The Use of Computer Simulation in Genetic Linkage Studies

JOHN RICE, PH.D.

The search for the gene or genes associated with alcoholism is complicated by the heterogeneous nature of the disease and the difficulty of selecting appropriate families for study. Computer programs that simulate genetic inheritance patterns may help make the task easier.

The author conducts a simulation at the computer. Photograph courtesy of John Rice, Ph.D.

Research over the past 20 years has provided compelling evidence that at least part of the vulnerability to alcoholism is inherited. This has led to a search for the gene or genes responsible for this vulnerability. An important strategy for determining the location of a disease gene on a chromosome is linkage analysis. Linkage analysis relies on the identification of a gene that may be completely unrelated to the disease, but that lies so close to the disease gene that the two tend to be transmitted to offspring as a unit. (For a more complete discussion of linkage, see the article by Crabb, pp. 197–203.)

The most powerful approach to linkage analysis is to study informative family pedigrees. The strategy is to determine from these pedigrees the manner in which the disease may be inherited and then to determine whether some other trait or characteristic is transmitted through the families along with the disease.

A neighboring gene can serve as a marker for the disease, because its presence suggests with high probability that

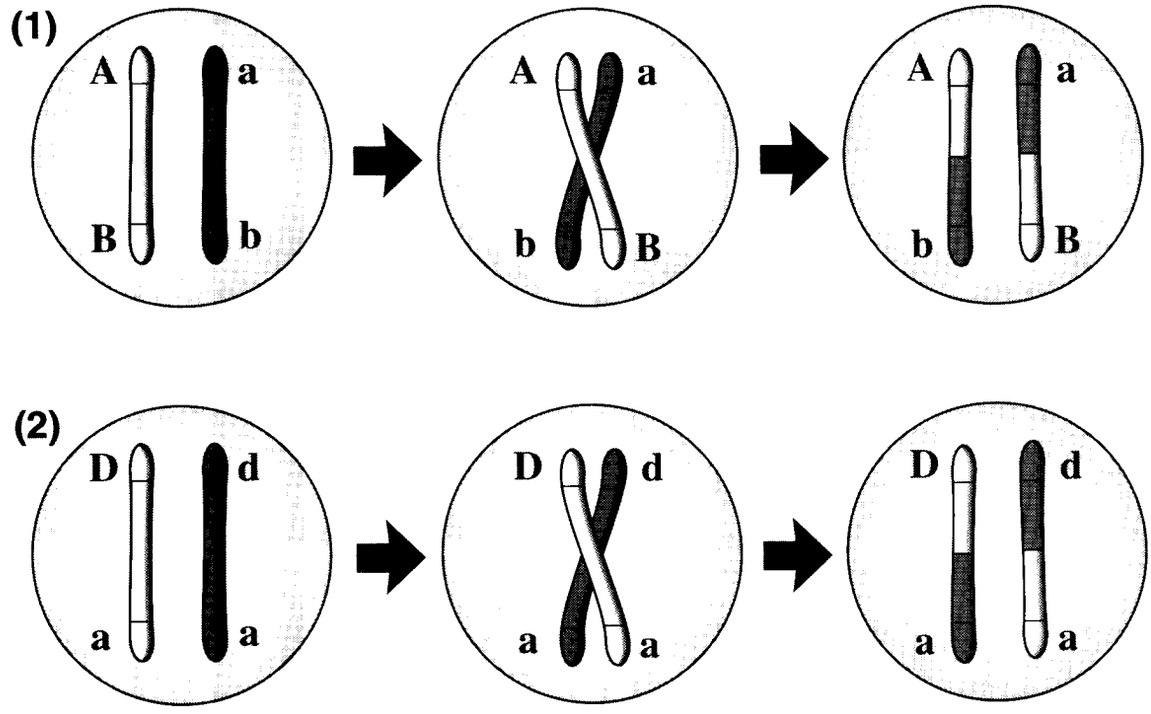

Figure 1 Crossing over between homologous chromosomes. (1) A crossover in an AB/ab parent leads to recombinants Ab and aB. The probability that a recombinant will occur is denoted as Θ. (2) A crossover between a disease locus and a marker when the parent is homozygous (for example, aa) at the marker. There is no way to determine whether the resulting gamete is a recombinant, and thus no information about linkage.

the disease is also present, even though the disease gene itself may not have been identified. Moreover, once a linkage is detected, markers closer to the gene may be found, and ultimately the disease gene itself may be cloned and its structure determined. The next step is to locate the structural defect within the gene, with the ultimate goal of curing or preventing the disease.

For some diseases that are caused by a defect in a single gene, such as sickle cell disease or Duchenne and Becker muscular dystrophies, the process of genetic analysis has been rather straightforward. Determining linkages is much more difficult for a condition such as alcoholism, in which several different genes may be involved. It is highly unlikely that the same gene or genes confers vulnerability in all families. The genes may be specific to alcoholism, or they may predispose to alcoholism through a more general effect on appetite, personality, mood, or behavior. Some cases of alcoholism appear to lack any genetic component; such cases may occur in the same families as genetically influenced alcoholism, further complicating the analysis.

One important approach to these problems is the use of computer simulations, in which a complex system or process is modeled based on data sampled at random from the system. By repeated sampling, information is gained to guide the design and to evaluate the feasibility of expensive, long-term clinical studies before field work is undertaken. In addition, these techniques can be used to develop statistical tests to be performed once data are collected.

For alcoholism, where cases are known to cluster within families but the mechanism by which the trait is inherited is uncertain, simulations may be used to model complexities such as heterogeneity, environmental resemblance, or age effects, and to evaluate their impact on the ability to detect linkage to a single gene. In addition, alternative sampling strategies may be explored to ensure that pedigrees are selected in an optimal way and have enough family members to provide a statistically adequate sample size.

This article will demonstrate how computer simulation can be used to estimate the information for a genetic linkage study. General approaches to simulating genetic data for a given set of pedigrees have been described by Boehnke (1986), Ploughman and Boehnke (1989) and Ott (1989), and a computer program, SIMLINK (available from Drs. Ploughman and Boehnke), is commonly used to implement these methods. Other applications of simulation studies can be found in Goldin et al. (1984), Cox et al. (1988),

JOHN RICE, PH.D., is professor of mathematics in psychiatry, Department of Psychiatry and Division of Biostatistics, Washington University School of Medicine, St. Louis, Missouri.

This research was supported in part by Public Health Service grants MH37685, MH31302, MH43028, MH25430, and AA08401, and the MacArthur Task Force on Analytic Strategies for Linkage Analysis of Psychiatric Disorders.

Martinez and Goldin (1989) and Neuman and Rice (1990). Much of this work is mathematical in nature, and rather than go into a detailed treatment here, we will start from first principles and use a basic example to demonstrate the utility of simulations.

METHODS

The Recombination Fraction Θ

Humans have 22 autosomal (nonsex) pairs of chromosomes plus 2 unpaired sex chromosomes (the X and the Y). The chromosomes of a pair are termed homologous, and a particular gene occupies the same position (or locus) on each. Many genes are polymorphic; that is, they occur in different forms called alleles. If a gene has two alleles, A and a, the possible genotypes at that locus are AA, Aa, and aa. Individuals who are AA or aa are known as homozygotes, and those who are Aa are heterozygotes. The expression of a particular genotype is known as the phenotype; the phenotype is the observable physical or biochemical trait produced by the genotype.

If allele D is dominant over d, then Dd and DD individuals express the same phenotype. However, there are cases where a given genotype, although present, is not expressed, and the corresponding phenotype is not observed. This phenomenon is called reduced penetrance and will be discussed later.

Gametes (eggs and sperm) are formed by a process known as meiosis, by which each gamete receives only one of each pair of chromosomes. In the formation of gametes, a parent passes one of his or her two alleles with a probability of ½. Thus, a child receives one allele from the father and one from the mother. In the case of the sex chromosomes, because a male is XY and a female XX, a male offspring receives his Y chromosome from the father, and a female offspring receives the father's X chromosome. This results in the distinctive pattern of inheritance for X-linked genetic disorders such as color blindness and hemophilia.

For two loci with, for example, alleles A, a at the first and alleles B, b at the second, four combinations (or haplotypes) are produced: AB, Ab, aB, and ab. If an individual with genotype AB/ab produces gametes of the four types, each with probability ¼, then the two loci are said to be unlinked.

In the process of meiosis during gamete formation, crossing over between homologous chromosomes can occur, as illustrated in Figure 1. If the loci are close physically on the same chromosome, they will tend to be transmitted as a unit, rather than be separated by a crossover. That is, if a parent is AB/ab (indicating that A and B are on one of the two homologous chromosomes, and that a and b are on the other), then the AB and ab gametes will be overrepresented. Haplotypes Ab and aB, the products of crossing over, are termed recombinants, and AB and ab are termed nonrecombinants. The probability that a parent will produce a recombinant is called the recombination fraction. It is traditionally denoted by the Greek letter Θ. If two loci are on different chromosomes, they are unlinked and $\Theta = \frac{1}{2}$.

As stated earlier, the goal in linkage analysis is to identify an (unknown) disease locus by identifying a known marker locus linked to the disease. In general, a battery of markers is available for linkage analysis, and we wish to test statistically for cotransmission of the marker with the disease phenotype in families (Figure 2).

The Lod Score

Consider a dominant disease with alleles D and d and suppose an affected parent is homozygous at the marker (a), so that he or she is, for example, Da/da (compare Figure 1). Recombinants would be Da and da, and nonrecombinants would also be Da and da. Therefore, there is no way to identify which children are recombinants and which are not. This illustrates the rule that we need a doubly heterozygous parent to provide information for linkage. Now consider a doubly heterozygous parent (Dd at the disease locus and Aa at the marker locus). If there is only one child, we do not have enough information to test for linkage, because we do not know the phase of the parent. That is, we do not know which two alleles occur together on the same chromosome; we do not know whether the parent is DA/da or Da/dA, so we cannot tell whether or not the child is a recombinant. Thus we need more than one child in order to estimate Θ .

In deciding whether there is evidence for linkage in a data set, investigators compute a number called the lod score, Z. This number reflects how strongly the data support evidence for linkage. A value of Z above 3 for a given recombination fraction is taken as evidence for linkage, a value

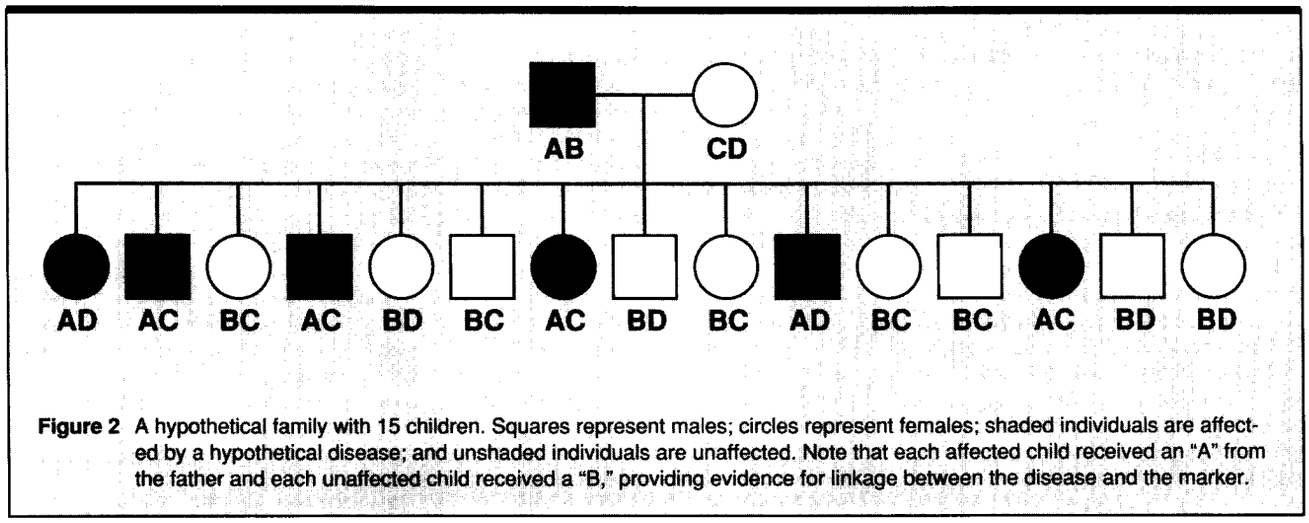

Figure 2 A hypothetical family with 15 children. Squares represent males; circles represent females; shaded individuals are affected by a hypothetical disease; and unshaded individuals are unaffected. Note that each affected child received an "A" from the father and each unaffected child received a "B," providing evidence for linkage between the disease and the marker.

Table 1 Simulation Results for Incomplete Marker Information

			1 Family		3 Families	
Phase[1]	Θ[2]	Number of Alleles	Mean Lod Score[3]	Power[4]	Mean Lod Score[3]	Power[4]
Known	0.0	Infinite	4.52	100%	4.52	100%
Unknown	0.0	Infinite	4.22	100%	3.62	100%
Unknown	0.0	4	2.92	64%	2.43	31%
Unknown	0.0	3	2.60	56%	2.11	20%
Unknown	0.0	2	1.67	27%	1.21	0%
Unknown	0.05	4	2.36	56%	1.77	13%
Unknown	0.05	3	2.05	46%	1.70	13%
Unknown	0.05	2	1.19	17%	0.85	1%
Unknown	0.1	4	2.02	43%	1.36	9%
Unknown	0.1	3	1.49	26%	1.17	5%
Unknown	0.1	2	0.84	12%	0.73	0%

[1]Phase = The genotype of the parent.
[2]Θ = Recombination fraction—the probability that a parent will produce a child who is a recombinant, or who has a different combination of alleles from that of the parent.
[3]Lod score = The logarithmic odds of linkage among loci; the evidence for linkage in a data set.
[4]Power = The power to detect linkage (which decreases as the Θ increases or is greatly reduced as penetrance is decreased, for example).

below −2 is taken as evidence against linkage, and values between −2 and 3 indicate that more data is needed before reaching a conclusion (Ott 1985). A value of 3 corresponds to odds favoring linkage of 1,000:1.

Investigators choose a particular set of families and simulate marker data and compute the average lod score in each situation. They also compute the power to detect linkage, where the power is the proportion of times the lod score is above 3. A power of 100 percent means that, if a linkage does exist, it will always be found (using the particular families and markers selected for study). A power of 10 percent means that there is only 1 chance in 10 that the families and markers selected will provide enough information to detect the linkage.

RESULTS

To illustrate these concepts, we chose a single nuclear family with 2 parents and 15 children. We chose 1 parent affected by a hypothetical disease and 1 parent unaffected, with 6 of the 15 children affected. We first considered a rare dominant trait, and assumed that the affected parent is of genotype Dd at the trait locus while the unaffected parent is of genotype dd. In addition, we considered 3 families, each with 5 children, 2 of whom were affected and 3 of whom were unaffected. Under these conditions, we know that each affected child is of genotype Dd and each unaffected child is of genotype dd.

If each affected parent is a double heterozygote and phase is known (see above), the lod score is calculated to be 4.52 for the case of Θ = 0.0. (A Θ of 0 indicates zero probability that a parent will produce a recombinant; it indicates that the genes are closely linked and always cotransmitted.) Moreover, since we have complete information, there is no difference between having 15 children in 1 family, or 5 children in each of 3 families. (By analogy, the odds of a coin toss are the same whether one flips 1 penny 15 times or 3 pennies 5 times each.) These lod scores are given in the first row of Table 1. Using a lod score above 3 as a cutoff, we have 100 percent power in this "ideal" situation; the linkage will always be detected. We have indicated an infinite number of alleles in the table to indicate that the marker is fully informative: Both parents are heterozygous at the marker, as in the example shown in Figure 2.

In general, there is no way to know the phase of the affected parent in a nuclear family. When Θ = 0.0, examination of the first child will, in fact, determine the phase of the parent (since there could not have been a crossover), so that the score with phase unknown is the same as the lod score for a family of 14 children with phase known. In this setting, the contribution of one child to the lod score is 0.3 when Θ = 0.0, so that the lod score is reduced by 0.3 in the second row of Table 1 for one family, and 0.9 for three families.

We then asked two important practical questions: What is the effect of having a marker that is not fully informative? What is the effect of Θ = 0.0? (A marker would not be fully informative if, due to chance, the person were homozygous at the marker; see Figure 1). We used the program SIMLINK to address these questions. We specified the family structure and phenotypes, and randomly selected marker data assuming 4, 3, and 2 alleles (that is, 4, 3, or 2 possible variants of the marker gene).

In each case, we assumed that the alleles occurred with equal frequency. Thus, we assumed frequencies of 0.25 for each of four alleles; 0.33 for each of three alleles; and 0.5 for each of two alleles. Equal frequencies can be shown to give the greatest power, and other combinations would give reduced lod scores and power. Simulations were performed for various combinations of Θ , number of alleles, and family structure, as shown in Table 1. One hundred replicates were generated for each combination, and the average lod score and power were determined and reported in Table 1.

Recall that if the affected parent is a ho-

mozygote at the marker locus, then the family contributes no information for linkage. This is more likely to happen as the marker becomes less polymorphic (as there are fewer possible alleles that can occur at the marker locus). Note that when $\Theta = 0.0$, the power decreases to 27 percent for the two-allele system with one large family, and 0 percent with the three small families (Table 1). These simulations underscore the importance of having highly polymorphic markers. Indeed, Table 1 shows that an experiment that could be definitive with one marker may be a waste of time (and money) with a less informative one. As expected, the power to detect linkage decreases as Θ increases.

In the above simulation, we assumed a one-to-one correspondence between the genotype and phenotype at the trait. This means that all affected people were Dd and all unaffected people were dd (complete penetrance). We next simulated 100 replicates of the large family with a marker with 4 alleles and a reduced penetrance at the trait locus. Reduced penetrance means that not all the children who are Dd will be affected. We simulated data with all affected people known to be Dd and with the unaffected parent to be dd. We allowed the unaffected children to be either dd or Dd according to the penetrances indicated in Table 2. For example, with a penetrance of 0.8, 20 percent of the children who are Dd will have the unaffected phenotype. Thus, it is uncertain whether an unaffected person has genotype dd and is a recombinant, or has genotype Dd and is a nonpenetrant nonrecombinant. Even when in reality $\Theta = 0.0$, the power to detect linkage is greatly reduced as penetrance is decreased.

In a linkage study, in addition to the power to detect linkage, it is important to be able to exclude areas of the chromosome where there is no trait locus. The generally accepted cutoff to exclude linkage to a marker is when the lod score is less than −2. The last column in Table 2 shows the average lod score when the true Θ is $1/2$ (i.e., no linkage) in each simulation, and the lod score is computed at $\Theta = 0$. Note that the reduced penetrance also affects our power to exclude linkage.

DISCUSSION

We created an example to illustrate the process of using simulations to explore the robustness of methods and to estimate the power of families for the detection of linkage. It must be emphasized that computation of power depends on assumptions concerning the true mechanism of inheritance of the disease. When the mode of inheritance is known, the family phenotypic data can be used with a program such as SIMLINK to determine power. When designing a study, the phenotypic data itself can be simulated to decide whether the intended sample size is large enough to provide adequate power for testing the primary hypotheses.

The above simulations are based on the assumption that the trait being studied is determined by a rare dominant gene; therefore, they may provide little guidance on the sample sizes needed for a study of alcoholism, which is not likely to be caused by such a gene. A major question we have not addressed is that of heterogeneity. If only a portion of families have the disease linked to a marker, then the overall lod score would be made up of families that make a positive contribution (the linked families) and those that make a negative contribution (the unlinked families). Accordingly, heterogeneity must be considered for linkage to a complex disease. Simulations by Martinez and Goldin (1989) have explored the effects of heterogeneity in this setting.

Table 2 Simulation Results for Reduced Penetrance

	$\Theta = 0.0$		$\Theta = 0.5$
Penetrance	**Mean Lod Score**	**Power**	**Mean Lod Score**
1.0	2.92	64%	−16.34
0.9	2.08	19%	−8.34
0.8	1.75	14%	−6.65
0.7	1.41	3%	−6.60
0.6	1.26	0%	−7.17
0.5	1.18	0%	−5.79

Another complicating factor is the presence of bilineal pedigrees, in which a trait occurs on both sides of the family. Alcoholism is a common disorder, and, moreover, there may be a tendency for assortative mating to increase the chances of finding families with alcoholism on both the maternal and paternal sides. If a disorder is heterogeneous, families may be sampled with heterogeneous forms of illness within the same pedigree. Although the strategy of identifying large extended pedigrees has proven successful in dissecting heterogeneity for a rare disorder such as retinitis pigmentosa, this approach may be problematic for a common disorder.

In summary, simulation studies will likely play a central role in genetic studies of disorders such as alcoholism. The availability of linkage maps of the human genome provides a technology which, if thoughtfully applied, can help unravel diseases that have eluded investigators seeking clues to the etiology and prevention of genetic illnesses. ■

REFERENCES

BOEHNKE, M. Estimating the power of a proposed linkage study: A practical computer simulation of approach. *American Journal of Human Genetics* 39(4):513-527, 1986.

COX, N.J.; HODGE, S.E.; MARAZITA, M.L.; SPENCE, M.A.; AND KIDD, K.K. Some effects of selection strategies on linkage analysis. *Genetic Epidemiology* 5(4):289-297, 1988.

GOLDIN, L.R.; COX, N.J.; PAULS, D.L.; GERSHON, E.S.; AND KIDD, K.K. The detection of major loci by segregation and linkage analysis: A simulation study. *Genetic Epidemiology* 1(3):285-296, 1984.

MARTINEZ, M.M., AND GOLDIN, L.R. The detection of linkage and heterogeneity in nuclear families for complex disorders: One versus two marker loci. *American Journal of Human Genetics* 44(4):552-559, 1989.

NEUMAN, R.J., AND RICE, J.P. A note on linkage analysis when the mode of transmission is unknown. *Genetic Epidemiology* 7:349-358, 1990.

OTT, J. *Analysis of Human Genetic Linkage*. Baltimore: The John Hopkins University Press, 1985.

OTT, J. Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences, USA* 86(11):4175-4178, 1989.

PLOUGHMAN, L.M., AND BOEHNKE, M. Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* 44(4):543-551, 1989.