the lod score at 0 is 10.225, and the lod score first falls below 7.225 (= 10.225 – 3) at θ = 0.19, where Z(0.19) = 7.12. So, we can see that in the 2-point analysis, we have a support interval for the location of our disease gene covering $-\frac{1}{2}\ln(1-2(0.19))=0.239\,M=23.9\,cM$.

Further, the support interval could extend on either side (proximal or distal) of our marker, so actually the support interval is twice this long, or 47.8 cM. Now, we should try and use multipoint analysis to see what effect it will have on the length of our support interval, and the accuracy with which we can map our disease gene.

Let us assume we have obtained a fixed map of our markers from the CEPH pedigrees as follows: 5-2-8-3-6-4-9-7, with intermarker θs as follows: 0.075 0.075 0.225 0.075 0.075 0.075 0.075. We must now find the location of our disease gene along this map of markers. To do this, we must use the LINKMAP program, and move the disease locus through each of the intervals along our fixed map of loci, as was done in the CMAP example. We will be unable to analyze all the loci jointly in a 9-point analysis with the general pedigrees version of LINKMAP, due to memory constraints. To circumvent this problem, we will use the technique of sliding our group of four loci down the map, and analyzing the disease only in the middle interval (i.e with 2 flanking markers on each side) of each set of markers. We will first analyze the disease against markers 5, 2, 8, and 3, allowing the disease locus to move over the region from the left of locus 5 up to locus 8. At that point we will switch to looking at the set of loci 2-8-3-6, when we move the disease in the middle interval of this set of loci, 8-3. We proceed like this until we have covered the entire length of our map, as shown in table 16-3.

```
====5====2====8----3
         2----8====3----6
              8----3====6----4
                   3----6====4----9
                        6----4====9====7====
```

Table 16-3: Demonstration of strategy for picking which markers to use in a multipoint LINKMAP analysis of disease (=) vs a map of 8 markers, when only 5-point analysis is possible.

In this way, we can compute multipoint lod scores across the map, using the nearest markers to each interval in which the disease will be placed. While these lod scores will not be the same as those obtained from a putative nine-point analysis, this type of analysis would be impossible on a PC with the LINKAGE programs. After computing all the multipoint lod scores across this map of markers, we can find our estimated map position for the disease, and a multipoint three-unit-of-lod-score support interval for the map position, as before.

To perform these multipoint analyses, invoke LCP, and this time choose the *LINKMAP* program. Since we are going to analyze the disease in each interval with a different set of markers, we need to specify the *Specific intervals* option, with *No sex difference*. Now, for our first few analyses we will be using the set of markers 5 2 8 and 3. Normally we only use a specific set of markers to analyze the disease in the middle interval, but in this case, we have no markers farther out, so we can get the best results from these four markers. Our test locus is the disease. This means that the disease will be moved throughout whatever interval we specify on the fixed map of loci 5-2-8-3. For the entry of recombination fractions, we must input the KNOWN, FIXED recombination fractions between each of the loci. In this case, they are 0.075 between 5 and 2, 0.075 between 2 and 8, and 0.225 between 8 and 3. The test interval must be specified next. It is imperative to have one evaluation for each set of markers with the disease fixed at θ = 0.50 to the left of the leftmost marker. This value of the loglikelihood at this point will have to be subtracted from the loglikelihoods at all points analyzed with this specific marker set to calculate multipoint lod scores and location scores. So to do this we specify test interval = 0, meaning to the left of the fixed map, and we request 1 evaluation, so it only does calculations at θ = 0.5, and possibly also at θ = 0 from the leftmost marker. Otherwise, test interval refers to which intermarker interval you want the disease (test locus) to move through. In our example, test interval 1 would cause the disease to be moved between marker 5 and 2, interval 2 would refer to that between 2 and 8, etc. Note that the largest value possible would be interval 4, meaning to the right of the rightmost (4th) locus. You may specify any number of points at which the likelihood should be calculated, but in general 5 is sufficient, unless you want a finer map, in which case you can raise it. For our purposes, we'll use five evaluations per interval. In other words to analyze the disease to the left of our map enter the following:

```
Test loci [] : 1
Order of fixed loci [] : 5 2 8 3
Recombination fractions [.1] : 0.075 0.075 0.225
Test interval [0] : 0
Number of evaluations [5] : 5
```

To analyze the disease between 5 and 2, just change test interval to 1. Further, analogously, to analyze the disease between loci 2 and 8, change the test interval to 2. Note that in this case I didn't have to separately specify an analysis to the left of the map of markers with one evaluation to get the value of the log likelihood at θ = 0.5, as that will be calculated anyway when the disease was moved through all points to the left of our map of loci. However, in the next interval, between 8 and 3, we will need to change our set of marker loci to 2 8 3 and 6, so in this case we will need to do a separate entry to calculate the value at θ = 0.5 to the left of this new set of markers as follows:

```
Test loci [] : 1
Order of fixed loci [] : 2 8 3 6
Recombination fractions [.1] : 0.075 0.225 0.075
Test interval [0] : 0
Number of evaluations [5] : 1
```

to get the value at θ = 0.5, and then

```
Test loci [] : 1
Order of fixed loci [] : 2 8 3 6
Recombination fractions [.1] : 0.075 0.225 0.075
Test interval [0] : 2
Number of evaluations [5] : 5
```

to move the disease between loci 8 and 3. Also note that as we changed the loci in the analysis, we also had to enter the correct recombination fractions for each interval, as this is crucial to the analysis. Please continue this process until you have prepared a batch file that will run the disease over every point on the map using this sliding window type of method. We would like you to calculate the location scores and lod scores across the map, using the output from LRP to help you if you wish. Everything is essentially the same as for the analysis of CMAP output in the previous chapter, and is presented in table 16-4, and the multipoint lod scores are illustrated graphically in Figure 16-2. The calculation of location scores is left as an exercise for the user.

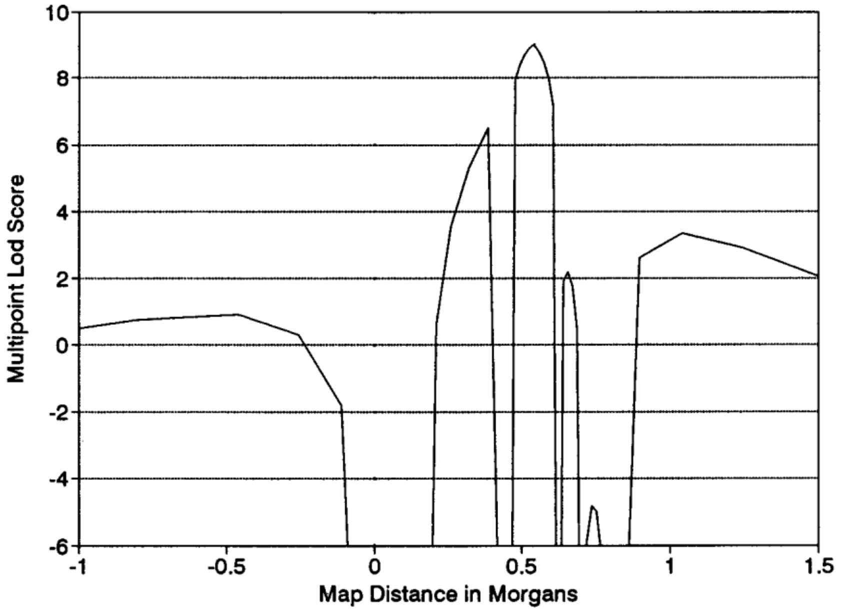


Figure 16–2. Graph of multipoint lod scores from Table 16–4

Intermarker Thetas				ln(like)	log(like)	Map distance from locus 5	Location Scores	Lod Scores
LOCI: 1 5 2 8 3								
0.5	0.075	0.075	0.225	-192.690	-83.684	-∞	0	0
0.4	0.075	0.075	0.225	-190.952	-82.929	-0.80471		0.755
0.3	0.075	0.075	0.225	-190.572	-82.764	-0.45814		0.920
0.2	0.075	0.075	0.225	-191.961	-83.367	-0.25541		0.317
0.1	0.075	0.075	0.225	-196.877	-85.502	-0.11157		-1.819
0	0.075	0.075	0.225	-∞	-∞	0	-∞	-∞
LOCI: 5 1 2 8 3								
0.015	0.062	0.075	0.225	-236.121	-102.545	0.015229		-18.862
0.03	0.048	0.075	0.225	-231.712	-100.631	0.030937		-16.947
0.045	0.033	0.075	0.225	-231.877	-100.702	0.047155		-17.018
0.06	0.017	0.075	0.225	-236.789	-102.836	0.063916		-19.152
0.075	0	0.075	0.225	-∞	-∞	0.081259	-∞	-∞
LOCI: 5 2 1 8 3								
0.075	0	0.075	0.225	-∞	-∞	0.081259	-∞	-∞
0.075	0.015	0.062	0.225	-237.803	-103.276	0.096489		-19.592
0.075	0.03	0.048	0.225	-231.880	-100.704	0.112197		-17.020
0.075	0.045	0.033	0.225	-230.579	-100.139	0.128414		-16.455

```
0.075  0.06 0.017 0.225  -233.556 -101.432  0.145176            -17.748
0.075 0.075     0 0.225    -∞        -∞     0.162518   -∞         -∞

LOCI: 1 2 8 3 6

0.500 0.075 0.225 0.075  -185.8052 -80.694

LOCI: 2 8 1 3 6

0.075     0 0.225 0.075    -∞        -∞     0.162518   -∞         -∞
0.075 0.045 0.198 0.075  -184.419  -80.092  0.209674              0.602
0.075  0.09 0.165 0.075  -177.598  -77.130  0.261744              3.564
0.075 0.135 0.123 0.075  -173.590  -75.389  0.319874              5.305
0.075  0.18  0.07 0.075  -170.796  -74.176  0.385662              6.518
0.075 0.225     0 0.075    -∞        -∞     0.461437   -∞         -∞

LOCI: 1 8 3 6 4

0.500 0.225 0.075 0.075  -183.599  -79.735

LOCI: 8 3 1 6 4

0.225     0 0.075 0.075    -∞        -∞     0.461437   -∞         -∞
0.225 0.015 0.062 0.075  -165.297  -71.787  0.476667              7.948
0.225  0.03 0.048 0.075  -164.244  -71.330  0.492375              8.405
0.225 0.045 0.033 0.075  -163.575  -71.039  0.508592              8.696
0.225  0.06 0.017 0.075  -163.117  -70.840  0.525354              8.895
0.225 0.075     0 0.075  -162.850  -70.724  0.542696              9.011

LOCI: 1 3 6 4 9

0.500 0.075 0.075 0.075  -176.860  -76.809

LOCI: 3 6 1 4 9

0.075     0 0.075 0.075  -156.197  -67.835  0.5427                8.974
0.075 0.015 0.062 0.075  -156.679  -68.044  0.5579                8.765
0.075  0.03 0.048 0.075  -157.363  -68.342  0.5736                8.467
0.075 0.045 0.033 0.075  -158.403  -68.793  0.5899                8.016
0.075  0.06 0.017 0.075  -160.302  -69.618  0.6066                7.191
0.075 0.075     0 0.075    -∞        -∞     0.6240     -∞         -∞

LOCI: 1 6 4 9 7

0.500 0.075 0.075 0.075  -178.324  -77.445

LOCI: 6 4 1 9 7

0.075     0 0.075 0.075    -∞        -∞     0.623956   -∞         -∞
0.075 0.015 0.062 0.075  -173.943  -75.542  0.639185              1.903
0.075  0.03 0.048 0.075  -173.317  -75.270  0.654894              2.174
0.075 0.045 0.033 0.075  -174.253  -75.677  0.671111              1.768
0.075  0.06 0.017 0.075  -177.254  -76.980  0.687873              0.465
0.075 0.075     0 0.075    -∞        -∞     0.705215   -∞         -∞

LOCI: 6 4 9 1 7

0.075 0.075     0 0.075    -∞        -∞     0.705215   -∞         -∞
0.075 0.075 0.015 0.062  -191.679  -83.245  0.720445             -5.800
0.075 0.075  0.03 0.048  -189.469  -82.285  0.736153             -4.840
0.075 0.075 0.045 0.033  -189.804  -82.430  0.752371             -4.986
0.075 0.075  0.06 0.017  -192.860  -83.757  0.769132             -6.313
0.075 0.075 0.075     0    -∞        -∞     0.786475   -∞         -∞

LOCI: 6 4 9 7 1

0.075 0.075 0.075     0    -∞        -∞     0.786475   -∞         -∞
0.075 0.075 0.075   0.1  -172.369  -74.858  0.898047              2.586
0.075 0.075 0.075   0.2  -170.619  -74.099  1.041888              3.346
0.075 0.075 0.075   0.3  -171.654  -74.548  1.244620              2.897
0.075 0.075 0.075   0.4  -174.283  -75.690  1.591194              1.755
```

Table 16-4: Results of LINKMAP analysis with disease against the fixed map of 8 markers.

In this example, we can see that our maximum multipoint lod score is 9.011, which is actually lower than that achieved in the 2-point analysis with marker 6, though the MLE of the disease location is still at marker 6. In this analysis, the 3-unit of lod score support interval would encompass only those regions with lod scores of greater than 6.011, which would extend, in this example, within the region between markers 3 and 4, and slightly to the left of marker 3, where a lod score of 6.518 was achieved at 0.07 to the left of marker 3. In this case, our support interval is disjoint, covering a range of 14.52 cM between the markers 3 and 4, and up to 7.58 cM to the left of marker 3, though the region is disjoint, having absolute length less than the entire 22.1 cM range. Remember that in the two-point example, the support interval of 47.8 cM is more than twice this length. If we had more tightly linked markers in our map, the properties would be different, and we could potentially even further isolate the location of our disease gene.

EXERCISE 16

Assume for the moment that the map was actually the following:

```
5-(0.02)-2-(0.02)-8-(0.04)-3-(0.01)-6-(0.01)-4-(0.02)-9-(0.03)-7
```

Now, rerun the LINKMAP analysis. In this case, what is the effect on the maximum lod score, and the 3-unit-of-lod-score support interval? How long is the support interval now, and how many intervals are covered? This points out the importance of having a good accurate map of your markers before doing the LINKMAP analysis, as the results are highly dependent on the initial map distance estimates used, in terms of the length of the 3-unit-of-lod-score support intervals, and sometimes the lod score values themselves. Remember that irrespective of the specific map of markers, the two-point support interval is still 47.8 cM long for this example, so if this were the true map of markers, multipoint analysis would have reduced the length of our support interval to a small fraction of its original length.

17 Exclusion Mapping

In this chapter, you will be introduced to the concept of exclusion mapping. Issues surrounding the usage of negative test results will be considered, including relative power of two-point and multipoint methods for excluding chromosomal regions, and various pitfalls in the interpretation of negative linkage test results. Two methods of exclusion mapping are in general use. The first is based on the log likelihood of a trait versus a marker or map of markers, whereas the second employs Bayesian arguments (prior/posterior probability of map position), and has been employed for two-point analysis in a program called EXCLUDE (Edwards, 1987). Here, only the likelihood-based approach is illustrated.

17.1 USING NEGATIVE TEST RESULTS

In linkage analysis, one is primarily interested in localizing putative disease genes relative to well-characterized marker loci. As we have discussed throughout, the foci of such an analysis are trying to obtain a positive test result for linkage (lod score > 3), and then trying to fine map the precise chromosomal location of the disease gene, as a prelude to isolating the precise genetic effect by molecular methods. However, with any given marker, the probability of finding a positive test result is quite low, as the human genome is very large, and most randomly selected markers will not be liked to the putative disease gene. Further, even if the selected marker is linked to the disease gene, there is no guarantee of a positive test result in any finite pedigree sample. In light of this, there is a need for some way to deal with negative test results as well, to try and eliminate various chromosomal regions from consideration. In doing this, one could concentrate the remainder of his genomic search without repeating redundant work, in areas where the gene most likely is not. One way of doing this is so-called exclusion mapping.

The methodology of exclusion mapping is quite different from that of the test and estimate approach to positive "inclusion" mapping. Obviously, if the test statistic one is applying is $Z_{\max} = \log_{10} \frac{L(\hat{\theta})}{L(\theta = \frac{1}{2})}$,

then $Z_{\max} \geq 0$ always. It is important to remember that the likelihood ratio test is a test of the hypothesis of no linkage, such that in the absence of a significant test result, you fail to reject H_0, meaning that there is no significant evidence for linkage. However, this does not mean that you accept H_0, and have proved by the failure to achieve a significant positive test result that there is no linkage. It is quite another thing to prove the absence of linkage, and can be statistically a very complicated problem.

It has been proposed (Morton, 1955) that one treat the test of linkage as a sequential likelihood ratio test (LRT) of a simple hypothesis, $\theta = \theta_1$. He proposed that one continue sampling new families until either one fulfills the criterion $Z(\theta_1) > 3$, in which case you would accept the hypothesis of linkage, or until $Z(\theta_1) < -2$, in which case you would reject the hypothesis of linkage. As long as $-2 < Z(\theta_1) < 3$, then no conclusions could be made. This concept has been extended to the general case, as described by Chotai (1984), such that the positive test is considered significant whenever $Z_{\max} > 3$, and the negative test is considered significant on $(\theta \mid Z(\theta) < -2)$, and the disease gene is said to be excluded from this region of the genome, where the lod scores fall below –2. The same criteria are routinely accepted for multipoint lod scores as well, though the theory is less well characterized in this case.

17.2 TWO POINT EXCLUSION MAPPING

Let us consider an example of exclusion mapping to see how it can be applied in practice. Go back to the dataset from the previous chapter (files MULTDIS1.*), and compute lod scores for the disease vs. marker 2 with MLINK, starting from $\theta = 0$, in steps of 0.01 up to $\theta = 0.5$. Then, examine the output using the LRP program. In this two-point analysis, there is clearly no significant positive test result, so one cannot do any estimation of the location of the putative disease gene. However, if you examine the lod scores, you will see that on the interval [0, 0.13), $Z(\theta) < -2$, so the disease gene can be excluded from being in this region. This exclusion covers a range of 13% recombination on either side of marker 2, which is a total genetic distance of 30.1 cM excluded.

Now, repeat the same analysis using marker 9 and the disease. In this case, there is a positive test result for linkage between these two loci, with $Z_{\max} = 4.698$ at $\theta = 0.15$. However, we can also note that $Z(\theta = 0) = -\infty$, so we also have a negative test result with the same marker. These two results can be reconciled if you go back to the original formulation of the sequential test as a test of each simple hypothesis, $\theta = \theta_1$. Thus, the hypothesis $\theta = 0$ is rejected, while the hypothesis $\theta = 0.15$ is accepted. The conclusion would be that there is linkage between the disease and marker 9, but that $\theta > 0$. In this case, the 3-unit-of-lod-score support interval for the estimate of θ would be $\theta \in (0.02, 0.43)2$, since $Z(\theta) > 1.698$ (= 4.698 – 3) at all values of $(\theta \mid 0.02 < \theta < 0.43)$. In general, exclusion mapping is most useful when there is no positive test result, since when there is a positive test result, one can develop a support interval for the location of the disease gene. Further exclusion information would not be of use in this situation, since the gene has already been localized to some degree, and now fine mapping can begin.

17.3 MULTIPOINT EXCLUSION MAPPING

Exclusion mapping is also routinely performed using multipoint analysis as well, based on the same criteria. In general, one can get a much greater exclusion map from multipoint analysis than one can get from two-point analysis, since under certain incorrect locus orders, obligate double recombinants can be quite prevalent, greatly lowering the multipoint lod score. Let us go back to the same example again, and look at the multipoint lod scores we generated in that analysis. If one were to look at all points with lod scores lower than –2, you would see that the regions from map positions [0, 0.162518] and [0.720445, 0.786475] could be excluded. Further, the disease gene has been mapped to the region (0.3856,0.6066), by the 3-unit-of-lod-score support criterion. Hence, we have the ability to exclude regions of the genome that are only 12 cM away from the region to which the disease was mapped, with such region covering a 7 cM range. The total exclusion region in this analysis is 23 cM in an analysis with a positive lod score of 9. This is impossible with two-point analysis in this pedigree set. We saw that you could exclude a 30 cM region with one two-point analysis, but when the two-point result contained a positive test result, the maximum exclusion was extremely small (< 1 cM). In fact, the two-point analysis with which the lod score of 10 was obtained allows for the exclusion of none of the genome. When there is no linkage, multipoint analysis can generally provide a much more complete exclusion map than two-point analysis.

17.4 MODEL ERRORS AND EXCLUSION MAPPING

It has been shown that using an incorrect model for your disease will not in general lead to an increased false positive rate (Clerget-Darpoux et al, 1986), although maximizing the lod score over models will (Weeks et al, 1990a). In other words, you will not spuriously obtain lod scores of 3 in the absence of linkage at a higher rate under the wrong model than the correct model. If there is linkage, however, there is lower power to detect it when the model parameters are misspecified. Further, the estimates of the recombination fraction are typically inflated.

Contrary to the lack of false positives, the false negative rate can be astronomical when an analysis is performed under an incorrect model. It is very simple to design cases where the disease can be "excluded" from its true location by the $Z(\theta) < -2$ criterion, when the analysis is done under an incorrect model. For this reason, when doing a linkage analysis with a complex disease, for which the model is not accurately known, it is not wise to do exclusion analysis, as the exclusion results obtained apply only to that specific model. You can only say that a given region was excluded if the analysis model were correct. Thus, we do not advocate any kind of exclusion analysis when the model is not known with a high degree of accuracy. There is of course an additional problem with exclusion mapping when there is the possibility of genetic heterogeneity, or diagnostic instability (see Part III). If there is linkage in only 20% of families, then summing the lod scores across families can easily lead to spurious exclusions. Similarly, if there is a significant rate of diagnostic uncertainty or instability, then it is again easy to induce false recombinants, leading to mistaken exclusion of the true disease map location.

EXERCISE 17

Let us consider again the disease pedigrees from chapter 16 (Files MULTDIS1.*). This time, however, repeat the same analysis using an incorrect model, assuming the disease to be autosomal recessive with 80% penetrance.

Next try analyzing the disease under the fully penetrant dominant model with the assumption that all unaffecteds are actually unaffected with only 75% certainty (see chapter 10). How does altering the model change your conclusions from the linkage analysis, including both positive and negative results?

18 Sex difference in recombination rates: Multipoint case!

In this chapter you will learn how to handle sex difference in recombination rates. You will be estimating them in CILINK, and then utilizing them in LINKMAP to help refine your estimates of gene location.

Gametes are produced in males and females by the extremely different processes of spermatogenesis and oogenesis. As a by-product of the different processes involved, the recombination rates in the two sexes are known to be quite different. In fact, with the exception of the telomeres, males tend to exhibit much lower rates of recombination than females. The precise biological reason for this is not important to the linkage analyst, but rather the quantification of the effect. If male and female recombination rates are really so different, then one should use this information in the linkage analysis to glean maximal information from the data at hand. One could obtain much finer recombination estimates, and gain higher power in multipoint analyses by using this information. The LINKAGE programs are equipped to estimate sex-specific recombination rates in ILINK and CILINK, and to utilize this information for fine mapping in LINKMAP and CMAP, as you will see below.

18.1 ESTIMATING SEX-SPECIFIC RECOMBINATION RATES

It is quite simple to estimate sex-specific rates of recombination in the ILINK or CILINK program. As an illustrative example, please reconsider the dataset from the locus ordering chapter. Call up the LCP program to analyze the appropriate pedigree and parameter files from chapter 14 (CEPH1.*). Then select the following options in order: *Three-generation pedigrees*, *CILINK*, *All Orders*, and *Varying sex difference*. The varying sex difference option means that separate male and female recombination fractions will be estimated for each interval. Then, set up the analysis exactly as you had done for the previous exercise, as follows:

```
Locus set [] : 1 2 3
Male recombination fractions [.1] : .1 .1
Female recombination fractions [.1] : .1 .1
```

Then, hit <Page Down> to save the analysis, and <Ctrl-Z> to exit the program. Run the analysis by typing *PEDIN* at the DOS prompt, and then examine the results in the LRP program, exactly as you did before. You should find the following results:

```
Order                 -2LN Like           Odds
--------------------------------------------
 .060 .060
3----1----2           -1.9418E+02       1.00E+00 <==
 .080 .020

 .060 .120
1----2----3           -1.8537E+02       8.21E+01
 .020 .060

 .060 .119
1----3----2           -1.6730E+02       6.90E+05
 .080 .060
```

In this analysis, there was even less significance than there was when the sex-difference in recombination rates was not allowed for.

18.2 CONSTANT FEMALE TO MALE MAP DISTANCE RATIO

Also, you will notice that in each case, the male recombination fraction estimates are presented on top, and the female recombination fraction estimates on the bottom. Note that under order 3-1-2, in interval 1, the female recombination rate is higher than the male rate, while in interval 2, the male rate is higher than the female rate. The same applies to order 1-3-2 as well. This is highly unlikely in reality, since the region is so small that one could reasonably expect the difference in recombination rates to be somewhat comparable. The programs allow you to fix such a restriction in the form of a constant female to male map distance ratio. What this means is that $x_{HALD}(\theta_f)/X_{HALD}(\theta_m) = R$, where R is held constant over the entire set of markers to be analyzed together. Now, instead of independently estimating four recombination fractions (two male and two female) for each locus order, only three free parameters are estimated jointly, two male θ's, and the constant female:male map distance ratio. In this way, the number of free parameters can be reduced, and the restriction is allowed for, that sex difference in recombination is held constant over the region under consideration, which is a reasonable thing to believe over a short map length. Further, in this way, you are using more data to estimate the constant ratio of recombination rates, making the estimated sex difference more precise. To do this, please repeat the above analysis, only choosing the *Constant sex difference* option, rather than the *Varying sex difference* option in LCP. Then, set up the following screen as follows:

```
 Locus set [] : 1 2 3
 Male recombination fractions [.1] : .1 .1
 Female/male distance ratio [1] : 1
```

Then, perform the analysis, and read the output into LRP. You should see the following results:

```
Order                    -2LN Like          Odds
---------------------------------------------------
 .076 .044
3----1----2              -1.9304E+02       1.00E+00 <==
 .064 .036

 .056 .124
1----2----3              -1.8529E+02       4.81E+01
 .024 .056

 .079 .101
1----3----2              -1.6630E+02       6.42E+05
 .061 .079
```

So, we can see that allowing for a constant sex difference in recombination rates even further reduces our power to order these three loci. From looking at the estimated recombination rates, the value of the constant sex ratio is not at all clear. Let us compute the value of this ratio for the order 3-1-2. In interval 1, we have $\theta_m = 0.076$, and $\theta_f = 0.064$. Converting these into map distances by Haldane mapping function (which in this case is mandatory, as the distance ratio is set to a constant in the program under this mapping function only!), we have $x_m = –0.5 \ln(1 – 2(0.076)) = 0.0824$; $x_f = –0.5 \ln[1 – 2(0.064)] = 0.0685$. From this, we can compute $R = x_f/x_m = 0.0685/0.0824 = 0.831$. Similarly, for interval 2, we have $R = x(0.036)/x(0.044) = 0.0374/0.0461 = 0.811$. Likewise, in the outer interval, $x_m(3,2) = 0.0824 + 0.0461 = 0.1285$; $x_f(3,2) = 0.0685 + 0.0374 = 0.1059$; $R = x_f/x_m = 0.1059/0.1285 = 0.824$. These are all very close to each other, and given the degree of rounding error present in these estimated θ's, and map distance conversions, they can be considered to be equal. In fact, a theoretical formulation of the female to male map distance can be written in one equation as $R = \ln(1 – 2\theta_f)/\ln(1 – 2\theta_m)$. Plugging in our values, for interval one, we get $\ln(1 – 2(0.064))/\ln(1 – 2(0.076)) = 0.831$, as above. Since we know that all these estimates of the female to male map distance ratio are approximate, subject to rounding error, let us use LRP to determine what value of R was really estimated by the program. To do this, call up LRP as before, only this time, select the Full format option. You should see the following screen for the constant sex ratio analysis:

```
 Initial Male Recomb. : 0.076 0.044
 Locus Order : 03-----01-----02
 Female Recombination : 0.064 0.036
 Constant Sex Ratio : 0.824

 Generalized LOD Score : +4.191800E+01

 Data : Valid
 Data Type : Autosomal

Maximum -2LN Likelihood : -1.930400E+02
 PTG : -6.665120E-05
 Number of Iterations : 8
 Likelihood Validity : Valid
 Gemini Exit Condition : Specified tolerance on normalized gradient met.
Iterated Parameter List : 1 1 1
 Final Parameter Values : +7.624180E-02 +4.373910E-02 +8.240760E-01
 Final Gradient Values : +0.000000E+00 +4.419420E-01 +0.000000E+00
```

This tells us that the estimated constant sex ratio was 0.824. Let us try and determine the female recombination fraction from the estimates of male recombination fraction and sex ratio. To do this (for interval one), one needs to first convert the male recombination fraction to map distance, according to $x_m = –0.5\ln(1 – 2(0.076)) = 0.0824$. Then, we know that $R = x_f/x_m$; so $x_f = Rx_m = (0.824)(0.0824) = 0.0679$. Converting this back to recombination fraction, you get $\theta_f = 0.5(1 – e^{-2(0.0679)}) = 0.0635$. To three decimal places, this estimate is 0.064, exactly that which was given by the CILINK program. So, you see how the rounding error caused all the deviations we observed in calculating the ratio for each interval. To see this, please do the same calculations for the second interval. You should again get results corresponding to those in the CILINK output file. Internally, the program estimates two parameters, the male recombination fractions, and the female to male map distance ratio (Haldane). From these values, it computes the corresponding female recombination fractions at the end. Hence, when we replicated this procedure exactly, the estimates presented in CILINK's output file are completely consistent.

Now, we need to look at our data again, to see if we had significant evidence in favor of either model for sex difference in recombination. To do this, consider the fixed order 3-1-2, and look at the values of –2LN Like under each of these three models of sex difference in recombination rate. You should have obtained the results summarized in table 18-1 in your analysis.

MODEL	df	-2LN Like	Δ(-2 LN Like)
Varying sex difference:	4	-194.18	0.00
Constant sex difference:	3	-193.04	1.14
No sex difference:	2	-192.94	1.24

Table 18-1: Results of analysis of sex difference in recombination rates with CILINK under locus order 3-1-2.

The difference in –2LN (Like) can be thought of as a chi-squared statistic with the number of degrees of freedom equal to the difference in degrees of freedom between the models being compared. Thus, in this study, a test of constant sex difference vs. no sex difference would have (3 – 2 =) 1 degree of freedom, and the value of the statistic would be (–192.94 – (–193.04)) = 0.10. Clearly this is insignificant. Likewise, if we wanted to test varying sex difference against no sex difference, the statistic would have (4 – 2 =) 2 degrees of freedom, and would have a value of (–192.94 – (–194.18)) = 1.24, which is again insignificant. The exact p-value, as computed with the CHIPROB program is 0.537944, which is completely insignificant, so you cannot reject the null hypothesis of no sex difference in this sample.

Please repeat the above analysis using all four loci in this dataset. At the end, you should have the results shown in table 18-2.

MODEL	df	-2LN Like	Δ(-2 LN Like)
Varying sex difference:	6	-328.91	0.00
Constant sex difference:	4	-326.04	2.87
No sex difference:	3	-325.94	2.97

Table 18-2: Results of analysis of sex difference in recombination rates with CILINK using all four loci jointly.

In this case, it is clear that the hypothesis of constant sex difference is not significantly better than that of no sex difference ($\chi^2_{(1)} = 0.10$), while that of varying sex difference is somewhat more supported ($\chi^2_{(3)} = 2.97$), but the p-value is still only 0.396277, so no significant evidence exists for sex difference in recombination rates in this pedigree set. It is interesting to examine more closely the results of the analysis with varying sex difference. The estimated recombination rates in this analysis were:

```
 .060 .020 .040
3----1----4----2
 .080 .021 .001
```

Computing the female to male map distance ratios for each intermarker region, we find $R_1 = 1.35$, $R_2 = 1.044$, $R_3 = 0.016$. In fact, the estimated recombination rate between loci 4 and 2 should have been 0, making this ratio 0. This shows that in this family set, every recombination event between markers 4 and 2 happened to have occurred in males. If you remember, this family had one hundred informative meioses, fifty in females, and fifty in males. Only four meioses showed a recombination between loci 4 and 2, and all of them seem to have occurred in males. Since this rate is so small anyway, it is not extremely rare to observe no recombinants by chance in a given sex. For this reason, it is often safer to estimate a constant sex ratio of map distances, to prevent having frequent estimates of θ = 0 in one sex, with a somewhat larger estimate in the other, since obviously if a recombination occurred in one sex between two markers, the recombination fraction must be greater than 0 in the other sex as well, since there is obviously some genetic distance between the markers.

18.3 SEX DIFFERENCE IN GENERAL PEDIGREE DATA

We now want to see how we can use information about sex difference in recombination rates in general pedigree data. Let us reconsider the autosomal dominant disease from before in two new pedigrees, structurally and phenotypically (at the trait locus) identical to those in figure 16-1, with different marker locus genotypes, as indicated in table 18-3, assuming the same markers have been typed here as well (make files MULTDIS2.* - Note that MULTDIS2.DAT should be identical to MULTDIS1.DAT as the markers and trait locus have the same parameters as before):

Ped	Ind	M1	M2	M3	M4	M5	M6	M7	M8
1	1	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
1	2	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	3	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	4	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	5	1 2	1 2	1 1	1 2	1 2	1 1	1 2	1 1
1	6	1 1	1 2	1 2	1 1	1 2	1 2	1 1	1 2
1	7	1 2	1 2	1 2	1 1	1 2	1 2	1 2	1 2
1	8	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	9	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	10	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	11	1 1	1 2	1 2	1 1	1 2	1 2	1 2	1 2
1	12	1 1	1 1	1 2	1 1	1 2	1 2	1 1	1 2
1	13	1 2	1 2	1 1	1 2	1 2	1 1	1 2	1 1
1	14	1 1	1 1	1 1	1 1	1 1	1 2	1 1	1 1
1	15	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
1	16	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
1	17	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
1	18	1 1	1 1	1 1	1 1	1 1	1 2	1 1	1 2
1	19	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	20	1 2	1 1	1 1	1 2	1 1	1 2	1 2	1 1
1	21	1 2	1 1	1 1	1 1	1 1	1 1	1 2	1 1
1	22	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
1	23	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
1	24	1 1	1 2	1 2	1 1	1 2	1 2	1 2	1 2
1	25	1 2	1 2	1 2	1 2	1 2	1 1	1 2	1 1
1	26	1 2	1 2	1 1	1 2	1 1	1 1	1 2	1 1
2	1	2 2	2 2	2 2	2 2	2 2	2 2	2 2	2 2
2	2	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	3	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	4	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	5	1 2	1 2	1 1	1 2	1 2	1 1	1 2	1 1
2	6	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	7	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	8	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	9	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	10	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	11	1 1	1 2	1 2	1 1	1 2	1 2	1 2	1 2
2	12	1 1	1 1	1 2	1 1	1 2	1 2	1 1	1 2
2	13	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	14	1 1	1 1	1 1	1 1	1 1	1 2	1 1	1 1
2	15	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
2	16	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	17	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	18	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	19	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1

2	20	1 2	1 1	1 1	1 2	1 1	1 1	1 2	1 1
2	21	1 2	1 1	1 1	1 1	1 1	1 1	1 2	1 1
2	22	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	23	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1
2	24	1 2	1 2	1 2	1 2	1 2	1 2	1 2	1 2
2	25	1 2	1 2	1 2	1 2	1 2	1 1	1 2	1 1
2	26	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1

Table 18-3: Genotypes for pedigrees with structure as shown in figure 16-1, to be entered in files MULTDIS2.*

Ignoring, for the moment, the disease, let us try and test whether there is a significant sex difference in recombination rates in these families. For the moment, let us only consider locus order 4-9-7, which we already know is the true order of these loci. Ideally, we would want to check for sex differences in recombination rates using the CEPH pedigrees, but in this case, it is not possible. Therefore, we will have to use the general pedigree version of ILINK, and a correspondingly reduced set of loci, in this case, three. Now, let us analyze these three loci in ILINK, using the *Specific order*, and each of the following options in turn: *No sex difference*, *Constant sex difference*, and *Varying sex difference*. Use starting values of *0.1* for all recombination fractions, and *1* for the female to male map distance ratio. When you finish running the analysis, you should examine the results in LRP, yielding the results shown in table 18-4.

MODEL	*n*	-2LN Like	Δ(-2 LN Like)
Varying sex difference:	4	181.90	0.00
Constant sex difference:	3	181.90	0.00
No sex difference:	2	181.25	-0.75

Table 18-4: Results (Incorrect) from first attempt at analysis of loci 4,9, and 7 on pedigrees from MULTDIS2.*; *n* = number of parameters estimated

Note that you have achieved an impossible result, in which the hypothesis of no sex difference has a higher likelihood than varying or constant sex difference, even though no sex difference is a nested hypothesis. This should tell you something is wrong. The other thing that should key you in to the fact that an error occurred is that the final estimates of the recombination fractions and female to male distance ratios are all the same as the starting values, meaning the program didn't do any maximization at all. ILINK has this peculiar property, that it does not always converge, depending on the initial values given to it. Whenever running this program, it is always advisable to examine the conditions under which the program finished the maximization of the likelihood. To examine the exiting conditions, call up the LRP program again, only this time select the option for *Full format*, rather than *Table format*. Look at the results there, and you will see the following message:

```
 Likelihood Validity : Not completely converged
 Gemini Exit Condition : Excessive cancellation in gradient
```

What this means is that the likelihood was not successfully maximized in this analysis, and that you should attempt the analysis again with different starting values. So, call up LCP, and set up another run with the starting values as follows:

No sex difference:

```
                     Locus order [] : 4 9 7
       Recombination fractions [.1] : 0.2 0.2
```

Constant sex difference:

```
                     Locus order [] : 4 9 7
  Male recombination fractions [.1] : 0.2 0.2
     Female/Male distance ratio [1] : 2
```

Varying sex difference:

```
                          Locus order [] : 4 9 7
   Male recombination fractions [.1] : 0.2 0.2
 Female recombination fractions [.1] : 0.05 0.05
```

Then, run the analysis, and examine the output in LRP. This time, first check the exit conditions and likelihood validity with the Full format option. This time the message should say the following:

```
 Likelihood validity : Valid
 Gemini exit condition : Specified tolerance on normalized gradient met
```

This should be the case for each sex difference option. Then, examine the likelihoods and you will see the results given in table 18-5.

MODEL	df	-2LN Like	Δ(-2 LN Like)
Varying sex difference:	4	180.52	0.00
Constant sex difference:	3	180.52	0.00
No sex difference:	2	181.25	0.63

Table 18-5: Correct results from analysis of loci 4, 9, and 7 on pedigrees from MULTDIS2.*

None of these comparisons are significant. So, for the purposes of testing on this set of data, there is nothing significant at all. However, there is always the option of using more loci together. On the PC it should be possible to analyze five loci together. So, try looking jointly at loci 3-6-4-9-7. In this analysis, when it is finished, the results should look like those given in table 18-6.

MODEL	df	-2LN Like	Δ(-2 LN Like)
Varying sex difference:	8	277.8569	0.0000
Constant sex difference:	5	277.8570	0.0001
No sex difference:	4	279.3242	1.4671

Table 18-6: Analysis of loci 3,6,4,9,7 from pedigrees in MULTDIS2.*

So, in this case, our test for constant sex difference vs. no sex difference is again insignificant, but has a p-value of only 0.2257, as opposed to the three-locus case, where the p-value was 0.4273. It turns out that if you can use all the loci jointly, the statistic becomes marginally significant, with an estimated female to male map distance ratio of about 2.113:1.

18.4 SEX DIFFERENCE IN RECOMBINATION FRACTION IN LINKMAP

It is very simple to take advantage of the sex differences in recombination fraction using the LINKMAP program. Analyze loci 3,6, and 4 vs. the disease, moving the disease across the entire length of the map, with 5 evaluations per interval. The map of this region was 3-(0.075)-6-(0.075)-4, under no sex difference; 3-(0.05)-6-(0.05)-4, with female to male map distance ratio of 2.113, and male thetas indicated for constant sex difference; 3-(0.05/0.10)-6-(0.05/0.10)-4, in format (θ_m/ θ_f), under varying sex difference. Let us reanalyze the data with LINKMAP now. First, call up LCP, and select the *LINKMAP* program, *All intervals* option, followed by *No sex difference*. Set up the next screen as:

```
                              Test loci [] : 1
                  Order of fixed loci [] : 3 6 4
      Recombination fractions [.1] : 0.075 0.075
 Number of evaluations in interval [5] : 5
```

Then, do the same thing with the *Constant sex difference* option, setting up the screen as follows:

```
                              Test loci [] : 1
                  Order of fixed loci [] : 3 6 4
      Male recombination fractions [.1] : 0.05 0.05
 Female/Male distance ratio [1] : 2.113
 Number of evaluations in interval [5] : 5
```

Finally, set up a third analysis with *Varying sex difference*, as follows:

```
                            Test loci [] : 1
                 Order of fixed loci [] : 3 6 4
      Male recombination fractions [.1] : 0.05 0.05
 Female recombination fractions [.1] : 0.10 0.10
 Number of evaluations in interval [5] : 5
```

Run these analyses, and when you have finished, you should obtain the results given in table 18-7.

			(Sex Averaged)	(No Sex Diff.)		(Sex Ratio=2.113)	
Thetas			Map Distance	Location	Lod	Location	Lod
			(in Morgans)	Scores	Scores	Scores	Scores
0.5	0.075	0.075		0	0	0	0
0.4	0.075	0.075	-0.80471	12.15294	2.638971	7.527224	1.634513
0.3	0.075	0.075	-0.45814	21.83398	4.741182	15.99249	3.472719
0.2	0.075	0.075	-0.25541	29.28252	6.358606	24.89760	5.406436
0.1	0.075	0.075	-0.11157	33.83958	7.348158	32.66120	7.092278
0	0.075	0.075	0	-4.0E+20	-8.7E+19	-4.0E+20	-8.7E+19
0.015	0.062	0.075	0.015229	45.33776	9.844952	45.46628	9.872858
0.03	0.048	0.075	0.030934	49.54470	10.75847	49.63480	10.77804
0.045	0.033	0.075	0.047157	52.06690	11.30616	52.12358	11.31847
0.06	0.017	0.075	0.063916	53.92788	11.71027	53.95502	11.71616
0.075	0	0.075	0.081257	55.45176	12.04117	55.45176	12.04117
0.075	0.015	0.062	0.096486	54.13072	11.75431	54.08710	11.74484
0.075	0.03	0.048	0.112193	52.46032	11.39159	52.37892	11.37391
0.075	0.045	0.033	0.128414	50.12484	10.88445	50.01056	10.85963
0.075	0.06	0.017	0.145174	46.10902	10.01242	45.96564	9.981292
0.075	0.075	0	0.162518	-4.0E+20	-8.7E+19	-4.0E+20	-8.7E+19
0.075	0.075	0.1	0.274090	33.83958	7.348158	32.66120	7.092278
0.075	0.075	0.2	0.417931	29.28252	6.358606	24.89760	5.406435
0.075	0.075	0.3	0.620664	21.83398	4.741181	15.99249	3.472719
0.075	0.075	0.4	0.967237	12.15293	2.638971	7.527224	1.634513

Table 18-7: LINKMAP results with and without allowing for sex difference in recombination rates. Disease vs. loci 3, 6, and 4.

In this example, the two situations, constant sex difference, and varying sex difference give identical results, since the recombination fractions are the same under both models in this case. However, you can see that there is a slight difference in the magnitude of some of the lod scores in the analyses with sex difference, and without, though in general, the differences are minimal. In general applications, however, one should always use the best supported model. If there is significant evidence for sex-difference in recombination fractions, then one should use this information when doing a LINKMAP analysis, as it will allow for a more accurate likelihood analysis. In any of the situations discussed thus far, the inclusion of sex specific recombination rates is quite simple and straightforward, so we will not go into specific details about how to include it in each possible application. Unless stated otherwise, all analyses to be discussed, using the LINKAGE programs can be adapted for sex-specific recombination rates with a minimum of additional effort.

EXERCISE 18

Go back to the set of CEPH pedigrees from exercise 14 (files USEREX14.*). Reanalyze the data allowing for sex difference in recombination under both the constant and varying sex difference options, using all five loci jointly.

Then, repeat the LINKMAP/ILINK analyses of sex difference in recombination fraction with our disease using the data in the pedigrees from chapter 16 (MULTDIS1.*). Is there strong evidence for sex difference in recombination fraction with loci 3-6-4-9-7? Which hypothesis is supported, constant or varying sex difference? What would happen if you were to combine the pedigrees in the two disease data sets with the dominant disease (MULTDIS1.* and MULTDIS2.*) and analyze them together? Is there more evidence for sex difference in recombination? Is there anything interesting about these two datasets?

19 Introduction to Interference

In this chapter, the concept of genetic interference will be introduced. While there is currently no overwhelming support for any specific model of interference in humans, it is believed to be present. Clearly, if an accurate model for positive interference were available, it should greatly increase the power of multipoint linkage analysis. In this chapter, we will introduce how one can allow for interference in 3-point analysis with the ILINK program, and will briefly discuss preliminary versions of the CLINKAGE programs which allow for interference according to certain models.

19.1 WHAT IS INTERFERENCE?

Interference is the phenomenon whereby crossovers do not occur independently along a chromosome, and the presence of a crossover at any given location affects the probability of finding another crossover in a nearby chromosomal region. It is a well known phenomenon in many species, eg. mice, drosophila, etc. Typically, such interference is positive, meaning that a crossover at one location decreases the probability of another crossover in a nearby region, perhaps due to a stearic hindrance, or other biochemical reaction. Researchers often postulate that positive interference exists in humans, though it has not yet been demonstrated by rigorous statistical examination (Sturt, 1975). The presence and characterization of such interference will be of great utility in linkage mapping. Presently, the LINKAGE programs perform multipoint likelihood calculations assuming absence of interference. If positive interference is present in humans, double recombinants in a small region are then accorded too large of a probability of occurrence, so allowing for it in the analysis could potentially allow greatly increased power for locus ordering, and accuracy in fine scale linkage mapping. The first problem, however, is to try and prove its existence, and develop a quantification of the phenomenon.

19.2 THREE-POINT ANALYSIS OF INTERFERENCE

The most simple and straightforward manner is based on three point analysis, as described in Ott (1991). Assume three colinear markers, A-B-C, and then estimate θ_{AB}, θ_{BC} and θ_{AC}, without restricting that there be no interference. One could reparametrize the analysis in terms of three parameters, as shown in table 19-1.

	Interval A-B		
Interval B-C	Recombinant	Non-Recombinant	Total
Recombinant	α	β	θ_{BC}
Non-Recombinant	γ	δ	$1 - \theta_{BC}$
Total	θ_{AB}	$1 - \theta_{AB}$	

Table 19-1: Reparametrization of probabilities of each meiosis type allowing for interference, where $\delta = 1 - \alpha - \beta - \gamma$

The ILINK program has an option to analyze three point data according to the above parametrization. Basically, the program estimates the probabilities α, β, and γ, with $\delta = 1 - \alpha - \beta - \gamma$. From these estimates, it determines the estimates of the three recombination fractions, $\theta_{AB} = \alpha + \gamma$; $\theta_{BC} = \alpha + \beta$; $\theta_{AC} = \beta + \gamma$, as should be clear from table 19-1 (Remember from chapter 14 that the recombination events in two intervals uniquely determine the third interval's recombination status in a three-point analysis). A more common, and meaningful parametrization of the results of this analysis is to express the results as estimates of θ_{AB}, θ_{BC}, and c, the coefficient of coincidence, which is a measure of the type and strength of interference.

Interference can be quantified in three point analysis by the coefficient of coincidence, $c = (\theta_{AB} + \theta_{BC} - \theta_{AC})/(2\theta_{AB}\theta_{BC})$. This quantity c can be interpreted as follows: if c = 0, then it is said that there is complete positive interference, meaning that a recombination in interval AB makes it impossible for a second recombination to occur in interval BC; if c = 1, there is no interference (and the Haldane mapping function applies), for then the recombination fractions are independent; if $0 < c < 1$, there is some positive interference, and presence of a recombination decreases the probability of further recombination to some degree in that immediate area; and if $c > 1$, there is negative interference, meaning that one crossover increases the probability of a second crossover in adjacent chromosomal regions. We are thus dealing with the three parameters, θ_{AB}, θ_{BC}, and c, and can perform a likelihood ratio test of the absence of interference, as $2\ln\frac{L(\hat{\theta}_{AB},\hat{\theta}_{BC},\hat{c})}{L(\hat{\theta}_{AB},\hat{\theta}_{BC},c=1)} \sim \chi^2_{(1)}$, where the θ's should be estimated separately in numerator and denominator.

Please consider the set of pedigrees from exercise 14, and use the LSP program to extract loci 3, 1, and 2 from the pedigree and parameter files you created in that chapter (userex14.*), in a form readable by ILINK, to estimate the recombination fractions θ_{31}, θ_{12}, and θ_{32}, allowing for the presence of interference. To do so, proceed as follows. First, call up the LSP program by typing *LSP* at the DOS prompt. Then, respond to its queries as follows:

```
Command [ILINK] : ILINK
Pedigree File [PEDIN.DAT] : USEREX14.PED
Parameter File     [DATAIN.DAT] : USEREX14.DAT
Number of Loci [] : 3
Locus Order                    [] : 3 1 2
Interference [0] : 1 (To allow interference)
Sex Difference           [0] : 0 (No sex difference)
Male Recombination Fractions [0.1] : 0.1 0.1 0.1
```

The LSP program will then construct new pedigree and parameter files for this analysis, with only the three specified loci in the new files. This program is called by the PEDIN.BAT files created in LCP in the performance of every linkage analysis done with LCP. However, as in this case, you can see where the LSP program can be a useful tool on its own, to prepare files for such analyses as this one. Your LSP-created parameter file (*DATAFILE.DAT*) and pedigree file (*PEDFILE.DAT*) are ready to be analyzed directly by the UNKNOWN and ILINK programs (YOU CANNOT USE LCP AT THIS POINT, since the files are named PEDFILE.DAT and DATAFILE.DAT!). The DATAFILE.DAT file should resemble the following:

```
3 0 0 3
0 0.00000000 0.00000000 0
 3 1 2

2 3
 0.53571430 0.41785710 0.04642857
3
 1 0 0
 0 1 0
 0 0 1

2 3
 0.34931510 0.54452060 0.10616440
3
 1 0 0
 0 1 0
 0 0 1

2 2
 0.74083770 0.25916230
3
 1 0 1
 0 1 1

0 1
 0.10000000 0.10000000 0.10000000
0
 1 1 1
```

The important features of this datafile relevant to the interference option are indicated on the last four lines. The fourth line from the bottom, *0 1,* has a 0 for *no sex difference*, and a *1* for *allow for interference*. The third line from the bottom contains starting values for the *three* recombination fractions. The second line from the bottom contains a *0,* since we do not wish to estimate other parameters for any of the loci, and the last line contains three *1*'s to tell the program to estimate all three recombination fractions in the manner described above. To use the files created by LSP to test and estimate interference, please call up the *UNKNOWN* program, followed by the *ILINK* program. Then, look at the FINAL.DAT file in your word processor. The file should look like the following:

```
CHROMOSOME ORDER OF LOCI :
 3 1 2
******************************************************
P VALUES:
 0.194 0.120 0.077
THETAS:
 0.270 0.197 0.314
******************************************************
-2 LN(LIKE) = 8.787537241097E+002
OTTS GENERALIZED LOD SCORE = 5.573634329867E+000
NUMBER OF ITERATIONS = 6
NUMBER OF FUNCTION EVALUATIONS = 33
PTG = -4.659531440578E-005
******************************************************
******************************************************
```

In this file, you are provided with estimates of *P-VALUES* and *THETAS*. First, the *THETAS* are respectively the estimates of θ_{31}, θ_{12}, and θ_{32}. The *P-VALUES* are not p-values in the statistical sense. They are merely the estimates of the parameters γ, β, and α from table 19-1, and can be used to recreate the recombination fraction estimates as described above. The other important value to compute in this case is *c*, the coefficient of coincidence. Using the formula given above, for this example, *c* = [0.270 + 0.197 – 0.314]/[2(.270)(.197)] = 1.44, indicating that in this example the estimated interference is negative, since 1.44 > 1. The next thing to consider is testing whether or not the evidence for interference is significant. Sometimes, only those values of ≤ 1 are tested against c = 1 (one-sided test). On the other hand, significant evidence for negative interference might be indicative of errors in marker typing in your pedigree data, or it might provide evidence for gene conversion at the middle locus. If one wanted to test for any deviation from c = 1, one would proceed as follows. The likelihood ratio test is of the form $-2\ln[L(\hat{\theta}_{AB},\hat{\theta}_{BC},c=1)/L(\hat{\theta}_{AB},\hat{\theta}_{BC},\hat{c})] \sim \chi^2_{(1)}$. The value of $-2\ln[L(\hat{\theta}_{AB},\hat{\theta}_{BC},\hat{c})]$is 878.75 from the FINAL.DAT file above. The value of $-2\ln[L(\hat{\theta}_{AB},\hat{\theta}_{BC},c=1)]$can be found by running ILINK on this locus order without allowing for interference. Perform this analysis, and you should find that $-2\ln[L(\hat{\theta}_{AB},\hat{\theta}_{BC},c=1)]$= 878.94, making our likelihood ratio statistic = 878.94 – 878.75 = 0.19, for a clearly non-significant p-value of 0.66. Hence, there is no evidence for interference in this sample. Terwilliger et al (1993b) examined a large number of such triples of loci in the CEPH Chromosome consortium data, and never found any significant evidence for interference in that dataset.

19.3 SEX-SPECIFIC INTERFERENCE ANALYSIS

Terwilliger et al. (1993b) also showed that when there is a sex difference in recombination rates, one must allow for it to validly test interference, as follows. They showed that the sex pooled coefficient of coincidence, c_T, is biased upward in general when there really is a sex-difference in recombination fractions. If we assume complete interference, c = 0 in both males and females, then they showed that c_T is asymptotically unbiased, but when there is absence of interference in both males and females separately, $c_T = 1 + \frac{(\theta_{f1} - \theta_{m1})(\theta_{f2} - \theta_{m2})}{(\theta_{f1} + \theta_{m1})(\theta_{f2} + \theta_{m2})}$ and therefore c_T is asymptotically biased, and inconsistent (as opposed to the consistent result when c = 0). Further, asymptotically $L(\hat{\theta}_1,\hat{\theta}_2,c=1) < L(\hat{\theta}_1,\hat{\theta}_2,c=\hat{c}_T(\neq 1))$ so $\frac{L(\hat{\theta}_1,\hat{\theta}_2,\hat{c}_T)}{L(\hat{\theta}_1,\hat{\theta}_2,c=1)} \rightarrow \infty$ as $n \rightarrow \infty$ and the chi-square test fails.

In general, c_T is always greater than the sex specific coefficients of coincidence. For example, if one estimated c_T = 1, we then know that $c_T > c_m$ (likewise for c_f). So when one estimates that there is no interference from sex-pooled data, and there is a sex difference in recombination fraction, this implies there is positive interference in each sex separately.

Since there is usually significant evidence for sex differences in recombination fraction, it is imperative to allow for sex difference in the recombination fractions (and coefficient of coincidence) in your ILINK estimation and testing for the presence of interference in your data. Let us reconsider the same dataset as above, loci 3-1-2 in USEREX14.*. And let us use LSP to extract these loci and prepare the files for the interference analysis under the different sex difference options. Repeat everything exactly as before, only now when you are prompted with *Sex Difference [0]:*, enter a *1* to specify a constant sex difference in recombination. Then, analyze the data with ILINK as described before. The FINAL.DAT file should resemble the following:

```
CHROMOSOME ORDER OF LOCI :
 3 1 2
******************************************************
P VALUES:
 0.171 0.087 0.000
FEMALE:
 0.304 0.164 0.095
THETAS:
 0.171 0.087 0.258
FEMALE:
 0.399 0.259 0.469
CONSTANT FEMALE/MALE DIST RATIO :
3.821
******************************************************
-2 LN(LIKE) = 8.730176055538E+002
OTTS GENERALIZED LOD SCORE = 6.819214051328E+000
NUMBER OF ITERATIONS = 14
NUMBER OF FUNCTION EVALUATIONS = 88
PTG = -6.854390155713E-004
******************************************************
******************************************************
```

In this situation, the estimated coefficient of coincidence is c_m = [0.171 + 0.087 – 0.258] / [2(0.171)(0.087)] = 0, indicating complete positive interference in males, and c_f = [0.399 + 0.259 – 0.469]/ [2(0.399)(0.259)] = 0.914, indicating very slight positive interference in females, assuming a constant sex ratio of 3.821. However, it is probably more appropriate to assume a varying sex ratio, since the sex ratio forces some constraints on the sex-specific coefficients of coincidence, so try analyzing the data with *Sex difference [0] :* set to *2* in LSP. Then, reanalyzing the data should yield the following FINAL.DAT file:

```
CHROMOSOME ORDER OF LOCI :
 3 1 2
******************************************************
P VALUES:
 0.146 0.107 0.000
FEMALE:
 0.348 0.115 0.130
THETAS:
 0.146 0.107 0.252
FEMALE:
 0.479 0.245 0.463
FEMALE/MALE DIST RATIO :
 9.161 2.802 3.701
******************************************************
-2 LN(LIKE) = 8.726082665176E+002
OTTS GENERALIZED LOD SCORE = 6.908100708333E+000
NUMBER OF ITERATIONS = 14
NUMBER OF FUNCTION EVALUATIONS = 117
PTG = -7.623469083546E-004
******************************************************
******************************************************
```

In this example, the value of c_m = [0.146 + 0.107 – 0.252]/[2(0.146)(0.107)] = 0.03, and c_f = [0.479 + 0.245 – 0.463]/[2(0.479)(0.245)] = 1.11. Again, there is almost complete positive interference in males, and slightly negative interference in females. We still must test the hypothesis of interference under each of these models, as before. So, we will need to use ILINK to compute the –2ln(LIKE) for this locus order under the two different sex difference options. The results of this analysis should show that for *Constant sex difference*, –2ln(LIKE) = 873.45, and for *Varying sex difference*, –2ln(LIKE) = 873.15. Therefore, subtracting off the values obtained with interference, our chi-square statistic for *Constant sex difference* = 873.45 – 873.02 = 0.43 (p = 0.51); *Varying sex difference* = 873.15 – 872.61 = 0.54 (p = 0.46). Thus we have still got no significant evidence for interference in this dataset, when sex difference in recombination fraction is properly allowed for.

In this chapter, the concept of interference was introduced on a basic level, and testing and estimating interference from three-point data was explained using the ILINK program of the LINKAGE package. To date, no conclusive evidence of interference has been found in man using this approach, and this can be due to either a lack of interference in humans, or simply that our sample sizes are too small to detect what interference may exist. Ott (1991) determined that for Kosambi level interference, with 3 equally spaced markers ($\theta = 0.15$), and fully informative phase known data, 847 meioses would be required to reject the null hypothesis of no interference at the 0.05 level with 80% power. With this in mind, it is not surprising that the results of the small analyses done in this chapter proved non-significant, whether or not interference is actually present. There are more sophisticated methods currently under development by Weeks et al (1991) to handle more than three loci at a time in a test for interference in humans. Their preliminary results suggest that the Sturt mapping function (Sturt, 1975) may be the best fitting model for interference in humans, and they have potentially significant evidence for interference of this nature in one six-point analysis, but further investigations will be necessary to prove that interference is a general phenomenon in human genetics. They are developing a modified version of the CILINK program, called CINTMAX (Weeks et al, 1991), which is capable of analyzing CEPH pedigree data under a variety of possible models of interference, which may become a useful tool in gene mapping in the future.

EXERCISE 19

Consider the data from MULTDIS2.*. Is there any evidence of interference in this dataset? Since Ott (1991) reported that the optimum intermarker spacing to detect interference would be approximately $\theta = 0.15 – 0.20$, assuming phase known data (which we have in this dataset for the most part). Choose all ordered triples of loci with $0.125 \leq \theta \leq 0.225$ (to allow us to have more possible triples to consider that are nearly optimal) for both adjacent θ's, assuming the sex-averaged map. 5-(0.075)-2-(0.075)-8-(0.225)-3-(0.075)-6-(0.075)-4-(0.075)-9-(0.075)-7. Then, test for the presence of interference on each such triple of loci, and report the value of the appropriate χ^2 statistic, and the coefficient of coincidence. Repeat this analysis under *no sex difference*, *constant sex difference*, and *varying sex difference*, providing the values of c_m and c_f separately, where applicable. What are your conclusions about interference in this dataset?

20 Solutions to Part II Problems

EXERCISE 14

The output from your CILINK analysis should resemble the results shown in Table 20-1. In order to order the loci conclusively, the best order must be at least 1000 times more likely than the second best order. In this case, you can see that the first four orders are all about equally well supported, and none of them can be excluded. Still, the remaining 56 possible orders can be excluded, as they are more than 1000 times less likely than the best order. Looking closely at the set of orders which cannot be excluded, you will notice that they all involve 3 basic groupings, as follows (3,4) – (5,1) – 2. Each order in the set of possible orders represents a simple inversion of the order of loci within one of the groups. All possible orders involving mere flips of 3 and 4 or 5 and 1 are approximately equally supported. Hence, the only firm conclusion is that the order is (3,4) – (5,1) – 2 with odds of over 1000:1. You will notice that LCP presents the user with an option *Inversions of adjacent loci*. This option would allow the user to see if flipping any pair of loci could increase the likelihood significantly, or to see if all inversions of adjacent loci are still excluded from being possible orders. If you wished to run this analysis with ILINK, instead of CILINK, it would be possible, only if your computer had the capacity to compile the program with the constant *maxneed* set to 782, and *maxhap* = 108 (See Appendix B for more information about program constants).

```
    Locus Order           Intermarker Θ's       -2LN Like       Odds

3----4----5----1----2  .049 .251 .049 .207  -1.0395E+02   1.00E+00 *
4----3----5----1----2  .058 .234 .049 .207  -1.0380E+02   1.08E+00 *
3----4----1----5----2  .052 .239 .036 .231  -1.0339E+02   1.33E+00 *
4----3----1----5----2  .056 .224 .035 .231  -1.0292E+02   1.68E+00 *
5----1----2----3----4  .049 .199 .350 .060  -8.9534E+01   1.35E+03
5----1----2----4----3  .049 .200 .365 .050  -8.8096E+01   2.78E+03
5----1----4----3----2  .046 .215 .055 .348  -8.7737E+01   3.32E+03
1----5----2----3----4  .036 .212 .347 .061  -8.7002E+01   4.80E+03
5----1----3----4----2  .045 .202 .051 .361  -8.5734E+01   9.05E+03
1----5----2----4----3  .036 .213 .363 .051  -8.5543E+01   9.95E+03
1----5----4----3----2  .044 .229 .054 .352  -8.4005E+01   2.15E+04
1----5----3----4----2  .045 .215 .051 .365  -8.2188E+01   5.33E+04
4----1----5----3----2  .221 .044 .122 .247  -8.0629E+01   1.16E+05
4----5----3----1----2  .204 .041 .140 .211  -7.9706E+01   1.84E+05
4----1----5----2----3  .249 .035 .214 .258  -7.7912E+01   4.52E+05
4----5----1----2----3  .264 .047 .200 .263  -7.7315E+01   6.09E+05
5----3----4----1----2  .207 .051 .240 .212  -7.7158E+01   6.59E+05
5----4----3----1----2  .226 .052 .227 .212  -7.6256E+01   1.03E+06
3----4----1----2----5  .049 .252 .188 .248  -7.4951E+01   1.99E+06
4----3----1----2----5  .060 .237 .187 .248  -7.4873E+01   2.06E+06
1----2----5----3----4  .182 .268 .231 .0581  7.2637E+01   6.32E+06
1----2----5----4----3  .182 .268 .255 .049  -7.1152E+01   1.33E+07
3----5----1----2----4  .166 .055 .190 .375  -7.0649E+01   1.71E+07
3----5----1----4----2  .135 .051 .193 .342  -7.0533E+01   1.81E+07
4----1----3----5----2  .213 .103 .044 .269  -6.8821E+01   4.26E+07
3----5----4----1----2  .064 .177 .229 .209  -6.7046E+01   1.03E+08
1----4----3----5----2  .204 .051 .201 .294  -6.4836E+01   3.12E+08
1----2----3----4----5  .197 .369 .053 .241  -6.4442E+01   3.80E+08
1----2----4----3----5  .197 .385 .053 .222  -6.3806E+01   5.22E+08
3----1----5----2----4  .225 .034 .211 .375  -6.2956E+01   7.99E+08
1----3----4----5----2  .194 .048 .223 .296  -6.2816E+01   8.57E+08
3----1----5----4----2  .185 .040 .209 .345  -6.2414E+01   1.05E+09
1----2----3----5----4  .186 .287 .065 .213  -6.1752E+01   1.46E+09
5----1----4----2----3  .044 .212 .317 .254  -6.0779E+01   2.37E+09
1----3----5----4----2  .125 .036 .183 .359  -5.9809E+01   3.85E+09
1----5----3----2----4  .048 .142 .216 .357  -5.9713E+01   4.04E+09
5----3----1----4----2  .051 .116 .189 .348  -5.8925E+01   6.00E+09
5----1----3----2----4  .046 .169 .199 .348  -5.7962E+01   9.71E+09
1----5----4----2----3  .042 .229 .320 .253  -5.6793E+01   1.74E+10
5----3----1----2----4  .056 .145 .194 .376  -5.6600E+01   1.92E+10
1----4----3----2----5  .213 .054 .333 .281  -5.4754E+01   4.83E+10
1----4----5----3----2  .196 .175 .069 .304  -5.3920E+01   7.32E+10
4----1----2----5----3  .271 .184 .259 .113  -5.3479E+01   9.13E+10
```

1----2----4----5----3	.198	.380	.196	.0651	5.3306E+01	9.96E+10
4----1----2----3----5	.270	.186	.263	.101	-5.2910E+01	1.21E+11
1----3----4----2----5	.202	.051	.345	.283	-5.2858E+01	1.25E+11
5----4----1----3----2	.221	.196	.150	.245	-4.7971E+01	1.43E+12
5----4----1----2----3	.228	.234	.201	.265	-4.7799E+01	1.56E+12
3----1----2----5----4	.221	.171	.246	.279	-4.7184E+01	2.13E+12
4----5----1----3----2	.100	.100	.100	.100	-4.6881E+01	2.47E+12
1----3----5----2----4	.119	.043	.246	.362	-4.6853E+01	2.51E+12
4----1----3----2----5	.224	.141	.218	.256	-4.5488E+01	4.96E+12
3----1----4----5----2	.171	.174	.209	.285	-4.3895E+01	1.10E+13
3----1----2----4----5	.220	.185	.363	.261	-3.7472E+01	2.73E+14
1----4----5----2----3	.197	.209	.258	.251	-3.6318E+01	4.86E+14
1----3----2----5----4	.175	.195	.246	.269	-3.6267E+01	4.99E+14
3----1----4----2----5	.170	.187	.327	.281	-3.4021E+01	1.53E+15
1----4----2----3----5	.218	.321	.275	.111	-3.2839E+01	2.77E+15
1----4----2----5----3	.218	.323	.283	.102	-3.2614E+01	3.10E+15
1----3----2----4----5	.181	.203	.335	.261	-2.9718E+01	1.32E+16

Table 20-1: Solutions to Exercise 14 (Files USEREX14.*). Orders followed by * are within 3 lod units of the best order, and cannot be excluded.

EXERCISE 15

The results of the CMAP analysis are presented in table 20-2, including multipoint lod scores and map distance from marker 1. In this analysis, it is impossible to localize new locus 4 to any one map interval, since the three unit support interval extends throughout the region between loci 1 and 3, and therefore cannot be uniquely placed with the required odds of 1000:1.

Locus Order	Loc. Score	-2LN Like	Odds		Lod Score	Map Distance
4====1----2----3						
.500 .040 .090	+0.0000E+00	-1.8316E+02	7.13E+28		0.00	-∞
.400 .040 .090	+3.4842E+01	-2.1800E+02	1.94E+21		7.57	-0.805
.300 .040 .090	+6.3905E+01	-2.4707E+02	9.46E+14		13.88	-0.458
.200 .040 .090	+8.8455E+01	-2.7162E+02	4.42E+09		19.21	-0.255
.100 .040 .090	+1.0877E+02	-2.9193E+02	1.71E+05		23.62	-0.112
.000 .040 .090	+6.6542E+01	-2.4970E+02	2.53E+14		14.45	0
1====4====2----3						
.000 .040 .090	-∞	∞	∞	-∞	0	
.008 .033 .090	+1.3115E+02	-3.1431E+02	2.37E+00		28.48	0.008
.016 .025 .090	+1.3281E+02	-3.1597E+02	1.03E+00		28.84	0.016
.024 .017 .090	+1.3287E+02	-3.1604E+02	1.00E+00	<==	28.85	0.025
.032 .009 .090	+1.3134E+02	-3.1451E+02	2.15E+00		28.52	0.033
.040 .000 .090	-∞	∞	∞	-∞	0.042	
1----2====4====3						
.040 .000 .090	-∞	∞	∞	-∞	0.042	
.040 .018 .075	+1.2875E+02	-3.1191E+02	7.87E+00		27.96	0.060
.040 .036 .058	+1.2769E+02	-3.1085E+02	1.34E+01		27.72	0.079
.040 .054 .040	+1.2398E+02	-3.0714E+02	8.53E+01		26.92	0.099
.040 .072 .021	+1.1595E+02	-2.9911E+02	4.73E+03		25.18	0.120
.040 .090 .000	-∞	∞	∞	-∞	0.141	
1----2----3====4						
.040 .090 .000	-∞	∞	∞	-∞	0.141	
.040 .090 .100	+8.6796E+01	-2.6996E+02	1.01E+10		18.85	0.253
.040 .090 .200	+7.4592E+01	-2.5775E+02	4.52E+12		16.20	0.396
.040 .090 .300	+5.5432E+01	-2.3859E+02	6.54E+16		12.04	0.599
.040 .090 .400	+3.0787E+01	-2.1395E+02	1.47E+22		6.69	0.946
.040 .090 .500	+0.0000E+00	-1.8316E+02	7.13E+28		0.00	∞

Table 20-2 : Results of CMAP analysis of CEPH1.*; Assuming locus order 1-(0.04)-2-(0.09)-3; adding new locus 4 to this map.

When you attempt to order the four loci with CILINK, the results should be similar to those in table 20-3. In this case, the odds for order were relatively quite good. Order 3-1-4-2 is the best order (consistent with our CMAP results of chapter 15), but the second best order is only 133 times less likely, so it cannot be excluded. That order is 1-4-2-3. All other orders can be excluded, leaving us with a definitive locus order of 1-4-2, with 3 either proximal to 1 or distal to 2. Again, however, we are left with no additional power to order loci 1, 2, and 3. The odds for ordering are identical, at 133:1 in favor of 3-1-2. However, we were able to conclusively order loci 1-4-2 relative to each other. You may verify with a three-point CILINK run that this order would still have been uniquely determined without including locus 3 in the analysis. Since all markers are typed, phase known, and fully informative in these families, the addition of further markers will not help in ordering the original loci. The only situation in which the addition of further markers can help is when some markers are phase unknown in some meioses, or markers are informative in some meioses, but not in others. In this case, neither situation obtained, and thus no net change in odds for ordering was possible.

Locus Order	Thetas	-2LN Like	Odds	
3----1----4----2	.070 .020 .020	-3.2594E+02	1.00E+00	<==
1----4----2----3	.020 .020 .089	-3.1616E+02	1.33E+02	
1----2----4----3	.040 .020 .070	-3.1196E+02	1.09E+03	
3----1----2----4	.070 .040 .020	-3.1196E+02	1.09E+03	
3----4----1----2	.070 .020 .040	-3.1196E+02	1.09E+03	
4----1----2----3	.020 .040 .090	-3.0218E+02	1.44E+05	
1----3----4----2	.070 .070 .020	-2.9482E+02	5.73E+06	
4----1----3----2	.020 .070 .090	-2.8505E+02	7.61E+08	
1----4----3----2	.020 .070 .090	-2.8505E+02	7.61E+08	
1----3----2----4	.070 .090 .020	-2.8505E+02	7.61E+08	
4----3----1----2	.070 .070 .040	-2.8084E+02	6.22E+09	
1----2----3----4	.048 .093 .078	-2.7081E+02	9.37E+11	

Table 20-3 : CILINK results for files CEPH1.*; Loci 1, 2, 3, and 4.

For the data in files USEREX14.*, the results from the CILINK analysis of loci 1, 2, and 3 are given in table 20-4. In this case, using only the three loci, it is impossible to exclude any of the possible locus orders at the 1000:1 level.

Locus Order	Thetas	-2LN Like	Odds	
1----2----3	.197 .294	-2.4795E+01	1.00E+00	<==
3----1----2	.249 .191	-2.4725E+01	1.04E+00	
1----3----2	.187 .226	-1.5197E+01	1.21E+02	

Table 20-4 : CILINK results for files USEREX14.*; Loci 1, 2, and 3.

So, let us try adding in locus 4 to the analysis, and see if it helps us order loci 1, 2, and 3. In this case, for the four loci jointly, the CILINK results are shown in table 20-5. When all four loci are analyzed together, we can see that the best 4-locus order has order 3-1-2 within it. The best order with 1-2-3 in it has odds of 145:1 against it, and the best order with 1-3-2 in it has odds of 2220:1, and can thus be excluded. In contrast to the situation in files CEPH1.*, this example has a great deal more uncertainty in it, and therefore, we can gain locus ordering ability by adding an additional locus as a sort of nuisance parameter.

Locus Order	Thetas	-2LN Like	Odds	
3----4----1----2	.049 .254 .210	-6.2575E+01	1.00E+00	<==
4----3----1----2	.059 .239 .209	-6.2475E+01	1.05E+00	
1----2----3----4	.199 .379 .058	-5.2629E+01	1.45E+02	
1----2----4----3	.199 .400 .051	-5.1394E+01	2.68E+02	
1----4----3----2	.217 .053 .351	-4.7168E+01	2.22E+03	
1----3----4----2	.206 .051 .365	-4.5381E+01	5.42E+03	
4----1----2----3	.272 .197 .266	-3.5169E+01	8.94E+05	
4----1----3----2	.223 .147 .244	-3.4711E+01	1.12E+06	
3----1----2----4	.225 .188 .391	-2.7989E+01	3.24E+07	
3----1----4----2	.170 .191 .347	-2.6190E+01	7.96E+07	
1----3----2----4	.183 .206 .356	-2.0028E+01	1.73E+09	
1----4----2----3	.221 .321 .253	-1.8631E+01	3.49E+09	

Table 20-5 : CILINK results for files USEREX14.*; Loci 1, 2, 3, and 4.

There is still no conclusive ordering of the three loci, so we should go to analyzing all five together, which we did in chapter 14. The best order had 3-1-2 in it, the best order with order 1-2-3 in it had odds of 1350:1 against it, and the best 5-locus order with 1-3-2 in it had odds of 3320:1 against it. Thus, we can establish the order 3-1-2 with the required 1000:1 odds criterion only by adding in the additional loci 4 and 5. Now, we need to compute the intermarker recombination fractions for this locus order from the 5-point ordering estimates, 3-(0.049)-4-(0.251)-5-(0.049)-1-(0.207)-2. We will need to find the recombination between loci 3 and 1 under this scenario. To do this, call up the MAPFUN program, which was introduced in chapter 15, and select the *MS* (summing Θ's) option. Then, enter successively the values of $\theta_{3,4} = 0.049$, $\theta_{4,5} = 0.251$, and $\theta_{5,1} = 0.049$. The program will then give the result that $\theta_{3,1} = 0.2974$ assuming the Haldane mapping function. This could also have been done by hand, by converting all the recombination fractions into map distance by the Haldane function, $x = –½ \ln(1-2\theta)$, adding the three map distances together, and then converting the total map distance back into a recombination fraction by the relation $\theta = ½(1 – e^{-2x})$. The result should come out the same, $\theta \approx 0.297$. Now, you can do the CMAP analysis requested with locus order 3-(0.297)-1-(0.207)-2. Try adding locus 4 and 5 to this constant map of markers with the CMAP program as an exercise (in general, since we used them to order the loci 3-1-2, we wouldn't bother with CMAP analysis at this point, but do it anyway, as an exercise with CMAP. The results are shown in Table 20-6 for locus 4, and table 20-7 for locus 5.

```
Locus Order/θ's     Loc. Score     -2LN Like       Odds      Lod Score   Map Distance
-------------------------------------------------------------------------------------
4====3----1----2
 .500 .297 .207    +0.0000E+00   -2.4286E+01   1.88E+08       0.00           -∞
 .400 .297 .207    +9.9910E+00   -3.4277E+01   1.27E+06       2.17         -0.805
 .300 .297 .207    +2.0518E+01   -4.4804E+01   6.58E+03       4.46         -0.458
 .200 .297 .207    +2.9457E+01   -5.3743E+01   7.54E+01       6.40         -0.255
 .100 .297 .207    +3.5852E+01   -6.0137E+01   3.08E+00       7.79         -0.112
 .000 .297 .207    -1.9271E+01   -5.0145E+00   2.87E+12      -4.18          0.000
3====4====1----2
 .000 .297 .207    -∞            ∞             ∞           -∞             0.000
 .059 .270 .207    +3.8103E+01   -6.2388E+01   1.00E+00 <==   8.27          0.063
 .119 .234 .207    +3.5723E+01   -6.0008E+01   3.29E+00       7.76          0.136
 .178 .185 .207    +2.9975E+01   -5.4261E+01   5.82E+01       6.51          0.220
 .238 .113 .207    +1.6710E+01   -4.0996E+01   4.42E+04       3.63          0.323
3----1====4====2
 .297 .000 .207    -∞            ∞             ∞           -∞          0.450
 .297 .041 .181    -3.8826E+01   +1.4540E+01   5.07E+16      -8.43          0.493
 .297 .083 .149    -3.0996E+01   +6.7104E+00   1.01E+15      -6.73          0.541
 .297 .124 .110    -3.7932E+01   +1.3646E+01   3.24E+16      -8.24          0.593
 .297 .166 .062    -6.4530E+01   +4.0244E+01   1.93E+22     -14.01          0.652
 .297 .207 .000    -∞            ∞             ∞           -∞          0.718
3----1----2====4
 .297 .207 .000    -∞            ∞             ∞           -∞          0.718
 .297 .207 .100    -4.4800E+01   +2.0514E+01   1.00E+18      -9.72          0.830
 .297 .207 .200    -1.2589E+01   -1.1697E+01   1.02E+11      -2.73          0.973
 .297 .207 .300    -3.5960E-01   -2.3926E+01   2.25E+08      -0.08          1.176
 .297 .207 .400    +2.6670E+00   -2.6953E+01   4.95E+07       0.58          1.523
 .297 .207 .500    +6.6360E-01   -2.4949E+01   1.35E+08       0.14             ∞
```

Table 20-6 : CMAP Results of locus 4 vs. map 3-(0.297)-1-(0.207)-2

```
Locus Order/θ's      Loc. Score     -2LN Like       Odds      Lod Score    Map Distance
──────────────────────────────────────────────────────────────────────────────────────
5====3----1----2
 .500 .297 .207    +0.0000E+00   -2.4285E+01   2.65E+08       0.00          -∞
 .400 .297 .207    +6.0046E+00   -3.0290E+01   1.32E+07       1.30        -0.805
 .300 .297 .207    +1.2257E+01   -3.6542E+01   5.78E+05       2.66        -0.458
 .200 .297 .207    +1.7746E+01   -4.2032E+01   3.71E+04       3.85        -0.255
 .100 .297 .207    +2.1638E+01   -4.5923E+01   5.31E+03       4.70        -0.112
 .000 .297 .207    -3.5068E+01   +1.0782E+01   1.09E+16      -0.76         0.000

3====5====1----2
 .000 .297 .207     -∞          ∞             ∞          -∞          0.000
 .059 .270 .207    +2.8783E+01   -5.3069E+01   1.49E+02       6.25         0.063
 .119 .234 .207    +3.2422E+01   -5.6707E+01   2.42E+01       7.04         0.136
 .178 .185 .207    +3.5563E+01   -5.9849E+01   5.02E+00       7.72         0.220
 .238 .113 .207    +3.8791E+01   -6.3077E+01   1.00E+00 <==   8.42         0.323

3----1====5====2
 .297 .000 .207     -∞          ∞             ∞          -∞          0.450
 .297 .041 .181    +3.2667E+01   -5.6953E+01   2.14E+01       7.09         0.493
 .297 .083 .149    +2.8794E+01   -5.3080E+01   1.48E+02       6.25         0.541
 .297 .124 .110    +2.0955E+01   -4.5241E+01   7.47E+03       4.55         0.593
 .297 .166 .062    +4.9642E+00   -2.9250E+01   2.22E+07       1.08         0.652
 .297 .207 .000     -∞          ∞             ∞          -∞          0.718

3----1----2====5
 .297 .207 .000     -∞          ∞             ∞          -∞          0.718
 .297 .207 .100    -5.1220E-01   -2.3773E+01   3.42E+08      -0.11         0.830
 .297 .207 .200    +9.7451E+00   -3.4031E+01   2.03E+06       2.12         0.973
 .297 .207 .300    +1.0337E+01   -3.4623E+01   1.51E+06       2.24         1.176
 .297 .207 .400    +6.2507E+00   -3.0536E+01   1.16E+07       1.36         1.523
 .297 .207 .500    +1.0000E-04   -2.4286E+01   2.65E+08       0.00           ∞
```

Table 20-7 : CMAP Results of locus 5 vs. map 3-(0.297)-1-(0.207)-2

The CMAP results for locus 4 are somewhat interesting. The entire region between loci 1 and 2 is excluded by the Z(x) < –2 criterion, as is much of the region to the right of locus 2. The maximum lod score of 8.27 occurs between loci 3 and 1, very close to locus 3, but the 3-unit support interval extends for about 25 cM on either side of locus 3, giving us no power to place locus 4 uniquely in an interval of this map. It is interesting to note that when the disease is unlinked to the map of markers (Θ=0.500) on the right hand side, the lod score is not 0, but 0.14. This is due to the fact that there is often some rounding error in that direction, and the recombination fraction at which that lod score was computed was not exactly 0.500, but slightly smaller. This phenomenon is similar to that observed when 4-(0.000)-3-(0.297)-1-(0.207)-2 has a lod score of –4.18, and 3-(0.000)-4-(0.297)-1-(0.207)-2 has a lod score of –∞, even though they are ostensibly the same point. This is, as well due to the fact that the first lod score is computed at θ > 0.000, and the value of θ we see in the output is just rounded to 3 decimal places. For this reason, we always use the value with θ = 0.500 to the left as our normalizing value in computing lod or location scores. The analysis with locus 5 allows for no exclusion region, but does show evidence for linkage, with maximum lod score of 8.42 between loci 3 and 1, very close to locus 1, with 3 unit support interval extending between loci 3 and 2, again not allowing us to order the locus with 1000:1 odds.

If we look back at our 4-point CILINK analysis with loci 1, 2, 3, and 4, we will see that the order 3-4-1-2 was the best order, 4-3-1-2 had odds of 1.05:1 against it (here as well, it was within the support interval), 3-1-2-4 had odds of 32,400,000:1 against it, and 3-1-4-2 had odds of 79,600,000:1 against it. This is compatible with the exclusion results we obtained in the CMAP analysis. Since the recombination fractions were estimated separately in each order with CILINK, the odds are naturally somewhat better than with CMAP, which gave odds against the same orders of 3.08:1, 49,500,000:1, and 1,010,000,000,000,000:1 respectively. In general, unless the map is known with a high degree of certainty, it is more conservative and reliable to use CILINK, rather than CMAP for such ordering problems. Similar results would be found if you were to run CILINK to order loci 1, 2, 3, and 5, and compare those results with the output from CMAP.

EXERCISE 16

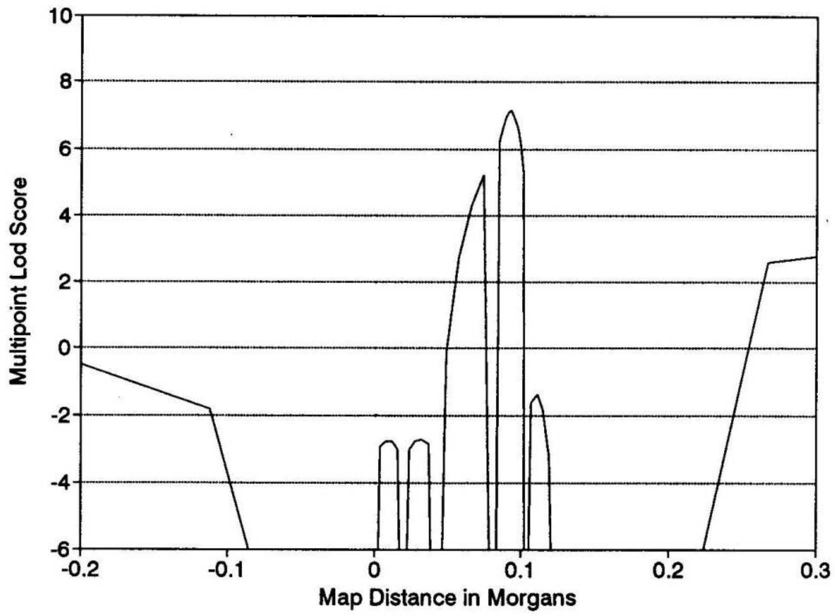


Figure 20–1. Graph of multipoint lod scores from Table 20–8

Results from the LINKMAP analysis are given in Table 20-8, and shown graphically in Figure 20-1 (This graph is based on the erroneous lod scores in Table 20-8. Therefore, the two humps immediately to the right of map location 0, with heights of about –3 each, should not be visible). First compare this graph with Figure 16-2, under the different intermarker recombination fractions. How much can you tell about the analysis from the picture alone? In this analysis, our maximum lod score is reduced to only 7.16 from 9.011 in the example from chapter 16. Further, our 3-unit support interval is also reduced to only 3.5 cM, on the interval (0.066, 0.101). So, you can see that having a finer marker map can help you to narrow the support interval for the location of your putative disease gene. In general, this was shown to be true, that the finer the grid of markers analyzed against a disease, the smaller the support interval will be in expectation, almost independent of marker heterozygosity (Terwilliger et al, 1992).

```
Locus Order/Θ's          Loc. Score     -2LN Like           Lod Score      Map Distance

1====5----2----8----3
 .500 .020 .020 .040    +0.0000E+00   +4.2881E+02             0.00               -∞
 .400 .020 .020 .040    +3.4775E+00   +4.2533E+02             0.76             -0.805
 .300 .020 .020 .040    +4.2394E+00   +4.2457E+02             0.92             -0.458
 .200 .020 .020 .040    +1.4615E+00   +4.2735E+02             0.32             -0.255
 .100 .020 .020 .040    -8.3698E+00   +4.3718E+02             -1.81            -0.112
5====1====2----8----3
 .000 .020 .020 .040       -∞       ∞              -∞         0.000
 .004 .016 .020 .040    -1.3563E+02   +5.6444E+02            -2.94              0.004
 .008 .012 .020 .040    -1.2733E+02   +5.5614E+02            -2.76              0.008
 .012 .008 .020 .040    -1.2822E+02   +5.5703E+02            -2.78              0.012
 .016 .004 .020 .040    -1.3865E+02   +5.6746E+02            -3.01              0.016
 .020 .000 .020 .040       -∞       ∞              -∞        0.020
5----2====1====8----3
 .020 .000 .020 .040       -∞       ∞              -∞         0.020
 .020 .004 .016 .040    -1.3883E+02   +5.6763E+02            -3.01              0.024
 .020 .008 .012 .040    -1.2746E+02   +5.5627E+02            -2.77              0.028
 .020 .012 .008 .040    -1.2539E+02   +5.5420E+02            -2.72              0.032
 .020 .016 .004 .040    -1.3193E+02   +5.6074E+02            -2.86              0.037
 .020 .020 .000 .040       -∞       ∞              -∞        0.041

1====2----8----3----6
 .500 .020 .040 .010    +0.0000E+00   +4.0999E+02
2----8====1====3----6
 .020 .000 .040 .010       -∞       ∞              -∞         0.041
 .020 .008 .033 .010    -6.1661E-01   +4.1061E+02            -0.13              0.049
 .020 .016 .025 .010    +1.2609E+01   +3.9738E+02             2.74              0.057
 .020 .024 .017 .010    +1.9822E+01   +3.9017E+02             4.30              0.066
 .020 .032 .009 .010    +2.4068E+01   +3.8592E+02             5.23              0.074
 .020 .040 .000 .010       -∞       ∞              -∞        0.083

1====8----3----6----4
 .500 .040 .010 .010    +0.0000E+00   +4.0455E+02
8----3====1====6----4
 .040 .000 .010 .010       -∞       ∞              -∞         0.083
 .040 .002 .008 .010    +2.8807E+01   +3.7574E+02             6.26              0.085
 .040 .004 .006 .010    +3.0757E+01   +3.7379E+02               6.68               0.087
 .040 .006 .004 .010    +3.1931E+01   +3.7262E+02               6.93               0.089
 .040 .008 .002 .010    +3.2653E+01   +3.7190E+02               7.09               0.091
 .040 .010 .000 .010    +3.2952E+01   +3.7160E+02               7.16               0.093

1====3----6----4----9
 .500 .010 .010 .020    +0.0000E+00   +3.7942E+02
3----6====1====4----9
 .010 .000 .010 .020    +3.2911E+01   +3.4651E+02               7.15               0.093
 .010 .002 .008 .020    +3.1948E+01   +3.4747E+02               6.94               0.095
 .010 .004 .006 .020    +3.0542E+01   +3.4888E+02               6.63               0.097
 .010 .006 .004 .020    +2.8388E+01   +3.5103E+02               6.16               0.099
 .010 .008 .002 .020    +2.4478E+01   +3.5494E+02               5.32               0.101
 .010 .010 .000 .020       -∞         ∞             -∞          0.103

1====6----4----9----7
 .500 .010 .020 .030    +0.0000E+00   +3.7765E+02
6----4====1====9----7
 .010 .000 .020 .030       -∞         ∞             -∞         0.103
 .010 .004 .016 .030    -7.4645E+00   +3.8511E+02              -1.62               0.107
 .010 .008 .012 .030    -6.3847E+00   +3.8403E+02              -1.38               0.111
 .010 .012 .008 .030    -8.4791E+00   +3.8613E+02              -1.84               0.115
 .010 .016 .004 .030    -1.4758E+01   +3.9241E+02              -3.20               0.119
 .010 .020 .000 .030       -∞         ∞             -∞         0.124
6----4----9====1====7
 .010 .020 .000 .030       -∞         ∞             -∞         0.124
 .010 .020 .006 .024    -4.9452E+01   +4.2710E+02             -10.74               0.130
 .010 .020 .012 .018    -4.5242E+01   +4.2289E+02              -9.82               0.136
 .010 .020 .018 .012    -4.6160E+01   +4.2381E+02             -10.02               0.142
 .010 .020 .024 .006    -5.2556E+01   +4.3021E+02             -11.41               0.149
6----4----9----7====1
 .010 .020 .030 .000       -∞         ∞             -∞         0.155
 .010 .020 .030 .100    +1.1874E+01   +3.6578E+02               2.58               0.267
 .010 .020 .030 .200    +1.5418E+01   +3.6223E+02               3.35               0.410
 .010 .020 .030 .300    +1.3358E+01   +3.6429E+02               2.90               0.613
 .010 .020 .030 .400    +8.0961E+00   +3.6955E+02               1.76               0.960
```

Table 20-8 : LINKMAP Results - Files MULTDIS1.*; remember: Lod score = location score divided by 4.6 (some of the lod scores in this table are too small by a factor of 10 but the location scores are accurate).

EXERCISE 17

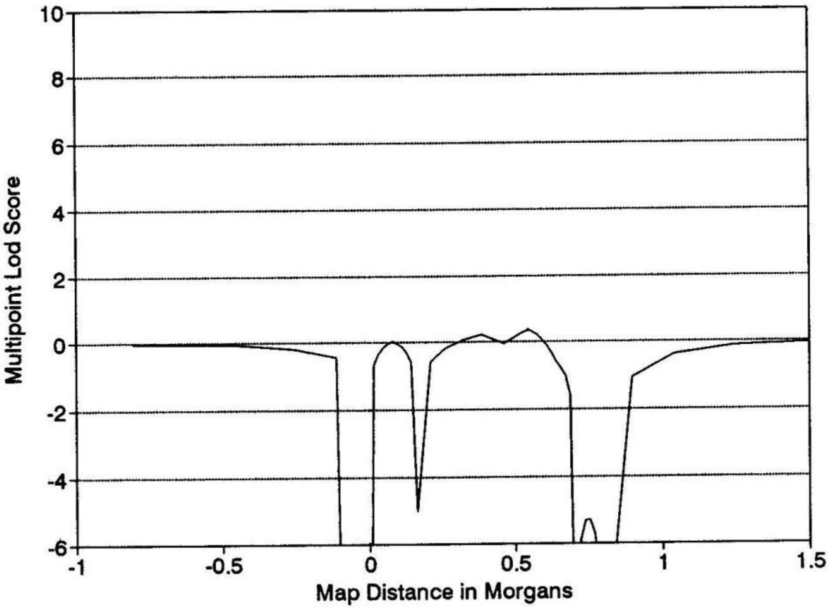


Figure 20–2. Graph of multipoint lod scores from Table 20–9

The results of the analysis under the assumption of a recessive disease with 80% penetrance are presented in table 20-9, and shown graphically in Figure 20-2 for purposes of visual comparison with figure 16-2. In this analysis, the maximum lod score was obtained at the same location as in the dominant model, only the maximum lod score was only 0.38 under this model. Similarly, there was much less exclusion power as well. Fortunately, in this case, due to the low penetrance for the disease, much ambiguity remained in terms of the genotypes for the unaffected individuals. Further, the analysis was using almost exactly the meioses which could not be used in the dominant analysis, since now all affected individuals were considered to be homozygous, so all the information about linkage comes from the heterozygous (at the disease locus) mothers. In light of this, it is clear that the fact that the maximum lod score occurred at the same point primarily by chance, and not through a strong correlation in the data between the dominant and recessive analyses, given the f these pedigrees.

```
Locus Order/θ's           Loc. Score     -2LN Like          Lod Score Map Distance
------------------------------------------------------------------------------------
1====5----2----8----3
 .500 .075 .075 .225    +0.0000E+00   +5.6631E+02           0.00      -∞
 .400 .075 .075 .225    -6.4885E-02   +5.6637E+02          -0.01       -0.80471
 .300 .075 .075 .225    -2.9353E-01   +5.6660E+02          -0.06       -0.45814
 .200 .075 .075 .225    -8.1105E-01   +5.6712E+02          -0.18       -0.25541
 .100 .075 .075 .225    -2.0020E+00   +5.6831E+02          -0.43       -0.11157
5====1====2----8----3
 .000 .075 .075 .225    -2.3051E+01   +5.8936E+02          -5.01       0.000000
 .015 .062 .075 .225    -3.0653E+00   +5.6937E+02          -0.67       0.015229
 .030 .048 .075 .225    -1.5937E+00   +5.6790E+02          -0.34       0.030937
 .045 .033 .075 .225    -7.4906E-01   +5.6706E+02          -0.16       0.047155
 .060 .017 .075 .225    -1.9528E-01   +5.6650E+02          -0.04       0.063916
 .075 .000 .075 .225    +1.6389E-01   +5.6614E+02           0.04       0.081259
5----2====1====8----3
 .075 .000 .075 .225    +1.6389E-01   +5.6614E+02           0.04       0.081259
 .075 .015 .062 .225    -1.4129E-01   +5.6645E+02          -0.03       0.096489
 .075 .030 .048 .225    -6.2496E-01   +5.6693E+02          -0.14       0.112197
 .075 .045 .033 .225    -1.3942E+00   +5.6770E+02          -0.30       0.128414
 .075 .060 .017 .225    -2.7949E+00   +5.6910E+02          -0.61       0.145176
 .075 .075 .000 .225    -2.3051E+01   +5.8936E+02          -5.01       0.162518

1====2----8----3----6
 .500 .075 .225 .075    +0.0000E+00   +5.5254E+02
2----8====1====3----6
 .075 .000 .225 .075    -2.3129E+01   +5.7567E+02          -5.02       0.162518
 .075 .045 .198 .075    -2.5648E+00   +5.5510E+02          -0.56       0.209674
 .075 .090 .165 .075    -8.5232E-01   +5.5339E+02          -0.19       0.261744
 .075 .135 .123 .075    +2.4066E-01   +5.5230E+02           0.05       0.319874
 .075 .180 .070 .075    +1.0938E+00   +5.5144E+02           0.23       0.385662

1====8----3----6----4
 .500 .225 .075 .075    +0.0000E+00   +5.4812E+02
8----3====1====6----4
 .225 .000 .075 .075    -2.2855E-01   +5.4835E+02          -0.05       0.461437
 .225 .015 .062 .075    +1.8114E-01   +5.4794E+02           0.04       0.476667
 .225 .030 .048 .075    +5.5754E-01   +5.4757E+02           0.12       0.492375
 .225 .045 .033 .075    +9.0693E-01   +5.4722E+02           0.20       0.508592
 .225 .060 .017 .075    +1.2339E+00   +5.4689E+02           0.27       0.525354
 .225 .075 .000 .075    +1.5421E+00   +5.4658E+02           0.33       0.542696

1====3----6----4----9
 .500 .075 .075 .075    +0.0000E+00   +5.3465E+02
3----6====1====4----9
 .075 .000 .075 .075    +1.7344E+00   +5.3291E+02           0.38       0.5427
 .075 .015 .062 .075    +1.4036E+00   +5.3324E+02           0.30       0.5579
 .075 .030 .048 .075    +9.9157E-01   +5.3366E+02           0.22       0.5736
 .075 .045 .033 .075    +4.5373E-01   +5.3419E+02           0.10       0.5899
 .075 .060 .017 .075    -3.0710E-01   +5.3496E+02          -0.07       0.6066
 .075 .075 .000 .075    -1.5850E+00   +5.3623E+02          -0.34       0.6240

1====6----4----9----7
 .500 .075 .075 .075    +0.0000E+00   +5.3758E+02
6----4====1====9----7
 .075 .000 .075 .075    -1.5850E+00   +5.3916E+02          -0.34       0.623956
 .075 .015 .062 .075    -2.8257E+00   +5.4040E+02          -0.61       0.639185
 .075 .030 .048 .075    -3.5801E+00   +5.4116E+02          -0.78       0.654894
 .075 .045 .033 .075    -4.6283E+00   +5.4220E+02          -1.01       0.671111
 .075 .060 .017 .075    -7.4094E+00   +5.4499E+02          -1.61       0.687873
 .075 .075 .000 .075    -6.8973E+01   +6.0655E+02         -14.98       0.705215
6----4----9====1====7
 .075 .075 .000 .075    -6.8973E+01   +6.0655E+02         -14.98       0.705215
 .075 .075 .015 .062    -2.7083E+01   +5.6466E+02            -5.88      0.720445
 .075 .075 .030 .048    -2.4572E+01   +5.6215E+02            -5.34      0.736153
 .075 .075 .045 .033    -2.4399E+01   +5.6198E+02            -5.30      0.752371
 .075 .075 .060 .017    -2.6332E+01   +5.6391E+02            -5.72      0.769132
 .075 .075 .075 .000    -5.1009E+01   +5.8859E+02           -11.08      0.786475
6----4----9----7====1
 .075 .075 .075 .000    -5.1009E+01   +5.8859E+02           -11.08      0.786475
 .075 .075 .075 .100    -4.9218E+00   +5.4250E+02            -1.06      0.898047
 .075 .075 .075 .200    -1.8085E+00   +5.3938E+02            -0.39      1.041888
 .075 .075 .075 .300    -5.6751E-01   +5.3814E+02            -0.12      1.244620
 .075 .075 .075 .400    -1.0589E-01   +5.3768E+02            -0.02      1.591194
```

Table 20-9: Analysis of MULTDIS1.PED under autosomal recessive model with 80% penetrance.

When the uncertainty of diagnosis is entered for the *unaffected* individuals in this pedigree, it is necessary to allow for an additional liability class, (since *affected* individuals are still affected with 100% certainty). So, we must go back into the parameter file, and add an additional liability class for the unaffected individuals. If you remember from chapter 10, to allow for uncertainty of diagnosis, the new penetrances should be computed as (p)P(affected | genotype) + ($1 - p$)P(unaffected | genotype); where in this case, p = the probability that the person really is affected. Thus in our situation, with a 75% chance that the people are unaffected, we have a 25% chance that they are affected, so $p = 0.25$. Then our penetrances should be f(+/+) = 0.25(0) + 0.75(1) = 0.75; f(D/+) = f(D/D) = 0.25(1) + 0.75(0) = 0.25. Thus our two liability classes would be as follows:

GENOTYPE:	+/+	D/+	D/D
Affecteds	0	1	1
Unaffecteds	0.75	0.25	0.25

Unaffecteds in the pedigree file would then have affection status phenotype *2 2*, and affecteds would have phenotype *2 1*. The resulting LINKMAP output is shown in table 20-10, and illustrated graphically in Figure 20-3. The maximum lod score in this example is only 6.80, with a support interval covering the range (0.3199,0.6066), for a 28.67 cM support interval. The exclusion regions are fairly similar to those obtained with 100% diagnostic certainty, though the magnitude of the lod scores is much smaller (i.e. the values are much *less* negative in this example, implying that many of the "obligate" recombination events occurred in unaffected individuals.

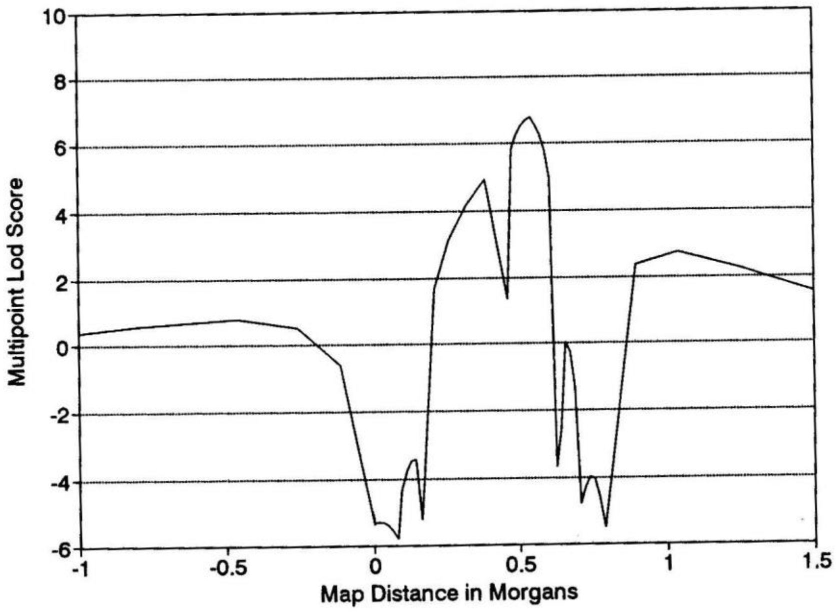


Figure 20–3. Graph of multipoint lod scores from Table 20–10

```
Locus Order/Θ's           Loc. Score    -2LN Like         Lod Score Map Distance
────────────────────────────────────────────────────────────────────────────────
1====5----2----8----3
 .500 .075 .075 .225    +0.0000E+00   +3.8998E+02             0.00        -∞
 .400 .075 .075 .225    +2.7396E+00   +3.8724E+02             0.59   -0.80471
 .300 .075 .075 .225    +3.7225E+00   +3.8626E+02             0.81   -0.45814
 .200 .075 .075 .225    +2.4881E+00   +3.8749E+02             0.54   -0.25541
 .100 .075 .075 .225    -2.8903E+00   +3.9287E+02            -0.63   -0.11157
5====1====2----8----3
 .000 .075 .075 .225    -2.4547E+01   +4.1453E+02            -5.33   0.000000
 .015 .062 .075 .225    -2.4432E+01   +4.1441E+02            -5.31   0.015229
 .030 .048 .075 .225    -2.4383E+01   +4.1436E+02             -5.29   0.030937
 .045 .033 .075 .225    -2.4731E+01   +4.1471E+02             -5.37   0.047155
 .060 .017 .075 .225    -2.5559E+01   +4.1554E+02             -5.55   0.063916
 .075 .000 .075 .225    -2.6670E+01   +4.1665E+02             -5.79   0.081259
5----2====1====8----3
 .075 .000 .075 .225    -2.6670E+01   +4.1665E+02             -5.79   0.081259
 .075 .015 .062 .225    -1.9969E+01   +4.0995E+02             -4.33   0.096489
 .075 .030 .048 .225    -1.7256E+01   +4.0724E+02             -3.75   0.112197
 .075 .045 .033 .225    -1.5866E+01   +4.0585E+02             -3.45   0.128414
 .075 .060 .017 .225    -1.5848E+01   +4.0583E+02             -3.44   0.145176
 .075 .075 .000 .225    -2.4010E+01   +4.1399E+02             -5.21   0.162518

1====2----8----3----6
 .500 .075 .225 .075    +0.0000E+00   +3.7621E+02
2----8====1====3----6
 .075 .000 .225 .075    -2.3982E+01   +4.0019E+02             -5.21   0.162518
 .075 .045 .198 .075    +7.7635E+00   +3.6845E+02              1.69   0.209674
 .075 .090 .165 .075    +1.4509E+01   +3.6170E+02              3.15   0.261744
 .075 .135 .123 .075    +1.9169E+01   +3.5704E+02              4.16   0.319874
 .075 .180 .070 .075    +2.2689E+01   +3.5352E+02              4.93   0.385662

1====8----3----6----4
 .500 .225 .075 .075    +0.0000E+00   +3.7180E+02
8----3====1====6----4
 .225 .000 .075 .075    +6.2405E+00   +3.6556E+02              1.36   0.461437
 .225 .015 .062 .075    +2.6736E+01   +3.4506E+02              5.81   0.476667
 .225 .030 .048 .075    +2.8731E+01   +3.4307E+02              6.24   0.492375
 .225 .045 .033 .075    +2.9977E+01   +3.4182E+02              6.51   0.508592
 .225 .060 .017 .075    +3.0818E+01   +3.4098E+02              6.69   0.525354
 .225 .075 .000 .075    +3.1299E+01   +3.4050E+02              6.80   0.542696

1====3----6----4----9
 .500 .075 .075 .075    +0.0000E+00   +3.5832E+02
3----6====1====4----9
 .075 .000 .075 .075    +3.1178E+01   +3.2714E+02              6.77   0.5427
 .075 .015 .062 .075    +3.0155E+01   +3.2817E+02              6.55   0.5579
 .075 .030 .048 .075    +2.8738E+01   +3.2958E+02              6.24   0.5736
 .075 .045 .033 .075    +2.6620E+01   +3.3170E+02              5.78   0.5899
 .075 .060 .017 .075    +2.2791E+01   +3.3553E+02              4.95   0.6066
 .075 .075 .000 .075    -1.7001E+01   +3.7532E+02             -3.69   0.6240

1====6----4----9----7
 .500 .075 .075 .075    +0.0000E+00   +3.6125E+02
6----4====1====9----7
 .075 .000 .075 .075    -1.7001E+01   +3.7825E+02             -3.69   0.623956
 .075 .015 .062 .075    -1.1794E+00   +3.6243E+02             -2.56   0.639185
 .075 .030 .048 .075    +1.6468E-01   +3.6108E+02              0.04   0.654894
 .075 .045 .033 .075    -1.3096E+00   +3.6256E+02             -0.28   0.671111
 .075 .060 .017 .075    -6.2153E+00   +3.6746E+02             -1.34   0.687873
 .075 .075 .000 .075    -2.2096E+01   +3.8335E+02             -4.79   0.705215
6----4----9====1====7
 .075 .075 .000 .075    -2.2096E+01   +3.8335E+02             -4.79   0.705215
 .075 .075 .015 .062    -1.9667E+01   +3.8092E+02             -4.27   0.720445
 .075 .075 .030 .048    -1.8273E+01   +3.7952E+02             -3.97   0.736153
 .075 .075 .045 .033    -1.8599E+01   +3.7985E+02             -4.04   0.752371
 .075 .075 .060 .017    -2.0947E+01   +3.8220E+02             -4.55   0.769132
 .075 .075 .075 .000    -2.5225E+01   +3.8647E+02             -5.48   0.786475
6----4----9----7====1
 .075 .075 .075 .000    -2.5225E+01   +3.8647E+02             -5.48   0.786475
 .075 .075 .075 .100    +1.0971E+01   +3.5028E+02              2.38   0.898047
 .075 .075 .075 .200    +1.2544E+01   +3.4871E+02              2.72   1.041888
 .075 .075 .075 .300    +1.0447E+01   +3.5080E+02              2.27   1.244620
 .075 .075 .075 .400    +6.1823E+00   +3.5507E+02              1.34   1.591194
```

Table 20-10: Analysis of MULTDIS1.* with LINKMAP assuming 75% certainty of diagnosis of unaffected individuals.

EXERCISE 18

The top 20 orders found with CILINK are presented in table 20-11 under the assumption of constant sex difference, and in table 20-12 under the assumption of varying sex difference. The same four orders are always the top four, under any of the three sex difference models. In order to test the significance of the evidence for sex difference in recombination, let us again consider the three hypotheses given the top ranked order, as outlined in table 20-13. From this table, we can see that the test of H_1 vs H_0 has a chi-squared value of 7.32 with (5 – 4) = 1df, with a corresponding p-value of 0.008. The test of H_2 vs H_1, however has a chi-squared value of 2.63 with (8 – 5) = 3 df, for a p-value of only 0.45, which is not at all significant. In light of this analysis, our best conclusion is that there is a constant sex difference, with a female/male map distance ratio of 2.573. It is interesting to note that some of the map distance ratios are enormous under the varying sex difference option. Usually this occurs when θ = 0.001, meaning that no recombinants were seen in this data set in males, even though many recombinants were seen in females. This phenomenon can lead to estimates of the map distance ratio as high as 641.752 in the small sample in table 20-12.

Locus Order	Male Thetas	x_f/x_m	-2LN Like	Odds
3----4----5----1----2	.033 .173 .031 .131	2.573	-1.1127E+02	1.00E+00
4----3----5----1----2	.034 .165 .031 .132	2.497	-1.1075E+02	1.29E+00
3----4----1----5----2	.035 .162 .021 .156	2.459	-1.0981E+02	2.07E+00
4----3----1----5----2	.035 .154 .022 .158	2.391	-1.0916E+02	2.87E+00
5----1----2----3----4	.022 .091 .269 .029	3.983	-9.8521E+01	5.85E+02
5----1----2----4----3	.022 .092 .274 .026	3.900	-9.7107E+01	1.19E+03
1----5----2----3----4	.014 .099 .266 .028	4.388	-9.5537E+01	2.60E+03
1----5----2----4----3	.015 .103 .271 .026	4.043	-9.4139E+01	5.24E+03
5----1----4----3----2	.022 .123 .032 .290	3.175	-9.3867E+01	6.00E+03
1----5----4----3----2	.022 .118 .028 .291	3.918	-9.1877E+01	1.62E+04
5----1----3----4----2	.023 .118 .027 .295	3.020	-9.1621E+01	1.84E+04
1----5----3----4----2	.022 .114 .023 .295	3.803	-8.9630E+01	4.99E+04
5----3----4----1----2	.112 .027 .163 .118	3.109	-8.7881E+01	1.20E+05
5----4----3----1----2	.118 .029 .156 .114	3.265	-8.7735E+01	1.29E+05
4----5----3----1----2	.126 .029 .101 .140	2.436	-8.6845E+01	2.01E+05
3----4----1----2----5	.028 .172 .097 .143	3.491	-8.6645E+01	2.22E+05
4----3----1----2----5	.030 .165 .095 .142	3.537	-8.6459E+01	2.44E+05
1----2----5----3----4	.083 .175 .132 .026	3.986	-8.4785E+01	5.63E+05
1----2----5----4----3	.081 .177 .132 .026	4.165	-8.3896E+01	8.77E+05
4----5----1----2----3	.179 .030 .125 .228	2.597	-8.3733E+01	9.52E+05

Table 20-11: Top twenty orders of loci in USEREX14.* - constant sex difference

Locus Order	Male Thetas	Female / Male Map Distance Ratios	-2LN Like	Odds
3----4----5----1----2	.001 .173 .001 .181	248.051 2.448 141.771 1.514	-1.1390E+02	1.00E+00
3----4----1----5----2	.001 .161 .001 .152	309.807 2.236 135.184 2.563	-1.1224E+02	2.29E+00
4----3----1----5----2	.001 .202 .001 .154	171.897 1.463 98.958 2.494	-1.1134E+02	3.60E+00
4----3----5----1----2	.001 .208 .001 .167	324.302 2.979 383.886 2.369	-1.0749E+02	2.47E+01
5----1----2----3----4	.001 .128 .262 .001	133.534 2.720 3.728 445.905	-9.9185E+01	1.57E+03
1----5----2----3----4	.001 .111 .283 .001	106.745 3.648 2.355 236.959	-9.6968E+01	4.76E+03
5----1----4----3----2	.001 .132 .001 .291	161.494 2.272 360.817 1.344	-9.5964E+01	7.86E+03
1----5----2----4----3	.001 .101 .285 .001	105.075 4.047 2.460 244.922	-9.5814E+01	8.47E+03
5----1----2----4----3	.001 .113 .355 .001	190.615 4.265 16.182 252.037	-9.5347E+01	1.07E+04
5----1----3----4----2	.001 .144 .001 .322	134.152 2.026 167.749 1.757	-9.4186E+01	1.91E+04
5----3----4----1----2	.047 .001 .166 .161	10.334 197.781 2.448 1.857	-9.1709E+01	6.60E+04
5----4----3----1----2	.046 .001 .166 .161	11.813 193.495 2.368 1.842	-9.1464E+01	7.46E+04
4----5----3----1----2	.001 .037 .131 .188	638.115 1.656 1.722 1.446	-9.1238E+01	8.35E+04
3----4----1----2----5	.001 .193 .106 .069	259.719 1.956 2.806 10.297	-9.0782E+01	1.05E+05
1----5----3----4----2	.001 .194 .001 .312	170.584 1.773 218.382 2.397	-9.0130E+01	1.45E+05
4----3----1----2----5	.001 .204 .105 .075	195.450 1.881 3.010 11.845	-9.0002E+01	1.55E+05
3----5----4----1----2	.029 .001 .169 .132	6.854 523.242 2.084 2.399	-8.8610E+01	3.11E+05
1----5----4----3----2	.001 .088 .001 .394	130.390 8.277 641.752 12.875	-8.6367E+01	9.54E+05
1----2----5----3----4	.074 .145 .187 .001	4.668 5.108 2.309 195.065	-8.5768E+01	1.29E+06
1----2----5----4----3	.088 .177 .161 .001	4.362 3.951 2.977 243.203	-8.4972E+01	1.92E+06

Table 20-12: Top twenty orders of loci in USEREX14.* - varying sex difference

	MODEL	df	-2LN(Like)	Δ-2LN(Like)
H_2	Varying Sex Difference:	8	-113.90	0.00
H_1	Constant Sex Difference:	5	-111.27	2.63
H_0	No Sex Difference:	4	-103.95	9.95

Table 20-13: Likelihoods of order 3-4-5-1-2 under different sex difference models

In the files MULTDIS1.*, the evidence for sex difference in recombination is outlined in table 20-14. As you can see, there is very little evidence for sex difference in recombination, with completely non-significant chi-squared test results. However, it is interesting to notice the estimated constant sex ratio for this set of loci. The ratio in this example was estimated to be 0.599, meaning that there were higher rates of recombination in males than females by almost 2:1. If you remember back in the original analysis with the dataset in MULTDIS2.*, the estimated sex ratio was 2.113, meaning that in that dataset, females showed twice the recombination rate of males. In light of this information, it is not surprising that when the two datasets are combined, the estimated map distance ratio was only 1.23, implying that there is a negligible sex difference. This is caused by the two datasets canceling each other out, in terms of recombination sex differences. It is therefore also not surprising that the combined datasets provided a much less significant test than either of the two datasets separately did, as shown in table 20-15.

	MODEL	df	-2LN(Like)	Δ-2LN(Like)
H_2	Varying Sex Difference:	8	309.76	0.00
H_1	Constant Sex Difference:	5	310.34	0.58
H_0	No Sex Difference:	4	310.83	1.07

Table 20-14: Likelihoods of order 3-6-4-9-7 under different sex difference models - files MULTDIS1.*

	MODEL	df	-2LN(Like)	_-2LN(Like)
H_2	Varying Sex Difference:	8	590.60	0.00
H_1	Constant Sex Difference:	5	590.61	0.01
H_0	No Sex Difference:	4	590.82	0.22

Table 20-15: Likelihoods of order 3-6-4-9-7 under different sex difference models - files MULTDIS1.* and MULTDIS2.* together.

EXERCISE 19

In this exercise, the only triples of loci meeting the criterion that $0.125 \le \theta_{AB}, \theta_{BC} \le 0.225$ are 5-(0.139)-8-(0.225)-3, 8-(0.225)-3-(0.139)-4, and 3-(0.139)-4-(0.139)-7. The results of the analyses of these triples of loci are presented in table 20-16, with the values of c_m, c_f, and –2ln(Likelihood) given for each analysis (θ's not shown). Does the great disparity in the estimates of c_m and c_f under the assumption of constant sex difference make sense to you? The reason is quite simple. You see, the "*constant sex difference*" option forces the recombination fractions in males and females to fit the constant map distance ratio criterion, which assumes the Haldane mapping function. Since the Haldane mapping function is incompatible with interference, when doing such an analysis, you are mixing apples and oranges, and the analysis simply doesn't make any sense. The analysis with loci 8-3-4 points it out quite nicely, as while c_m is essentially 0 (to fit the constant map distance criterion), the female θ's are computed such that $\theta_f = 0.60$, when the *variable sex difference* and *no sex difference* options both yielded an estimate of $c_m = c_f = 0$. Clearly, it is impossible for this to hold under the constant sex difference option unless the sex ratio was 1. Try and prove this to yourself as an exercise.

Loci	Sex Difference	c_	c_	-2ln L(c = 1)	-2ln L(c =)	χ^2	p-value
5-8-3	None	1.48	1.48	215.10	214.66	0.44	0.50
	Constant	2.32	1.31	213.00	212.15	0.85	0.36
	Varying	3.33	0.83	212.99	211.35	1.64	0.20
8-3-4	None	0.00	0.00	215.10	211.75	3.35	0.07
	Constant	0.03	0.60	213.00	211.16	1.84	0.17
	Varying	0.00	0.00	212.99	209.04	3.95	0.05
3-4-7	None	0.00	0.00	206.26	204.14	2.12	0.15
	Constant	0.00	0.58	204.67	203.41	1.26	0.26
	Varying	0.00	0.00	204.67	202.20	2.47	0.12

Table 20-16 : Interference analysis of loci in MULTDIS2.* with ILINK

To further demonstrate the poor fit of the constant sex difference model to the interference analysis, take a look at the likelihood ratio χ^2 values shown in table 20-16. Notice that with the exception of the first triple of loci (for which no evidence of interference was found), the *constant sex difference* option performed the worst of all, due to its internal incompatibility. In the set of loci 8-3-4, we have our strongest evidence for interference, with a p-value of 0.05 for the test. This is pretty convincing evidence, given the small dataset we had to work with, and the fact that our estimated coefficients of coincidence were all 0 under the *Varying sex difference* model. Since we have already established that this dataset shows sex difference in recombination rates, and the *Constant sex difference* option is invalid for interference calculations in ILINK, the only valid test to consider is the *Varying sex difference* test. So, we would conclude that there is some evidence for positive interference in this dataset, although not very powerful evidence. Since we looked at three separate samples, we need to correct for multiple testing. Since the one triple of loci that gave a significant test result was only marginally significant, after applying a correction for the three tests we carried out, this value would no longer be significant, but combined with the reasonably powerful test result from locus order 3-4-7, and the fact that it also yielded estimates of $\hat{c} = 0$, you could safely say that there is suggestive evidence of positive interference in this dataset, but not conclusively so.

Part III: Advanced Topics in Linkage Analysis

21 Mutation Rates and the LINKAGE programs

In this chapter, you will be introduced to the concept of mutation, and the way in which it can be utilized in the LINKAGE programs. You will see why it is sometimes obligatory to use this option, especially when analyzing sexlinked recessive lethal diseases.

21.1 MUTATIONS

It is often possible for a child to inherit an allele from one of his parents which is not found in the parent in question. In other words, it is sometimes seen that a 1/1 father and a 2/2 mother might have a child with genotype 2/2. Clearly, this is inconsistent with normal Mendelian laws as introduced in chapter 1. However, the process of *mutation* can cause one parental allele to be mutated or changed into a different allele, and this mutant allele can then be transmitted to the offspring. Other mechanisms can lead to this situation, as discussed in Ott (1991, p.256, Malcolm et al, 1990), but in LINKAGE the only such phenomenon that can directly be dealt with is mutation. Many biochemical explanations can be given for the existence of such genetic mutations, and they are the primary source of the genetic variability and genetic diseases that exist. It is often essential to allow for the occurrence of mutations in performing a linkage analysis. A situation like the one described above would cause the LINKAGE programs to complain about a genetic inconsistency in the data if the mutation rate were not allowed for. However, if one were to perform the linkage analysis assuming some fixed rate of mutation, then the Mendelian inconsistency would be explained away as a newly mutated allele.

Frequently, when performing a linkage analysis with certain genetic diseases, one might want to assume a certain rate of mutation creating new disease alleles, and when one is analyzing a sex-linked recessive disease which is lethal, it is mandatory to allow for a certain fixed mutation rate, as you will see below.

21.2 ALLOWING FOR MUTATION RATES IN LINKAGE

In general, one could define separate mutation rates to and from each different allele at a locus. For example, if we had a locus with two alleles, there could be separate mutation rates $\mu_{1\to2}$ and $\mu_{2\to1}$, for the two types of mutation possible. In general, these will not likely be equal. Further, it is often the case that there are different rates of mutation in males and females, since the processes of spermatogenesis and oogenesis are very different.

In the LINKAGE programs, however, the use of mutation rates are extremely limited. The restriction is so severe that only one locus (of any locus type) is allowed to have mutations, and at one constant rate for males, and another for females. Further, the mutation can only be specified in a unidirectional manner, from any allele to the last allele. For this reason, one usually sets up an affection status locus with the disease allele as allele *2*, so mutation can be allowed to the disease allele, but not from disease to normal. Typically, the disease allele frequency is so small that any such back-mutation would occur at a negligible frequency.

The primary use of this option will be when you are considering disease loci, with mutation occurring from the normal allele to the disease allele at the disease locus. This is the purpose with which this option was originated, and it remains a potentially major restriction in the applicability of LINKAGE with highly mutable CA repeat loci being used with more regularity.

Let us return to the phase known example from chapter 5. In this pedigree, we shall now assume that there is a mutation rate μ, where μ is the probability that any normal allele mutates to a disease allele in one meiosis. Please read the parameter file from that example into PREPLINK, and select the option *(d) Mutation*. You will then see a screen like the following:

```
**********************************************
(a) MUTATION LOCUS      : 0
(b) MUTATION RATE MALES       : 0.00000E+00
(c) MUTATION RATE FEMALES     : 0.00000E+00
(d) MUTATION      : N
(e) RETURN TO MAIN MENU
**********************************************
enter letter to modify values
```

At this point, you should select option (a) and respond to the questions as follows:

```
 ENTER NEW MUTATION LOCUS
1           (The disease locus)
 ENTER NEW MUTATION RATE MALES
.000001     (Here you could enter whatever the value is for μ)
 ENTER NEW MUTATION RATE FEMALES
.000001     (If we assume equal rates in males and females, type μ)
```

Then, write this datafile, and rerun the analysis of the pedigree in question. Please try varying values for the mutation rate, to visualize its effect on our analysis. The results should resemble the results given in table 21-1.

μ	$\hat{\theta}$	$Z(\hat{\theta})$
$< 10^{-5}$	0.798	0.4185
10^{-4}	0.798	0.4183
10^{-3}	0.799	0.4168
10^{-2}	0.797	0.4015
10^{-1}	0.775	0.2746
0.5	0.655	0.0456

Table 21-1: Effects of mutation rate on the M.L.E. of θ, and the corresponding lod score on a phase known pedigree

In this example, it is clear that allowing for mutation has little effect on the likelihood, implying that there are most likely no new mutations in this pedigree. This does make sense, in light of the pedigree structure and phenotypes. Let us look at this analytically to see that this is so.

We know that the disease locus genotype of *father* is D/+, and that *mother* is +/+. Because of the low gene frequency, one can assume that *fgrandma* has genotype D/+, and we know that *fgrandpa* has genotype +/+. Because of the mutation rate, however, father could have gotten the D allele from either parent, since a mutation could have occurred coming from *fgrandma*. Hence, there are two possible phases for *father*. We must look at the likelihood of *father* having received genotype D/+ given the D allele came from *fgrandma*, as compared with the likelihood given the D allele came from *fgrandpa*. The likelihood of *father* receiving the D allele from *fgrandma* and the + allele from *fgrandpa* is just [(1/2) + (1/2)μ][1 – μ] = (1/2)(1 – μ)(1+μ), and the likelihood of him receiving the + allele from *fgrandma* and the D allele from *fgrandpa* is just μ(1-μ)/2. Thus the probability that *father* received the D allele from *fgrandpa* is just [(1/2)(1 – μ)μ]/[(1/2)(1-μ)(1+2μ)] = μ/(1+2μ). Clearly, for μ = 0, this probability is 0, and the phase is known with certainty, and we have the same situation as described in chapter 5.

The overall likelihood of this pedigree is now more analogous to the phase unknown situation in the lower generation of the pedigree. There are 2 possible phases for *father*, with Phase I = D 1/+ 2, with P(Phase I) = (1+μ)/(1+2μ); Phase II = D 2/+ 1, with P(Phase II) = μ/(1+2μ). Now, let us consider the bottom generation. First, *dau1* has disease locus genotype +/+, having received the *2* allele from *father*. Under Phase I, this would happen with likelihood 1 – θ, and likelihood θ under Phase II. *dau2* and *son1* are identical, having either disease locus genotype D/+ or D/D. There are three possible scenarios to consider here. Either they received the D allele from *father,* and the + allele from *mother,* with likelihood [(1/2) θ + (1/2)μ(1 – θ)](1 – μ) under Phase I, and [(1/2)(1 – θ) + (1/2)μ θ](1 – μ) under phase II; or they received the + allele from *father*, and a mutated D allele from *mother* with likelihood (1/2)(1 – μ)(1 – θ)μ under Phase I, and (1/2)(1 – μ) θμ under Phase II; or they received the D allele from both parents with likelihood [(1/2) θ + (1/2)μ(1 – θ)]μ under Phase I, and [(1/2)(1 – θ) + (1/2)μθ]μ under Phase II. Finally, for *dau3* and *son2,* the received genotype is just +/+. Clearly, they also received the *1* allele from *father*, making their likelihood just θ under Phase I, and 1 – θ under Phase II. Putting all of this together, we can get the overall likelihood of the pedigree to be P(Pedigree | Father is Phase I)P(Phase I) + P(Pedigree | Father is Phase II)P(Phase II). In this case, it works out to be

$$L = \frac{1+\mu}{1+2\mu}[1-\theta][\theta + \mu(2-\mu)(1-\theta)]^2\theta^2 + \frac{\mu}{1+2\mu}\theta[(1-\theta) + \mu(2-\mu)\theta]^2[1-\theta]^2 .$$

Again, if you set $\mu = 0$, this likelihood is the same as the likelihood for the same pedigree obtained in section 3.1. The lod score is then $\log_{10}[L(\hat{\theta}, \mu)/L(\theta = ½, \mu)]$. To verify the calculations done by the LINKAGE programs above, please plug in the value of 0.01 for μ, and 0.797 for θ. The resulting lod score should turn out to be 0.4015, just as computed with LINKAGE. It is important to realize that 0.01 is an extremely high value to assume for μ; however, since the mutation rate must fit with certain conditions regarding mutation-selection equilibrium, μ must always be less than or equal to p, the frequency of the disease allele, as you will see in the following sections.

21.3 MUTATION-SELECTION EQUILIBRIUM FOR AUTOSOMAL TRAITS

In the previous example, we looked at the effect of various mutation rates on the lod score in our pedigree. However, we failed to consider whether they were meaningful estimates or not. In determining the mutation rate for a given disease, we need to select a value which is appropriate. If we assume that the disease allele frequency is stable, and equal to p, then the mutation rate and selection coefficient must balance out, such that the frequency of the disease gene remains constant from generation to generation. Most genetic diseases have negative effects on the fitness of an individual, and thus a certain proportion, s, of affected individuals will not reproduce in a given generation (s can also be interpreted as the relative reproductive fitness of any one individual). In order to have equilibrium of the allele frequency, these lost disease alleles must be replaced somehow. The only way to increase the frequency of the disease would be through new mutations. At equilibrium, therefore, the number of new mutations must equal the number of lost alleles in a given generation (Wright, 1968). For example, if we are dealing with an autosomal recessive disease, and a proportion s, of affected individuals will be lost due to selection, then sp^2 genes will be lost in each generation, and the new gene frequency would be $p_2 = p_1(1 - sp_1)$. If there were no new mutations, then over time, the gene frequency would continually decrease. These alleles lost due to selection must be replaced continuously for the gene frequency to be maintained. Clearly, then, the mutation rate must replace all of the alleles lost from selection, so $\mu = sp^2$, whenever p is presumed to be small. Therefore, the general equilibrium condition can be specified by $p = \sqrt{\mu/s}$, for a recessive disease. Similarly, it can be shown that for an autosomal dominant disease, $p = \mu/s$. For more details about mutation-selection equilibrium, consult Cavalli-Sforza and Bodmer (1971) and [22, 23].

21.4 SEX-LINKED LETHAL RECESSIVE DISEASE

The most simple and important situation is that of sex-linked recessive diseases which are lethal, i.e. no affected people live to reproduce. This situation is complete selection against the disease phenotype. Unchecked, it will lead inexorably to the loss of the disease allele. However, many such diseases are known to exist with constant prevalence. The best explanation is that the alleles are maintained in mutation-selection equilibrium (Haldane, 1935). In linkage analysis, the important allele frequency is that of the mating population. Clearly, since the disease is lethal at young age, and sex-linked recessive, all males will carry the normal allele, and the gene frequency in females will be p. Let us consider the distribution of genotypes in the next generation. Clearly, the possible matings (and offspring) will be (+/+♀ × +♂) → 1/2 +/+♀, 1/2 +♂ and (D/+♀ × +♂) → 1/4 D/+♀, 1/4 +/+♀, 1/4 D♂, 1/4 +♂. In this generation, half of the disease alleles went to D/+ females, and the other half went to the D males. Since all of the D males are lost due to selection, half of the disease alleles are lost in this generation, and $p_2 = p_1/2$ in the mating population, where p_1 is the gene frequency in the parental generation, and p_2 in the offspring generation. In order to maintain a stable equilibrium, a mutation rate of $p_1/2$ is required, to replace the lost alleles (Haldane, 1935). In practice, therefore, whenever analyzing a sex-linked recessive lethal disease, it is imperative to allow for $\mu = p_1/2$ in the linkage analysis, since it means that in a nuclear family with one affected child, there is therefore a 1/3 chance of his being a new mutation (the gene frequency in the parental generation being p, and the mutation rate being $p/2$). Then, the probability of any given disease allele in the next generation being a new mutation is just $(p/2)/(p + p/2) = 1/3$. This property can play a significant role in the linkage analysis. Consider the simple pedigree from chapter 6 with an X-linked recessive disease segregating. Please alter the marker locus genotype of *son4* to be *1* instead of *2*, remembering that in allele numbers format hemizygous male phenotypes must be entered as if they were homozygous with 2 copies of the allele (i.e. *1* would be coded as *1 1*). Now analyze this pedigree as in that chapter with this one marker typing change (no mutation rate). Then analyze the pedigree allowing for a mutation rate equal to $p/2$. The results should be as shown in table 21-2.

θ	Z(θ,μ=0)	Z(θ,μ = p/2)
0	-infinity	-1.2365
0.1	-0.2289	-0.2094
0.2	-0.0602	-0.0549
0.3	-0.0113	-0.0097
0.4	-0.0007	-0.0004

Table 21-2: Analysis of sex-linked disease with and without mutation rate (two affected sons).

In this example, there was little effect of allowing for the theoretical value for the mutation rate. However, if we were to make *son3* unaffected, and redo the analysis, you would see a much more pronounced effect. Clearly in this example, since there are two affected sons, it is much more likely that there was one gene segregating in the family (likelihood p = 0.01), than two mutations (likelihood = μ^2 = 0.005^2 = 0.000025), or one gene and one mutation (likelihood = $p\mu$ = (0.005)(0.01) = 0.00005). However, in the pedigree with only one affected son, the likelihood is of the order p = 0.01 for having the disease caused by a gene, and μ = 0.005 for having the disease caused by a new mutation. For this reason, μ has a much greater effect on the second one-affected child pedigree than in the two-affected child pedigree, as shown in table 21-3.

θ	Z(θ,μ=0)	Z(θ,μ = p/2)
0	-infinity	-0.0980
0.1	-0.8874	-0.0837
0.2	-0.3876	-0.0549
0.3	-0.1514	-0.0265
0.4	-0.0355	-0.0069

Table 21-3: Analysis of sex-linked disease with and without mutation rate (one affected son).

This demonstrates that the mutation rate is much more important in pedigrees with small numbers of affected individuals, as explained above. The take home message is simply that you must always allow for mutation when there is selection against the disease and in the case of sexlinked recessive lethal diseases, the mutation rate should be p/2. It is important to remember that for most disease, one should not try and estimate μ directly, but should use mutation-selection equilibrium conditions to select an appropriate value for μ.

EXERCISE 21

Reanalyze the pedigree from exercise 6, assuming the disease is fully lethal, and no affecteds live to reproduce. Then, try and work out the appropriate mutation rate for the pedigree in exercise 7, assuming that 50% of all affected individuals do not live to reproduce. Compute the appropriate mutation rate to ensure a stable mutation-selection equilibrium, and then perform an appropriate linkage analysis on this pedigree.

Design a computer experiment to show which alleles are allowed to mutate into which other alleles given the implementation of mutation in the LINKAGE programs.

22 Gene frequencies and LINKAGE

In the previous chapter, we discussed issues of mutation rates and population genetics. Now, we will discuss the relevance of gene frequency information in linkage analysis, methods for estimating them, and the consequences of using incorrect gene frequency models in a linkage analysis.

22.1 HOW ARE GENE FREQUENCIES USED IN THE LINKAGE PROGRAMS?

In the LINKAGE programs, the population frequency of each marker and disease allele is required for the computation of the likelihood. If every genotype at a locus in a pedigree is uniquely known, then the gene frequencies for that locus have no effect on the value of the lod score. However, as soon as there is one founder whose genotype cannot be uniquely determined, the gene frequencies begin to have an impact on the lod scores. This holds for both marker and disease loci - at disease loci, the gene frequencies are always important, since there rarely is a 1:1 correspondence between genotype and phenotype. The frequency of each allele can play a significant role in the analysis. For this reason, it is imperative to have good estimates for the gene frequencies of each allele at each locus, since in practice there is almost always some ambiguity in the genotypes of some individuals in most every pedigree. Further, it is imperative for investigators to uniquely identify each allele, such that for example, the *1* allele is the same allele in each pedigree, whenever some individuals are untyped at a locus, since otherwise the gene frequency isn't really appropriate.

Any deviations from the true gene frequencies can have major effects on the results and conclusions drawn from any given linkage analysis. To demonstrate this dependency, let us reconsider the example pedigree from chapter 7 about homozygosity mapping (Figure 7-2). In this pedigree there was a recessive disease with only one affected individual typed, with genotype *1 1* at the marker locus. Based on the information obtained from the consanguinity in this family, we were able to achieve a maximum lod score of 1.14 at $\Theta = 0$. However, this result was certainly heavily dependent on the gene frequencies we selected. Please reanalyze this pedigree with various gene frequencies for both disease and marker loci. Let the frequency of the disease allele, and the *1* allele at the marker locus be 0.9, 0.5, 0.1, 0.01, and 0.00001, considering all possible combinations of disease and marker allele frequencies. The resulting maximum lod scores should be as shown in table 22-1 (all with = 0).

	Frequency of Disease Allele				
Frequency of *1* allele	0.9	0.5	0.1	0.01	0.00001
0.9	0.00012	0.00338	0.01945	0.03859	0.04274
0.5	0.00102	0.02802	0.14323	0.25348	0.27468
0.1	0.00622	0.14860	0.53034	0.76720	0.80614
0.01	0.01473	0.29571	0.82723	1.10035	1.14338
0.00001	0.01706	0.32902	0.88315	1.16052	1.20401

Table 22-1: Maximum lod scores obtained for homozygosity mapping problem with different disease and marker gene frequencies

It is obvious that when few individuals in a pedigree are typed, the gene frequency plays an important role. If one looks at table 22-1, when the gene frequency of the disease allele is less than 90%, the gene frequency of the *1* allele has more effect on the analysis than the frequency of the disease allele. This is true because the penetrances at the disease locus tell us more about the untyped individuals' disease locus genotypes than we know about their marker locus genotypes. We know that the parents of the affected boy are each heterozygous for the disease allele, and that one of each of their parents also must be heterozygous. However, at the marker locus, the parents could just as easily be homozygous for the *1* allele as they could be heterozygous for it. This fact that we have less information about the genotypes of the untyped individuals makes the relevance of the gene frequency maximal. This example is quite extreme, and in practice, if your families can show such drastically different lod scores when the gene frequencies alone are altered, then the significance of your results must be highly questionable. In such a situation, the only reasonable solution may be to go out and type additional pedigree members, to reduce the dependency on gene frequency. In situations where this is not possible, you should report the lod scores for all sets of p_i

within their confidence intervals as done by Hsiao et al1 (1989), for example. As an illustration of the lack of dependency on gene frequency in fully known pedigrees, reconsider the pedigree from chapter 2, the phase unknown nuclear pedigree (figure 2-2). Analyze the pedigree with whatever gene frequencies you like, and the maximum lod score will always be 0.124929, at θ = 0.21. The absolute values of the log likelihoods themselves will change, but the value of the lod score will not.

22.2 CONSEQUENCES OF USING INCORRECT GENE FREQUENCIES

From the results of the analysis above, it is clear that the gene frequencies can play an important role in linkage analysis. Still, if one were to randomly select gene frequencies it is not clear if there would be a systematic effect on the lod score if one were to select the gene frequencies to be used in the analysis beforehand. Often investigators working with multiallelic CA repeat markers will just assume that all alleles have equal gene frequencies for the purposes of the linkage analysis. This is typically done because these markers have not been sufficiently well characterized, and accurate gene frequency estimates are unavailable. Also, it is often difficult to characterize the alleles on a population sense. For example allele *1* in family 1 may not in general correspond to allele *1* in family 2. This type of situation will further complicate the gene frequency estimation problem, since allele *1* may have a different meaning, and a correspondingly different gene frequency in the different families. In general, the only good solution to this problem would be to characterize each allele uniquely. In this way, the *1* allele would always refer to a specific allele, which should be invariant in different pedigrees. Only in this manner can one appropriately deal with the problem of gene frequency estimation, and avoid the bias associated with incorrect gene frequency modelling.

The effects of arbitrarily choosing to use equal gene frequencies in a linkage analysis was shown to lead to a systematic bias in favor of linkage when some individuals in a pedigree are untyped, or genotypes cannot be uniquely determined (Ott, 1992). In other words, setting the gene frequencies to be equal for all alleles will tend to give false positive evidence of linkage, and even to positive expected lod scores when there really is no linkage. To see this, look at the results of the homozygosity mapping exercise above. In this example, let us assume the actual gene frequency of the *1* allele was 0.9, and the gene frequency of the disease allele was 0.01. The maximum lod score with the assumed actual gene frequencies would be only 0.03859. However, if there were 10 alleles at this locus, and we assumed erroneously that their frequencies were equal, the lod score would have jumped to 0.76720. In general, there is a systematic bias, which has been extensively studied by Ott (1992), through simulations and analytical means. The take home message here is that it is always important to have accurate gene frequency estimates when there are untyped individuals in your pedigrees, or there is not a 1:1 correspondence between genotype and phenotype (i.e. dominance, recessivity, etc.).

22.3 ESTIMATION OF GENE FREQUENCIES

For many markers there are published estimates available for their gene frequencies, based on a random sample of unrelated individuals. As a first approximation one may use these values, which are readily available. However, the gene frequencies may differ strongly between populations at selectively neutral markers. Thus, it is advisable to estimate marker allele frequencies on your own from a cohort of unrelated individuals taken from the same genetic population as your disease pedigrees. If you were to consistently type 50-100 random individuals (depending on the number of alleles in your system) for each marker, and estimate the allele frequencies from these observations, you would have an accurate source of information about your specific population. This would be a very good approach to take to resolve this problem. Still, there are often situations in which this additional work is unfeasible, or in which an investigator wishes to estimate the gene frequencies based on his family data. Typically in a large pedigree there will be several founder individuals who are unrelated. These individuals are known to be from the appropriate genetic population, and can be used as a cohort for investigating the frequency of marker alleles (though certainly not for the disease allele due to ascertainment problems...). To do this, one can either simply treat them as an independent sample, and apply counting methods, as has been described elsewhere (e.g., in Hartl, 1988 or Weir, 1990). Another approach which is more powerful when some of the founders have not been typed is to use the ILINK program of the LINKAGE package to estimate the allele frequencies from the pedigree data (Boehnke, 1991).

To estimate the gene frequencies with the ILINK program, it is essential to modify the parameter file manually, since PREPLINK is not equipped for this option. Let us start with a simple example. Consider the pedigree from exercise 2. In this pedigree, there are eight founders, two of whom are untyped. If you were to directly estimate the allele frequencies based on the six typed founders, you would have four copies of the *1* allele, two copies of the *2* allele, five copies of the *3* allele, and one copy of the *4* allele, giving us gene frequency estimates of 0.3333, 0.1667, 0.4167, and 0.0833 for the four alleles respectively. However, there is some information in the pedigree about the genotypes of the two untyped founder individuals. To take advantage of it, we will use the ILINK program. Please prepare the parameter file for this example, assuming the disease locus to be fully penetrant autosomal dominant with gene frequency for the disease allele equal to 0.00001. At the marker locus, there are four alleles, and you can assume the above estimated values for their gene frequencies. Now make this parameter file in ILINK format, and write the file. It should resemble the following:

```
 2  0  0  3 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0  0.0  0.0  0 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
 1  2
1  2 << AFFECTION, NO. OF ALLELES
 9.99990E-01  1.00000E-05 << GENE FREQUENCIES
 1 << NO. OF LIABILITY CLASSES
 0  1.0000  1.0000 << PENETRANCES
3  4 << ALLELE NUMBERS, NO. OF ALLELES
0.3333  0.1667  0.4167  0.0833 << GENE FREQUENCIES
 0  0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.1000 << RECOMBINATION VALUES
 1 << THIS LOCUS MAY HAVE ITERATED PARS
 0
```

Now, you must make a few changes to this file. First, the next to last line reads :

```
1 << THIS LOCUS MAY HAVE ITERATED PARS
```

Since you wish to iterate the parameters (gene frequencies) for the second locus, change the *1* to a *2*. Next consider the bottom line of this file, which now contains a *0*. In general, the final line of this file tells the program which parameters to iterate, and which ones to keep fixed. *0* means you should keep that parameter fixed, and a *1* would mean you should estimate the corresponding parameter. If there are *n* loci in your parameter file, then the first $n - 1$ parameters will be the $n - 1$ recombination fractions. So, if you wished to estimate all recombination fractions, you would need to have $n - 1$ *1*'s on the last line. If you only wanted to estimate the first recombination fraction, and keep the others fixed, you would have a *1* followed by $n - 2$ *0*'s on the last line, etc. If you have specified a constant male/female map distance ratio, the next parameter would correspond to this ratio. If you want to estimate it, you must add an additional *1*. Otherwise, add a *0* to keep it fixed. Similarly, if you have specified variable male-female map distance ratios in each interval, then you must add an additional $n - 1$ *1*'s to estimate (or *0*'s to fix) the $n - 1$ female recombination fractions for each interval as well. If there are *m* alleles at the locus specified on the next to last line of the file, the next $m - 1$ parameters are the gene frequencies of the first $m - 1$ alleles at that locus (the last gene frequency being set equal to $1 - \Sigma p_i$; $i < m$).

So, for our problem, we would like to fix the recombination fraction between the two loci to be 0.079, as found in exercise 2 (modify the third line from the bottom to be sure the recombination value is correct), and estimate the gene frequencies for the four alleles at the second locus. Thus the last line of the parameter file should be *0 1 1 1*, since we want to fix the recombination fraction, there is no sex-difference parameter, and we wish to estimate the three free gene frequencies, with $p_4 = 1 - p_1 - p_2 - p_3$. When you have made these changes, please save this file as DATAFILE.DAT. It should look like this:

```
 2  0  0  3 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0  0.0  0.0  0 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
 1  2
1  2 << AFFECTION, NO. OF ALLELES
 9.99990E-01  1.00000E-05 << GENE FREQUENCIES
 1 << NO. OF LIABILITY CLASSES
 0 1.0000 1.0000 << PENETRANCES
3 4 << ALLELE NUMBERS, NO. OF ALLELES
0.3333 0.1667 0.4167 0.0833 << GENE FREQUENCIES
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.079 << RECOMBINATION VALUES
 2 << THIS LOCUS MAY HAVE ITERATED PARS
 0 1 1 1
```

Then, call up the UNKNOWN program. It is important that the version of UNKNOWN you use ***MUST*** be dated after July 1993 since there was a bug in earlier program versions which affected allele frequency estimation and all analyses with linkage disequilibrium (we modified the UNKNOWN program by setting the program variable *makehomozygous = false*). It is imperative to run the UNKNOWN program immediately before doing such an analysis. This is necessary for valid estimation of allele frequencies in general pedigree datasets.

Then call up the ILINK program to analyze the pedigree and estimate the allele frequencies for locus 2. It is important to note again that LCP cannot be used, since it rewrites the bottom of the parameter file. The FINAL.DAT file should look like this:

```
CHROMOSOME ORDER OF LOCI:
 1 2
****************** FINAL VALUES ***************************
PROVIDED FOR LOCUS 2 (CHROMOSOME ORDER)
***********************************************************
GENE FREQUENCIES :
0.333383 0.199991 0.399935 0.066691
***********************************************************
THETAS:
0.079
***********************************************************
-2 LN(LIKE) = 1.192555040143E+002
LOD SCORE = 1.820979412697E+000
NUMBER OF ITERATIONS = 6
NUMBER OF FUNCTION EVALUATIONS = 37
PTG = -1.875727167222E-006
***********************************************************
***********************************************************
```

So you can see that the gene frequency estimates were somewhat refined. Try and rerun the ILINK program using the newly refined estimates of gene frequency as starting values to see if they can be further refined. In this case, they cannot be further refined, and the program should return the same estimates again. At this point, two questions immediately pop into mind. First of all, we did this estimation conditional on there being linkage between marker and disease. What would happen to the estimates if we assumed that the recombination fraction between disease and marker was 50%? This would involve estimating marker allele frequencies ignoring all information about linkage. To do this, just alter the DATAFILE.DAT such that the recombination value is set to 0.5, and rerun the ILINK program. This time, the estimates have changed slightly, to 0.366901, 0.200645, 0.365811, and 0.066643 respectively, which can be further refined to be 0.366830, 0.200045, 0.366430, and 0.066695. The final thing to be considered is the possibility of jointly estimating recombination fraction with the gene frequencies, which can be done by setting the bottom line of the parameter file to be *1 1 1 1*, such that all four parameters be estimated. Using the original starting values, as shown in the parameter file above, the first estimates should be θ = 0.078, p_i = 0.333419, 0.200082, 0.399984, 0.066515, which can be further refined to be θ = 0.078, p_i = 0.333366, 0.200032, 0.399933, and 0.066669. To summarize the results of these analyses, consult table 22-2.

	θ = 0.079	θ = 0.500	θ = $\hat{\theta}$	Counting
p_1	0.333383	0.366830	0.333666	0.333333
p_2	0.199991	0.200045	0.200032	0.166667
p_3	0.399935	0.366430	0.399933	0.416667
p_4	0.066691	0.066695	0.066669	0.083333

Table 22-2: Gene frequency estimates under different hypotheses.

Here we can see that estimating the gene frequencies based solely on the marker genotypes can lead to slightly different estimates than when the gene frequencies are estimated jointly with linkage to a second locus (here the disease). Fortunately the difference is not huge, though it may have a significant influence on the lod scores in some situations. One way to correct for this difference would be to treat the marker allele frequencies as nuisance parameters in the analysis, and compute your lod score as

$$Z(\hat{\theta}) = \log_{10}[L(\hat{\theta}, \hat{p}_i)/L(\theta = ½, \hat{p}_i)],$$

where the p_i are estimated separately under linkage (in the numerator), and under no linkage (in the denominator). The numerator and denominator can be separately determined from the ILINK output. In this case, if you look at the FINAL.DAT files created by ILINK for the two appropriate analyses, you will see that $–2\ln[L(\hat{\theta}, \hat{p}_i)] = 119.255$, so $\log_{10}[L(\hat{\theta}, \hat{p}_i)] = –25.925$. Similarly, $–2\ln[L(\theta = ½, \hat{p}_i)] = 127.559$, so $\log_{10}[L(\theta = ½, \hat{p}_i)] = –27.730$. Therefore, $Z[\hat{\theta}] = –25.925 – (–27.730) = 1.8052$. This is not much different from the original lod score (assuming equal gene frequencies) of 1.78, or the lod score when gene frequencies were estimated assuming linkage (1.82), with said estimates used in both numerator and denominator of the lod score (of course, this statistic now has two degrees of freedom, since two parameters are estimated in the numerator, and none in the denominator). Since most pedigree members were typed in this example, the gene frequencies were not very crucial, while in other examples, the results may vary dramatically. If there is sufficient data, it is safest, and most conservative to ***estimate the gene frequencies separately in numerator and denominator*** of the likelihood ratio. This feature is implemented in the Pseudomarker program (section 24.7).

In conclusion, we have learned how to estimate gene frequency parameters in the LINKAGE programs, why it is important to do so, and we have examined the ramifications of using improper gene frequencies to do a linkage analysis in practical situations.

EXERCISE 22

Go back to exercise 8, and estimate gene frequencies for the ABO blood group in this same pedigree. Does the lod score change when these frequencies are estimated, instead of using population gene frequency estimates? Then, consider the incomplete penetrance model used in exercise 9 on this same family. Does incorporating this reduced penetrance affect your estimates of marker allele frequencies? How does the gene frequency information affect the lod score between ABO and the disease?

23 Linkage Disequilibrium Between Alleles at Marker Loci

Linkage disequilibrium is another population genetic phenomenon which can be useful in gene mapping. When the frequencies of pairs of alleles at different loci occurring on the same haplotype are not independent, the deviation from independence has been termed linkage disequilibrium. In this chapter, we will introduce various methods for detecting and quantifying such linkage disequilibrium, and then we will demonstrate its use in linkage analysis with the LINKAGE programs.

23.1 WHAT IS LINKAGE DISEQUILIBRIUM

For now, we will restrict ourselves to the simplest case of linkage disequilibrium between alleles of two loci with two alleles each. In general, linkage disequilibrium is usually seen as an association between one specific allele at one locus and one other specific allele at the other, so this formulation is fairly general, quantifying the association between specific alleles at each locus. If you will look at table 23-1, you will see a 2 × 2 table with each cell in the table representing one of the four possible haplotypes created by the two marker loci. The rows refer to the first marker genotype, while the columns refer to the second marker genotype. X1, X2, X3, and X4 represent the number of observations in each cell, where X1+X2+X3+X4 = *n*. The probabilities given in each cell are the probabilities of any random individual having the haplotype indicated in that cell. D_{11} (often denoted by δ), the coefficient of gametic linkage disequilibrium between allele *1* at locus *1*, and allele *1* at locus *2*, therefore is defined as $E[\, X1X4 - X2X3 \mid n = 1 \,]$.

	Marker 2	
Marker 1	Allele 1	Allele 2
Allele 1	X1 $p_1p_2 + D_{11}$	X2 $p_1(1-p_2) - D_{11}$
Allele 2	X3 $(1-p_1)p_2 - D_{11}$	X4 $(1-p_1)(1-p_2) + D_{11}$

Table 23-1: Linkage disequilibrium coefficient definitions, where p_1 = gene frequency of allele *1* at marker *1*, and p_2 = gene frequency of allele *1* at marker *2*, D_{11} = coefficient of linkage disequilibrium between allele *1* at locus *1*, and allele *1* at locus *2*.

In general, one could extend this concept to multiple alleles, and estimate haplotype frequencies for the n_1n_2 possible haplotypes, but in general, this requires a much larger sample size than the two-marker two-allele case, and the corresponding analytical approaches are analogous.

23.2 POPULATION BASED SAMPLING AND THE EH PROGRAM

The first thing one needs to do is to select a random cohort of individuals from one genetic population. What is meant by this is that the individuals should be from one (hopefully) randomly mating interbreeding unit. For example, one could assume that all of the individuals on a given Pacific island with minimal immigration comprise one homogeneous interbreeding population. Likewise, one could consider French-Canadians, or Bavarian Germans, or Transylvanian Magyars to be approximately homogeneous. Often people extend this concept with reasonable accuracy to larger groups which appear to be randomly mating.

	Locus 1		
Locus 2	AA	Aa	aa
BB	k_1	k_2	k_3
Bb	k_4	k_5	k_6
bb	k_7	k_8	k_9

Table 23-2: Table of all possible two-locus genotypes.

Let us assume we wish to test the absence of disequilibrium between allele A at locus 1, and allele B at locus 2, ($D_{AB} = 0$). Our sample of individuals consists of genotypic data, however, making it typically impossible to fully distinguish all of the haplotypes in each individual. Each individual can be classified uniquely in terms of his two-locus genotype, and can be placed into one of the cells of table 23-2. If the individual falls into cell 1, then you know it is made up of two identical A B haplotypes. Similarly, a person in cell 4 has one A B haplotype, and one A b haplotype. In almost every cell, this haplotype determination can be done uniquely, with the notable exception of cell 5, in which case there can be either of two phases, A B/a b, or A b/a B. These individuals are then rather difficult to deal with. However, it can be shown that omitting these individuals from consideration in an analysis of this type can lead to a bias, and a loss of information for the test, despite what you might think. Methods have been developed, however, to allow for these individuals to be used in the analysis, by using likelihood methods. One can simply try and maximize the log likelihood of the data observed, where $\ln [L(data)] = \sum_{i=1}^{a_1a_2} k_i \ln(p_i)$; k_i = number of observations of two-locus genotype *i*, p_i = probability of observing two-locus genotype *i*. The only remaining question is how to compute the p_i. For most cases this is quite straightforward. In cell 1, we know that there are two A B haplotypes, so p_1 = [P(A B)]^2. Similarly, in cell 4, there is one A B haplotype, and one A b haplotype, so the probability of being in this cell is just p_4 = 2P(A B)P(A b). For cell 5, however, there is some ambiguity about the phase, so P(Aa,Bb) = P(A B/a b) + P(A b/a B) = 2P(A B)P(a b) + 2P(A b)P(a B). The computation of all cell probabilities is shown in table 23-3.

	Locus 1		
Locus 2	AA	Aa	aa
BB	p(A B)^2	2p(A B)p(a B)	p(a B)^2
Bb	2p(A B)p(A b)	2p(A B)p(a b) + 2p(a B)p(a b)	2p(A b)p(a B)
bb	p(A b)^2	2p(A b)p(a b)	p(a b)^2

Table 23-3: Table of probabilities of each cell in table 23-2, parametrized by haplotype frequency.

Then, one could maximize the likelihood above, over the possible haplotype frequencies, or equivalently over the three parameters, p(A), p(B), and D_{AB}, which make up the haplotype frequencies in the two allele - two locus case. This likelihood can then be compared to the maximum likelihood when D_{AB} is set equal to 0 (i.e. absence of linkage disequilibrium). This can form the basis of a test of linkage equilibrium, and has been implemented in the linkage utility program EH (for Estimate Haplotype frequencies). Similar programs are given in Weir (1990).

Let us assume the following dataset, in the notation of table 23-2, $k_1 = 10$, $k_2 = 10$, $k_3 = 3$, $k_4 = 15$, $k_5 = 50$, $k_6 = 13$, $k_7 = 5$, $k_8 = 13$, and $k_9 = 10$. In every case, except k_5, all the haplotypes can be uniquely determined, and if you were to just count, you would find 45 A B haplotypes, 29 a B haplotypes, 38 A b haplotypes, and 46 a b haplotypes. If we were to assume that this was an exhaustive population sample of haplotypes, we could perform a chi-square test of independence of the 2 × 2 table shown below:

	A	a
B	45	29
b	38	46

In this case, $\chi^2_{(1)} = 3.83$, with a corresponding p-value of 0.05. The corresponding haplotype frequency estimates are as follows:

	A	a
B	0.284810	0.183544
b	0.240506	0.291140

Parametrizing these haplotype frequencies in terms of p(A), p(B), and D_{AB}, clearly p(A) = 0.284810 + 0.240506 = 0.525316; p(B) = 0.284810 + 0.183544 = 0.468354; D_{AB} = p(A B)-p(A)p(B) = 0.284810 – (0.525316)(0.468354) = 0.038776. However, this sample was biased due to the elimination of the fifty observations in k_5. The EH program can be used as described above to perform the appropriate analysis on all the data together. Let us set up our input file for EH as follows:

Line 1: Number of alleles at each of the 2 loci
Line 2: k_1 k_4 k_7
Line 3: k_2 k_5 k_8
Line 4: k_3 k_6 k_9

Create such a file, and name it EH.DAT. The file should look like this:

```
2 2
10 15 5
10 50 13
 3 13 10
```

Then, call the EH program by entering *EH* at the DOS prompt. Be sure to respond *n*o when the program asks if you wish to use the case-control sampling option. Then, specify the input file to be EH.DAT, and the output file to be EH.OUT, as the defaults are already setup. Then, after the program runs, you should get an output file similar to the following:

```
Estimates of Gene Frequencies (Assuming Independence)
----\----------------------------
locus \ allele 1 2
---------\------------------------
 1 | 0.5155 0.4845
 2 | 0.4806 0.5194
---------------------------------
# of Typed Individuals: 129

There are 4 Possible Haplotypes at These 2 Loci.
They are Listed Below, with their Estimated Frequencies:

-------------------------------------------------
| Allele    Allele  |   Haplotype Frequency     |
| at        at      |                           |
| Locus 1   Locus 2 | Independent w/Association |
-------------------------------------------------
 1          1           0.247762    0.327684
 1          2           0.267742    0.187820
 2          1           0.232859    0.152937
 2          2           0.251638    0.331560
-------------------------------------------------
# of Iterations = 16

                                    df    Ln(L)       Chi-square
----------------------------------------------------------------
H0: No Association                  2     -252.68     0.00
H1: Allelic Associations Allowed    3     -248.23     8.89
```

The statistic of interest here is the chi-square statistic, which is just the difference in 2*ln*(likelihood), which is 8.89. In this case, the chi-square statistic has 1 degree of freedom, because under the hypothesis of allelic association, there are three free parameters, the haplotype frequencies, while under the hypothesis of no allelic association, there are only two free parameters, the two gene frequencies. The difference in free parameters is one, so the distribution has one degree of freedom. The p-value associated with $\chi^2_{(1)} = 8.8928$ is 0.002873. Further, in comparing the haplotype frequency and δ estimates from the two approaches, with and without censoring the k_5 individuals, it is clear that they contribute significant information about disequilibrium, even though the phase cannot be uniquely determined a priori, as shown in table 23-4.

	Haplotype Frequencies			
	Without k_5		With k_5	
Haplotype	Indep.	Assoc.	Indep.	Assoc.
A B	0.246034	0.284810	0.247762	0.327684
A b	0.279282	0.240506	0.267742	0.187820
a B	0.222320	0.183544	0.232859	0.152937
a b	0.252364	0.291140	0.251638	0.331560
p(A)	0.525316		0.515504	
p(B)	0.468354		0.480621	
δ	0.038776		0.079922	

Table 23-4: Table of haplotype frequency estimates based on both methods, first by censoring the individuals with ambiguous haplotypes, and secondly be using all the data, with the EH program.

The EH program also is capable of estimating haplotype frequencies for loci with greater than two alleles, however the format for data entry is more complicated. In general, for two loci in the EH program, you must enter the data as follows:

Line 1: Number of alleles at each locus
Subsequent Lines: The number of observations of each genotype as per table 23-5 (just the numbers of observations, not the rest of the table.

	LOCUS 2					
LOCUS 1	1/1	1/2	2/2	1/3	2/3	3/3...
1/1	a1	b1	c1	d1	e1	f1
1/2	a2	b2	c2	d2	e2	f2
2/2	a3	b3	c3	d3	e3	f3
1/3	a4	b4	c4	d4	e4	f4
2/3	a5	b5	c5	d5	e5	f5
3/3	a6	b6	c6	d6	e6	f6
...						

Table 23-5 : Format for entering multiallelic genotype information in the EH program for two loci.

The table, and column/row headers are given to indicate the format in which you should enter the numbers of observations, not as something to be entered as well). The EH program could also be used analogously for estimating haplotype frequencies at more than two loci. To do this, the appropriate genotype entry format for three loci is demonstrated in table 23-6 (Of course, on line 1 you would now have three numbers of alleles, instead of two, to tell the program there were three loci in the datafile). Additional loci would be added in an analogous manner.

		LOCUS 3		
LOCUS 1	LOCUS 2	1/1	1/2	2/2
1/1	1/1	a1	b1	c1
	1/2	a2	b2	c2
	2/2	a3	b3	c3
1/2	1/1	a4	b4	c4
	1/2	a5	b5	c5
	2/2	a6	b6	c6
2/2	1/1	a7	b7	c7
	1/2	a8	b8	c8
	2/2	a9	b9	c9

Table 23-6 : Format for entering multilocus genotypic data in the EH program (here assuming two alleles per locus. Additional alleles could be dealt with in the manner outlined in table 23-5).

23.3 ESTIMATING DISEQUILIBRIUM FROM PEDIGREE DATA

The EH program incorporates a powerful and robust way to test and estimate deviations from linkage equilibrium between alleles at marker loci, based on a random sample of unrelated individuals in a fixed homogeneous population. It allows the user to include those individuals with no phase information in a disequilibrium analysis. However, if one has collected family pedigree information, it is possible to incorporate the phase information obtained from using such data in your analysis.

In the literature, some attempts have been made to look at family data, by using the founders and married-ins as the set of unrelated individuals for conducting the disequilibrium analysis (Kerem et al., 1989). The advantage here is that one could determine the phase in the doubly heterozygous individuals, allowing one to directly count haplotypes without relying on the EH algorithm. However, in these cases, in determining the phase of the haplotypes, they generally "...have assumed that there were no recombinants in our family material."(Chakravarti et al, 1984). As pointed out by Chakravarti, this is going to lead to biased results, since there is no basis for making such an assumption in all circumstances. The probability of a recombination is equal to the recombination fraction between the two loci, which is usually not zero! However, the phase information available from using such family information could be taken advantage of by using the ILINK program to estimate haplotype frequencies from the pedigree data, which would still be using the founders, but would take into account the recombination fraction between the loci, and other factors. Let us consider the pedigree drawn in Figure 23-1. *Note*: An updated and expanded version of ILINK is the Pseudomarker program [24, 25]; section 24.7.

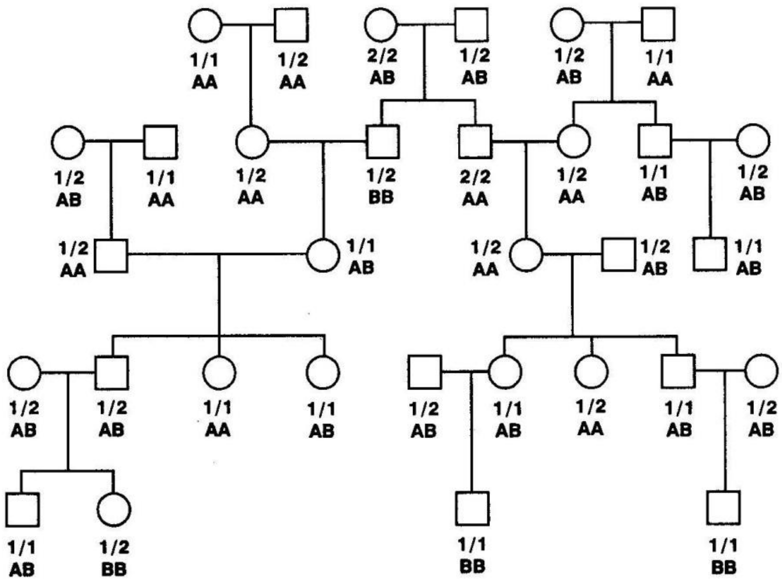


Figure 23–1. Pedigree for association analysis between two marker loci

Let us first consider the founders and married-in individuals as a cohort of unrelated individuals from this population, and analyze them with the EH program. The results should resemble the following:

```
Estimates of Gene Frequencies (Assuming Independence)
----\----------------------------
locus \ allele 1 2
---------\------------------------
 1 | 0.5769 0.4231
 2 | 0.6538 0.3462
---------------------------------
# of Typed Individuals: 13

There are 4 Possible Haplotypes of These 2 Loci.
They are Listed Below, with their Estimated Frequencies:

----------------------------------------------------
| Allele Allele | Haplotype Frequency |
| at at | |
| Locus 1 Locus 2 | Independent w/Association |
----------------------------------------------------
 1 1 0.377219 0.576921
 1 2 0.199704 0.000002
 2 1 0.276627 0.076925
 2 2 0.146450 0.346152
----------------------------------------------------
# of Iterations = 8
 df Ln(L) Chi-square
-----------------------------------------------------------------
H0: No Association 2 -22.01 0.00
H1: Allelic Associations Allowed 3 -16.00 12.02
```

Clearly, there is overwhelming evidence for disequilibrium in this example ($\chi^2_{(1)} = 12.02$, p = 0.0005), with a strong association between the 1 and A alleles. Now, let us use the full amount of information at our disposal to estimate haplotype frequencies with the ILINK program. Please enter the pedigree above in a

LINKAGE format pedigree file (both markers in allele numbers format). Then, call up the PREPLINK program, and specify the two allele numbers loci (NO DISEASE LOCUS!) with the gene frequency estimates as obtained from the EH program above. Now, try and refine the gene frequency estimates using the ILINK program, as explained in chapter 22. Do this first for locus 1, and then for locus 2, updating the frequencies for locus 1, as per your ILINK estimates. In this case, since there are no untyped people at either locus, it may be advisable to estimate the gene frequencies independently of the recombination fraction, since Θ should have no effect on the gene frequency estimates. Then, compute the lod score for this pedigree. You should find that the gene frequency estimates do not change, and are already maximized by the EH program. The maximum lod score for this pedigree, then, occurs at $\theta = 0$, with $Z(\theta = 0) = 1.2$, with corresponding –2ln(like) value of 99.48.

Now, let us use ILINK to maximize the likelihood over haplotype frequencies, to see if there is significant evidence for disequilibrium when the entire pedigree is used to establish phase for the doubly heterozygous individuals. We can further see what the effect is on the haplotype frequency estimates. First, read the parameter file back into the PREPLINK program, and this time select option (e) Haplotype Frequencies. You will be prompted with the following screen:

```
*****************************************
(a) SEE HAPLOTYPE FREQUENCIES
(b) CHANGE HAPLOTYPE FREQUENCIES
(c) HAPLOTYPE FREQUENCIES DEFINED : N
(d) RETURN TO MAIN MENU
*****************************************
```

Select option (c) to define haplotype frequencies, and for starting values, please give the haplotype frequencies estimated by the EH program (under independence, since under the hypothesis of allelic association, the frequency of the 1 B haplotype is too close to zero) as follows:

```
ENTER NEW FREQUENCY AFTER EACH "?"
NOTE THAT HAPLOTYPES ARE GIVEN USING CHROMOSOME ORDER OF LOCI
LOCUS : 1 2
ALLELES : 1 1 0.000000E+00
 ?
0.377219
ALLELES : 1 2 0.000000E+00
 ?
0.199704
ALLELES : 2 1 0.000000E+00
 ?
0.276627
ALLELES : 2 2 0.000000E+00
 ?
0.146450
ENTER c TO CONTINUE
```

Now, return to the main menu, and write the new parameter file. Then, bring the file into your word processor, and examine it. It should now resemble the following:

```
 2 0 0 3 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0 0.0 0.0 1 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
 1 2
3 2 << ALLELE NUMBERS, NO. OF ALLELES
3 2 << ALLELE NUMBERS, NO. OF ALLELES
 0.377219 0.199704 0.276627 0.146450 << HAP FREQ
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.5000 << RECOMBINATION VALUES
 2 << THIS LOCUS MAY HAVE ITERATED PARS
 0
```

Now, you should alter the bottom of the parameter file by adding three *1*'s after the *0* on the last line, to indicate that you wish to estimate the three haplotype frequencies (the fourth being equal to $1-p_1-p_2-p_3$), while fixing the recombination fraction at θ = ½. Run UNKNOWN (again, you must have the version dated after July 1993, which sets the variable *makehomozygous = false*), and then run the ILINK program, and your FINAL.DAT file should resemble the following:

```
CHROMOSOME ORDER OF LOCI:
 1 2
****************** FINAL VALUES **************************
PROVIDED FOR LOCUS 2 (CHROMOSOME ORDER)
***********************************************************
HAPLOTYPE FREQUENCIES:
 0.576458 0.000691 0.077266 0.345585
***********************************************************
THETAS:
 0.500
***********************************************************
-2 LN(LIKE) = 9.302258802337E+001
LOD SCORE = 0.00000000000E+000
NUMBER OF ITERATIONS = 10
NUMBER OF FUNCTION EVALUATIONS = 59
PTG = -7.504932736675E-005
***********************************************************
***********************************************************
```

Naturally, these haplotype frequency estimates are almost identical to those obtained from the EH program, since no information about phase is available from the rest of the family when the markers are assumed to be unlinked. However, if we were to jointly estimate the recombination fraction with the haplotype frequencies, there may be additional phase information available. To do this, read the parameter file into your word processor, and modify the bottom three lines to look like the following:

```
 0.1000 << RECOMBINATION VALUES
 2 << THIS LOCUS MAY HAVE ITERATED PARS
 1 1 1 1
```

In this way, the recombination fraction will be estimated along with the three haplotype frequencies. Now, rerun the ILINK program, and you should get the following quite different results: haplotype frequencies = 0.316321, 0.257770, 0.340819, 0.085091; θ = 0; –2ln(Like) = 97.994.

These results are very different, since upon looking at the family, it appears that in most of the heterozygous founders, the 1 allele is on the same haplotype as the B allele, while this is never seen in the individuals with unambiguous phase. To summarize the results of the haplotype frequency estimates, consult table 23-7.

	Haplotype				
Method of Estimation	1 A	1 B	2 A	2 B	D_{1A}
1) EH: Independence	0.377219	0.199704	0.276627	0.146450	0
2) EH: Association	0.576921	0.000002	0.076925	0.346152	0.200
3) ILINK: θ = 0.5	0.576458	0.000691	0.077266	0.345585	0.199
4) ILINK: θ = $\hat{\theta}$ = 0	0.316321	0.257770	0.340819	0.085091	-0.06

Table 23-7 : Haplotype frequency estimates obtained from EH and from ILINK based on figure 23-1.

Interestingly, the gene frequency estimates are almost constant in each example, and the differences in haplotype frequency can be explained entirely in terms of D_{1A}, which is indicated in the last column below. It shows that while under the hypothesis of no linkage there is evidence for a strong association between the 1 and A alleles, when analyzed under the assumption of linkage, this association disappears, and in fact a slight association is noted between the 1 and B alleles. The question still remains as to how to use this output to develop a test for the presence of linkage disequilibrium. We have four likelihoods to consider, as indicated in table 23-8.

D_{1A}	θ	-2ln(Likelihood)
0	0.5	105.01
0	$\hat{\theta}$ = 0	99.48
$\hat{D}_{1A}$ = 0.199	0.5	93.02
$\hat{D}_{1A}$ = -0.06	$\hat{\theta}$ = 0	97.99

Table 23-8 : Table of -2ln(Likelihoods) under different hypotheses about θ and D_{1A}.

Clearly, the bottom line should have a value of –2ln(Likelihood) that is less than or equal to the value on the second to last line, since these are nested hypotheses. It means that our maximization must have gotten stuck in some local maximum. Let us maximize the likelihood at various fixed values of the recombination fraction, θ. Please set up a series of ILINK analyses to compute the values of –2ln(Likelihood) for values of θ between 0 and 0.5 in steps of 0.05. Please also record what the FINAL.DAT file indicates as the "LOD SCORE" for each point. The results you obtain should be approximately the same as those in table 23-9.

Theta	D_{1A}	-2ln(Likelihood)	"LOD SCORE"	Lod Score
0	-0.060	97.98	1.98	-1.08
0.05	-0.060	99.13	1.73	-1.33
0.10	-0.061	100.15	1.52	-1.55
0.15	-0.067	101.05	1.36	-1.74
0.20	-0.072	101.86	1.21	-1.92
0.25	-0.076	102.64	1.07	-2.05
0.30	+0.200	101.61	-1.87	-1.87
0.35	+0.200	99.09	-1.31	-1.32
0.40	+0.200	96.88	-0.84	-0.84
0.45	+0.200	94.88	-0.40	-0.40
0.50	+0.200	93.02	0.00	0.00

Table 23-9 : Table of likelihoods maximized over D_{1A} for a set of fixed values of θ in the pedigree from figure 23-1.

This is a very interesting result, since it shows a dichotomy of sorts between two very different local maxima for the likelihood over D_{1A}. For small θ, D_{1A}=-0.07 is about optimal, while for large θ, D_{1A}=0.200 is optimal. It stands to reason that at some point the two must give an approximately equal likelihood, and that point should be somewhere between θ = 0.25, and θ = 0.3. It turns out that at θ = 0.275, the value of –2ln(Likelihood) with D_{1A} = –0.07 is 103.05, and with D_{1A} = 0.200, it equals 103.05, so at this point is where the global maximum over D_{1A} switches over. This is a very interesting phenomenon, and is due to the fact that in the founder individuals of this pedigree, there are seven observed 1 A haplotypes, two 2 A haplotypes, and one 2 B haplotype. This provides some evidence for a population association between the 1 and A alleles, as would be the case when D_{1A} = 0.2. However, upon closer examination of the pedigree, we can find that in the eight doubly heterozygous individuals, under the hypothesis of tight linkage, the 1 allele occurs predominantly in association with the B allele. This apparent contradiction leads us to this dichotomy between two potential estimates of the disequilibrium coefficient. If one looks at the "LOD SCORES" in table 23-9, it is clear that they too follow a similar dichotomy, being very positive until θ = 0.275, and suddenly becoming very negative. This is because these "LOD SCORES" are computed with the estimated haplotype frequencies used in both numerator and denominator. The more appropriate way to compute lod scores in such a situation would be as $\log_{10}[\ L(\hat{\theta}, \hat{D}_{1A})/L(\theta = ½, \hat{D}_{1A})]$, the values of which are indicated in the last column of table 23-9 (with D_{1A} estimated separately in numerator and denominator). In this analysis, then, while ILINK tells you there is a "LOD SCORE" of 1.98, when one correctly treats the haplotype frequencies as nuisance parameters, the lod score at θ = 0 is actually –1.08, when the analysis is done properly. This exercise points out just how important it is to do the linkage analysis carefully, and when using haplotype frequency information, one MUST be careful about re-estimating the frequencies under the hypotheses of linkage and no linkage respectively. In this case, it is also interesting to consider the effect of using the haplotype frequency estimates obtained from the EH program, and computing lod scores based solely on these estimates. In this case, using those estimates will give you a lod score at θ = 0 of –∞, with an exclusion region extending through θ = 0.275, whereas in the appropriate ILINK analysis, the lod scores started increasing again from θ = 0.275 to θ = 0, indicating that in this analysis, you make false exclusions, while by following the "LOD SCORE" values given in the ILINK program, you make **false assumptions of a positive linkage finding**.

EXERCISE 23

Compute the genotype probabilities for all possible genotypes for use in the EH program, assuming one 4-allele locus, and one 3-allele locus. Then, analyze the observations in table 23-10 with the EH program, and separately by censoring individuals with ambiguous haplotype phase.

	Locus 2									
Locus 1	1/1	1/2	1/3	1/4	2/2	2/3	2/4	3/3	3/4	4/4
1/1	10	5	6	4	1	2	3	1	2	0
1/2	6	3	3	3	1	2	1	1	2	1
2/2	12	9	8	11	3	2	5	1	0	3
1/3	1	2	2	1	1	1	1	0	4	2
2/3	0	2	2	8	2	2	9	3	6	8
3/3	8	6	4	10	3	3	8	5	9	13

Table 23-10: Table of observations for analysis with the EH program in exercise 23.

Next, consider the pedigree from exercise 8, and look for linkage disequilibrium between alleles of the ABO blood group and the other marker locus. Analyze this pedigree with ILINK to estimate haplotype frequencies for each θ in steps of 0.1 from θ = 0.1 to θ = 0.5. (*Hint*: Be sure to eliminate the disease locus from the pedigree and parameter files before you commence this analysis).

24 Linkage Disequilibrium and Disease Loci

When looking for linkage disequilibrium between a disease allele and a marker allele, one cannot apply the methods as explained in the previous chapter, due to ascertainment problems. The straight EH approach is not practical, since the population frequency of most genetic diseases is so small that in a random sample of individuals, you would not likely encounter a single affected person. It is therefore imperative to devise directed ascertainment schemes, in which your sample is enriched for the disease. In this chapter, we will introduce some basic approaches to this problem, and the ways such disequilibrium can be used in a linkage analysis with the MLINK and ILINK programs.

24.1 CASE-CONTROL SAMPLING

Perhaps the most obvious approach is to simply take a sample of unrelated individuals affected with a certain genetic disease, and compare the frequency of certain alleles with their frequency in a sample of unrelated normal individuals. If the mode of inheritance for the disease is known, then it would be possible to estimate haplotype frequencies from this type of data. However, the primary object of interest is the test of whether or not there is an allelic association. If one were to collect a sample of cases, and a sample of normal individuals from the same population, one could simply perform a chi-square test of the equality of the gene frequencies in the case and control samples. For example, consider the following set of observations:

	Marker allele	
	1	2
Case	60	40
Control	40	60

The test would be a simple chi-square test on this two by two table, and if you analyze it using the CONTING program, you will find that the value of the chi-square statistic is 8.00, with a corresponding two-sided p-value of 0.004. There is another useful linkage utility program, called 2BY2. It performs Fisher's exact test on a 2 × 2 table. In this case, you can call up the 2BY2 program, and enter the data as in the table above. You will find that the exact one-sided p-value in this case is 0.0035. Of course, this p-value is one sided, and the other is two-sided, so for a fair comparison, this p-value should be doubled, for an approximate comparison of the two approaches.

This is a simple approach, and if one was to assume that the disease were fully penetrant recessive, with very rare disease allele frequency, we can estimate haplotype frequencies for the disease-marker haplotypes as follows. Clearly, we know that everyone in the disease sample carries two copies of the disease allele, so we can say that P(allele 2 │ disease allele) = 0.4, and P(allele 1 │ disease allele) = 0.4. Assuming that we have a population-based estimate for the disease allele frequency (in this case, p = 0.001), we can compute the disease-marker haplotype frequencies as just P(1 D) = P(allele 2 │ disease allele)P(disease allele) = (0.4)(0.001) = 0.0004. Similarly, P(2 D) = 0.0006. Since the disease allele is so rare, we can safely assume that the control population consists solely of homozygous normal individuals. In this case, then, by the logic outlined above, P(1 +) = 0.5994 and P(2 +) = 0.3996.

24.2 MORE COMPLICATED PENETRANCE MODELS

If the disease were dominant, or there were phenocopies allowed for, the situation would be more complicated. To allow for these types of diseases, we will use the case-control option of the EH program, to more accurately compute haplotype frequencies, and to test linkage equilibrium given various specific disease models, when the data are sampled according to a case-control strategy, as above. The basic idea behind this program is that one would separately collect a sample of individuals with the disease and without the disease. Then, according to the penetrances and gene frequencies (which must be user specified for the disease locus only), each individual is assigned a probability of having each possible disease locus genotype. For example, if we had $f_1 = P(\text{Aff} \mid \text{DD})$; $f_2 = P(\text{Aff} \mid \text{Dd})$; $f_3 = P(\text{Aff} \mid \text{dd})$; and $p = P(\text{D})$; then, in the population, we would have prevalence, $\varphi = f_1 p^2 + 2f_2(1 - p)p + f_3(1 - p)^2$. Then, for each affected individual, we could compute $P(\text{DD} \mid \text{Aff}) = f_1 p^2/\varphi$, etc. Then, every affected individual observation among affecteds would be partitioned among the three possible genotypes according to these conditional probabilities. For example, if we had one affected individual with marker genotype *1/1*, he would be partitioned into three observations, P(DD | Aff) observations of disease locus genotype DD, marker locus genotype *1/1*, P(Dd | Aff) observations of disease locus genotype Dd, marker locus genotype *1/1*, and P(dd | Aff) observations of disease locus genotype dd, marker locus genotype *1/1*. Similar decomposition is done with the unaffecteds (who have penetrances P(NA | DD) = 1 – P(Aff | DD), etc.), with the resulting data combined across disease locus phenotypes. Thus if we had m_1 observations of affected with marker genotype *1/1*, and m_2 observations of unaffected with marker genotype *1/1*, then we would have genotype-decomposed observations of m_1P(DD | Aff) + m_2P(DD | NA) observations of disease locus genotype DD, marker locus genotype *1/1*, and so on, for the other possible disease locus genotypes. Then, this genotype-based data can be analyzed with the EH program, with the restriction that the disease allele frequency must remain equal to *p* throughout, since if we were to estimate it from the data, it would in general be vastly overestimated because of the case-control ascertainment scheme involved here.

To use this version of the EH program, you must prepare your data in exactly the same form as in the chapter 23, with the exception that you must have separate files for the case data (CASE.DAT), and the control data (CONTROL.DAT). In these files, you would indicate the numbers of observations of each genotype at the marker locus (or loci), in exactly the same format as shown in the previous chapter. The one difference is that you must now specify additionally the gene frequency of the disease allele, and penetrance values for each disease locus genotype (always assuming the disease-predisposing allele to be the first allele at the disease locus). As an example, let us consider the data shown in table 24-1.

	Marker locus genotype									
Sample	1/1	1/2	1/3	1/4	2/2	2/3	2/4	3/3	3/4	4/4
Case	13	5	11	4	0	3	2	8	10	9
Control	2	4	5	7	2	6	5	5	18	14

Table 24-1 : Dataset for Case-Control study of disequilibrium with EH.

The CASE.DAT file, for example, should resemble the following:

```
4
13 5 0 11 3 8 4 2 10 9
```

To run the program, just type EH at the DOS prompt, and when the program prompts you with

```
Do you wish to use the case-control sampling option? [Y/N]
```

Respond by entering *Y*, to invoke this option. Then, you will need to tell the program the names of the separate input files for the CONTROL sample genotypes and the CASE sample genotypes. Since we have created these two files with the appropriate default names, you need only hit the [Enter] key when prompted with

```
Enter control data file [CONTROL.DAT], and
Enter case data file [CASE.DAT].
```

The output file can also be left at the default, EH.OUT. The next phase requires us to specify various parameters about the disease locus. In this case, you need to tell the program that the gene frequency of the disease allele is *0.01*, and then you will need to specify the penetrances for each of the three possible disease-locus genotypes, in this case assuming a dominant disease with 80% penetrance, and 0.1% penetrance for phenocopies, as follows: *+/+* (= *0.001*), *+/D* (= *0.80*), *D/D* (= *0.80*). The output file, EH.OUT should resemble the following:

```
Estimates of Gene Frequencies (Assuming Independence)
(Disease gene frequencies are user specified)
----\-----------------------------------------------------
locus \ allele    1           2           3           4
---------\------------------------------------------------
Disease |         0.9900      0.0100
 1      |         0.2481      0.1090      0.2970      0.3459
----------------------------------------------------------
# of Typed Individuals: 133

There are 8 Possible Haplotypes of These 2 Loci.
They are Listed Below, with their Estimated Frequencies:

-----------------------------------------------
| Allele  Allele  | Haplotype Frequency        |
|   at      at    |                            |
| Disease Marker1 | Independent w/Association  |
-----------------------------------------------
      +     1           0.245602    0.189403
      +     2           0.107895    0.138809
      +     3           0.294060    0.286668
      +     4           0.342443    0.375120
      D     1           0.002481    0.004357
      D     2           0.001090    0.000058
      D     3           0.002970    0.003217
      D     4           0.003459    0.002368
-----------------------------------------------
# of Iterations = 23

                                   df    Ln(L)       Chi-square
-------------------------------------------------------------
H0: No Association                 3     -538.62      0.00
H1: Markers and Disease Associated 6     -532.73     11.78
```

The likelihood ratio test of linkage equilibrium between the marker and disease would then be –2ln[L(H_0)/L(H_1)] ~ χ^2, with 6 – 3 = 3 degrees of freedom. In this case, this statistic has a value of 11.78, which has an associated p-value of 0.003, which is significant evidence for linkage disequilibrium between the disease and marker alleles, with estimated haplotype frequencies shown above. Apparently the strongest association is between the disease allele and allele *1* at the marker locus.

One additional point of interest is that contrary to the situation in which sampling is random with respect to both markers, in this case, the estimated frequencies of the marker alleles is different under the assumption of linkage disequilibrium. For example, under the hypothesis of no disequilibrium, the frequency of the *1* allele is estimated to be 0.2481, while under the hypothesis of linkage disequilibrium, its frequency is estimated to be P(*1*) = P(*1 +*) + P(*1 D*) = 0.189403 + 0.002368 = 0.191771, which is much smaller. This decrease is a result of the fact that the *1* allele is associated with the *D* allele which is overrepresented in the sample, due to the case-control sampling scheme, which overrepresents haplotypes which carry the *D* allele.

This approach can be used just as well to estimate haplotype frequencies with two or more marker loci, but it is important to be careful in what you consider to be a significant result. While linkage disequilibrium at more than two loci is beyond the scope of this book, it is important to point out a few minor things, so that the unwary user doesn't misuse this method in such situations. Consider the following simple example with two marker loci. Let us assume that we observe the genotypes shown in table 24-2.

	Marker 1					
	Case			Control		
Marker 2	1/1	1/2	2/2	1/1	1/2	2/2
1/1	10	5	2	10	5	2
1/2	5	10	5	5	10	5
2/2	2	5	10	2	5	10

Table 24-2: Dataset for Case-Control study of disequilibrium with a disease and two marker loci

Clearly, there is no association between any of the marker alleles and the disease, yet there is a strong association between the alleles of the two markers. If you run the EH program on these data, you should find that there are now three likelihood values given at the bottom, $\chi^2(H_0) = 0.00$; $\chi^2(H_1) = 29.67$; and $\chi^2(H_2) = 29.67$. These values are just $-2\ln[L(H_0)/L(H_i)] = 2\ln[L(H_i)] - 2\ln[L(H_0)]$. Thus, they provide the appropriate test statistic for comparing either hypothesis against the overall null hypothesis of no association between any alleles at any of the loci. Thus, there is highly significant ($p < 0.000001$) evidence for an association between alleles of the two markers, independent of the disease, and highly significant evidence for an overall association ($p = 0.000006$). However, the most appropriate test for association between disease and alleles at one or more of the markers would be to compare H_2 with H_1. The reader may question the use of H_1 as the appropriate null hypothesis in general, especially if there were no significant evidence for rejecting H_0 in favor of H_1, but in fact, we are not interested in whether the markers are associated with each other (and typically in such a study, we would assume this to be the case), but are solely interested in whether the disease is associated with one or more of the markers. In this case the desired test is $-2\ln[L(H_1)/L(H_2)]$. This is equivalent to $\chi^2(H_2) - \chi^2(H_1) = [2\ln L(H_2) - 2\ln L(H_0)] - [2\ln L(H_1) - 2\ln L(H_0)] = 2\ln L(H_2) - 2\ln L(H_1) = -2\ln[L(H_1)/L(H_2)]$, and can thus be easily determined. In this case, clearly, there is no significance, since $\chi^2(H_2) - \chi^2(H_1) = 29.67 - 29.67 = 0$, so there is absolutely no evidence for any association between the disease allele and any allele at any of the marker loci.

24.3 THEORY BEHIND THE HAPLOTYPE RELATIVE RISK

In many studies of allelic association of the case control variety, people often question the meaning of the results, since it can be difficult to find well-matched case and control samples from the same genetic population. As a possible remedy to this problem, Rubinstein et al (1981) proposed their genotype-based Haplotype Relative Risk (GHRR) design to obtain matched case and control samples in an association study. The basic idea of their method is to collect a sample of random affected individuals, and their parents, and base the analysis solely on these small nuclear families. One would consider the affected child's marker genotype as the "case" sample, and the two parental alleles which were not transmitted to the affected child as an artificial "control" sample, obviously well matched from the same genetic population. For example, consider a family with parental genotypes G/H, and H/J, with affected son H/H, as shown in Figure 24-1.

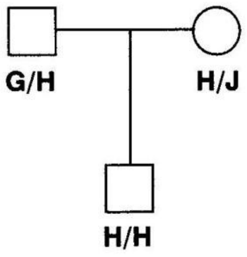


Figure 24–1. Sample HRR pedigree

In this family, the "case" genotype would be H/H, and the artificial "control" genotype would be G/J, the two alleles (one from each parent) that were not transmitted to the affected child. In the original Rubinstein et al formulation of this approach, they looked at whether or not a given allele was present or absent from each "genotype". For our case, let us look at the H allele. Clearly, there are two H alleles in the transmitted (case) sample, and no H alleles in the non-transmitted sample (control). Hence, this family would contribute one observation of H transmitted, and one observation of not-transmitted. One could then collect n such nuclear families, and obtain two such observations from each family (one transmitted, and one not-transmitted).

	H	$\overline{H}$	Total
Transmitted	W	X	(W + X)
Not-Transmitted	Y	Z	(Y + Z)
	(W+Y)	(X+Z)	2N

Table 24-3: Haplotype Relative Risk 2x2 table

Filling in table 24-3 with these observations, one can test linkage equilibrium by a simple chi-square test of independence on this table, as $\chi^2 = \frac{2N(WZ-XY)^2}{(W+X)(W+Y)(X+Z)(Y+Z)}$. It is important to point out that the theory that has been developed for the HRR is based on the rather restrictive assumption that families are singly ascertained, and that the affected child in the analysis must be the proband in each family. It is unclear what the effect of violating these rather strict rules may be, so it is suggested that if you are planning to start an association study using this approach, that you collect families based solely on this criteria. It is granted that for many complex diseases with high rates of sporadic cases, you may have a large number of sporadic cases in your sample, but if some association exists, then if you collect a sufficiently large sample, you should hopefully be able to detect it anyway. Please refer to Terwilliger and Ott (1992b) for further details about sample size and power considerations.

24.4 APPLICATION OF THE HRR AND THE CONTING PROGRAM

This test of equilibrium is very simple to apply, as any fully typed family of this type can be uniquely classified according to whether or not a given allele (here denoted H) is transmitted and not-transmitted. This can easily be done by hand, and no complicated computer software is needed. Further, there is a Linkage Utility Program called CONTING which can be used to compute the chi-square statistic for any such 2 × 2 table. Let us assume that we have collected 50 such families, with W = 40, X = 10, Y = 20, Z = 30. Now, call up the CONTING program. It will first prompt you with

```
Interactive use? [Y/n]
```

to which you should respond *Y*. Then it will ask

```
NEW TABLE:
Number of rows (0 to stop) =
```

to which you should respond that there are *2* rows. Similarly, tell the program that there are *2* columns, since we have a 2 × 2 table. You will then be asked (Your responses are indicated in italics):

```
Enter observed numbers rowwise
Row 1- 40 10
Row 2- 20 30
```

The program will then recreate the 2 × 2 table, to give you a chance to verify that you have entered the data correctly, as follows:

```
The observed figures are:
             1    2
1           40.  10.
2           20.  30.

Use Yates' correction for continuity (y/N)?
```

If the sample size is large, as in this case, you do not need to corect for continuity, so just hit enter (the upper case N indicates that this is the default). You will then see the following screen:

```
Frequencies expected under independence:
         1           2
1       30.00      20.00
2       30.00      20.00

To continue, press return key...
```

These values are the expected counts in each cell, assuming independence of the two by two table. When you hit the enter key, the results of the chi-square calculation will appear as follows:

```
The contributions to chi-square are:
      1     2
1     3.33  5.00
2     3.33  5.00

Chi-square = 16.67 1 degree of freedom
2-sided p-value = 0.000045

Collapse some rows or columns to form a new table (screen input)?
```

to which you should respond *N*, followed by

```
NEW TABLE:
Number of rows (0 to stop) =
```

You may now enter *0* to exit the program. As you can see, the null hypothesis of equilibrium is rejected by this test at the 0.000045 level, so we have very significant evidence for linkage disequilibrium. Just as an exercise, please use your calculator, and plug in the appropriate values of W, X, Y, and Z, in the chi-square formula above to verify that this result is correct. If you do this, you should get

$$\chi^2 = \frac{2(50)([40][30] - [20][10])^2}{(40+10)(40+20)(10+30)(20+30)} = 16.67,$$

which matches the result obtained using the CONTING program.

24.5 PAIRED SAMPLING AND THE CHIPROB PROGRAM

It is also possible to consider each family as contributing one observation to table 24-4, in which each family is classified in terms of both its transmitted and non-transmitted genotypes. For the sample pedigree described above, the one observation would be in cell B (H transmitted, not-transmitted). If one looks further at this table, it can be seen that the marginals of table 24-4 provide the data on which the haplotype relative risk statistic is based.

	Not Transmitted		
Transmitted	H	$\overline{H}$	Total
H	A	B	W
$\overline{H}$	C	D	X
	Y	Z	N

Table 24-4 : Paired Sampling HRR Table

One can base his tests of equilibrium on this table as well, by performing a McNemar test. Clearly, the null hypothesis of the HRR test, from the previous section, is that W = Y. But, if you look at the actual familial source of the data in the paired sampling case, you will see that this hypothesis is equivalent to (A + B) = (A + C), which implies B = C. A simple and straightforward test of this sort is a McNemar test on this paired sampling table. The McNemar test is simply $\chi^2 = \frac{(B-C)^2}{B+C}$, as described in Terwilliger and Ott (1992b).

While this test is easy to apply, and intuitively appealing, it can be shown to have uniformly lower power than tests of the HRR variety. However, as pointed out by Spielman et al. (1993), the McNemar test has the advantage that one needn't assume the presence of Hardy-Weinberg Equilibrium, which would not be present when there is population stratification. Further, the power of this test is minimally lower than the HRR tests in general, making it a useful option, when population stratification is likely.

Let us apply this test to our sample data from section 24.4. Let us fill in the missing data in Table 24-4 as A = 15, B = 25, C = 5, D = 5 (Note that W = A+B = 40, etc). Applying the more simple McNemar test, we can simply use our calculators to obtain the value $\chi^2 = \frac{(25-5)^2}{25+5} = 13.33$.

To determine the associated p-value, we must either consult a chi-square table, or use the Linkage Utility Program CHIPROB. Please call up this program now, and you will be prompted with:

```
Enter x² and df
```

At this point, you must merely enter *13.33 1*, since the value of your statistic is 13.33, and there is 1 degree of freedom. The program will then provide you with the appropriate two-sided p-value of 0.000263. To exit the CHIPROB program, just enter a 0 at the next prompt. This p-value is still highly significant, but less significant than the HRR statistic applied to the same data.

24.6 HAPLOTYPE-BASED HAPLOTYPE RELATIVE RISK

Terwilliger and Ott (1992b) developed a way to use this same HRR experimental design which would glean much additional information from the same dataset. In the case of the original genotype-based HRR (GHRR) statistic of Rubinstein et al (1981), they lumped together *H*/*H* homozygotes, and $H/\overline{H}$ heterozygotes as *H* genotypes. However, Terwilliger and Ott noticed that since under the null hypothesis, the two parental genotypes are independent, the transmitted and non-transmitted alleles from each parent can be treated as independent observations, and thus supply us with four observations per family, in what they termed the haplotype-based HRR (HHRR) statistic, as opposed to the two observations obtained in the GHRR approach. Returning to the sample pedigree above, we can see that it would now contribute two observations of *H* transmitted (*H*/*H*), and two observations of $\overline{H}$ not-transmitted (*G*/*J*). In this case, the same statistic can be applied, only now *N* refers to the total number of parents, as opposed to the total number of families, and is thus twice as large as it was under the GHRR method described above. Going back again to our sample dataset of 50 families, if we break it down into haplotypes, we might have found the following data (in form of Table 24-4 above): A = 19; B = 42; C = 10; D = 29. This would mean that W = 61; X = 39; Y = 29; Z = 71. If we then were to compute the chi-square statistic associated with this table (as described above for the GHRR), we would find that the HHRR was equal to $\chi^2 = \frac{2[100]([61][71] - [39][29])^2}{(61+39)(61+29)(39+71)(29+71)} = 20.69$. The p-value obtained from CONTING (or equivalently from CHIPROB) is 0.000005 which is much stronger than the result obtained from the GHRR approach. Analogously, one could apply the McNemar test to this data as well, and the haplotype-based McNemar test would give a result of $\chi^2 = \frac{(42-10)^2}{42+10} = 19.69$. Again, this test gives somewhat lower significance than the HHRR statistic, but higher than the genotype-based McNemar test. In general, the HHRR approach has been shown by Terwilliger and Ott (1992b) to provide better power than the GHRR approach almost uniformly, and is thus more useful to apply in general, for both HRR and McNemar type statistics.

It is also important to be careful about multiple comparisons (cf. Ott (1991), sec. 4.7), in evaluating the significance of any given test result, since people often consider each marker allele separately against the disease allele. In these cases, one should divide the critical p-value by the number of comparisons done, to adequately allow for the multiple testing problem (Anderson and Sclove, 1986).

24.7 USING ILINK TO ESTIMATE LD WITH DISEASE, Pseudomarker

Of course, the ILINK program can be used to estimate haplotype frequencies, as we saw in chapter 23. However, there are problems with this estimation when one of the loci involved is the disease locus. The basic problem is that the disease allele will necessarily be overrepresented in the pedigree dataset. One could estimate haplotype frequencies, ($\underline{D\ i}$), and then normalize them, a posteriori, to the known population disease allele frequency as P($\underline{D\ i}$) = $p_D[(\underline{D\ i})/\Sigma_i(\underline{D\ i})]$, for example. Still, the estimates may not be accurate, since the constraint on disease allele frequency was made after the maximization process, and not before it. This limitation of the ILINK program makes this approach somewhat unreliable. Further, it is generally better to estimate your haplotype frequencies from one dataset, and then use this information in your subsequent pedigree analysis. The method for using the ILINK program to estimate haplotype frequencies in general was illustrated in chapter 23, and the normalization process is analogous to what was done in the EH program.

A specialized version of ILINK is the Pseudomarker program [24, 25]. It allows for the joint testing for linkage and/or disequilibrium, properly accounting for ascertainment of disease. See its online user manual.

24.8 USING LINKAGE DISEQUILIBRIUM IN THE LINKAGE PROGRAMS

It is possible to use information about linkage disequilibrium in the LINKAGE programs, to do a standard linkage analysis. The use of such haplotype frequency information can have a strong effect on the linkage analysis results, since it makes the prior probabilities of the possible parental phases unequal in an otherwise phase-unknown mating. Let us consider the two pedigrees shown in figure 7-3, from the section on marriage loops. If you remember correctly, these pedigrees were completely uninformative for linkage, since they are both phase-unknown matings with only one offspring each. What would happen to this analysis, if we were to allow for the presence of linkage disequilibrium?

In these pedigrees, we know the disease is recessive, and that each parent is heterozygous at the disease locus. So, let us first consider the first pedigree, with parents *2/3*, and *1/1*. In this pedigree, the only potentially informative meiosis is from *father*, the *2/3* individual. There are two possible phases for this parent, 2 D/3 +, and 2 +/3 D. The likelihood of these two phases are L_1 = P(2 D)P(3 +) and L_2 = P(2 +)P(3 D) respectively. Under phase I, the affected daughter would be a recombinant, and under phase II, she would be a non-recombinant, so the likelihood of this pedigree is just $L_1\theta + L_2(1 - \theta)$. Under the assumption of no linkage disequilibrium, $L_1 = L_2$ by definition, so the overall likelihood of this pedigree is just L_1, which is independent of θ, and thus provides no information about linkage. However, whenever there is linkage disequilibrium, and $L_1 \neq L_2$, this likelihood is a function of θ, and therefore provides information about linkage. For the second pedigree, by analogy, we have phases for *father* of 1 D/2 + or 1 +/2 D, with corresponding likelihoods L_3 = P(1 D)P(2 +) and L_4 = P(1 +)P(2 D), and overall pedigree likelihood of $L_3(1 - \theta) + L_4\theta$. Our lod score, for the two pedigrees together, is therefore equal to $Z(\theta) = \log_{10}[L_1\theta + L_2(1 - \theta)] - \log_{10}[½(L_1 + L_2)] + \log_{10}[L_3(1 - \theta) + L_4\theta] - \log_{10}[½(L_3 + L_4)]$. Let us consider the haplotype frequency information given in table 24-5.

Haplotype	Model 1	Model 2	Model 3	Model 4
D 1	0.25*p*	0.01*p*	0.98*p*	0.01*p*
D 2	0.40*p*	0.98*p*	0.01*p*	0.01*p*
D 3	0.35*p*	0.01*p*	0.01*p*	0.98*p*
+ 1	0.25(1-*p*)	0.25(1-*p*)	0.25(1-*p*)	0.25(1-*p*)
+ 2	0.40(1-*p*)	0.40(1-*p*)	0.40(1-*p*)	0.40(1-*p*)
+ 3	0.35(1-*p*)	0.35(1-*p*)	0.35(1-*p*)	0.35(1-*p*)
L_1	k(0.40)(0.35)	k(0.98)(0.35)	k(0.01)(0.35)	k(0.01)(0.35)
L_2	k(0.40)(0.35)	k(0.40)(0.01)	k(0.40)(0.01)	k(0.40)(0.98)
L_3	k(0.25)(0.40)	k(0.01)(0.40)	k(0.98)(0.40)	k(0.01)(0.40)
L_4	k(0.25)(0.40)	k(0.25)(0.98)	k(0.25)(0.01)	k(0.25)(0.01)

Table 24-5 : Haplotype frequency models for analysis of pedigrees from figure 7-3, where *k* = *p*(1-*p*).

Under these four models (the first of which represents the situation where there is no linkage disequilibrium, and the other models represent extremely strong associations between the disease allele and one of the marker alleles), the lod scores are very different, due to the incorporation of phase information. Let us compute the lod scores at various values of the recombination fraction. Note that the constant *k* can be dropped from each of the L_i, since they are likelihoods, each divided by the same constant $k = p(1 - p)$. Under model 1, the lod score is just $\log_{10}[(0.40)(0.35)\ \theta + (0.40)(0.35)(1 - \theta)] - \log_{10}[½((0.40)(0.35) + (0.40)(0.35))] + \log_{10}[(0.25)(0.40)\ \theta + (0.25)(0.40)(1 - \theta)] - \log_{10}[½((0.25)(0.40) + (0.25)(0.40))] =$

$\log_{10}[(0.40)(0.35)] - \log_{10}[(0.40)(0.35)] + \log_{10}[(0.25)(0.40)] - \log_{10}[(0.25)(0.40)] = 0$, for all θ. For the other models, the analytically computed lod scores are given in table 24-6.

	Lod Scores			
θ	Model 1	Model 2	Model 3	Model 4
0	0.000	-3.130	0.326	0.387
0.1	0.000	-1.307	0.275	0.325
0.2	0.000	-0.761	0.219	0.258
0.3	0.000	-0.428	0.156	0.182
0.4	0.000	-0.188	0.084	0.097

Table 24-6 : Lod Scores computed for various values of θ, on pedigrees from figure 7-3, with haplotype frequency models given in table 24-5.

To analyze the same pedigrees with the LINKAGE programs, it is necessary to specify haplotype frequencies in the parameter file. To do this, use the pedigree and parameter files used in chapter 7 to analyze these two pedigrees, files EX6A.*. Read the EX6.DAT file into PREPLINK, and select the *(e) Haplotype frequencies* option, followed by *(c) HAPLOTYPE FREQUENCIES DEFINED*. Then, you will be prompted with the following screen:

```
ENTER NEW FREQUENCY AFTER EACH "?"
NOTE THAT HAPLOTYPES ARE GIVEN USING CHROMOSOME ORDER OF LOCI
LOCUS : 1 2
ALLELES : 1 1 0.0000000000000E+00
 ?
```

Then, you will have to input the appropriate frequency for the haplotype containing allele *1* at locus *1*, and allele *1* at locus *2*. In this case, locus *1* is the disease locus, and at the disease locus, allele *2* is the disease allele, with frequency 0.00001. From table 24-5, you can see that this haplotype frequency *1 1* (under model 1) is just (0.25)(0.99999) = 0.2499975. Then, enter the appropriate frequencies for each of the other haplotypes. It is imperative to realize that the order of loci (for haplotype frequency computation purposes) is not the order of loci in the parameter file, but is rather the user-specified locus order. For example, if we had specified locus order *2 1*, then we would treat the marker as locus *1*, and the disease as locus *2*, even though the disease was still the first locus in the parameter file. When you have entered the appropriate haplotype frequencies, set up the file to analyze the pedigrees in MLINK format starting a recombination fraction of 0, in steps of 0.1, stopping at θ = 0.4, and then save this file as EX6A1.DAT. It is important to note that LCP cannot be used when you are using linkage disequilibrium in the analysis. Therefore, you must copy EX6A.PED to PEDFILE.DAT, and EX6A1.DAT to DATAFILE.DAT, and run the UNKNOWN and MLINK programs. Again, UNKNOWN versions from Columbia University dated after July 1993 must be used, especially when dealing with multipoint data, as homozygous *1 1* individuals would have different disease locus genotype probabilities than unknown individuals, based on the linkage disequilibrium information. If one is considering multipoint analysis, and the other loci are informative, making all individuals in a pedigree *1 1* will affect the results of the multipoint linkage analysis. Further, in risk calculations, the homozygosity at the marker locus *can and will* affect the genetic risk to the proband in the presence of linkage disequilibrium. All the lod scores, for each of the four models should be identical to those shown in table 24-6, with the analytical computations.

In this small example, you can see the potential effect of allowing for linkage disequilibrium in your linkage analyses. One example for which this has been applied with greatly increased power was in a recent investigation of myelin basic protein and multiple sclerosis (Tienari et al, 1992), in which case the lod score rose from 1.64 to 3.42, when linkage disequilibrium was allowed for. Of course they first had to prove that linkage disequilibrium existed, and then had to estimate the haplotype frequencies as well.

EXERCISE 24

Let us consider the data from the HHRR analysis above. Suppose we wanted to estimate haplotype frequencies on this dataset. Let us assume the disease to be a fully penetrant recessive disorder, with gene frequency 0.01 for the disease allele. Try and design an experiment to use this parametric information to help you test linkage equilibrium, and to estimate the haplotype frequencies in this dataset.

25 Parametric Analysis of Complex Diseases

In this chapter, we will be introducing the most basic approaches to linkage analysis of complex diseases. We will briefly consider how one could go about selecting a model to use in the analysis. Further, we will be discussing methods of affecteds-only analysis, and the benefits of doing same, and finally we will introduce the problem of inflation of the maximum lod score due to maximizing the lod score over different models for the disease.

25.1 COMPLEX DISEASES

Complex disease is a broad designation which basically covers any disease which we cannot accurately define or explain. This covers a large gamut of possible problems or complexities. One generally thinks of diseases with unknown mode of inheritance, especially polygenic models, or other modes of inheritance which cannot be simply fit to a reasonable single locus model. Another thing which can make a disease "complex" would be the case where one doesn't really know who is "affected" with the disease, or at least who would be potential carriers of a specific genetic defect. This is commonly the case with psychiatric disorders, where one cannot really distinguish one genetically relevant phenotype from another with any great accuracy. It is often hoped that discovery of the genetic cause(s) of these diseases will allow for better definition of the disease phenotypes, and allow researchers to better determine what affection really means. This is in a sense "reverse genetics" carried to an extreme. Other things which may make a disease "complex" are genetic heterogeneity (both allelic and non-allelic), which will be discussed specifically later, the possibility of large rates of sporadic non-genetic causes of the same (or similar) disease phenotypes. The list goes on and on, and all of these things have the basic effect of making fully parametrized likelihood analysis very difficult and error-prone. A large variety of possible methods of dealing with these problems has been proposed, but none of them is completely satisfactory, and much work will likely be required to develop more efficient approaches to these problems, but at least these approaches give us a starting point towards gross-scale localization of genes which play some sort of role in the etiology of these diseases.

In this chapter, we will consider some simple approaches to the problem. The analyses will be done on a well-known dataset, the schizophrenia pedigrees of Sherrington et al (1988), with two markers on chromosome 5, as shown in Figure 25-1, with disease and marker phenotypes indicated in table 25-1. We will use this example to explain various potential analysis techniques for linkage with complex diseases. We wish to point out that we have selected this dataset to illustrate some techniques for the analysis of complex diseases, and not to be critical of the analyses performed in the study in question. We will make a point of not trying to recreate the analyses that were performed by Sherrington et al, but rather to start from scratch, and illustrate some potential methods for the analysis of such a dataset. The authors feel that it is useful to demonstrate these techniques with a real dataset, and further, in this example, the disease is truly complex, in that we have no real knowledge about the mode of inheritance, or the correct diagnostic criteria, and thus it serves as a useful vehicle to illustrate the primitive approaches we will be considering in this chapter.

		Diagnosis Under Scheme			Marker Phenotype	
Ped	Person	1	2	3	M 1	M 2
1	A	Aff	Aff	Aff	1 2	1 3
1	B	N/A	N/A	N/A	1 2	3 3
1	C	N/A	N/A	Aff	2 2	3 3
1	D	Aff	Aff	Aff	2 2	3 3
1	E	Aff	Aff	Aff	2 2	3 3
1	F	Aff	Aff	Aff	2 2	3 3
1	G	???	???	???	0 0	0 0
1	H	???	???	???	0 0	0 0
1	I	N/A	N/A	N/A	1 2	1 3
1	J	N/A	N/A	N/A	1 2	3 3
1	K	N/A	N/A	N/A	1 2	3 3
1	L	Aff	Aff	Aff	1 2	3 3
1	M	Aff	Aff	Aff	1 2	3 3
1	N	N/A	N/A	N/A	2 2	1 3
1	O	N/A	N/A	N/A	1 2	1 3
1	P	Aff	Aff	Aff	2 2	1 3

2	A	Aff	Aff	Aff	0 0	0 0
2	B	N/A	N/A	N/A	0 0	0 0
2	C	N/A	Aff	Aff	1 1	0 0
2	D	N/A	N/A	N/A	1 1	0 0
2	E	Aff	Aff	Aff	1 2	2 3
2	F	N/A	N/A	N/A	1 2	1 3
2	G	N/A	N/A	N/A	1 2	2 3
2	H	Aff	Aff	Aff	1 1	1 3
2	I	Aff	Aff	Aff	1 1	1 3
2	J	N/A	N/A	N/A	0 0	0 0
2	K	N/A	Aff	Aff	1 1	2 3
2	L	Aff	Aff	Aff	1 1	1 3
2	M	N/A	N/A	N/A	1 2	1 3
2	N	N/A	N/A	N/A	0 0	0 0
2	O	N/A	Aff	Aff	0 0	0 0
2	P	N/A	Aff	Aff	0 0	0 0
2	Q	N/A	N/A	N/A	0 0	0 0
3	A	N/A	N/A	N/A	1 1	1 3
3	B	N/A	N/A	N/A	0 0	0 0
3	C	Aff	Aff	Aff	1 1	1 3
3	D	N/A	Aff	Aff	1 1	1 3
3	E	N/A	N/A	N/A	1 2	1 3
3	F	Aff	Aff	Aff	1 1	1 1
3	G	N/A	N/A	N/A	1 2	1 1
3	H	N/A	N/A	Aff	1 1	1 3
3	I	N/A	N/A	N/A	1 1	3 3
3	J	N/A	Aff	Aff	1 1	1 3
3	K	N/A	N/A	N/A	1 1	1 2
3	L	N/A	Aff	Aff	1 1	1 3
3	M	N/A	N/A	N/A	1 1	1 3
3	N	N/A	N/A	N/A	1 1	1 3
3	O	N/A	N/A	Aff	1 1	3 3
3	P	N/A	N/A	Aff	1 1	1 3
3	Q	Aff	Aff	Aff	1 1	3 3
3	R	N/A	N/A	N/A	1 1	1 1
3	S	N/A	N/A	N/A	1 1	1 1
3	T	N/A	N/A	Aff	1 2	1 1
3	U	Aff	Aff	Aff	1 1	3 3
3	V	N/A	N/A	N/A	1 1	1 3
3	W	Aff	Aff	Aff	0 0	0 0
3	X	N/A	N/A	N/A	1 1	1 3
3	Y	Aff	Aff	Aff	0 0	0 0
3	Z	N/A	N/A	N/A	1 1	1 3
3	AA	Aff	Aff	Aff	1 1	3 3
3	BB	Aff	Aff	Aff	1 1	3 3
3	CC	N/A	Aff	Aff	1 1	3 3
3	DD	Aff	Aff	Aff	1 1	1 3
3	EE	N/A	N/A	N/A	1 1	2 3
3	FF	N/A	N/A	Aff	1 1	2 3
3	GG	N/A	N/A	N/A	1 1	1 3
3	HH	N/A	N/A	Aff	1 1	2 3
3	II	N/A	N/A	N/A	1 1	2 3
4	A	N/A	Aff	Aff	0 0	0 0
4	B	N/A	N/A	N/A	0 0	0 0
4	C	N/A	N/A	N/A	0 0	0 0
4	D	N/A	N/A	Aff	0 0	0 0
4	E	N/A	N/A	Aff	0 0	0 0
4	F	N/A	N/A	N/A	0 0	0 0
4	G	Aff	Aff	Aff	1 1	3 3
4	H	N/A	N/A	N/A	1 2	3 3
4	I	Aff	Aff	Aff	1 2	3 3
4	J	Aff	Aff	Aff	1 1	3 3
4	K	Aff	Aff	Aff	0 0	0 0
4	L	N/A	N/A	N/A	1 2	3 3
4	M	N/A	N/A	N/A	1 2	3 3

4	N	N/A	N/A	N/A	1 2	3 3
4	O	N/A	N/A	N/A	2 2	3 3
4	P	Aff	Aff	Aff	1 2	1 3
4	Q	Aff	Aff	Aff	1 2	1 3
5	A	N/A	N/A	N/A	0 0	0 0
5	B	N/A	N/A	N/A	0 0	0 0
5	C	N/A	N/A	N/A	1 2	2 3
5	D	Aff	Aff	Aff	0 0	0 0
5	E	N/A	N/A	N/A	1 2	2 2
5	F	N/A	N/A	N/A	1 2	2 2
5	G	N/A	N/A	N/A	0 0	0 0
5	H	Aff	Aff	Aff	1 1	2 3
5	I	N/A	N/A	N/A	0 0	0 0
5	J	N/A	N/A	N/A	0 0	0 0
5	K	N/A	N/A	N/A	1 1	2 3
5	L	N/A	N/A	N/A	1 1	2 3
5	M	N/A	N/A	N/A	1 1	3 3
5	N	Aff	Aff	Aff	1 2	2 3
5	O	Aff	Aff	Aff	1 1	2 2
5	P	N/A	N/A	N/A	1 2	2 3
5	Q	N/A	N/A	N/A	1 2	2 3
5	R	Aff	Aff	Aff	1 1	2 3
6	A	N/A	N/A	N/A	0 0	0 0
6	B	N/A	N/A	N/A	0 0	0 0
6	C	N/A	N/A	N/A	0 0	0 0
6	D	Aff	Aff	Aff	1 1	1 3
6	E	Aff	Aff	Aff	1 1	1 3
6	F	N/A	N/A	N/A	1 1	1 3
6	G	N/A	N/A	N/A	0 0	0 0
6	H	N/A	N/A	N/A	1 1	1 1
6	I	N/A	N/A	N/A	1 1	1 3
6	J	N/A	N/A	N/A	1 2	3 3
6	K	N/A	N/A	N/A	1 2	1 2
6	L	Aff	Aff	Aff	1 2	2 3
6	M	Aff	Aff	Aff	1 2	2 3
6	N	Aff	Aff	Aff	1 2	3 3
6	O	N/A	N/A	N/A	1 2	1 3
6	P	Aff	Aff	Aff	1 2	1 3
6	Q	N/A	N/A	N/A	1 2	1 2
6	R	N/A	N/A	Aff	1 2	1 3
6	S	N/A	N/A	N/A	0 0	0 0
6	T	N/A	Aff	Aff	1 1	3 3
6	U	Aff	Aff	Aff	1 2	3 3
6	V	N/A	N/A	N/A	0 0	0 0
6	W	N/A	N/A	N/A	1 2	1 3
6	X	N/A	N/A	N/A	1 1	1 3
6	Y	N/A	N/A	N/A	1 2	1 3
6	Z	N/A	N/A	N/A	1 1	1 3

Table 25-1: Disease and marker phenotypes for people in schizophrenia pedigrees from figure 25-1. At the disease locus, ??? = Unknown; N/A = Not affected; Aff = Affected. Marker locus phenotypes are given in allele numbers format.

25.2 ENTERING DATA FOR MULTIPLE DIAGNOSTIC SCHEMES

The first task that presents itself is to come up with a way to enter the data on these pedigrees in LINKAGE format. The problem is that in this dataset there are three possible diagnostic schemes to be considered. In order to make the analysis as simple as possible, we recommend entering the data in such a way that the different diagnostic schemes can be taken into account solely by making modifications to the parameter file, without modifying the pedigree file. (It may be useful for you to review the affection status locus type before continuing, to make sure you fully understand the definition of the *2* phenotype [see chapter 10]).

You must define the trait as an affection status locus in such a way that the same trait definition for each individual can be used under each model. This can be difficult, for example, when you wish to consider an individual to be affected under one diagnostic criteria, and unaffected under another. To this end, we propose the following approach. In this case, we are dealing with three diagnostic schemes (and any number of penetrance models). We can set up an affection status locus with multiple liability classes to handle this situation. For example, consider the 8 possible cross-scheme phenotype vectors for each individual as shown in table 25-2.

Scheme 1	Scheme 2	Scheme 3	Affection	Liability Class
Aff	N/A	N/A	2	1
Aff	Aff	N/A	2	2
Aff	N/A	Aff	2	3
Aff	Aff	Aff	2	4
N/A	Aff	N/A	2	5
N/A	Aff	Aff	2	6
N/A	N/A	Aff	2	7
N/A	N/A	N/A	2	8

Table 25-2 : List of all possible diagnostic categories for a disease with three diagnostic schemes, and its liability class representation.

Under this classification scheme, if an individual was considered to be affected under diagnostic scheme 1, yet unaffected under diagnostic schemes 2 and 3, he would be coded as *2 1*, or presence of the phenotype defined in liability class 1. Then, you would define the penetrances accordingly for each model, as you will see below. In general, it is not necessary to have liability classes corresponding to phenotypes that do not exist in your dataset. Since affected and not affected are complementary phenotypes (i.e. P(Aff│genotype) = 1 – P(NA│genotype), you can immediately cut the number of liability classes required in half. Consider the rearrangement of table 25-2 shown in table 25-3.

			Original		New	
Scheme 1	Scheme 2	Scheme 3	Affection	L.C.	Affection	L.C.
Aff	N/A	N/A	2	1	2	1
N/A	Aff	Aff	2	6	1	1
Aff	Aff	N/A	2	2	2	2
N/A	N/A	Aff	2	7	1	2
Aff	N/A	Aff	2	3	2	3
N/A	Aff	N/A	2	5	1	3
Aff	Aff	Aff	2	4	2	4
N/A	N/A	N/A	2	8	1	4

Table 25-3 : All possible diagnostic categories for three diagnostic schemes expressed in terms of four liability classes.

As you can see in that table, we have matched up pairs of complementary phenotype definitions, so we could define both of them with only one liability class. Consider liability class 1. The phenotype defined by this is (Aff, N/A, N/A) [The ordered triple refers to the diagnosis of an individual in this liability class under (diagnostic scheme 1, diagnostic scheme 2, diagnostic scheme 3)]. Therefore under each diagnostic class, the phenotype *2* would mean presence of the indicated phenotype (Aff or N/A) for each scheme. Similarly, the phenotype *1* would mean absence of the indicated phenotype (Aff, or N/A) for EACH scheme. Thus, since Aff is the indicated phenotype for diagnostic scheme 1, N/A would be the absence of the indicated phenotype in diagnostic scheme 1. This same rule must hold for ALL diagnostic schemes. In this case, the complementary phenotype to (Aff, N/A, N/A) would be (N/A, Aff, Aff) as indicated in table 25-3. Remember that the phenotype *2* DOES NOT, IN GENERAL, MEAN AFFECTED, but merely indicates that the phenotype is defined by the given penetrances, while *1* means the phenotype is defined by the complement of the given penetrances.

Let us go back to figure 25-1. You will note that there are three diagnostic schemes to be considered, with the phenotypes given in separate columns for each individual for each diagnostic scheme, in table 25-1. In this case, the ONLY four categories to be considered (because only four occur) are (Aff, Aff, Aff), (NA, Aff, Aff), (NA, NA, Aff), and (NA, NA, NA). We must now determine the number of liability classes we will need to define these models. Since there are four categories present in the data, let us start out by allowing for the four classes as shown in table 25-3 (You may, of course, allow for eight liability classes, but it is more efficient to streamline the analysis and allow for the minimum required number).

At first glance, we can see that it is clear that diagnostic classes 1 and 4 are complementary, so we can eliminate liability class 4, and just code those individuals as *1 1*. You should only have three diagnostic classes remaining, 1 = (Aff, Aff, Aff), 2 = (Aff, Aff, NA), and 3 = (Aff, NA, NA). Then (Aff, Aff, Aff) is the complement of (NA, NA, NA), which would be coded as phenotype *1 1*, as shown in table 25-4.

Scheme 1	Scheme 2	Scheme 3	Phenotype	Liability Class
Aff	Aff	Aff	2	1
NA	Aff	Aff	2	2
NA	NA	Aff	2	3
NA	NA	NA	2	4 (or 1 1)

Table 25-4: All observed diagnostic categories for the Schizophrenia pedigrees of Sherrington et al, and their liability class representation.

Now, please create the required pedigree file for this pedigree, SCHIZO.PED. Remember that for each individual in figure 25-1, the following information is given in table 25-1: their diagnostic status under each diagnostic criterion, in order (first column = scheme 1 diagnosis, second column = scheme 2 diagnosis, third column = scheme 3 diagnosis, fourth column = marker 1 phenotype, fifth column = marker 2 phenotype). Again, we are not trying to emulate the original analysis, and the diagnostic assignments in table 25-1 may not correspond exactly to what was used in the original study.

The next difficult task is the definition of the penetrances, since we have different meanings for our phenotypes under each diagnostic scheme. For the sake of illustration, let us consider a dominant disease with penetrance 0.6, and no phenocopies. The penetrances for such a trait would be as shown in table 25-5.

	Penetrances for Genotypes		
Phenotype	D/D	D/+	+/+
Affected	0.6	0.6	0.0
Unaffected	0.4	0.4	1.0

Table 25-5 : Penetrances for different phenotypes for a dominant disease with 60% penetrance.

As you can see, affected and unaffected are complementary phenotypes, with “unknown” being model-independent. Now, how do we combine penetrance models and diagnostic schemes to create an appropriate datafile? In diagnostic scheme 1, in liability classes 1, 2, and 3 the *2* phenotype defines the *affected* phenotype. Thus, our penetrances for the three liability classes would be as shown in table 25-6. Similarly, for diagnostic scheme 2, and a recessive disease with 40% penetrance for homozygous gene carriers, and 1% penetrance for everyone else, the liability class penetrance definitions are shown in table 25-7. Finally, for diagnostic scheme 3, the penetrances for a dominant disease with 20% penetrance for gene carriers, and a 5% penetrance for non-gene carriers are shown in table 25-8.

Liability Class	D/D	D/+	+/+
1	0.6	0.6	0.0
2	0.4	0.4	1.0
3	0.4	0.4	1.0

Table 25-6 : Liability class definitions for the dominant disease with 60% penetrance in diagnostic scheme 1.

Liability Class	D/D	D/+	+/+
1	0.4	0.01	0.01
2	0.4	0.01	0.01
3	0.6	0.99	0.99

Table 25-7 : Liability class definitions for a recessive disease with 40% penetrance for genetic cases, and 1% penetrance for phenocopies in diagnostic scheme 2.

Liability Class	D/D	D/+	+/+
1	0.2	0.2	0.05
2	0.2	0.2	0.05
3	0.2	0.2	0.05

Table 25-8: Liability class definitions for a dominant disease with 20% penetrance for genetic cases, and 5% penetrance for phenocopies in diagnostic scheme 3.

25.3 CHOOSING AN APPROPRIATE ONE-LOCUS PARAMETRIC MODEL

In general, the selection of model parameters is best left to the segregation analyst. We do not wish to discuss this complicated topic in this book about linkage analysis, but would rather refer you to other more appropriate sources for segregation analysis modelling. (Elston et al, 1986; Elandt-Johnson, 1971) We will assume that you have knowledge of some simple population parameters from other sources, including segregation analyses, etc, in our attempts to help you select analysis models in a primitive sense. It has been pointed out that if there is linkage, and your model is not completely out of whack, you should be able to detect the linkage, with somewhat reduced power, of course, and also, that there is no increase in type I error rates when an analysis is done under an incorrect model (Clerget-Darpoux et al, 1986), with few exceptions (Terwilliger et al, 1991). In general, with a complex disease, people usually want to try a dominant model, and a recessive model, since one is typically unclear about the overall mode of inheritance, and since there is always the possibility that multiple loci (some dominant, some recessive) are working epistatically to cause some disease phenotype, and you are interested in detecting any of the loci involved. Therefore, one typically tries at least one model of each variety. The selection of the penetrance values is the only remaining variable. To choose an appropriate penetrance model, the most important thing to know is the ratio of penetrances for phenocopies to genetic cases ($k = f_p/f$ from chapter 9). If the penetrance is assumed to be age dependent, then the penetrance ratio would most likely be variable with respect to age as well. (Typically, one assumes that those with later age of onset have a greater ratio than those with low age of onset, who are more likely to be genetic cases.) For the moment, let us assume that the ratio, k, is constant (if it is not, you should use some lifetime penetrance ratio for the remainder of the computations in this chapter). If this is the case, then our population prevalence, φ, of the disease should satisfy the equation $\varphi = f[\text{P(susceptible genotype)} + k\text{P(non-susceptible genotype)}]$ (see chapter 10). If we were considering a dominant disease, $\varphi_d = f[p(2 - p) + k(1 - p)^2]$, and if the disease is recessive, $\varphi_r = f[p^2 + k(1 - p^2)]$. The gene frequency, p, and the overall penetrance for susceptible genotypes, f, are the only parameters to be specified. For any given value of p, f can be uniquely determined, and vice versa. Quantities like f can often be approximately obtained from segregation analysis, while p is typically more easily estimable from population data. The value of k can either be estimated through segregation analysis, or through some population based analysis. If one estimates, for example, that 50% of all cases of a disease are non-genetic, then one could use this information as well, since it would mean that $kf[\text{P(non-susceptible genotype)}] / \varphi = R = 0.50$, so $kf\text{P(non-susceptible genotype)} = \varphi R$, and $\varphi = f\text{P(susceptible genotype)} + \varphi R$, or $\varphi = f\text{P(susceptible genotype)}/(1 - R)$, which can be another useful parametrization of the prevalence, in which $f_p = \varphi R/\text{P(Non-susceptible genotype)}$. In most cases, it is easier to obtain an estimate of R, the proportion of all cases in the population due to non-genetic causes (eg. Merette et al, 1992). Given this value, and the overall prevalence of the disease, one can either determine the gene frequency, p, from a given value of f, or the penetrance, f, from a given value of p, by the equations above.

We will assume certain values of p, and then determine f from them. Clearly, when the diagnostic criteria are changed, then the prevalence values, φ, and prevalence ratios, R, will change as well. In our example, let us assume values of φ for each diagnostic level as follows: $\varphi_1 = 0.01$, $\varphi_2 = 0.015$, $\varphi_3 = 0.025$, with prevalence ratios of $R_1 = 0.35$, $R_2 = 0.5$, and $R_3 = 0.65$, and $p_{dom} = 0.01$, and $p_{rec} = 0.1$. For each diagnostic model, please determine the appropriate penetrance values for the analysis from the equations above assuming the disease to be alternatively dominant (*p* = 0.01) and recessive (*p* = 0.1). The analysis parameters should match those in table 25-9. Then make parameter files, SCHIZO#.DAT, where # ranges from 1-6 for the six models to be considered from table 25-9. For the first marker, use gene frequencies of 0.33 for the *1* allele, and 0.67 for the *2* allele, and at the second locus, use gene frequencies of 0.32 for the *1* allele, 0.16 for the *2* allele, and 0.52 for the *3* allele. Then perform the appropriate two-point linkage analyses with the disease versus each of the two markers separately. The analysis results should match those in table 25-10. It is important to remember that one should NOT do multipoint analysis with a complex trait (see chapter 18), due to the increased propensity for false negative results when there are model misspecifications (as there always will be with analysis of a complex trait).

Diagnostic	Scheme	R	φ	p	k	f_{DD}	f_{D+}	f_{++}
Dominant	1	0.35	0.01	0.01	0.010909	0.33	0.33	0.0036
	2	0.50	0.015	0.01	0.020263	0.38	0.38	0.0077
	3	0.65	0.025	0.01	0.037727	0.44	0.44	0.0166
Recessive	1	0.35	0.01	0.1	0.005385	0.65	0.0035	0.0035
	2	0.50	0.015	0.1	0.010133	0.75	0.0076	0.0076
	3	0.65	0.025	0.1	0.018636	0.88	0.0164	0.0164

Table 25-9 : Penetrance models based on prevalence, and ratio of prevalences for genetic cases and non-genetic cases.

	Diag. Scheme 1		Diag. Scheme 2		Diag. Scheme 3	
Model	θ	Lod Score	θ	Lod Score	θ	Lod Score
Dominant						
Marker 1	0.0	1.640431	0.0	2.266583	0.0	2.652390
	0.1	1.255017	0.1	1.831763	0.1	2.122298
	0.2	0.815333	0.2	1.216497	0.2	1.433496
	0.3	0.395659	0.3	0.609274	0.3	0.742590
	0.4	0.098284	0.4	0.159752	0.4	0.203170
ILINK:	0.001	1.638094	0.001	2.265304	0.001	2.650069
Marker 2	0.0	-0.590675	0.0	0.942746	0.0	1.815745
	0.1	0.633358	0.1	1.666246	0.1	2.197464
	0.2	0.760761	0.2	1.422284	0.2	1.752781
	0.3	0.506081	0.3	0.854349	0.3	1.008521
	0.4	0.161641	0.4	0.255371	0.4	0.294816
ILINK:	0.171	0.778730	0.104	1.666898	0.078	2.217621
Recessive						
Marker 1	0.0	0.155426	0.0	0.812548	0.0	1.658007
	0.1	0.940115	0.1	1.765145	0.1	2.124254
	0.2	0.751805	0.2	1.371401	0.2	1.652050
	0.3	0.393696	0.3	0.740799	0.3	0.923150
	0.4	0.101407	0.4	0.205094	0.4	0.267005
ILINK:	0.105	0.940433	0.090	1.770275	0.077	2.148463
Marker 2	0.0	-3.638244	0.0	-2.800846	0.0	-3.993544
	0.1	-1.409702	0.1	-0.843994	0.1	-1.311147
	0.2	-0.540226	0.2	-0.255292	0.2	-0.494319
	0.3	-0.171925	0.3	-0.067066	0.3	-0.164924
	0.4	-0.033334	0.4	-0.013796	0.4	-0.035273
ILINK:	0.5	0.000000	0.5	0.000000	0.5	0.000000

Table 25-10: Results of analysis of schizophrenia pedigrees under six selected penetrance models.

25.4 INTERPRETING MAXIMIZED OVER MODELS LOD SCORES

In this exercise, our "maximized-over-models" maximum lod score was 2.65 between marker 1 and disease (under dominant model with diagnostic scheme 3), and 2.22 between marker 2 and disease (with the same model). It has been repeatedly demonstrated that although analysis of a pedigree set under one fixed wrong model does not lead to an increased false positive rate (Clerget- Darpoux et al, 1986), maximizing the lod score over different models does lead to an "inflation" of the maximum lod score (Weeks et al, 1990a). It has been suggested that an appropriate correction factor for this maximization would be to no longer accept a lod score of 3 as a critical value for declaring a linkage result significant, but rather to use 3 + log(n), where n is the number of models tested (Kidd and Ott, 1984). Weeks et al (1990a) did a complex simulation study on this same set of pedigrees, and found that the inflation of the maximum lod score corresponded almost exactly to this theoretical approximation. In light of this, it seems prudent to adopt this criterion for the declaration of linkage in a complex disease. Thus, for our example, we would have needed a lod score greater than 3 + log(6) = 3.78.

An additional problem remains, however. In a normal, well-characterized Mendelian disease, the critical value of $Z_{max} > 3$ as a test for linkage is robust to multiple testing, since as one finds negative test results with more markers, the prior probability of linkage to the remaining markers is increased sufficiently to offset the increased probability of finding a significant result by chance. In other words, if one has eliminated 50% of the genome, the prior probability of linkage is twice as high for the remaining markers than it would have been before any of the genome had been excluded, since it is known with certainty that the gene is somewhere, and that the model is correct, so that the disease would be detected if you examined a truly linked marker. This increased prior probability of linkage offsets the effect of testing multiple markers. Of course if you test 20 markers, and each one of them has a probability of 0.001 of having a significant linkage result by chance, then the probability that at least one of the 20 markers has a significant result by chance would be approximately 0.02. However, while the prior probability of linkage of one marker is low, the probability that one of 20 markers is truly linked is somewhat larger, to such a degree that the phenomena of increased prior probability of linkage and multiple testing tend to offset each other. In a complex disease, in contrast to the situation above, there is no guarantee that there is truly a disease gene, nor is there truly a guarantee that it would be detected in our analysis, since we are knowingly using incorrect models in the analysis. As a matter of fact, linkage analyses of complex traits are often carried out with the stated purpose of showing the existence of major genes by virtue of significant evidence for linkage (how could one have linkage if there isn't really a gene...). In light of this, it may be prudent to allow for some correction for multiple testing. Given the past history of linkages with psychiatric diseases that have had very "significant" lod scores which disappeared under further scrutiny, it may be prudent to insist on such a correction for multiple markers. On average, 100 independent markers should be enough to cover most of the genome. If the genome is assumed to be approximately 4-5000 cM long (Weissenbach et al, 1992), and markers which are 40-50 cM apart are presumed to be approximately independent, then 100 independent markers would cover most of the genome. Of course more markers may be used in an analysis, but additional markers are no longer independent of one another. If one were to apply the correction for 100 markers by the 3 + log(*m*) criterion, where *m* now refers to the number of markers, we would be starting out, for a genome wide search, with a critical value of 3 + log(100) = 5 (for current thinking on this see [26]). If we were to adequately allow for the multiple models as well, we would have a conservative critical value of 5 + log(*n*), where *n* is the number of models tested. This may seem like a very strict criterion to declare a linkage test significant in these diseases where the power of the test is going to be reduced significantly to begin with because of the complexity of the diseases involved, but given the past history of psychiatric genetics, and the multiple sources of random error involved, it is arguable that this is a reasonable correction factor to apply in general for complex diseases, since when one starts a linkage study, they are typically planning to go until they find the gene, meaning that basically they would test all 100 markers, barring a significant finding. For example, for the Sherrington et al study, if we allow for the 18 models used in the analysis, and use a base line threshold of 5, the critical value for declaring a linkage significant would be 5 + log(18) = 6.25. The actual maximum lod score in that study was 6.49, which would still be marginally significant, but much less so than when it was compared with a critical value of 3. A better solution than using such stringent lod score criteria is to perform computer simulation under the null hypothesis of no linkage.

25.5 COMBINING DIAGNOSTIC CRITERIA IN A SINGLE MODEL

It may be possible to combine the three diagnostic criteria in a single analysis model to reduce the number of models used in the analysis (it is *very important* to note that any analysis attempted must be included in the total number of models considered, even if its results are not reported in the publication!). Looking back to section 10.4, we can see how we modelled the penetrances for a disease with uncertainty of diagnosis. We could combine these three diagnostic schemes by saying that people in diagnostic class *1* are definitely affected, those in diagnostic class *2* have a certain certainty, p_2, of being affected (this should be based on some prior belief about the true nature of these spectrum phenotypes), and those in diagnostic class *3* have another probability, p_3, of being caused by the same genetic defect (cf. Ott, 1993b). In this way, we could construct a penetrance model for the disease giving greater weight to those in the first diagnostic level, and lower weight to the diagnoses for the remaining individuals. Let us assume that $p_2 = 0.85$, and $p_3 = 0.70$, and consider the dominant and recessive penetrance models obtained in table 25-9 under diagnostic scheme 2. In this case, for our three liability classes, we would have penetrances

$$f = p_i \mathrm{P}(\mathrm{Aff} \mid \mathrm{genotype}) + (1 - p_i)\mathrm{P}(\mathrm{N/A} \mid \mathrm{genotype})$$

for each genotype. Our final liability class models are indicated in table 25-11. Note that the basic effect is to increase the penetrance ratios f_p/f towards 1, as more diagnostic uncertainty is introduced. In these models, the results given in the text assumed that $p_3 = 0.6$ under the dominant model, and $p_3 = 0.70$ in the recessive model. Making this change should give you the penetrance values given in table 25-11

Clearly if there were a 50% chance that an individual were affected, the penetrance ratio would be 1. Can you show this mathematically? The results of this analysis are presented in table 25-12. In this analysis, our maximum lod score with marker 1 was only 2.16, and with marker 2 it was 1.43. This may seem to be a loss of information, but let us consider this value relative to the critical limit for declaring a linkage significant under each situation. In the first case, we had maximum lod scores of 2.65 and 2.22 respectively. If we assume the critical value to be 5 + log(n), the critical limit would have been 5 + log(6) = 5.78, while in the example with multiple diagnostic criteria combined in one analysis model, the critical limit would have been only 5 + log(2) = 5.30. Thus, with marker 1, we gained 0.49 units of lod score by trying three diagnostic models. However, the critical value when all three models were used was 0.48 units higher. Therefore, you can see that when the diagnoses are combined into one model, in this specific weighting of the different diagnostic models, there is no overall loss in significance, and yet the actual analysis is simplified substantially.

	Dominant Model			Recessive Model		
Liability Class	DD	D+	++	DD	D+	++
1	0.38	0.38	0.0077	0.75	0.0076	0.0076
2	0.416	0.416	0.1554	0.675	0.1553	0.1553
3	0.476	0.476	0.4015	0.55	0.4015	0.4015

Table 25-11 : Penetrance model for multiple diagnostic schemes in one analysis based on probability (from prior belief) that people in a certain diagnostic class are truly affected by the same genetic disease.

	Dominant Model		Recessive Model	
Marker	θ	Lod Score	θ	Lod Score
1	0.0	2.162044	0.0	1.095595
	0.1	1.747951	0.1	1.807159
	0.2	1.164273	0.2	1.365761
	0.3	0.583685	0.3	0.730016
	0.4	0.152360	0.4	0.200834
ILINK:	0.001	2.160694	0.080	1.828390
2	0.0	0.743953	0.0	-3.568465
	0.1	1.428744	0.1	-1.103271
	0.2	1.221847	0.2	-0.428977
	0.3	0.726209	0.3	-0.153162
	0.4	0.218206	0.4	-0.033850
ILINK:	0.108	1.430856	0.5	0.000000

Table 25-12 : Analysis results of penetrance models in table 25-11.

25.6 AFFECTEDS ONLY ANALYSIS

It has been thought that many of these so-called complex diseases are actually produced by a combination of different genes working together to produce a phenotype. If this is the case, then many unaffected individuals may possess the disease-predisposing genotype at one of the loci, but lack the required second-disease locus genotype necessary for development of the disease. For this reason, it is often advisable to do these linkage analyses considering all unaffected individuals to actually have "unknown" phenotype. In this way, one bases the linkage analysis solely on the marker status of the affected individuals in the pedigree, and doesn't apply any disease-locus genotypic information whatsoever to the unaffected individuals. There are two ways in which one could go about performing such a linkage analysis. The first way is obviously to go back and alter your pedigree file such that all unaffected individuals are given the unknown phenotype. In the parameter files, then, for phenotypes that would've corresponded to unaffected (i.e. in diagnostic scheme 2, people with phenotype *2 3* would be unaffected), we replace the penetrances with those for unknown individuals (0.5 for all three genotypes, for example). To do this, you must do an astronomical amount of file manipulation, changing all *1 1* individuals in the pedigree file to *0 1*, and changing the penetrances for liability classes *2* and *3*, such that when these individuals are unaffected, the penetrances in the parameter files are equal for all three genotypes. There is a simpler way to do it, however, which is to simply reduce the maximum penetrance values for affection to 0.001, and keep the penetrance ratios the same as they were in the original files, as shown in table 25-13 for the penetrance values shown in table 25-9.

Diagnostic Class		k	f_{DD}	f_{D+}	f_{++}	$(1-f_{DD})$	$(1-f_{D+})$	$(1-f_{++})$	$(1-f_p)/(1-f)$
Dominant	1	0.010909	0.001	0.001	(0.001)k	0.999	0.999	0.99999	1.001
	2	0.020263	0.001	0.001	(0.001)k	0.999	0.999	0.99998	1.001
	3	0.037727	0.001	0.001	(0.001)k	0.999	0.999	0.99996	1.001
Recessive	1	0.005385	0.001	(0.001)k	(0.001)k	0.999	0.999995	0.999995	1.001
	2	0.010133	0.001	(0.001)k	(0.001)k	0.999	0.99999	0.99999	1.001
	3	0.018636	0.001	(0.001)k	(0.001)k	0.999	0.99998	0.99998	1.001

Table 25-13 : Affecteds only penetrances (for affecteds, *f*, and unaffecteds, *1-f*) for the models outlined in table 25-9, with penetrance ratios for affected and unaffected individuals indicated.

In this way, for affected individuals, the likelihoods will remain the same, down to a constant multiplier, which will disappear in the likelihood ratio (see section 9.2) for each individual affected. However, for unaffected individuals, the penetrances are essentially equal for all three genotypes, with a penetrance ratio in each case of 1.001. If you remember, the penetrance ratio for unknown individuals is 1.000, so this parametrization is essentially equivalent to making all unaffected individuals unknown in the pedigree file, as described above. To see this, please apply both of these methods to the analysis of the schizophrenia pedigrees under diagnostic scheme 3, with both the dominant and recessive models. The results of these analyses are shown in table 25-14. There are some differences in the lod scores between these two methods, but the change is only in the second or third decimal place, and as such has no important effect on the interpretation of the results. There are of course two possible sources of error here. The first is that our penetrance ratio in unaffecteds is 1.001, not 1.000. Similarly, we have rounded off our penetrances to five significant digits for the non-susceptible genotypes, making some alteration in this penetrance ratio as well. One potential method for making the two penetrance ratios even more accurate would be to just divide the penetrances by 1000, such that, for the dominant model in diagnostic scheme 3, our penetrances would be 0.00044, 0.00044, and 0.0000166. In this case we would preserve the penetrance ratio among affecteds exactly, and reduce the penetrance ratio for unaffecteds to only 0.9999834/0.99956 = 1.0004. Dividing the penetrances by 1000 will provide more than sufficient accuracy in any realistic situation.

	TRUE Affecteds Only		TABLE 25-13 Penetrances	
Model	θ	Lod Score	θ	Lod Score
Dominant				
Marker 1	0.0	2.222397	0.0	2.223299
	0.1	1.683718	0.1	1.684594
	0.2	1.092772	0.2	1.093428
	0.3	0.544293	0.3	0.545285
	0.4	0.144506	0.4	0.144608
ILINK:	0.001	2.219163	0.001	2.220066
Marker 2	0.0	1.389189	0.0	1.389983
	0.1	1.355322	0.1	1.356693
	0.2	1.008209	0.2	1.009340
	0.3	0.549278	0.3	0.549901
	0.4	0.156570	0.4	0.156734
ILINK:	0.039	1.431084	0.039	1.432282
Recessive				
Marker 1	0.0	1.066889	0.0	1.068376
	0.1	1.019994	0.1	1.021295
	0.2	0.697922	0.2	0.698752
	0.3	0.351886	0.3	0.352275
	0.4	0.094701	0.4	0.094798
ILINK:	0.036	1.115353	0.036	1.116852
Marker 2	0.0	0.234950	0.0	0.233835
	0.1	0.280799	0.1	0.280290
	0.2	0.203024	0.2	0.202794
	0.3	0.104595	0.3	0.104506
	0.4	0.028112	0.4	0.028090
ILINK:	0.064	0.290322	0.064	0.289653

Table 25-14 : Comparison of results of the two methods of affecteds-only analysis using diagnostic scheme 3.

EXERCISE 25

Please consider these pedigrees again, using different starting information. Repeat the linkage analyses in this chapter assuming prevalences $\varphi_1 = 0.005$, $\varphi_2 = 0.01$, $\varphi_3 = 0.03$, prevalence ratios of $R_1 = 0.10$, $R_2 = 0.35$, $R_3 = 0.50$, and gene frequencies of $p_{dom} = 0.01$, $p_{rec} = 0.125$. Please determine the appropriate penetrance models for an affecteds-only analysis with each model (divide the true penetrances by 1000), and perform the linkage analysis with it.

Consider weighting the diagnostic criteria accordingly such that those in class *1* are considered to be affected with 99% certainty, those in diagnostic class *2* are affected with 80% certainty, and those in diagnostic class *3* are affected with 65% certainty. Then perform a regular linkage analysis, using as base-line penetrances those derived from the population characteristics of the intermediate diagnostic model *2* above, under both recessive and dominant models. Do whatever analyses are needed to compare these results with the results obtained when the lod score is maximized over models.

26 Non-parametric Methods of Linkage Analysis

In this chapter you will be introduced to non-parametric approaches to linkage analysis, so-called affected sib-pair and affected pedigree member methods. These methods are based on the concepts of identity-by-descent and identity-by-state. We will introduce the basic methods of sib-pair and affected pedigree member analysis, and then the extended sib-pair analysis (ESPA) computer program (no longer available). Further, power considerations will be discussed as well, in terms of study design.

26.1 IDENTITY BY DESCENT vs. IDENTITY BY STATE

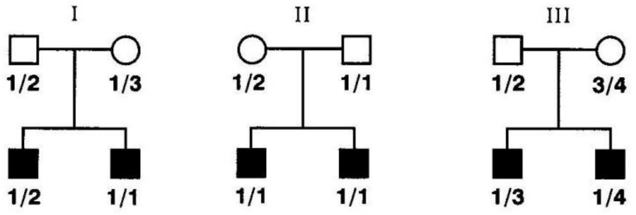


Figure 26–1. Example pedigrees for sib-pair analysis

Any two copies of allele *1* at a given locus are considered to be *identical by state* (IBS), but only copies of allele *1* that are inherited from a common ancestral source are said to be *identical by descent* (IBD). Of course, if two alleles are IBD, then they are definitely also IBS, but the inverse is not necessarily true. Consider pedigree I in Figure 26-1, with father *1/2*, and mother *1/3*. If they had children with marker genotypes *1/2*, and *1/1*, then clearly there is one allele identical by descent, since the latter child had to receive a *1* allele from each parent, one of which is IBD with the *1* allele the first child received. However, if the children were *1/2*, and *1/3*, then there are zero alleles IBD, since the first child received the *1* allele from his mother, while the second child received the *3* allele from his mother. Similarly, the first child had to have received the *2* allele from his father, while the second child received the *1* allele from his father. Thus, while the two children share a *1* allele IBS, the *1* alleles in question came from different ancestral sources, and thus are not IBD. In pedigree II in figure 26-1, you see mother *1/2*, and father *1/1*, with sons *1/1*, and *1/1*. Clearly the sons share two alleles IBS. Also, they must both have received the same *1* allele from the mother, and therefore share one allele IBD. However, there is no information about IBD status of the paternally derived alleles, as there is no way to tell which *1* allele is which. So, there is a 50% chance that the *1* alleles are IBD, and a 50% chance that they are not. One could either delete the paternal information (scoring the sib-pair as one IBD out of one opportunity), or give it a 50-50 weighting, calling the sib-pair as having 1.5 IBD alleles out of two opportunities. Clearly there are advantages to each approach. The former allows us to give the most accurate interpretation of what we know, but the latter allows us to use our full sample size. Even further complications can arise in situations where you have two parents, each with genotype *1/2*. If the children are homozygous, then IBD counts can easily be determined accurately. However, if the children are BOTH heterozygous, then one cannot tell whether there are two alleles IBD, or none. As long as at least one sib in each pair is homozygous, there is no ambiguity, but when they are heterozygous, there is ambiguity, and effectively no information is available about the sib-pair's IBD status.

26.2 AFFECTED SIB-PAIR ANALYSIS

The concept of affected sib pair analysis is that if a given marker is co-segregating with a disease predisposing allele, then affected siblings of affected persons are more likely to have received the same allele identical by descent at a closely linked marker locus than if the marker locus was segregating independently (i.e. is unlinked) to the disease predisposing allele. While this may seem very similar to the parametric idea of counting recombinants and non-recombinants, the main difference is that in this type of analysis, no assumptions are required about the mode of inheritance (however, some mode of inheritance is *implied* [27]). In this sense, sib-pair type methods are more robust than parametric methods, since one does not have to rely on as many potentially erroneous model assumptions in the analysis. Further, the problem of trying multiple models, and correcting for inflation of the lod score as is often required in such cases, is avoided in sib-pair approaches, although multiple diagnostic schemes must still be corrected for in sib-pair analyses. The basic fundamental idea is that any two siblings are expected to have one allele identical by descent (IBD). However, when the sibs are both affected with a given disease, and the analysis is done with a marker tightly linked to the disease predisposing gene, then one would expect them to share more than one allele IBD. Let us consider the third example shown in figure 26-1.

In this family, there are two parents with marker genotypes *1/2*, and *3/4*. They have two sons affected with a given disease, with the first son having marker genotype *1/3*, and the second having genotype *1/4*. Then, it is clear that they both got the *1* allele from their father, while one son got the *3*, and the other got the *4* from their mother. Hence, these two children share one allele IBD. If we consider all possible combinations, we can see the possible outcomes shown in table 26-1.

SON1	SON2	IBD	SON1	SON2	IBD	SON1	SON2	IBD	SON1	SON2	IBD
1/3	1/3	2	1/4	1/4	2	2/3	2/3	2	2/4	2/4	2
1/3	1/4	1	1/4	1/3	1	2/3	2/4	1	2/4	2/3	1
1/3	2/3	1	1/4	2/4	1	2/3	1/3	1	2/4	1/4	1
1/3	2/4	0	1/4	2/3	0	2/3	1/4	0	2/4	1/3	0

Table 26-1 : Possible arrangements of marker genotypes of 2 affected sons with parents 1/2 and 3/4

If there was no association between the disease and the marker locus, then each of these genotype combinations would be equally likely. There are four combinations with two alleles shared IBD, eight combinations with one allele shared IBD, and four combinations with no alleles shared IBD. If we compute the expected (average) number of alleles shared IBD, we would get $(4{\cdot}2 + 8{\cdot}1) / 16 = 1$. However, if there was complete linkage, at $\theta = 0$, and the disease were dominant, then assuming, without loss of generality, that the disease came from the father, and was on the same haplotype with the *1* allele, we would have the only possibilities being cases in which both sons carry a *1* allele. So, we would have (*1/3*,*1/3*), (*1/3*,*1/4*), (*1/4*,*1/3*), and (*1/4*,*1/4*) as possible genotypes for the two sons. Each of these would be equally likely, and our expected number of alleles IBD would be $(2{\cdot}2 + 1{\cdot}2\ ) / 4 = 1.5$. Clearly, this is greater than 1, and could be detected in a sib-pair analysis with a large enough sample size. Similarly, if the disease were recessive, with the maternal disease allele on the same haplotype with the *3* allele, and the paternal disease allele in coupling with the *1* allele, and $\theta = 0$, we would only have the combination (*1/3*,*1/3*) possible. Thus we would expect to find two alleles IBD. This is a complete association which could be picked up in the analysis. Clearly most real situations are somewhere in the middle between these two extremes, and often we do not even know what the true disease model is. It can be demonstrated that when the model is known correctly, that parametric analysis is more powerful than sib-pair methods, but when the disease model is not known, and we have some complex disease, that sib-pair methods can perform as well as or better than linkage analysis with an incorrect model, though the power of a sib-pair test is always greatest when the disease is recessive. Simple sib-pair analysis can be done by hand, while more complicated variations of the technique can be analyzed with such programs as SAGE (Elston et al, 1986), for example.

26.3 AFFECTED PEDIGREE MEMBER METHOD

In the affected sib-pair method, we based our statistical analysis on how many alleles a given sib-pair shared IBD. Since each sib-pair shares either 0, 1, or 2, and this can often be determined with little difficulty, the statistical analysis was quite simple. This is true because regardless of the gene frequencies, or the mode of inheritance of the disease, under absence of linkage, you expect every sib-pair to share, on average, one allele IBD. In extended pedigrees, however, the situation is much more complicated, and it has been proposed that one look not at IBD relationships but rather at IBS relationships for distantly related individuals (Weeks and Lange, 1988). Again, they would tend to share alleles IBS at loci linked to disease-predisposing loci, because of an increased probability that they inherit the marker locus IBD with the disease locus, or because the marker locus itself is contributing to the disease phenotype in some manner. Since some individuals are more distantly related, it is not typically possible to determine whether alleles are inherited IBD or not, so one must resort to comparing IBS status. This can be much more complicated, however, in a statistical sense.

The statistic proposed by Weeks and Lange (1988) is called the similarity statistic (Z_{ij}) for two affected relatives. Basically for any individual, you can order their alleles without loss of generality in either direction. Do this for both affected relatives, such that the genotype of the first affected relative would be (A_1,A_2), and the genotype of the second affected relative would be (B_1,B_2). Then, looking at pairs of alleles from the two relatives, there are 4 options, (A_1,B_1), (A_1,B_2), (A_2,B_1), and (A_2,B_2). One can then define a function δ(x,y) such that if x and y represent the same allele IBS, then δ(x,y) = 1, otherwise δ(x,y) = 0. For example, if the first affected relative had ordered genotype (2,3) and the second relative had ordered genotype (1,3), then $\delta(A_1,B_1) = \delta(2,1) = 0$; $\delta(A_1,B_2) = \delta(2,3) = 0$; $\delta(A_2,B_1) = \delta(3,1) = 0$; and $\delta(A_2,B_2) = \delta(3,3) = 1$. The similarity statistic Z_{ij} is then defined as the average of these 4 possible δ functions for the two affected relatives in question; $Z_{ij} = \frac{1}{4}\sum_{a=1}^{2}\sum_{b=1}^{2} \delta(A_a, B_b)$. For our sample affected relative pair, the similarity statistic would be simply ¼(0 + 0 + 0 + 1) = ¼. One can then sum the similarity statistics for all possible affected relative pairs in an extended pedigree to compute an overall similarity statistic

$$Z = \sum_{i=1}^{s-1} \sum_{j=(i+1)}^{s} Z_{ij} .$$

This statistic is simple and easy to compute for any given pedigree. However, it lacks one desirable property. Clearly, it is much more striking for two relatives (especially distant relatives) to share a very rare allele IBS than it is for them to share a very common allele IBS. Some allotment for this should be made, and it was proposed to consider using some sort of a weighted average to make this contrast possible within the APM method. Weeks and Lange (1988) proposed altering the statistic by adding in a weight term as follows: $Z_{ij} = \frac{1}{4}\sum_{a=1}^{2}\sum_{b=1}^{2}\delta(A_a, B_b)\, f(A_a)$, where $f(A_a)$ is the weight, and is based on the gene frequency of allele a in person A. Three main weights have been proposed by Weeks and Lange, $f_1(A_a) = 1$ (equal weights); $f_2(A_a) = \frac{1}{p_{A_a}}$; $f_3(A_a) = \frac{1}{\sqrt{p_{A_a}}}$. These latter two weights will give more strength to any observed sharing of rare alleles than sharing of common alleles, and will alleviate the intuitive problem of the sharing of common alleles being equally significant with the sharing of rare ones. In our sample statistic, assuming gene frequency of 0.3 for allele 3 (the only one shared IBS by our two affected relatives), our weighted similarity statistics would be Z_1 = ¼(0 + 0 + 0 + 1) = 0.25; Z_2 = ¼(0 + 0 + 0 + 1(1/0.3)) = 0.833; Z_3 = ¼(0 + 0 + 0 + $1/\sqrt{0.3}$ 3) = 0.456.

The similarity statistics are simple and easy to compute, as seen above. However, it is more difficult to determine the null hypothesis expected value of each of these similarity statistics. It is necessary to go through a complicated series of calculations, based on the theory of extended pedigree IBD relationships, to determine the null hypothesis mean and variance for each weighted similarity statistic. Then, one must convert these similarity statistics into standard normal random variables, so that they can be easily combined across pedigrees, and the results can be interpreted in a straightforward manner. The details of these transformations and calculations are beyond the scope of this book, and the reader is referred to Weeks and Lange (1988) for a more complete description of the mathematics of this non-parametric method of linkage analysis. The reliability and robustness of this method is unclear, since its dependency on good gene frequency estimates is much more significant than the dependency of standard linkage analysis on gene frequency, which has been shown to be of major concern in chapter 22. Given the strong dependency of this method on gene frequency, and the complexities involved in using the program, we will not go into its usage, but we felt the method should be introduced, since it is a popular approach to non-parametric analysis in pedigrees with structures not conducive to affected sib-pair analysis (i.e., not many affected sib pairs...).

26.4 WHEN SHOULD ONE USE NONPARAMETRIC METHODS?

Nonparametric methods can be very powerful tools in linkage analysis, especially when one is trying to localize a disease gene to a given region, and has little reliable information about the mode of inheritance. Further, they are very rapid, and affected sib-pair methods are simple to both apply and interpret. Although they have lower power than parametric analyses when the model is well characterized, they are not as susceptible to possible modelling errors. In general, if one is analyzing a complex disorder, it would be advisable to try using sib-pair or other non-parametric methods in the analysis of your data, either in lieu of or in addition to parametric analyses, to be more confident that an observed result is not spurious.

26.5 HOW TO DO SIB-PAIR ANALYSIS

In many cases, it is possible to do sib-pair analysis analytically, without the need for complicated computer programs. One needs merely to count the number of IBD alleles in all affected sibling pairs in your sample.

One interesting thing which needs to be pointed out is that all possible pairs of affected sibs in a large sibship can be treated separately, with no effect on the mean or variance of the statistic under the null hypothesis as was shown by Suarez and Van Eerdewegh (1984). Further, according to a result of Blackwelder and Elston (1984), this approach of forming all possible pairs from any given sibship is in most cases the most powerful approach, although they recommended that an appropriate weighting of the contribution of such large sibships might be obtained by dividing the contribution of such a large sibship by s(s-1)/2, although the relative power of such weighted vs. unweighted measures is dependent on the particular model for the disease in question. A good overview of the various test statistics for affected sib-pair analyses is given in Blackwelder and Elston (1984). The ESPA program employs the mean test which is just a chi-square test comparing the number of observed alleles IBD with the number expected under the null hypothesis of no linkage. In this test, the number of shared alleles is compared with the number expected to be shared under the null hypothesis (50% shared, 50% unshared). The statistic, therefore is just a standard chi-square test, as

$$X^2 = 2\frac{(S - E[S])^2}{E[S]} = 4\frac{(S - \frac{S+NS}{2})}{S + NS},$$

where S = number of observed alleles shared IBD, and NS = number of observed alleles not shared IBD.

26.6 EXTENDED SIB-PAIR ANALYSIS AND THE ESPA PROGRAM

This program was developed by Sandkuyl (1989) but is no longer available.

27 Genetic Heterogeneity

In this chapter, you will be introduced to the most fundamental and basic approaches to linkage analysis under genetic heterogeneity. The two most important questions asked are 1) given a positive linkage test result, is there significant evidence for a proportion of the families segregating the linked gene, and another proportion segregating a putative unlinked gene for the same disease?; and 2) although I do not have significant evidence for linkage assuming homogeneity of the disease gene, if I allow for a certain percentage of my families to be segregating an unlinked gene, is there significant evidence for linkage of a disease gene in a proportion of my families? It is typically assumed that only one of the two disease genes segregates per family, which is a reasonable assumption for rare diseases. A general overview of some of the methods available to address these two questions will be presented, followed by an introduction to how to use the HOMOG program to do the simple analyses discussed above.

27.1 WHAT IS GENETIC HETEROGENEITY?

When people talk of genetic heterogeneity, they refer to a situation in which any of a number of genes can independently cause the identical disease phenotype. There are many different types of heterogeneity. The simplest type would be *allelic heterogeneity*, in which multiple separate disease alleles at the same locus can each cause the same disease phenotype. An example of this would be in cystic fibrosis, for which there have been a number of different mutated alleles each of which can contribute to the CF phenotype. In general, these do not provide any significant hardship for the linkage analyst, with the exception of analyses involving linkage disequilibrium or genetic risk calculations.

Another form of heterogeneity, which is of greater significance to the linkage analyst, is *non-allelic (or locus) heterogeneity*, in which disease alleles at two or more independently acting loci could each cause the same disease phenotype. For example, this could be a situation in which a biochemical pathway could be disrupted by a defect in any of the enzymes required for the pathway to be completed. Typically, either mutation would cause the phenotype that none of the end-product of that biochemical pathway can be synthesized, causing the disease. An example of this would be Charcot Marie-Tooth disease (Chance et al, 1990). This non-allelic heterogeneity can be further subdivided into diseases with multiple genetic causes, each with the same mode of inheritance, and those with different modes of inheritance. For example, retinitis pigmentosa shows both forms of heterogeneity, with at least two separate loci for an X-linked recessive form of RP, an autosomal recessive form, and an autosomal dominant form (McKusick, 1990). Heterogeneity is much easier to detect when there is a different mode of inheritance in some families from others, since this can typically be seen without the need for linkage analysis. However, when two forms of the disease share a common mode of inheritance (i.e. autosomal dominant, etc.), the only way to determine that more than one genetic locus is involved would be through some special form of linkage analysis, which would treat the heterogeneity as an additional parameter to be dealt with in the analysis of the data. The remainder of this chapter will deal with the application of some of the methodology in current usage for linkage analysis under non-allelic heterogeneity. For a comprehensive discussion of various theoretical approaches to this problem, please consult Ott (1991).

27.2 TEST FOR HOMOGENEITY GIVEN LINKAGE

The simplest thing to test for is given a significant linkage test result (i.e. a lod score greater than 3), is there significant evidence to support the hypothesis that some of the families in our pedigree set might be segregating a different unlinked gene for the same disease, and not the gene that is linked to the marker in question in this analysis. The method we will be applying in this chapter is the so-called A-test (Smith, 1961), or admixture test. In this test, the underlying assumption is that there are two categories of families in the data, some with $\theta = ½$, and some with $\theta = \theta_1 < ½$, with a proportion α of families segregating the linked gene (i.e. $\theta = \theta_1$ in proportion α of the families, and $\theta = ½$ in proportion $(1 - \alpha)$ of the families). The additional assumption is that unequivocal assignment of any family to one class or the other is impossible a priori. This is almost always the case, since we have no way other than through linkage analysis of assigning any given family to one category or the other. So, the likelihood for any given family can be written as $L(\alpha, \theta) = \alpha L(\theta) + (1 - \alpha)L(\theta = ½)$, which could be rewritten as $\alpha L(\theta)/L(\theta = ½) + (1 - \alpha)$, since $L(\theta = ½)$ is a constant. Thus, for *n* families, the total likelihood is just $\Pi[\alpha L(\theta)/L(\theta = ½) + (1 - \alpha)]$ If one wanted to test the null hypothesis of linkage homogeneity (i.e. that all families are of the linked type), you could form a likelihood ratio as $L(\alpha, \theta)/L(\alpha = 1, \theta) = \Pi\ [\alpha + (1 - \alpha)L(\theta = ½)/L(\theta)]$. Then, $2\ln \Pi_i\ [\ \alpha + (1 - \alpha)L_i(\theta = ½)/L_i(\theta)] \sim \chi^2_{(1)}$ However, since the test is of the form, $L(\alpha < 1, \theta)/L(\alpha = 1, \theta)$, i.e. H_0: ($\alpha = 1$) vs. H_1: ($\alpha < 1$), the test is carried out in a one-sided manner, and therefore one must adjust the p-values accordingly. In other words, one would find the p-value associated with the $\chi^2_{(1)}$ value computed from the likelihood ratio test, and divide the corresponding p-value in half to compute the appropriate p-value for this chi-squared test of homogeneity. This is true for reasons outlined in Ott (1985).

27.3 TEST FOR LINKAGE GIVEN HETEROGENEITY

It is possible to consider another null hypothesis for the likelihood ratio test of homogeneity described above. One could consider the situation where one had a null hypothesis of no linkage, against an alternative hypothesis of linkage and heterogeneity. Such a test would be parametrized as $L(\alpha, \theta)/L(\alpha = 1, \theta = ½)$. However, when considering this test, there are two fundamental problems in interpretation. First of all, we need to realize that we are trying to declare a linkage significant while allowing for heterogeneity as a sort of nuisance parameter. Hence, one would need to have a significance at least equivalent to that of the lod score of three criterion in a straight linkage analysis. If one were to consider the maximum "lod score with heterogeneity" to be $\log_{10}[L(\hat{\alpha}, \hat{\theta})/L(\alpha = 1, \theta = ½)]$, and require that this value exceed three as a test of linkage, we would be slightly non-conservative, since there is an additional free parameter, α, in the numerator of this lod score. In order to allow for this, one could add approximately $\log_{10}(2) = 0.30$, to the critical value to allow for an additional degree of freedom. This would then make a critical value of 3.30 for declaring the linkage test significant (corresponding to a likelihood ratio of 2000:1).

However, there is another problem with this likelihood ratio test. Under the null hypothesis, $\theta = ½$ in all families, the parameter α disappears. Notice that the likelihood of any given family is $L(\alpha, \theta) = \alpha L(\theta) + (1 - \alpha)\ L(\theta = ½)$. This would make $L(\alpha, \theta = ½) = \alpha L(\theta = ½) + (1 - \alpha)L(\theta = ½) = L(\theta = ½)$. One could also parametrize this likelihood making $\alpha = 0$, in which case θ would disappear as a parameter, since $L(\alpha = 0, \theta) = 0L(\theta) + (1 - 0)\ L(\theta = ½) = L(\theta = ½)$. Therefore, we have a completely degenerate situation under the null hypothesis, where $L(\alpha = 0, \theta) = L(\alpha, \theta = ½) = L(\theta = ½)$, and thus there is one parameter under H_0, while under H_1, there are two (α and θ). This leads to a problem with the asymptotic distribution of the likelihood ratio, and $-2\ln[L(\theta = ½)/L(\alpha, \theta)] \sim \chi^2$. Hence, we have even further troubles. Several people have considered the asymptotic distribution of this statistic (M. Shoukri, Personal communication; Davies, 1977; Faraway, 1993), which can be extremely complicated. In light of all this, it seems that one should not in general apply any asymptotic theory to this test statistic, but use a criterion of a likelihood ratio > 2000:1 to declare significant evidence exists for linkage in some of the families in your dataset. This 2000:1 odds criterion being based on the normal 1000:1 odds required in a normal linkage test, and the allowance for the second free parameter (α) in the numerator of the odds ratio.

27.4 USING THE HOMOG PROGRAM

Both of these tests are incorporated in the HOMOG program (Ott, 1991). To use this program, you must first compute lod scores at a large number of recombination fractions in each pedigree separately. Let us consider the schizophrenia pedigrees from chapter 25, under the recessive model with diagnostic scheme 3. Compute two point lod scores between marker 2 and the disease, at Θ values ranging from 0 to 0.45 in steps of 0.05. Please use LCP, as you did earlier to perform this analysis on all the pedigrees together. Then examine the FINAL.OUT file obtained from this analysis. The $\log_{10}$(likelihood) for each pedigree at each recombination are shown in table 27-1.

	Pedigree					
Theta	1	2	3	4	5	6
0.00	-27.088124	-11.855281	-28.836341	-9.981559	-14.330920	-22.606300
0.05	-26.760831	-11.849109	-28.469527	-10.008431	-14.209290	-21.585746
0.10	-26.564485	-11.830223	-28.337708	-10.033840	-14.145533	-21.104337
0.15	-26.431280	-11.795900	-28.292208	-10.057423	-14.112566	-20.823790
0.20	-26.335704	-11.749105	-28.289131	-10.078831	-14.097507	-20.649023
0.25	-26.265375	-11.696470	-28.307058	-10.097721	-14.093468	-20.538287
0.30	-26.213481	-11.645108	-28.333387	-10.113758	-14.096308	-20.467862

0.35	-26.176015	-11.600649	-28.360206	-10.126625	-14.103046	-20.422618
0.40	-26.150575	-11.566824	-28.382478	-10.136040	-14.111042	-20.393293
0.45	-26.135765	-11.545810	-28.397042	-10.141784	-14.117674	-20.375567
0.50	-26.130888	-11.538696	-28.402098	-10.143715	-14.120392	-20.369191

Table 27-1 : $\log_{10}$ (Likelihood) for each pedigree at each value of Θ for schizophrenia pedigrees with recessive model, diagnostic scheme 3.

For each pedigree separately, you should then compute the lod scores at each value of θ, by the formula $Z(\theta) = \log_{10}L(\theta) - \log_{10}L(\theta = ½)$. These lod scores are required for the input file for the HOMOG program. The format for the input file is as follows:

Line 1 : Title of the problem

Line 2 : *N STEPSIZE LDIFF* , where *N* = the number of values of θ at which the lod scores are given in the input file (note that values should not be given for θ = 0.5, where Z(θ = ½) = 0); *STEPSIZE* = the step size in which α should be incremented. By default this value should be set to 0.05, meaning that the likelihood would be evaluated for values of α = 0, 0.05, 0.1, 0.15,...; *LDIFF* is optional, and can be omitted without a problem. What it stands for is the difference in natural log likelihood to be used as the basis for the support interval for α around its MLE. In other words if this value is equal to *2*, the program will give you a 2-unit support interval for all parameters (remember that these intervals are given in terms of *natural* log, not *common* log!).

Line 3: *OUT ALOW LL*, where *OUT* refers to the output option. The program can give you, in the output file, a table of values of lnL(α, θ) over all α and θ combinations. It could also give you a list of lod scores for each family separately. The value of the variable *OUT* identifies what combination of these options you wish to have in your output file, as indicated in table 27-2; *ALOW* is the smallest value of α to be considered. For example, if you wished to only consider values of α ≥ 0.1, you would set *ALOW* = 0.1. In most situations, however, it is safest to set *ALOW* = 0, and maximize the likelihood over all possible values of α; *LL* denotes the line length of the output file, and is an optional variable. If it is omitted, then the line length is assumed to be set to 80 characters. In some situations, you may wish to allow for longer lines, especially if *OUT* = 2 or 3, since the table of values of lnl(α,Θ) can become very long as the number of values of α and θ considered increases.

OUT	Table of lnL(α,Θ)	Lod scores for families
0	no	no
1	no	yes
2	yes	no
3	yes	yes

Table 27-2 : Table of definitions of output options for HOMOG programs in variable *OUT*.

Line 4: $\theta_1, \theta_2, ..., \theta_N$; where the θ_i are the values of θ for which the lod scores are going to be provided below. These should in general be in ascending order. Naturally, the finer the grid of points at which lod scores are computed, the more powerful the homogeneity test will be, since the maximization of the likelihood will be more accurate. (There is no need to provide the actual values of θ, since there is *no* interpolation scheme used. One could just as easily provide integer values or map distance measures, as long as there are *N* real numbers provided.)

Line 5 : *NFAM*; where *NFAM* is the number of families for which lod scores will be provided.

Line 6 : $Z(\theta_1), Z(\theta_2),... ,Z(\theta_N)$ in Family 1

Line 7 : $Z(\theta_1), Z(\theta_2),... ,Z(\theta_N)$ in Family 2

...

Line (5 + *NFAM*) : $Z(\theta_1), Z(\theta_2),... ,Z(\theta_N)$ in Family *NFAM*

It is important to note that any lod score less than –80 is assumed to represent a lod score of –∞, and in the output (and in the input files), log likelihoods of –∞ should be indicated as –99. The input file for this dataset should be as follows:

```
Schizophrenia pedigrees - Recessive model - Diagnostic scheme 3
10 0.05
0 0
0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45
6
-0.957236 -0.629943 -0.433597 -0.300392 -0.204816 -0.134487 -0.082593 -0.045127 -0.019687 -0.004877
-0.316585 -0.310413 -0.291527 -0.257204 -0.210409 -0.157774 -0.106412 -0.061953 -0.028128 -0.007114
-0.434243 -0.067429 0.064390 0.109890 0.112967 0.095040 0.068711 0.041892 0.019620 0.005056
 0.162156 0.135284 0.109875 0.086292 0.064884 0.045994 0.029957 0.017090 0.007675 0.001931
-0.210528 -0.088898 -0.025141 0.007826 0.022885 0.026924 0.024084 0.017346 0.009350 0.002718
-2.237109 -1.216555 -0.735146 -0.454599 -0.279832 -0.169096 -0.098671 -0.053427 -0.024102 -0.006376
```

Save this file as HOMOG.DAT, the required input file name for the HOMOG programs, and type *HOMOG* at the DOS prompt to run the program and do the analysis. The output file should resemble the following:

```
Program HOMOG version 3.33 J. Ott
Heterogeneity: two family types, one with linkage, one without
 alpha = proportion of families with linkage (θ < 1/2)
 1-alpha= proportion of families without linkage (θ = 1/2)

>> Schizophrenia pedigrees - Recessive model - Diagnostic scheme 3 <<

Table of conditional max. Ln(L) over α's, given θ or x
      θ or x      α max       Max.Ln(L) Lik. ratio
 1    0.0000      1.0000      0.0000      1.0000
 2    0.0500      1.0000      0.0000      1.0000
 3    0.1000      1.0000      0.0000      1.0000
 4    0.1500      1.0000      0.0000      1.0000
 5    0.2000      1.0000      0.0000      1.0000
 6    0.2500      1.0000      0.0000      1.0000
 7    0.3000      1.0000      0.0000      1.0000
 8    0.3500      1.0000      0.0000      1.0000
 9    0.4000      1.0000      0.0000      1.0000
 10   0.4500      1.0000      0.0000      1.0000
```

The output above shows the maximum natural log likelihood for each value of θ, maximized over α, normalized such that ln L(θ = ½) = 0. In this case you note that for all α, when θ < ½, lnl(θ, α) < 0, implying that the M.L.E. of θ is 0.5 (or equivalently that the M.L.E. of α is 1), meaning that all families would be considered to be unlinked.

```
                                            Estimates of
Hypotheses                    Max.lnL     Alpha       Theta
H2: Linkage, heterogeneity    0.0000      1.0000      99.0000
H1: Linkage, homogeneity      0.0000      (1)         99.0000
H0: No linkage                (0)         (0)         (0.5)

Components of chi-square
Source                        df Chi-square     L ratio
H2 vs. H1 Heterogeneity       1     0.000       1.0000
H1 vs. H0 Linkage             1     0.000       1.0000
H2 vs. H0 Total               2     0.000       1.0000
```

The values in these tables are the crux of the analysis. The top table indicates the maximum natural log likelihood of the entire family set under each of the three hypotheses. The first line corresponds to lnL($\hat{\alpha}$,$\hat{\theta}$); the second to lnL(α = 1,$\hat{\theta}$); and the third to lnL(θ = ½). These three hypotheses are denoted H_2, H_1, and H_0 respectively, with the meaning that in H_2 there is linkage and heterogeneity, H_1 has linkage and homogeneity, and under the true null hypothesis, H_0, there is neither linkage nor heterogeneity. When an estimate of θ is given as 99.0000, it means the disease is unlinked (θ = ½). This value is used because when multipoint lod scores are used in a HOMOG analysis, the value of 0.5 may have some other meaning, since then map distances are used instead of recombination fractions. The bottom table summarizes the three possible likelihood ratio tests possible with these three hypotheses. The first is the test of heterogeneity given linkage (H_2 vs. H_1), and the second is the standard test of linkage assuming homogeneity (H_1 vs H_0), or just L($\hat{\theta}$)/L(θ = ½). The third is the joint test of linkage and heterogeneity, L($\hat{\alpha}$,$\hat{\theta}$)/L(θ = ½), the properties of which are described above. Each of these values can be computed by finding the difference between the ln(Likelihood) values for each hypothesis, and multiplying it by two. This then provides the quantity found under the heading *Chi-Square*. The number of degrees of freedom is given in the first column, though the number of degrees of freedom in the test of H_2 vs H_0 is not really two, and this statistic is not really distributed as a χ^2 random variable. The likelihood ratio in the last column provides the exact odds for the alternative hypothesis against the null hypothesis. By the criterion described above for considering the test of H_2 vs H_0 to be significant, this quantity would have to exceed 2000. In this example, there is absolutely no evidence for heterogeneity in the data, and as such the value of each likelihood ratio is 1, with $\chi^2 = 0$.

```
Family no.  Conditional prob. of linked type
 1          1.0000
 2          1.0000
 3          1.0000
 4          1.0000
 5          1.0000
 6          1.0000
```

Finally, the values in the table above are the conditional probabilities that each family is segregating the linked gene, assuming the values of α and θ estimated under H_2, from our observed family data. This conditional probability is described in Ott (1991). In general, the values found here should be taken with a grain of salt, and they cannot ever be validly used to separate families for the remainder of a linkage study. It should be required that any further marker typings be done on all the families combined, and analyzed with these HOMOG programs, to compute the appropriate log likelihood ratio statistics on the entire data set, in order to preserve the validity of the results, and not to induce any potential bias which could lead to an increased false positive rate. If one were to selectively type further linked markers solely in those families with high posterior probabilities of segregating a linked gene, we would be selecting families for further linkage analysis conditional on there being few observed recombinants, which could easily lead to false positive evidence of linkage.

In general, all of the information required for a complete analysis of your dataset, allowing for the presence of heterogeneity, is available from this output file. One should as a matter of practice always perform such an analysis with any complex disease, since it is usually assumed that there is non-allelic heterogeneity involved in the etiology of these diseases, and performing homogeneity tests can allow you to more completely extract information from your dataset. Other programs are available, based on the same algorithm, to handle more complicated heterogeneity situations; HOMOG2, HOMOG3, HOMOG3R, etc., but discussion of them is beyond the scope of this book.

EXERCISE 27

For the four pedigrees shown in Figure 27-1, please perform a complete linkage analysis, including homogeneity testing, assuming the disease to be inherited as an autosomal dominant disease with gene frequency of 0.00001 for the disease allele, and a marker with two alleles with gene frequencies of 0.65, and 0.35 respectively. *Note*: In HANDDATA.ZIP, the HOMOG.DAT file is the datafile to these families, not to be confused with HOMOG.DAT as an input file to the HOMOG program.

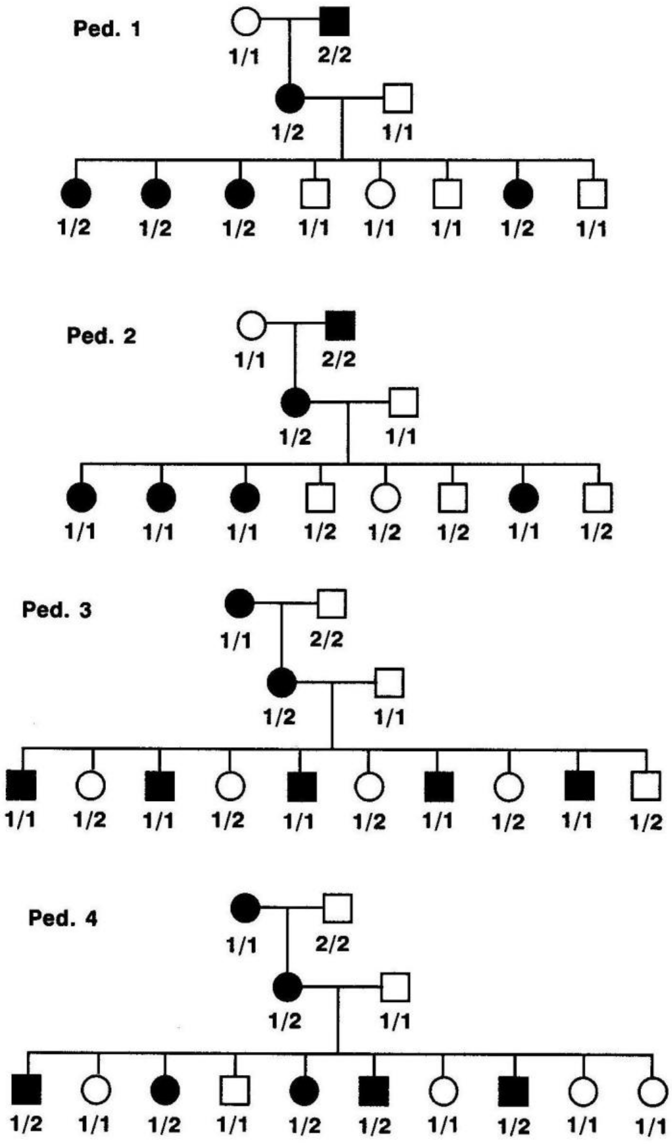


Figure 27–1. Example pedigrees for homogeneity test

28 Computer Simulation Methods

In this chapter you will get a brief overview of computer simulation, and some of its many applications in human genetic linkage analysis. The various approaches to pedigree data simulation will be introduced, as well as the various types of statistical information one can obtain from a simulation.

28.1 RANDOM NUMBERS AND SIMULATION

It is possible to write a computer program that will generate so-called "pseudo-random" numbers, based on an initial seed provided by the user. These routines basically approximate randomness with any of a large number of complicated mathematical functions designed to generate a series of numbers on the interval (0,1) which are approximately uniformly distributed, but see [28]. What this means is that it is equally likely for any real number on the interval (0,1) to be chosen independent of the previously selected number. Then, let us assume that we wished to simulate one replicate of a coin toss. Since heads and tails have equal probability of 50%, we would divide the interval in half, and say that if a given random number is less than 0.5, then it is heads, and if the number is greater than 0.5, it is tails. In this way, we can simulate the flipping of a fair coin by computer. If we think about it intuitively, flipping a coin can also be thought of as a primitive random number generator, and a sample simulation can be done solely by tossing a coin. As an illustration of this technique, we will start by considering the simple case of simulating a marker unlinked to the disease (θ = ½).

28.2 PEDIGREE SIMULATION BY TOSSING A COIN

Let us consider the pedigree shown in Figure 28-1. Assume the disease is dominant with full penetrance and two alleles, *T* (trait) and + (normal), such that the disease phenotypes uniquely determine the corresponding disease locus genotypes. Then, let us simulate a 2-allele marker with equal gene frequency for each allele, which is unlinked to the disease locus (θ = 0.5). In this simple situation, we can simulate this pedigree solely by flipping a coin. As an example of how this works, let us first simulate marker locus genotypes for the parents in this pedigree (with phase). Since the marker locus has two alleles with equal gene frequency of 50%, we can simulate the alleles by coin toss. Let us assume that on each flip, for the founder individuals, if the coin comes up heads, then we select allele *1*, and if the coin comes up tails, we select allele *2*. We first simulate the marker allele in phase with the disease allele in person *A*. The first flip of my coin came up tails, so allele *2* is selected. Next, we simulate a marker allele in phase with the + allele in person *A*. The coin came up heads, so this haplotype is now assigned to carry allele *1*. Thus, this individual has genotype *1 2* at the marker locus. Similarly we must simulate the two marker alleles in individual *B*. Assume the coin came up heads the first time, and tails the second time, giving this individual genotype *1 2*.

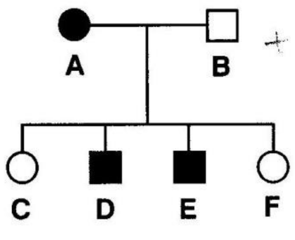


Figure 28–1. Pedigree for manual simulation example

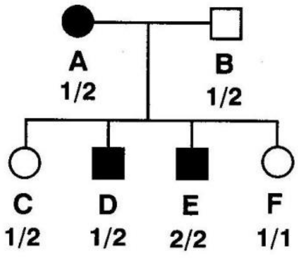


Figure 28–2. Manually simulated replicate of pedigree from Figure 28–1

Next, we simulate the children of these parents. Let us first simulate the transmission of marker alleles from person *A* to her children. In this case, we know which disease locus allele was transmitted to each child. Thus we need only simulate the recombination process from mother to child. Clearly, if the marker is unlinked to the disease, the recombination probability is ½ by definition. When we flip our coin, let us assume that if the coin comes up heads, a recombination occurred, and if it comes up tails, no recombination event occurred. In this case, for individual *C*, the coin came up heads, so a recombination occurred in the meiosis from mother to child. Since this individual received the + allele from her mother, at the marker locus she must have received the *2* allele, which was in phase with the *D* allele in the mother, and a recombination event occurred. For individuals *D*, *E*, and *F*, assume the coin came up tails. Then, these individuals received marker alleles from their mother without recombination, meaning that they received alleles *2*, *2*, and *1* respectively. Finally, we need to simulate the alleles passed from individual *B* to his children. Since he is homozygous +/+ at the disease locus, we cannot simulate recombination events. However, we still must simulate the segregation of the alleles from father to children. According to Mendelian laws, there is a 50% chance that either allele was inherited at the marker locus from the father. Therefore, let us assign heads to the inheritance of allele *1*, and tails to the inheritance of allele *2* from father. Flipping our coin, we come up with heads, heads, tails, and heads, meaning the alleles transmitted to individuals *C*, *D*, *E*, and *F* will be *1*, *1*, *2*, and *1* respectively. Thus the final simulated replicate of our family would be that shown in Figure 28-2. The linkage analysis of this family would yield a maximum lod score of 0.301029 at θ = 0. Although we did simulate one recombinant in this family, because the analysis cannot make the phase assumption we made in our simulation, the recombination event goes unnoticed because of the marker alleles inherited from the unaffected parent. In fact, individuals *C* and *D* provide essentially no linkage information since they are heterozygous, and their parents are identical and heterozygous. It is equally likely that they received the *1* allele from *A*, and the *2* allele from *B* as it is that they received the *1* allele from *B*, and the *2* allele from *A*. Hence, there is essentially no information from such a situation. Thus, the only individuals which provide unequivocal information about recombination are individuals *E* and *F*, who are homozygous at the marker, making it obvious which alleles were inherited from *A* with the disease. Since they were both non-recombinants, we are left with a phase-unknown pedigree with two non-recombinants, which yields a maximum lod score of 0.301029 = $\log_{10}(2)$. What do you suppose would have happened had the two individuals in question both have been recombinants? Since the family is phase unknown, the result of the analysis would be the same as in this situation, since the inheritance pattern would be completely consistent with no recombination. In any event, this is a primitive example of how a pedigree simulation is conducted with the computer software we will be introducing in this chapter. The only difference is that our computer software will allow us to simulate random numbers over the range of (0,1), such that probabilities other than 0.5 can be accurately simulated, which is not so simple in the coin toss random number generator.

28.3 THE SIMULATE PROGRAM

It is a very simple matter to simulate pedigree data based only on pedigree structure, and not on any already-known phenotypes, for in this case, one need not condition your simulation on anything other than the simple straightforward population genetics and Mendelian laws. When you are simulating a group of markers which are unlinked to your disease locus, it is equivalent to completely disregarding whatever information is known about the trait locus, and is therefore model-free simulation.

Basically, the simulation is performed in a very straightforward way, using the concept of pseudo-random number generation introduced in the opening section of this chapter. One must first simulate genotypes with phase at each marker independently in all founders (i.e. people without parents in the pedigree) according to the laws of population genetics, assuming Hardy-Weinberg Equilibrium (HWE). The next step is to simulate marker segregation from the already simulated parents to their children. For the first locus to be simulated, there is a 50-50 chance of receiving either allele from each parent to each child independently, according to the law of independent segregation. However, if you are simulating multiple linked loci, the next locus no longer segregates independently, so you must simulate it according to the recombination fraction between itself and the first locus. If the recombination fraction (θ) is 0.1, then if the random-number generator selects a number below 0.1, a recombination is simulated, and the allele to be inherited from the parent in question must be selected from the other parental haplotype (this is why we must insist on simulating all genotypes with phase in the first step). We continue like this, switching parental haplotypes whenever a recombination is simulated across the chromosome, until we reach the final marker to be simulated. This process is repeated for each individual, from each of his parents. When all individuals have been simulated, it constitutes one simulated replicate of your linear set of multiple linked markers segregating through your pedigree (or set of pedigrees) independent of your predetermined trait locus (if one exists). These pedigree replicates are then saved in a form ready for analysis with the MSIM, LSIM, or ISIM programs of the SLINK package, as will be described below. The technical details of using SIMULATE will be outlined in section 28.5.

28.4 POSSIBLE APPLICATIONS OF SIMULATION UNDER H_0

This simulation method can be useful in determining the p-value associated with a given observed maximum lod score in a pedigree (set). In other words, let us say that you found a lod score of 2.5 in your pedigree set with a given marker vs. disease. In and of itself, this is an insignificant result according to the traditional lod-score-of-three criterion. However, it may be very unusual to observe such an extreme lod score in your particular family set. You could, for example, simulate the marker(s) in question in a large number of replicates of your family set, and see how often the observed lod score is reached or exceeded by chance. If this is very, very rare, then you might want to report the simulated p-value associated with this result. However, it may also be very easy to get such a high lod score in your pedigree set, in which case the significance should also be reported, and used to guide your future plans. Knowing the p-value associated with given lod scores has many uses and applications. For the molecular biologist it is useful to know the null hypothesis distributions of maximum lod scores, as this can serve as a useful guide in their search for a new gene. If a certain pedigree almost never (1/1000 replicates, for example) gives a lod score as high as 1, yet you observe such a high lod score, this lod score would be much more significant than finding the same result in a pedigree that exceeds this threshold 10% of the time by chance. This kind of information can help the investigator decide whether to invest further energy in typing additional markers in a given region or not to try and find the gene causing the disease in question. So, knowing these null hypothesis lod score distributions can be a very useful tool for the linkage analyst throughout a given investigation. Further, in the case of maximizing the lod score over many different sets of genetic model parameters, one needs to compute the probability of exceeding a given lod score given no linkage, and the exact number of models investigated, as will be described in a later section. To compute these p-values, one needs to simulate marker data independent of the disease, and then analyze it under various models to compute the p-values needed. This unconditional method is ideally suited to these types of investigations.

Further, if you are interested in ordering a new marker against a map of markers in a given pedigree set, you can use this rapid simulation approach to simulate the genotypes of any given set of markers segregating in your families to determine the power of your collection of pedigrees to order the markers. This is a very useful thing to do if you are trying to decide how many families you would need to type a given new marker in order to accurately order it with the required 1000:1 odds. If for a given marker set you find that you have 80% power to accurately order the markers with 1000:1 odds using half of your total available pedigrees, and you have 90% power if you type them all, it may be advisable to only type half of your families to begin with, and then type the remaining families later only if no significant result is obtained from the first half. You might also find that you have no chance to order the markers with 1000:1 odds in your pedigree set, in which case, it might not be advisable to bother trying until you have collected more or better families.

For the statistical geneticist, testing the properties of new statistical methods is perhaps the most common application of simulation. Whenever a new method is derived, and one desires to compute its power and other properties, one often needs to rely on simulation, since theoretical computations involving pedigrees can often be prohibitive in realistic situations. Therefore, one often needs a rapid way of generating data to do a particular study. This SIMULATE program can simulate any type of locus describable in LINKAGE format, so one could simulate a trait locus linked to a number of markers, and by modifying the source code slightly, could incorporate whatever ascertainment restrictions the investigator desires. In this way, one can test out the properties of new methods, and novel statistics in situations where theoretical analysis cannot be easily done.

28.5 HOW TO USE THE SIMULATE PROGRAM

Let us consider the set of schizophrenia pedigrees from chapter 25, and simulate an unlinked marker with the properties of markers *1* and *2* from that dataset, with the dominant model for the disease in diagnostic scheme *3*. Our goal in this case is to determine the probability (when there really is no linkage) of getting a lod score at least as large as the 2.65 we observed, assuming we analyzed the pedigree under only this one model. Similarly, we want to test the linkage of disease to marker *2* and determine the probability of getting a lod score at least as large as the 2.22 we observed in our example for that comparison. To do this we must simulate the two markers linked to each other at recombination fraction 0.12 (Sherrington et al, 1988), and unlinked to the putative disease locus.

Three input files are required for the SIMULATE program, they are called SIMPED.DAT (a pedigree file), SIMDATA.DAT (a locus parameter file), and PROBLEM.DAT (a file containing simulation parameters). The basic format for each file will now be introduced in the context of the schizophrenia pedigrees.

The SIMPED.DAT file is a linkage format pedigree file (processed by MAKEPED), with some slight alterations. The first locus should be the disease, (if there is one). If no affection status locus is present, the program will simulate just a set of linked markers, but if there is a disease present, it will simulate the remaining loci, and keep the disease locus phenotype as given in this file. The only restriction is that the disease must be the first locus in the file, and it must be an affection status locus. Instead of providing marker locus phenotypes for the remaining loci, one has to enter a *1* if the first marker locus is to be simulated, and a *0* if it is to be left unknown. The next column would then contain a *1* or a *0* to tell the program whether the second marker is to be simulated, or left untyped, etc. Thus, each line of the SIMPED.DAT file would contain an affection status locus for the disease, followed by a series of *0*'s and *1*'s to tell whether each marker locus is to be simulated or left untyped. If an individual was untyped for a given marker in the original dataset, please assume that he is unavailable for the simulation as well (for the same marker), to make sure the simulation results are consistent with our real dataset. For the schizophrenia pedigree set, the SIMPED.PRE file should resemble the following (information shown for pedigree *1* only, to save space, but you should complete the file for all pedigrees):

```
1 A 0 0 2 2 1 1 1
1 B 0 0 1 1 1 1 1
1 C B A 2 2 3 1 1
1 D B A 1 2 1 1 1
1 E B A 1 2 1 1 1
1 F B A 1 2 1 1 1
1 G 0 0 1 0 1 0 0
1 H 0 0 2 0 1 0 0
1 I G H 2 1 1 1 1
1 J B A 2 2 1 1 1
1 K G H 1 1 1 1 1
1 L F I 1 1 1 1 1
1 M F I 1 1 1 1 1
1 N F I 1 2 1 1 1
1 O K J 1 1 1 1 1
1 P K J 1 2 1 1 1
```

Process this file with MAKEPED, to make the SIMPED.DAT file needed for the analysis. Further, there needs to be a header in the first two lines of this file. The first line must indicate the number of pedigrees in the SIMPED.DAT file (in this case, there are six); and the second line must contain the number of individuals in each pedigree (including doubled individuals). In this case, the first two lines should be as follows:

```
6
17 17 35 17 18 26
```

Further, there is one more important alteration that must be made to this input file. It is imperative that the individuals be numbered consecutively from *1* through *n*, where *n* is the number of individuals in the pedigree. They must also be presented in numerical order in the file. In pedigree 1, however, there is a marriage loop, and when individual *6* is doubled, a new individual with id number *17* is added directly after individual *6*. You must move this individual to the end of the first pedigree, directly following individual *16*, such that the SIMPED.DAT would look like the following (pedigree *1* only shown).

```
6
17 17 35 17 18 26

1 1 0 0 3 0 0 2 0 2 1 1 1
1 2 0 0 3 0 0 1 0 1 1 1 1
1 3 2 1 0 4 4 2 0 2 3 1 1
1 4 2 1 0 5 5 1 0 2 1 1 1
1 5 2 1 0 6 6 1 0 2 1 1 1
1 6 2 1 0 10 10 1 2 2 1 1 1
1 7 0 0 9 0 0 1 1 0 1 0 0
1 8 0 0 9 0 0 2 0 0 1 0 0
1 9 7 8 12 11 11 2 0 1 1 1 1
1 10 2 1 15 0 0 2 0 2 1 1 1
1 11 7 8 15 0 0 1 0 1 1 1 1
1 12 17 9 0 13 13 1 0 1 1 1 1
```

```
1 13 17 9 0 14 14 1 0 1 1 1 1
1 14 17 9 0 0 0 1 0 2 1 1 1
1 15 11 10 0 16 16 1 0 1 1 1 1
1 16 11 10 0 0 0 1 0 2 1 1 1
1 17 0 0 12 0 0 1 2 2 1 1 1
```

The second required file is called SIMDATA.DAT, and provides the locus parameters for the simulation, including penetrance, gene frequencies, locus types and definitions, and intermarker recombination fractions. This file should be in standard MLINK format. For the schizophrenia pedigrees, in diagnostic scheme *3*, with the dominant model, the file should be as follows:

```
 3 0 0 5 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0 0.0 0.0 0 << MUT LOCUS, MUT RATE, HAPLOTYPE FREQUENCIES (IF 1)
 1 2 3
1 2 << AFFECTION, NO. OF ALLELES
 0.010000 0.990000 << GENE FREQUENCIES
 3 << NO. OF LIABILITY CLASSES
 0.4400 0.4400 0.0166
 0.4400 0.4400 0.0166
 0.4400 0.4400 0.0166 << PENETRANCES
3 2 << ALLELE NUMBERS, NO. OF ALLELES
 0.330000 0.670000 << GENE FREQUENCIES
3 3 << ALLELE NUMBERS, NO. OF ALLELES
 0.320000 0.160000 0.520000 << GENE FREQUENCIES
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
 0.50000 0.12000 << RECOMBINATION VALUES
 1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

This is basically the same as the LINKAGE parameter file we used in chapter 25, with the recombination fractions correctly specified on the next-to-last line of the file, to indicate the recombination fractions required for the simulation to work. Since this program assumes the disease is unlinked to the markers, if a number other than 0.5 is given for the first recombination fraction, the program will ignore it, and still set this θ to 0.5.

The final input file required is called PROBLEM.DAT, and should contain two lines. The first line should contain three seeds for the random number generator (these should be numbers between 1 and 30000, with higher numbers providing better (i.e. more random) results). The second line should contain the number of replicates of the set of pedigrees desired. In general, this number should be very large, when you are attempting to compute p-values. The normal p-value associated with a lod score of three, for example, is always less than 0.001, so you would need a large number of replicates to get an accurate estimate of such small p-values. For our purposes, however, let us simulate only 50 replicates, in the interest of saving time. It is important to remember that the larger the number of replicates, the greater the accuracy. In general, 50 replicates is **way** too few to obtain reliable estimates of small p-values, but in the interests of time, we will limit our study to this level for this instructional example. For our purposes, please use seeds of 24553, 29773, and 20142. In practice, whenever you use the same seeds, the results will be identical, since the sequence of pseudo-random numbers will be the same for the same starting values. Whenever these values change, however, the entire sequence will be altered as well. The PROBLEM.DAT file should look like the following:

```
24553 29773 20142
50
```

Next, please run the SIMULATE program by typing *SIMULATE* at the DOS prompt (assuming the SIMULATE program is accessible to DOS either by being in the same directory where you are working, or being in the path). The program will then simulate 50 replicates of the pedigree with both linked markers segregating in it.

28.6 ANALYZING THE SIMULATED REPLICATES WITH MLINK AND ILINK

The file containing the simulated replicates is called PEDFILE.DAT. However, we will be only interested in the two-point lod scores, so we will need to use LCP (which calls LSP) to extract two loci at a time from this file. Hence, we must change the name of the output file. Let us call the new file SCHIZSIM.PED, by typing *RENAME PEDFILE.DAT SCHIZSIM.PED*. Then, you may also rename the SIMDATA.DAT file to SCHIZSIM.DAT.

First, let us confirm that the two markers which were simulated are actually separated by approximately $\theta = 0.12$, by evaluating the lod score over the entire set of $50 \times 6 = 300$ pedigrees simulated. To do this, use LCP as you have done throughout the book, selecting loci 2 and 3 with the MLINK program, starting from $\theta = 0.1$, in steps of 0.01, stopping at $\theta = 0.15$. The resulting maximum lod score should be found at $\theta = 0.13$, with $Z(0.13) = 121.01$. The lod score at $\theta = 0.12$ is only 0.04 lower, which is highly insignificant on such a large dataset. So, we have verified that the two markers were simulated at the appropriate recombination fraction. The remaining thing to verify is that the disease locus is, in fact, simulated unlinked to both marker loci. To do this, please use MLINK to compute lod scores for the disease vs. each marker at recombination fractions 0.45 through 0.5, in steps of 0.01, to verify that the maximum lod score occurs at the true value of $\theta = 0.5$. For disease vs. each of the markers, the maximum lod score is 0 at $\theta = 0.5$, as expected. When the ILINK program is used, the estimated recombination fractions are as follows: $\hat{\theta}_{D,M1} = 0.526$, $\hat{\theta}_{D,M2} = 0.516$, $\hat{\theta}_{M1,M2} = 0.125$. So, we have produced a set of pedigrees consistent with our simulation parameters. Further, one can compute the *expected lod score* for the original set of six pedigrees quite simply from this data. The formula for approximating the expected lod score at a given value of the recombination fraction is just $E[Z(\theta)] \approx (1/n) \sum Z_i(\theta)$; where n is the total number of replicates simulated, and the sum goes from i=1 to n, where $Z_i(\theta)$ is the lod score at recombination fraction θ in replicate i. Clearly, then, we have computed the lod score for the entire family set, and since the lod score is additive over families, the results we have obtained from MLINK are just $\sum Z_i(\theta)$. If we then divide by the number of replicates (50), we would get, for markers 2 vs 3, $E[Z(\theta = 0.12)] = 119.97/50 = 2.40$. Since 0.12 was the simulated recombination fraction, this quantity is of particular interest, and is denoted the ELOD. Another statistic of particular interest which can be computed from this data is the maximum expected lod score, or MELOD. This is the lod score at the value of Θ for which the expected lod score is maximized, or $\max_{\Theta}(E[Z(\theta)])$. Asymptotically, the maximum will always occur at the true value of θ (at which the simulation was carried out), but in small samples (like 50 replicates), there will likely be some deviation. In our case, the maximum occurs at $\theta = 0.125$, as determined by ILINK. The corresponding MELOD is just $121.07/50 = 2.42$, which is slightly larger than the ELOD. In general, MELOD $\geq$ ELOD, and asymptotically, they are equal. The ELOD is a very important measure, as it tells you what lod score on average you can expect to get from your dataset at the true recombination fraction (if it were 0.12, in this case). The MELOD has much less utility, since it is larger than the ELOD solely due to random fluctuations. For this reason, when simulating all the marker information in a pedigree set, we discourage the use of MELOD as it is not a very meaningful measure.

28.7 ANALYZING THE SIMULATED REPLICATES WITH MSIM

To analyze the entire set of pedigrees at once takes an enormous amount of computer power, and the programs must be compiled to handle enormous numbers of pedigrees, etc, which can rapidly eat up your memory. Further, it is not a trivial thing to manipulate the output from MLINK such that you can determine the properties of each pedigree (from the original pedigree set) individually, to see the relative contributions of each to the ELOD. Further, there are reasons for which you might like to compute some other statistical measures on your dataset before starting a linkage study. For the purposes of analyzing simulated data, Weeks et al developed modified versions of ILINK, MLINK, and LINKMAP, called ISIM, MSIM, and LSIM respectively. These programs will compute the lod scores one replicate at a time, and keep track of various pieces of information about each replicate of each family in the original pedigree set. To use these programs, you will need to have the simulated replicates in a file called PEDFILE.DAT (the same name as the output file from SIMULATE or LSP), and a parameter file called DATAFILE.DAT, specifying the analysis parameters. Keeping in mind that we still need to extract only two loci at a time from our original pedigree and parameter files (SCHIZSIM.*), use the LSP program to extract the disease and marker 1, with the parameter file set up for MLINK (for use in our MSIM analysis), with starting recombination fraction of

0, in steps of 0.10, stopping at θ = 0.5, with no sex difference in recombination rates. The MSIM program will also read the file SIMOUT.DAT, which was created by the SIMULATE program, to identify the number of replicates simulated, etc. One additional file, LIMIT.DAT, is required. The MSIM program will compute the probability of exceeding a given lod score in any replicate of the pedigree set, P(Z(θ) > x), where the user can select three values of x. These three lod score thresholds must be entered in a file called LIMIT.DAT. By default, one often chooses *1*, *2*, and *3*. But, if you were trying to evaluate the significance level of a given observed maximum lod score, you would input that value as one of the three thresholds. So, for the analysis of disease vs. marker 1, in our original analysis, we found a lod score of 2.65, so we might want to use thresholds of *1*, *2.65*, and *3*, for this analysis. In this way, we can find the p-value associated with our lod score of 2.65 in this pedigree set. The LIMIT.DAT file should then look like the following:

```
1 2.65 3
```

To run the analysis, you must first run the UNKNOWN program, and then the MSIM program by simply typing *MSIM* at the DOS prompt after UNKNOWN has completed. Then, examine the MSIM.DAT file, which contains the results of the analysis. Below is shown the segment of the output corresponding to θ = 0.10:

```
--------------------------------------------------------
THETAS 0.100
--------------------------------------------------------
Pedigree    Average     StdDev      Min         Max
 1          -0.165047   0.283531    -0.781806   0.428780
 2          -0.027733   0.223751    -0.361404   0.977586
 3          -0.099215   0.264000    -0.609994   0.753716
 4          -0.094414   0.197641    -0.458361   0.438553
 5          -0.048003   0.241983    -0.643913   0.561918
 6          -0.130423   0.281867    -0.770454   0.536066
Study       -0.564834   0.695094    -2.409748   1.274584
--------------------------------------------------------
```

The column headed *Average* contains the average lod score in each family, averaged over all replicates. In practice, this average lod score is usually just referred to as the expected lod score, which it approximates as shown above. The rows correspond to each of the pedigrees in the initial pedigree set, and the row headed *Study* provides the information on the entire set of pedigrees taken together. So, in this case, our overall expected lod score at Θ = 0.10 is –0.564834. The next column provides the standard deviation of the expected lod score, which gives you some idea of the variability of the lod scores across replicates. The last two columns indicate the smallest and largest lod scores found over all replicates in the study. These values are indicated to provide an additional measure of the variability of the lod score across replicates. It is important for you to consider that the value under *maximum* has no easily interpretable statistical importance. If you found a lod score of 1.2 in the original pedigrees, it would not have any increased significance because the largest observed lod score in the set of simulated replicates of 1.27. You must still obtain a lod score of three to declare a linkage significant.

The next segment of the MSIM.DAT file contains the following information:

```
------------------------------------------------------------
 Average Maximum Lod Scores based on quadratic interpolation
------------------------------------------------------------
Pedigree    Average     StdDev      Min         Max
 1          0.046081    0.127706    0.000000    0.523855
 2          0.072394    0.221885    0.000000    1.254722
 3          0.082413    0.198594    0.000000    0.995219
 4          0.058039    0.143175    0.000000    0.586752
 5          0.098635    0.172257    0.000000    0.758645
 6          0.064638    0.137281    0.000000    0.557189
Study       0.125075    0.297165    0.000000    1.339499
-----------------------------------------------------------
```

This table contains a somewhat different piece of information, the average maximum lod score based on quadratic interpolation. The average maximum lod score is an approximation to the expected maximum lod score, $E[\max Z(\theta)] = (1/n)\sum \max Z_i(\theta)$, where the lod score is maximized over θ separately in each replicate. To remain consistent with the notation ELOD, and MELOD, this might appropriately be called the EMLOD, for *E*xpected *M*aximum *LOD* score. The minimum for this value is always 0, since $\max Z_i(\theta) \geq 0$, since $Z(\theta = 0.5) = 0$ by definition. Further, in any given replicate, $\max Z_i(\theta) \geq Z(\theta_0)$, by definition, so EMLOD ≥ MELOD ≥ ELOD. For this reason, many people like this number better, since it is always larger than the ELOD. This measure is an indicator of what maximum lod score you can expect to find in your dataset on average, without respect for where it occurs. In some sense, this quantity provides some measure of power, in that it provides the average value of the test statistic across the set of replicates. However, one nice property of the ELOD is that it is additive across pedigrees, while the EMLOD is not. If your pedigree gives you an ELOD of 1, then you will know that if you had two additional pedigrees of equivalent size and structure, you would have an overall ELOD of 3. However, if you had an EMLOD of 1, there is no direct way of knowing how many additional pedigrees would be required for an EMLOD of 3.

In this case, our unlinked marker still gives an EMLOD of 0.125 (it will not typically be zero, as there is usually some variability in the value of the maximum lod score even under no linkage - hence the possibility of false positives...), while the ELOD is 0.000 (at θ = 0.5). It is intuitively clear that if you had 60 pedigrees in your set, you would not have an EMLOD of $10 \cdot 0.125 = 12.5$, and yet the ELOD would be $10 \cdot 0.000 = 0.000$. In this case, the EMLOD's had to be determined by quadratic interpolation. With the MSIM program, the lod scores were only computed at a predefined set of recombination fractions. However, by a technique called quadratic interpolation, the maximum of any lod score curve can be approximated. (Ott, 1991, p.183) To avoid the approximation element, the ISIM program can be applied, as you will see below.

Finally, the remainder of the file contains information about the probability of finding a lod score greater than the constants you specified in the LIMIT.DAT file. In this example, there were only two replicates with $\max Z(\theta) > 1$, and none with $\max Z(\theta) > 2.65$. Since 2.65 was the observed lod score, the estimated p-value associated with that lod score would be 0/50 = 0, but to determine a confidence interval for the p-value, you can use the linkage utility program BINOM. To do this, select the *CONFIDENCE INTERVALS (2)* option. Then the program will ask you to

```
Enter observed k and n [+no. of decimal places] (-1 exits, ENTER repeats k,n)
```

In this case, you should enter *0 50*, since there were 0 observation of $Z(\theta) > 2.65$, out of 50 opportunities. Next, the program will ask you for the upper error probability. Since the estimate is $\hat{p} = 0$, you need only enter the significance level here. If you want a 95% confidence interval, enter (1 – 0.95) = *0.05*. The confidence interval would be [0, 0.0582]. The 99% confidence interval would be [0, 0.0880]. The basic message is that from a simulation of only 50 replicates, all we can conclude is that $P(\max Z(\theta) > 2.65) \leq 0.0582$.

28.8 ISIM

There is an additional analysis program of use here, the ISIM program, which accurately computes the EMLOD, as described above. To use this program, you need to modify the DATAFILE.DAT file, such that it is in ILINK format. To do this, just read the DATAFILE.DAT into PREPLINK, and put the file in ILINK format, specifying that the recombination fraction should be iterated. Then, first run the UNKNOWN program, and then the ISIM program by just typing *ISIM* at the DOS prompt. The ISIM program is much slower than the MSIM program, so in many cases, it may be advisable to use MSIM, computing the lod scores at a minimum of three recombination fractions, and allowing it to approximate the EMLOD by quadratic interpolation, although it can be a good bit less accurate, as you can see from the following excerpt from the ISIM.DAT output file.

```
--------------------------------------------------
 Average Maximum   StdDev      Min         Max
 0.148284          0.307969    -0.000478   1.339499
--------------------------------------------------
```

Comparing these results with those obtained from MSIM, you can see that the EMLOD rose from 0.125075 to 0.148284 when the appropriate analysis was done with ISIM. The other quantities are comparable. Note that the *Min* column gives a value that is less than zero. This is the same situation that we observed with the ILINK program, where the lod score is not 100% maximized, but only to a specified tolerance. Clearly -0.0005 is almost the same as zero, so there is nothing to worry about from such results.

28.9 SLINK

The SIMULATE program is a very fast way of simulating pedigree data in a completely random fashion, but it does not allow the user to simulate marker loci conditional on previously known information, like disease locus phenotypes, or partially known marker information. To do a simulation under these more general conditions requires the use of far more sophisticated computer algorithms. The most versatile such program, available to date is the SLINK program of Weeks et al (1990b). This program employs an algorithm of Ott (1989), to simulate marker data, by the use of complicated likelihood methods. The basic approach is that the SLINK program uses the MLINK program to compute the conditional probabilities of each multilocus genotype (genotype risk) for any individual in a pedigree, given the known genotypes and phenotypes of the other pedigree members,

$$P(g_i \mid x_1,\ldots,x_n) = P(x_1,\ldots, x_i, g_i, x_{i+1},\ldots, x_n)/P(x_1,\ldots, x_n),$$

where the denominator is the likelihood of the entire set of pedigree data, and the numerator is the likelihood of the pedigree data given that individual i has genotype g_i. Both of these values can be computed easily with the MLINK program. The SLINK program uses the algorithm from the MLINK program to compute these probabilities for each multilocus genotype to be simulated in each individual, based on the model specified for each locus (including the disease and markers), and the recombination fractions between adjacent pairs of loci. Then, based on these conditional probabilities, the SLINK program selects pseudorandom numbers to simulate multilocus genotypes for each pedigree member, one at a time, each time conditioning the results on everything that has already been simulated; i.e. for the second individual, the appropriate conditional probability would be $P(g_j \mid x_1,\ldots,x_i,g_i,x_{i+1},\ldots,x_n)$. This conditioning procedure would then be continued iteratively for each successive individual, until the last individual was simulated according to the distribution of $P(g_n \mid x_1,g_1, x_2,g_2,\ldots,x_{n-1},g_{n-1},x_n)$. In this way, the SLINK program can simulate multilocus genotypes conditional on all previously simulated or previously known genotypic information in a pedigree. Unfortunately, as a consequence of this repeated conditioning algorithm, the program can be very slow, as you will see later. Other approaches to simulating marker genotypes conditional on disease phenotypes have been implemented in the SIMLINK program (Boehnke, 1986), and the CHRSIM program (Terwilliger et al, 1993; Speer et al, 1992). These programs are much faster than SLINK, but are not as general, since SLINK allows for the presence of partial marker typing, and the other methods do not. However, the CHRSIM program allows the user to perform simulations assuming map functions other than the Haldane function (i.e. allowing for interference), which is not possible with SIMLINK or SLINK. Further, CHRSIM, SIMULATE, and SLINK allow the user to simulate the data under one model, and analyze it under a different model, while the SIMLINK program does not.

The SLINK program is easy to use, and follows the same basic file format as the SIMULATE program with a couple of small differences. Since the SLINK program performs a simulation conditional on known phenotypes at any of the loci (including disease and/or marker loci, the pedigree file requires that you specify a genotype for each locus in each individual (specifying the *unknown* phenotype for all individuals to be simulated, and the known marker or disease phenotype for individuals whom you wish to predetermine the phenotypes). Then, at the end of each line in the pedigree file, you must input a so-called *availability code*. This is an integer between 0 and 4, the meanings of which are indicated in table 28-1.

Code	Trait	Markers
0	As indicated	Unknown (even if phenotypes are given)
1	Simulate	Simulate or use given phenotypes
2	As indicated	Simulate or use given phenotypes
3	Simulate	Unknown (even if phenotypes are given)

Table 28-1 Table of definitions of availability codes for SLINK

When the trait is selected to be *as indicated*, the trait phenotypes will be fixed as they are indicated in the SIMPED.DAT file. Otherwise, they will be simulated according to the parameters in the SIMDATA.DAT file. Typically, the trait is assumed to be *as indicated*, since you want to examine properties of the pedigrees as they have been collected assuming linkage to a trait locus with known phenotypes. At the marker locus, there are two options as well. Either an individual will be considered unknown at *ALL* marker loci, including those for which phenotypes are given in SIMPED.DAT, or the individual will be assigned a phenotype at all marker loci. If a phenotype other than *unknown* is indicated in the SIMPED.DAT file for a given locus, that phenotype will be used. Otherwise, the phenotype will be simulated for that locus. These codes should be indicated after the last locus in the SIMPED.DAT file. For the schizophrenia data, if we wished to simulate the disease being between the two marker loci, we could make the disease locus the second locus in our SIMPED.DAT and SIMDATA.DAT files. Let us assume that the locus order is Marker 1-(0.08)-Disease-(0.057)-Marker 2, and that we wish to simulate conditional on known disease locus phenotypes, and to simulate marker genotypes for individuals who had *both* markers typed in the original dataset, and to make leave all the markers unknown for individuals with zero or one marker typed in the original pedigrees. It is a limitation of SLINK that either all markers must be known or all markers unknown for a given individual, as opposed to SIMULATE, in which each marker can be specified independently. The individuals should have genotype *0 0* at each of the marker loci (such that all genotypes would be simulated by SLINK). The SIMPED.DAT file also should not have the header lines required by the SIMULATE program, as they are not needed by SLINK. The SIMPED.DAT file should resemble the following (for pedigree 1):

```
1 1 0 0 3 0 0 2 0 0 0 2 1 0 0 2
1 2 0 0 3 0 0 1 0 0 0 1 1 0 0 2
1 3 2 1 0 4 4 2 0 0 0 2 3 0 0 2
1 4 2 1 0 5 5 1 0 0 0 2 1 0 0 2
1 5 2 1 0 6 6 1 0 0 0 2 1 0 0 2
1 6 2 1 0 10 10 1 2 0 0 2 1 0 0 2
1 7 0 0 9 0 0 1 1 0 0 0 1 0 0 0
1 8 0 0 9 0 0 2 0 0 0 0 1 0 0 0
1 9 7 8 12 11 11 2 0 0 0 1 1 0 0 2
1 10 2 1 15 0 0 2 0 0 0 2 1 0 0 2
1 11 7 8 15 0 0 1 0 0 0 1 1 0 0 2
1 12 17 9 0 13 13 1 0 0 0 1 1 0 0 2
1 13 17 9 0 14 14 1 0 0 0 1 1 0 0 2
1 14 17 9 0 0 0 1 0 0 0 2 1 0 0 2
1 15 11 10 0 16 16 1 0 0 0 1 1 0 0 2
1 16 11 10 0 0 0 1 0 0 0 2 1 0 0 2
1 17 0 0 12 0 0 1 2 0 0 2 1 0 0 2
```

Similarly, the SIMDATA.DAT file should be in standard MLINK format as shown below:

```
 3 0 0 5 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0 0.0 0.0 0 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
 1 2 3
3 2 << ALLELE NUMBERS, NO. OF ALLELES
 0.33 0.67
1 2 << AFFECTION, NO. OF ALLELES
 0.010000 0.990000 << GENE FREQUENCIES
 3 << NO. OF LIABILITY CLASSES
0.44 0.44 0.0166
0.44 0.44 0.0166
0.44 0.44 0.0166
3 3 << ALLELE NUMBERS, NO. OF ALLELES
 0.32 0.16 0.52
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
0.08 0.057
 1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

There is one final input file required for the SLINK program, called SLINKIN.DAT, which contains the analysis parameters (note that SLINK does not require the PROBLEM.DAT file of SIMULATE, but rather this SLINKIN.DAT file). The file should consist of four lines as follows:

Line 1 - *Three* seeds for the random number generator between 1 and 30323. As in SIMULATE, numbers over 25000 perform better.

Line 2 - The number of replicates to be simulated

Line 3 - The locus number of the trait locus (in our example, the trait locus is locus 2. You would input a *0* if there were no trait locus.)

Line 4 - The proportion of unlinked families (if you want to allow for heterogeneity, you should input the value of $(1 – \alpha)$ here, where α is the proportion of linked families as defined in section 6.3). Typically, this value is set to be 0, to assume homogeneity in the simulation. (The properties of simulation under heterogeneity are beyond the scope of this book and will not be discussed here.)

For our sample data set, let us again simulate 50 replicates of our pedigree set (though in practice, you would typically want to simulate a much larger set of replicates to get more reliable results). The SLINKIN.DAT file should look like the following:

```
28733
50
2
0
```

To perform the simulation, just type *SLINK* at the DOS prompt. Note that SLINK is incredibly slow, taking 33 hours of CPU time on a 486, due to the amount of computation required, as can be seen in the algorithm described above, since each time a new individual has to be simulated, the MLINK program must compute the conditional probability of each possible multilocus genotype for the individual to be simulated. In fact, it is precisely because of the incredibly lengthy amounts of computer time required to do an SLINK analysis that the SIMULATE program was developed. If you remember, to simulate 50 replicates of the two linked markers (unlinked to the disease), it took a matter of only several seconds, while to do the identical simulation with SLINK would again take on the order of 33 hours on a 486. As a tradeoff for the increased computing time, however, much additional flexibility is available, allowing for both simulation conditional on the disease phenotypes, and on previously typed marker data. For practice, you may want to try simulating only five replicates for now. We will present the analysis results for both five and fifty replicates.

The file produced by SLINK is also called PEDFILE.DAT, as was the case with SIMULATE. Again, we have three loci in our simulated data, this time with the disease being locus *2*, and the markers being loci *1* and *3*. Since we initially want to perform two-point analysis of the data, you should rename the PEDFILE.DAT to SIM.PED. To analyze the data under the model we simulated for the disease and markers, copy the SIMDATA.DAT file to SIM.DAT. Then, use LSP to extract loci *1* and *2* from these files, setting up the parameter file for MLINK, with θ ranging from 0 through 0.5 in steps of 0.05. Then, run UNKNOWN, followed by MSIM, as you did in the previous section. Then, alter the DATAFILE.DAT file to put it in the proper format for ISIM, with starting value of $\theta = 0.1$. The results should match those found in table 28-2 (results given for both cases, five replicates and fifty replicates). Next, repeat the process for marker 2 (extract loci *2* and *3* from the SIM.PED file with LSP), the results for which are indicated in table 28-3.

θ	$E[Z(\theta)]_{50}$	StdDev_{50}	$E[Z(\theta)]_{5}$	StdDev_{5}
0	0.735496	1.166530	0.414042	1.062897
0.05	0.836342	0.977525	0.622942	0.863761
0.10	0.816879	0.820165	0.655215	0.738885
0.15	0.742267	0.675951	0.617683	0.622972
0.20	0.634117	0.540453	0.540525	0.508303
0.25	0.505767	0.412221	0.439205	0.394128
0.30	0.368617	0.292287	0.324957	0.282457

0.35	0.234871	0.184363	0.208760	0.177726
0.40	0.118823	0.094987	0.103985	0.087554
0.45	0.036177	0.032759	0.028154	0.023666
0.50	0.000000	0.000000	0.000000	0.000000

	$E[maxZ(\theta)]_{50}$	StdDev$_{50}$	$E[maxZ(\theta)]_{5}$	StdDev$_{5}$
MSIM	1.045271	0.822056	0.819046	0.582468
ISIM	1.059299	0.808330	0.818405	0.584770

	Observed Lod Scores Greater Than a Given Constant			
Constant	Number$_{50}$	Percent$_{50}$	Number$_{5}$	Percent$_{5}$
1	20	40%	1	20%
2.65	2	4%	0	0%
3	0	0%	0	0%

Table 28-2: Results of MSIM and ISIM analyses of disease vs. marker 1 based on both 50 and 5 replicates (all pedigrees)

θ	$E[Z(\theta)]_{50}$	StdDev$_{50}$	$E[Z(\theta)]_{5}$	StdDev$_{5}$
0.00	1.344803	1.477235	0.926615	1.352813
0.05	1.578782	1.267425	1.276471	1.165699
0.10	1.555690	1.103191	1.351794	0.985503
0.15	1.428120	0.944638	1.299463	0.822139
0.20	1.236861	0.784552	1.165164	0.670369
0.25	1.004403	0.621787	0.974942	0.525907
0.30	0.749474	0.458352	0.749062	0.386712
0.35	0.493028	0.299932	0.508250	0.253790
0.40	0.261602	0.158164	0.278780	0.133944
0.45	0.087424	0.052783	0.096245	0.044466
0.50	0.000000	0.000000	0.000000	0.000000

	$E[maxZ(\theta)]_{50}$	StdDev$_{50}$	$E[maxZ(\theta)]_{5}$	StdDev$_{5}$
MSIM	1.751465	1.171508	1.471560	1.028230
ISIM	1.750680	1.169886	1.467988	1.031358

	Observed Lod Scores Greater Than a Given Constant			
Constant	Number$_{50}$	Percent$_{50}$	Number$_{5}$	Percent$_{5}$
1	36	72%	2	40%
2.65	11	22%	1	20%
3	7	14%	1	20%

Table 28-3: Results of MSIM and ISIM analyses of disease vs. marker 2 based on both 50 and 5 replicates.

28.10 LSIM

There is also a version of the LINKMAP program designed to analyze simulated data from SLINK or SIMULATE. In both cases, we simulated multilocus data, so it may be of value to see the properties of the multipoint analysis, in terms of expected multipoint lod score, etc. To do this, you must use the original SIM.PED file (containing the disease and both markers), by copying it to PEDFILE.DAT. Then, you should modify the SIMDATA.DAT file, such that it is in LINKMAP format. The file should specify that the locus order is *2 1 3*, meaning that the disease will start outside the set of linked marker loci on the right, as is usually the case in a LINKMAP analysis. Then, the recombination fractions should be set to *0.5* (since the disease should start out unlinked to the markers), and *0.128* (the recombination fraction between the two markers). Finally, you should specify that the trait locus is locus *2*, with *5* evaluations per interval, and finishing value of *0* (this number is irrelevant to the analysis, but is required in the parameter file). The final version of this parameter file should look like this:

```
 3 0 0 5 << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
 0 0.0 0.0 0 << MUT LOCUS, MUT MALE, MUT FEM, HAP FREQ (IF 1)
 2 1 3
3 2 << ALLELE NUMBERS, NO. OF ALLELES
 0.33 0.67
1 2 << AFFECTION, NO. OF ALLELES
 0.010000 0.990000 << GENE FREQUENCIES
 3 << NO. OF LIABILITY CLASSES
0.44 0.44 0.0166
0.44 0.44 0.0166
0.44 0.44 0.0166
3 3 << ALLELE NUMBERS, NO. OF ALLELES
 0.32 0.16 0.52
 0 0 << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
0.50 0.128
 2 0 5 << LOCUS VARIED, FINISHING VALUE, NU OF EVALUATIONS
```

At this point, you can save this file as DATAFILE.DAT and run the UNKNOWN program, followed by the LSIM program. The resulting expected lod scores are given in table 28-4. If there is linkage, then the multipoint lod scores are typically much higher than the two-point lod scores from the same dataset. In this situation, our ELOD is 2.25, which is substantially higher than the ELOD from either two-point linkage analysis. This is because there are many more informative meioses for linkage when both markers are typed. It is perhaps more striking to look at the power of these pedigrees. If we examine the probability of finding a lod score greater than 3, we will note that with the more informative marker, marker *2*, the probability of exceeding a lod score of 3 in two-point analysis was 0.14, and with marker one, the probability was approximately 0. However, in the three-point LSIM analysis, the probability of getting a multipoint lod score greater than 3 was raised to 0.36. While this would still not be considered to be enough power to encourage one to invest in linkage analysis with these pedigrees and these markers alone, it is substantially better than the power of 0.14 with just marker *2*. This increase in information can also be taken advantage of through the polylocus method of Terwilliger and Ott (1993), in the case of complex disease analysis, without the need to rely on the non-robust and time consuming process of multipoint analysis.

Locus Order	$E[Z(x)]_{50}$	$StdDev_{50}$	$E[Z(x)]_{5}$	$StdDev_{5}$
2-0.500-*1*-0.128-*3*	0.000000	0.000000	0.000000	0.000000
2-0.400-*1*-0.128-*3*	0.254666	0.133885	0.236215	0.106462
2-0.300-*1*-0.128-*3*	0.767525	0.405632	0.674440	0.293006
2-0.200-*1*-0.128-*3*	1.330319	0.719721	1.109016	0.484413
2-0.100-*1*-0.128-*3*	1.783796	1.039690	1.378166	0.694069
2-0.000-*1*-0.128-*3*	1.888853	1.399177	1.175593	1.087273
1-0.000-*2*-0.128-*3*	1.888853	1.399177	1.175593	1.087273
1-0.026-*2*-0.108-*3*	2.104250	1.369755	1.422980	0.997804
1-0.051-*2*-0.086-*3*	2.208696	1.376830	1.530529	1.034229
1-0.077-*2*-0.060-*3*	2.249394	1.405504	1.561322	1.107357
1-0.102-*2*-0.032-*3*	2.211023	1.457537	1.502997	1.208236
1-0.128-*2*-0.000-*3*	1.961298	1.549443	1.251705	1.329041
1-0.128-*3*-0.000-*2*	1.961298	1.549443	1.251705	1.329041
1-0.128-*3*-0.100-*2*	2.047632	1.148514	1.630032	1.027003
1-0.128-*3*-0.200-*2*	1.574862	0.817627	1.356377	0.700748
1-0.128-*3*-0.300-*2*	0.932305	0.477947	0.848989	0.405352
1-0.128-*3*-0.400-*2*	0.316440	0.164488	0.305904	0.143177
1-0.128-*3*-0.500-*2*	0.000000	0.000000	0.000000	0.000000

	Observed Lod Scores Greater Than a Given Constant			
Constant	Number$_{50}$	Percent$_{50}$	Number$_{5}$	Percent$_{5}$
1	46	92%	4	80%
2.65	21	42%	1	20%
3	18	36%	1	20%

Table 28-4: Results of LSIM analysis of disease versus fixed map of two linked markers in replicates simulated with SLINK in file SIM.PED.

It is also the case that when there is no linkage in reality, the multipoint lod scores tend to be lower than the two-point lod scores, and one's ability to do exclusion mapping is enhanced by multipoint analysis, but *only* if the model is known with accuracy. In this case, we have been doing simulation on a dataset that is segregating a complex disorder, for which the model is not accurately known. For that reason, we will not be considering multipoint results for the simulation with the disease unlinked to the markers, since it would be of little meaning with a truly complex disease (cf. section 25.3).

EXERCISE 28

Use SLINK to simulate marker 1 from the schizophrenia dataset under the recessive model, diagnostic scheme 3, linked to the disease at θ = 0.05. Then, analyze it under the dominant model for diagnostic scheme 3. Next, simulate the same marker, again at θ = 0.05 from the disease, under the dominant model for diagnostic scheme 3, and analyze the data assuming the recessive model for the same diagnostic class. Do the results of this experiment confirm what was discussed in chapter 17 about analysis under the incorrect model?

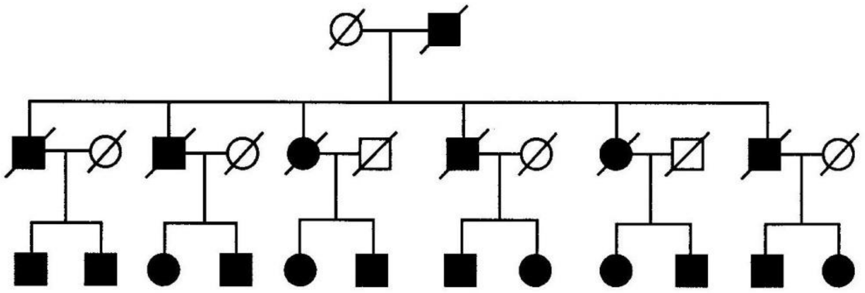


Figure 28–3. Pedigree to be simulated to show the effect of gene frequencies

Finally, to demonstrate the effect of gene frequencies on a linkage analysis, please use the SIMULATE program to simulate 100 replicates of an unlinked marker to the fully penetrant dominant disease (p = 0.00001) in the family in Figure 28-3. Simulate the marker under the assumption that there are 5 alleles, with gene frequencies 0.05, 0.05, 0.05, 0.05, and 0.80. All persons with a "/" through them are dead, and therefore should not be simulated (i.e. they should all be assigned genotype *0 0* at the marker locus). Then, analyze the pedigree assuming equal gene frequencies for all five alleles. Is there false positive evidence for linkage in this example?

29 Solutions to Part III Problems

EXERCISE 21

Under the assumption that the disease in the pedigree from exercise 6 is a fully penetrant sex-linked recessive lethal disorder, the mutation rate would have to be $p/2 = 0.01/2 = 0.005$ in this example. Reanalyzing this pedigree allowing for $\mu = 0.005$ should yield the results shown in table 29-1. Compare these results with those obtained when mutation was not allowed for (see table 12-6), and in this instance, you will see that there is negligible difference between the two situations. This is due primarily to the large density of affecteds in this pedigree, which makes it much more likely that the disease allele is segregating than that a new mutation occurred at some point in the pedigree. Hence, allowing for mutation has little effect on the analysis, as explained in chapter 21.

θ	Lod Score
0.0	-0.234382
0.1	-0.170833
0.2	-0.021145
0.3	0.010842
0.4	-0.019488
0.5	0.000000

ILINK: $\hat{\theta}$ = 0.862; Z($\hat{\theta}$) = 0.554451

Table 29-1: Results of analysis of sex-linked recessive disease pedigree from exercise 6, allowing for $\mu = 0.005$.

In the pedigree from exercise 7, there is a fully penetrant recessive disease with $p = 0.001$. As explained in this chapter, at equilibrium, $\mu = sp^2$, so in this case, $\mu = sp^2 = (0.5)(0.001)^2 = 0.0000005$. Incorporating this information in the linkage analysis leads to the results shown in table 29-2. Again, due to the high density of affecteds in this pedigree, there is little effect of allowing for the possibility of mutation, with extremely slight reductions in lod score. Further, in this case, the mutation rate is 2000 times smaller than the gene frequency, which is minute, as is typically the case with autosomal recessive disorders, in contrast to the sex-linked case, when the mutation rate is as large as half the gene frequency.

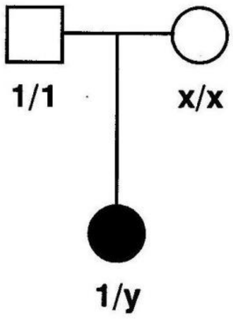


Figure 29–1. Pedigree with internal inconsistency for mutation experiment

Theta	Lod Score
0.0	4.210082
0.1	3.300826
0.2	2.305048
0.3	1.246533
0.4	0.349473
0.5	0.000000

ILINK: $\hat{\theta}$ = 0.0; Z($\hat{\theta}$) = 4.210082

Table 29-2 : Results of analysis of autosomal recessive disease pedigree from exercise 7, assuming $\mu = 0.0000005$.

To design an experiment to determine the types of mutation allowed for in the LINKAGE programs, we could do the following: Set up a pedigree with an internal inconsistency, as shown in Figure 29-1. Now, let us assume that the disease is autosomal recessive with full penetrance and gene frequency of 0.01; the marker locus has four alleles, with equal frequencies. In figure 29-1, the mother has marker genotype *x/x*, and her daughter has genotype *1/y*, meaning that she had to have received the *y* allele from her mother (since her father was homozygous *1/1*. Clearly, whenever $x \neq y$, there is an inconsistency, and our lod scores will all be $-\infty$. However, if we allow for mutation rate μ at the **marker** locus, then we should get lod scores of 0 everywhere (since there is obviously no information for linkage in this little pedigree, yet the inconsistency would be eliminated), whenever allele *x* is allowed to mutate into allele *y* with probability μ. If mutation were allowed in all directions, then we would always get lod scores of 0 in this pedigree, and never get lod scores of $-\infty$. To determine which types of mutation are allowed, please try analyzing this pedigree, allowing for mutation rate $\mu = 0.001$ at the marker locus (locus *2* in this example), and assuming all possible combinations of *x* and $y \in (1, 2, 3, 4)$. Then compute the lod score at recombination fraction $\theta = 0.1$ (without loss of generality). The results should match those shown in table 29-3. This shows us that the only types of mutation allowed in the LINKAGE programs is from any allele to the last allele at the locus. You can try and verify this for other locus types, and other numbers of alleles, but it is the general solution that mutations to the last (i.e. highest numbered) allele at any locus are the only ones permitted in this implementation of mutation. It is primarily for this reason that we always specify the second allele to be the disease predisposing allele at an affection status locus, so we could allow for mutations to the disease allele, and not away from it.

	y			
x	1	2	3	4
1	0.0	-∞	-∞	0.0
2	-∞	0.0	-∞	0.0
3	-∞	-∞	0.0	0.0
4	-∞	-∞	-∞	0.0

Table 29-3 : Lod Scores, Z(θ = 0.1), for pedigree in figure 29-1, assuming that μ = 0.001 at the marker locus.

EXERCISE 22

In this pedigree, the first thing you will need to do is to extract the disease locus and the ABO blood group locus from the files USEREX8.* with LSP. Then, you must modify the DATAFILE.DAT file as explained in chapter 22, to estimate allele frequencies for the ABO locus. There are two separate ways to do this. The first would be to set the recombination fraction between disease and ABO to 0.5, and estimate allele frequencies, in which case, you should get estimates of 0.288, 0.343, and 0.369 for the A, B, and O alleles respectively. When estimated jointly with the recombination fraction, the estimates are revised slightly to be 0.277, 0.341, and 0.382 respectively, with recombination fraction = 0.001. Now, there are three possible ways to compute lod scores, given these gene frequency estimates. The first approach would be to simply consider $Z(\hat{\theta}) = \log_{10}[\ L(\hat{\theta}, \hat{p}_{\ i})/L(\theta = ½,_{i})]$, where the $\hat{p}_{\ i}$ are estimated jointly with the recombination fraction in the numerator, and kept the same in the denominator. In this case, the lod score (at θ = 0), is equal to 3.459960. Another possible approach would be to compute lod scores using the gene frequency estimates obtained when the disease was considered to be unlinked to the marker (i.e. do the estimation of allele frequencies in the denominator and use these same estimates in the numerator), in which case the lod score is Z(θ = 0) = 3.454484. The best way to compute this lod score, however, would be to treat the gene frequency estimates as nuisance parameters, and compute the lod score with separate estimates of the p_i in numerator and denominator, as $\log_{10}[L(\hat{\theta}, \hat{p}_{\ i})] - \log_{10}[L(\theta = ½, \hat{p}_{\ i})]$, where the $\hat{p}_{\ i}$ are separately estimated in each term. In this case, the resulting lod score is (–23.299604) – (–26.756902) = 3.457298. In this example, the lod scores are not greatly affected by the changes in gene frequency estimates at ABO, but in some cases they can be quite significant, so this procedure is usually advisable, when accurate allele frequency estimates are unavailable.

If we consider the example from exercise 9, when a reduced penetrance model is applied to the same pedigree, and re-estimate the allele frequencies, we obtain identical gene frequency estimates of 0.288, 0.343, and 0.369 for the A, B, and O alleles respectively, when the disease and marker are assumed to be unlinked. This is true because the estimation of allele frequency is done essentially independently of the disease phenotypes in the pedigree. However, when we estimate the allele frequencies jointly with recombination fraction, we obtain estimates of 0.277, 0.341, and 0.382, which are again identical to those estimated with the full penetrance model. However, this is not typically going to be the case unless there is little ambiguity as to the disease locus genotypes of the founder individuals, which is clearly the case in this pedigree. The three lod scores, as computed above, are now as follows: using the gene frequency estimates obtained when $\theta = \hat{\theta}$, we get Z(θ = 0) = 2.172223; using the gene frequency estimates obtained under θ = ½, we get Z(θ = 0) = 2.166747; and finally, separately estimating the gene frequencies in each term, we get Z(θ = 0) = (–25.497910) – (–27.667471) = 2.169561, which is again right between the two lod scores computed with fixed gene frequency estimates.

EXERCISE 23

The first problem is to compute the genotype probabilities in terms of haplotype probabilities for use in the EH program, under the assumption of one four-allele locus, and one three-allele locus. These frequencies are given in table 29-4.

	1/1	1/2	2/2	1/3	2/3	3/3
1/1	$P_{1\,1}^2$	$2P_{1\,1}P_{1\,2}$	$P_{1\,2}^2$	$2P_{1\,1}P_{1\,3}$	$2P_{1\,2}P_{1\,3}$	$P_{1\,3}^2$
1/2	$2P_{1\,1}P_{2\,1}$	$2[P_{1\,1}P_{2\,2} + P_{1\,2}P_{2\,1}]$	$2P_{1\,2}P_{2\,2}$	$2[P_{1\,1}P_{2\,3} + P_{1\,3}P_{2\,1}]$	$2[P_{1\,2}P_{2\,3} + P_{1\,3}P_{2\,2}]$	$2P_{1\,3}P_{2\,3}$
2/2	$P_{2\,1}^2$	$2P_{2\,1}P_{2\,2}$	$P_{2\,2}^2$	$2P_{2\,1}P_{2\,3}$	$2P_{2\,2}P_{2\,3}$	$P_{2\,3}^2$
1/3	$2P_{1\,1}P_{3\,1}$	$2[P_{1\,1}P_{3\,2} + P_{1\,2}P_{3\,1}]$	$2P_{1\,2}P_{3\,2}$	$2[P_{1\,1}P_{3\,3} + P_{1\,3}P_{3\,1}]$	$2[P_{1\,2}P_{3\,3} + P_{1\,3}P_{3\,2}]$	$2P_{1\,3}P_{3\,3}$
2/3	$2P_{2\,1}P_{3\,1}$	$2[P_{2\,1}P_{3\,2} + P_{2\,2}P_{3\,1}]$	$2P_{2\,2}P_{3\,2}$	$2[P_{2\,1}P_{3\,3} + P_{2\,3}P_{3\,1}]$	$2[P_{2\,2}P_{3\,3} + P_{2\,3}P_{3\,2}]$	$2P_{2\,3}P_{3\,3}$
3/3	$P_{3\,1}^2$	$2P_{3\,1}P_{3\,2}$	$P_{3\,2}^2$	$2P_{3\,1}P_{3\,3}$	$2P_{3\,2}P_{3\,3}$	$P_{3\,3}^2$
1/4	$2P_{1\,1}P_{4\,1}$	$2[P_{1\,1}P_{4\,2} + P_{1\,2}P_{4\,1}]$	$2P_{1\,2}P_{4\,2}$	$2[P_{1\,1}P_{4\,3} + P_{1\,3}P_{4\,1}]$	$2[P_{1\,2}P_{4\,3} + P_{1\,3}P_{4\,2}]$	$2P_{1\,3}P_{4\,3}$
2/4	$2P_{2\,1}P_{4\,1}$	$2[P_{2\,1}P_{4\,2} + P_{2\,2}P_{4\,1}]$	$2P_{2\,2}P_{4\,2}$	$2[P_{2\,1}P_{4\,3} + P_{2\,3}P_{4\,1}]$	$2[P_{2\,2}P_{4\,3} + P_{2\,3}P_{4\,2}]$	$2P_{2\,3}P_{4\,3}$
3/4	$2P_{3\,1}P_{4\,1}$	$2[P_{3\,1}P_{4\,2} + P_{3\,2}P_{4\,1}]$	$2P_{3\,2}P_{4\,2}$	$2[P_{3\,1}P_{4\,3} + P_{3\,3}P_{4\,1}]$	$2[P_{3\,2}P_{4\,3} + P_{3\,3}P_{4\,2}]$	$2P_{3\,3}P_{4\,3}$
4/4	$P_{4\,1}^2$	$2P_{4\,1}P_{4\,2}$	$P_{4\,2}^2$	$2P_{4\,1}P_{4\,3}$	$2P_{4\,2}P_{4\,3}$	$P_{4\,3}^2$

Table 29-4 : Table of genotype probabilities for one three-allele marker and one four-allele marker. $P_{i\,j}$ is the haplotype frequency for allele i at the four-allele marker and allele j at the three-allele marker.

The dataset from table 23-10 should yield the following results from the EH program:

```
Estimates of Gene Frequencies (Assuming Independence)
----\-------------------------------------------------------
locus \ allele    1           2           3           4
---------\--------------------------------------------------
 1      |         0.2236      0.3650      0.4114
 2      |         0.3439      0.1857      0.1730      0.2975
------------------------------------------------------------
# of Typed Individuals: 237

There are 12 Possible Haplotypes of These 2 Loci.
They are Listed Below, with their Estimated Frequencies:

-------------------------------------------------
| Allele Allele   |   Haplotype Frequency        |
|   at      at    |                              |
| Locus 1 Locus 2 | Independent w/Association    |
-------------------------------------------------
    1       1           0.076902    0.106861
    1       2           0.041518    0.038783
    1       3           0.038687    0.045654
    1       4           0.066522    0.032331
    2       1           0.125510    0.149313
    2       2           0.067760    0.077541
    2       3           0.063140    0.048721
    2       4           0.108570    0.089404
    3       1           0.141470    0.087708
    3       2           0.076377    0.069330
    3       3           0.071169    0.078621
    3       4           0.122376    0.175734
-------------------------------------------------
# of Iterations = 7
```

```
                                            df     Ln(L) Chi-square
-------------------------------------------------------------------
H0: No Association                           5    -983.22      0.00
H1: Allelic Associations Allowed            11    -965.53     35.37
```

In this case, there were 5 free parameters under the assumption of linkage equilibrium, and 11 under the assumption of allelic association. Therefore, our chi-square statistic, 35.37, has (11 – 5) = 6 df, and is significant at the 0.000004 level, indicating that there is some association between the alleles of these two loci. If we were to consider only the unambiguous haplotypes, and simply count them up, we would fill a 4 × 3 table like that shown in table 29-5. When one does a chi-square test of independence on this table, the chi-square statistic with 6 df is 30.07, corresponding to a p-value of 0.000038, which is still highly significant. However, our new haplotype frequency estimates (computed as k_i/n for each cell) are given in table 29-6. These estimates are somewhat different from those obtained using the EH program, but not dramatically so.

	Allele at Locus 2			
Allele at Locus 1	1	2	3	4
1	42	14	13	12
2	58	25	16	31
3	37	26	29	63

Table 29-5 : Haplotypes unequivocally determinable from genotype data given in table 23-10.

	Allele at Locus 2			
Allele at Locus 1	1	2	3	4
1	0.115	0.038	0.036	0.033
2	0.158	0.068	0.044	0.085
3	0.101	0.071	0.079	0.172

Table 29-6 : Haplotype frequencies estimated from data in table 29-5.

In the pedigree from exercise 8, there are only six founder individuals, two of whom are untyped at the marker locus. Therefore, there is not likely to be much power to estimate haplotype frequencies. For simplicity, assume that all twelve haplotype frequencies are 0.08, as starting values (with the last frequency set to 0.12, such that they sum up to 1). The estimated haplotype frequencies and corresponding lod scores are given in table 29-7, with lod scores provided both assuming the haplotype frequencies are estimated only in the numerator (then fixed in the denominator) of the likelihood ratio, and again with the frequencies estimated separately in numerator and denominator.

	Haplotype Frequency					
Haplotype	θ=0.5	θ=0.4	θ=0.3	θ=0.2	θ=0.1	$\hat{\theta}$=0.22
1 A	0.090	0.090	0.090	0.090	0.091	0.090
1 B	0.094	0.094	0.094	0.093	0.092	0.094
1 O	0.095	0.095	0.095	0.095	0.093	0.095
2 A	0.090	0.090	0.088	0.090	0.091	0.090
2 B	0.095	0.095	0.094	0.094	0.093	0.094
2 O	0.082	0.082	0.084	0.082	0.081	0.082
3 A	0.092	0.092	0.094	0.092	0.093	0.092
3 B	0.087	0.087	0.088	0.087	0.088	0.087
3 O	0.102	0.102	0.102	0.101	0.102	0.102
4 A	0.089	0.089	0.087	0.089	0.091	0.089

4 B	0.084	0.084	0.085	0.084	0.084	0.084
4 O	0.000	0.000	0.000	0.002	0.001	0.000
Z(θ)	0.000	0.416	0.760	0.886	0.459	0.898
-2lnL(θ)	126.9	125.0	123.3	123.0	125.0	122.8
Z(Θ,δ)	0.000	0.413	0.782	0.848	0.413	0.891

Table 29-7 : Haplotype frequency estimates from pedigree in exercise 8, for marker 1 vs. ABO blood group. Lod scores are provided both given the estimated haplotype frequencies and given haplotype frequencies are estimated separately in numerator and denominator.

EXERCISE 24

With the simplifying assumption that all parents are unaffected, and the disease is fully penetrant recessive, we know that every parent is heterozygous for the disease allele, and we can therefore attempt to estimate the haplotype frequencies with the EH program. To do this, do not use the case-control option, but merely enter every parent as if he was heterozygous (1/2) at the first locus, and treat the H allele and the allele as alleles *1* and *2* respectively at the second locus. Then, we would have A 1/1 haplotypes, (B + C) 1/2 haplotypes, and D 2/2 haplotypes in our parental sample. When we run the EH program using the data from section 24.6, we have A = 19, (B + C) = 52, and D = 29. Our output should tell us that there is no evidence for any allelic association whatsoever. This is because everyone is heterozygous 1/2 at the disease locus, so it is not possible to discern the phase at all. However, if we were to try and construct pedigrees which match the data, and analyze the data with the ILINK program, we might be able to use the offspring genotypes to reconstruct the phases in the parents. Remember that we showed in chapter 24 that nuclear pedigrees with one offspring **do** contain linkage information when there is disequilibrium, but we haven't discussed the reverse situation. Let us try it now. Although we had separate information for the GHRR and HHRR tests, collected from the same pedigree set, since the disease is fully penetrant recessive, and the parents are unaffected, we only need to be sure we have the data in our pedigrees match that provided for the HHRR test. For simplicity's sake, let us assume that there are nine pedigrees of the structure HH × HH → HH, one pedigree of the form HH × H → HH, twenty pedigrees of the form H × H → HH, one pedigree of the form H x → H, five pedigrees of the form H × H → , and fourteen pedigrees of the form x → . Then you can run this analysis with ILINK, estimating haplotype frequencies and recombination fraction jointly. First assume starting values of 0.4, 0.1, 0.1, and 0.4 for the four haplotype frequencies, and 0.1 for the recombination fraction. Then, use the new estimates as starting values to refine them. Repeat this process two additional times, and the resulting haplotype frequency estimates should be 0.1448, 0.3548, 0.3055, and 0.1949, with Θ = 0.001, and a lod score of 4.58, and –2ln(Like) = 531.44. This lod score, however, is not meaningful to us, as it assumes the same haplotype frequencies under the null hypothesis as well. To be fully accurate, we should re-estimate them when Θ = 0.5. When this is done, the corresponding haplotype frequency estimates should be 0.2807, 0.2192, 0.1692, and 0.3308, with –2ln(Like) = 552.26. The overall lod score should therefore be (552.26 – 531.44)/4.6 = 4.53, which is actually almost the same as with the other haplotype frequency estimates. This is true, because under the null hypothesis of free recombination, the haplotype frequencies have much less influence on the pedigree likelihood. So far, we have determined a lod score, evaluating our evidence for linkage, but we have yet to test the significance of our disequilibrium. To do this, we would need to maximize the likelihood over allele frequencies for each locus, assuming absence of allelic association. When you do this, you should find that the best estimate for the disease allele frequency is 0.5 (which makes sense, since we know that every founder is heterozygous for the disease allele), and 0.45 for the H allele at the marker locus (coded as the *1* allele in the datafiles). The likelihood is now going to be independent of the recombination fraction, since phase unknown pedigrees with only one offspring are uninformative for linkage in the absence of linkage disequilibrium. Therefore, we can just calculate the appropriate likelihood with the ILINK program, fixing θ = 0.1, without loss of generality, to get –2ln(Like) = 552.51. Now, we know that the difference in –2ln(Like) is distributed as a chi-square statistic, in this case with 1 degree of freedom. We have 552.51 – 531.44 = 21.07, which is significant at the 0.000004 level, so there is highly significant evidence for linkage disequilibrium in this parametric analysis. One problem remains, however, which is that we have drastically overestimated the disease allele frequency in every case. Let us consider the haplotype frequencies estimated with ILINK, as summarized in table 29-8.

	$\theta = \hat{\theta}$		θ = 0.5	
Haplotype	Estimated	Normalized	Estimated	Normalized
d H	0.3055	0.006105	0.1692	0.003384
d $\bar{H}$	0.1949	0.003985	0.3308	0.006616
+ H	0.1448	0.2869	0.2807	0.5558
+ $\bar{H}$	0.3548	0.7031	0.2193	0.4342

Table 29-8 : Haplotype frequency estimates from HHRR pedigrees computed with ILINK, and normalized to fit the known population gene frequency of the disease allele, *p* = 0.01.

In this table, the "normalized" haplotype frequencies are also provided. These are computed as explained in section 24.7, given the known disease allele frequency of 0.01. It is always required to then reevaluate the likelihoods, and significance level of your test given this constraint, which you have applied a posteriori. In this case, when you reanalyze the data, it should have no effect on the significance level of either test, since all genotypes are known with certainty, but to demonstrate this, the lod score at θ = 0.001 is still 4.58, with –2ln(Like) = 1177.23. At θ = 0.5, –2ln(Like) = 1198.04. The difference is just 20.81, corresponding to a lod score of 4.5, as before. The test for linkage disequilibrium as well maintains its significance. To test this, compute the likelihood assuming the disease allele frequency to be 0.01, and the H allele frequency to be 0.45, again at θ = 0.1, without loss of generality. In this case, –2ln(Like) = 1198.30, for a chi-square statistic of 1198.30 – 1177.23 = 21.07, exactly as before. In general pedigrees, however, the significance of your statistics may change dramatically after normalization of the haplotype frequencies, and they may no longer be optimal either, given the additional constraint on the disease allele frequency.

EXERCISE 25

The analysis parameters for the six models to be considered are computed as shown in table 29-9. The results of the linkage analysis of the schizophrenia pedigrees under these six models are shown in table 29-10. As you can see, the maximized over models maximum lod score is quite a lot larger, at 2.937, which would be almost significant in a single-model analysis with a simple Mendelian disorder. However, we have to use a cutoff point of 5 + $\log_{10}(m)$, where *m* is the number of models tested (this is too stringent). Here, again, *m* = 6, so our cutoff point is 5.78, and we are only halfway to a significant lod score in this analysis.

Model	Diagnosis	φ	R	p	f	f_p	k
Dominant	Narrow	0.005	0.1	0.01	0.23	0.0005	0.0023
	Medium	0.01	0.35	0.01	0.33	0.0036	0.0109
	Broad	0.03	0.50	0.01	0.75	0.015	0.0203
Recessive	Narrow	0.005	0.1	0.125	0.288	0.0005	0.0018
	Medium	0.01	0.35	0.125	0.416	0.0036	0.0085
	Broad	0.03	0.50	0.125	0.96	0.0152	0.0158

Table 29-9: Penetrance models used in exercise 25.

		Dominant Lod Scores		Recessive Lod Scores	
Diagnosis	θ	Marker 1	Marker 2	Marker 1	Marker 2
Narrow	0	1.649	-0.580	0.584	-1.503
	0.1	1.266	0.630	0.751	-0.639
	0.2	0.819	0.720	0.546	-0.269
	0.3	0.395	0.469	0.275	-0.093
	0.4	0.098	0.147	0.072	-0.020
ILINK:	0.001	1.646 0.164	0.743 0.071	0.765 0.500	0.000

Medium	0	2.239		0.770		0.945		-0.993
	0.1	1.818		1.613		1.200		-0.351
	0.2	1.206		1.389		0.874		-0.117
	0.3	0.603		0.835		0.447		-0.034
	0.4	0.158		0.249		0.119		-0.008
ILINK:	0.001	2.238	0.109	1.615	0.072	1.227	0.500	0.000
Broad	0	2.937		1.143		0.970		-5.767
	0.1	2.518		2.584		2.066		-1.800
	0.2	1.776		2.234		1.746		-0.666
	0.3	0.963		1.406		1.016		-0.222
	0.4	0.276		0.462		0.301		-0.048
ILINK:	0.003	2.937	0.107	2.586	0.108	2.069	0.500	0.000

Table 29-10: Results of linkage analysis of schizophrenia pedigrees using the penetrance models in table 29-9.

The affecteds-only analysis was done assuming the penetrances, P(Aff | Genotype), were all divided by 1000, making "unaffected" individuals essentially unknown in the analysis, as the penetrance ratio for unaffecteds f_P/f, would be approximately equal to 1. The results of this analysis are shown in table 29-11, and are typically less significant than the regular analysis done before (table 29-10). The most striking change occurs under the recessive model, with marker 2, in which highly negative lod scores, and estimates of = 0.5, were changed to slightly positive lod scores in the affecteds only analysis, with corresponding reductions in the estimate of . This is especially noticeable under the broad diagnostic class, which had a penetrance of 0.96 for susceptible genotypes in the regular analysis. This translated into a penetrance ratio in unaffecteds of *k* = 24.5, which made unaffecteds particularly informative for linkage. Eliminating this information led to slightly positive lod scores, and $\hat{\theta}$ = 0.05. Thus you can see the potential perils of putting too much emphasis on unaffected individuals, when you are dealing with a disease of uncertain diagnosis and mode of inheritance.

		Dominant Lod Scores				Recessive Lod Scores		
Diagnosis	θ	Marker 1		Marker 2		Marker 1		Marker 2
Narrow	0	1.625		0.095		0.675		-0.651
	0.1	1.231		0.693		0.625		-0.280
	0.2	0.782		0.632		0.421		-0.119
	0.3	0.372		0.377		0.208		-0.041
	0.4	0.092		0.112		0.055		-0.009
ILINK:	0.001	1.623	0.127	0.709	0.029	0.693	0.500	0.000
Medium	0	2.003		0.951		0.909		-0.168
	0.1	1.530		1.212		0.763		-0.017
	0.2	0.982		0.954		0.494		0.016
	0.3	0.476		0.529		0.239		0.013
	0.4	0.121		0.146		0.062		0.003
ILINK:	0.001	2.000	0.079	1.118	0.003	0.909	0.238	0.017
Broad	0	2.249		1.335		0.996		0.268
	0.1	1.721		1.360		0.924		0.279
	0.2	1.124		1.025		0.625		0.195
	0.3	0.563		0.564		0.312		0.099
	0.4	0.150		0.162		0.083		0.026
ILINK:	0.003	2.246	0.050	1.415	0.030	1.025	0.050	0.297

Table 29-11 : Results of affecteds-only linkage analysis of schizophrenia pedigrees based on the penetrance models in table 29-9.

The analysis parameters for the diagnostic uncertainty model are presented in table 29-12, and the output from that linkage analysis is given in table 29-13. In this case, our maximum lod score dropped to 2.054 (a loss of 0.883), while the number of models was reduced by two, causing our critical value to drop to 5 + $\log_{10}(2)$ = 5.3, which is a drop of 0.4 units. While more information was lost in this example from the combining of the diagnostic criteria, there is still a great reduction in computing time, and roughly the same degree of significance in the results.

		Dominant			Recessive		
Diagnostic Class	p_i	f	f_p	k	f	f_p	k
1	0.99	.333	.0135	.0405	.418	.0135	.0323
2	0.80	.398	.2022	.5080	.450	.2022	.4493
3	0.65	.449	.3511	.7820	.475	.3511	.7392

Table 29-12 : Penetrance models for dominant and recessive analysis with certainty of diagnosis parameters, p_i, from exercise 25.

		Dominant Lod Scores				Recessive Lod Scores		
	θ	Marker 1		Marker 2		Marker 1		Marker 2
	0	2.057		0.948		1.338		-0.950
	0.1	1.606		1.327		1.217		-0.433
	0.2	1.051		1.088		0.823		-0.185
	0.3	0.518		0.629		0.406		-0.068
	0.4	0.133		0.185		0.106		-0.015
ILINK:	0.001	2.054	0.093	1.328	0.025	1.360	0.500	0.000

Table 29-13 : Results of linkage analysis of schizophrenia pedigrees with penetrance models outlined in table 29-12.

EXERCISE 27

When you run the linkage analysis on these fully penetrant dominant disease pedigrees, you should get the lod scores shown in table 29-16. The best estimate of θ in the entire family set together is obviously $\hat{\theta}$ = 0.5, yet it appears that in families 1 and 3 there are no recombinants, which might be indicative of some heterogeneity. There is the problem, however, that we have no evidence for linkage, so we must try to use the techniques outlined in section 27.3, and test for linkage and heterogeneity jointly (since proving one exists obligates the other to exist).

	Lod Scores				
θ	Family 1	Family 2	Family 3	Family 4	Total
0	2.41	$-\infty$	3.01	$-\infty$	$-\infty$
0.05	2.23	-8.00	2.79	-10.00	-12.98
0.10	2.04	-5.59	2.55	-6.99	-7.99
0.15	1.84	-4.18	2.30	-5.23	-5.26
0.20	1.63	-3.18	2.04	-3.98	-3.49
0.25	1.41	-2.41	1.76	-3.01	-2.25
0.30	1.17	-1.77	1.46	-2.22	-1.36
0.35	0.91	-1.24	1.14	-1.55	-0.74
0.40	0.63	-0.78	0.79	-0.97	-0.32
0.45	0.33	-0.37	0.41	-0.46	-0.08

Table 29-16: Results, by family, of linkage analysis with pedigrees in Figure 27-1.

After running these lod scores through the HOMOG program (assuming STEPSIZE = 0.05, and ALOW = 0), the output should resemble the following:

```
                                            Estimates of
Hypotheses                    Max.lnL     Alpha        Theta
H2: Linkage, heterogeneity    9.7077      0.5000        0.0000
H1: Linkage, homogeneity      0.0000       (1)         99.0000
H0: No linkage                  (0)        (0)         (0.5)

Source                  df  Chi-square          L Ratio
H2 vs. H1 Heterogeneity  1    19.415             16440
H1 vs. H0 Linkage        1     0.000                 1
H2 vs. H0 Total          2    19.415             16440
```

Since there is absolutely no evidence for linkage (H1), the test of H2 vs. H1 is not meaningful, since the null hypothesis is not a valid null hypothesis. Instead, we must consider the comparison of H2 vs H0 (the correct null hypothesis of no linkage). If you remember, we suggested that this likelihood ratio test must exceed 2000 for there to be significant evidence for linkage and heterogeneity. In this case, we have a likelihood ratio of 16520, which is highly significant evidence for linkage and heterogeneity, so in this case, while there is no evidence for linkage whatsoever under the assumption of locus homogeneity, as soon as we allow for heterogeneity, we have significant evidence for linkage (in at least some of the families), with $Z(\hat{\theta},\hat{\alpha}) > 4$.

EXERCISE 28

For the first problem, we simulated 50 replicates of the pedigree set using a seed of 27801 for the random number generator. If you used a different seed, or a different number of replicates, your results will be somewhat different, but similar in their interpretation, if everything was done correctly. The results of the analyses are presented in table 29-17.

	Expected Lod Scores	
	Simulated Recessive	Simulated Dominant
θ	Analyzed Dominant	Analyzed Recessive
0	0.172	-1.704
0.1	0.436	-0.119
0.2	0.374	0.245
0.3	0.220	0.239
0.4	0.066	0.092
EMLOD	0.740	0.466
$P(Z_{max} > 3)$	0.02	0.000

Table 29-17: Expected lod scores in the schizophrenia pedigrees when simulated under one model, and analyzed under the wrong model.

As was explained earlier, in part I of the book, if you analyze something as a dominant disease, when it is truly a recessive disease, you are basically throwing away information, yet you should still see positive lod scores. This is demonstrated in this instance, where the EMLOD is about 0.74, and the probability of getting a lod score over three is about 0.02, which is not bad, considering the model is wrong, and roughly half of the meioses are being thrown away by treating the recessive condition as if it were dominant. On the other hand, when something is really dominant, and mis-analyzed as a recessive condition, then you are going to upwardly bias your recombination fraction estimates dramatically. In this case, while one meiosis to each affected offspring truly does contain the disease allele, the other one doesn't, so from one parent, the disease and marker would appear to be linked (in this case, at θ = 0.05), and from the other parent, they would appear to segregate independently (since there really is no disease allele there). So, the overall recombination fraction estimate should be somewhere around the average of 0.05 and 0.5, which is 0.275. In this example, the MELOD occurs somewhere between 0.2 and 0.3, consistent with this theoretical prediction. Thus, you can see the importance of having an accurate model, and also why it is that analyzing under a dominant model is typically more robust than a recessive model, when it is actually an incorrect model.

θ	θ_0 = 0.5	θ_0 = 0.05
0	-∞	-∞
0.1	0.894	1.013
0.2	0.541	0.594
0.3	0.246	0.265
0.4	0.061	0.065
EMLOD	1.370	1.517
P(Z_{max} > 1)	0.800	0.900

Table 29-18: Expected lod scores for the pedigree in figure 28-3, when simulated assuming a marker with one common allele, and four rare ones, and analyzed under the assumption of five equally frequent alleles, simulated both under absence of linkage (θ_0 = 0.5), and under tight linkage (θ_0 = 0.05).

For the example with the SIMULATE program, the results are presented in table 29-18, given seeds for the random number generator of 27801, 29721, and 24562. In this case, you can see that even though the disease and marker are truly unlinked, by using equal gene frequencies for the marker alleles, when this is not the true state of nature, you get positive **expected** lod scores, with an EMLOD of 1.37! This, you remember is in a pedigree where the marker is actually **unlinked** to the disease. In this example, you can see just how important it is to have good gene frequency estimates for your markers, since the false positive rate can easily lead an investigator chasing a lot of wild gooses unnecessarily. Some people have remarked that they don't mind a few false positives, if it means they will have increased power to detect a true linkage. The problem is that in these situations, the expected lod scores are not much different with or without linkage. To illustrate this, try simulating 50 replicates of this pedigree with the SLINK program, at θ = 0.05 between disease and marker (again with frequencies 0.05, 0.05, 0.05, 0.05, and 0.20 for the five marker alleles), and analyzing them assuming equal gene frequencies at the marker locus. The results of this should be approximately the same as those shown in table 29-18, in the right-hand column. As you can see, the expected lod scores are almost identical, whether or not there really is linkage, when the analysis is done under such an incorrect model for the marker allele frequencies. The expected lod scores are thus essentially independent of the true recombination fraction, when these data are analyzed assuming equal gene frequencies. Further, if one analyzes the simulated replicates under the correct model, the expected lod scores are naturally lower (the MELOD for the unlinked replicates is 0.00), yet the power is somewhat increased, as now there is a 6% chance that Z_{max} > 2, where it was 0 when the data were analyzed under the equal gene frequencies model. All of this together should dissuade people from using erroneous gene frequency information, and it should point out the importance of getting accurate estimates of the gene frequencies, in terms of saving much time and effort tracking down false positive findings.

Appendix A: The Linkage Utility Programs

The Linkage Utility Programs (http://www.jurgott.org/linkage/util.htm) are a collection of small programs that often prove useful in a linkage analyst's everyday life. For example, there is a program (CHIPROB) that computes the *p*-value associated with an observed chi-square with *n* degrees of freedom (df), NORPROB carries out analogous calculations for the standard normal distribution, and NORINV does the reverse, i.e. computes standard normal deviates from given p-values, etc. Some of these programs have already been used in this book. Please read through the documentation of these programs; you may find useful hints. Below, a few selected programs will be applied to examples one might find in practical applications. These programs make use of published formulas [29].

A.1 The BINOM program

Conventionally, linkage is declared significant when the lod score attains or exceeds the value 3, which is associated with a p-value of at most 0.001. If your data consists of counts of recombinants and nonrecombinants you may compute the p-value directly and declare linkage significant if p ≤ 0.001. The p-value is defined as the probability, given the null hypothesis (of no linkage, in this case), of finding a maximum lod score as large or larger than the one actually observed.

Assume that in an experiment, k = 4 recombinants are observed in a total of n = 20 opportunities for recombination (phase-known meioses). Does this result represent significant evidence for linkage? What about k = 2 recombinants in n = 20 meioses? To find out, we compute the p-value as the probability, given a (true) recombination fraction of r = ½, of observing four or fewer recombinants because decreasing the number of recombinants leads to an increase in the maximum lod score. Call up the BINOM program and choose the *binomial probabilities* option. Then, select n = 20 and p = 0.5 and have the program calculate the binomial probabilities of k = 0...4 (choose $k_1 = 0$ and $k_2 = 4$). What do you get? Repeat this analysis assuming k = 2 recombinants were observed. Which of the two outcomes is significant? You should find that for k = 4, the p-value is equal to 0.006 (not significant), and for k = 2 it is 0.0002 (significant).

A result of k = 4 or k = 2 recombinants in 20 phase known meioses leads to recombination fraction estimates of $\hat{\theta}=0.20$ and $\hat{\theta}=0.10$, respectively. To assess the accuracy of such point estimates one constructs support intervals or confidence intervals for the (true) parameter for which an estimate was obtained. Support intervals are generally easy to obtain (see below for an example). However, calculating a confidence interval can be difficult. One usually works with the normal approximation to the binomial distribution but this has the disadvantage that the confidence intervals are forced to be symmetric about the estimate of the recombination fraction (the lower bound may become negative). Using the BINOM program we can calculate proper confidence intervals (it does the calculations numerically using an iterative procedure to solve the relevant equations). For further explanations on confidence intervals, please consult section 3.6 in Ott (1991). Before proceeding, take a guess at the 95% confidence intervals for *r* based on k = 4 and k = 2. What are its lower and upper bounds?

For each of the two observations, k = 4 and k = 2, calculate the 95% confidence interval for *r*. Call up the BINOM program, choose the confidence interval option and, for each of k = 4 and k = 2, have it compute the two-sided confidence interval using 0.025 each for the lower and upper error probabilities. Write down the resulting 95% confidence intervals with three decimal places. You should find [0.057, 0.437] for k = 4 and [0.012, 0.317] for k = 2.

It is unclear which is more appropriate, support or confidence intervals. Statisticians are divided about this issue, and different schools of thought have different preferences. In linkage analysis, it is customary to compute support intervals rather than confidence intervals for *r*. The support interval often applied is the 1-lod-unit support interval, which is constructed by finding the maximum lod score, Z_{max}, and determining those points of θ for which the lod score is at least $Z_{max\,–}$ 1. these θ points form the desired support interval. As outlined in section 1.3.5, however, 3-lod-unit support intervals are more appropriate in linkage analysis. Compute the 3-lod-unit support interval for the observation k = 2 recombinants in n = 20 meioses. One way to achieve this task is to use a spreadsheet program, fill one column with values of θ from 0.001 through 0.5 in steps of 0.001, and fill another column with the formula for the lod score corresponding to k = 2 and n = 20. Then, simply find the maximum lod score, subtract 3 from it, and find the θ values with associated lod scores closest to $Z_{max\,–}$ 3. You may also proceed as follows. Create a family with 20 offspring, two parents, and the parents of one of the parents. Of the grandparents, one has genotype 1/1 at each of two loci, and the other is 2/2 at each locus. Their offspring (a parent) is 1/2 at each locus, and this parent's mate is 1/1 at each locus. Then, one assumes two types of offspring, 1 2/1 1 (recombinants) and 1 1/1 1 (nonrecombinants). This way one can create a family with exactly 2 recombinants and 18 nonrecombinants. Now, run the MLINK program and have it compute lod scores at θ between 0 and 0.40 in steps of 0.001. You should find a 3-lod-unit support interval of [0.002, 0.485], which is similar to the 99.9% confidence interval found above.

A.2 The PIC and HET programs

Genetic marker loci may be more or less polymorphic depending on the number of alleles and their population frequencies. The degree of polymorphism of a marker may be assessed by the proportion of individuals in the population who are heterozygous for that marker. In other words, the probability that a random individual is heterozygous is used as a measure of the degree of polymorphism. This probability may be estimated in two principal ways, each based on a random sample of unrelated individuals.

The first measure is the amount of heterozygosity observed, $\hat{h}$, and is simply the proportion of heterozygous individuals observed in the sample (Weir, 1990). It is an unbiased estimate of the proportion h of heterozygous individuals in the population.

In human genetics, a more precise estimate is usually used. It rests on the assumption that the genotypes are in Hardy-Weinberg equilibrium (HWE), which is the reason for its increased precision. This expected heterozygosity (or, in human genetics, just heterozygosity) is defined as $H = 1 - \Sigma p_i^2$, where the sum is taken over all alleles, with p_i denoting the frequency of the i-th allele. The maximum likelihood estimate of H is given by $\hat{H}_M = 1 - \Sigma(\hat{p}_i)^2$ and is slightly biased. An unbiased estimate is $\hat{H}_U = \hat{H}_M n/(n-1)$, where n is the number of alleles observed in a sample (Ott, 1992). $\hat{H}_U$ is preferable over $\hat{H}_M$ because it is unbiased and has smaller mean squared error than $\hat{H}_M$.

An older measure of heterozygosity is the PIC value (Botstein et al. 1980), which is defined as

$$PIC = 1 - \sum_{i=1}^{a} p_i^2 - \sum_{i=1}^{a-1} \sum_{j=i+1}^{a} 2\, p_i^2\, p_j^2 .$$

where a is the number of alleles at the given locus. In the PIC value, a quantity is subtracted from the heterozygosity that corresponds to the probability that offspring are uninformative, because if both parents are identically heterozygous, on average, half of their children (the homozygotes) will be informative and half (the heterozygotes) will be uninformative. For family data, PIC may be somewhat more appropriate, whereas the heterozygosity is more general. The maximum likelihood estimate of the PIC value is obtained by replacing the gene frequencies by their estimates.

Genotype	1/1	1/2	1/3	2/2	2/3	3/3
No. of individuals	2	23	2	13	9	1

As an application, assume that you find a new DNA polymorphism. You type 50 unrelated individuals to estimate the number of alleles, their population frequencies, and the heterozygosity of the system. The genotyping results are summarized in the table above. We easily estimate the observed heterozygosity (h) by the proportion of heterozygous individuals, that is, as $\hat{h} = (23 + 2 + 9)/50 = 0.68$. Also, we easily find the 95% confidence interval for the proportion of heterozygous individuals in the population. Call up the BINOM program and select the confidence intervals option. Enter 34 (for k) and 50 (for n) and choose 0.025 each for the lower and upper error probabilities. You should find a confidence interval of [0.533, 0.805].

The expected heterozygosity (H) is more difficult to calculate by hand. Thus we make use of the PIC program to compute it. In addition to the biased (maximum likelihood) estimate, it will also furnish the unbiased heterozygosity estimate and the PIC value. Since these quantities are based on allele counts rather than genotype counts, we first make a list of the different alleles and how often they occur. For our alleles 1, 2, and 3, we find 29, 58, and 13 copies in the 50 individuals, respectively (the number of alleles must sum to 2×50 = 100). Now call up the PIC program and choose the Count alleles option. You should find the following estimates: $\hat{H}_M$ = 0.5626 (maximum likelihood, biased), $\hat{H}_U$ = 0.5683 (unbiased), and PIC = 0.4918. These values happen to be quite a bit lower than $\hat{h}$=0.68.

As the observed proportion of heterozygous individuals is an estimate for the corresponding proportion in the total population, so are the expected heterozygosities, $\hat{H}_U$ and $\hat{H}_M$, estimates for the population value, *H*. A support interval for *H* can be found with the help of the HET program (Shugart and Ott 1992). HET works in terms of *m*-unit support intervals, where *m* refers to the number of units of natural log likelihood. Thus, *m* = 2 corresponds to a likelihood ratio of 7.4. Asymptotically, it is equivalent to a 95% confidence interval.

Call up the HET program and follow the directions to calculate a 2-unit support interval for *H*. You should find [0.492, 0.622], which is less than half as long as the 95% confidence interval for *h* computed above. This difference clearly demonstrates the gain in precision when one can rely on HWE, which is generally reliable for a stable population.

A.3 The CHIPROB program and testing HWE

In the previous section, the observed heterozygosity of 0.68 deviates quite a bit from the value of 0.57 expected under HWE, and we wonder whether the assumption of HWE might not be violated here. To investigate this, we carry out a test of HWE, that is, we test whether the genotype frequencies are compatible with HWE. Before proceeding, we take a hint from the confidence interval for the population heterozygosity - it includes the point estimate under the assumption of HWE, $\hat{H}_U$ = 0.57, so we suspect that the genotype frequencies will be compatible with HWE.

We test the assumption of HWE for the observed genotype frequencies given in section A.2 using a chi-square test. We first calculate expected genotype frequencies assuming HWE and do this on the basis of the allele frequencies already obtained, that is, with $\hat{p}_1$=0.29, $\hat{p}_2$=0.58, and $\hat{p}_3$=0.13. Our expected genotype frequencies are then estimated, for example, as P(1/1) = $(\hat{p}_1)^2$ and P(1/2) = 2 $\hat{p}_1 \hat{p}_2$. These estimates are biased. Unbiased or less biased genotype frequency estimates have been derived but are not generally used in practice. For our chi-square test, we will use the (biased) maximum likelihood estimates. We multiply each of the expected genotype frequencies by 50 to obtain the expected number of individuals with the respective genotypes. Please carry out these calculations for all genotypes. You should find, in the order of genotypes given in the table in section A.2, 4.205, 16.82, 3.77, 16.82, 7.54, and 0.845. As a check, the sum of these figures should be equal to 50. Then, for each of the genotypes we calculate its contribution to chi-square in the usual manner as $(O – E)^2/E$ where O stands for the observed number and E for the expected number of individuals. For example, the contribution from the 1/1 genotype is $(2 – 4.205)^2/4.205$ = 1.156. Please calculate all six contributions and sum them up. You should obtain a chi-square value of 5.44. The number of df associated with this chi-square is given by 6 – 1 – 2 = 3, where 6 is the number of classes in which observed and expected number of observations are contrasted. We subtract 1 from the number of classes because the total number of observations is fixed and the numbers in the sixth class are given once we know the numbers in the first five classes. We estimated three gene frequencies but only two of them represent independent estimates. the third is again given once we know the first two. Thus, we subtract 2 to arrive at a number of 3 df.

Is the chi-square of 5.44 on 3 df significant? Instead of looking up critical values for chi-square in a table, we calculate the empirical significance level *p* associated with this result, that is, the probability that, under the null hypothesis (HWE in this case), the observed chi-square value is exceeded by chance, and declare the result significant if $p \leq 0.05$, and highly significant if $p \leq 0.01$. Call up the CHIPROB program and simply enter the two values, 5.44 and 3. You'll quickly find that *p* = 0.14, which is larger than 0.05, so there is no significant evidence for a deviation from HWE.

In addition to the application of CHIPROB shown above, this program also allows combining p-values from different independent investigations into one overall p-value, where the individual p-values may result from any statistical test furnishing a p-value. The approach, based on a method by R.A. Fisher [30], specifies that one should transform each value of *p*, which has a uniform distribution under the null hypothesis, into $c = –2 \times \text{LN}(p)$, which has a chi-square distribution on 2 df. Assume that *n* independent p-values should be combined. The corresponding *n* c-values are then added together. Their sum, Σ(c), represents a chi-squared variable with 2*n* df. So, if Σ(c) is entered in the CHIPROB program with 2*n* df, the p-value returned is the desired overall empirical significance level. As an example, assume that three independent tests (not necessarily chi-square tests) have furnished the respective p-values 0.011, 0.047 and 0.35. The corresponding c-values are 9.02, 6.12 and 2.10. Their sum, 17.24, with 6 df, yields a combined p-value of 0.008. This procedure is implemented in the PVALUES program.

Appendix B: Practical considerations

B.1 Overview of linkage programs

Below is the original text of this book. For an up to date list of programs, see our review on linkage in the age of sequencing [31].

Linkage programs may be divided into two groups. The major programs belonging to the first group are LIPED (Ott, 1974), PAP (Hasstedt and Cartwright, 1981), LINKAGE (MLINK, LINKMAP, ILINK, etc. (Lathrop et al., 1984), MENDEL (Lange et al., 1988) [20], and GRONLOD (te Meerman, 1993) (see section B.4 and appendix C for ordering information). They are able to carry out linkage analyses for families of an arbitrary structure, with possibly incomplete penetrance and other complicating factors. The second group of programs, comprising MAPMAKER (Lander et al., 1987), CRI-MAP (P. Green, personal communication), and the CEPH version of LINKAGE (CLINKAGE), is applicable only to special loci and/or pedigree types. For example, a version of MAPMAKER and the CEPH version of LINKAGE work with codominant markers in three-generation pedigrees with a nuclear family and up to four grandparents (figure 13-1). The CRI-MAP program originally was also written for this type of application but has been extended to some other pedigree structures and loci. Also, other programs (for example, special versions of MAPMAKER and LINKAGE) work with quantitative traits observed on experimental crosses. MAPMAKER is very user friendly and has many options for automated linkage analysis and map building.

The LINKAGE package, like most of the other linkage analysis programs, are available for a variety of computer systems. They were developed by Mark Lathrop with contributions by Jean-Marc Lalouel, Cécile Julier, and Jurg Ott. Peter Cartwright has made major contributions to the development of the shell programs, which have greatly increased the usefulness of LINKAGE. Mark Lathrop regularly updates these programs.

The LINKAGE and MENDEL programs are very similar in their focus. They both handle pedigrees of arbitrary structure and various phenotypes. MENDEL is more flexible in the problems it can address, for example, the user can impose linear constraints on the parameters to be estimated, which is not possible in ILINK. On the other hand, MENDEL is more demanding both of the user (it requires fluency in FORTRAN) and of computer resources (it requires more memory than LINKAGE for the same problem). On the PC many linkage problems that cannot be handled by MENDEL can successfully be carried out by the LINKAGE programs.

A special version of the LINKAGE programs, TLINKAGE [32-34], allows for two loci jointly leading to disease (Lathrop and Ott, 1990). This possibility has also been incorporated in the MENDEL program (Schork et al., 1993).

GRONLOD is the newest member of these programs. It has been written in Prolog and takes advantage of Prolog's possibilities to represent abstract objects and work with dependencies among them.

B.2 Database and pedigree drawing programs

How should pedigree data be kept in a computer? Many people simply keep it in files (in ASCII or other format) with rows corresponding to individuals and columns corresponding to phenotypes at different loci. In addition to the phenotypes, there are columns containing the sex of an individual and pointers to the two parents (analogous to a LINKAGE pedigree file). This approach is probably easiest and sufficient if the amount of data is small enough to be manageable in this form.

Another method is to enter the data into a database using one of the commercially available database programs such as dBASE, Foxbase, Paradox, etc. Databases have various advantages. For example, some columns can be singled out for special consideration, or a printout of the database may be made suppressing some columns containing sensitive data. On the other hand, the user must learn commands specific for that database. Specialized databases have been developed for family pedigree data. They are typically capable of writing output in a format suitable for analysis by a linkage program or of graphically displaying the pedigree. Nowadays, many people keep linkage and association data in PLINK format [35], which actually is the LINKAGE format, which in turn is the LIPED format.

B.3 Sources of information

Nowadays, most information for scientists is online. For example, for linkage analysis, see http://lab.rockefeller.edu/ott/, which contains links to other sources of information. Also, the “McKusick catalog,” or *Mendelian Inheritance of Man* (McKusick, 1990) contains information on loci with an established mode of onheritance.

The section number B.4 contained information on how to obtain programs by mail, which is now obsolete.

B.5 PC hardware and operating systems

Some of the newer program versions run only on Linux. I have been using Ubuntu Linux for a while. It is easy to use and provides a large number of add-on programs. My personal favorite is Kubuntu, but I also had good experiences with openSUSE. To log into a Linux server remotely, putty (SSH) and filezilla are very useful. Direct access to a client’s desktop is possible with TeamViewer.

B.6 Program constants and recompiling the LINKAGE programs

In the Pascal version of LINKAGE (the version considered in this book), array bounds such as maximum number of alleles are given as program constants. They may be changed and set to a user's needs. Once such constants are changed, the programs need to be recompiled for the new constant values to take effect. We first discuss the most important constants for the general analysis programs. Deviating constant definitions for CEPH and other programs will be noted. Then we outline the steps necessary to recompile the programs. The Users' Guide to Analysis Programs, which comes with the LINKAGE programs, also contains explanations on these constants. Ideally, all constants are set to high values such that no recompiling of the programs is required. However, to keep the size of the program down to a manageable level, the program constants should be set to small values whenever possible. My favorite Pascal compiler is FreePascal, which seems to be derived from Turbo Pascal. It comes with extensive documentation.

MAXNEED sets the upper bound to an array containing various recombination probabilities. Its value depends only on the number of loci. For locus numbers from 2 through 8, the minimum values of MAXNEED are 7, 32, 157, 782, 3907, 19532, and 97657 respectively. If the array size required by the program is larger than MAXNEED, the program will terminate with an error. If MAXNEED is larger than necessary, the program will write on the screen the minimum value of MAXNEED it requires in this run.

MAXCENSOR determines the length of an array, which holds intermediate results for individuals. In a given run, there is an optimal value for MAXCENSOR. If MAXCENSOR is smaller than this optimal value, the program will run less efficiently and will print on the screen that it would benefit from an increase in MAXCENSOR. If MAXCENSOR is larger than the optimal value, the program will not run faster than with the optimal value and will print on the screen the optimal value for MAXCENSOR.

MAXLOCUS simply determines the maximum number of loci allowed. This and all other constants below represent upper limits. If these upper limits are insufficient for a particular run, the program will stop with an error message.

MAXSEG should be set equal to 2 to the power of (maxlocus-1). For example, MAXLOCUS=5 calls for MAXSEG=16.

MAXALL specifies the maximum number of alleles at a single locus.

MAXHAP is the maximum number of multilocus haplotypes the program can handle. To be safe, set MAXHAP to a value as large as the product of the number of alleles at all loci used in a particular run. In Pascal, MAXHAP must not be larger than 126 or else the program cannot be compiled.

MAXFEM is the maximum number of female multilocus genotypes the program can handle, and MAXMAL is the analogous quantity for males. These two values should be set to a value at least as large as MAXHAP·(MAXHAP+1)/2. Larger values are wasteful of memory space. Some Pascals allow setting constants as simple functions of other constants. In Free Pascal, for example, the user need not set MAXFEM and MAXMAL because they are determined by the constant MAXHAP as given above. Note that in Free Pascal, the product MAXFEM·MAXPED must not be larger than 65,536, where MAXPED = maximum number of pedigrees. For example, with MAXHAP=64 (MAXFEM=2080), no more than MAXPED=31 pedigrees can be analyzed in a single run.

MAXIND is the maximum number of individuals in all pedigrees combined.

AFFALL is an integer number indicating which allele is the disease allele. In most applications, a low allele frequency together with the penetrances imply which allele is the disease allele. In these cases the value of AFFALL is irrelevant. At sex-linked quantitative trait loci, the phenotype in males is assumed to be of a fully penetrant affection status type and no penetrances must be specified. In this case, AFFALL is used to identify the disease allele. Also, for rare dominant diseases, homozygotes may be disregarded in the analysis (see MINFREQ below). In this case, too, AFFALL is used to identify the disease allele.

MINFREQ represents a gene frequency limit for the AFFALL allele in the following sense: if the specified frequency of the AFFALL allele is smaller than MINFREQ, then homozygotes for the AFFALL allele will not be considered. This is only meaningful for rare dominant diseases where AFFALL is the disease allele. In practice, one usually has MINFREQ=0. Homozygotes may be eliminated from consideration with the use of liability classes.

MAXTRAIT is the maximum number of (quantitative) variables at a quantitative trait locus type. In most applications this is equal to 1 (univariate phenotype). Some compilers require MAXTRAIT to be at least 2.

MAXFACT is the maximum number of binary codes at a locus. MAXFACT must be at least as large as MAXALL.

SCALE and SCALEMULT are used to increase the likelihood such that it does not grow to too small a number thereby causing an underflow. If problems with underflow or overflow occur, SCALE and SCALEMULT should be changed. The best protection against underflow and overflow, however, is to work with double precision instead of single precision variables.

FITMODEL is either true or false. If true, the program will calculate likelihoods whether or not a family is informative. If FITMODEL=false, uninformative families are skipped.

DOSTREAM should be set to true if the locus report program, LRP, is to be used to interpret program output.

BYFAMILY is either true or false. If set to true, likelihoods (not lod scores) will be output for each family, otherwise no individual family results will be provided.

NBIT indicates the precision of real variables and should be set to a value as large as the mantissa length of real variables. For example, it is equal to 23 for single precision and 52 for double precision variables. In some versions of LINKAGE, a procedure "precision" is furnished, which calculates NBIT as a variable such that the user need not set it as a constant.

MAXN is the maximum number of parameters that ILINK can iteratively estimate for that locus identified at the bottom of the datafile by "this locus may have iterated parameters." This number of parameters includes the penetrances in all liability classes.

In the CEPH programs, MAXSYSTEM, MAXIND, and MAXPED refer to the respective maximum number of loci, individuals, and pedigrees after the transformation step. These values may be considerably larger than the corresponding values in the pedigree data and cannot be determined before the CFACTOR program has completed. CFACTOR produces two output files, TEMPDAT.DAT and TEMPPED.DAT, which contain the numbers of loci, individuals, and pedigrees to which the three constant above refer.

The Pascal version of the LINKAGE program has been compiled with Free Pascal, which is available for Windows and Linux.

B.7 How to set up a linkage study with sequence data

B.7.1 Disease phenotype

Assume that you want to investigate the genetics of a particular disease (yes-no trait, affected and unaffected individuals). One of the first questions is how much of the variability of the phenotype is due to genes as opposed to effects of common family environment or random environment. Generally, the fact that a disease "runs in families" is taken as evidence for a genetic component, and it usually is. But well-known exceptions exist. For example, Kuru (a disease earlier quite prevalent in Papua-New Guinea) was once thought to be due to a dominant gene but was later found to be transmitted by a virus through cannibalism (Lindenbaum, 1979).

Classical methods of addressing the question of genetic involvement are comparing the concordance rate among monozygotic siblings (who share 100% of their genes) with the concordance rate among dizygotic twins (who on average share 50% of their genes). To avoid the confounding effects of common family environment, siblings reared apart are often also investigated. However, such analyses only show genetic effects but these may be due to one or several genes. Specialized analyses can help discriminate between these possibilities. For example, the major gene statistic of Jayakar et al. (1984) is designed to be sensitive to the presence of single major genes influencing quantitative characters. It has been applied to obesity and gave weak evidence for the presence of a major gene influencing body weight (Zonta et al., 1987).

A method often used to dissect effects of major genes from other effects is complex segregation analysis (Morton and MacLean 1974; Bonney et al, 1988). Also, the comparison of recurrence risks among different types of relatives (Risch, 1990b) appears to be a powerful tool for detecting major genes.

Segregation analysis typically is very sensitive to ascertainment, and false assumptions on ascertainment may easily invalidate a segregation analysis (Greenberg, 1986). Linkage analysis, on the other hand, does not suffer from this problem, that is, one may select family members on the basis of the phenotypes at one locus in any way one wants to and the linkage analysis will still be valid and will furnish unbiased estimates of the recombination fraction. For this reason, many investigators skip formal segregation analysis and make a small number of "reasonable" assumptions on the mode of inheritance of the disease to be investigated. The problems addressed here are beyond the scope of this book. You may also want to consult Ott (1999).

If multiple genes jointly have an effect on a disease, several investigations have shown that in many cases analysis under a single-gene model retains much of the linkage information of the multilocus situation but the recombination fraction tends to be biased upwards. This implies that in testing a disease versus a map of markers, a complex trait tends to be localized outside of the map even if in reality a gene exists inside the map. Therefore, for complex traits, one should stick to twopoint analysis (disease versus one marker at a time).

Now, we will assume the presence of a single gene responsible for a disease. The first task is to find a number of suitable families willing to collaborate and donate small amounts of blood for marker typing. Ideally, there should be several affected individuals in one sibship. For dominant diseases, extended families are more useful than nuclear families. Also, parents should be available for marker typing whenever possible. Once a set of families has been ascertained and affection status has been established at least in a preliminary manner, before marker typing has even begun, you may want to estimate the expected lod score or the power for detecting linkage with the given family data. This is usually done by computer simulation under the assumption that it will be possible to find a marker close to the disease gene, say, at a distance of 2-5 cM. The simulation may also be carried out with two flanking markers but that takes much more computer time. Therefore, one usually works with a single “virtual” marker tightly linked (1 cM) with the disease gene.

At the true (simulated) recombination fraction, *r*, an expected lod score of, say, 3 means that on average the lod score you will find at $\theta = r$ will be equal to 3. Approximately, there is a 50% chance that the lod score in your study will be equal to 3 or higher. Usually, a power larger than 50% is desired. Of course, these simulations are reliable only when carried out with reasonable parameter values. For example, if penetrance for the disease at high age is 50% but you assume 80% in the simulation, the simulation will indicate more power than is available in the data, and the results of the linkage analysis tend to be disappointing. Also, at least in complex traits, be sure to allow for phenocopies (nongenetic cases) and make their penetrance age dependent if the penetrance of genetic cases is also age dependent.

B.7.2 Large numbers of markers and sequence data

Genotyping with SNP chips is now relatively inexpensive, so this is a good way to start a linkage study. You may also do exome or whole genome sequencing on family members and extract SNPs from sequence data but this is a more expensive undertaking.

With several 100,000 SNPs, the most important task is to verify family relationships. It is good practice to run a suitable program, for example, *plink* version 1.9 (plink2) with the `--mendel` parameter, which will furnish, for example, numbers of mendel errors by individual. The program will not detect all mendel errors; this is only possible by actually performing likelihood calculations, but such a *plink* run will be very informative. It may well show that one or two individuals stand out as causing unusually large numbers of errors. This may mean that such individuals do not belong into the pedigree and their marker genotypes should then preferably all be set equal to unknown.

A more precise way to scutinize family relationships is to compute pairwise IBD relationships, for example, with the `--genome` option in the *plink* program.

Appendix C: List of programs, and where to obtain them

The following list is not intended to be exhaustive, but rather to provide an overview of some of the programs available for human genetic analysis. The programs listed are mostly as in the original book version (and some may no longer exist), augmented with some newer programs. For a detailed list of programs see http://lab.rockefeller.edu/ott/geneticsoftware. See also our upcoming review article (Ott & Leal, submitted).

Segregation Analysis

PAP (Hasstedt and Cartwright, 1981)
POINTER (Lalouel and Yee, 1980)
REGRESS (Bonney et al, 1988)
SAGE (Elston et al, 1986)

Database programs

CYRILLIC (Chapman, 1990)
dbLINK (Sarfarazi, 1990)
dGENE (Lange et al, 1988)
KINDRED - Epicenter Software
LABMAN/LINKMAN (Adams et al, 1990)
LIPIN (Trofatter et al, 1986)
Megabase (Fenton et al, 1990)
MEGADATS (Gersting, 1987)
PEDSYS (Dyke and Mamelka, 1987)

Pedigree Drawing Programs

FTREE - Rodney Go
GENETREE - Ellen Wijsman
KINDRED - Epicenter Software
PEDIGREE/DRAW (Dyke and Mamelka, 1987)
PEDRAW (Curtis, 1990)
PLOT2000 (Wolak and Sarfarazi, 1987)
SCHESIS (Round et al, 1990)

Linkage Analysis

CINTMAX (Weeks et al, 1991)
CRI-MAP (Lander and Green, 1987)
EXCLUDE (Edwards, 1987)
GRONLOD (te Meerman, 1993)
LINKAGE (Lathrop et al, 1984) - (See section B.4)
LIPED (Ott, 1974)
MAPMAKER (Lander et al, 1987)
MAP90 (Morton and Andrews, 1989)
MDMAP (Falk, 1991)
MENDEL [20]

PAP (Hasstedt and Cartwright, 1981)
PROGRAMS FOR PEDIGREE ANALYSIS (MENDEL/FISHER/SEARCH) [20]
PSEUDOMARKER [24, 25]
RHMAP (Boehnke et al, 1992)
SAGE (Elston et al, 1986)
SCHESIS (Round et al, 1990)
TLINKAGE [32-34] See manual

Simulation Programs

CHRSIM (Speer et al, 1992)
MOM (Ott and Terwilliger, 1992)
SIMLINK (Boehnke, 1986)
SIMULATE (Ott and Terwilliger, 1992)
SLINK (Including **MSIM/ISIM/LSIM)** (Weeks et al, 1990)
TYPENEXT (Ott et al, 1992)

Non-Parametric Analysis Programs

APM (Weeks and Lange, 1988)
SAGE (Elston et al, 1986)

Heterogeneity Testing

B-TEST (Risch, 1988)
HOMOG (Ott, 1991)
MTEST (Ott, 1991)
C-GEN (MacLean et al, 1993)

Miscellaneous Programs

EH - [36]
LINKAGE UTILITY PROGRAMS (Ott, 1991)
MULTIMAP (Cox et al, 1992)
Miscellaneous Population Genetics Programs (Weir, 1993) - (Source Code given in Weir, 1993)
PLINK version 1.9, highly updated and very fast, also reads vcf files;
https://www.cog-genomics.org/plink2
SENSEN - Sensitivity Analysis (Hodge and Greenberg, 1992)
VARYPHEN - (Xie et al, 1991)

GNU Free Documentation License

Version 1.3, 3 November 2008

Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.
<http://fsf.org/>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed.

0. PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "publisher" means any person or entity that distributes copies of the Document to the public.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.
- I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
- J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
- K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
- L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
- M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
- N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

- O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give

Original References

Adams, P.B., J.D. Lish, N. Freimer, L.M. Brzustowicz, and V. Vieland. 1990. Pedigree and DNA Marker management: integrated system for molecular and family-genetic studies. Am J Med Genet 46:A171.

Anderson, T. W., and S. L. Sclove. 1986. The statistical analysis of data. 2nd ed. Palo Alto, Calif.: Scientific Press.

Attwood, J., and S. Bryant. 1988. A computer program to make linkage analysis with LIPED and LINKAGE easier to perform and less prone to input errors. Ann Hum Genet 52:259.

Ayala, F. J., and J. A. Kiger. 1984. Modern genetics. Menlo Park, Calif.: Benjamin/Cummings.

Bailey, N. T. J. 1961. Introduction to the mathematical theory of genetic linkage. Oxford: Clarendon Press.

Blackwelder, W. C., and R. C. Elston. 1985. A comparison of sib-pair linkage tests for disease susceptibility loci. Genet. Epidemiol. 2:85-97.

Boehnke, M. 1986. Estimating the power of a proposed linkage study: a practical computer simulation approach. Am. J. Hum. Genet. 39:513-27.

Boehnke, M. 1991. Allele frequency estimation from data on relatives. Am. J. Hum. Genet. 48:22-25.

Boehnke, M. 1992. Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. Cytogenet. Cell Genet. 59:119-121.

Bonney, G.E., G.M. Lathrop, and J.-M. Lalouel. 1988. Combined linkage and segregation analysis using regressive models. Am J. Hum Genet. 43:29-37.

Botstein, D., R.L. White, M.H. Skolnick, and R.W. Davies. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314-31.

Brzustowicz, L.M., C. Mérette, X. Xie, L. Townsend, T.C. Gilliam, and J. Ott. 1993 Detecting marker inconsistencies in human gene mapping. Hum Hered **43**:25-30.

Cavalli-Sforza, L. L., and W. F. Bodmer. 1971. The genetics of human populations. Paperback reprint 1977. San Francisco: Freeman.

Chakravarti, A., C. C. Li, and K. H. Buetow. 1984. Estimation of the marker gene frequency and linkage disequilibrium from conditional marker data. Am. J. Hum. Genet. 36:177-86.

Chance, P.F., T.D. Bird, P. O'Connell, H. Lipe, J.-M. Lalouel, and M. Leppert. 1990. Genetic linkage and heterogeneity in type I Charcot-Marie-Tooth disease (hereditary motor and sensory neuropathy type I). Am. J. Hum. Genet. 47:915-925.

Chapman, C.J. 1990. A visual interface to computer programs for linkage analysis. Am J Med Genet 36:155-160.

Chotai, J. 1984. On the lod score method in linkage analysis. Ann. Hum. Genet. 48:359-378.

Clerget-Darpoux, F., C. Bonaïti-Pellié, and J. Hochez. 1986. Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393-99.

Conneally, P. M., J. H. Edwards, K. K. Kidd, J.-M. Lalouel, N. E. Morton, J. Ott, and R. White. 1985. Report of the committee on methods of linkage analysis and reporting. Cytogenet Cell Genet 40:356-9.

Cottingham, R.W., R.M. Idury, A.A. Schaeffer. 1993. Faster sequential genetic linkage computations. Am. J. Hum. Genet. **53**:252-263.

Cox, T.K., M. Perlin, A. Chakravarti. 1992. Multimap : Automatic construction of linkage maps. Am. J. Hum. Genet. 51:A33

Crow, J. F. 1966. The quality of people: human evolutionary changes. Bioscience 16:863-7.

Curtis, D. 1990. A program to draw pedigrees using LINKAGE or LINKSYS data files. Ann. Hum. Genet. 54:365-367.

Dausset, J., H. Cann, D. Cohen, M. Lathrop, J.-M. Lalouel, and R. White. 1990. Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome. Genomics 6:575-7.

Davies, R.B. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247-54.

Dyke, B., and P. Mamelka. 1987. A computer program that draws pedigrees. Am. J. Hum. Genet. 41:A253.

Edwards, A.W.F. 1992. Likelihood, expanded edition. Baltimore, Maryland: Johns Hopkins University Press.

Edwards, J. H. 1987. Exclusion mapping. J. Med. Genet. 24:539-43.

Elandt-Johnson, R. C. 1971. Probability models and statistical methods in genetics. New York: Wiley.

Elston, R. C. 1993. P-values, power and pitfalls in the linkage analysis of psychiatric disorders. In: Genetic approaches to mental disorders, edited by E.S. Gershon and C.R. Cloninger. American Psychiatric Press, Washington, D.C.

Elston, R.C., J.E. Bailey-Watson, G.E. Bonney, B.J. Keats, and A.F. Wilson. 1986. A package of computer programs to perform statistical analysis for genetic epidemiology. Presented at the Seventh International Congress of Human Genetics, Berlin

Falk, C.T. 1991. A simple method for ordering loci using data from radiation hybrids. Genomics **9**:120-123.

Falk, C. T., and P. Rubinstein. 1987. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann. Hum. Genet. 51:227-33.

Faraway, J.J. 1993. Distribution of the admixture test for the detection of linkage under heterogeneity. Genet. Epid. 10:75-83.

Fenton, I., L.A. Sandkuijl, J.R. Sampson, A. Williams, and J. Myring. 1990. MEGABASE : A pedigree-based computer program for genetic data management which facilitates risk assessment. Proceedings of the Clinical Genetics Society, Newcastle.

Gersting, J.M. 1987. Rapid prototyping of database systems in human genetics data collection. J. Med. Syst. 11:177-189.

Greenberg, D.A. 1986. The effect of proband designation on segregation analysis. Am. J. Hum. Genet. 39:329-339.

Greenberg, D. 1993. Linkage analysis of 'necessary' disease loci versus 'susceptibility' loci. Am. J. of Hum. Genet. **52**:135-143.

Haines, J.L. 1992. CHROMLOOK: An interactive program for error detection and mapping in reference linkage data. Genomics 14:517-519.

Haldane, J.B.S. 1935. Spontaneous mutation of a human genome. J.Genet. 31:317-326.

Hartl, D. L. 1988. A primer of population genetics. Sunderland, Mass.: Sinauer Associates, Inc.

Hasstedt, S. J., and P. E. Cartwright. 1981. PAP - pedigree analysis package, University of Utah, Department of Medical Biophysics and Computing, Technical Report No. 13. Salt Lake City, Utah.

Hsiao, K., H.F. Baker, T.J. Crow, M. Poulter, F. Owen, J.D. Terwilliger, D. Westaway, J. Ott, S.B. Prusiner. 1989. Linkage of a prion protein missense variant to Gerstmann-Straussler syndrome. Nature **338**:342-344.

Jayakar, S.D., J.A. Williamson, and L. Zonta-Sgaramella. 1984. A nonparametric and parametric version of a test for the detection of the presence of a major gene applicable on data for the complete nuclear family. Hum. Genet. 67:143-150.

Karlin, S., and U. Liberman. 1978. Classifications and comparisons of multilocus recombination distributions. Proc. Natl. Acad. Sci. USA 75:6332-36.

Kerem, B., J.A. Buchanan, P. Durie, M.L. Corey, H. Levison, J.M. Rommens, M. Buchwald, and L.-C. Tsui. 1989. DNA Marker Haplotype association with pancreatic sufficiency in cystic fibrosis. Am. J. Hum. Genet. 44:827-834.

Kidd, K. K., and J. Ott. 1984. Power and sample size in linkage studies. HGM 7:510-11.

Kong, A., M. Frigge, M. Irwin, and N. Cox. 1992. Importance sampling (I) : computing multimodel p-values in linkage analysis. Technical Report No. 337 - University of Chicago Dept. of Statistics.

Kwan, S.-P., J. Terwilliger, R. Parmley, G. Raghu, L.A. Sandkuyl, J. Ott, H. Ochs, R. Wedgwood, and F. Rosen. 1990. Identification of a closely linked DNA marker, DXS178, to further refine the X-linked agammaglobulinemia locus. Genomics. 6:238-242.

Lalouel, J.-M., and S. Yee. 1980. POINTER: A computer program for complex segregation analysis with pointers. Tech Rep Population Genetics Laboratory, University of Hawaii, Honolulu.

Lander, E. S., and D. Botstein. 1986. Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harbor Symp. Quant. Biol. 51:49-62.

Lander, E.S., and D. Botstein. 1987. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567-1570.

Lander, E., and P. Green. 1987. Construction of multilocus genetic linkage maps in humans. PNAS USA 84:2363-2367.

Lander, E. S., P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. 1987. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Lange, K., D. Weeks, and M. Boehnke. 1988. Programs for Pedigree Analysis: MENDEL, FISHER, and dGENE. Genet. Epidemiol. 5:471-472

Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott. 1984. Strategies for multilocus linkage analysis in humans. Proc. Natl. Acad. Sci. USA 81:3443-3446.

Lathrop, G.M., and J. Ott. 1990. Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am. J. Hum. Genet. 47:A188.

Liberman, U., and S. Karlin. 1984. Theoretical models of genetic map functions. Theor. Popul. Biol. 25:331-346.

Lindenbaum, S. 1979. Kuru sorcery. Palo Alto: Mayfield.

MacLean, C.J., L.M. Ploughman, S.R. Diehl, and K.S. Kendler. 1992. A new test for linkage in the presence of locus heterogeneity. Am. J. Hum. Genet. **50**:1259-1266.

Malcolm, S., J. Clayton-Smith, H. Nichols, S. Robb, T. Webb, J.A.L. Armour, A.J. Jeffreys, and M.E. Pembrey. 1990. Uniparental Paternal Disomy in Angelman's Syndrome. Lancet. **337**:694-697.

McKusick, V. A. 1990. Mendelian inheritance of man. 9th Edition. Baltimore: Johns Hopkins University Press.

Merette, C, M.C. King, and J. Ott. 1992. Heterogeneity analysis of breast cancer families by using age at onset as a covariate. Am. J. Hum. Genet. **50**:515-519.

Mills, K.A., K.H. Buetow, Y. Xu, J.L. Weber, M.R. Altherr, J.J. Wasmuth, and J.C. Murray. 1992. Genetic and physical maps of human chromosome 4 based on dinucleotide repeats. Genomics 14:209-219.

Morton, N.E. 1955. Sequential tests for the detection of linkage. Am. J. Hum. Genet. 7:277-318.

Morton, N.E., and V. Andrews. 1989. MAP, an expert system for multiple pairwise linkage analysis. Ann. Hum. Genet. 53:263-9.

Morton, N.E., and C.J. MacLean. 1974. Analysis of family resemblance. III. Complex segregation of quantitative traits. Am. J. Hum. Genet. 26:489-503.

Ott, J. 1974. Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. Am. J. Hum. Genet. 26:588-97.

Ott, J. 1976. A computer program for linkage analysis of general human pedigrees. Am. J. Hum. Genet. 28:528-529.

Ott, J. 1985. Analysis of Human Genetic Linkage. 1st Edition. Baltimore: Johns Hopkins University Press.

Ott, J. 1989. Computer-simulation methods in human linkage analysis. Proc. Natl. Acad. Sci. USA 86:4175-8.

Ott, J. 1991. Analysis of Human Genetic Linkage, 2nd Edition. Baltimore: Johns Hopkins University Press.

Ott, J. 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. Am. J. Hum. Genet. 51:283-290.

Ott, J. 1993. Molecular and statistical approaches to the detection and correction of errors in genotype databases. Am J Hum Genet (in press).

Ott, J. 1993b. Choice of Genetic Models for Linkage Analysis of Psychiatric Traits. In :Genetic Approaches to Mental Disorders, American Psychiatric Press (In Press).

Ott, J., and J.D. Terwilliger. 1992. Assessing the evidence for linkage in psychiatric genetics. In: Genetic Research in Psychiatry, edited by J. Mendlewicz and H. Hippius. Berlin: Springer-Verlag, pp. 245-249.

Ott, J., J.D. Terwilliger, and X. Xie. 1992. Determining the informativeness of untyped individuals in a pedigree analysis. Am. J. Hum. Genet. 51:A197

Petrukhin, K.E., M.C. Speer, E. Cayanis, M. de Fátima Bonaldo, U. Tantravahi, M.B. Soares, S.G. Fischer, D. Warburton, T.C. Gilliam, and J. Ott. 1993. A microsatellite genetic linkage map of human chromosome 13. Genomics **15**: 76-85.

Risch, N. 1988. A new statistical test for linkage heterogeneity. Am. J. Hum. Genet. 42:353-64.

Risch, N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. 46:222-28.

Risch, N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. 46:229-241.

Round, A.P. 1990. Computerized pedigree drawing in the SCHESIS risk calculation and linkage package. Am. J. Hum. Genet. 46:A75.

Rubinstein, P., M. Walker, C. Carpenter, C. Carrier, J. Krassner, C. Falk, and F. Ginsberg. 1981. Genetics of HLA disease associations. The use of the haplotype relative risk (HRR) and the "haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. Hum. Immunol. 3:384 (abstr.).

Safarazi, M. 1990. A database management system for linkage analysis. Proceedings of the European Society of Human Genetics Meeting, Corfu.

Schork, N.J., M. Boehnke, J.D. Terwilliger, and J. Ott. 1993. Two trait locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am. J. Hum. Genet. (submitted)

Seuchter, S.A., and M.H. Skolnick. 1988. HGDBMS: A human genetics database management system. Comput. Biomed. Res. 21:478-487.

Sherrington, R., J. Brynjolfsson, H. Petursson, M. Potter, K. Dudleston, B. Barraclough, J. Wasmuth, M. Dobbs, and H. Gurling. 1988. Localization of a susceptibility locus for schizophrenia on chromosome 5. Nature 336:164-7.

Shugart, Y.Y., and J. Ott. 1992. Significance tests relating to heterozygosity. Am. J. Hum. Genet. 51:A159.

Smith, C.A.B. 1953. The detection of linkage in human genetics. J. Roy. Statist. Soc. 15B:153-84.

Smith, C.A.B. 1963. Testing for heterogeneity of recombination fraction values in human genetics. Ann. Hum. Genet. 27:175-82.

Speer, M., J.D. Terwilliger, and J. Ott. 1992. A chromosome-based method for rapid computer simulation. Am. J. Hum. Genet. 51:A202.

Spence, M.A., D.T. Bishop, M. Boehnke, R.C. Elston, C.T. Falk, S.E. Hodge, J. Ott, J. Rice, K. Merikangas,and D. Kupfer. 1993. Methodological issues in linkage analyses for psychiatric disorders: secular trends, assortative mating, bilineal pedigrees. Report of the MacArthur Foundation Network I Task Force on Methodological Issue. Hum. Hered. 43:166-172.

Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. **52**:506-16.

Sturt, E. 1976. A mapping function for human chromosomes. Ann. Hum. Genet. 40:147-63.

Suarez, B.K., and P. Van Eerdewegh. 1984. A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. Am. J. Med. Genet. 18:135-146.

te Meerman, G.J. 1993. A logic programming approach to pedigree analysis, with applications to lod score, minimum recombinant and general genetic model computation. Hum. Hered. (in press)

Terwilliger, J.D., Y. Ding, and J. Ott. 1992. On the relative importance of heterozygosity and intermarker distance in gene mapping. Genomics **13**:951-56.

Terwilliger, J.D., G.M. Lathrop, and J. Ott. 1993a. Multipoint analysis to detect and quantify interference on CEPH chromosome 10 consortium data. (In preparation).

Terwilliger, J.D., T. Lehner, and J. Ott. 1991. Differential sex dependent penetrances of autosomal dominant diseases mimic linkage to the boundary of the pseudoautosomal region. Abstract to International Congress of Human Genetics.

Terwilliger, J.D., and J. Ott. 1990. Laboratory errors in the reading of marker alleles cause massive reductions in lod score and lead to gross overestimation of the recombination fraction. Am. J. Hum. Genet. **47**:A201.

Terwilliger, J.D., and J. Ott. 1992a. A multi-sample bootstrap approach to the estimation of maximized-over-models lod score distributions. Cytogenetics and Cell Genetics **59**:142-144.

Terwilliger, J.D., and J. Ott. 1992b. A haplotype-based haplotype relative risk statistic. Hum. Hered. **42**:337-346.

Terwilliger, J.D., and J. Ott. 1992c. A novel approach to combining data from multiple linked loci into a maximally heterozygous "super-locus" yields greatly increased power in 2-point linkage and sib-pair analysis. Am. J. Hum. Genet. **51**:A202.

Terwilliger, J.D., and J. Ott. 1993. A novel polylocus method for linkage analysis using the lod score or affected sib-pair methods. GAW 8 Proceedings (In Press).

Terwilliger, J.D., M.C. Speer, and J. Ott. 1993b. A chromosome-based method for rapid computer simulation in human genetic linkage analysis". Genet. Epid. (In press).

Thompson, M.W., R.R. McInnes, and H.F. Willard. 1991. Thompson and Thompson: Genetics in Medicine. Fifth Edition. Philadelphia: W.B. Saunders Company.

Tienari, P.J., J. Wikstrum, A. Sajantila, J. Palo, and L. Peltonen. 1992. Genetic susceptibility to multiple sclerosis linked to myelin basic protein gene. Lancet 340:987-991.

Trofatter, J.A., J.L. Haines, and P.M. Conneally. 1986. LIPIN: an interactive data entry and management program for LIPED. Am. J. Hum. Genet. 39:147-148.

Vogel, F., and A.G. Motulsky. 1986. Human genetics. New York: Springer.

Weeks, D.E. 1991. Human linkage analysis: strategies for locus ordering. In Advanced techniques in chromosome research, edited by K.W. Adolph, 297-330. New York: Marcel Dekker.

Weeks, D.E., and K. Lange. 1988. The Affected-Pedigree-Member method of linkage analysis. Am. J. Hum. Genet. 42:315-326.

Weeks, D.E., T. Lehner, E. Squires-Wheeler, C. Kaufmann, and J. Ott. 1990a. Measuring the inflation of the lod score due to its maximization over model parameter values in human linkage analysis. Genet. Epidemiol. 7:237-243.

Weeks, D.E., J. Ott, and G.M. Lathrop. 1990b. SLINK: a general simulation program for linkage analysis. Am J Hum Genet 47:A204.

Weeks, D.E., J. Ott, and G.M. Lathrop. 1991. Multipoint mapping under different models of interference using the LINKAGE programs. Am. J. Hum. Genet. 49:372.

Weir, B.S. 1990. Genetic data analysis. Sunderland, Mass.: Sinauer Associates, Inc.

Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morisette, P. Millasseau, G. Vaysseix, G.M. Lathrop. 1992. A second-generation linkage map of the human genome. Nature **359**:794-801.

White, R.L., J.-M. Lalouel, Y. Nakamura, H. Donis-Keller, P. Green, D.W. Bowden, C.G.P. Mathew, D.F. Easton, E.B. Robson, N.E. Morton, J.F. Gusella, J.L. Haines, A.E. Retief, K.K. Kidd, J.C. Murray, G.M. Lathrop, and H.M. Cann. 1990. The CEPH consortium primary linkage map of human chromosome 10. Genomics 6:393-412.

Wolak, G.R., and M. Sarfarazi. 1987. Plot 2000: A Universal Pedigree Plotting Program. J. Med. Genet. 24:246-247.

Wright, S. 1968. The Genetics of Human Populations: a treatise in four volumes. Paperback edition 1984. Chicago: University of Chicago Press.

Xie, X., and J. Ott. 1992. Finding all loops in a pedigree. Am J. Hum. Genet. 51:A206.

Zonta, L.A., S.D. Jayakar, M. Bosisio, A. Galante, and V. Pennetti. 1987. Genetic analysis of human obesity in an Italian sample. Hum. Hered. 37:129-39.

New References

1. Ott, J., *Analysis of human genetic linkage*. 3rd ed. 1999, Baltimore: Johns Hopkins University Press. xxiii, 382.
2. Terwilliger, J.D. and J. Ott, *Handbook of human genetic linkage*. 1994, Baltimore: Johns Hopkins University Press. x, 307.
3. Pulst, S.M., *Genetic linkage analysis.* Arch Neurol, 1999. **56**(6): p. 667-72.
4. Ott, J., Y. Kamatani, and M. Lathrop, *Family-based designs for genome-wide association studies.* Nat Rev

Genet, 2011. **12**(7): p. 465-74.

5. Mather, K., *Crossing-over*. Biol Reviews (Cambridge Philosophical Society), 1938. **13**: p. 252-292.

6. Ott, J. and H. Donis-Keller, *Statistical methods in genetic mapping.* Genomics., 1994. **22**(2): p. 496-7.

7. Kamphans, T., et al., *Filtering for compound heterozygous sequence variants in non-consanguineous pedigrees.* PLoS One, 2013. **8**(8): p. e70151.

8. Hong, Y.B., et al., *A compound heterozygous mutation in HADHB gene causes an axonal Charcot-Marie-tooth disease.* BMC Med Genet, 2013. **14**: p. 125.

9. McMillan, H.J., et al., *Compound heterozygous mutations in glycyl-tRNA synthetase are a proposed cause of systemic mitochondrial disease.* BMC Med Genet, 2014. **15**: p. 36.

10. Bell, J. and J.B.S. Haldane, *The linkage between the genes for colour-blindness and haemophilia in man.* Proc Roy Soc Ser B, 1937. **123**: p. 119-150.

11. Ott, J., *Y-linkage and pseudoautosomal linkage.* Am J Hum Genet., 1986. **38**(6): p. 891-7.

12. Elston, R.C. and J. Stewart, *A general model for the genetic analysis of pedigree data.* Hum Hered, 1971. **21**(6): p. 523-42.

13. Goldschmidt, R., *Gen und Ausseneigenschaft (Untersuchungen an Drosophila) I.* Z Indukt Abstamm Vererbungsl, 1935. **69**(1): p. 38-69.

14. Ming, J.E. and M. Muenke, *Multiple hits during early embryonic development: digenic diseases and holoprosencephaly.* Am J Hum Genet, 2002. **71**(5): p. 1017-32.

15. Savage, D.B., et al., *Digenic inheritance of severe insulin resistance in a human pedigree.* Nat Genet, 2002. **31**(4): p. 379-84.

16. Schaffer, A.A., *Digenic inheritance in medical genetics.* J Med Genet, 2013. **50**(10): p. 641-52.

17. Marini, C., et al., *Childhood absence epilepsy and febrile seizures: a family with a GABA(A) receptor mutation.* Brain, 2003. **126**(Pt 1): p. 230-40.

18. Thompson, E., *The structure of genetic linkage data: from LIPED to 1M SNPs.* Hum Hered, 2011. **71**(2): p. 86-96.

19. Gelernter, J., et al., *Assignment of the 5HT7 receptor gene (HTR7) to chromosome 10q and exclusion of genetic linkage with Tourette syndrome.* Genomics, 1995. **26**(2): p. 207-9.

20. Lange, K., et al., *Mendel: the Swiss army knife of genetic analysis programs.* Bioinformatics, 2013. **29**(12): p. 1568-70.

21. Matise, T.C., et al., *A second-generation combined linkage physical map of the human genome.* Genome Res, 2007. **17**(12): p. 1783-6.

22. Hartl, D.L., *A primer of population genetics*. 3rd ed. 2000, Sunderland, Mass.: Sinauer Associates. xvii, 221 p.

23. Hartl, D.L. and A.G. Clark, *Principles of population genetics*. 4th ed. 2007, Sunderland, Mass.: Sinauer Associates. xv, 652 p.

24. Hiekkalinna, T., et al., *PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals.* Hum Hered, 2011. **71**(4): p. 256-66.

25. Gertz, E.M., et al., *PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD.* BMC Bioinformatics, 2014. **15**: p. 47.

26. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nat Genet, 1995. **11**(3): p. 241-7.

27. Knapp, M., S.A. Seuchter, and M.P. Baur, *Linkage analysis in nuclear families. 2: Relationship between affected sib-pair tests and lod score analysis.* Hum Hered, 1994. **44**(1): p. 44-51.

28. Marsaglia, G., *Random numbers fall mainly in the planes.* Proc Natl Acad Sci U S A, 1968. **61**(1): p. 25-8.

29. Abramowitz, M. and I.A. Stegun, *Handbook of mathematical functions, with formulas, graphs, and mathematical tables*. Ninth printing ed. 1972, New York: Dover Publications. xiv, 1046 p.

30. Fisher, R.A., *Statistical methods for research workers*. 14th ed. 1970, New York: Oliver and Boyd. 362.

31. Ott, J., J. Wang, and S.M. Leal, *Genetic linkage analysis in the age of whole-genome sequencing.* Nat Rev Genet, 2015. **16**(5): p. 275-84.

32. Dietter, J., et al., *Efficient two-trait-locus linkage analysis through program optimization and parallelization: application to hypercholesterolemia.* Eur J Hum Genet, 2004. **12**(7): p. 542-50.

33. Schork, N.J., et al., *Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits.* Am J Hum Genet., 1993. **53**(5): p. 1127-36.

34. Wu, C.C. and S. Shete, *Analysis of genes for alcoholism using two-disease-locus models.* BMC Genet, 2005. **6 Suppl 1**: p. S149.

35. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

36. Xie, X. and J. Ott. *Testing linkage disequilibrium between a disease gene and marker loci*. in *43rd Annual Meeting*. 1993. New Orleans LA, October 5, 1993: Am J Hum Genet.