

Power for detecting genetic divergence: differences between statistical methods and marker loci

NILS RYMAN,* STEFAN PALM,* CARL ANDRÉ,† GARY R. CARVALHO,‡ THOMAS G. DAHLGREN,† PER ERIK JORDE,§ LINDA LAIKRE,* LENA C. LARSSON,* ANNA PALMÉ* and DANIEL E. RUZZANTE¶

*Division of Population Genetics, Department of Zoology, Stockholm University, S-106 91 Stockholm, Sweden, †Department of Marine Ecology, Göteborg University, Tjärnö Marine Biological Laboratory, S-452 96 Strömstad, Sweden, ‡Göteborg University, Department of Zoology, Box 463, S-405 30 Göteborg, Sweden, §Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, PO Box 1066 Blindern, N-0316 Oslo, Norway, and Institute of Marine Research, Flødevigen, N-4817 His, Norway, ¶Department of Biology, Dalhousie University, Halifax, NS, Canada B3H 4J1

Abstract

Information on statistical power is critical when planning investigations and evaluating empirical data, but actual power estimates are rarely presented in population genetic studies. We used computer simulations to assess and evaluate power when testing for genetic differentiation at multiple loci through combining test statistics or *P* values obtained by four different statistical approaches, viz. Pearson's chi-square, the log-likelihood ratio *G*-test, Fisher's exact test, and an F_{ST} -based permutation test. Factors considered in the comparisons include the number of samples, their size, and the number and type of genetic marker loci. It is shown that power for detecting divergence may be substantial for frequently used sample sizes and sets of markers, also at quite low levels of differentiation. The choice of statistical method may be critical, though. For multi-allelic loci such as microsatellites, combining exact *P* values using Fisher's method is robust and generally provides a high resolving power. In contrast, for few-allele loci (e.g. allozymes and single nucleotide polymorphisms) and when making pairwise sample comparisons, this approach may yield a remarkably low power. In such situations chi-square typically represents a better alternative. The *G*-test without Williams's correction frequently tends to provide an unduly high proportion of false significances, and results from this test should be interpreted with great care. Our results are not confined to population genetic analyses but applicable to contingency testing in general.

Keywords: chi-square, Fisher's exact test, Fisher's method, genetic differentiation, *G*-test, statistical power

Received 30 October 2005; revision accepted 14 November 2005

Introduction

A steadily growing number of studies in conservation and evolutionary biology are based on statistical analysis of genotypic data. A basic question in many such studies is one of genetic homogeneity, i.e. whether or not two or more samples are likely to represent the same population or gene pool. Depending on the outcome of the statistical test — significance or not — genetic divergence or homogeneity

is concluded. This conclusion, in turn, may be used for inference on evolutionary processes or for decisions on management and conservation (e.g. Waples 1998; Hedrick 1999, 2001).

The typical way of addressing the question of genetic homogeneity is to assay a series of gene loci, and to test statistically the null hypothesis of identical allele frequencies. Here, all polymorphic loci scored are considered potentially informative regarding overall genetic homogeneity. A common approach is to first test for allele frequency homogeneity at each locus separately, and then to combine the information from these single locus tests into an overall

Correspondence: Nils Ryman, Fax: + 46-8 154041;

E-mail: nils.ryman@popgen.su.se

P value relating to the joint null hypothesis of no allele frequency difference among populations at any locus.

A variety of procedures for testing the hypothesis of identical allele frequencies is available in the literature and through different software, but it is frequently difficult to see their pros and cons. In particular, there has been little discussion on the relative merits of alternative approaches for how to combine the test results at separate loci into an overall P value (but see Ryman & Jorde 2001 and Petit *et al.* 2001).

A central issue in all hypothesis testing refers to the question of statistical power, i.e. the probability of rejecting the null hypothesis (H_0) when it is false. Knowledge of statistical power is important in two basic situations (e.g. Peterman 1990; Erdfelder *et al.* 1996). First, when planning a survey or an experiment it is critical to design it in a way that makes it likely to detect differences of a magnitude considered relevant for the specific question at hand. Second, in studies where no significance has been obtained, assessment of power makes it possible to evaluate the magnitude of true differences that may have gone unnoticed, given the particulars of the investigation conducted.

The concept of power and its importance is frequently mentioned in studies dealing with genetic divergence. For example, it is often claimed that statistical power (or 'resolving power') is expected to be 'high', or higher than in other studies, because of the use of large sample sizes or some particular type of genetic marker. In cases of no significance, it is frequently implied that a relevant amount of divergence should have been detected with a high probability. Alternatively, a lack of significance is in some cases discussed in terms of a potentially low power. Quantitative assessments of these claims are typically lacking, though.

An obvious reason for this lack of power estimates is the large number of factors affecting it, including the number of samples, their size, the magnitude of the true divergence, the number and type of loci assayed, the number of alleles (or haplotypes), and their frequency distributions. In addition, different statistical techniques are not expected to yield identical power (e.g. Goudet *et al.* 1996; Ryman & Jorde 2001).

We use computer simulations to examine the relative importance of various factors determining power when testing for genetic homogeneity at multiple loci, focusing on commonly applied statistical testing methods, sample sizes, and numbers and types of genetic markers. We have previously shown that power may vary considerably between frequently used statistical methods. In some situations when combining exact P values by Fisher's method, power may actually decrease with a growing number of loci, rather than increase as would typically be expected (Ryman & Jorde 2001). Here, we expand on those results for biologically more realistic scenarios, examining power when sampling from divergent populations where the amount of hetero-

geneity among loci varies stochastically. We also present power estimates for some empirical sets of commonly used genetic markers (microsatellites and allozymes) with different numbers of alleles and frequency distributions.

Basic concepts

Important characteristics of a statistical test may be described in terms of the associated α and β (or type I and II) errors. Here, α signifies the probability of erroneously rejecting the null hypothesis (H_0) when it is true, and β is the probability of accepting H_0 when it is false. The quantity $1-\beta$ is referred to as the power of the test, indicating the probability of rejecting a false H_0 , equivalent to accepting the alternative hypothesis (H_1). In the context of testing for genetic homogeneity when sampling from two or more populations, H_0 typically specifies that all populations have identical allele frequencies, whereas H_1 implies that they have not. Thus, in case of a true divergence, the power ($1-\beta$) specifies the probability of detecting genetic structuring through a significant outcome of the test.

For any chosen (pre-assigned) level of the α error, a 'good' statistical test should be characterized by as small a β error as possible. The α and β errors are not independent, though, and reducing α generally results in a larger β and thereby a reduced power ($1-\beta$), and vice versa. The α error is typically set to 0.05 in biological testing, and for any particular experimental or sampling situation this generally accepted error level results in a specific power, which is frequently unknown but should be assessed.

Checking that the realized α level is reasonably close to the intended one represents a natural starting point when evaluating the characteristics of a test. An inflated realized α implies that the test is 'unreliable' producing an unduly high proportion of false significances, whereas a too small realized α implies that the test may be conservative and inefficient in detecting true differentiation. Without knowledge of realized α levels, power comparisons between tests may not be very informative because higher power of one test (when H_0 is false) may merely reflect an excessive proportion of false significances (when H_0 is true).

Simulations

We mimicked sampling from populations at various levels of expected divergence through random number computer simulations under a classical Wright-Fisher model without migration or mutation (program written in TURBOPASCAL 7.0). An infinitely large base population segregating for a specified number of independent, selectively neutral loci with defined allele frequencies was divided into s subpopulations of equal effective size (N_e) through random sampling of $2N_e$ genes. Each of the subpopulations of size N_e was allowed to drift for t generations, and the expected degree

of divergence in generation t is then $F_{ST} = 1 - (1 - 1/2N_e)^t$ (e.g. Nei 1987 p. 359).

In generation t a random sample of $2n$ genes (corresponding to n diploids) was drawn (with replacement) at each locus from each subpopulation, and the H_0 of identical allele frequencies in all the s populations was tested using different statistical approaches. The entire process of splitting the base population, random drift, sampling, and testing was repeated 1000 times (runs) for each set of starting allele frequencies, s , F_{ST} (combination of N_e and t), and n . The proportion of significant outcomes ($P < 0.05$) was taken as the estimate of power. Estimates of α error were obtained in a similar way, using samples drawn directly from the base population and omitting the drift steps (i.e. $F_{ST} = 0$).

At each locus the H_0 of identical allele frequencies was tested using four different statistical methods, i.e. Pearson's traditional contingency chi-square, the standard log-likelihood ratio G -test (e.g. Sokal & Rohlf 1995; chapter 17), an extension of Fisher's exact test for general $r \times c$ tables (as implemented in the STRUC routine in GENEPOP 3.3; Raymond & Rousset 1995a, b), and a randomized permutation test evaluating the probability of obtaining a sample F_{ST} of the magnitude observed or larger when permuting genes among samples (with F_{ST} calculated according to Weir & Cockerham 1984, equation 6, assuming Hardy-Weinberg proportions within populations). For Fisher's exact test we used the default values of GENEPOP 3.3 for the number of dememorizations (1000) and Markov chain length (100 000), and 10 000 randomizations was employed in the permutation tests.

With respect to the chi-square and G statistics we did not include their respective corrections, i.e. Yates's for 2×2 chi-square tables and Williams's for G (e.g. Sokal & Rohlf 1995). First, these corrections are frequently not applied in population genetic studies (most likely because software like POPGENE (Yeh *et al.* 1997) only provide uncorrected chi-square and G values). Further, our previous analysis (Ryman & Jorde 2001) indicated that Williams's and Yates's corrections yield results fairly similar to those of chi-square and Fisher's exact test, respectively. The tendency of Williams's G (G_W) to behave largely as chi-square was also observed when computing G_W in the present study, although this was not done systematically (not shown).

When combining the information from multiple single-locus tests we used the same techniques as Ryman & Jorde (2001). That is, for chi-square and G we summed the test statistics (χ^2 or G) from the separate tests (loci) along with their associated degrees of freedom (d.f.), assuming that under the joint null hypothesis of no difference at any locus this sum approximately follows a χ^2 distribution with d.f. = $\sum d.f._i$, where i is the number of single-locus contingency tests. Similarly, for P values produced by Fisher's exact test, the quantity $-2 \sum \ln P_i$ was assumed to be distributed as

χ^2 with d.f. = $2i$ when the joint H_0 is true ('Fisher's method' as implemented in, e.g. GENEPOP; see Ryman & Jorde 2001 for details on null hypotheses in this type of testing).

We focus primarily on 'traditional' contingency tests, only including one out of several possible alternatives for randomization or permutation by using F_{ST} to rank the permuted contingency tables (cf. Goudet *et al.* 1996). When combining the P values obtained by multiple permutation tests at each of several loci the same strategy as for Fisher's exact test was used, evaluating the joint H_0 using Fisher's method. Note that 'Fisher's method' and 'Fisher's exact test' are two different things. Similarly, the G -test examined here should not be confused with the test implemented in FSTAT (Goudet 1995, 2001) that evaluates permuted contingency tables on the basis of the G statistic.

Accuracy of the simulated drift process was checked by comparing simulated F_{ST} distributions with expected values using the recurrence equations of Nei (1975, pp. 121–24), and through comparing means and variances of the observed allele frequency distributions of the base populations with those expected theoretically (Choy & Weir 1978). The sampling from the base populations was tested in a similar manner, and the results of our statistical test routines (chi-square, G , and Fisher's exact test) were compared with those of several software including GENEPOP, POPGENE, STATXACT (Mehta & Patel 1997), and BIOM (Rohlf 1987).

A few comments may be warranted on the simulation scheme. First, sampling is performed from populations characterized by an *expected* amount of differentiation as described by a particular F_{ST} , thus representing the F_{ST} that would result had an infinite number of selectively neutral loci been scored. This strategy was chosen because we believe that in practical situations the most adequate question on power typically refers to the probability of detecting divergence among populations exhibiting some particular level of 'true' differentiation across a large number of loci affected by the same evolutionary forces.

Further, a specified expected F_{ST} may be obtained through many combinations of N_e and t . We have used fairly large N_e s, typically in the range of 2000–4000, to maintain overall allele frequencies reasonably similar to those of the base population, and to prevent excessive loss of low-frequency alleles. To permit evaluation of the effect of losses and fixations, our program presents separately the results from simulations (runs) maintaining all the alleles in the base population and those where loss has occurred. This distinction turned out to be of little importance within the parameter space employed, however, and we only present the pooled results. Thus, when providing results for loci with different number of alleles (e.g. Fig. 1c, d) we consistently refer to the number in the base population, regardless of the number that was actually observed in the contingency tables used for statistical testing (i.e. those tests occasionally differed somewhat with respect to their degrees of freedom).

There is clearly an infinite number of scenarios that can be addressed with the present approach. We have focused on the effect on power of combining multiple loci with different levels of variation meant to represent commonly employed markers such as allozymes, microsatellites, and SNPs (single nucleotide polymorphisms). Most results refer to designs using frequently reported sample sizes that are of the same magnitude, but the effect of unbalanced sampling was also evaluated. We have not dealt with the Bonferroni technique for evaluating the joint H_0 for multiple loci, because our previous results indicate that this approach may not be appropriate in the present context (Ryman & Jorde 2001; but see Neuhäuser 2003, 2004 for discussions on alternatives).

Results

We studied the effect on α and power of the number of loci and alleles, sample size and number of samples, balanced vs. unbalanced sampling, and whether or not the number of 'independent' alleles is a good indicator of power. Most of the simulations refer to 'hypothetical' loci, but authentic data sets are also included.

In a first step power differences between hypothetical loci with few or several alleles, and uniform or skewed allele frequency distributions, were examined. We focused on loci with 2 or 10 (and in some cases up to 50) alleles, where uniform indicates that all alleles occur at the same frequency in the base population (0.5 and 0.1 for 2 and 10 alleles per locus, respectively), whereas for the skewed distributions we let all but one allele be of a 'low' frequency (0.01). Thus, for a di-allelic locus the skewed distribution represents the allele frequencies of 0.99 and 0.01, and for a 10-allelic one the frequencies are 0.91, 0.01, 0.01 ... 0.01. The power for various numbers of these types of loci was evaluated for different combinations of F_{ST} , s (number of sampled populations), and n (number of 'diploids' sampled from each population).

A single locus

Figure 1 (a and b) depict the proportion of significances as a function of F_{ST} when sampling 50 individuals (100 genes) from each of five populations with an expected F_{ST} ranging from 0 (no differentiation) to 0.05. The proportion of significances (out of 1000 runs) signifies the α error when $F_{ST} = 0$, and otherwise power.

Clearly, power may be reasonably high even for a single locus with the actual value depending on the true F_{ST} , the number of alleles, their distribution, and the testing method. Multiple alleles and uniform frequency distributions are 'better' than few alleles occurring at skewed frequencies. The different statistical testing methods provide very similar results at uniform allele frequencies, but there are some

obvious power differences with many low-frequency alleles. In particular, the permutation test performs worse than the others, unless F_{ST} is so large that power for all tests approaches unity (Fig. 1b). The α error stays fairly close to the intended level of 5% for all tests, with the exception of the G -test applied to 10 alleles with a skewed distribution, where the realized α is elevated (c. 14%).

The dependence of α and power on the number of alleles (2–50) is further detailed in Fig. 1 (c and d) for $s = 2$ and $n = 50$. Here, the unduly large α error of the G -test becomes more pronounced as the number of alleles increases (exceeding 35% for the skewed allele frequency distribution), whereas this error stays close to or is below the intended 5% for the other three tests (Fig. 1c).

With respect to power (Fig. 1d), we first note that the G -test consistently yields the highest proportion of significances, but the perceived 'superiority' of this test must not be taken at face value, because it merely reflects its inappropriately large α . With uniform allele frequencies, the Fisher's exact, chi-square, and permutation tests all provide similarly increasing power that approaches unity at 50 alleles. At skewed frequencies, however, there are marked power differences between these testing methods, with Fisher's exact test generally outperforming the two others (chi-square and permutations). The exception refers to the case of two alleles, where chi-square exhibits a somewhat higher power (cf. below).

Multiple loci

Figure 2 exemplifies results obtained when combining the separate test statistics (or P values) for an increasing number (1–40) of the four types of loci for $F_{ST} = 0$ and 0.01, $s = 2$, and $n = 50$. With respect to α (left column of plates), the most obvious observation is that the G -test frequently yields an unduly high proportion of significances in the case of no differentiation, also when combining the information from multiple loci. This is particularly so for loci where low frequency alleles are common, including situations with several uniformly distributed alleles. For example, for ten 10-allele loci with skewed allele frequencies (Fig. 2g) the G -test results in a probability of erroneously rejecting the true H_0 of about 65%. Clearly, such a high realized α is unacceptable and should disqualify this test from use in many situations.

In contrast to the unduly high realized α for the G -test, the other three tests occasionally result in an α that is markedly smaller than the intended 5% (Fig. 2, left column). The most striking observations for di-allelic loci are that combining P values from Fisher's exact or permutation tests by means of Fisher's method results in an α that (i) may decline with an increasing number of loci (Fig. 2a), and (ii) may even approach zero regardless of the number of loci (Fig. 2c). For 10-allele loci, however, the realized α for these two tests stays fairly close to the intended 5% (Fig. 2e, g).

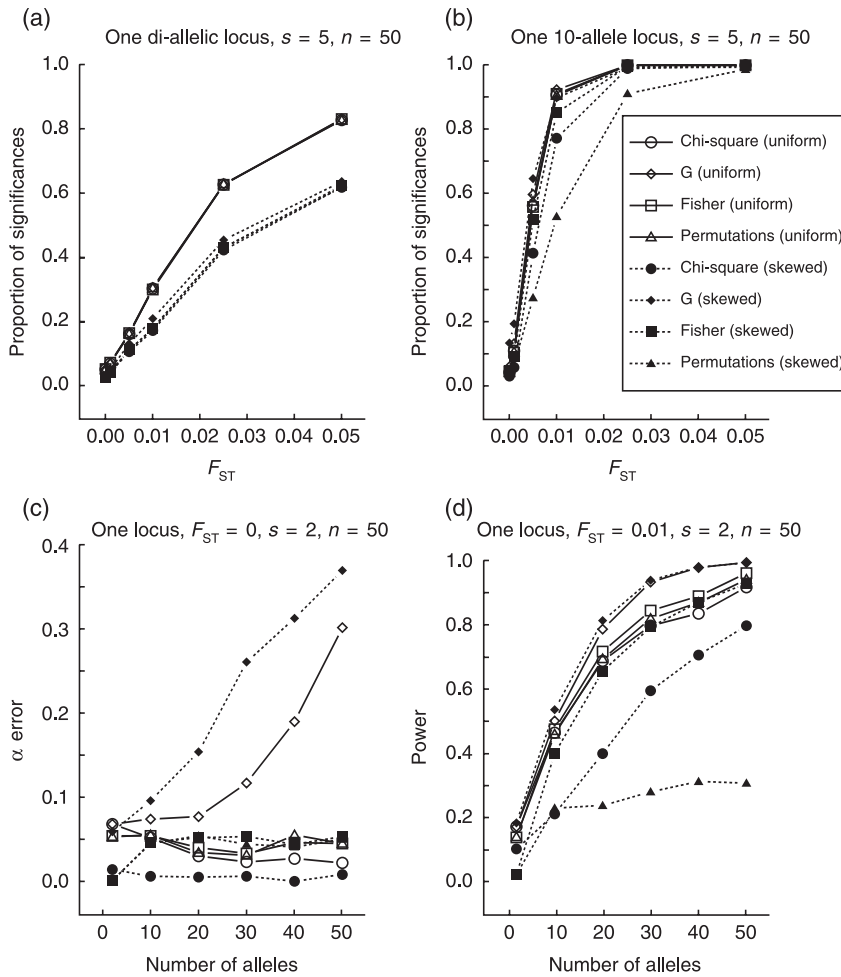


Fig. 1 Simulated α error and power for a single locus. Plates (a) and (b): proportion of significances ($P < 0.05$) obtained through different statistical methods as a function of the true level of differentiation (F_{ST}) when testing the hypothesis of identical allele frequencies at a di-allelic (a) or a 10-allelic (b) locus with uniform (solid lines) or skewed (dotted lines) allele frequency distributions. The number of populations sampled (s) is 5, and each sample size (n) is 50 diploids. The proportion of significances corresponds to α when $F_{ST} = 0$ and to power when $F_{ST} > 0$. Plates (c) and (d): α error (c) and power (d; $F_{ST} = 0.01, s = 2$, and $n = 50$) as a function of the number of alleles (2–50). Note the different scales of the y -axes. See text for details.

When combining multiple chi-squares the α error is low in cases with low-frequency alleles, but it is never zero, and there is no obvious relation to the number of loci. Thus, with respect to α , the Pearson chi-square generally appears to perform better than the other tests considered here (Fig. 2, left column).

Power curves for the same scenarios as presented for the α errors are shown in the right column of Fig. 2. Clearly, the proportion of significances depends on the number of loci, the number of alleles per locus and their frequency distributions, and the statistical method. As noted previously for a single locus, the G-test consistently yields a high power also for multiple loci, but again, the tendency of the G-test to produce excessive α errors indicates that this test is unreliable in many situations (in realistic settings we do not know if H_0 is true or false).

Power is generally higher for 10-allele loci than for di-allelic ones, and as expected, it also increases with the number of loci. A striking and counter-intuitive exception refers to the permutation and the Fisher's exact tests in the di-allelic low-frequency case (Fig. 2d). Here, power initially

is very low and decreases further as the number of loci goes up (in contrast to that of the chi-square that steadily increases). This small and decreasing power largely seems to be a reflection of the realized α being essentially zero in these cases. A situation where $\alpha = 0$ implies that the test method applied cannot produce a significant result regardless of how extreme the observations in a sample may be, and power is then necessarily zero.

The chi-square performs better than the permutation and the Fisher's exact tests for di-allelic loci with more uniform frequency distributions (Fig. 2b). Again, this behaviour seems to be related to the reduced levels of realized α of the two latter tests.

For 10-allele loci power is high and similar for all approaches when allele frequencies are uniformly distributed (Fig. 2f). The segregation of infrequent alleles reduces power somewhat when the number of loci is not very large, but the reduction is in most cases fairly modest (Fig. 2h). When power differs between methods (skewed frequencies, few loci), Fisher's exact test consistently yields the highest power and the permutation test the lowest, whereas

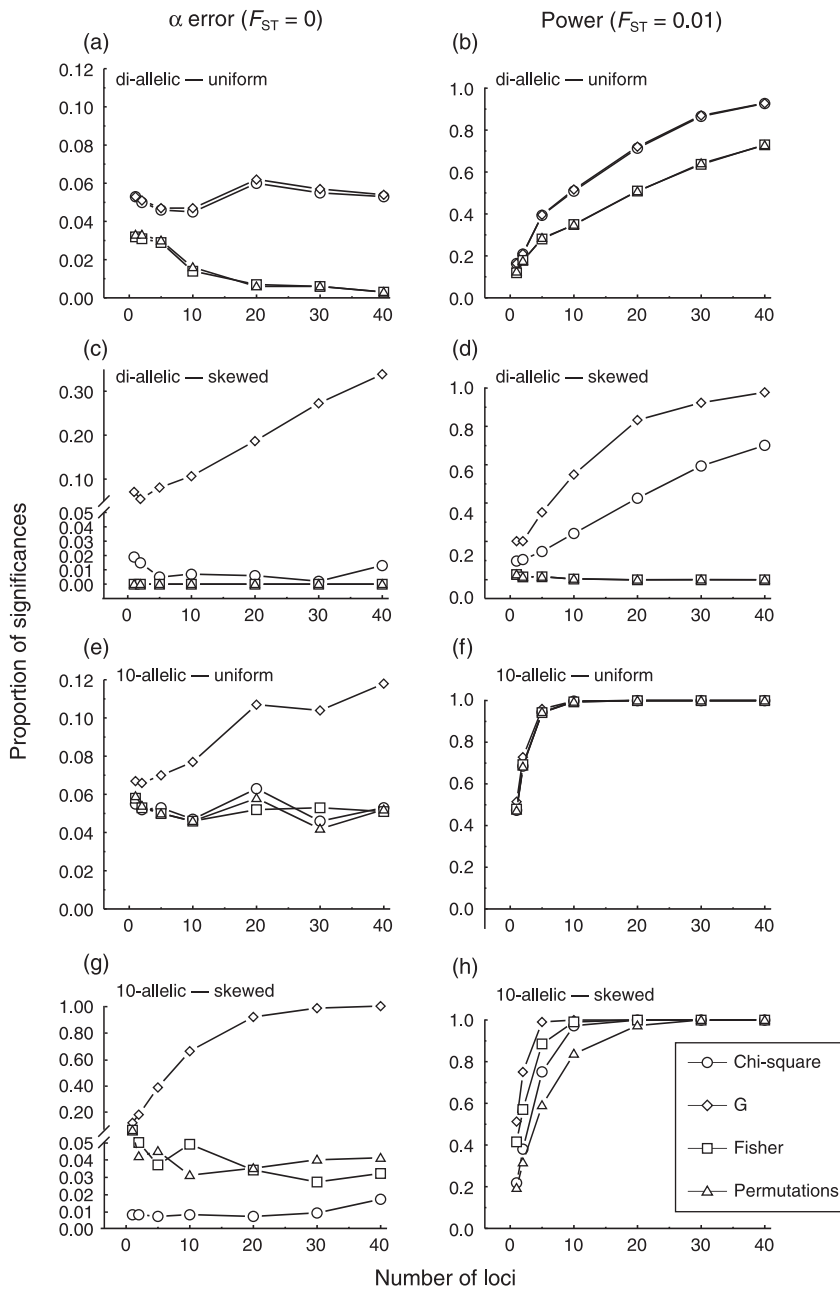


Fig. 2 The α error ($F_{ST} = 0$; left column of plates) and power at $F_{ST} = 0.01$ (right column) for hypothetical di- and 10-allelic loci with uniform or skewed allele frequency distributions when combining the information from 1 to 40 loci of the same type. $s = 2$, $n = 50$. Note the different scales of the y-axes.

chi-square is intermediate. This is in contrast to the di-allelic cases (Fig. 2b, d) where chi-square always performed best.

Examples of the impact on power from increasing the sample size and the number of samples are shown in Fig. 3. Here, we focus on the effects of combining 10 loci of a particular type for various combinations of n and s , with $F_{ST} = 0.01$ or 0.001 (F_{ST} was reduced to 0.001 for the 10-allele loci not to consistently provide a power close to unity). The results from the G-tests are omitted from this figure because of its inflated power.

Regardless of statistical method, increasing total sample size ($s \times n$) through a larger s or n consistently yields a higher power. Expressed differently, for a fixed s , power increases with a larger n , and vice versa. We have also examined the effect of allocating between various combinations of s and n resulting in the same total number of individuals ($s \times n = 100$ or 250 ; Fig. 3). For a fixed total sample size the general effect of scoring few large samples or several smaller ones seems to be minor at di-allelic loci. For 10-allele loci it appears as if it is most efficient to analyse many individuals from each of few samples, although the absolute power

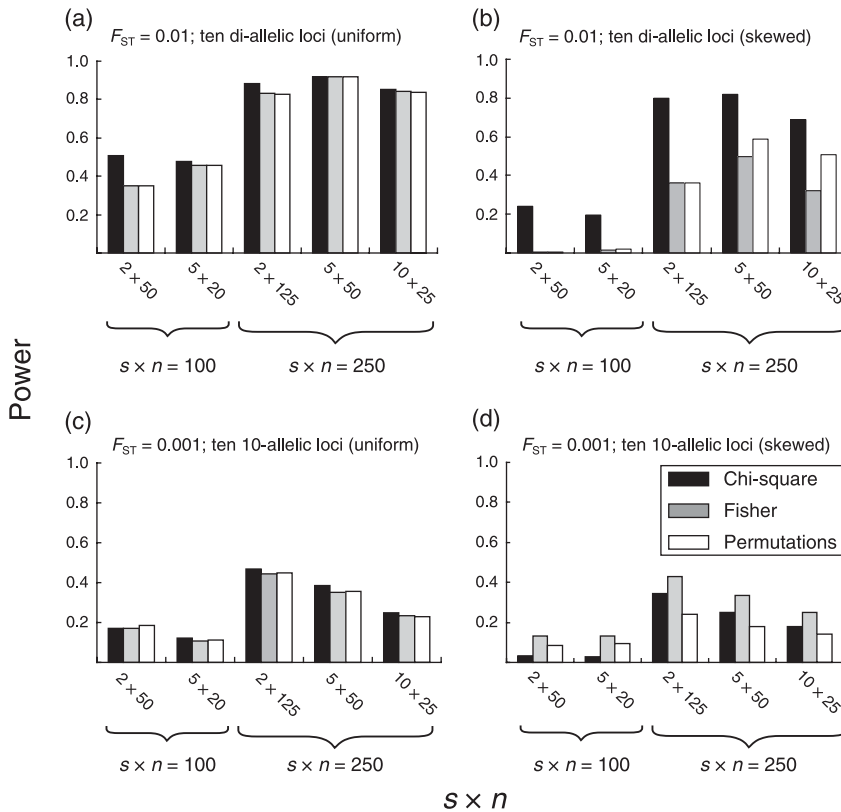


Fig. 3 Power when combining the information from 10 hypothetical loci with uniform or skewed allele frequency distributions for selected combinations of s and n ($s \times n =$ total sample size). Plates (a) and (b): 10 di-allelic loci; $F_{ST} = 0.01$. Plates (c) and (d): 10 10-allelic loci; $F_{ST} = 0.001$.

difference is small. The latter observation is similar to that of Goudet *et al.* (1996) for single loci in power simulations where all individuals in the base population were uniquely heterozygous, seemingly implying a fairly large number of alleles in the populations sampled for testing.

An important observation from Fig. 2 was the conspicuous power difference between test methods for di-allelic loci with skewed allele frequencies for $s = 2$ and $n = 50$ (Fig. 2d). When increasing the total sample size through a larger n or s , however, the power of the Fisher's exact and permutation tests is no longer close to zero (Fig. 3b). Nevertheless, the chi-square test consistently performs best and appears to be the method of choice when dealing with few skew-frequency alleles. In contrast, with few alleles at more uniform frequencies (Fig. 3a) the difference between tests is less pronounced or nonexistent, except for small values of both s and n . With 10 alleles (Fig. 3c, d) differences between methods are also minor, although Fisher's exact test performs slightly better than the others in cases with skewed allele frequencies (Fig. 3d). It should be noted that in no case are these power differences between test methods associated with an unduly large α error (all α values corresponding to the simulations of Fig. 3 are around 5% or less; not shown).

Small or unbalanced sample sizes

It is frequently suggested that chi-square may not be suitable

for testing contingency tables with small expected values, as is commonly encountered for loci with infrequent alleles, because of a presumed risk of an excessive proportion of false significances (e.g. Weir 1996; Agresti 2001). The present results, however, indicate that chi-square in some cases performs as well as, or better than, the other approaches. A similar observation was made in the simulation study by Ryman & Jorde (2001).

To address the issue of a potentially inflated α for chi-square in greater detail we conducted a series of simulations analysing more sparse contingency tables with several cells exhibiting considerably smaller expectancies than previously. Specifically, we focused on unbalanced sampling with $s = 3$, letting one sample be large ($n_1 = 100$) and the other two of a different but mutually identical size ($n_2 = n_3$ ranging in the interval 2–100), and loci with 2, 10, and 50 alleles.

Resulting α errors from simulations with unbalanced sample sizes at one locus are shown in Fig. 4. Several observations can be made from this figure. First, it is clear that chi-square may result in an inflated α when small samples are compared with much larger ones. However, this phenomenon only appears to be an issue when the smaller samples comprise about 20–30 diploids or fewer, and at loci segregating for low-frequency alleles (i.e. loci with skewed frequencies or with so many uniformly distributed alleles that they all occur at low frequencies). Further, the chi-square statistic may not be as sensitive to small

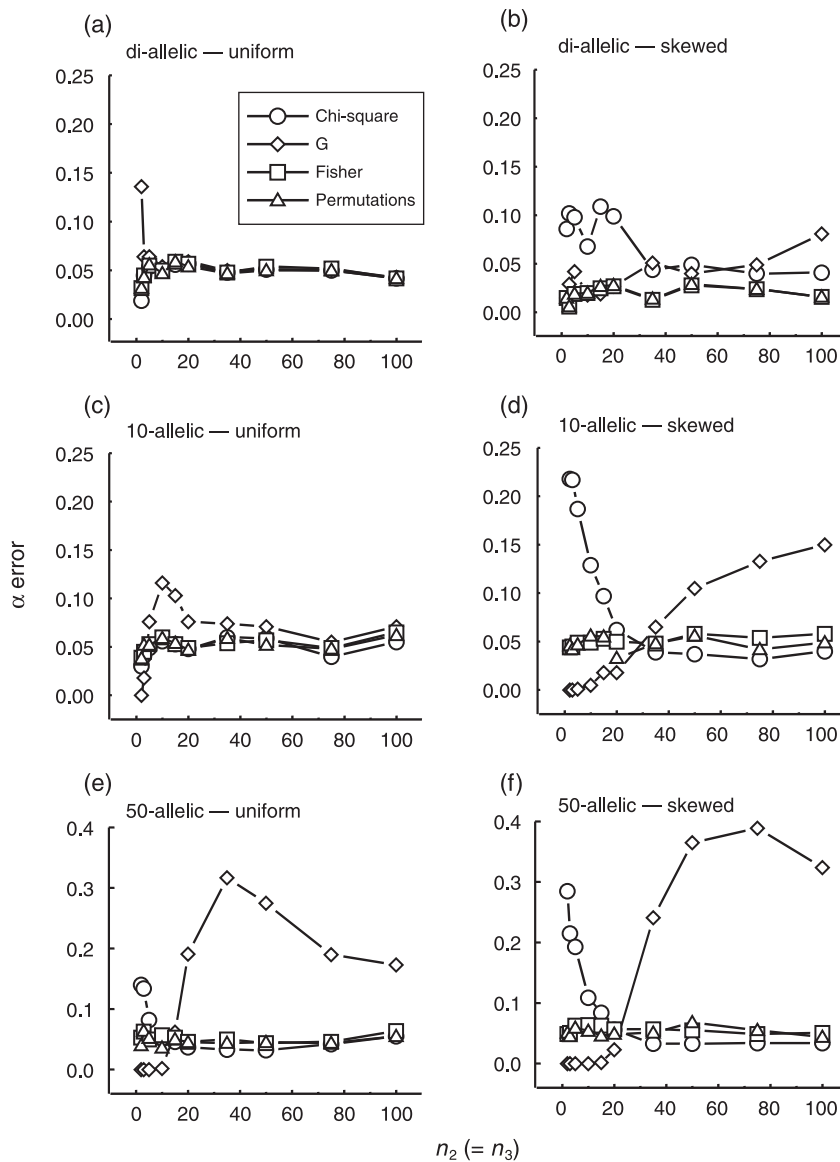


Fig. 4 The α error ($F_{ST} = 0$) at a single locus with different number of alleles (2, 10, and 50) when drawing three samples ($s = 3$) of different size (n_1, n_2 , and n_3). One sample is fixed at 100 diploids ($n_1 = 100$), and the other two are of a mutually equal size ($n_2 = n_3$) that varies in the interval 2–100. The left and right columns of plates refer to uniform and skewed allele frequency distributions, respectively. Note the different scales of the y-axes.

expectancies as is commonly presumed. As an example, we may consider the Fig. 4(f) scenario (50-allelic skewed) for $n_2 = n_3 = 50$, where all the 1000 simulated 3×50 tables contain cells with expectancies < 1 , and where the mean number of cells with an expected value of less than 5 and less than 1 was 144 and 44 (out of 150), respectively. Nevertheless, realized α for chi-square stays very close to the intended 5%.

Figure 4 does not include cases where all samples are small, say $n < 20$, but some simulations were conducted to cover that scenario also. Interestingly, small but balanced samples resulted in an α for chi-square that was very close to or below the intended level of 5% ($s = 2$ or 10 and $n = 10$; not shown). That is, the problem with small expectancies appears primarily to be a matter of concern when dealing

with highly unbalanced sample sizes in combination with low-frequency alleles.

In the case of unbalanced sampling, α for the Fisher's exact and permutation tests always stays close to the intended 5%. In contrast, the G-test behaves quite erratically. In some situations α for G is too large, and in others it is too small, but it is difficult to discern a clear pattern for how G will deviate in either direction. For example, with one 10-allelic locus with skewed allele frequencies (Fig. 4d) α for G increases continuously from zero to c. 15% as n_2 and n_3 goes from 2 to 100. In contrast, for a 10-allelic locus with uniform allele frequencies, α starts out at zero, increases rapidly to about 12%, and then goes down and levels out just above 5% (Fig. 4c). Similar but much more pronounced deviations from the intended α of 5% are seen for 50 alleles

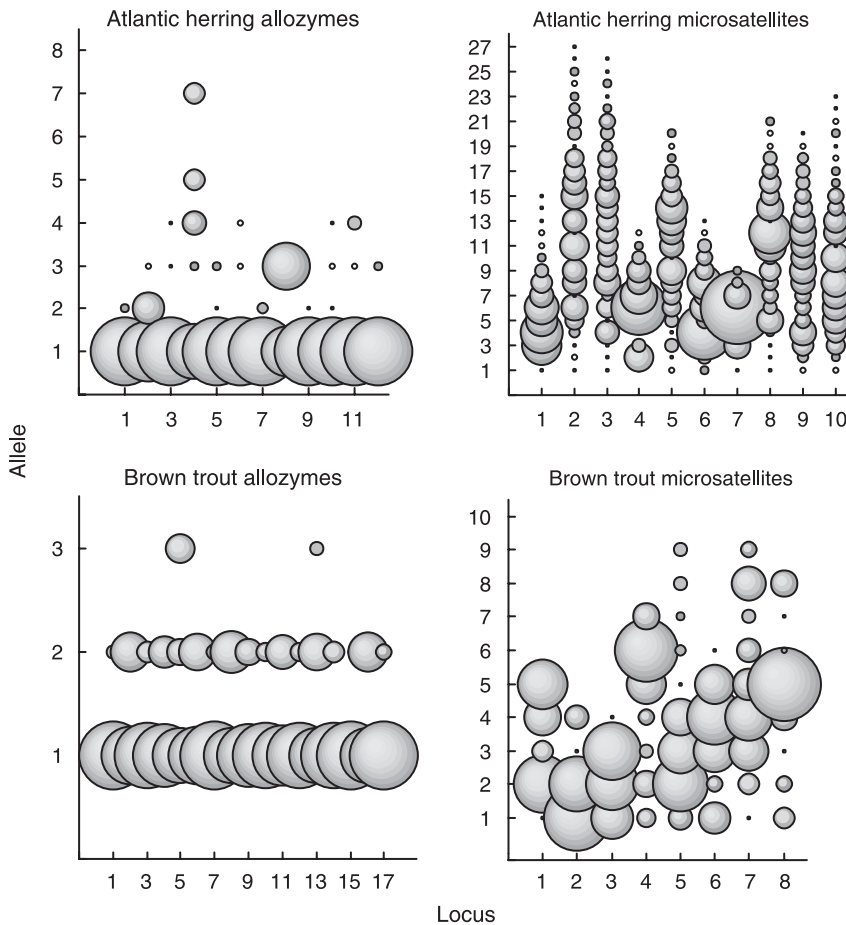


Fig. 5 Bubble diagram representation of number of loci and average allele frequency distributions in authentic allozyme and microsatellite data sets for Atlantic herring (Ryman *et al.* 1984; André, unpublished) and brown trout (Palm *et al.* 2003) used as base population data in power simulations (Fig. 6). Note that all loci are variable, but that some variant alleles occur at fairly low frequencies (resulting in small 'bubbles').

(Fig. 4e, f). Combining information from multiple loci results in α error graphs resembling those in Fig. 4, except that the deviations from intended α are reinforced (not shown).

Authentic data

An authentic data set typically consists of a combination of loci with many or few alleles segregating at various frequencies. As indicated by the above results, the actual power and the performance of various tests under such circumstances may be difficult to predict because of the many factors that must be considered. As examples of power for different sets of loci from authentic studies, we have conducted simulations based on allozyme and microsatellite data from our own research on Atlantic herring (*Clupea harengus*) and brown trout (*Salmo trutta*). Figure 5 depicts, by means of bubble diagrams, the number of loci and their allele frequency distributions for the four combinations of locus type and species used as starting points for the simulations.

There are obvious differences both between and within species and marker types regarding the number of loci, the number of alleles, and their frequency distributions (Fig. 5).

For the herring it appears reasonable that the large number of microsatellite alleles and their fairly uniform distribution would yield a higher power than for the less variable allozymes, although the magnitude of this difference is difficult to anticipate. This difficulty seems even more pronounced for the brown trout, where the number of allozyme loci is more than twice that of the microsatellites (17 vs. 8), and where the difference between marker types with respect to the number of alleles per locus is less pronounced.

Power curves for the four combinations of species and genetic markers are shown in Fig. 6 for $n = 50$, $s = 2$ or 5, and F_{ST} ranging from 0 to 0.01 ($F_{ST} = 0$ representing the α error). Maybe most important, power may be quite high also at fairly low levels of divergence, regardless of statistical approach and the set of loci examined. For the herring microsatellites, for example, the probability of obtaining a significant result when sampling from five populations approaches 100% for an F_{ST} of about 0.003, and the corresponding probability is around 50% for an F_{ST} as small as 0.001 (Fig. 6b). Similarly, power may be high also for the less polymorphic allozymes when a reasonably large number of loci is scored (Fig. 6a, c).

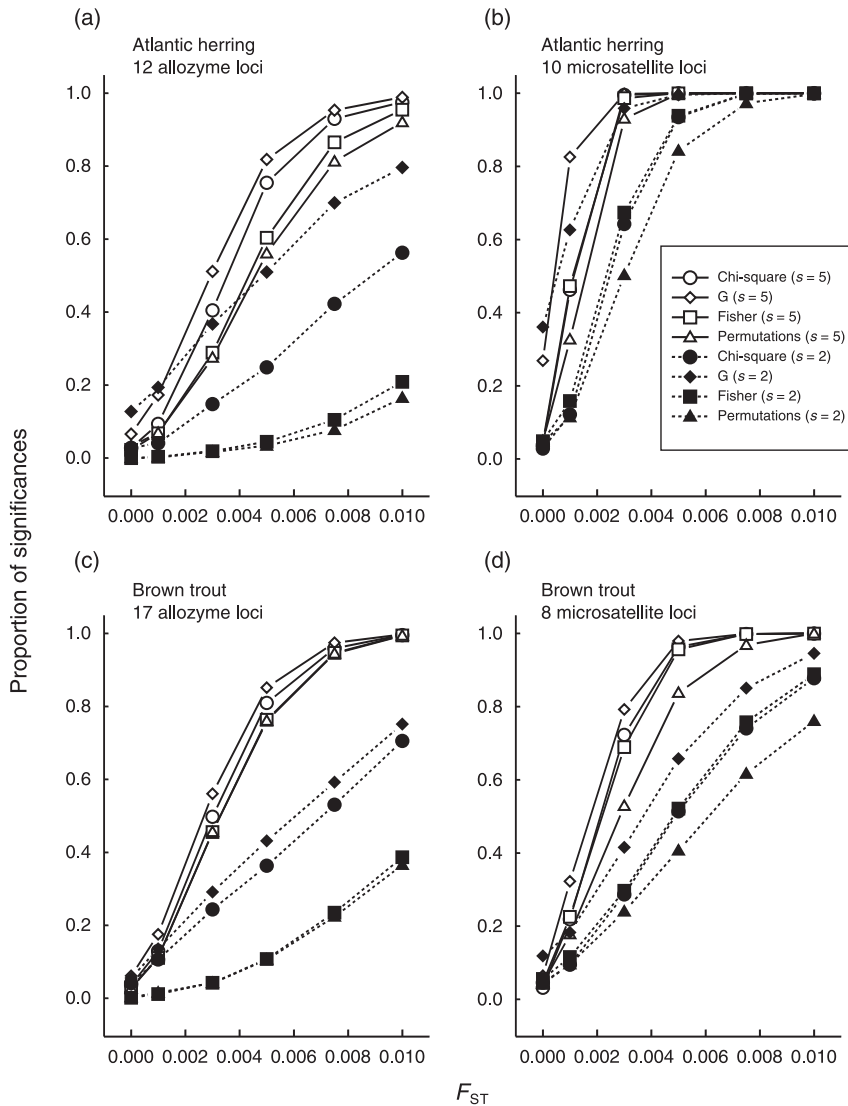


Fig. 6 The α error ($F_{ST} = 0$), and power as a function of true F_{ST} , for the four sets of genetic markers depicted in Fig. 5. $n = 50$ and $s = 2$ or 5 (dotted and solid lines, respectively).

A large number of alleles per locus generally results in an increased power, but for loci with few alleles this may be compensated for by using a larger number of loci. For the brown trout, for example, the 17 allozyme loci may be about as powerful as the eight microsatellites when $s = 5$ (Fig. 6c, d). In contrast, for the herring, where the number of allozyme and microsatellite loci is similar, the power is markedly higher for the microsatellites (Fig. 6a, b).

In Fig. 6 we have included the G-test to examine its behaviour in more realistic data sets than in Figs 1 and 2. Here, this test consistently yields the highest α error, and in several cases the magnitude of α is unduly large, e.g. for the herring microsatellites where this error is about 30–40% for both examples of s (Fig. 6b). As a consequence, the G-test also provides the highest power, but as pointed out previously, this fact does not justify the use of this test.

Ignoring the G-test, there are obvious power differences among the three other tests, and these differences are most

pronounced with $s = 2$ and for the allozymes that are characterized by few alleles per locus (Fig. 6a, c, dotted lines). Here, chi-square outperforms the Fisher's exact and the permutation tests, as was previously seen when dealing with 'hypothetical' loci (cf. Fig. 2b, d). For a larger number of samples or for the microsatellites (with more alleles per locus), i.e. when combining P values from larger contingency tables, the difference between methods is less pronounced. The chi-square and Fisher's exact tests are the best and yield very similar results, but whenever a distinction can be discerned chi-square has the higher power.

Total number of independent alleles

In a real situation the investigator frequently has the option to select which loci to include. In a recent simulation study, Kalinowski (2002) evaluated how the number of alleles per locus affects the precision of various estimates of genetic

distance, including F_{ST} . He showed that similar precision may be obtained either by using a large number of loci with few alleles or fewer loci with many alleles, and that the total number of independent alleles is a good indicator of how precise the estimates will be. The number of independent alleles at a locus is then defined as the total number of alleles minus one, and the total number of independent alleles is obtained by summing over loci.

Kalinowski (2002) did not evaluate statistical power, and his findings are not directly applicable to this issue. This is because power is affected, among other things, by the allele frequency distributions and not only by the number of alleles (e.g. Fig. 2). It is nevertheless interesting to see to what extent the total number of independent alleles may be an indicator of power.

We conducted a restricted set of simulations aimed at addressing this question, focusing on a total of 48 independent alleles distributed over 1–48 loci (corresponding to a number of alleles per locus ranging from 2 to 49; Fig. 7). Under the conditions of the present simulations ($s = 2$, $n = 50$, and $F_{ST} = 0.005$) it appears that power may be about the same when examining 48 independent alleles regardless of their distribution over loci, provided that (i) all the loci considered are characterized by the same ‘type’ of allele frequency distributions (uniform vs. skewed), and (ii) the test method is chi-square (Fig. 7). In contrast, when using Fisher’s exact test (or permutations), the total number of independent alleles is not a good indicator of power. As before, combining P values from those tests through Fisher’s method may provide an unduly low power for multiple loci with few alleles. When dealing with many alleles per locus, however, Fisher’s exact test yields a reasonably good power that may be higher than for chi-square at loci segregating for many low-frequency alleles (Fig. 7b; cf. Fig. 1d). The generality of the extent to which the number of independent alleles is a good predictor of power for chi-square warrants further study though.

Discussion

Our analysis has followed three major lines, i.e. (i) assessing the magnitude of actual power in some empirical and basic hypothetical settings, (ii) comparing the efficiency of a few statistical approaches for testing for heterogeneity, and (iii) evaluating in greater detail the phenomenon of a reduced power when combining multiple exact tests by means of Fisher’s method, particularly in relation to chi-square (cf. Ryman & Jorde 2001).

The major observations may be summarized as follows. First, and as expected, power generally increases with the level of differentiation (F_{ST}), sample size, number of samples, and the number of loci and alleles, and uniform allele frequency distributions are better than skewed ones. Second, regardless of the statistical method employed, the

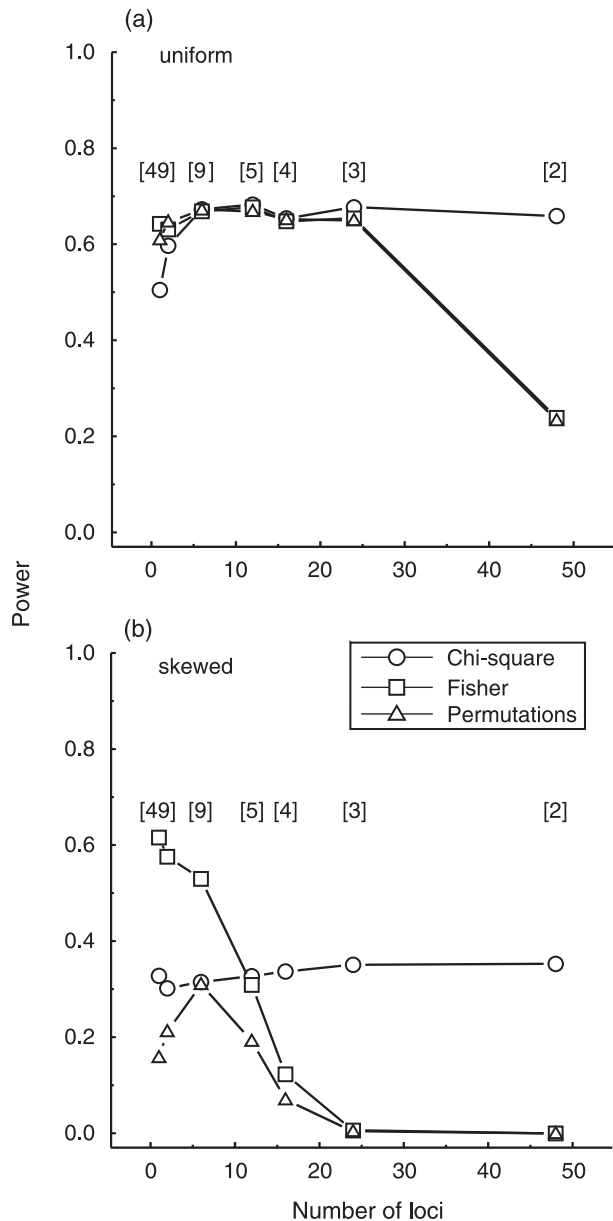


Fig. 7 Power when scoring 48 independent alleles distributed over different numbers of loci for true $F_{ST} = 0.005$, $s = 2$, and $n = 50$. The 48 independent alleles may correspond to a single locus with 49 alleles, 2 loci with 25 alleles each, etc. (as exemplified by numbers in brackets). See text for details. (a) Uniform allele frequency distributions; (b) skewed allele frequency distributions. All corresponding α errors are around 5% or less (not shown).

power for detecting divergence may be substantial for frequently used sample sizes (say 50–100 individuals) and sets of genetic markers (say 5–20 allozyme or microsatellite loci) also at quite low values of F_{ST} (e.g. $F_{ST} < 0.01$). Actually, in many studies using large numbers of highly polymorphic markers, the differences that are likely to be detected are so

small that it may be questionable if they should be considered biologically meaningful for the issue at hand (cf. Waples 1998; Hedrick 1999, 2001).

Third, we show that in some situations the choice of statistical method may be critical for detecting an existing genetic structure. Specifically, with a restricted number of alleles per locus summation of chi-square may outperform Fisher's exact and permutation tests, when the single locus P values from these tests are combined by Fisher's method. This distinction between methods is most pronounced at skewed allele frequencies and few samples, and sometimes the already poor performance of the latter methods may decrease further with an increasing number of loci. Single or combined P values from Fisher's exact test never display elevated α errors, though, and with large numbers of infrequent alleles the power of this approach may be higher than for chi-square (e.g. Figs 1d and 7b).

Finally, an interesting observation refers to the good performance and the apparent robustness of the traditional chi-square test in situations when it is generally considered to be unreliable, i.e. for tables with small expected frequencies. On the other hand of the spectrum, the uncorrected G -test (without Williams's correction) frequently tends to yield an unduly high proportion of false significances at loci segregating for low-frequency alleles.

Our results extend and support our previous observations (Ryman & Jorde 2001), but here we have examined a wider range of parameter settings representing more biologically realistic scenarios. In particular, we have focused on average (expected) F_{ST} over loci rather than on the same fixed allele frequency difference at all loci, and dealt with a larger assortment of hypothetical and authentic loci including those with fairly skewed frequencies. We have also included unbalanced sampling, and show that the comparison of quite small samples with large ones represents a situation where the chi-square approach may produce an excessive rate of false significances at skewed allele frequencies.

Matters not dealt with include tests based on genotypic rather than allelic contingency tables, as suggested for situations with deviations from random mating (e.g. Goudet *et al.* 1996; Weir 1996). Our general results on the behaviour of the different test procedures should be valid for any set of contingency tables, however, keeping in mind that genotypic tables only contain half as many items and that the number of columns (genotypic classes) may be considerably larger than for those underlying our present results. Similarly, it remains to be evaluated to what extent other methods for conducting permutation tests for multiple loci may influence power. We permuted F_{ST} at each locus separately and combined the resulting P values using Fisher's method. We recognize, though, that other permutation approaches are available for multiple loci (e.g. Goudet 1995, 2001; Schneider *et al.* 2000).

We have focused exclusively on power and not addressed the question of precision of estimates quantifying the amount of genetic divergence. In general, one should expect a higher precision in situations of high power and vice versa. For example, the recent observations of Kalinowski (2005), showing that highly variable loci tend to yield more precise estimates of genetic distance and F_{ST} at similar sample sizes, are in line with our observations that a higher power is typically obtained with multi-allelic markers.

The tendency of the uncorrected G to provide an inflated α error has been noted previously (e.g. Sokal & Rohlf 1995). The reason for this phenomenon appears to be that the distribution of the G statistic frequently exhibits a somewhat larger mean and variance than the theoretical χ^2 distribution it is expected to follow asymptotically. Although not necessarily constituting a large problem for a single G value, summation of multiple G s (as when combining information from multiple loci) may yield a distribution located far to the right of the corresponding one expected for the summed χ^2 distribution, thus resulting in a dramatically increased α error (cf. Ryman & Jorde 2001).

The reason why combining P values from Fisher's exact (or the permutation) test sometimes yields a poor power is discussed by Ryman & Jorde (2001). In brief, Fisher's exact test is conditional on both the row and column margins of the contingency table evaluated, restricting the number of possible P values that can occur for a given set of margins. This discreteness may result in a smaller *realized* α than expected from the *intended* one, and this phenomenon is most pronounced for small contingency tables with large differences between the row or column totals (as with pronounced unbalanced sampling or highly skewed allele frequencies). Expressed differently, when H_0 is true the proportion of significances (e.g. $P < 0.05$) occurs in a frequency of less than 5% (cf. Fig. 2 in Ryman & Jorde 2001), and the difficulty to obtain a significant exact P value is amplified when combining multiple values by means of Fisher's method.

Although Fisher's exact test is commonly applied when testing for allele frequency differences, it is arguable whether the underlying model is the most appropriate in this context. There is an ongoing discussion among statisticians (e.g. Mehta & Patel 1997; Agresti 2001) on when to apply different procedures for conducting an 'exact' contingency probability test. For example, Sokal & Rohlf (1995; p. 727 and onwards) discuss three models (I–III) where the Fisher test is of type III, conditional on both row and column totals being fixed. The type II model has either row or column totals fixed, whereas type I only has the grand total sample size fixed. Type I and II model tests are expected to provide a smaller degree of discontinuity due to the larger number of tables being evaluated. Thus, it is possible that the problem with small realized α errors, that in some situations characterizes the Fisher test, is less pronounced for type I and

II tests. The issue of the exact test procedure most suitable for genetic data warrants further investigation, though.

Implications for data analysis

Our present estimates of α and power may serve as guidelines for investigators evaluating existing data or planning a study. For example, the results for brown trout and herring (Fig. 6) should be possible to use as crude approximations for other authentic data sets with similar numbers of loci and allele frequencies. Still, we have only evaluated power for a very restricted combination of parameter values. Thus, we recommend power simulations of the type conducted here whenever there is uncertainty regarding the most appropriate choice of loci, statistical method, etc. In fact, this is the only way to address central questions raised by the present paper. Such questions include assessing, for a particular situation, what is to be considered a 'small' contingency table, 'few' alleles or samples, or 'skewed' allele frequencies. To this effect we have developed and made available a reasonably user friendly version of the simulation program used for this study (POWSIM; Ryman & Palm in press).

Assessment of power does not only provide a means for designing a study with 'high resolution'. It may also help to avoid wasting resources on studies that are unnecessarily powerful for the question at hand. Considering the herring microsatellite results in Fig. 6(b), for example, it is clear that scoring as much as 10 loci may be excessive for an investigator sampling five populations and only being interested in detecting divergence corresponding to an F_{ST} of 1% or more. Similarly, when scoring these 10 loci it would be enough to only sample two populations rather than five (or to reduce the number of individuals per sample), for maintaining a power close to unity when addressing this particular question. On the other hand, in situations where the interest is focused on detecting very small differences, such as when testing for temporal changes within a population (where the temporal 'drift signal' is inversely proportional to $2N_e$), not even the power provided by the 10 herring microsatellites may be sufficient.

The primary focus of this study refers to the question how to combine the information from multiple loci. With respect to a single locus (contingency table), it is clear that an exact test should be preferred before an approximate one. With multiple loci, however, the method of choice may vary from case to case, and it is frequently not clear what approach is 'best'. Most importantly, we have shown that Fisher's exact test (with multiple loci) is sometimes associated with a remarkably low power. Investigators should be aware of this fact and act accordingly when considered appropriate, for example through conducting additional tests such as chi-square. We have recently made a computer program available that tests for genetic differ-

entiation using both chi-square and Fisher's exact test (CHIFISH; Ryman 2006).

The apparent superiority of the chi-square approach in many situations when dealing with few alleles per locus may seem of minor importance considering the dominance of highly variable DNA markers in contemporary studies. It must be noted, though, that the generally preferred type of marker changes continuously, and that loci with few alleles may again be widely used. For example, di-allelic SNPs are foreseen to be commonly applied in the future (e.g. Brumfield *et al.* 2003; Morin *et al.* 2004). Thus, our results on remarkably low power when using Fisher's method (e.g. as in GENEPOP) to combine exact or permuted P values from multiple loci with two alleles (Fig. 2b, d) seem highly pertinent. This is particularly so because a small number of alleles per locus, as for SNPs, is likely to be compensated for by screening an increasingly larger number of loci (cf. Morin *et al.* 2004).

Similarly, the fact that the lack of power when combining P values with Fisher's method is most pronounced when comparing two samples may also seem of little importance when current studies on population structure are commonly based on a considerably larger number of samples. However, the extensive use of pairwise sample comparisons as a basis for data analysis and presentation justifies a serious concern regarding a potential lack of power in studies dealing with few allele loci.

Concluding remarks

On the basis of the present observations we recommend the following:

- 1 Assess power and α before launching a study. This can be done either through simulation, or more crudely through comparisons with estimates from similar investigations. For example, investigators should seriously consider whether the sample sizes and numbers of loci and alleles commonly used are really necessary for the particular question at hand.
- 2 When only considering a single locus (e.g. mitochondrial DNA) or contingency table, an exact test should be applied (rather than an approximation such as chi-square).
- 3 The uncorrected G -test should be avoided because of its tendency to produce excessive rates of false significances. Although not evaluated systematically in this study, it appears that application of Williams's correction reduces this rate substantially (to levels similar to those of chi-square).
- 4 Be aware that combining exact P values from multiple contingency tables (loci) by means of Fisher's method may under some circumstances result in a very low power. The risk for this phenomenon is most pronounced in pairwise sample comparisons when dealing with di-allelic loci with skewed allele frequencies.

- 5 When the results from multiple Fisher's tests tend to yield an unduly low power, summation of Pearson's traditional chi-square may constitute a good alternative. Chi-square seems to be more robust than commonly appreciated. The risk for an inflated α error should still be considered, though, particularly in situations when comparing small samples (say $n < 20$ – 30) with larger ones at loci with skewed allele frequencies.
- 6 In cases with several multi-allelic loci (e.g. 10 or more alleles), using Fisher's method to combine P values obtained by Fisher's exact test seems to constitute the method of choice, because of a generally high power and an apparent tendency not to exceed the intended α level.
- 7 When in doubt of whether to apply the Fisher's exact or chi-square test, power (and α) evaluations should be conducted. If this is not feasible we recommend that both tests are performed and their outcomes reported.
- 8 The relative merits of various statistical approaches when testing for homogeneity at multiple loci need to be examined in greater detail. In particular, the appropriateness of applying Fisher's exact test (that assumes that both column and row totals are fixed) should be reviewed. Similarly, the performance of G with Williams's correction and of various multilocus permutation procedures need further evaluation.
- 9 Several conclusions of this paper are not restricted to tests for genetic heterogeneity; they are valid in general when analysing contingency tables separately or in combination.

Acknowledgements

We thank the subject editor Phil Hedrick and three anonymous reviewers for comments on earlier versions of this paper. This work is part of the research project HERGEN (www.hull.ac.uk/hergen/) funded by the European Union within the Framework Programme 5. The study was also supported by grants to N.R. from the Swedish Research Council. Additional funding to C.A and L.L. was provided from the Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (Formas).

References

Agresti A (2001) Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*, **20**, 2709–2722.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution*, **18**, 249–256.

Choy SC, Weir BS (1978) Exact inbreeding coefficients in populations with overlapping generations. *Genetics*, **89**, 591–614.

Erdfelder E, Faul F, Buchner A (1996) GPOWER: a general power analysis program. *Behavior Research Methods, Instruments and Computers*, **28**, 1–11.

Goudet J (1995) FSTAT (version 1.2): a computer program to calculate F -statistics. *Journal of Heredity*, **86**, 485–486.

Goudet J (2001) FSTAT: A program to estimate and test gene diversities and

fixation indices (version 2.9.3). Available from www.unil.ch/izea/software/fstat.html. Updated from Goudet (1995).

Goudet J, Raymond M, de Meeüs T, Rousset F (1996) Testing differentiation in diploid populations. *Genetics*, **144**, 1933–1940.

Hedrick PW (1999) Highly variable loci and their interpretation in evolution and conservation. *Evolution*, **53**, 313–318.

Hedrick PW (2001) Conservation genetics: where are we now? *Trends in Ecology & Evolution*, **16**, 629–636.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity*, **88**, 62–65.

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity*, **94**, 33–36.

Mehta C, Patel N (1997) STATXACT for Windows. *Statistical Software for Exact Nonparametric Inference. User Manual*. CYTEL Software Corporation, Cambridge, Massachusetts.

Morin PA, Luikart G, Wayne RK, the SNP workshop Group (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, **19**, 208–216.

Nei M (1975) *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam, The Netherlands.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Neuhäuser M (2003) Tests for genetic differentiation. *Biometrical Journal*, **45**, 974–984.

Neuhäuser M (2004) Testing whether any of the significant tests within a table is indeed significant. *Oikos*, **106**, 409–410.

Palm S, Dannewitz J, Järvi T, Petersson E, Prestegard T, Ryman N (2003) Lack of molecular genetic divergence between sea-ranched and wild sea trout (*Salmo trutta*). *Molecular Ecology*, **12**, 2057–2071.

Peterman RM (1990) Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 2–15.

Petit E, Balloux F, Goudet J (2001) Sex-biased dispersal in a migratory bat: a characterization using sex-specific demographic parameters. *Evolution*, **55**, 635–640.

Raymond M, Rousset F (1995a) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity*, **86**, 248–249.

Raymond M, Rousset F (1995b) An exact test for population differentiation. *Evolution*, **49**, 1280–1283.

Rohlf FJ (1987) BIOM. A Package of Statistical Programs to Accompany the Text of Biometry. Applied Biostatistics, Inc., New York.

Ryman N, Palm S (in press) POWSIM – a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology Notes*.

Ryman N (2006) CHIFISH: a computer program testing for genetic heterogeneity at multiple loci using chi-square and Fisher's exact test. *Molecular Ecology Notes*, **6**, 285–287.

Ryman N, Jorde PE (2001) Statistical power when testing for genetic heterogeneity. *Molecular Ecology*, **10**, 2361–2373.

Ryman N, Lagercrantz U, Andersson L, Chakraborty R, Rosenberg R (1984) Lack of correspondence between genetic and morphologic variability patterns in Atlantic herring (*Clupea harengus*). *Heredity*, **53**, 687–704.

Schneider S, Roessli D, Excoffier L (2000) ARLEQUIN: A software for population genetics data analysis, version 2.000. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland.

Sokal RR, Rohlf FJ (1995) *Biometry*. 3rd edn. W.H. Freeman, New York.

- Waples RS (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity*, **89**, 438–450.
- Weir BS (1996) *Genetic Data Analysis II*. Sinauer Associates, Sunderland, Massachusetts.
- Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.
- Yeh FC, Yang R-C, Boyle T *et al.* (1997) *POPGENE (version 1.31)*, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre. University of Alberta, Canada. <http://www.ualberta.ca/~fyeh/info.htm>.

All coauthors share an interest in understanding the mechanisms that drive population structure in aquatic organisms, both marine and freshwater. This paper resulted from interactions within a European Union funded project, HERGEN, the primary aim of which was to examine the genetic population structure of Atlantic herring in the North Sea and adjacent areas, and to use the information in the context of management and sustainable use of aquatic resources.
