Heredity (2006) 96, 290–297

www.nature.com/hdy

Testing for segregation distortion in genetic scoring data from backcross or doubled haploid populations

J Nixon
Molecular Genetics, Agriculture and Agri-Food Canada, Saskatoon Research Centre, 107 Science Place, Saskatoon, SK, Canada S7N 0X2

It is important that breeders have the means to assess genetic scoring data for segregation distortion because of its probable effect on the design of efficient breeding strategies. Scoring data is usually assessed for segregation distortion by separate nonindependent χ^2 tests at each locus in a set of marker loci. This analysis gives the loci most affected by selection if it exists, but it cannot give a statistically correct test for the presence or absence of selection in a linkage group as a whole. I have used a combined test based on the statistic, which is the most significant *P*-value from the above tests, called the *single locus* test. I have also derived mathematically a new combined statistical test, the *overall* test, for segregation distortion that requires genetic scoring data for a single linkage group. This test also takes genetic linkage into account. Using a range of marker densities and population sizes, simulations were carried out, to compare the power of these two statistical tests to detect the effect of selection at one or two loci. The *single locus* test was always found to be more powerful than the *overall* test, but the *single locus* test required a more complicated *P*-value correction. For the *single locus* test, approximate correction factors for the *P*-values are given for a range of marker densities and genetic lengths.
Heredity (2006) **96,** 290–297. doi:10.1038/sj.hdy.6800797; published online 1 March 2006

Keywords: statistical test; segregation distortion; genetic scoring data; selection; viability; duplicate loci

Introduction

Backcross or doubled haploid populations descended from a single F_1 genotype are commonly generated in plant breeding programmes and plant genetics experiments. There is evidence that selection can occur during these experiments, for example, selection occurring on the gametes. This has often been reported particularly during the process of culturing male gametes and anthers to produce double haploid plantlets (Cloutier *et al*, 1995; Foisset *et al*, 1997; Haitham *et al*, 2002; Knox and Ellis, 2002; Lu *et al*, 2002). Differential selection will manifest itself in segregation distortion that is departures from the expected 1:1 ratio of allele pairs in the progeny of the F_1 individual at heterozygous loci. Whether or not selection happens is likely to be important in the interpretation of the results of such experiments and the design of breeding programmes. Appropriate data are frequently generated from some of the many types of highly reproducible and relatively inexpensive molecular marker systems (Botstein *et al*, 1980; McCouch *et al*, 1988; Williams *et al*, 1990; Vos *et al*, 1995) that are capable of providing information on the parental origin of alleles at a closely spaced set of marker loci.

A related issue is to identify markers for which it is hard to score the presence of one or both of the parental alleles (whether this is suspected by the experimenter at the time or not) that could lead to falsely distorted segregation ratios; these unreliable observations should be removed from the data at an early stage because they can artificially increase the length of genetic maps (Knox and Ellis, 2002) that are used in gene cloning and gene introgression. A related issue is when a marker locus is masked by another copy of the locus with the same alleles but at a different location on the genetic map. There is much evidence (Frisch *et al*, 2004) that duplicated segments of chromosomes are common in eukaryotes, particularly in plants. If a molecular marker is based on a part of a duplicated sequence this can result in a duplicate pair of marker loci. As the bands for the markers cannot be distinguished, the result is, in the initial phase of genetic map construction, a single 'ghost marker' instead of the pair of loci that will show segregation distortion that is related to recombination between the duplicate loci.

The usual way in which segregation distortion is assessed for marker data is by chi-square tests on a locus by locus basis, (Haitham *et al*, 2002; Lu *et al*, 2002; Ruiz and Asins, 2003). The χ^2 test is a convenient large sample approximation of the exact binomial test with $P = 0.5$. In this test, the *P*-value to be associated with an observation of the number of individuals with one of the alleles is twice the sum of the probabilities in the binomial distribution from the observed value to the tail of the distribution (Brunk, 1975). Although this analysis can give an indication of the loci most affected by selection if it exists, and the direction in which selection is acting, it cannot give a statistically correct overall test for the presence or absence of selection in a single linkage group because the many separate statistical tests have not been combined into a single overall statistical test. One simple way to do this is to carry out the χ^2 test separately at each locus as above to assess the significance of departures from the expected 1:1 ratio of allele frequencies, and record the

Correspondence: J Nixon, Molecular Genetics, Agriculture and Agri-Food Canada, Saskatoon Research Centre, 107 Science Place, Saskatoon, SK, Canada S7N 0X2. E-mail: nixonj@agr.gc.ca
Received 25 November 2004; accepted 13 December 2005; published online 1 March 2006

smallest P-value obtained as the test statistic. I shall refer to this as the *single locus* test. By simulating scoring data based on probability models of recombination, without selection, and specified marker placement on the genetic map, I have empirically determined the relationship between the smallest P-value for the single locus tests and the effective P-value with which it should be associated in the test for the presence of segregation distortion over the whole linkage group. This will allow an approximate correction to be made to the *single locus* test.

There are other methods to construct a combined test for the case when the separate statistical tests are carried out on n independent datasets and the aim is to test the null hypothesis which states that the null hypothesis holds for each of the datasets. Under this composite null hypothesis, the joint distribution of all the P-values is uniform on the hypercube $[0.1]^n$. To construct a combined test of the null hypothesis, a suitable rejection region in this space must be specified. If the combined test is such that it rejects the composite null hypothesis if and only if any of the n individual tests reject the null hypothesis (each with the same probability α), then the combined test rejects the null hypothesis with probability $1-(1-\alpha)^n$. This probability is the corrected α value for the combined test. However, such methods of constructing combined tests are not applicable here because the tests for the different loci are not independent because a small recombination fraction between one locus and the next limits the amount by which their respective segregation ratios differ. To address this problem, a novel statistical test of segregation distortion for genetic scoring data for a single linkage group that takes both multiple testing and genetic linkage into account was derived. As this test takes segregation distortion into account at all the marker loci chosen for the analysis simultaneously, I shall call it the *overall* test. This procedure allows for any pattern of missing data. It was derived mathematically in an attempt to avoid the need for an empirical correction derived from computer simulations. Simulations were, however, carried out for both these tests to compare their power (ie the probability that the test will detect a true selection effect), with selection at one or two loci, using a range of marker densities. The results described here will allow users of these tests to (i) decide which test to use based the power to detect selection effects and ease of use, for example, whether a correction to the nominal P-value is required, (ii) decide what marker set to use, and (iii) correct the nominal P-value obtained, if necessary.

Many investigators have considered the problem of testing for segregation distortion based on fitting models describing selection (ie viability) at loci whose distance from the nearest marker(s) in the genetic map are included in parameters to be fitted. Much of this work has been recently reviewed (Luo *et al*, 2005). In their paper that extends earlier work, the formulation of the selection model for F2 data also includes the two degrees of freedom (eg selection coefficient and degree of dominance) to describe selection at a single locus and the concept of liability that allows for an environmental variable. As the authors demonstrate, the 'covariate' can allow removal of some of the residual error and therefore increase the power to detect selection effects.

However, the approach that I am using in this work applies to backcross data (but could be extended to other mating designs) and amounts to restricting the search for models in parameter space such that selection loci are coincident with any of the marker loci. Advantages of this approach are that it is relatively easy to implement, runs quickly, and it could be used to screen large amounts of data before any further analysis of segregation distortion is applied. It is appropriate for marker maps that are dense, because then it is less likely to be worth the effort (and would be a waste of time if the confidence interval for the selection locus was comparable with the marker spacing) to use an interval mapping strategy to map the selection locus between markers.

Methods

Testing the statistical tests using simulated populations of chromatids

Before any statistical test is applied to experimental data, it is necessary to check (1) that by applying the test repeatedly to data that are randomly generated under the assumptions of the null hypothesis, a uniform distribution of P-values is obtained, and (2) that there is a high probability that the P-value obtained from such simulated data under the alternative hypothesis is less than a given level say 0.05. This probability is known as the power of the statistical test because it is the probability that the test will reject the null hypothesis when it is false, where the procedure is to reject the null hypothesis if and only if the P-value obtained is less than the previously agreed value known as the level of the test. Ideally for a statistical test, the power should be as large as possible over a wide range of alternative hypotheses.

Using computer simulation, it is possible to check assumption (1), and if it holds, to make an empirical estimate of the power of the test against any given alternative hypothesis by examining the distribution of P-values in this case. To do this, the P-values are recorded, and afterwards sorted into increasing order to plot their cumulative probability distribution. This cumulative probability is an estimate of the probability that a given P-value or less will be observed, so the plot is a graph of the power of the test against the level of the test.

Furthermore it is possible to make an empirical observation of the relationship between the power and the level α of the test, for any given alternative hypothesis, after correcting for any departure of the distribution of P-values under the null hypothesis from the uniform distribution. In this method, the P-value obtained is treated like any other test statistic. This method uses the empirical distributions of P-values under both the alternative and null hypotheses. For each P-value generated under the alternative hypothesis, the proportion of the P-values generated under the null hypothesis that are smaller than this value (ie the cumulative probability) gives the empirical corrected P-value to be associated with it. Therefore, to give the corrected plot, the two lists of P-values for the alternative and null hypotheses are sorted into increasing order to give $p_1,\ldots p_N$ and $q_1,\ldots q_N$, respectively. Then the power, $y=i/N$ is plotted against the level, $x=1/N$ (position of p_i in $\{q_1,\ldots q_N\}$) for points that differ in x or y by at least some small number δ.

This corrected power function is the relationship between the empirically determined power of the statistical test (probability that the test statistic is less than a specified value under the alternative hypothesis) and the empirically determined α value (probability that the test statistic is less than the same specified value under the null hypothesis).

For simulating the scoring data, chromatids (determined by the parental origin of the zero end of the chromatid, the number and positions of crossovers) from an F_1 individual were simulated according to the random model (Haldane, 1919), which was the same model used in the derivation of the *overall* test. In this model, the number of crossovers is determined by the Poisson distribution with mean equal to the genetic length of the chromatid in Morgans, and the crossover positions are uniformly and independently distributed on the model chromatid.

For simulating a population showing segregation distortion, such randomly generated chromatids were subjected to a selection process that consisted of a possible culling. If the chromatid was culled, a new chromatid was randomly generated in the same way to take its place and was subjected to the same selection process, and the procedure was repeated until a chromatid survived. The selection process consisted of one or two stages. At each stage, the chromatid was culled with a fixed probability if its genotype at a specific locus was '+'.

This procedure was repeated until the population had been generated. The simulated scoring data was generated from each population of randomly generated chromatids using the specified marker positions and assuming that no scoring errors occur. The statistical test for the presence of segregation distortion was applied to each such set of simulated scoring data, and the whole process was repeated to generate the *P*-values from *N* such populations. Using repeated sets of simulations, it was found that (for Figures 1, 2 and 3) sets of $N = 10\,000$ simulated populations with $\delta = 0.002$ gave adequate reproducibility and resolution in the resulting graphs, but for Figure 4 the values $N = 10^5$ and $\delta = 10^{-4}$ were used.

The two statistical tests were compared for a range of different situations using (1) the corrected power function and (2) the empirically determined *P*-value as a function of the nominal *P*-value, both determined as outlined above.

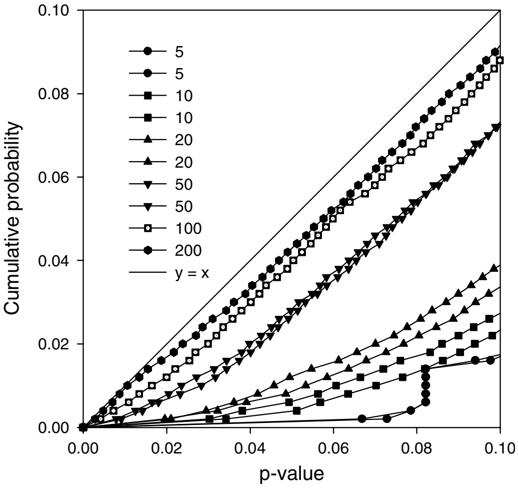

Figure 2 The probability distribution of *P*-values in the range 0–0.1 from the *overall* test of segregation distortion using 10 000 randomly generated populations of chromosomes with genetic length 1 M satisfying the random model with no selection. The 11 markers were placed at 10 cM intervals. The legend shows the population sizes. For several population sizes, sets of 10 000 populations were repeatedly generated and analysed to give an indication of the reproducibility of the curves. The curves rapidly approximate to the line $y = x$ (particularly for the larger population sizes) for larger *P*-values, consistent with the behaviour of a correctly constructed statistical test under the null hypothesis.

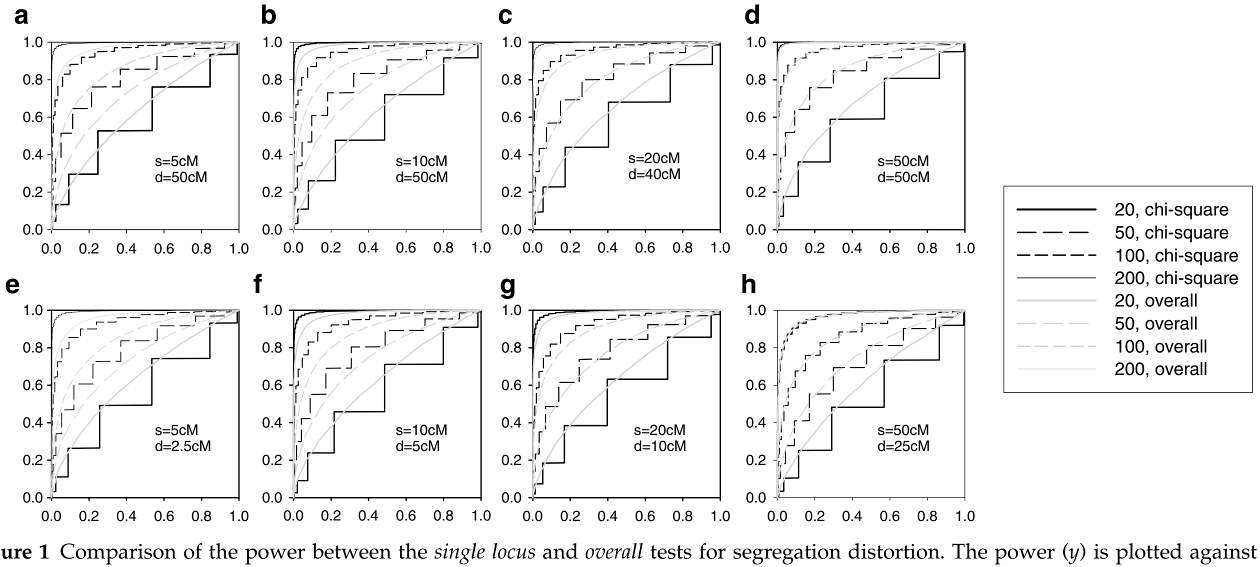

Figure 1 Comparison of the power between the *single locus* and *overall* tests for segregation distortion. The power (y) is plotted against the level (x) (often denoted by α) of the test using simulated data from 10 000 populations of chromatids assayed with equally spaced markers that include the endpoints. The random model of a 1 M chromosome is assumed with 50% survival of one allele at different loci. The population sizes are indicated in the legend. The marker spacing *s* and the genetic distance *d* from one end of the chromosome to the locus of selection are indicated in the figures.

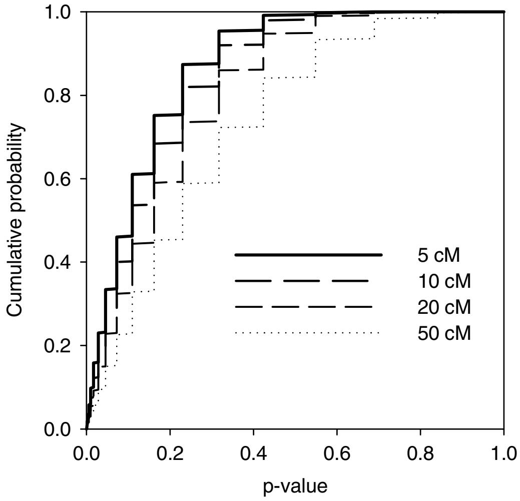

Figure 3 The probability distribution of P-values from the *single locus* test of segregation distortion using 10 000 randomly generated populations of 100 chromosome pairs with genetic length 1.0 (M) satisfying the random model and no selection. The markers were placed at both ends of the model chromosome and at equal intervals between them, with their spacing shown in the legend.

Derivation of the statistical tests

In these derivations, the scorings will be designated as '+' and '−' according to the parent of the F_1 individual that donated the allele to the F_1 individual.

The single locus test

The probability distribution of the number of individuals scored as '+', n^+, under the null hypothesis of no segregation distortion, is binomial with $P = 0.5$ and so has mean $n/2$ and variance $n/4$, where n is the number of successfully scored individuals. Approximating this by the Normal distribution with the same mean and variance gives that $(n^+ - n/2)/\sqrt{n/4}$ has a standard Normal distribution or equivalently that its square $(n^+ - n/2)^2/(n/4)$ has a χ^2 distribution with one degree of freedom. The P-value is obtained from this for each locus in the set, and the most significant (smallest) value is recorded as the result.

Calculation of the covariance between the proportions of individuals scored as + at two linked loci in a backcross population assuming that there is no segregation distortion.

Suppose that m_1 and m_2 are two marker loci with a recombination probability r between them and let s_1 and s_2 be the corresponding scorings at those loci for an

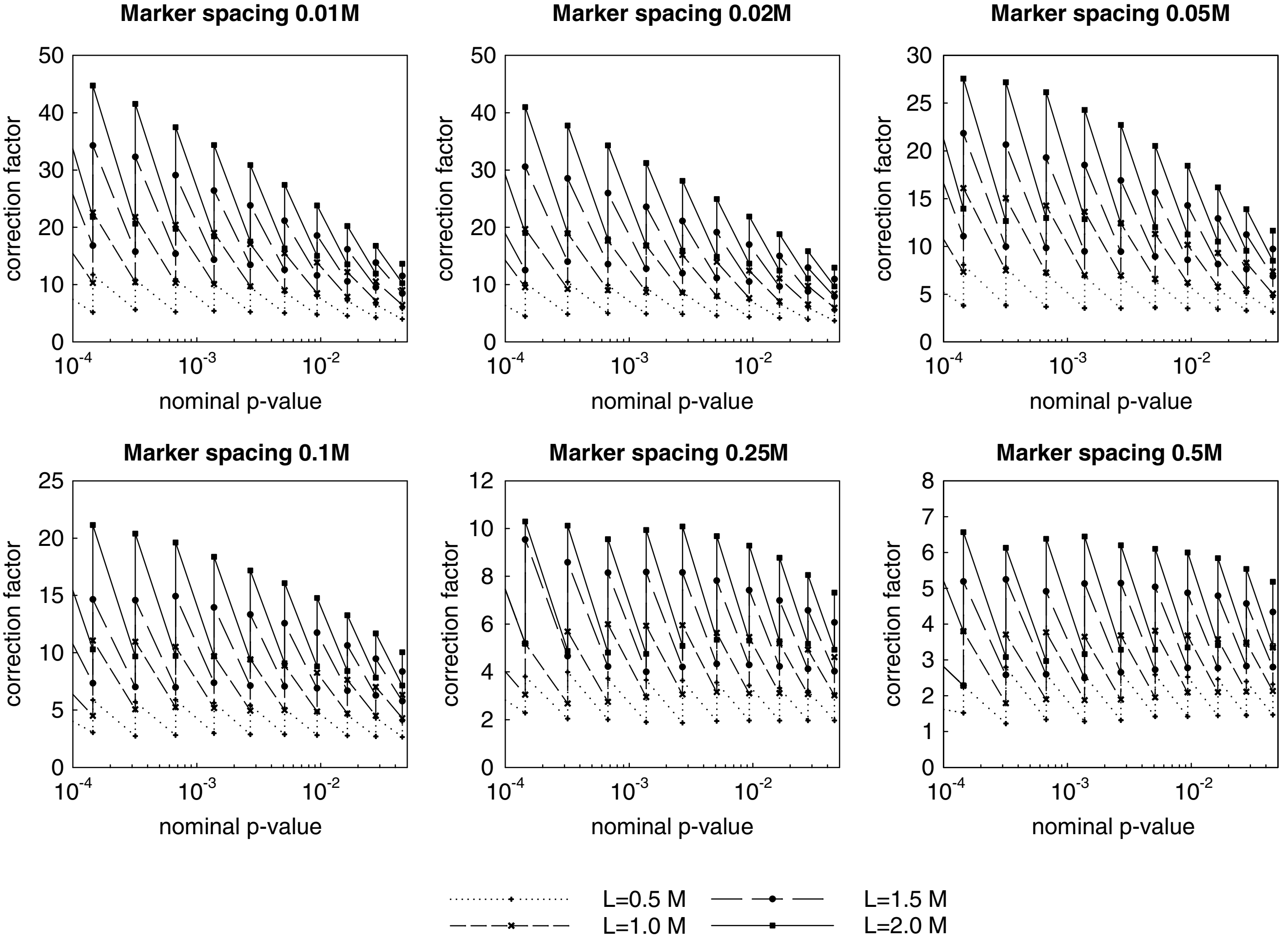

Figure 4 Approximate correction factors for the *single locus* test of segregation distortion for simulated populations of 100 random model chromosomes with a range of genetic lengths (L), assayed with equally spaced markers that include the endpoints. The most significant P-value found from the single-locus chi-square tests of segregation distortion should be multiplied by this factor, regardless of the population size, to get an approximate corrected overall P-value for the statistical test for segregation distortion.

individual. Suppose also that out of N individuals in the whole population, n_1 have scoring data at locus m_1 but not at m_2, and n_1 have scoring data at locus m_2 but not at m_1, and n have data at both loci, and the remainder have no data at either locus. In any set of individuals S, the number in the subset of S that have a specified scoring will be denoted by the symbol for the number of individuals in S, with a superscript specifying the scoring. For example, n_1^+ is the number of individuals that have scoring at locus m_1 but not at m_2 and have the scoring '+'. P will denote probability. Then $P(s_1 = `+') = P(s_1 = `-') = 0.5$ and likewise for s_2. Also the probability that $s_2 = `+'$, given that $s_1 = `+'$ which is denoted by $P(s_1 = `+' \mid s_1 = `+')$ is $1-r$ and is the probability that there is no recombination between s_1 and s_2. Similarly, $P(s_2 = `+' \mid s_1 = `-') = r$. If the two-locus genotypes are designated by $s_1 s_2$ that is $++$, $+-$, $-+$, or $--$ then $P(++) = P(--) = (1-r)/2$ and $P(+-) = P(-+) = r/2$ and since the n individuals are independent, the joint frequency distribution of (n^{++}, n^{+-}, n^{-+}, n^{--}) has a multinomial form and hence the covariance of the frequencies of two different outcomes is minus n times the product of their separate probabilities for example, $\mathrm{Cov}(n^{++}, n^{+-}) = \mathrm{Cov}(n^{++}, n^{-+}) = -n(\frac{1-r}{2})(\frac{r}{2})$ and $\mathrm{Cov}(n^{+-}, n^{-+}) = -n(\frac{r}{2})^2$. Also $\mathrm{Cov}(n^{++}, n^{++}) = \mathrm{Var}(n^{++}) = n(\frac{1-r}{2})(\frac{1+r}{2})$ ($= npq$ where p is the probability of $++$ and q is $1-p$). Hence the covariance between the proportions of + scorings at the two loci for the n completely scored individuals is

$$\begin{aligned} &\mathrm{Cov}\left(\frac{n^{++} + n^{+-}}{n}, \frac{n^{++} + n^{-+}}{n}\right) \\ &= \frac{1}{n^2}\mathrm{Cov}(n^{++} + n^{+-}, n^{++} + n^{-+}) \\ &= \frac{1}{n^2}\left[\frac{n}{4}(1-r)(1+r) - \frac{n}{2}(r - r^2) - \frac{nr^2}{4}\right] \\ &= \frac{1}{n}\left(\frac{1}{4} - \frac{1}{2}r\right) \end{aligned} \qquad (1)$$

This formula applies to two loci for which the recombination fraction between them is r, there is no selection leading to segregation distortion, and the loci are scored (without errors) with no missing data points for n individuals. This result can be extended to the situation where there are missing data points with n, n_1, n_2 and N all fixed. Then the required covariance is

$$\begin{aligned} &\mathrm{Cov}\left(\frac{(n^{++} + n^{+-}) + n_1^+}{n + n_1}, \frac{(n^{++} + n^{-+}) + n_2^+}{n + n_2}\right) \\ &= \frac{1}{(n + n_1)(n + n_2)} \times [\mathrm{Cov}(n^{++} + n^{+-}, n^{++} + n^{-+}) \\ &\quad + \mathrm{Cov}(n_1^+, n^{++} + n^{-+}) + \mathrm{Cov}(n^{++} + n^{+-}, n_2^+) \\ &\quad + \mathrm{Cov}(n_1^+, n_2^+)] \end{aligned} \qquad (2)$$

The last 3 terms are zero, because in each of these cases the two arguments of the covariance refer to disjoint sets of individuals. Using equations (1) and (2) can be written as

$$\begin{aligned} &\mathrm{Cov}(\text{proportion} + \text{on locus } 1, \text{ proportion} + \text{on locus } 2) \\ &= \frac{1}{(n + n_1)(n + n_2)} n\left(\frac{1}{4} - \frac{1}{2}r\right) \end{aligned} \qquad (3)$$

The derivation of the *overall* test also needs the following procedure:

From a set of m observed values, the procedure to test the hypothesis that specifies mean values for each of these observations, assuming that the observations come from a multivariate Normal distribution with a known variance-covariance matrix.

Given an m-dimensional vector $\underline{x}_0$ of observations, and an m-dimensional multivariate Normal distribution $N(\underline{\mu}, \Sigma)$ with unknown mean μ and known variance-covariance matrix Σ, which is the assumed probability distribution from which the observation $\underline{x}_0$ came, I want to derive a statistic to test for deviations from the null hypothesis $\mu = \mu_0$. Under this hypothesis, the probability density for $\underline{x}$ is

$$f(\underline{x}) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underline{x} - \underline{\mu_0})'\Sigma^{-1}(\underline{x} - \underline{\mu_0})\right] \qquad (4)$$

If the new vector variable $\underline{y}$ is introduced by $\underline{y} = \Sigma^{-1/2}(\underline{x} - \underline{\mu_0})$ then its transpose is $\underline{y'} = (\underline{x} - \underline{\mu_0})'\Sigma^{-1/2}$, and $\partial y_j/\partial x_i = (\Sigma^{-1/2})_{ji}$, so the Jacobian of the transformation is $J = |\Sigma^{-1/2}|$ and the probability density of $\underline{y}$, $g(\underline{y})$, is given by

$$\begin{aligned} g(\underline{y}) = J^{-1}f &= \frac{1}{(2\pi)^{m/2}} \exp\left[-\frac{1}{2}y'y\right] \\ &= \frac{1}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\Sigma y_i^2\right). \end{aligned}$$

This is the probability density for m independent standardised Normal random variables, that is, y_1, y_2 … y_m are independent Normal random variables with mean zero and variance 1. Hence, the test statistic should be based on the distance from the transformed point $\underline{y}_0 = \Sigma^{-1/2}(\underline{x}_0 - \underline{\mu}_0)$ to the mean in $\underline{y}$ space, which is zero. The square of this distance is

$$\begin{aligned} \underline{y}'_0\underline{y}_0 &= (\underline{x}_0 - \underline{\mu}_0)'\Sigma^{-1}(\underline{x}_0 - \underline{\mu}_0) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}(x_{0i} - \mu_{0i})\Sigma_{ij}^{-1}(x_{0j} - \mu_{0j}) \end{aligned} \qquad (5)$$

which has a χ^2 distribution with m degrees of freedom.

The overall test for segregation distortion

This test is based on the following approximations:

(1) the proportions of + scoring, x_i, whose variance-covariance matrix can be computed as above, are regarded as continuous variables.

(2) the joint distribution of the x_i in the null hypothesis is approximated by the multivariate Normal distribution with the same variance-covariance matrix and means (which are all 0.5). The parameters in this distribution are then regarded as known exactly but in fact the variance-covariance matrix is only estimated; this is because the recombination fractions used to estimate the covariance of xs between different loci, are themselves estimates based on the calculation of map distances, which also have error.

The calculation procedure was as follows for each set of simulated scoring data representing a linkage group:

(1) To avoid a singular variance-covariance matrix, each pair of loci was checked for recombination, and if none was found, one of the pair of loci was deleted after any scoring information it carried had been transferred to the other locus.
(2) For each of the m distinct loci in the linkage group, the fraction x of $+$ scorings was computed. These values constitute the vector $\underline{x}_0 = \{x_1, x_2 \ldots x_m\}$ in equation (5).
(3) Estimates of the recombination fraction r between each pair of loci in the linkage group were obtained, using the map distances between them computed using the Haldane mapping function and summing the genetic distances between adjacent loci. For the purpose of the simulation tests, it was necessary to avoid chance infinite values of map distances by arbitrarily assigning the large map distance 2 M to recombination fractions r that exceeded 0.5.
(4) From the recombination fractions, the corresponding covariances between pairs of members of $\underline{x}_0$ were estimated from equation (3). The variance of each member of $\underline{x}_0$ is $p(1-p)/n$ that is $1/(4n)$ where $P = 0.5$ and n is the number of individuals without missing data at that locus. In this way the complete expected variance-covariance matrix was constructed.
(5) The inverse of the variance-covariance matrix was calculated.
(6) The right-hand side of equation (5) was calculated where $\mu_{0i} = 0.5$ and $\underline{x}_0$ was as above; this expression has a χ^2 distribution with m degrees of freedom from which the P-value can be computed by standard methods.

Results

The *single locus* test and *overall* test for testing for the presence of segregation distortion due to selection were compared by simulation for a range of different situations to determine

(1) which statistical test is more powerful for detecting selection effects,
(2) under which conditions the power is greatest for the same level, and
(3) whether either of the tests requires the correction to ensure that the level of the test is correct (for example, a test performed at the nominal level 0.05 may in fact be a test performed at level 0.01), and if so,
(4) whether there is a simple procedure for making the correction.

If very few errors can be tolerated (probability α) when the no-selection case (null hypothesis) is true, it is often not possible to make a decision in favour of the alternative hypothesis when this is true (the probability of this is the power), but conversely if relatively frequent errors in the identification of the no-selection case are acceptable, a greater frequency of decisions in favour of the alternative hypothesis can be made with the same data. Thus the power of a test always increases with α. In fact, the power of a test always increases from 0 to 1 as α increases from 0 to 1.

The greatest interest is in the behaviour of the power for small values of α. The graphs (Figure 1) of power against α for the *single locus* test are stepped because there are only a discrete set of possible P-values. This set of values is the same for every locus and they correspond to the segregation ratios $n/2{:}n/2$, $n/2-1{:}n/2+1$, $n/2-2{:}n/2+2$ etc. where n is the population size. This effect is not seen in the *overall test* because the segregation ratios at all loci are taken into account simultaneously.

The initial investigations were carried out with chromosomes 1 Morgan in length using the random model for chiasma distributions and with marker loci equally spaced such that one marker locus was at each end of the chromosome. The marker spacings were 5, 10, 20 and 50 cM and the population sizes were 20, 50, 100 and 200. One stage selection with 50% survival of the '+' allele at single loci was used (Figure 1), with loci of selection (1) at marker loci nearest the centre of the chromosome where the tests should be the most powerful or (2) midway between markers at an end of the chromosome where the tests should be the least powerful. This was verified for markers spaced at 50 cM (data not shown). As expected, the power of both tests always increased with population size and the power was greater for case (1) (Figure 1a–d) than for case (2) (Figure 1e–h).

The *overall* test

The power of this test increased with increasing distance between markers for the case (1) above (Figure 1a–d), but for case (2) the test had optimum power for a marker spacing near 20 cM (Figure 1g). For case (1) the effect of increasing the marker spacing from 0.05 M (Figure 1a) to 0.5 M (Figure 1d) was roughly equivalent to doubling the population size and the corresponding comparison in case (2) revealed little difference. Also, the power of the test for markers spaced at 0.05 M in case (1) (Figure 1a) was almost the same as the power of the test for markers spaced at 0.5 M in case (2) (Figure 1h), so for markers spaced at 0.5 M the effect of moving the locus of selection from the marker locus at 0.5 M (Figure 1d) to midway between marker loci (0.25 M) (Figure 1h) was to effectively halve the population size. A similar effect was observed to a lesser degree for a marker spacing of 0.2 M, and for closer marker spacings, there was almost no difference between the results for both cases.

The *single locus* test

For this test a different picture emerges. The power of the test depends very little on the marker spacing for case (1) above (Figure 1a–d), but it increases with decreasing marker spacing in case (2) (Figure 1e–h) because the locus of selection then becomes closer to a marker locus. But as before, there is a noticeable reduction in power for case (2) compared with case (1) particularly for marker spacings 20 cM and greater.

Comparison of power between the two tests

For both cases when the locus of selection was at a marker locus near the centre of a chromosome or midway between marker loci towards the end of a chromosome, and for populations <50, the power of both tests was low. For both cases, for populations of 50 and above, and for markers spaced at less than 50 cM the *single locus* test was more powerful than the *overall* test, with this difference becoming greater as the marker spacing decreased. This effect was more clearly seen as the population size increased. However, at 50 cM spacing for both cases the power of the two tests were extremely close (with the *single locus* test having a slightly greater power) when the convex envelope of the set of 'risk points' (Dudewicz and Mishra, 1988) that make up stepped graphs for the *single locus* test are compared with the corresponding graphs for the *overall* test.

The two tests were compared for power for chromosomes of lengths 0.2, 0.6, 1.0, 1.4 and 1.8 M, with markers spaced at 0.1 M, using populations of 100 and selection with a survival probability of 0.5 for the '+' allele at one end of the chromosome (data not shown). In all cases the *single locus* test had greater power, but the tests had very similar power for 0.2 M, and as the length of the chromosome increased, the difference in power between the tests increased.

The two tests were also compared for power for the situation in which there were two loci of selection, each with a survival probability of $1/\sqrt{2}$ for the '+' allele at the ends of the chromosome (data not shown). The chromosome had lengths of 0.6, 1.0, 1.4 and 1.8 M, and populations of 100 were used. The *single locus* test had greater power in all these cases, but in this case the tests converged in power as the length of the chromosome increased.

Corrections to the level of the tests

For the *overall* test extensive corrections to the *P*-values reported by the test are required for small population sizes and closely spaced markers (Figure 2). For example, from Figure 2 you can read off the empirical probability that under the null hypothesis the *P*-value obtained from the test will be less than, for example, 0.05 corresponding to a rejection of the null hypothesis.

However, these are the conditions where *overall* test is least powerful. For chromosomes 1 M in length, and for marker spacings in the range 0.05–0.5 M, and if x, which denotes the population size multiplied by the marker spacing in Morgans, has the value 10, then a test performed at a nominal α level of 0.05 was in fact a test performed at $\alpha \approx 0.04$. If x was smaller than 10, more of a correction was needed, but if x was larger than 10 the correction required was very small indicating that the approximations made in the *overall* test are accurate under these conditions.

For the *single locus* test, the empirical correction for the *P*-value is greatest when the marker spacing is least, and the length of the chromosome is greatest, but it appears to be almost independent of the population size. Unfortunately the correction factor does depend on the marker spacing even if the total number of markers is kept fixed and the chromosome length varies. Also there is only an approximate linear correction that can be applied to the *P*-value given by the test, because there is considerable curvature in the cumulative probability distribution of *P*-values for the single locus test after 'smoothing out' the steps. This is shown, for example, for a population of 100 (Figure 3). Despite this, correction factors equal to the corrected *P*-value (ie cumulative probability) divided by the nominal *P*-value were obtained as a function of the nominal *P*-value for different marker spacings and genetic lengths of the chromosome (Figure 4) to allow approximate corrections to be carried out.

Discussion

Two statistical tests for segregation distortion used to detect selection in plant breeding were described and studied by simulation. The first is the commonly used *single locus* test that simply assesses each locus separately for segregation distortion from the expected 1:1 segregation ratio and reports the most significant *P*-value, and the second, the *overall* test, is a test that was derived mathematically in an attempt to take information from all the loci of a chromosome into account simultaneously and to correct for the fact that the loci are genetically linked that is the tests based on individual loci are not independent.

In all the cases studied, after the correction for the empirical level of the test had been performed, the *single locus* test was found to be more powerful than the *overall* test that is it had a greater probability of detecting selection. For both tests there is an unavoidable loss of power when the locus of selection is between markers. Another advantage of the *single locus* test over the *overall* test, apart from its simplicity and that the marker data do not have to be put into genetic map order, is that it is appropriate for detecting systematic scoring errors (SSE) that affect individual loci and do not affect adjacent loci as a result of linkage. If the most distorted locus is removed from the data which is then re-analysed, then loss of segregation distortion would suggest that an SSE was responsible, otherwise selection is indicated. However the *overall* test might be less likely to be able to make this distinction because a single highly distorted locus could have a similar effect to several less strongly distorted loci.

However, as expected, the corrections required for the *single locus* test (the correction factors given here are valid only for single chromosomes) were more extensive than those required for the *overall* test, especially for the larger population sizes where the *overall* test is quite accurate. This is probably because the derivation of the *overall* test took into account sampling errors due to a finite population size and the correlation between the segregation ratios of linked markers, although the population was considered to be large enough for the validity of the multivariate Normal approximation. The results suggest that if the *overall* test is to be used, the test should be used with markers as nearly evenly spaced as possible and spaced such that the marker spacing in Morgans multiplied by the population size is larger than about 10. Failing to do this will probably result in considerable errors and loss of power, which can only be accurately assessed by doing simulations.

Because the derivation of the *overall* test is essentially independent of the mapping function used, if this test is to be used with scoring data that do not appear to follow the random model for the frequency distribution of crossovers among chromatids (as is usually the case) it is probably better to replace the Haldane mapping function by the Kosambi, 1944 mapping function that takes into account some of the effects of the non-random distribution of chiasmata (interference) on the calculation of genetic map distances. This change will have very little effect if the markers are closely spaced.

A possible advantage of the *overall* test is that its derivation can be easily extended to a test for segregation distortion for the genome as a whole. This can be shown by extending the analysis here to apply to a whole dataset in which the recombination frequency between different linkage groups is expected to be 0.5 and hence the covariance is expected to be zero for pairs of loci in different linkage groups. This analysis implies that the matrix Σ and hence its inverse Σ^{-1} are in block diagonal form if loci from each linkage group are together. Hence the combined χ^2 statistic is the sum of the χ^2 statistics calculated as described here for each linkage group separately and total number of degrees of freedom is equal to the total number of distinct loci. The procedure should be to calculate the combined χ^2 statistic, to test the hypothesis of no segregation distortion unless there is another reason to look at a particular linkage group or locus. Only if it is significant should the χ^2 tests for individual linkage groups or loci be carried out.

References

Botstein D, White RL, Skolnick M, Davis RW (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**: 314–331.

Brunk HD (1975). *An Introduction to Mathematical Statistics*. Xerox College Publishing: Lexington, Massachusetts, Toronto.

Cloutier S, Cappadocia M, Landry BS (1995). Study of microspore-culture responsiveness in oilseed rape (*Brassica napus* L.) by comparative mapping of a F2 population and two microspore-derived populations. *Theor Appl Genet* **91**: 841–847.

Dudewicz EJ, Mishra SN (1988). *Modern Mathematical Statistics*. John Wiley and Sons: New York.

Foisset N, Delourme R, Lucas MO, Renard M (1997). *In vitro* androgenesis and segregation distortion in *Brassica napus* L.: Spontaneous *versus* colchicine-doubled lines. *Plant Cell Reports* **16**: 464–468.

Frisch M, Quint M, Lübberstedt T, Melchinger AE (2004). Duplicate marker loci can result in incorrect locus orders on linkage maps. *Theor Appl Genet* **109**: 305–316.

Haitham S, Kayyal H, Ramsey L, Ceccarelli S, Baum M (2002). Segregation distortion in doubled haploid lines of barley (*Hordeum vulgare* L.) detected by simple sequence repeat (SSR) markers. *Euphytica* **125**: 265–272.

Haldane JBS (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *J Genet* **8**: 299–309.

Knox MR, Ellis THN (2002). Excess heterozygosity contributes to genetic map expansion in pea recombinant inbred populations. *Genetics* **162**: 861–873.

Kosambi DD (1944). The estimation of map distances from recombination values. *Ann Eugen* **12**: 172–175.

Lu H, Romero-Severson J, Bernardo R (2002). Chromosomal regions associated with segregation distortion in maize. *Theor Appl Genet* **105**: 622–628.

Luo L, Zhang YM, Xu S (2005). A quantitative genetics model for viability selection. *Heredity* **94**: 347–355.

McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, Coffman WR *et al* (1988). Molecular mapping of rice chromosomes. *Theor Appl Genet* **76**: 815–829.

Ruiz C, Asins MJ (2003). Comparison between *Poncirus* and *Citrus* genetic linkage maps. *Theor Appl Genet* **106**: 826–836.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M *et al* (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* **23**: 4407–4414.

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res* **18**: 6531–6535.