Comparative Genomics, Marker Density and Statistical Analysis of Chromosome Rearrangements

Daniel J. Schoen

Department of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada

Manuscript received February 12, 1999
Accepted for publication October 22, 1999

ABSTRACT

Estimates of the number of chromosomal breakpoints that have arisen (*e.g.*, by translocation and inversion) in the evolutionary past between two species and their common ancestor can be made by comparing map positions of marker loci. Statistical methods for doing so are based on a random-breakage model of chromosomal rearrangement. The model treats all modes of chromosome rearrangement alike, and it assumes that chromosome boundaries and breakpoints are distributed randomly along a single genomic interval. Here we use simulation and numerical analysis to test the validity of these model assumptions. Mean estimates of numbers of breakpoints are close to those expected under the random-breakage model when marker density is high relative to the amount of chromosomal rearrangement and when rearrangements occur by translocation alone. But when marker density is low relative to the number of chromosomes, and when rearrangements occur by both translocation and inversion, the number of breakpoints is underestimated. The underestimate arises because rearranged segments may contain markers, yet the rearranged segments may, nevertheless, be undetected. Variances of the estimate of numbers of breakpoints decrease rapidly as markers are added to the comparative maps, but are less influenced by the number or type of chromosomal rearrangement separating the species. Variances obtained with simulated genomes comprised of chromosomes of equal length are substantially lower than those obtained when chromosome size is unconstrained. Statistical power for detecting heterogeneity in the rate of chromosomal rearrangement is also investigated. Results are interpreted with respect to the amount of marker information required to make accurate inferences about chromosomal evolution.

EVOLUTIONARY change in the macrostructure of individual chromosomes occurs largely by reciprocal translocation and inversion. During the course of the independent evolutionary histories separating two species from their common ancestor, divergence in chromosome structure arising from chromosomal rearrangement is manifested as the progressive fractionation of the genome into increasingly smaller conserved chromosome segments (Nadeau and Taylor 1984). For example, comparative mapping studies often show that closely related organisms share large portions of chromosome segments in which the identities and linear orders of genes are conserved, while more distantly related taxa exhibit shorter conserved chromosome segments (Paterson *et al.* 1996; Ehrlich *et al.* 1997).

The discovery of conserved segments of chromosomes among taxa suggests that it may be possible to construct unified genetic maps for a number of organismal groups (*e.g.*, the grasses, higher plants, fishes, and mammals; Ahn and Tanksley 1993; Paterson *et al.* 1996; Nadeau and Sankoff 1997; Gale and Devos 1998). This could have important consequences for genetic and biotechnical applications. For instance, the detailed information on genome structure gained from sequencing and mapping efforts with model systems such as *Arabidopsis thaliana* may assist in the identification of agriculturally important genes in domesticated plant species and help facilitate marker-based introgression from exotic germ plasm, marker-assisted selection, and positional cloning. If the chromosomal locations of one or more genes of interest are known with reference to the positions of a set of marker genes in the model species, the probabilities of linkage between the markers and genes of interest in the target species can be calculated (Nadeau and Taylor 1984). Information on the overall amount of the chromosomal rearrangement separating two species may also help to detect conserved gene blocks (*i.e.*, blocks that are larger than expected given the overall amount of rearrangement observed between two genomes). As well, estimates of the extent and type of chromosomal rearrangement may be useful for reconstructing evolutionary history or for testing specific evolutionary hypotheses about rates of chromosome evolution (Ohno 1967; Charlesworth 1992).

Currently there are few statistical tools for comparing genetic maps, and most studies are based on visual inspection of shared syntenies and conserved gene arrangements. One of the more useful analytical approaches for interpreting comparative map data was

Address for correspondence: Department of Biology, McGill University, 1205 Ave. Docteur Penfield, Montreal, Quebec H3A 1B1 Canada. E-mail: dan_schoen@maclan.mcgill.ca

initiated by Nadeau and Taylor (1984) and subsequently expanded by Sankoff and colleagues (Sankoff and Nadeau 1996; Ehrlich *et al.* 1997; Sankoff *et al.* 1997; Nadeau and Sankoff 1998). These researchers proposed the use of a probabilistic model to infer the amount of chromosomal rearrangement from the lengths and numbers of conserved chromosomal segments detected in a comparative genetic mapping investigation. They express the amount of chromosomal evolution between two species as the number of chromosomal breakpoints separating their genomes (Sankoff and Nadeau 1996). The underlying model is referred to as the "random-breakage model" of chromosomal evolution, because it assumes a uniformly distributed probability that any given chromosomal location will experience a breakpoint (*e.g.*, arising from translocation or inversion) during divergence from a common ancestor.

In practice, the information required to apply the random-breakage model to the estimation of chromosomal evolution comes from the comparative mapping of homologous marker loci such as conserved expressed sequence tags (ESTs; Paterson *et al.* 1996; Van Deynze *et al.* 1998). Estimates based on the model are expected to depend on the validity of the model assumptions, and on the amount and quality of the comparative map data. The amount of chromosomal evolution separating the species in question may also influence the accuracy of the estimates. Given the increasing interest in comparative genomic investigations, it is surprising that there have been no studies of how the amount of mapping effort influences the quality of inferences obtained from the comparative maps. In this article I investigate the estimation of chromosome evolution based on the random breakage model, and (1) how estimates of chromosomal breakpoints and their variances are influenced by the density of markers used in comparative mapping; (2) how estimates of chromosomal breakpoints and their variances are influenced by the amount and type of chromosomal rearrangement; and (3) how ability to detect heterogeneity in the rate of chromosomal rearrangement is influenced by the density of markers and extent of chromosomal evolution.

METHODS

Maximum-likelihood estimates of chromosomal divergence and their variances (numerical solutions): In analytical studies of the random-breakage model, the genome is represented as a single interval of unit length 1.0, broken at n randomly placed positions (*e.g.*, by translocations and inversions) as well as by chromosome endpoints (Sankoff and Nadeau 1996). This results in $n + 1$ segments in which gene order is conserved with reference to another genome of interest. When there are m homologous marker genes distributed uniformly on the interval 0–1, the probability that an arbitrary segment contains r marker genes is (Sankoff and Nadeau 1996)

$$P(r) = \frac{n}{n+m}\binom{n}{m}\Big/\binom{n+m-1}{r}. \qquad (1)$$

In a comparative mapping study one uncovers conserved segments containing $r \geq 1$ marker genes. These are referred to as "nonempty segments." There will also be a certain number of conserved segments that do not contain markers and that thus remain undetected (empty segments). Comparison of the maps of the two species provides information on the number of nonempty segments, each containing r marker genes (s_r, where $r \geq 1$). The sum total of the nonempty segments, $a = \Sigma s_r$, is sufficient for calculation of the likelihood that there are n chromosomal breakpoints separating the two species (Sankoff *et al.* 1997). This likelihood is

$$L(n \mid m,a) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}. \qquad (2)$$

Numerical analysis of Equation 2 allows one to determine the maximum-likelihood estimate (MLE) of (hereafter $\hat{n}$). The estimated asymptotic variance of $\hat{n}$ can be calculated as

$$\sigma^2(\hat{n}) = -\left(\frac{\partial^2 L(n \mid m,n)}{\partial n^2}\bigg|_{n=\hat{n}}\right)^{-1} \qquad (3)$$

(Elandt-Johnson 1971). To solve for $\hat{n}$ and its variance estimate, we require the value of a expected when there have been n chromosomal breakpoints and m markers. This expected value, a^*, is derived from Equation 1 as

$$a^* = [1 - P(0)](n + 1) = m\,(n + 1)/(n + m). \qquad (4)$$

This solution allows one to obtain numerical solutions to Equation 3 under different combinations of n, m, and a^*, and thereby examine how the numbers of markers used and the actual amount of chromosomal evolution influence the estimates of chromosomal rearrangement and their variances.

Maximum-likelihood estimates of chromosomal divergence and their variances (simulation studies): Results obtained with the methods outlined above give one picture of the relationship of the mean and variance of $\hat{n}$ to the numbers of markers used and the amount of chromosomal evolution separating the species. These results, however, may differ from those obtained with actual genomes for several reasons. First, the analytical model described above (and the associated likelihood estimator) assumes that all conserved segments arising from chromosomal rearrangement will be detected provided they contain one or more markers. As illustrated below, this need not be true in general, especially in the case of chromosomal inversions. Second, the random-breakage model assumes that the genome is comprised of a single long interval with uniformly distributed breakpoints arising from both chromosomal segment reshuffling as well as from the chromosome end points. True chromosome size variation, however, is constrained (Stebbins 1971), and so there the assumption that chromosomal ends are uniformly distributed along a single interval will be violated. This will not influence the expected value of $\hat{n}$ (Sankoff and Nadeau 1996), but it will influence its variance; *i.e.*, there is more variation in *a* under the analytical methods compared with the case where chromosome size variation is constrained. Third, the numbers of markers employed in a comparative mapping study may be insufficient for the asymptotic approximation in Equation 3 to yield an accurate variance estimate.

To extend the investigation to more realistic genomes, chromosome evolution was modeled by computer simulation. A fixed ancestral genome size of T length units was assumed such that each chromosome was of equal length T/c. The m homologous marker genes were assigned to random positions along the chromosomes. Starting with this ancestral genome, t random translocation and i random inversion events were distributed at random to two isolated lineages. For each translocation, chromosome segment exchange involved two randomly chosen chromosomes and two randomly chosen breakpoints (separated by the same distance on each of the two chromosomes). For each inversion, one chromosome was chosen at random, and two breakpoints within it were randomly chosen. Following the $e = t + i$ chromosome rearrangement events, the chromosomes of the two species were compared, and the number of conserved chromosome segments (*i.e.*, the number of segments containing identical runs of one or more marker genes when compared in forward or reverse order) was counted. The total number of conserved segments containing one or more marker genes was recorded to obtain the value of *a*, which together with *m* was used to calculate the probabilities in Equation 2. The value of *n* that maximized the probability was retained as $\hat{n}$.

To restrict the number of different simulation conditions, it was assumed that chromosome numbers remain constant following divergence from the common ancestor. Chromosome evolution involving duplication of chromosomes, followed by divergence of the duplicated chromosomes, is thus outside the realm of the results presented below.

Simulations were conducted for a variety of different combinations of *m*, *c*, *t*, and *i*. The choice of values for these parameters was guided by results from published investigations (Tanksley *et al.* 1992; Ahn and Tanksley 1993; Paterson *et al.* 1996; Nadeau and Sankoff 1997). For each combination of these parameters, the mean and variance of $\hat{n}$ was calculated over 500 simulation trials. A copy of the simulation program (written in FORTRAN) is available from the author on request.

Detection of heterogeneity in the rate of chromosomal evolution: Studies have shown that different lineages may undergo different rates of chromosomal rearrangement (Ehrlich *et al.* 1997), though there are few statistical tools for examining rate heterogeneity. Likelihood estimation as outlined above can be extended to the detection of heterogeneity in the rate of chromosomal rearrangement. One approach is to compare the estimated rate of chromosomal rearrangement for the taxa of interest with the rate(s) reported in studies of other taxa (Paterson *et al.* 1996; Lagercrantz 1998).

Let the MLE of chromosomal rearrangement occurring between two species, species A and B, be denoted as $\hat{n}_{AB}$. The estimate of chromosomal rearrangement reported between two other species C and D (scaled for the same estimated amount of time separating species A and B) is denoted as n_{CD}. The log-likelihood ratio test statistic follows from Equations 1 and 2 as

$$\Phi = -2[L(n_{\text{constrained}} \mid a, m) - L(\hat{n}_{AB} \mid a,m)], \quad (5)$$

where $n_{\text{constrained}}$ is the likelihood when n is constrained to a given value (*e.g.*, that of n_{CD}). The test statistic is distributed as χ^2 with 1 d.f. (Weir 1996). The sensitivity of this test depends on the number of markers used as well as on the amount of chromosomal evolution separating the species of interest from the reference species.

The approximate statistical power (probability of rejection of the null hypothesis) of the test was examined by simulation. Simulations were conducted, as described above, under a variety of input parameters (different combinations of *m*, *t*, *i*, and *c*). For each combination of input parameters, 500 simulations were conducted, and for each set of simulated data, the value of Φ was calculated for null hypotheses of $n_{\text{constrained}} = k\,\hat{n}_{AB}$ (where k is a constant that defines the null hypothesis in question). The proportion of cases where the value of the test statistic exceeded the critical value at the $P < 0.05$ and 0.01 levels gives an approximation of the statistical power of the test.

RESULTS

Maximum-likelihood estimates of chromosomal divergence and their variances (numerical analysis): The maximum-likelihood estimator returns the value of $\hat{n}$ expected for a^*. The likelihood peak becomes progressively sharper with increases in *m* (Figure 1). The asymptotic variance estimate of $\hat{n}$ is seen to be a decreasing curvilinear function of *m*. The effect on the estimation of *n* can be seen most clearly by examining the relationship of the coefficient of variation (CV) of $\hat{n}$ to *m* (Figure 2). The largest reductions in CV occur in the initial stages of mapping effort, but as *m* is increased beyond

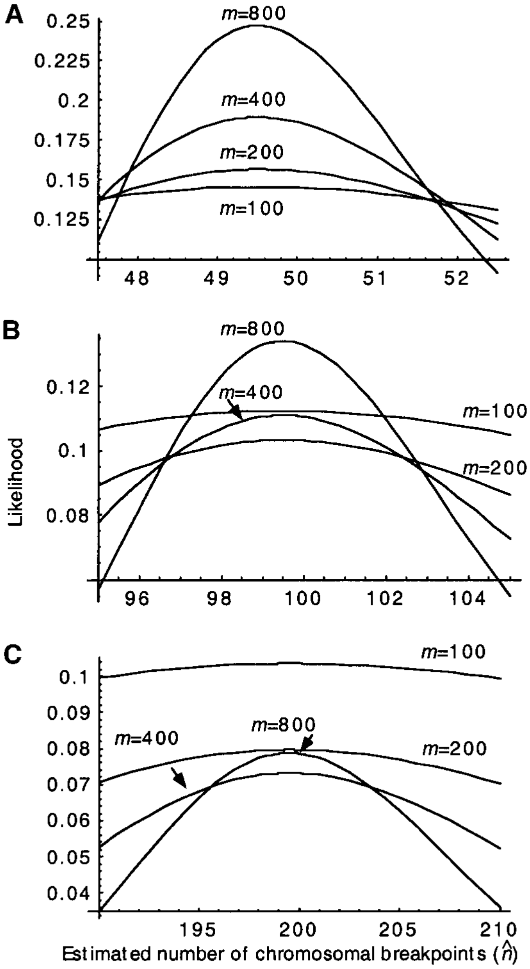

Figure 1.—Likelihood function under the random-breakage model evaluated numerically for different values of n and m. True values of n are as follows: (A) $n = 50$; (B) $n = 100$; and (C) $n = 150$.

several hundred markers, reductions in the variance of $\hat{n}$ become progressively smaller. For any given value of m, the CVs are larger when there are more chromosomal breakpoints separating the species in question (*i.e.*, true value of n large), but only marginally so (Figure 2).

Maximum-likelihood estimates of chromosomal divergence (simulation results): When chromosome evolution occurs via t translocation and i inversion events in a pair of species each having c chromosomes, the number of chromosomal breakpoints expected is $n = 2(t + i) + c$ (Sankoff and Nadeau 1996). Results obtained from the application of likelihood Equation 2 to the estimation of n with simulated data are shown in Figure 3.

When chromosome evolution occurs by translocation alone, and the density of markers is high relative to the number of chromosomes, the MLEs are close to their expected values (Figure 3, a and b). In the case where chromosome evolution occurs by translocation alone, and the number of markers is low relative to the number of chromosomes, there is significant departure of the MLEs from their expected values (Figure 3c). The occurrence of inversions contributes further to underestimation of the true value of n by the MLEs. It is in the range of 5–50% when inversions account for half of the rearrangements (Figure 3, d–f) and rises to nearly 70% when inversions account for all rearrangements (Figure 3, g–i). Again, underestimation is most pronounced when marker number is low relative to chromosome number. The basis for this underestimation of $\hat{n}$ is discussed below.

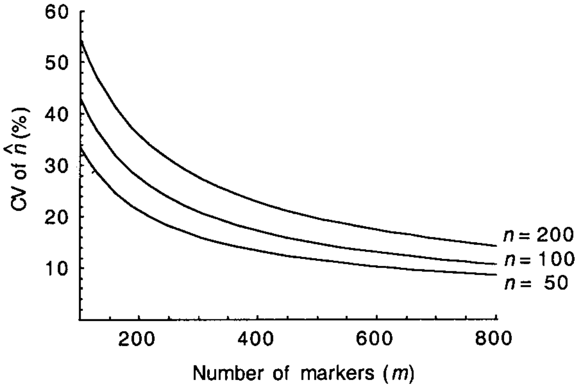

Figure 2.—Coefficient of variation of $\hat{n}$ under the random-breakage model evaluated numerically with different numbers of markers (m).

As seen in the numerical analysis, the CVs of the MLEs decrease with m in an accelerating manner (Figure 4). The effect of increasing the number of markers is most pronounced for the first few hundred markers. For instance, with $c = 20$ chromosomes, progressing from 100 to 200 markers reduces the CV by ∼50–60%, from 200 to 400 markers by 25%, and from 400 to 800 markers by ∼10%. For any given value of m, the CVs are slightly larger when there are more rearrangement events separating the species, but the difference becomes almost nil for $m \geq 400$. The relation of the CVs with m are similar for translocations and inversions (Figure 4).

Heterogeneity in the rate of chromosomal evolution (simulation results and illustration using published data): The power of the log-likelihood ratio test to detect heterogeneity in rates of chromosomal reshuffling increases as the number of markers placed on the maps increases, but for tests involving rate heterogeneity of >10%, the rate of gain in statistical power diminishes rapidly with marker number. These results are shown in Figure 5 for $c = 20$ chromosomes and $e = 90$ rearrangements. Nearly identical results were obtained for $c = 10$ and $c = 30$ (results not shown). The increase in power is roughly linear when rate heterogeneity between the lineages being compared is in the vicinity of 10% or less; but above 20% rate heterogeneity, the increase in power is decelerating with increasing marker numbers. When marker numbers are $>m = 200$, there are rapidly diminishing returns in power per marker

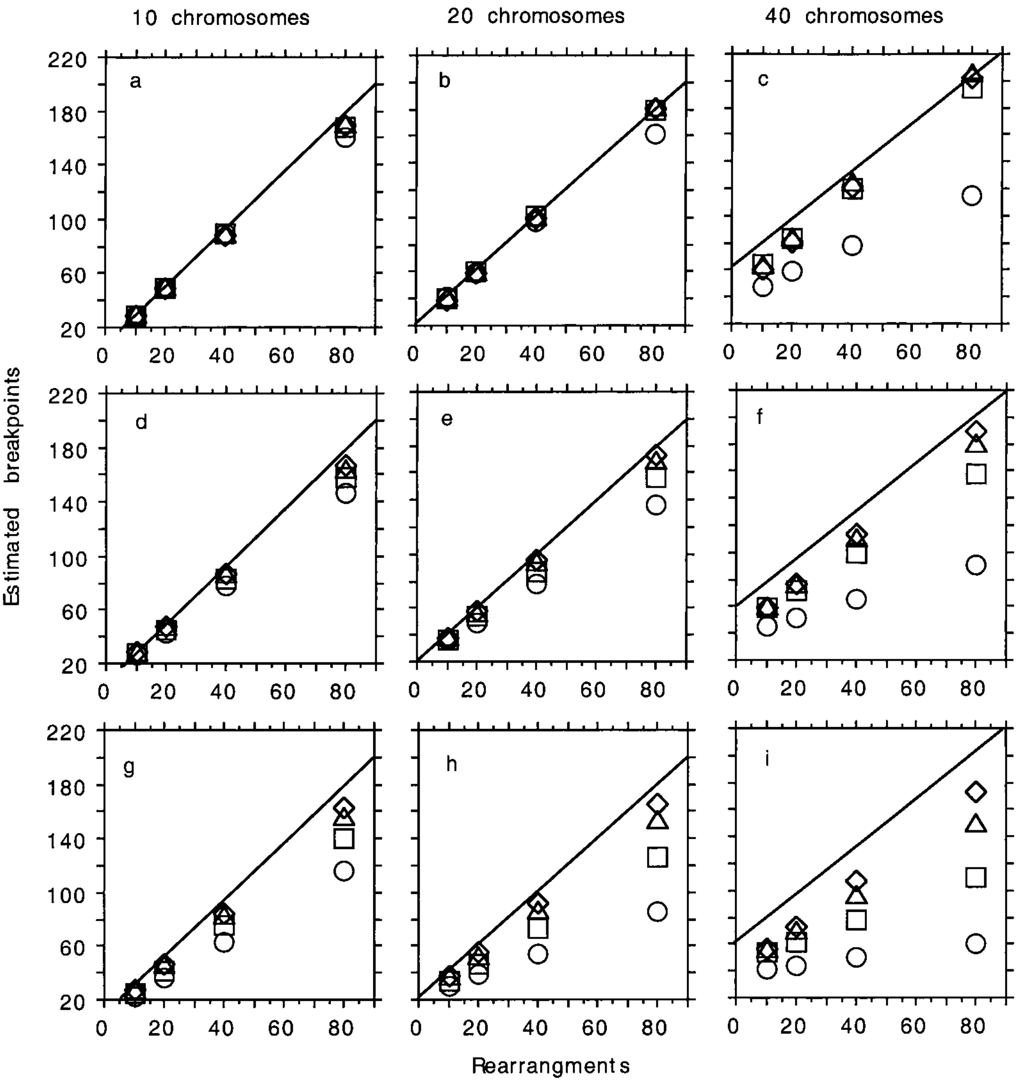

Figure 3.—Mean values of $\hat{n}$ obtained by application of the random-breakage model to estimation of chromosomal evolution. (a–c) Rearrangements by translocation only; (d–f) half of the rearrangements by translocation, half by inversion; (g–i) rearrangements by inversion only. Circles, $m = 100$; squares, $m = 200$; triangles, $m = 400$; diamonds, $m = 800$.

added to the maps. There is relatively little difference in the shapes of the power curves across the range of e values studied ($e = 15$–90) regardless of whether rearrangments were due to translocations or inversions (results not shown).

To illustrate the application of the log-likelihood ratio test, published comparative mapping data from *A. thaliana* and *Brassica nigra* were examined (Lagercrantz 1998). In this study, comparative mapping based on 284 markers uncovered 87 conserved segments. Estimation of n based on numerical evaluation of the likelihood Equation 2 gives an estimate of 124 breakpoints separating the species. This is higher than the estimate obtained by Lagercrantz (1998), who used the more conservative procedure of Nadeau and Taylor (1984) that does not consider segments marked by single loci. Lagercrantz (1998) notes that the two mustard family species may have diverged ~35 million years ago, and that the rate at which chromosomal rearrangement has occurred since their divergence is significantly greater than that seen in other plants and animals. Comparison of the results for the *A. thaliana-B. nigra* rate estimate with those obtained in other comparative mapping investigations (Paterson *et al.* 1996; Lagercrantz 1998) lend qualitative support to this conclusion. For instance, the next highest rate of chromosome rearrangement currently reported is the 13 rearrangements between Triticum and Secale that are estimated to have occurred over 6 million years. (Paterson *et al.* 1996). Scaling the Triticum-Secale estimate for the divergence time assumed above for Arabidopsis and Brassica gives 76 breakpoints. As this number is smaller than the 87 conserved segments observed by Lagercrantz (1998) in the Arabidopsis-Brassica comparison, a $L(n_{\text{constrained}} \mid a^*,m)$, where $n_{\text{constrained}} = 76$ cannot be calculated. If instead we take $n_{\text{constrained}}$ to be equal to 87, Φ can be calculated as

$$\Phi = -2[L(n_{\text{constrained}} = 87 \mid\mid a = 87, m = 284) - L(\hat{n}_{\text{AB}} = 124 \mid a = 87, m = 284)] = 42.73.$$

This result is highly significant ($P < 0.001$) and supports

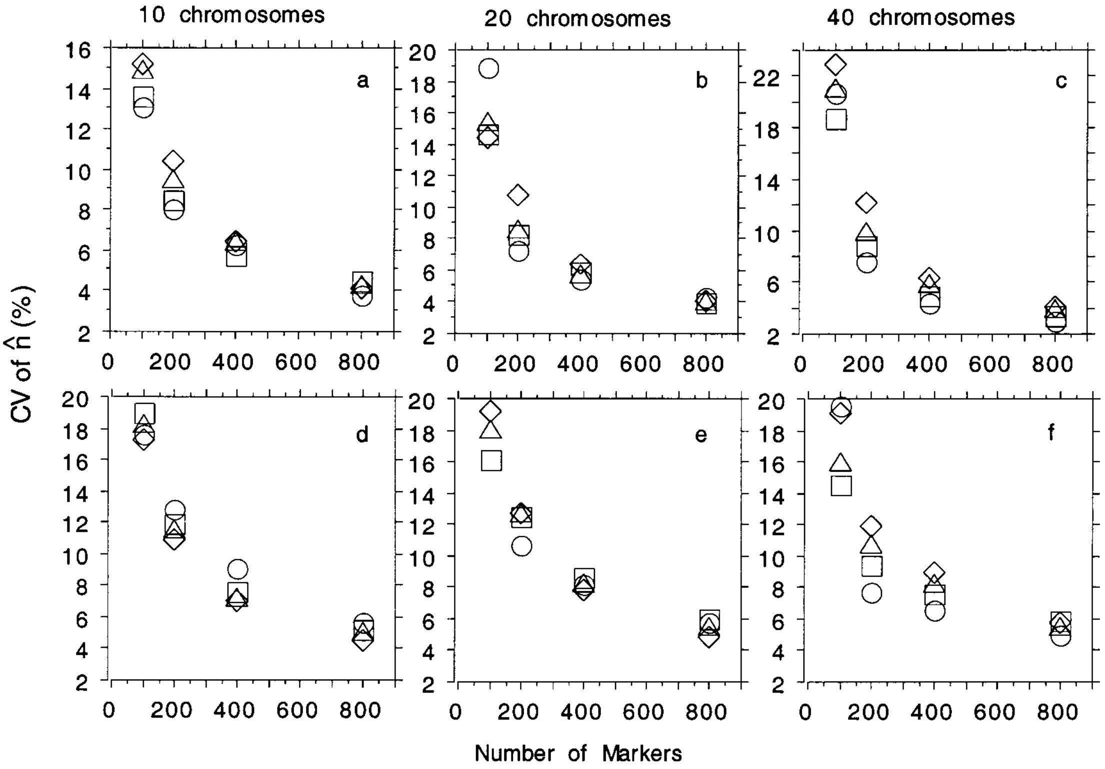

Figure 4.—Coefficient of variation of $\hat{n}$ *vs.* number of markers used in comparative mapping (m). Results are from the analysis of simulated data (see text). The coefficient of variation was calculated over 500 replicate simulations for each value of m, t, and i. Graphs shown here are for c = 20 chromosomes. (a–c) Translocations only; (d–f) inversions only. Symbols are as follows: circles, e = 50; squares, e = 90; triangles, e = 140; diamonds, e = 200.

Lagercrantz's conclusion that the rate of chromosomal rearrangement following divergence of Arabidopsis and Brassica has been unusually high.

DISCUSSION

Underestimates of *n*: The simulation results illustrate that when marker number is low relative to the number of chromosomes, or when rearrangements occur by both translocation and inversion, the number of breakpoints is underestimated under the random breakage model. This can be understood by considering Equations 1 and 2 together with some of the possible relationships that may arise between marker positions and chromosome rearrangement events, as illustrated in Figure 6. As noted, the MLE of n is a function of the total number of nonempty fragments (detected conserved fragments), a, observed in the comparative mapping study. Under the random-breakage model, probabilities of observing such fragments are defined by Equation 1. If, however, one considers the biological mechanisms by which chromosome breakpoints are generated (Figure 6), it becomes clear that there are several types of rearrangements of nonempty segments that may go undetected. Accordingly, the value of a obtained will be lower than expected under the random-breakage model.

One type of undetected rearrangement is an inversion that occurs in a segment containing a single marker (Figure 6a). Such an event is effectively "invisible" to the investigator, and the extent of underestimation can, in fact, be quantified when rearrangements arise only by inversion. Note that undetected inversions will occur with probability $P(r = 1)$ as defined by Equation 1. The expected number of rearranged segments, a^*, will, therefore, be reduced by the fraction $[1 - P(0) - P(1)]/[1 - P(0)]$. From Equation 4, the number of nonempty fragments (when all rearrangements occur by inversion) becomes

$$a_{\mathrm{I}}^* = [1 - P(0) - P(1)](n + 1). \quad (6)$$

Comparison of the MLE of n based on a_{I}^* reveals a relationship with m and a level of underestimation similar to that observed with simulation (Figure 7).

Translocations located in nonempty segments may also go undetected as illustrated in Figure 6b. When these types of events occur, a is underestimated, and the MLE of n is again underestimated. Compared with undetected inversions, however, undetected translocations are less likely to lead to underestimation of n, because they involve the sequential progression of several events. The problem is expected to occur most frequently when the number of markers per chromosome is low, a result that is supported by the simulations (Figure 3, a–c).

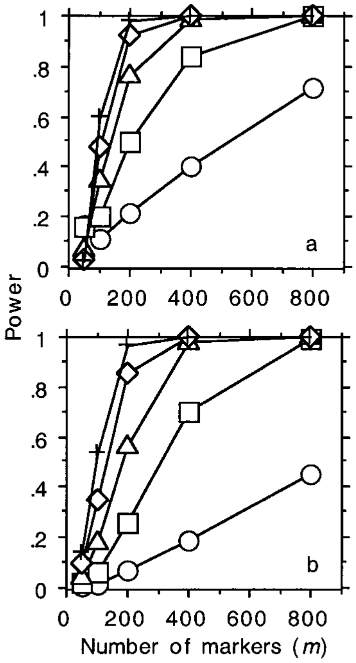

Figure 5.—Power curves for the log-likelihood ratio statistic applied to the detection of heterogeneity in the rate of chromosomal rearrangement. Results are for $c = 20$ chromosomes. Symbols are as follows: circles, $k = 1.1$; squares, $k = 1.2$; triangles, $k = 1.3$; diamonds, $k = 1.4$; and crosses, $k = 1.5$; see text for definition of k. (a) $t = 90$ translocations; significance level = 0.05. (b) $t = 90$ translocations; significance level = 0.01.

Variances of the MLE estimate of *n*: As progressively more conserved fragments are detected through comparative mapping of new markers, additional mapping effort will only marginally reduce the variance of the estimate of chromosome evolution (Figure 4). A similar relationship was found between mapping effort and ability to detect heterogeneity in the rate of chromosomal evolution (Figure 5). The actual values of the variances and CVs are significantly smaller (by 50% or more) than those obtained via numerical analysis based on the random-breakage model (Figure 2). This qualitative difference is not unexpected given that the simple random breakage model studied by numerical analysis assumes that chromosomal boundaries are distributed at random on the interval 0–1 (see above). While the variances seen using simulation are likely to be more representative than those calculated under the random-breakage model, nonrandom distributions of translocations and inversions could act to inflate variances obtained with actual data.

Marker numbers, estimation of *n*, and the detection of conserved functional gene blocks: There has been some discussion that blocks of genes found in conserved chromosome segments may represent gene combinations that interact functionally to produce important organismal characteristics (*e.g.*, blocks of genes that interact to produce characteristics closely related to organismal fitness; Bodmer 1975; Lundin 1979; Paterson *et al.* 1996). But because all genomes are interrelated, most colinear groups of genes detected in a comparative genomic investigation are likely to reflect nothing more than the limited number of genomic rearrangements following descent from a common ancestor. To move beyond the simple observation of large, conserved genome segments in the search for functionally related gene blocks, one requires knowledge of the "null" distribution of conserved segment lengths (*i.e.*, that expected from random chromosome reshuffling and descent from a common ancestor). If the number of breakpoints separating the species in question is known, along with the total lengths of their genomes (in centimorgans or base pairs), the mean number of rearrangements per unit genome length n/L (where L is the total genome length) can be calculated. Given this information, the probability distribution of no rearrangements in a segment of length x can be derived from the Poisson distribution as $P(x) = \exp(-nx/L)$ (see Nadeau and Taylor 1984). This distribution provides a benchmark against which to compare the observed distribution of conserved segment sizes. One may then ask whether there are segments that appear longer than expected given $\hat{n}$ and L. Such segments may have been selectively conserved due to their function. But because n is estimated, the distribution $P(x)$ is not known with certainty. The question arises, therefore, of how many markers are needed to compare observed with predicted segment length distributions. Applying the method of statistical differentials (Elandt-Johnson 1971) to obtain a variance estimate and 95% confidence interval around the calculated $P(x)$ (Figure 8), it is apparent that in the region of the distribution that one may wish to explore (*e.g.*, large and relatively rare segments, 20–30 cM and above in the example shown), the upper 95% confidence limit does *not* fall off as sharply with increasing marker density (as it does in the case of the CV of $\hat{n}$). These results suggest that in contrast to the other applications discussed above, a comparative genomics investigation that aims to detect selectively conserved chromosome segments by examining segment size distribution may benefit from the mapping of larger numbers of markers. Moreover, the comparison of observed and expected chromosome segment length requires that the true segment lengths and the total genome length be known. The segment lengths can be estimated from the observed distances between the outermost markers on each segment (see Nadeau and Taylor 1984), and genome length can be estimated given knowledge gained from recombination between markers (Chakravarti *et al.* 1991). These additional sources of variance have not been addressed here.

Marker numbers, the estimation of *n*, and exploratory surveys of genomic evolution: The results of this investigation have implications for applied studies and comparative evolutionary work based on comparative mapping. They suggest that studies of chromosome evolution based on low densities of markers (*e.g.*, <100–200

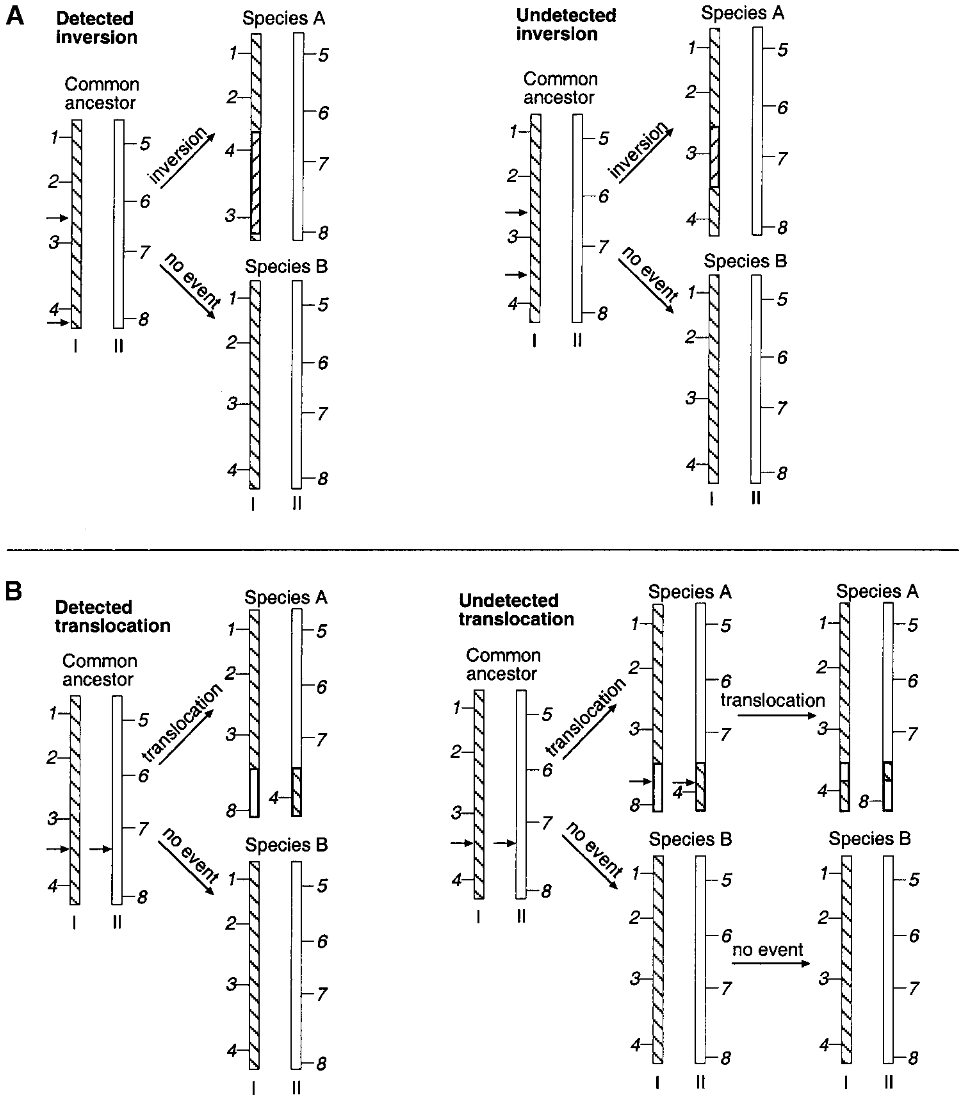

Figure 6.—Detection of chromosome rearrangements via comparative mapping in two species. (A) Detected and undetected inversions in nonempty chromosome segments. Nondetection occurs when segments contain only one marker. (B) Detected and undetected translocations in nonempty chromosome segments. Nondetection occurs when chromosome ends containing the same set of markers are translocated back and forth between a pair of chromosomes. Arabic numbers indicate positions of marker loci. Roman numerals indicate chromosomes. Short arrows mark the breakpoints resulting from inversion and translocation.

per genome) may underestimate the amount of chromosomal rearrangement, especially between taxa that are distantly related or in instances where inversion has played a large role in restructuring the chromosome. Ehrlich *et al.* (1997), who have used the random-breakage model to study chromosomal evolution in mammals, estimated that interchromosomal rearrangements have occurred roughly four times as often as intrachromosomal rearrangements following the divergence of humans and mice from their common ancestor. This is unexpected given the apparent strong selection against translocations. A relatively high ratio of translocations to inversions has also been reported in other investigations (Lagercrantz 1998). It is possible that some of the observed high ratios of translocations to inversions may be due to the inherent bias against detection of inversions as noted above.

Another issue is how many markers are required to obtain a low variance estimate of chromosomal rearrangement. When comparative mapping is used to examine the prospects of applying genetic map information from a well-characterized model species to a less well-characterized target species, the emphasis is often on uncovering candidate regions containing quantitative trait loci (*e.g.*, for genes contributing to yield or disease resistance; Lin *et al.* 1995; Paterson *et al.* 1995; Pereira and Lee 1995). The first objective is not a fine-scale comparative map, but rather the rough evaluation of the extent of conservation of synteny and gene order in the target group. Once a picture of this emerges, the investigator can determine whether additional map detail would greatly enhance the prospects of finding conserved segments containing the gene(s) of interest and marker(s). For the initial task, our results suggest that several hundred markers per species are sufficient. This means that if other (*e.g.*, related species) are of interest, comparative mapping effort could be allocated over more members of the target group. This has relevance to efforts aimed at uncovering and evaluating the potential of nontraditionally used germ plasm (*e.g.*, wild relatives of crop plants) as sources of useful genetic variation (Tanksley and McCouch 1997). In other

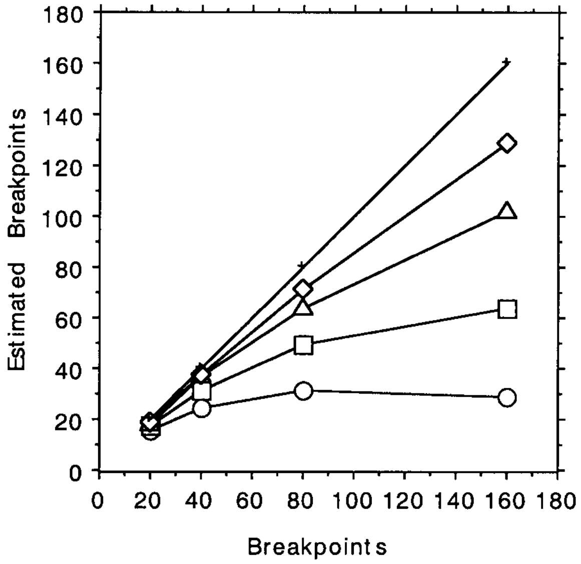

Figure 7.—Analysis of the underestimation of breakpoints arising when actual breakpoints (due to inversions containing only one marker) are not detected. Solutions obtained by numerical maximization of the likelihood Equation 2 when a^* is set equal to a_1^*. Circles, $m = 100$; squares, $m = 200$; triangles, $m = 400$; diamonds, $m = 800$; and crosses, expected estimate values.

words, a more useful division of comparative mapping effort in these types of investigations may be to spread effort across a larger number of candidate species rather than to pursue an ever more detailed comparative study of one or two species. A similar argument may hold when one is interested in using information on chromosomal rearrangement to construct a phylogeny or to compare rearrangement rates in different lineages (Ehrlich *et al.* 1997).

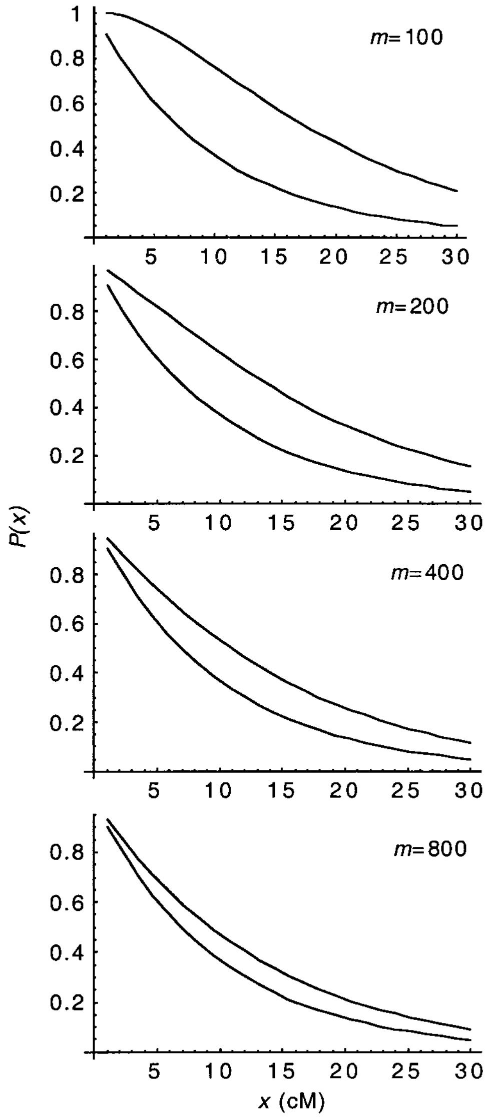

Figure 8.—Probability distribution of no rearrangements in a segment of length x (bottom line in each graph), and its top 95% confidence interval (top line) as obtained by normal approximation. This illustration assumes a total genome length $L = 2000$ cM and $n = 200$ breakpoints. The variance of $\hat{n}$ was estimated by the numerical analysis of the random-breakage model. Results shown in each graph are for different numbers of markers (m).

Conclusions: Apart from the bias against detection of inversion, the results presented in this investigation accord well with those of other studies in suggesting that estimation of numbers of chromosome breakpoints is robust to relatively small numbers of markers. For example, Nadeau and Sankoff (1998) have shown that as additional markers are included in a comparative mapping effort, the undetected but conserved segments become progressively smaller in number and in length. As well, estimates of genome rearrangement obtained with few markers have not changed substantially when many more markers are added (Nadeau and Taylor 1984; Copeland *et al.* 1993). It seems reasonable to conclude that much can be learned about the amount of gross chromosomal rearrangement from comparative mapping studies that use a moderate number of markers. In some species, however, factors such as many inversions, small-scale deletions and transpositions (*e.g.*, below the resolution provided by the marker density used), and large-scale duplications of entire chromosomes may render the task more difficult, and more research will be required to deal with the complications resulting from such events.

I thank Steve Wright, Andrew Paterson, and two anonymous reviewers for commenting on the material presented in this article. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

LITERATURE CITED

Ahn, S., and S. D. Tanksley, 1993 Comparative linkage maps of the rice and maize genomes. Proc. Natl. Acad. Sci. USA **90:** 7980–7984.

Bodmer, W. F., 1975 Analysis of linkage by somatic cell hybridization and its conservation by evolution, pp. 53–61 in *Chromosome Varia-*

tions in Human Evolution, edited by A. Boyce. Taylor and Francis, London.

Chakravarti, A., L. K. Lasher and J. E. Reefer, 1991 A maximum likelihood method for estimating genome length using genetic linkage data. Genetics **128:** 175–182.

Charlesworth, B., 1992 Evolutionary rates in partially self-fertilizing species. Am. Nat. **140:** 126–148.

Copeland, N. G., N. A. Jenkins, D. J. Gilbert, J. T. Eppig, L. J. Malatais *et al.*, 1993 A genetic linkage map of the mouse: current applications and future prospects. Science **262:** 57–66.

Ehrlich, J., D. Sankoff and J. H. Nadeau, 1997 Synteny conservation and chromosome rearrangements during mammalian evolution. Genetics **147:** 289–296.

Elandt-Johnson, R. C., 1971 *Probability Models and Statistical Methods in Genetics*. John Wiley & Sons, New York.

Gale, M. D., and K. M. Devos, 1998 Comparative genetics in the grasses. Proc. Natl. Acad. Sci. USA **95:** 1971–1974.

Lagercrantz, U., 1998 Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics **150:** 1217–1228.

Lin, Y. R., K. F. Schertz and A. H. Paterson, 1995 Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics **141:** 391–411.

Lundin, L.-G., 1979 Evolutionary conservation of large chromosomal segments reflected in mammalian gene maps. Clin. Genet. **16:** 72–81.

Nadeau, J. H., and D. Sankoff, 1997 Landmarks in the Rosetta Stone of mammalian comparative maps. Nat. Genet. **15:** 6–7.

Nadeau, J. H., and D. Sankoff, 1998 The lengths of undiscovered conserved segments in comparative maps. Mamm. Genome **9:** 491–495.

Nadeau, J. H., and B. A. Taylor, 1984 Lengths of chromosomal segments conserved since divergence of man and mouse. Proc. Natl. Acad. Sci. USA **81:** 814–818.

Ohno, S., 1967 *Sex Chromosomes and Sex-linked Genes*. Springer-Verlag, New York.

Paterson, A. H., Y.-R. Lin, Z. Li, K. F. Schertz, J. F. Doebley *et al.*, 1995 Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science **269:** 1714–1718.

Paterson, A. H., T. H. Lan, K. P. Reischmann, C. Chang, Y. R. Lin *et al.*, 1996 Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. Nat. Genet. **14:** 380–382.

Pereira, M. G., and M. Lee, 1995 Identification of genomic regions affecting plant height in sorghum and maize. Theor. Appl. Genet. **90:** 380–388.

Sankoff, D., and J. H. Nadeau, 1996 Conserved synteny as a measure of genomic distance. Disc. Appl. Math. **71:** 247–257.

Sankoff, D., M.-N. Parent, I. Marchand and V. Ferretti, 1997 On the Nadeau-Taylor theory of conserved chromosome segments, pp. 262–274 in *Combinatorial Pattern Matching*, edited by A. Apostolico and J. Hein. Springer-Verlag, New York.

Stebbins, G. L., 1971 *Chromosomal Evolution in Higher Plants*. Addison-Wesley, Reading, MA.

Tanksley, S. D., and S. R. McCouch, 1997 Seed banks and molecular maps: unlocking genetic potential from the wild. Science **277:** 1063–1065.

Tanksley, S. D., M. W. Ganal, J. P. Prince, M. C. De Vicente, M. W. Bonierbale *et al.*, 1992 High density molecular linkage maps of the tomato and potato genomes. Genetics **132:** 1141–1160.

Van Deynze, A. E., M. E. Sorrells, W. D. Park, N. M. Ayres, H. Fu *et al.*, 1998 Anchor probes for comparative mapping of grass genera. Theor. Appl. Genet. **97:** 356–369.

Weir, B. S., 1996 *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer, Sunderland, MA.

Communicating editor: A. G. Clark