Constrained maximum likelihood estimation for ordering three genetic loci with a small-sample robustness result for the no-interference model*

Mary Sara McPeek

TECHNICAL REPORT NO. 410

Department of Statistics
The University of Chicago
Chicago, Illinois 60637

April 1995

* This research was supported in part by NIH Grant R01-HG01093-01 and NSF Grant DMS 90-05833.

Constrained maximum likelihood estimation for ordering three genetic loci with a small-sample robustness result for the no-interference model

by

Mary Sara McPeek

Department of Statistics, University of Chicago,

Chicago, Illinois 60637

Running Head: Ordering three genetic loci

Corresponding author:

Mary Sara McPeek

address: The University of Chicago Department of Statistics.

5734 University Ave., Chicago, IL 60637

phone: (312) 702-7554

FAX: (312) 702-9810

Key words and phrases. linkage analysis: interference: genetic mapping: small-sample robustness: constrained maximum likelihood estimation.

AMS 1991 subject classifications. Primary 62P10; secondary 62F11.

Supported in part by NIH Grant R01-HG01093-01 and NSF Grant DMS 90-05833.

SUMMARY

Linkage analysis is a powerful technique by which genes may be mapped to chromosome regions. It is based on observation of the pattern of inheritance relative to known markers of the traits that the genes cause. Creation of the initial marker map is also by linkage analysis. Most linkage analyses rely on the assumption of no interference (NI), although this assumption is known to be grossly violated in nearly all organisms studied. We consider the much weaker assumption of no chromatid interference (NCI), and in the three-locus case, we characterize the maximum likelihood estimates of order and recombination probabilities under NCI. We show that in the case of three loci with complete recombination data, the estimation of their linear order along the chromosome by maximum likelihood under NI gives the same estimate of order as under the considerably more general assumption of NCI, in the case when the NI estimate is unique. When the NI estimate is not unique, the set of NI estimates contains the set of NCI estimates. Speed, McPeek, and Evans showed that estimation of order by maximum likelihood under NI with complete data is consistent for any number of loci, even when interference is present, as long as NCI holds. Here we establish a nonasymptotic result for ordering three loci under the false assumption of NI. We show that the result does not hold for four or more loci.

orders only. ?

1. Introduction.

1.1 Recombination and Interference. Genetic mapping involves ordering a set of genetic loci on a chromosome and finding genetic distances between them. One way this may be done is through analysis of data on meiotic recombination among the loci. A recombination between two loci is said to occur when the alleles or versions of the loci inherited by an offspring from its parent were inherited from different grandparents. Meiotic recombination between loci on the same chromosome is believed to be the result of crossing over between nonsister chromatids during the pachytene phase of meiosis. A recombination will occur if a particular chromatid passed on in meiosis was involved in an odd number of crossovers between two loci.

It is important to keep in mind that crossing over takes place in the four-stranded state, when each chromosome has duplicated to form two sister chromatids, and all four chromatids are lined up in a tight bundle. Crossovers occur along the four-strand bundle, with each crossover involving only two of the four chromatids, one from each of the sister pairs. We refer to the occurrence of crossovers along the bundle of four chromatids as the *chiasma process.* Any given chromatid will be involved in some subset of the crossovers of the full chiasma process. The occurrence of crossovers along a given chromatid will be referred to as the *crossover process.*

In linkage analysis, two assumptions are commonly made about the occurrence of crossovers. First, it is assumed that the chiasma process is a (possibly inhomogeneous) Poisson process. Violation of this assumption is known as *chiasma interference.* Second, it is assumed that each pair of non-sister chromatids is equally likely to be involved in a crossover, independent of which were involved in other crossovers. This assumption is equivalent to specifying that the crossover process is obtained from the chiasma process by independently deleting each point with chance 1/2. Violation of this assumption is known as *chromatid interference*, and the assumption itself is referred to as no chromatid interference (NCI). This pair of assumptions specifies a model for the occurrence of crossovers known as the *no-interference (NI) model.* Deviation from this model is known as interference, which encompasses both chiasma interference and chromatid interference.

The phenomenon of interference was first detected in *Drosophila* by Sturtevant (1915) and Muller (1916) and is well-documented in a wide range of organisms. Although it may be difficult to distinguish between chiasma and chromatid interference in many cases, the existence of chiasma interference has been well-established in experimental organisms (see e.g. Mortimer and Fogel 1974). There is little consistent evidence for strong chromatid interference (Zhao, McPeek, Speed 1995). In the most general model we consider, the *no-chromatid interference (NCI) model*, we permit an arbitrary amount of chiasma interference, but assume no chromatid interference.

1.2 Recombination Data. Complete recombination data, the type of data considered here, arises in breeding experiments with organisms such as fruit flies, mice, or tomatoes. In contrast, the rule in human studies is incomplete data, which are much more difficult to analyize and are not discussed here.

Suppose there are m loci under study. In complete recombination data, for each meiosis considered, the parent will have two distinguishable *alleles*, or differing versions of DNA, at each locus, one inherited from the grandmother and one from the grandfather. The offspring will inherit from the parent one allele at each locus. Some of these alleles will be of grandmaternal and some of grandpaternal origin, due to recombination. In complete recombination data, there are 2^m possible outcomes of meiosis, corresponding to grandmaternal or grandpaternal inheritance at each of the m loci. Note that each outcome has a complementary outcome which is assumed equally probable and therefore equivalent in terms of recombination, namely the one in which all the grandmaternal and grandpaternal alleles are switched. Therefore, there are 2^{m-1} possible recombination outcomes, and the number of times each occurs out of a total of n meioses would be recorded. Each possible recombination event is assumed to occur with fixed probability, independently across meioses. Thus, complete recombination data are multinomial.

If the m loci are ordered on a chromosome, we let I_j denote the interval between loci j and $j+1$. We let $x = (x_1, ..., x_{m-1})$, $x_j = 0$ or 1, $j = 1, ..., m-1$, denote the event of a recombination in each interval I_j for which $x_j = 1$, and no recombination in each interval I_j for which $x_j = 0$. Then each of the 2^{m-1} possible recombination outcomes would correspond to one of the 2^{m-1} possible x's. Changing the order of the loci permutes the x's relative to the observed recombination outcomes.

1.3 The No-Chromatid Interference Model. The NCI model is a multinomial model in which the multinomial probabilities satisfy a set of constraints that depends on the order of the loci. Assume an order for the loci. Recall that event x occurs when the crossover process of a given chromatid results in an odd number of crossovers in each of the intervals I_j for which $x_j = 1$ and an even number of crossovers in each of the intervals I_j for which $x_j = 0$. A set of related events will be denoted $y = (y_1, ..., y_{m-1})$, $y_j = 0$ or 1, $j = 1, ..., m-1$. y is the event that the chiasma process has at least one crossover in each of the intervals I_j for which $y_j = 1$ and no crossovers in each of the intervals I_j for which $y_j = 0$. We let p_x denote the probability of the event x and q_y denote the probability of the event y. The assumption of NCI gives a correspondence between these two sets of probabilities (Speed, McPeek, and Evans 1992; see also Weinstein 1936), namely

$$(1) \qquad p_x = \sum_{y:\, 1 \geq y \geq x} \frac{1}{2^{y \cdot 1}} q_y \qquad \text{for all } x$$

and inverting,

$$q_y = 2^{y \cdot 1} \times \sum_{x:\, 1 \geq x \geq y} (-1)^{(x-y) \cdot 1} p_x \qquad \text{for all } y,$$

where, for example, $y \cdot 1 = \sum_{j=1}^{m-1} y_j$, and $1 \geq y \geq x$ means $1 \geq y_j \geq x_j$ for all j. Since the q_y's must all be nonnegative, we have the constraints

$$\sum_{x:\, \mathbf{1} \geq x \geq y} (-1)^{(x-y)\cdot \mathbf{1}} p_x \geq 0 \qquad \text{for all } y.$$

Speed, McPeek, and Evans (1992) showed that these constraints on the multinomial p's are necessary and sufficient for them to be compatible with at least one simple point process model for the chiasma process under the assumption of NCI. Thus, they characterize the general NCI model. In the language of the point process literature, the NCI constraints are equivalent to the requirement that the avoidance function for the chiasma point process be completely monotone on the set of intervals with endpoints in the set of m loci (see, e.g. Daley and Vere-Jones (1988) pp. 215-219).

Maximum likelihood estimation of multinomial probabilities under quasi-order restrictions is known to be solved by isotonic regression with equal weights, that is, the constrained maximum likelihood estimate is the closest point, in terms of Euclidean distance, in constraint space to the unconstrained estimate (see e.g. Barlow et al. 1972). Similarly, maximum likelihood estimation of normal means under homogeneous linear constraints is also solved by isotonic regression with equal weights (Raubertas and Lee 1986). In the case of the NCI constraints, since the p_x's sum to 1, elimination of the redundant parameter results in a set of inhomogeneous inequalities, to which the above literature does not apply. In fact, isotonic regression with equal weights does not yield the maximum likelihood estimate (MLE) in this case. The constrained MLEs of the p_x's may be easily approximated numerically by parametrizing the likelihood in terms of the q_y's, thinking of the knowledge of the observed events x as being incomplete data, the knowledge of the unobserved events y as being complete data, and applying the EM algorithm of Dempster, Laird, and Rubin (1977). In what follows, exact calculations of the MLEs are made for the case $m = 3$.

Note that under the assumption of NCI, the chance of recombination across an interval increases monotonically as the interval is enlarged, with an upper bound of $\frac{1}{2}$. This is clear from the two-locus version of (1), due to Mather (1935): $p_1 = .5q_1$. Here, q_1 must increase as the interval is widened, and q_1 has an upper bound of 1. It is this property of the recombination probability that underlies genetic mapping: if chance of recombination increases with distance, then recombination data provide information on order and distances between markers.

1.4 The No-Interference Model. The NI model is a special case of NCI in which the chiasma process is assumed to be Poisson (not necessarily homogeneous). The parameters in the model would then be the nonhomogeneous Poisson intensity function λ_t and the positions of the loci, $L_1, ..., L_m$, all relative to some origin. These are not actually identifiable. One conventional set of identifiable parameters is $d_1, ..., d_{m-1}$ where

$$d_i = .5 \times \int_{L_i}^{L_{i+1}} \lambda_t dt.$$

Known as the *genetic distance* associated with interval I_i, d_i is the expected number of crossovers in the crossover process in interval I_i. The other conventional set of identifiable parameters is $\theta_1, ..., \theta_{m-1}$, where $\theta_i = (1 - \exp(-2d_i))/2$ is the chance of recombination in interval I_i. For the NI model,

$$p_x = \prod_i \theta_i^{x_i}(1 - \theta_i)^{1-x_i}.$$

1.5 Estimation. For either model, order is estimated by maximizing the likelihood under each order and choosing the order whose maximized likelihood is highest. In terms of the p_x's, the log-likelihood for a particular order is, up to an additive constant, $\sum_x a_x log(p_x)$, where $x = (x_1, ..., x_{m-1})$, $x_i = 0$ or 1, and a_x is the number of times the event x occurs in the data out of a total of n meioses. Recall that the vector x associated with a given observed recombination event changes depending on the order assumed, and so the probabilities associated with the observed cell counts are subject to different constraints under different orders. The log-likelihood would then be maximized over the unknown parameters subject to the NCI constraints imposed under the assumed order. In the case of the NI model, the MLEs of the unknown parameters under a given order can easily be written down explicitly, namely

$$\hat{\theta}_i = \min(\sum_{x:x_i=1} a_x/n, 1/2).$$

In what follows, we first derive explicitly the MLEs of the p_x's in the NCI model under any given order when the number of loci $m = 3$. We then compare the maximized likelihoods under the three different orders to obtain the MLE (or MLEs) of order under the NCI model. Finally, we compare these order estimates to those that we obtain for the much simpler NI model. The NI model is known to be consistent for the estimation of order by maximum likelihood even when an arbitrary amount of interference is actually present, under the assumption of NCI (Speed, McPeek, and Evans 1992). Our complementary non-asymptotic result is that in the three-locus case, the MLE of order under the NI model is actually identical to that under the NCI model when there is a finite amount of data and when the MLE is unique. When the MLE is not unique, the set of maximum likelihood orders under the NI model contains the set of maximum likelihood orders under the NCI model. We give a counterexample for the case $m = 4$.

2. Maximum Likelihood Estimation of Recombination Parameters for Three-Locus NCI Model. Suppose we have three marker loci, A, B, and C. Assume we observe the recombination outcomes of n meioses, which can be summarized as in TABLE 1, with $a+b+c+d = n > 0$.

[Insert TABLE 1 about here]

Let the putative order of the three loci be A-B-C. Under the assumption of NCI, we derive the MLEs of the recombination parameters $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, where p_{ij} is the probability of i recombinations in interval A-B and j recombinations in interval B-C in a given meiosis. Under NCI, (a, b, c, d) are multinomial(n.p), with $\mathbf{p}$ satisfying the following constraints in addition to the constraint that they sum to 1:

1. $0 \leq p_{11}$
2. $p_{11} \leq \min(p_{10}, p_{01})$
3. $p_{10} + p_{01} \leq 1/2$.

Note that these conditions imply $\max(p_{10}, p_{01}) \leq p_{00}$.

LEMMA 1. Assume $a + d > 0$ and $b + c > 0$ (see REMARKS below). Then the MLE of $(p_{11}, p_{10}, p_{01}, p_{00})$ under the NCI model, when the data are as given in TABLE 1, with putative order A-B-C is as follows:

1. If $a \leq b$ and $a \leq c$ and $b + c \leq n/2$, then the MLE $(\hat{p}_{11}, \hat{p}_{10}, \hat{p}_{01}, \hat{p}_{00}) = (a, b, c, d) \times n^{-1}$
2. If $b + c > n/2$, then the MLE lies in the region $(p_{10} + p_{01} = 1/2)$:
   (a) if also $ab \leq cd$ and $ac \leq bd$, then the MLE is $(a/(a+d), b/(b+c), c/(b+c), d/(a+d)) \times 2^{-1}$.
   (b) if also $b \leq c$ and $ac > bd$ and $a+b \leq c+d$, then the MLE is $(a+b, a+b, c+d, c+d) \times (2n)^{-1}$.
   (c) if also $c \leq b$ and $ab > cd$ and $a+c \leq b+d$, then the MLE is $(a+c, b+d, a+c, b+d) \times (2n)^{-1}$.

(d) if also $ac > bd$ and $ab > cd$ and $a + b > c + d$ and $a + c > b + d$, then the MLE is $(.25, .25, .25, .25)$.

3. If $b < a$ and $b \leq c$ and $b + c \leq n/2$, then the MLE lies in the region $(p_{11} = p_{10})$:

(a) if also $c \leq d$ and $a + b \leq 2c$, then the MLE is $((a+b)/2, (a+b)/2, c, d) \times n^{-1}$.

(b) if also $a + b > 2c$ and $a + b + c \leq 3n/4$, then the MLE is $((a+b+c)/3, (a+b+c)/3, (a+b+c)/3, d) \times n^{-1}$.

(c) if also $c > d$ and $a + b \leq c + d$, then the MLE is $(a + b, a + b, c + d, c + d) \times (2n)^{-1}$.

(d) if also $[c > d$ or $a + b > 2c]$ and $a + b + c > 3n/4$ and $a + b > c + d$, then the MLE is $(.25, .25, .25, .25)$.

4. The solution for $c < a$ and $c \leq b$ and $b + c \leq n/2$ is similar. Interchange b and c and p_{10} and p_{01} in case 3.

PROOF. See appendix A1.

REMARKS. If $a + d = 0$, then the MLEs of p_{10} and p_{01} are $b/(2n)$ and $c/(2n)$, respectively, while p_{11} and p_{00} do not have unique MLEs. Likewise, if $b + c = 0$, then the MLEs of p_{11} and p_{00} are $a/(2n)$ and $d/(2n)$, respectively, while p_{10} and p_{01} do not have unique MLEs.

Let $C \subset I\!R^4$ be the three-dimensional solid determined by the NCI constraints and by the constraint that the p's sum to 1. If $(a/n, b/n, c/n, d/n) \notin C$, then the MLE over C must lie on the boundary of C. This is so because the log-likelihood is a smooth function of the parameters whose gradient is zero only at $(a/n, b/n, c/n, d/n)$, and C is a compact set in R^4. We now describe the boundary of C and how we maximize the likelihood over it.

Let $G_0 \subset I\!R^4$ be the region determined by the constraints ($p_{11} = 0$ and p's nonnegative and sum to 1), let G_1 be the region determined by ($p_{10} + p_{01} = 1/2$ and p's nonnegative and sum to 1), let G_2 be the region determined by ($p_{11} = p_{10}$ and p's nonnegative and sum to 1), and let G_3 be the region determined by ($p_{11} = p_{01}$ and p's nonnegative and sum to 1). For $i = 0, 1, 2, 3$, let $A_i = G_i \cap C$. Then A_0, A_1, A_2, A_3 are polygons in $I\!R^4$ and $A_0 \cup A_1 \cup A_2 \cup A_3 = \partial(C)$ (the boundary of C). We shall not be concerned with A_0, because if the MLE over C lies in $A_0 \setminus A_1$ ($= A_0 \cap A_1^c$), then it must be the unconstrained MLE $(a/n, b/n, c/n, d/n)$. If the MLE over C lies in $A_0 \cap A_1$, then we must have $a = 0$, and it can be seen from TABLE 2 row 2 that in this case, the MLE over C is the MLE over G_1. Thus A_0 need not be considered when we maximize the likelihood over the boundary of C, as long as we are already considering the unconstrained MLE and the MLE over G_1.

Note that one way to find the MLE over C in case $(a/n, b/n, c/n, d/n) \notin C$ is to let g_1 be the MLE over the region G_1, g_2 the MLE over G_2, g_3 the MLE over G_3, g_4 the MLE over $G_4 = G_2 \cap G_3$, g_5 the MLE over $G_5 = G_1 \cap G_2$, g_6 the MLE over $G_6 = G_1 \cap G_3$, and finally, let $g_7 = (.25, .25, .25, .25) = G_1 \cap G_2 \cap G_3$. Then determine which of these seven points gives the highest likelihood, among the ones that lie in C. Of course, whenever possible, one uses the fact that the maximized likelihood over X dominates the maximized likelihood over Y, when $Y \subset X$. Note that if $g_i \in C$, then g_i is the MLE of $\mathbf{p}$ over A_i, for $i = 1, 2, 3$. If $g_4 \in C$, then g_4 is the MLE of $\mathbf{p}$ over $A_2 \cap A_3$. Similarly, if $g_5 \in C$, then g_5 is the MLE of $\mathbf{p}$ over $A_1 \cap A_2$, if $g_6 \in C$, then g_6 is the MLE of $\mathbf{p}$ over $A_1 \cap A_3$, and g_7 is always the MLE of $\mathbf{p}$ over $A_1 \cap A_2 \cap A_3 = g_7$.

The MLE on A_2 is the closest point on A_2 to $(a/n, b/n, c/n, d/n)$ in Euclidean distance. Similarly for A_3. This is, in general, false on A_1 (Counterexample: $a = 1$, $b = c = d = 2$. MLE on A_1 is $(1/6, 1/4, 1/4, 1/3)$. Closest point on A_1 is $(5/28, 1/4, 1/4, 9/28)$.), and hence false on C since the same counterexample holds there. Note that it is true on $B \subset I\!R^4$, $B \supset C$ defined by $(0 \le p_{11} \le \min(p_{10}, p_{01})$ and p's sum to 1). This property of A_2, A_3, and B follows from Example 2.1 on pp. 65-66 of Barlow et al. (1972), in which it is shown that the MLEs for a multinomial in which the multinomial parameters are subject to partial ordering constraints are obtained by isotonic regression with equal weights. That is, the MLE is the closest point in constraint space to the point representing the observed frequencies $(a/n, b/n, c/n, d/n)$. This result is used in the proof. It does not apply to the regions C and A_1 because they have the additional constraint $p_{10} + p_{01} \le 1/2$ that does not correspond to a partial ordering constraint. In fact, we find that the MLEs over A_1 are obtained by isotonic regression of $p^* = (a/n, b/n, c/n, d/n)$ on A_1 with weights $w_{11} = d/n$, $w_{10} = c/n$, $w_{01} = b/n$, $w_{00} = a/n$. That is, they are obtained by minimizing $\sum_i w_i(p_i - p_i^*)^2$ over $p \in A_1$.

One interesting property of the solution is that the MLE over C is the closest point, in Euclidean distance, to $(a/n, b/n, c/n, d/n)$ among the $g_1, ..., g_7$ that lie in C, although there may be other points in C, not among $g_1, ...g_7$, that lie closer to $(a/n, b/n, c/n, d/n)$. We find that if $\hat{p}$ is the MLE over B and $\hat{p} \in C$, then $\hat{p}$ is the MLE over C. If $\hat{p} \notin C$, and $\check{p}$ is the MLE over A_1, then $\check{p}$ is the MLE over C. $\hat{p}$ and $\check{p}$ are both obtained by isotonic regression, but with different weights, as described above. In what follows, let g_0 denote the unconstrained MLE $(a/n, b/n, c/n, d/n)$. In TABLE 2, we list $g_0, g_1, ..., g_7$ along with their defining constraints (in addition to $p_{11} + p_{10} + p_{01} + p_{00} = 1$ and $p_i \ge 0$ for all i) and the conditions under which they lie in C. Using TABLE 2 and the REMARKS, LEMMA 1 is proved in Appendix A1.

[Place TABLE 2 here]

3. Maximum Likelihood Estimation of Order for Three-Locus NCI Model. Suppose we have the recombination data displayed in TABLE 1, as before, but now instead of assuming an order for the loci, we wish to estimate order.

LEMMA 2. First suppose $a \leq b \leq c$ in TABLE 1. Let order A-B-C be denoted by O_1, B-A-C by O_2, A-C-B by O_3. Order O_1 is always an MLE of order under NCI in this case, though it is not necessarily unique. In particular, the MLEs of order under the NCI model are as follows in this case:

1. If $a < \min(b, c, d)$ and $ac < bd$, then the MLE of order is O_1.

2. If $a < \min(b, c, d)$ and $ac \geq bd$, then the MLEs of order are O_1 and O_2.

3. If $a = b < \min(c, d)$ then the MLEs of order are O_1 and O_2.

4. If $d \leq a$ and $a + b < n/2$, then the MLEs of order are O_1 and O_2.

5. If $d \leq a$ and $a + b \geq n/2$, then all three possible orders are MLEs.

6. If $a = b = c$, then all three possible orders are MLEs.

Let $c_1 = a, c_2 = b, c_3 = c$. In the case where $c_i \leq c_j \leq c_k$, where (i, j, k) is some ordering of 1,2,3, replace a by in the above solution by c_i, b by c_j, and c by c_k. Similarly, replace O_1 in the above solution by O_i, O_2 by O_j and O_3 by O_k, to get the MLE of order under the NCI model.

PROOF. See Appendix A2.

4. Maximum Likelihood Estimation of Order for Three-Locus NI Model. Wilson (1988) pointed out that the orders for which the product of the pairwise recombination fractions between adjacent loci is minimized are the MLEs under the NI model. Following is a more explicit formulation in the three-locus case.

LEMMA 3. With recombination data as in TABLE 1, assuming $a \leq b \leq c$, then O_1 is always an MLE for order under the NI model, though it is not necessarily unique. In particular, the MLEs of order under the NI model are as follows in this case:

1. If $b + c \leq n/2$, then O_1 is always an MLE, O_2 is an MLE if and only if $a = b$, and O_3 is an MLE if and only if $a = c$.

2. If $a + c \leq n/2$ and $b + c > n/2$, then O_1 is always an MLE, O_2 is an MLE if and only if $a + c = n/2$, and O_3 is an MLE if and only if $a + b = n/2$.

3. If $a + b \leq n/2$ and $a + c > n/2$, then O_1 and O_2 are always MLEs and O_3 is an MLE if and only if $a + b = n/2$.

4. If $a + b > n/2$, then O_1, O_2, and O_3 are all MLEs.

As in LEMMA 2 above, in the case where $c_i \leq c_j \leq c_k$, where (i, j, k) is some ordering of 1,2,3 and

$c_1 = a, c_2 = b, c_3 = c$, replace a, b, c and the O's in the above solution accordingly.

PROOF. Assume without loss of generality that $a \leq b \leq c$. First consider order O_1. There are two recombination parameters for this order under the NI model, θ_{AB} and θ_{BC}, where θ_{AB} is the probability of a recombination between markers A and B and θ_{BC} is the probability of a recombination between markers B and C. The MLE for $(\theta_{AB}, \theta_{BC})$ is $(\min(\frac{a+b}{n}, \frac{1}{2}), \min(\frac{a+c}{n}, \frac{1}{2}))$. Letting $f(x, y) = x \log(y/n) + (n - x)\log(1 - y/n)$, we can write the maximized log-likelihoods under each of the three orders, in the case when $a \leq b \leq c$, as follows:

1. Maximized log-likelihood $\log(\hat{L})$ under order O_1:

(a) If $a + c \leq n/2$, $\log(\hat{L}) = f(a + b, a + b) + f(a + c, a + c)$.

(b) If $a + b \leq n/2$ and $a + c > n/2$, $\log(\hat{L}) = f(a + b, a + b) + f(a + c, n/2)$.

(c) If $a + b > n/2$, $\log(\hat{L}) = f(a + b, n/2) + f(a + c, n/2)$.

2. Maximized log-likelihood under order O_2:

(a) If $b + c \leq n/2$, $\log(\hat{L}) = f(a + b, a + b) + f(b + c, b + c)$.

(b) If $a + b \leq n/2$ and $b + c > n/2$, $\log(\hat{L}) = f(a + b, a + b) + f(b + c, n/2)$.

(c) If $a + b > n/2$, $\log(\hat{L}) = f(a + b, n/2) + f(b + c, n/2)$.

3. Maximized log-likelihood under order O_3:

(a) If $b + c \leq n/2$, $\log(\hat{L}) = f(a + c, a + c) + f(b + c, b + c)$.

(b) If $a + c \leq n/2$, and $b + c > n/2$, $\log(\hat{L}) = f(a + c, a + c) + f(b + c, n/2)$.

(c) If $a + c > n/2$, $\log(\hat{L}) = f(a + c, n/2) + f(b + c, n/2)$.

From this, the result is easily verified. □

5. Comparison of MLEs of Order for Three-Locus NI and NCI Models.

PROPOSITION: If, the recombination data are as given in TABLE 1 with $a \leq b \leq c$, then the maximum likelihood estimates of order under the NI and NCI models are identical with the following exception:

If $a < \min(b, d)$ and $ac < bd$, then the unique MLE of order under the NCI model is O_1, while under the NI model, we have the following three subcases:

1. When $c > d$ and $a + c = n/2$, then O_1, O_2, and O_3 are all MLEs of order.

2. When $c > d$ and $a + c > n/2$, then O_1 and O_2 are both MLEs of order.

3. Otherwise, O_1 is the unique MLE.

Thus, when the MLE of order under the NI model is unique, it is the same as the MLE of order under the NCI model. When there is no unique MLE of order under the NI model, the set of NI MLE orders contains the set of NCI MLE orders. The proposition follows from LEMMAS 1,2, and 3.

6. Counterexample in the Case of Four Loci. We give a counterexample to show that in the case of four marker loci, the NI MLE of order and the NCI MLE of order may be different. This clearly implies that they may differ for any number of markers greater than three.

Suppose we are given complete recombination data on four loci, A, B, C, and D, which may be summarized as in TABLE 3, where a '1' between two markers that are adjacent in the table indicates that they recombined and a '0' indicates that they did not recombine.

Then the estimates of the recombination fractions obtained by considering each pair of loci separately are $\hat{\theta}_{AB} = 4/100, \hat{\theta}_{AC} = 6/100, \hat{\theta}_{AD} = 5/100, \hat{\theta}_{BC} = 4/100, \hat{\theta}_{BD} = 5/100, \hat{\theta}_{CD} = 1/100$. Using Wilson's (1988) result that the MLE of order under the NI model in the four-locus case is the one that minimizes the product of the adjacent recombination fractions, it is clear that the unique MLE of order under the NI model is A-B-C-D.

The MLE for the recombination probabilities of order A-B-C-D under the NCI model lies on the surface $p_{001} + p_{111} = p_{101} + p_{011}$, and the maximized log-likelihood is -5.46. The MLE for the recombination probabilities of order A-D-C-B lies on the surface $p_{001} + p_{111} = p_{101} + p_{011}$, where these probabilities are expressed in terms of the new order, A-D-C-B, and the maximized log-likelihood is -5.29. Thus, A-B-C-D is not an MLE of order under the NCI model.

APPENDIX

APPENDIX A1. PROOF OF LEMMA 1. To prove the lemma, we assume, without loss of generality, that $b \leq c$. By interchanging b and c and p_{10} and p_{01}, we can obtain the solution under $b > c$. We prove the following:

1. If $b \leq c$ and $b + c > n/2$, then the MLE lies in G_1:

   (a) if also $ac \leq bd$ then g_1 is the MLE.

(b) if also $ac > bd$ and $a+b \le c+d$, then g_5 is the MLE.

(c) if also $ac > bd$ and $a+b > c+d$, then g_7 is the MLE.

2. If $b < a$ and $b \le c$ and $b+c \le n/2$, then the MLE lies in G_2:

(a) if also $c \le d$ and $a+b \le 2c$, then g_2 is the MLE.

(b) if also $a+b > 2c$ and $a+b+c \le 3n/4$, then g_4 is the MLE.

(c) if also $c > d$ and $a+b \le c+d$, then g_5 is the MLE.

(d) if also $[c > d$ or $a+b > 2c]$ and $a+b+c > 3n/4$ and $a+b > c+d$, then g_7 is the MLE (Note that $c > d$ and $a+b > c+d \rightarrow a+b+c > 3n/4$. Also, $a+b > 2c$ and $a+b+c > 3n/4 \rightarrow a+b > c+d$.

Clearly, in all of these cases, the unconstrained MLE g_0 does not lie in C, since either $b+c > n/2$ or $b < a$.

Case 1.(a): $b \le c$ and $b+c > n/2$ and $ac \le bd$.

Show g_1 is the MLE. In this case, $g_1 \in C$. This is because $b \le c$ and $ac \le bd$ imply $[a \le d$ or $b = c = 0]$. Since $b+c > n/2$, we must have $a \le d$, so then also $ab \le cd$. Thus, the necessary conditions listed in TABLE 2 are satisfied.

Let L be defined as follows: $L(w,x,y,z) = w^a x^b y^c z^d$.

(i) In case $c \le d$ and $a+b \le 2c$, i.e. when $g_2 \in C$, we need to show that $L(g_1) \ge L(g_2)$, i.e. $2^{-n}(a/(a+d))^a(b/(b+c))^b(c/(b+c))^c(d/(a+d))^d \ge ((a+b)/(2n))^{a+b}(c/n)^c(d/n)^d$, which is equivalent to $((2a)/(a+b))^a((2b)/(a+b))^b \ge ([2(a+d)]/n)^{a+d}([2(b+c)]/n)^{b+c}$ or $a\log(1+(a-b)/(a+b)) + b\log(1+(b-a)/(a+b)) \ge (a+d)\log(1-(b+c-a-d)/n) + (b+c)\log(1+(b+c-a-d)/n)$

Taking a Taylor expansion of the LHS gives

$$\text{LHS} = \sum_{i=1}^{\infty}(a-b)^{2i}/[2i(2i-1)(a+b)^{2i-1}] \ge (a-b)^2/[2(a+b)].$$

On the RHS, note that

$$(a+d)\log(1-(b+c-a-d)/n) = -(a+d)\sum_{i=1}^{\infty}(b+c-a-d)^i/(in^i) \le -(a+d)(b+c-a-d)/n$$

and

$$(b+c)\log(1+(b+c-a-d)/n) = (b+c)\sum_{i=1}^{\infty}(-1)^{i+1}(b+c-a-d)^i/(in^i) \le (b+c)(b+c-a-d)/n,$$

so RHS $\le (b+c-a-d)^2/n$. Thus, we need only show $(a-b)^2/[2(a+b)] \ge (b+c-a-d)^2/n$ in order to prove $L(g_1) \ge L(g_2)$. This is equivalent to $n/[2(a+b)] \ge (b-a+c-d)^2/(a-b)^2$. We must have $a \le b$, since otherwise $c \le d$ would imply $a+d \ge b+c$ which contradicts $b+c > n/2$. Thus, $0 < b-a+c-d \le b-a$ since $c \le d$ and $b+c > a+d$. So $(b-a+c-d)^2/(a-b)^2 \le 1$. Also, $a+b \le a+d$ since $b \le c \le d$, and $a+d < n/2$ because $a+d < b+c$. These imply $a+b < n/2$, so $n/[2(a+b)] > 1$. Thus, $L(g_1) \ge L(g_2)$ in the case when $g_2 \in C$.

(ii) In case $b \le d$ and $a+c \le 2b$, i.e. when $g_3 \in C$, we need to show that $L(g_1) \ge L(g_3)$, i.e. $2^{-n}(a/(a+d))^a(b/(b+c))^b(c/(b+c))^c(d/(a+d))^d \ge ((a+c)/(2n)^{a+c}(b/n)^b(d/n)^d$, which is equivalent to $((2a)/(a+c))^a((2c)/(a+c))^c \ge ([2(a+d)]/n)^{(a+d)}([2(b+c)]/n)^{(b+c)}$. We have proved above that $\log(\text{LHS}) \ge (c-a)^2/[2(a+c)]$ and $\log(\text{RHS}) \le (b+c-a-d)^2/n$. Showing $(c-a)^2/[2(a+c)] \ge (b+c-a-d)^2/n$ is equivalent to showing $n/[2(a+c)] \ge (b+c-a-d)^2/(c-a)^2$, which is true because $0 < b+c-a-d \le c-a$ since $b+c > a+d$ and $b \le d$; also $a+c \le 2b \le b+d$ so $a+c \le n/2$. Thus, $L(g_1) \ge L(g_3)$ in the case when $g_3 \in C$.

(iii) In case ($c > d$ or $a+b > 2c$) and ($b > d$ or $a+c > 2b$) and $a+b+c \le 3n/4$, i.e. when $g_4 \in C$ and $g_2, g_3 \notin C$ (Recall: $L(g_4) \le L(g_2), L(g_3)$, so we do not consider g_4 unless $g_2, g_3 \notin C$), we need to show that $L(g_1) \ge L(g_4)$. Note that $a+b > 2c$ may be eliminated from the above conditions because $a+b > 2c$ and $a+b+c \le 3n/4$ and $b+c > n/2$ would imply $b > c$ which contradicts an assumption. Note also that $L(g_1)$ is automatically greater than or equal to $L(g_5)$, because $G_5 \subset G_1$. If we can show that $L(g_5) \ge L(g_4)$ under the given conditions, then we're done. Recall that the

MLE over the region A_2 is the closest point on A_2 to g_0 in Euclidean distance (see REMARKS). We show that g_5 is the closest point on A_2 to g_0 in Euclidean distance. Then since $g_4 \in C$, we have $g_4 \in A_2$ (see REMARKS) so $L(g_5) \geq L(g_4)$. To find the closest point on A_2 to g_0 in Euclidean distance, we first note that the closest point to g_0, in Euclidean distance, on the plane $G_2 \supset A_2$ is g_2 (see REMARKS). We have $g_2 \notin C$ and hence $g_2 \notin A_2$. Then, since the Euclidean distance to g_0 from a point in G_2 is a smooth function whose gradient is zero only at g_2, the nearest point on $A_2 \in G_2$ to g_0 must lie on the boundary of A_2, i.e. $A_2 \cap (A_1 \cup A_3)$. We find that the closest points to g_0 on the lines G_4 (which contains $A_2 \cap A_3$) and G_5 (which contains $A_2 \cap A_1$) are g_4 and g_5, respectively. We have already seen that $g_4 \in A_2$. To see that $g_5 \in A_2$, i.e. that $(a+b)/(2n) \leq (c+d)/(2n)$, note that we have the conditions $a+b+c \leq 3n/4$ and $c > d$, so $n/4 \leq d < c$ which implies $c + d > n/2$ which implies $a + b < c + d$. It remains to show that g_5 is closer to g_0 than g_4 is, i.e. that $2[(a-b)/(2n)]^2 + 2[(c-d)/(2n)]^2 \leq [(2a-b-c)/(3n)]^2 + [(2b-a-c)/(3n)]^2 + [(2c-a-b)/(3n)]^2$. The RHS $= [(a-b)^2 + (b-c)^2 + (a-c)^2]/(3n^2)$, so the inequality becomes $3(a-b)^2 + 3(c-d)^2 \leq 2(a-b)^2 + 2(b-c)^2 + 2(a-c)^2$ which is equivalent to $3(c-d)^2 \leq (2c-a-b)^2$. Since $a+b < c+d < 2c$ and $d < c$, then the statement $3(c-d)^2 \leq (2c-a-b)^2$ is equivalent to the statement that $\sqrt{3}(c-d) \leq 2c-a-b$, which is equivalent to $a+b \leq (2-\sqrt{3})c + \sqrt{3}d$. The RHS $> 2d \geq n/2$ (since $c > d$ and also $(a+b+c)/n \leq 3/4$, the latter implying $d \geq n/4$). The LHS $< n/2$ since $a+b < c+d$. This shows that g_5 is closer to g_0 than g_4 is, hence that g_5 is the closest point on A_2 to g_0 and so is the MLE on A_2. Thus, by the argument above, $L(g_1) \geq L(g_4)$ in the case when $g_4 \in C$.

(iv) We do not need to consider g_5, g_6, and g_7, because the likelihoods at these points will always be dominated by the likelihood at g_1.

Case 1.(b): $b \leq c$ and $b + c > n/2$ and $ac > bd$ and $a + b \leq c + d$.

Show g_5 is the MLE. $g_5 \in C$ since $a + b \leq c + d$.

(i) In this case, $g_1 \notin C$ since $ac > bd$.

(ii) We show that $g_2 \notin C$. To do this, we show that $c > d$. Note that $a < c$, because $a + b \leq c + d$, and $b + c > n/2$ implies $a + d < b + c$. First suppose $b \leq a < c$. Then $c > d$ because otherwise $c \leq d$ would imply $b + c \leq n/2$, which contradicts an assumption. Second, suppose $a < b \leq c$. Then $c > d$ because otherwise $c \leq d$ would imply $bd > ac$ which contradicts an assumption. Thus, $g_2 \notin C$ in the case 1.(b).

(iii) We show that $g_3 \notin C$. To do this, we show that either $b > d$ or $a + c > 2b$. Assume $b \leq d$. Then $ac > bd \geq b^2$ implies that the geometric mean of a and c is greater than b, but the arithmetic mean of a and c cannot be smaller than the geometric mean, so $a + c > 2b$. ($\sqrt{ac} \leq (a + c)/2 \iff ac \leq a^2/2 + ac + c^2/2 \iff 0 \leq (a^2 + c^2)/2$). Thus, $g_3 \notin C$ in the case 1.(b).

(iv) In case $a + b + c \leq 3n/4$, i.e. when $g_4 \in C$, we need to show that $L(g_5) \geq L(g_4)$. In fact, this follows from the proof given in case 1.(a)(iii).

(v) In case $a + c \leq b + d$, i.e. when $g_6 \in C$, we need to show that $L(g_5) \geq L(g_6)$, i.e. that $[(a + c)/(2n)]^{a+c}[(b + d)/(2n)]^{b+d} \leq [(a + b)/n]^{a+b}[(c + d)/(2n)]^{c+d}$, which is equivalent to $[(a + c)/(2n)]^{a+c}[1 - (a + c)/n]^{n-a-c} \leq [(a + b)/n]^{a+b}[1 - (a + b)/n]^{n-a-b}$, where we have $a + b \leq a + c \leq b + d \leq b + c$ which implies that $0 \leq a + b \leq a + c \leq n/2$. This is true because the function $f(x) = (x/n)^x(a - x/n)^{n-x}$ is decreasing in $0 \leq x \leq n/2$. Thus, $L(g_5) \geq L(g_6)$ in the case 1.(b).

(vi) We do not need to consider g_7 because the likelihood at this point will always be dominated by the likelihood at g_5.

Case 1.(c): $b \leq c$ and $b + c > n/2$ and $ac > bd$ and $a + b > c + d$.

Show g_7 is the MLE. Note that $g_7 = (.25, .25, .25, .25) \in C$.

(i) In this case, $g_1 \notin C$ since $ac > bd$.

(ii) In this case, $g_2 \notin C$ since $g_2 \in C$ would require $c \leq d$ and $a + b \leq 2c$, but we have $a + b > c + d$

and assuming $c \le d$, this would give $a + b > 2c$.

(iii) In this case, $g_3 \notin C$ since $g_3 \in C$ would require $b \le d$ and $a + c \le 2b$. Since we have $b \le c$ and $a + b > c + d$, having $b \le d$ would require $a > \max(c, d) \ge c \ge b$, so $a + c > 2b$.

(iv) In this case, $g_4 \notin C$ since $g_4 \in C$ would require $a+b+c \le 3n/4$, which implies $d \ge n/4$. Since also $c \ge b$ and $b - c > n/2$, we have $c \ge n/4$, which, with $d \ge n/4$, implies $c + d \ge n/2$, giving $c + d \ge a + b$, which contradicts an assumption.

(v) In this case, g_5 is clearly not in C since $a + b > c + d$.

(vi) In this case, $g_6 \notin C$ since $a + b > c + d$ and $c \ge b$ imply $a + c > b + d$.

Case 2.(a): $b < a$ and $b \le c$ and $b + c \le n/2$ and $c \le d$ and $a + b \le 2c$.

Show g_2 is the MLE. $g_2 \in C$ since $c \le d$ and $a + b \le 2c$. Consider $B \supset C$ as defined in REMARKS. If we can show that g_2 is the MLE over B, then it is the MLE over C and we're done. Note that to find the MLE over B, one may first find the unconstrained MLE, g_0. If this does not lie in B, then one should consider the MLEs over the planes bounding the region B, namely g_2 and g_3 (the MLE over A_0 need not be considered (see REMARKS). The larger of these two lying in B is the MLE over B. If neither lies in B, then g_4 is the MLE over B. Clearly $g_0 \notin B$ since $b < a$. We already have that $g_2 \in C \subset B$. To see that $g_3 \notin B$, note that $a + c > 2b$.

Case 2.(b): $b < a$ and $b \le c$ and $b + c \le n/2$ and $a + b > 2c$ and $a + b + c \le 3n/4$.

Show g_4 is the MLE. We reason as in **Case 2.(a)** . g_2 and g_3 do not lie in B since $a + b > 2c$ and $a + c > 2b$. $g_4 \in C \subset B$ since $a + b + c \le 3n/4$. Thus, g_4 is the MLE over B and hence over C.

Case 2.(c): $b < a$ and $b \le c$ and $b + c \le n/2$ and $c > d$ and $a + b \le c + d$.

Show g_5 is the MLE. $g_5 \in C$ since $a + b \le c + d$.

(i) In this case, $g_1 \notin C$ since $ac > bd$.

(ii) In this case, $g_2 \notin C$ since $c > d$.

(iii) In this case, $g_3 \notin C$ since $a + c > 2b$.

(iv) In case $a+b+c \le 3n/4$, i.e. when $g_4 \in C$, we need to show $l(g_5) \ge L(g_4)$. Apply the argument of **Case 1.** (a) (iii).

(v) In this case, $g_5 \notin C$ since $a+c > b+d$.

Since g_5 always dominates g_7, g_5 is the MLE in this case.

Case 2.(d): $b < a$ and $b \le c$ and $b+c \le n/2$ and $[c > d$ or $a+b > 2c]$ and $a+b+c > 3n/4$ and $a+b > c+d$.

Show g_7 is the MLE. Note that $g_7 = (.25, .25, .25, .25)$ is always in C.

(i) In this case, $g_1 \notin C$ since $ac > bd$. This is because $c \ge b$ and $a+b > c+d$ imply $a > d$, and since $b \le c$, and all counts are assumed positive, we have $ac > bd$.

(ii) In this case, $g_2 \notin C$ since $c > d$ or $a+b > 2c$.

(iii) In this case, $g_3 \notin C$ since $a+c > 2b$.

(iv) In this case, $g_4 \notin C$ since $a+b+c > 3n/4$.

(v) In this case, $g_5 \notin C$ since $a+b > c+d$.

(vi) In this case, $g_6 \notin C$ since $a+c > b+d$. □

APPENDIX A.2 PROOF OF LEMMA 2. Assume without loss of generality that $a \le b \le c$. First we find the MLEs of the multilocus recombination parameters in the different orders, by applying LEMMA 1. The multilocus recombination parameters are listed in order corresponding to the following order of their associated cell counts: (a, b, c, d).

1. MLE of multilocus recombination parameters in order *A-B-C*:

Applying LEMMA 1, we get the following solution:

If $b+c \le n/2$ then MLE is $(a, b, c, d) \times n^{-1}$.

If $b+c > n/2$ and $ac \le bd$, then MLE is $(a/(a+d), b/(b+c), c/(b+c), d/(a+d)) \times 2^{-1}$ unless $a+d = 0$ (see REMARKS after LEMMA 1).

If $b + c > n/2$ and $ac > bd$ and $a + b \leq n/2$, then MLE is $(a + b, a + b, c + d, c + d) \times (2n)^{-1}$.

If $a + b > n/2$, then MLE is $(.25,.25,.25,.25)$.

2. MLE of multilocus recombination parameters in order B-A-C:

Let $A' = B$, $B' = A$, $C' = C$, $a' = b$, $b' = a$, $c' = c$, $d' = d$. Applying LEMMA 1 to loci A', B', C' and counts a', b', c', d', we get:

If $a = b$ and $a + c \leq n/2$, then MLE is $(a, b, c, d) \times n^{-1}$.

If $a < b$ and $a + c \leq n/2$ and $c \leq d$, then MLE is $((a + b)/2, (a + b)/2, c, d) \times n^{-1}$.

If $a < b$ and $a + c \leq n/2$ and $c > d$, then MLE is $(a + b, a + b, c + d, c + d) \times (2n)^{-1}$.

If $a + c > n/2$ and $a + b \leq n/2$ (Note $d < c$ since otherwise $a + c \leq n/2$. Thus, $bc > ad$ since all counts are positive), then MLE is $(a + b, a + b, c + d, c + d) \times (2n)^{-1}$.

If $a + b > n/2$ (Note that here, $d < a \leq b \leq c$.), then MLE is $(.25,.25,.25,.25)$.

3. MLE of multilocus recombination parameters in order A-C-B:

Let $A' = A$, $B' = C$, $C' = B$, $a' = c$, $b' = b$, $c' = a$, $d' = d$. Applying LEMMA 1 to loci A', B', C' and counts a', b', c', d', we get:

If $a = b = c \leq d$, then MLE is $(a, b, c, d) \times n^{-1}$.

If $a < c$ and $b \leq d$ and $a + c \leq 2b$, then MLE is $((a + c)/2, b, (a + c)/2, d) \times n^{-1}$.

If $a + b \leq n/2$ and $a + c > 2b$ and $a + b + c \leq 3n/4$, then MLE is $((a + b + c)/3, (a + b + c)/3, (a + b + c)/3, d) \times n^{-1}$.

If $a < c$ and $a + c \leq n/2$ and $b > d$, then MLE is $(a + c, b + d, a + c, b + d) \times (2n)^{-1}$.

If $a + b \leq n/2$ and $[b > d$ or $a + c > 2b]$ and $a + b + c > 3n/4$ and $a + c > n/2$, then MLE is $(.25,.25,.25,.25)$.

If $a + b > n/2$ (here $d < a$), then MLE is $(.25,.25,.25,.25)$.

MLE of order, applying the above results:

Case 1: Suppose $b + c \leq n/2$. Then A-B-C is clearly an MLE of order, since the maximized likelihood in that case is the same as the unconstrained maximized likelihood.

(a) If also $a < b$, then neither of the maximized likelihoods for the other orders is the same as the unconstrained maximized likelihood, so A-B-C is the unique MLE of order.

(b) If also $a = b < c$, then B-A-C is also an MLE of order, because its maximized likelihood is the same as the unconstrained maximized likelihood. A-C-B is not, for its maximized likelihood is not the same.

(c) If also $a = b = c$, then all three orders have the same maximized likelihood, so all are MLEs.

Case 2: Suppose $b + c > n/2$ and $ac \leq bd$. Let C_1 be the region defined by the NCI constraints under order A-B-C, consisting of all points $(w, x, y, z) \in I\!R$ satisfying $(0 \leq w \leq \min(x, y)$ and $x + y \leq 1/2$ and $w + x + y + z = 1)$. The MLE for the recombination parameters under order A-B-C is the MLE over C_1 which is $(a/(a+d), b/(b+c), c/(b+c), d/(a+d)) \times 2^{-1}$ unless $a + d = 0$. We shall show that A-B-C is always an MLE of order by showing that under the other orders, the MLEs for the multilocus recombination parameters also lie in C_1. Then the maximized likelihood under order A-B-C will never be smaller than the maximized likelihoods under the other orders.

(a) We show that the MLE under order B-A-C lies in C_1, and we give conditions under which this order is an MLE in addition to order A-B-C. The cases below follow the cases in the above list of MLEs for recombination parameters for order B-A-C.

(i)Note that $a = b$ in this case would require $a = b = 0$, for then $b + c > n/2$ would imply $c > d$ which would imply $ac > bd$ if $a > 0$, contradicting an assumption. However, $a = b = 0$ and $c > d$ do not allow $a + c \leq n/2$.

(ii) If $a < b$ and $a + c \leq n/2$ and $c \leq d$, then the MLE given in 2) above lies in C_1, because $c \leq d$ and $a + b \leq 2c$. The maximized likelihood under B-A-C is always less than that under A-B-C. They could be equal only in case $a = b$ and $c = d$, but this is not possible in this case.

(iii) If $a < b$ and $a + c \leq n/2$ and $c > d$, then the MLE given in 2) above lies in C_1 since

$a+b \leq a+c \leq n/2$ implies $a+b \leq c+d$. B-A-C is also an MLE of order if and only if $ac = bd$, because the maximized likelihood is the same in this case as for A-B-C.

(iv) If $a+c > n/2$ and $a+b \leq n/2$, then the MLE given in 2) clearly lies in C_1. B-A-C is also an MLE of order if and only if $ac = bd$, because the maximized likelihood is the same in this case as for A-B-C.

(v) Note that it is impossible to have $a+b > n/2$ because we also have $c \geq b$, so this would imply $a > d$ which contradicts $ac \leq bd$, since $0 \leq d < a \leq b \leq c$.

(b) We show that the MLE under order A-C-B lies in C_1, and we give conditions under which this order is an MLE in addition to order A-B-C. The cases below follow the cases in the above list of MLEs for recombination parameters for order A-C-B.

(i) As in (a)(i), we cannot have both $a = b$ and $a+c \leq n/2$ in this case.

(ii) If $a < c$ and $a+b \leq n/2$ and $b \leq d$ and $a+c \leq 2b$, then the MLE given in 3) above clearly lies in C_1. A-C-B is never an MLE for order in this case, because this would require $b+c = n/2$ and $a = c$.

(iii) If $a+b \leq n/2$ and $a+c > 2b$ and $a+b+c \leq 3n/4$, then the MLE given in 3) above clearly lies in C_1. A-C-B is never an MLE for order in this case, because this would require $a = b = c = d = n/4$.

(iv) If $a < c$ and $a+c \leq n/2$ and $b > d$, then the MLE given in 3) above clearly lies in C_1. A-C-B is an MLE for order in the case when $ab = cd$.

(v) If $a+b \leq n/2$ and $[b > d$ or $a+c > 2b]$ and $a+b+c > 3n/4$ and $a+c > n/2$, then the MLE given in 3) above clearly lies in C_1. A-C-B is never an MLE for order in this case, because this would require $a = b = c = d = n/4$.

(vi) Note that it is not possible to have $a+b > n/2$ in this case because we have $c \geq b$, which with $a+b > n/2$ implies $a > d$, which contradicts $ac \leq bd$.

Case 3: Suppose $b+c > n/2$ and $ac > bd$ and $a+b \leq n/2$. We show A-B-C and B-A-C are both

MLEs of order, and give conditions under which A-C-B is also an MLE of order. The MLE for the recombination parameters under order A-B-C is $(a+b, a+b, c+d, c+d) \times (2n)^{-1}$.

(a) Consider the cases in 2) above giving the MLE of recombination parameters under order B-A-C. Note that $ac > bd$ implies $(c-d)b > (b-a)c$ which implies $c - d > (b-a)c/b \geq b - a$ since $0 < a \leq b \leq c$. Thus, $c - d > b - a$ implies $a + c > n/2$, so the first two cases under B-A-C do not apply. The last case does not apply either. In the only two remaining cases, the MLE for the recombination parameters is the same under order B-A-C as under order A-B-C, so these are both MLEs of order.

(b) Consider the cases in 3) above giving the MLE of recombination parameters under order A-C-B.

(i) We have already shown $a + c > n/2$, i.e. $a + c > b + d$, so the first case does not apply.

(ii) If $b \leq d$, then $a + c > 2b$, so the second case does not apply.

(iii) In case, $a + c > 2b$ and $a + b + c \leq 3n/4$, then the MLE given in 3) above lies on C_1 since $a+b+c \leq 3n/4$. A-C-B is never an MLE for order in this case since this would require $c = d = n/4$ which would contradict $ac > bd$.

(iv) Since $a + c > n/2$, the fourth case does not apply.

(v) In case $[b > d$ or $a + c > 2b]$ and $a + b + c > 3n/4$, then the MLE $(.25, .25, .25, .25) \in C$. A-C-B will also be an MLE for order in this case if and only if $a + b = n/2$.

(vi) Since $a + b \leq n/2$, the sixth case does not apply.

Case 4: Suppose $a + b > n/2$. A-B-C, B-A-C, and A-C-B are all MLEs of order, because the MLEs for the recombination parameters under all three orders are equal to $(.25, .25, .25, .25)$. □

REFERENCES

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions.* Wiley, New York.

Daley, D. J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes.* Springer, New York.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1-22.

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

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

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

Raubertas, R. F. and Lee, C. C. (1986). Hypothesis tests for normal means constrained by linear inequalities. *Communications in Statistics Theory and Methods* **15** 2809-2833.

Speed, T. P., McPeek, M. S., and Evans, S. N. (1992). Robustness of the no-interference model for ordering genetic markers. *Proceedings of the National Academy of Sciences of the United States of America* **89** 3103-3106.

Sturtevant, A. H. (1915). The behavior of the chromosomes as studied through linkage. *Zeitschrift fur induktive Abstammungs- und Vererbungslehre* **13** 234-287.

Weinstein, A. (1936). The theory of multiple-strand crossing over. *Genetics* **21** 155-199.

Wilson, S. R. (1988). A major simplification in the preliminary ordering of linked loci. *Genetic Epidemiology* **5** 75-80.

Zhao, H., McPeek, M. S., Speed, T. P. (1995). A statistical analysis of chromatid interference.

Genetics **139** 1057-1065.

TABLE 1

	recombination between B & C	no recombination between B & C
recombination between A & B	*a*	*b*
no recombination between A & B	*c*	*d*

LEGEND: Counts of recombination events among three loci.

TABLE 2

r	defining constraints	MLE (g_r)	conditions for $g_r \in C$
0	none	$(a, b, c, d) \times n^{-1}$	$a \leq b,\ a \leq c,$ $b + c \leq n/2$
1	$p_{10} + p_{01} = 1/2$	$(a/(a+d), b/(b+c),$ $c/(b+c), d/(a+d)) \times 2^{-1}$	$ab \leq cd,\ ac \leq bd,$ $[b + c > 0$ or $a \leq d]$
2	$p_{11} = p_{10}$	$((a+b)/2,$ $(a+b)/2, c, d) \times n^{-1}$	$c \leq d,\ a + b \leq 2c$
3	$p_{11} = p_{01}$	$((a+c)/2, b,$ $(a+c)/2, d) \times n^{-1}$	$b \leq d,\ a + c \leq 2b$
4	$p_{11} = p_{10} = p_{01}$	$((a+b+c)/3, (a+b+c)/3,$ $(a+b+c)/3, d) \times n^{-1}$	$a + b + c \leq 3n/4$
5	$p_{11} = p_{10} =$ $1/2 - p_{01}$	$(a+b, a+b, c+d,$ $c+d) \times (2n)^{-1}$	$a + b \leq c + d$
6	$p_{11} = p_{01} =$ $1/2 - p_{10}$	$(a+c, b+d, a+c,$ $b+d) \times (2n)^{-1}$	$a + c \leq b + d$
7	$p_{11} = p_{10} =$ $p_{01} = 1/2 - p_{01}$	$(.25, .25, .25, .25)$	always

LEGEND: Some constrained MLEs of NCI recombination probabilities and the conditions under which they lie the the NCI constraint space C. g_0 is the unconstrained MLE, g_r, $r = 1, 2, 3$ are MLEs over planes containing faces of ∂C (the boundary of C), g_r, $r = 4, 5, 6$ are MLEs over lines containing edges of ∂C, and g_7 is a vertex of ∂C. Column 2 contains the constraints which, in addition to the constraints ($p_i \geq 0$ and $\sum_i p_i = 1$), define the region G_r over which g_r is the MLE.

TABLE 3

A	B	C	D	count
	1	0	0	2
	0	1	0	3
	0	0	1	0
	1	1	0	1
	1	0	1	1
	0	1	1	0
	1	1	1	0
	0	0	0	93
				100

LEGEND: Counts of recombination events among four loci. A '1' between two markers that are adjacent in the table indicates that they recombined and a '0' indicates that they did not recombine.