**Multilocus Linkage Analysis in Humans: Detection of Linkage and Estimation of Recombination**

G. M. LATHROP, J. M. LALOUEL, C. JULIER, AND J. OTT

**SUMMARY**
Multilocus linkage analysis is investigated from the viewpoint of the efficiency of recombination estimates under different strategies for detecting linkage and determining gene order within a linkage group. We consider the appropriateness of assuming no interference with data available in human genetic studies. Examples are given to show the significance of multilocus analysis in humans. A computer program package, LINKAGE, for multilocus linkage analysis is described.

**INTRODUCTION**
The motivation for contemplating multilocus linkage analysis arises from the rapidly increasing number of genetic markers available for linkage studies in humans [1, 2]. With few exceptions [3, 4], linkage analysis in humans has been based on two-point tests by standard methods reviewed in [5, 6]. With the prospect of denser linkage groups and finer genetic maps, we may anticipate that multilocus analyses should prove more efficient than classical two-locus methods [7]. However, the merits of multilocus analysis must be evaluated in relation to purpose and strategy prior to deciding on analytical methods and algorithms for a solution. The relevance of multilocus analysis for the calculation of genetic risks is clear; for the detection of linkage, the estimation of recombination, and the construction of a genetic map, it remains to be established.
MULTILOCUS LINKAGE ANALYSIS

The construction of a linear map from recombination data requires a functional relation between recombination and genetic distance. The latter reflects the probability of occurrence of crossing-overs in an interval and is expressed in Morgans, where a Morgan is that distance for which, on average, one crossing-over is observed over a large number of meioses. Genetic distance is additive, that is, the distance between two loci is the sum of the distances in contiguous segments involving an intervening locus. Although recombination yields order constraints on the location of loci in situations that are amenable to analysis in humans, recombination values are not additive over contiguous segments for two major reasons.

First, recombination between two loci results from the occurrence of an odd number of crossing-overs in the interval they define. If such events were independent, the recombination rates between three loci, $A$, $B$, and $C$, in that order, would satisfy the relation $\theta_{AC} = \theta_{AB} + \theta_{BC} - 2\theta_{AB}\theta_{BC}$, approximately additive only for closely linked genes.

Second, considerable evidence has accumulated documenting the nonindependence of crossing-overs, or interference. This evidence stems from multipoint crosses in laboratory animals [8]. Various tests have been devised to detect interference [8, 9], and many mapping functions have been proposed to express the relationship between recombination and map distance (e.g., [8, 10–14]). It should be noted, however, that the experimental conditions of these early studies, which involve controlled crosses and very large sample sizes, do not apply to humans. To deal with this issue, we must take into account the nature and limitations of feasible study designs.

In an earlier publication [15], particular attention was given to the problem of detecting linkage between a disease locus or a new marker and a pre-established linkage group, for which we presented a convenient graphical method of location scores. The earlier version of our Pascal computer program LINKAGE [16] has been extended to deal with a variety of situations in multilocus linkage analysis; the implementation is described in APPENDIX A.

The present paper is devoted to statistical aspects underlying the approaches we have proposed and to consideration of some illustrative examples. We first recall concepts and notations necessary for the study of three-point linkage. Then we evaluate the relative efficiency of three-point and two-point estimates for a variety of strategies when interference is not present and discuss the appropriateness of assuming no interference for such data as are available in human genetics. This is followed by a discussion of the inference of gene order. Two examples are given to underline the significance of multilocus linkage analysis in humans.

THREE-POINT LINKAGE

Let $A$, $B$, and $C$ represent linked loci, in any order, with recombination rates $\theta_{AB}$, $\theta_{BC}$, and $\theta_{AC}$. Segregating chromosomes belong to one of four recombination classes: (1) recombination between both $AB$ and $BC$ (hence, no recombination between $AC$); (2) recombination between $AB$ and no recombination between $BC$ (hence, recombination between $AC$); (3) no recombination between
AB and recombination between BC (hence, recombination between AC); and
(4) no recombination. Let the probabilities of these events be $p_1, p_2, p_3,$ and $p_4,$ respectively, with $\Sigma p_i = 1.$ The maximum likelihood estimates of $p_1, p_2,$ and $p_3$ yield, in turn, maximum likelihood estimates of the recombination rates, using the identities [9]:

\[
\begin{align*}
\theta_{AB} &= p_1 + p_2 \\
\theta_{BC} &= p_1 + p_3 \\
\theta_{AC} &= p_2 + p_3
\end{align*}
\]

(1)

Solving for the $p$'s gives:

\[
\begin{align*}
p_1 &= (\theta_{AB} + \theta_{BC} - \theta_{AC})/2 \\
p_2 &= (\theta_{AB} + \theta_{AC} - \theta_{BC})/2 \\
p_3 &= (\theta_{AC} + \theta_{BC} - \theta_{AB})/2
\end{align*}
\]

Since the probabilities for the recombination classes ($p$'s) must be nonnegative, this shows that the true recombination rates among any three loci satisfy the triangular inequality, that is, the sum of any two recombination rates is greater than or equal to the third. Estimates from separate two-locus analyses may violate this requirement and thus be inconsistent. The three-locus analysis, with the constraint $p_i > 0$ for all $i,$ always leads to consistent estimates of recombination. Equation (1) requires no assumption either on the gene order or on interference.

If it is assumed that recombination between flanking loci is always greater than or equal to that between adjacent loci, then the most likely gene order can be inferred from the estimated recombination rates. If the inferred order is A-B-C, that is, $\theta_{AC}$ is the largest estimated recombination, then interference can be measured in terms of a coefficient of coincidence, $c,$ which has a nonnegative value:

\[
c = (\theta_{AB} + \theta_{BC} - \theta_{AC})/2 \theta_{AB}\theta_{BC}
\]

(2)

Solving for $\theta_{AC}$ gives the equivalent formulation: $\theta_{AC} = \theta_{AB} + \theta_{BC} - 2 c \theta_{AB} \theta_{BC}.$ By expressing the $p$'s in terms of $\theta_{AB}, \theta_{BC},$ and $c,$ estimates of the latter can be obtained directly.

The upper bound for $c$ under the order A-B-C is $1/[2 \min (\theta_{AB}\theta_{BC})], which is obtained when $\theta_{AC} = \max(\theta_{AB}, \theta_{BC}).$ Note, however, that $c > 1$ implies negative interference, that is, that crossing-over in one segment increases the probability of recombination in an adjoining segment. Logically, if we allow negative interference, there is no reason to give an upper bound on $c.$ Conceivably,
MULTILOCUS LINKAGE ANALYSIS

485

therefore, recombination between flanking loci may be less than that between adjacent loci. In this case, identical estimates of the p’s are obtained for all gene orders, implying that the likelihood of the orders are equivalent. This simplifies calculations as the parameters need be estimated only once. On the other hand, if it is assumed that interference cannot be negative, c must be restricted to the interval [0,1].

An alternative is to assume a specific mapping function that specifies coincidence from the values of recombination in adjacent segments. Under the assumption of no interference, crossing-over events are independent, so that \( \theta_{AC} = \theta_{AB} + \theta_{BC} \) and \( c = 1 \). With complete interference, crossing-over in one segment inhibits crossing-over in the adjacent segment; hence, \( \theta_{AC} = \theta_{AB} + \theta_{BC} \) and \( c = 0 \). With Kosambi’s mapping function, coincidence varies between 0 and 1, depending on the recombination rates in the adjacent segments, so that: \( \theta_{AC} = (\theta_{AB} + \theta_{BC})(1 + 4\theta_{AB}\theta_{BC}) \). With this, solving equation (2) gives \( c = 2(\theta_{AB} + \theta_{BC})(1 + 4\theta_{AB}\theta_{BC}) \). Felsenstein [14] discusses a general family of mapping functions that allows calculation of the recombination rate between flanking markers from the rates in two adjacent segments.

Once a mapping function is assumed, the probabilities of the recombination classes in the offspring depend on gene order; hence, estimation of the recombination rates using a mapping function requires comparison of the likelihood of various gene orders. A number of considerations, discussed below, lead us to propose the assumption of no interference for most human genetic studies.

**Relative Efficiency of Three-Point and Two-Point Linkage Estimates**

The relative merits of three-point and two-point linkage tests can be assessed in terms of the precision of the estimates of relative locations of the loci, which depends on the variances of the estimated recombination estimates. For a fixed study design, the relative efficiency of three-locus vs. two-locus linkage estimates of a recombination parameter is defined as the inverse of the ratio of the corresponding variances; the larger this ratio, the greater the gain to be expected from three-point over two-point data.

The relative efficiency is a function of the strategy used for estimating recombination. Depending on prior knowledge and assumptions regarding interference, we shall consider four possible strategies of linkage analysis: (1) no assumption regarding interference; (2) assuming no interference; (3) assuming knowledge of recombination between two of the three loci but making no assumption on interference; and (4) assuming knowledge of one recombination rate and no interference. (As discussed in [16], strategies (3) and (4) are of interest for the analysis of a new locus when prior studies have established a linkage map for other markers.)

In addition, the relative efficiency depends on the recombination rates, the mode of inheritance of the traits, and the type of family data available. For simplicity, we will consider only family units in which phenotype information is available for the three loci on all individuals. Families that are informative for linkage range from parents and two offspring when phase is unknown to pedi-
degrees in which parental phase can be inferred. Note that the phase-known case is approached as the number of children in a single family increases. In sibships from phase-known parents, each child contributes independent information on recombination; hence, for the latter case, we consider phase-known families with one child.

For loci $A$ and $C$, we assume codominant alleles with equal gene frequencies and either two alleles each or a number sufficiently large so that parents can be considered to carry four different alleles at a locus (distinct allele case). The locus $B$ may be either a codominant system similar to $A$ and $C$ or a rare disease trait with either dominant or recessive inheritance, possibly with reduced penetrance. For a disease trait, we assume truncate selection through affected children and known penetrance. The phase for a rare recessive disease generally cannot be determined with pedigree data; in this case, the results for phase-known families apply only for large sibship size.

Large-sample variances of the estimated recombination rates are calculated by inverting a numerical approximation to the expected information matrix as discussed in APPENDIX B. When assuming no interference, the estimates are calculated for the true gene order. We have verified that the extent to which the variances of the estimates would be altered when gene order is simultaneously inferred is negligible for the number of offspring observations typical in human studies.

Table 1 shows the relative efficiencies of three-point vs. two-point linkage analysis with $\theta_{AB} = 0.1$ and $\theta_{BC} = 0.1$ for the order $ABC$ and with $\theta_{AC} = 0.05$ and $\theta_{CB} = 0.1$ for the order $ACB$. The latter values have been chosen to assure that the locus $A$ is close enough to the disease locus $B$ to contribute some information in the linkage analysis. The true recombination rates between flanking markers are derived assuming no interference.

For codominant loci with two markers at each locus, we find that the three-point linkage estimates are 1.12 to 1.30 times more efficient for the estimates of the recombination rates in the two contiguous segments in phase-unknown and phase-known families, respectively; for the flanking loci, the increase in efficiency is 1.30 and 1.61 in these two cases. Thus, some improvement of the estimates can be obtained using the three-locus analysis. As the number of alleles with approximately equal frequencies increases, the gain in efficiency is reduced; for the distinct allele case, the relative efficiencies are less than 1.05.

When the assumption of no interference is introduced, the relative efficiencies are similar to those above for the rates within each segment. By contrast, the efficiency of the three-point estimate of recombination between the flanking markers is between 3.7 and 3.8 times greater than the efficiency of the two-point estimate. When the recombination rate between flanking markers is fixed, the three-locus estimates of the other rates have greater efficiency than in the previous cases. In particular, for strategy (4), the efficiency is twice that of the two-locus estimates in all the codominant examples given in table 1.

It can be seen from table 1 that a rare dominant disease with complete or reduced penetrance yields relative efficiencies similar to those for the codominant case. For a rare recessive, the relative efficiencies are generally larger,
MULTILOCUS LINKAGE ANALYSIS

TABLE I
RELATIVE EFFICIENCY OF THREE-LOCUS VS. TWO-LOCUS ESTIMATES OF RECOMBINATION RATES

<table>
<thead>
<tr>
<th>Case</th>
<th>Two alleles strategy (1)</th>
<th>Two alleles strategy (2)</th>
<th>Two alleles strategy (3)</th>
<th>Two alleles strategy (4)</th>
<th>Distinct alleles strategy (1)</th>
<th>Distinct alleles strategy (2)</th>
<th>Distinct alleles strategy (3)</th>
<th>Distinct alleles strategy (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codominant: Order: ABC</td>
<td>1.12 1.27 1.20 2.00</td>
<td>1.00 1.00 1.41 2.00</td>
<td>1.29 1.59 1.52 2.00</td>
<td>1.00 1.00 1.64 2.00</td>
<td>1.29 3.77 Fixed Fixed</td>
<td>1.00 1.91 Fixed Fixed</td>
<td>1.61 3.70 Fixed Fixed</td>
<td></td>
</tr>
<tr>
<td>Order: ABC</td>
<td>1.04 1.11 1.16 1.95</td>
<td>1.00 1.00 1.00 1.00</td>
<td>1.07 1.15 1.37 1.92</td>
<td>1.00 1.00 1.00 1.00</td>
<td>Penetrance at B locus: 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: ABC</td>
<td>1.10 1.75 1.11 2.00</td>
<td>1.00 1.00 1.00 1.00</td>
<td>1.27 1.76 1.40 2.67</td>
<td>1.00 1.00 1.00 1.00</td>
<td>Penetrance at B locus: 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: ABC</td>
<td>1.26 1.38 1.28 1.42</td>
<td>1.00 1.00 1.00 1.00</td>
<td>1.19 1.28 1.25 1.54</td>
<td>1.00 1.00 1.00 1.00</td>
<td>Penetrance at B locus: 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: ABC</td>
<td>1.18 1.44 1.28 1.44</td>
<td>1.00 1.00 1.00 1.00</td>
<td>1.31 1.32 1.33 1.33</td>
<td>1.00 1.00 1.00 1.00</td>
<td>Penetrance at B locus: 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order: ABC</td>
<td>1.22 1.34 1.23 1.35</td>
<td>1.00 1.00 1.00 1.00</td>
<td>1.46 1.51 1.47 1.53</td>
<td>1.00 1.00 1.00 1.00</td>
<td>Penetrance at B locus: 0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Relative efficiency is measured by the ratio of the large sample variance of the two-locus and three-locus estimates. For the order ABC, θAC is assumed to be 0.05. Strategy (1): no assumption on interference; strategy (2): no interference; strategy (3): θAC fixed, strategy (4): θAC fixed and no interference.
especially when phase is known or can be inferred from the study of large families; three- to 10-fold increases in the efficiencies are found in the cases examined in table 1, with the largest gains observed for reduced penetrance. Again, the most substantial increase in efficiency is found with strategy (4).

Increased efficiency is also found when the disease locus flanks the two marker loci, that is, under the order A-C-B. This is readily understandable if we consider the limiting case $\theta_{AC} = 0$. Then the combined information from the A and C loci permits the definition of parental haplotypes, increasing the number of informative matings.

Similar gains in efficiency are found for other values of recombination within the range 0.05–0.15. When the flanking loci are not symmetrical with respect to the central locus, the gain in efficiency is greatest for the segment with the largest recombination rate.

**Interference**

The results of the previous section pertain to the situation of no interference, that is, where the true value of the coincidence is 1. In view of the accumulated evidence for some positive interference in laboratory animals, the relevance of these results remains to be established in the present context. A test for the presence of interference, based on a normal approximation to the estimate of the coincidence coefficient, has been discussed by Bailey [8] and Elandt-Johnson [9]. However, as shown by Ott [6], to obtain substantial power in the range of recombination amenable to analysis in humans, much larger samples are required than presently can be achieved.

Moreover, we have seen that without interference, more efficient estimates are obtained by assuming independence of recombination in adjacent segments. If this assumption is made when interference actually exists, the recombination estimates can be biased, but even with substantial sample sizes, they may have smaller mean-square error (i.e., expected variation about the true values) than the estimates obtained with the coincidence estimated simultaneously. In this case, the bias is offset by the smaller variance, and the former estimates may be preferred because of the increased precision in determining genetic locations.

We have examined this possibility for the codominant distinct allele case in phase-known and phase-unknown families, with interference at the Kosambi level. Table 2 shows the number of offspring observations necessary to achieve a power of 80% in a test of size 5% of the hypothesis $c = 1$, the bias in the estimate of the recombination between flanking markers under the assumption of no interference, and the number of offspring observations needed to obtain equal accuracy under strategies (1) and (2) (i.e., without and with the assumption of no interference). For sample sizes below the latter number, the mean-square error for the estimate of recombination between the flanking markers is smaller when no interference is assumed. As discussed in APPENDIX B, where the method of calculation is given, estimates of recombination in the adjacent segments are unbiased in this case and have the same variance under strategies (1) and (2).
MULTILOCUS LINKAGE ANALYSIS

TABLE 2
BIAS AND NO. OFFSPRING OBSERVATIONS, n, REQUIRED TO TEST FOR INTERFERENCE AND TO OBTAIN EQUAL MEAN-SQUARE ERRORS OF THE ESTIMATE OF θ_{AC} UNDER STRATEGY (1) (NO ASSUMPTION ON INTERFERENCE) AND STRATEGY (2) (NO INTERFERENCE)

<table>
<thead>
<tr>
<th>Recombination Value θ_{AB} = θ_{BC}</th>
<th>n for Power of 0.8</th>
<th></th>
<th>n for Equal Mean-Square Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phase known</td>
<td>Phase unknown</td>
<td>Bias (%) of True Value</td>
</tr>
<tr>
<td>0.05       .</td>
<td>1,163</td>
<td>3,872</td>
<td>4.4%</td>
</tr>
<tr>
<td>0.10       .</td>
<td>507</td>
<td>2,056</td>
<td>6.4%</td>
</tr>
<tr>
<td>0.15       .</td>
<td>423</td>
<td>2,334</td>
<td>7.3%</td>
</tr>
<tr>
<td>0.20       .</td>
<td>475</td>
<td>4,024</td>
<td>7.1%</td>
</tr>
<tr>
<td>0.30       .</td>
<td>1,147</td>
<td>38,711</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

NOTE: The true recombination values are calculated with Kosambi level interference. Bias and equal mean-square error sample sizes, n, are reported for θ_{AC}.

As shown in table 2, under no interference, the bias of the estimate of the recombination between flanking markers is always less than 10% of the true value. In phase-known families, more than 400 offspring are required to detect interference in the best case (θ = 0.15); in phase-unknown families, more than 1,000 offspring are required (θ = 0.1). For smaller values of recombination, the required samples sizes are still larger. Similarly, for θ < 0.1, which is the range of primary interest in human linkage studies, the estimates obtained under the assumption of no interference have smaller mean-square error than unrestricted estimates for less than 500 offspring in phase-known families and 900 offspring in phase-unknown families.

Since the situation examined here is likely to be the most informative for the estimate of interference, larger samples will be required in most circumstances to obtain accurate inference without restrictions on interference. In addition, an accurate assessment of interference will be increasingly difficult if interference is less than predicted by the Kosambi function.

Comparing Gene Orders

For a fixed gene order, maximum likelihood estimates of recombination and interference can be obtained with c constrained to lie between 0 and 1 (assuming interference cannot be negative). The evidence in favor of a particular order over an alternative can be assessed using the relative odds, that is, the ratio of the probabilities at the two maxima.

As discussed above, the unconstrained maximum likelihood gene order is chosen so that the estimated coincidence is between 0 and 1. In practice, this estimate is so imprecise that the likelihood is very close to the maximum if c is fixed at 1. Under the alternate orders, the likelihood equations usually have a solution at the upper bound. Generally, therefore, the support for the various gene orders can be assessed by comparing the maximum likelihood estimates under the assumption of c = 1, that is, no interference. For gene orders other than the maximum likelihood order, the likelihood may change considerably
when \( c \) varies between 0 and 1; this is because the likelihood function is not near the global maximum. Consequently, the assumption of a mapping function with coincidence different from 1 may inflate the apparent relative odds in favor of the maximum likelihood gene order. The assumption of \( c = 1 \) is thus conservative. This will be illustrated by the first example below. Of course, this problem does not occur when using a parametric family of mapping functions that includes the case of no interference (e.g., \([13, 14]\)).

**STRATEGIES OF MULTILOCUS LINKAGE ANALYSIS**

Multilocus linkage analysis (i.e., more than three loci) is likely to be more efficient than three-locus analysis in estimating recombination for the following reasons. First, as in the case of three-locus analysis, some families or parts of pedigrees will be informative for different subsets of loci; multilocus analysis allows the cumulation of information on genetic locations from all available data while correctly accounting for nonindependence. Second, our results for three-locus linkage analysis shows that the efficiency of the recombination estimates is often increased relative to two-locus analysis even when codominant marker loci lie on the same side of a disease locus, as would be the case in studying a disease in relation to three or more marker loci. Moreover, the study of several markers on each side of the disease locus will increase the number of families that are informative for multilocus linkage involving at least two flanking markers.

Multilocus linkage analysis requires the calculation of the theoretical frequencies of all possible recombination classes; for \( n \) loci, there are \( 2^n - 1 \) such classes. Therefore, some simplifying assumptions will be needed to reduce the number of parameters. As discussed in Bailey \([8]\), most mapping functions do not allow calculation of recombination class frequencies. Assumptions on interference mechanisms, chiasma formation, or limits on the number of crossing-overs considered may be used to reduce the complexity of the problem \([5, 17-21]\). However, our results indicate that the assumption of no interference may be warranted for the methodology currently applied in humans. We have proposed earlier \([16]\) a method of location scores for the detection of linkage between a locus and a known linkage group that involves only one free parameter: the genetic location of this locus relative to the known genetic map.

**EXAMPLES**

**Example 1**

Antonarakis et al. \([4]\) report restriction fragment length polymorphisms detected with probes for the \( \beta \)-globin, insulin, and parathyroid (PTH) hormone genes on the short arm of chromosome 11 in 17 nuclear families and one family with four grandparents. Fourteen of these families are informative for two or three of the recombination parameters. Because of the simple family structure and codominance at each locus, exact analytic equations for the likelihood were obtained and agreed with the numerical evaluation using the LINKAGE programs.
With two-locus analysis of the data, we obtain the same estimates of recombination and lod scores as reported in [4]. The recombination estimates are 0.07 between PTH and β-globin, 0.11 between insulin and β-globin, and 0.25 between PTH and insulin. Three-point linkage analysis using the parametrization (1) gives the same estimates of recombination for PTH–β-globin and insulin–β-globin; for PTH-insulin, the estimate is 0.18. Therefore, the most likely order of the genes is PTH–β-globin–insulin. There is no evidence for interference under this order ($\chi^2 = 0.35$). If we assume $c = 1$, the estimates of PTH–β-globin and the β-globin–insulin recombination rates are unchanged, but the estimate between the flanking loci becomes 0.17.

With the coefficient of coincidence restricted to the range 0 to 1, the relative odds for the three orders—PTH–β-globin–insulin, insulin–β-globin–PTH, and insulin–PTH–β-globin—at the maximum likelihood solutions are 140:4:1. The estimate of coincidence is $c = 0$ under the first order and $c = 1$ under the other two. If no interference is assumed for all three orders, the relative odds become 118:4:1. In contrast, the relative odds reported in [4] for the same set of data are 891:2:1. The reason for this difference is unknown.

Assuming a specific interference level other than $c = 1$ has the effect of increasing the apparent relative odds for the order PTH–β-globin–insulin. Kosambi level interference gives odds of 505:4:1. Similarly, the assumption of complete interference leads to odds of 3,498:3:1. These results illustrate the importance of assuming either no interference or estimating coincidence when comparing the likelihoods of various gene orders.

**Example 2**

Analysis of the pedigree shown in figure 1 illustrates the difference that may arise using two- and three-locus recombination estimates. Two codominant

---

**Fig. 1.**—Pedigree and marker data used for the analysis of linkage in example 2
alleles are assumed at each of the three loci. The grandparental phenotypes determine the phase of the alleles in the parents (individuals 5 and 6).

We first consider the two-locus analyses. Since the mother is a homozygote at the B locus, only the chromosome inherited from the father can be informative for linkage with this locus. For loci A and B, the children 7, 8, and 9 are noninformative. Among the remaining children, 11 and 12 have received nonrecombined chromosomes from the father, while 10 is a recombinant. Therefore, the likelihood is proportional to \( \theta_{AB}(1 - \theta_{AB})^2 \) and \( \hat{\theta}_{AB} = 1/3 \).

For loci B and C, the only noninformative child is individual 12. The remaining children have received nonrecombined chromosomes from the father; hence, the likelihood is proportional to \((1 - \theta_{BC})^5 \) and \( \theta_{BC} = 0 \).

The analysis for loci A and C is more complicated since both parents are informative for linkage. Four children (7, 8, 9, and 12) have received either two nonrecombined chromosomes or two recombined chromosomes. Each of these children contributes a term proportional to \( \theta_{AC}^2 + (1 - \theta_{AC})^2 \) in the likelihood. The children 10 and 11 have each received one recombed and one nonrecombined chromosome; they contribute terms proportional to \( \theta_{AC}(1 - \theta_{AC}) \). Hence, the likelihood for all data is proportional to: \( L_{AC} = [\theta_{AC}^2 + (1 - \theta_{AC})^2]^4 \). Numerical maximization of this equation gives \( \hat{\theta}_{AC} = 0.21 \).

For the three-locus linkage analysis, we make no assumptions on interference and the order of the loci. Table 3 shows the terms contributed to the likelihood for each child in the three-locus analysis. The total likelihood is proportional to: \( \theta_{ABC} = [p_2 (p_2 + p_3) + p_4 (p_1 + p_4)]^3 p_2 (p_1 + p_4) p_4 (p_2 + p_3) [p_3 (p_2 + p_3) + p_4 (p_1 + p_4)] \). Estimates of the parameters are obtained by numerical maximization of this equation. This gives \( \hat{\rho}_1 = 0.0, \hat{\rho}_2 = 0.0, \hat{\rho}_3 = 0.19, \) and \( \hat{\rho}_4 = 0.81 \). Applying equation (1), we obtain \( \hat{\theta}_{AB} = \hat{\theta}_{AC} = 0.19 \) and \( \hat{\theta}_{BC} = 0.0 \). Note that the three-locus estimates of both \( \theta_{AB} \) and \( \theta_{AC} \) are different from the two-locus estimates. The essential point, however, is that individuals 7, 8, and 9 are informative for linkage of A with B only when all loci are considered simultaneously.

Now consider the test for linkage of A with B or C. The two-point tests are made by calculating the odds in favor of linkage, that is, the ratio of the

<table>
<thead>
<tr>
<th>Individuals</th>
<th>Phase</th>
<th>Likelihood contribution</th>
<th>Phase</th>
<th>Likelihood contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>7, 8, and 9</td>
<td>AbCaBC</td>
<td>( p_2(p_2 + p_3) )</td>
<td>abCaBC</td>
<td>( p_4(p_1 + p_4) )</td>
</tr>
<tr>
<td>10</td>
<td>AbCaBC</td>
<td>( p_2(p_1 + p_4) )</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>11</td>
<td>abCaBC</td>
<td>( p_4(p_2 + p_3) )</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>12</td>
<td>abc/aBC</td>
<td>( p_3(p_2 + p_3) )</td>
<td>abCaBC</td>
<td>( p_4(p_1 + p_4) )</td>
</tr>
</tbody>
</table>

*Note:* When two phases are possible, the likelihood term is the sum of the two contributions.
MULTILOCUS LINKAGE ANALYSIS

likelihood at the estimated recombination value to the likelihood under the assumption of free recombination ($\theta = 0.5$). In this case, the odds in favor of linkage are 1.184:1 (lod score of 0.074) for $A-B$ and 1.402:1 (lod score of 0.227) for $A-C$. To perform a similar test for linkage of $A$ with the group $B-C$ using three-point data, we first note that under the assumption of free recombination between $A$ and $B$ or $C$, $\hat{\Theta}_{BC}$ is equal to its value from the two-locus estimate. Therefore, in this example, $\hat{\Theta}_{BC}$ is unchanged by the constraint $\theta_{AB} = \theta_{AC} = 0.5$. Under this constraint, we obtain $\hat{\rho}_1 = \hat{\rho}_3 = 0.0$ and $\hat{\rho}_2 = \hat{\rho}_4 = 0.5$. The odds in favor of linkage of $A$ to the group $B-C$ is the ratio of the unconstrained and constrained likelihood, or 2.239:1 (lod score of 0.420).

Although the three-point test gives higher odds in favor of linkage, the likelihood ratio test statistic, twice the natural logarithm of the odds, is approximately distributed as a chi-square variate with 2 degrees of freedom (df) for the three-point test and 1 df for the two-point tests. To provide a comparison of the tests, we use the fact that the likelihood-ratio statistics are additive over independent families. This allows us to calculate the number of families of this form that would be needed to obtain a test significant at a given level; this number is obtained by dividing the corresponding critical value of the chi-square (with appropriate degrees of freedom) by the value of the likelihood-ratio test statistic. The likelihood-ratio statistic for the three-point test is 1.932. For the two-point tests, the likelihood-ratio statistics are 0.338 for $A-B$ and 0.677 for $A-C$. Using the 0.05 critical values of the chi-square distribution, we find that approximately three families would be required for a significant three-point test and 11 or six families, respectively, for a significant two-point test. Thus, in this example, the three-point test improves the detection of linkage between loci $A$ and $B$ or $C$ by a factor of two or more. Note that if we were to assume no interference, the three-point test statistic would have 1 df and the improvement would be correspondingly greater.

DISCUSSION

We have investigated the relative efficiency of three-locus vs. two-locus linkage analysis to delineate circumstances where multilocus analysis in humans is indicated. Multilocus linkage analysis raises the issue of interference in crossing-over even though this is not the primary concern in establishing the genetic map. For practical relevance, limitations in both experimental control and feasible sample sizes characterizing human linkage studies must be taken into account.

Our results suggest that an hypothesis of no interference in crossing-over may be satisfactory for most linkage analyses in humans. This conclusion should not be interpreted as denying the existence of interference. Extensive empirical evidence documents this phenomenon in laboratory animals [8]; however, statistical significance is reached in such studies through controlled crosses and very large samples.

Our willingness to entertain a hypothesis of no interference is based on the following results. Studies of the power of the classical test for interference show that even in the best of cases large sample sizes are required to detect
interference operating at the Kosambi level. The best case reported in table 2 requires 423 offspring from phase-known families with four distinct parental alleles at each of three loci when recombination between flanking markers is 0.28. Note that if interference could be detected at this distance, the recombination rate between flanking markers would be of little significance for the genetic map. For smaller recombination rates, and more realistic study designs, the detection of interference is probably not feasible. Moreover, we find that even for flanking markers the recombination estimate has little bias, and often smaller mean-square error, when obtained under the assumption of no interference. Finally, while it is possible to assume a mapping function implying interference in three-point linkage analysis, this may inflate support for one gene order over alternatives.

We have investigated the relative efficiency of three point vs. two-point linkage tests for simple situations readily amenable to calculation of amounts of information by analytic or numerical methods. Although not exhaustive, our investigation gives bounds on the gains in efficiency that can be anticipated in some typical situations. The efficiency of the recombination estimates can be substantially increased when the number of alleles at the test loci is small and heterozygosity is moderate. This gain is greater when assuming no interference. The most significant increase in efficiency can be expected when analyzing a rare recessive condition as information on phase increases. For codominant systems, the rare dominant disease traits, there is less incentive to estimate recombination from three-point tests when the heterozygosities at the test loci are large.

Aside from the calculation of genetic risk, it is probably for detection of linkage that multilocus analysis has its greatest appeal. This is supported by our investigation of efficiency when assuming partial knowledge of a linkage map [strategies (3) and (4) in table 1]. Our results emphasize the value of establishing genetic maps of linked marker loci for the efficiency analysis of genetic disease [1]. Once such a map is determined to a sufficient accuracy, the detection of linkage between a disease locus, or a new marker, and various linkage groups is possible by the method of location scores proposed earlier [15]. This analytic strategy will become feasible with the current effort being undertaken to stimulate collaboration in human gene mapping through the availability of a common panel of test families from the Human Polymorphism Study Center (Pr. Dausset, Hôpital St. Louis, Paris, France).

Finally, multipoint tests allow evaluation of the relative likelihood of alternate gene orders. As the genetic map develops, such tests will be essential to resolve possible inconsistencies exhibited by classical two-point tests. Only practical experience will permit an assessment of their usefulness in constructing the human gene map.

APPENDIX A
THE COMPUTER PROGRAM PACKAGE LINKAGE

LINKAGE is a series of PASCAL programs for multilocus linkage analysis, whether detection of linkage, estimation of recombination, or calculation of genetic risk. An arbitrary
MULTILOCUS LINKAGE ANALYSIS

number of loci can be considered. At each locus, the phenotype can be defined as: (1) a simple Mendelian trait treated as a factor-union system [22], appropriate for classical markers and DNA polymorphisms; (2) a classification of disease status with variable penetrances specified for different classes of risk; or (3) one or several quantitative measurements with a multivariate normal distribution of errors. A special provision has been made to allow a mixture of affection status and quantitative measurements for sex-linked traits. Allowance is made for linkage disequilibrium and mutation. Pedigree data are treated with a recursive algorithm [23–25] allowing up to one inbreeding loop [25]. Recursive formulation also allows the treatment of multiple loci such that the number of loci that can be considered on any given problem depends only on limits imposed by the computer used. Further details on the implementation are given in [16].

The programs perform either likelihood calculations or the computation of genetic risks for specified individuals. Likelihood calculations can be made for incremental values of one recombination rate with other rates fixed, giving either a lod-score table for the case of two loci or a location-score graph for multiple loci. Alternatively, maximum likelihood estimation of recombination rates, and possibly other parameters, can be performed iteratively. In the three-locus case, estimates can be obtained without constraint on interference or by assuming one of several mapping functions. When more than three loci are considered, calculations are carried out under the assumption of no interference for specified orders of the loci. When a location score is calculated, the program automatically varies the location of the main locus with respect to the linkage group against which it is tested, and a curve is generated.

The linkage programs are available from the authors upon request.

APPENDIX B

STATISTICAL CONSIDERATIONS

The large sample covariance matrix of the maximum likelihood estimates of recombination is the inverse of the information matrix I defined by

\[
I_{ij} = \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 E}{\partial \theta_i \partial \theta_j},
\]

where \( \theta_i \) and \( \theta_j \) are recombination parameters to be estimated, \( L \) is the likelihood, and the expectation is evaluated at the true recombination rates. The second equality follows from regularity of the likelihood function as defined in ([26] p. 500). The expectation of the logarithm of the likelihood was calculated by summing the likelihood over all possible genotype configurations compatible with the family and size and weighting by the probability of the configuration given the gene frequencies, the sampling scheme, and the true values of the recombination rates. Numerical approximations to the second derivatives were calculated by finite difference approximation.

To obtain the results in table 2, we first note that under the distinct allele model with codominant loci the likelihood factors into contributions from each triple backcross parent. Let

\[
q_1 = p_1^2 + p_2^2 + p_3^2 + p_4^2, \quad q_2 = 2(p_1p_2 + p_3p_4), \quad q_3 = 2(p_1p_4 + p_2p_3), \quad q_4 = 1 - q_1 - q_2 - q_3.
\]

Thus, \( q_1 \) is the probability that the maternal or paternal haplotypes in the two sibs are identical or differed simultaneously at all three loci. \( q_2 \) is the probability they are identical at locus A and B but different at locus C, or different at locus A and B but identical at locus C, etc. In families with two children, the likelihood is proportional to:

\[
L \sim q_1^{m_1} q_2^{m_2} q_3^{m_3} q_4^{m_4}
\]

and \( q_4 = 1 - q_1 - q_2 - q_3 \), where \( m_1, m_2, m_3, \) and \( m_4 \) are the number of observed sib-pairs in each of these classes. Each sib-pair is counted twice, once for the genes re-
ceived from the father and once for those received from the mother. Under strategy (1), the maximum likelihood equations, found by setting the partial derivatives of equation (A1) equal to 0, can be solved to give:

\[
\hat{\theta}_{AB} = \frac{1 - (2\hat{r} - 1)^2}{2}
\]
\[
\hat{\theta}_{BC} = \frac{1 - (2\hat{s} - 1)^2}{2}
\]
\[
\hat{\theta}_{AC} = \frac{1 - (2\hat{t} - 1)^2}{2}
\]

where \(\hat{r} = \hat{q}_1 + \hat{q}_2, \hat{s} = \hat{q}_1 + \hat{q}_4,\) and \(\hat{t} = \hat{q}_1 + \hat{q}_3.\) The variance for these estimates, derived using a large-sample approximation ([27], p. 246) are:

\[
\text{var}(\hat{\theta}_{AB}) = \frac{s(1 - s)}{4(2s - 1)}
\]
\[
\text{var}(\hat{\theta}_{BC}) = \frac{t(1 - t)}{4(2t - 1)}
\]
\[
\text{var}(\hat{\theta}_{AC}) = \frac{r(1 - r)}{4(2r - 1)}
\]

Under strategy (2), the estimates of \(\theta_{AB}\) and \(\theta_{BC}\) given the order ABC are unchanged, while the estimate of \(\theta_{AC}\) becomes \(\hat{\theta}_{AC} = (1 - \hat{\theta}_{AB})\hat{\theta}_{BC} + (1 - \hat{\theta}_{BC})\hat{\theta}_{AB},\) with large sample variance:

\[
\text{var}(\hat{\theta}_{AC}) = \frac{[(1 - q_1)q_1 - q_4q_1 - q_4q_3 - q_3q_1]/2 \sqrt{(2s - 1)(2t - 1)}}{2}
\]

In phase-known families, the likelihood is proportional to:

\[
L \sim p_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4},
\]

where \(n_i\) is the observed number in the \(i\)th recombination class. The estimates for strategy (1) are:

\[
\hat{\theta}_{AB} = \frac{n_1 + n_3}{n}
\]
\[
\hat{\theta}_{BC} = \frac{n_1 + n_2}{n}
\]
\[
\hat{\theta}_{AC} = \frac{n_2 + n_3}{n}
\]

where \(n = \Sigma n_i.\) For strategy (2), the estimates of \(\theta_{AB}\) and \(\theta_{BC}\) are unchanged, while \(\hat{\theta}_{AC} = (1 - \hat{\theta}_{AB})\hat{\theta}_{BC} + (1 - \hat{\theta}_{BC})\hat{\theta}_{AB}.\) The estimate for \(\theta_{AC}\) has variance

\[
\text{var}(\hat{\theta}_{AC}) = \frac{(1 - \theta_{AC})\theta_{AC}}{n}
\]

for strategy (1), and
MULTILOCUS LINKAGE ANALYSIS

\[ \text{var}(\theta_{AC}) = [(1 - \theta_{BC})^2(1 - \theta_{AC})\theta_{AC} + (1 - \theta_{AC})^2(1 - \theta_{BC})\theta_{BC} - (1 - \theta_{BC})(1 - \theta_{AC})(p_1 [1 - p_1] - p_1 p_3 - p_2 p_3 - p_1 p_2)]/n \]  

(A5)

for strategy (2).

The large sample expectations of \( \theta_{AB} \) and \( \theta_{BC} \) are the true values even when no interference is assumed. The large-sample bias for \( \theta_{AC} \) is just the difference between the true value with interference at the Kosambi level, and \( \hat{\theta}_{AC} = \hat{\theta}_{AB}(1 - \hat{\theta}_{BC}) + \hat{\theta}_{BC}(1 - \hat{\theta}_{AB}) \). The mean-square error for \( \theta_{AC} \) is then:

\[ \text{mse}(\hat{\theta}_{AC}) = \text{var}(\hat{\theta}_{AC}) \]

\[ + [\theta_{AC} - \theta_{AB}(1 - \theta_{BC}) - \theta_{BC}(1 - \theta_{AB})]^2 \]

(A6)

For small values of \( n \), the mean-square error for the estimate of recombination between flanking loci is smaller under strategy (2). To obtain the value of \( n \) for which the mean-square errors are equal, \( \text{var}(\theta_{AC}) \) in equation (A2) or (A4) is substituted into equation (A6), and the result equated to \( \text{var}(\theta_{AC}) \) from equation (A3) or (A5); we then solve for \( n \).

The test statistic for \( c = 1.0 \) is based on the asymptotic normal approximation to the maximum likelihood estimate \( \hat{c} = (\hat{\theta}_{AB} + \hat{\theta}_{BC} - \hat{\theta}_{AC})/2\hat{\theta}_{AB}\hat{\theta}_{BC} \) for the gene order ABC [9, 10]. The formula for the asymptotic variance of \( \hat{c} \), denoted by \( V_c \), is given in Bailey [8] p. 149). Approximate sample sizes required to achieve a power of \( \beta \) in a test of size \( \alpha \) are

\[ n = (Z_{1-\alpha}V_1 + Z_{\beta}V_L)^2/(c - 1)^2 \]

where \( V_1 \) is the variance of \( \hat{c} \) at \( c = 1 \) and \( Z_{1-\alpha}(Z_{\beta}) \) is the \( 1 - \alpha(\beta) \) level of the standard normal distribution.

REFERENCES

The Southwest Foundation Forum will host an international symposium entitled GENETIC RESEARCH WITH NONHUMAN PRIMATES: SERVING THE NEEDS OF MANKIND, in San Antonio, Texas, March 2-5, 1986. Invited speakers will present papers in the areas of biochemical genetics, cytogenetics, immunogenetics, molecular genetics, population genetics, and genetic predisposition to common diseases. Contributed papers also will be accepted for presentation. A Distinguished Scientist Award in Genetics and a cash prize of $1,000 will be made to a geneticist who has already made significant contributions in health-related basic research and has demonstrated great potential for future achievement. The recipient, who need not necessarily have worked with nonhuman primates, will be invited to present a keynote address at the symposium. Please send requests for information and letters of nomination for the Distinguished Scientist Award to: John L. VandeBerg, Director, Department of Genetics, Southwest Foundation for Biomedical Research, P.O. Box 28147, San Antonio, TX 78284. Letters of nomination must be received by July 1, 1985.