SC/F2K/J5

Evaluating the performance of different statistics to detect population subdivision

Barbara L. Taylor and Susan J. Chivers
Southwest Fisheries Science Center, 8604 La Jolla Shores Drive, La Jolla, CA 92038 U.S.A.

Abstract

We use statistical power to compare five different statistics commonly used to detect population subdivision: the haplotypic statistics χ^2, H_{ST}, F_{ST}, and the sequence statistics K^*_{ST} and ϕ_{ST}. The p-values for all statistics were estimated using randomization procedures. We evaluated these statistics at a high dispersal rate (1%/year) that is of interest in conservation applications. At this dispersal rate χ^2 always performed best and haplotypic statistics always outperformed sequence statistics, which differs from previous investigations that did not examine such high rates of dispersal. We show that the reason for poor performance of the sequence statistics is that when dispersal is high relative to genetic drift the phylogeographic signal is either low or non-existent. We suggest using simple diagnostics, such as regressing genetic differences against geographic distance, to choose the appropriate statistic. We advise that χ^2 should be used whenever possible, but that if the situation occurs where most individuals have unique haplotypes, K^*_{ST} is preferred to ϕ_{ST} because its performance is improved by down-weighting the phylogeographic signal relative to the frequency differences.

Introduction

Numerous statistics use genetic data in hypothesis testing to estimate whether population subdivision is present. Because sample size is often limited and research funding is usually scarce, it behooves scientists to use the most powerful statistics possible. We evaluate the performance of different commonly used statistics to detect population structure using mitochondrial DNA (mtDNA). We are particularly interested in the range of dispersal often of interest to conservation biologists. When dispersal is so low that genetic differences are significant on an evolutionary scale, population structure is easy to detect. However, when dispersal is high enough to muddy or eradicate the phylogeographic signal, yet low enough that subpopulations should be managed differently, population structure is likely to be difficult to detect. Therefore, we focus our investigation of the performance of different statistics within the dispersal range where we expect that statistical power will be low.

In a similar comparison, Hudson et al. (1992) found that performance, as measured by statistical power, depended on the migration rate, the mutation rate, and whether the marker(s) under consideration recombined. Statistics to evaluate population subdivision were divided into two types: "haplotypic statistics", where each haplotype is treated as a categorical variable and frequencies are compared, and "sequence statistics" that utilize the magnitude of differences between different haplotypes. They suggested a strategy to maximize power that switched from using haplotypic statistics to sequence statistics when haplotypic diversity (H_T) became high (when H_T is greater than approximately 0.95). The logical explanation for switching is that as diversity increases haplotypic frequencies decrease. Eventually most individuals have unique haplotypes and the resolving power of frequency-based statistics becomes poor.

Taylor et al. (SC/F2K/J4) developed a method to estimate power that used a Monte Carlo birth-and-death model that allowed 40 variable base-pairs of mtDNA to evolve. Although their primary interest was in temporally sampling the stepping-stone populations to estimate statistical power, the simulations yielded the opportunity to examine a higher dispersal rate than formerly examined and to see whether a model that was quite different from the coalescent approach used by Hudson et al. (1992) would yield the same result. Here we present the results of our performance evaluation of the statistics previously found to be most powerful for non-recombining markers: the haplotype statistics χ^2 and H_{ST}, plus we added the commonly used F_{ST}, and the sequence statistics K^*_{ST} plus the newer and commonly used ϕ_{ST} (Excoffier et al. 1992).

Methods

Power was estimated as described in Taylor et al. (SC/F2K/J4). The type of simulation model used to generate haplotypic frequencies influences the performance of the statistics so we provide some detail here. Because we wanted a model where we could specify abundances, include some simple spatial dynamics and capture the behavior of a commonly used genetic marker for population structure studies, we chose a Monte Carlo model arranged in a stepping-stone pattern where mitochondrial DNA (mtDNA) was allowed to evolve (Taylor et al. In Press; Taylor et al. SC/F2K/J4). The model allows annual dispersal to nearest neighbor populations. We chose a dispersal rate of 1%/year to represent a difficult case to detect population structure that will therefore contrast performance of the different statistics. Initially all individuals in the five populations had a single haplotype. We ran the model until the distributions for a number of parameters remained essentially constant: haplotypic diversity (H_T), the number of haplotypes and the measures of genetic differentiation. This stochastic equilibrium had occurred after 100,000 years. Once populations were in stochastic equilibrium we gathered genetic data every 25 generations (100 years). We gathered the following data at each discrete time interval: haplotype frequencies, haplotypic diversity, the actual measures of population differentiation (χ^2, H_{ST}, F_{ST}, K^*_{ST}, ϕ_{ST}), and the p-values for those measures for different sample sizes (n = 20, 40). For each statistic of differentiation, the p-values were estimated by performing 5,000 randomizations (Hudson et al. 1992). Thus, the null distribution for panmixia was formed by randomly assigning each individual to either population A or population B and calculating the statistics of differentiation. The p-value was the proportion of this null distribution that was equal to or greater than the observed value calculated for the sampled individuals. The simulation was run for 50,000 years yielding 500 sets of statistics. Statistical power is calculated as the proportion of time that H_0 is correctly rejected.

Results

The performance differed dramatically between different statistics (Table 1).

Abundance	Median H_T	χ^2	H_{ST}	F_{ST}	K_{ST}	ϕ_{ST}
100	0.70	0.87	0.81	0.81	0.79	0.71
1,000	0.95	0.71	0.53	0.53	0.36	0.29

Table 1. Haplotypic diversity (H_T) and statistical power at $\alpha = 0.05$ *for different measures of genetic differentiation for simulations run with two different abundances. Sample size was 40 from each population and the dispersal rate was 1%/year.*

The highest power was obtained using the randomization version of χ^2 (Rolf & Bentzen 1989), lower but similar values were obtained for H_{ST}, F_{ST}, and the sequence statistics (K_{ST} and ϕ_{ST}) performed most poorly. Surprisingly, ϕ_{ST} performed better for cases with low abundance and diversity but even so, Fig. 1 illustrates that much greater power (1 - Type 2 error) for a given level of α (Type 1 error) and sample size can be obtained simply by choosing a statistic that performs better at the task of differentiation.

Discussion

The different performance of statistics of differentiation was well documented previously (Hudson et al. 1992) but a number of our results were surprising. For the highest abundance we simulated (N = 1,000) H_T was right at the borderline (0.95) when sequence statistics (K^*_{ST}, ϕ_{ST}) should have been equal to or outperformed frequency statistics (χ^2, H_{ST}, and F_{ST}). Yet, the sequence statistics performed relatively more poorly than when diversity was low (Table 1). Presumably the decline in performance of frequency based statistics with high diversity occurs because the more haplotypes there are, the lower the mean frequency and the higher the chance that only a few individuals will be represented in any of the haplotypes. Fig. 2 shows an example of the haplotypic frequency distributions for N = 100 and N = 1,000. Although diversity is clearly much higher for N = 1,000, most individuals have common haplotypes (50% of the population have haplotypes with frequencies >5%, i.e. fairly common). Therefore, the frequency statistics continue to work well. As abundance continues to increase, however, the frequency distribution will become increasingly flat making it necessary to get higher sample sizes in order to use haplotypic statistics.

The skewed distribution of haplotype frequencies even when N = 1,000 and H_T = 0.95 explains why the haplotypic statistics continue to perform well, but it does not explain the decreasing performance of the sequence statistics K^*_{ST} and ϕ_{ST}. The underlying premise for why sequence statistics should add information to our picture of population structure is that the magnitude of differences (in our case the number of base-pair differences) should increase as geographic distance increases. In other words, an individual should on average be more closely related to her/his geographic neighbor than to an individual that is more geographically distant. If dispersal is occurring at a level that would allow meaningful genetic differences to accumulate (at less than one disperser/generation) then we would expect a strong correlation between genetic differences and geographic distance. However, as dispersal increases we expect to rapidly lose that phylogeographic signal. The strength of the signal will depend not only on dispersal but also on abundance because genetic drift allows small populations to differentiate more rapidly than large ones.

A simple way to examine patterns of genetic relatedness is simply to view how the mean base-pair difference relates to geographic distance. Recall that our stepping-stone model had five populations. Thus, we can calculate the mean base-pair difference between individuals for five within population comparisons (geographic distance of zero), four nearest neighbor comparisons (geographic distance of one), and so on to the final comparison of the first to the last population in the stepping stone series (geographic distance of four). For the small abundance case (Fig. 3a) there is a clear relation between genetic distance and geographic distance even though there is a good deal of noise. Note, for example, that there is one within population mean base-pair difference that is greater than between population differences that are two or even three populations distant. Thus, the sequence statistic can be expected to contribute some signal and much noise to the picture of population subdivision. In contrast, the larger abundance case (Fig. 3b) shows no relation between genetic differences and geographic distance. Thus, there is no signal plus a great deal of noise. It should not surprise us then than a statistic designed to clarify the population structure picture by incorporating a phylogeographic signal would instead make the picture fuzzier than if we had merely looked at frequencies. We suggest this technique as a simple diagnostic to suggest when sequence statistics are appropriate. We note, however, that even in the case with high drift because of a very low abundance (N = 100), the χ^2 performed best.

The dispersal rate we examined was much higher than Hudson's highest rate (Table 2). Roughly, our rates for 4Nm were 16 and 160 for the N = 100 and 1,000 cases respectively. Yet even for the lower range that Hudson et al. (1992) examined, we can see that χ^2 is consistently outperforming the best performing sequence statistic K^*_{ST}.

4Nm	K^*_{ST}	χ^2
1	0.99	1.00
2	0.94	0.99
5	0.67	0.86
10	0.45	0.59

Table 2. Statistical power from Hudson et al. (1992) Table 2 when sample size from both populations is 25.

Table 2 is not inconsistent with our results (Table 1) except in the magnitude of increased performance of χ^2 over $K^*_{ST.}$ It is interesting that Hudson et al. compared K_{ST}, which uses the number of base-pair differences to K^*_{ST}, which uses the log of the differences. Thus, for K^*_{ST} the magnitude of the differences, which is the phylogeographic signal, is down-weighted. They found that K^*_{ST} always yielded higher power than K_{ST}. That is, the sequence statistic performed better when the strength of the phylogeographic signal was reduced making it perform more like a haplotypic statistic. In light of our findings, we believe that this suggests that the potential of resolving population structure by adding phylogeographic properties (through using sequence data) is small and in cases of high dispersal adding phylogeographic data will greatly reduce our ability to detect population structure.

Thus, we suggest modifying Hudson et al.'s (1992) previous suggested strategy, particularly for cases where the dispersal rates of interest are high (relative to an evolutionary perspective): χ^2 is always the best strategy if sufficient sample size can be obtained. It is likely that if sample sizes are too low to run χ^2 that power will be so low as to make the value of any analysis of population structure questionable. It is possible, however, that even though a dispersal rate of greater than one disperser/generation would be of interest, the actual dispersal rate may turn out to be very low. In such a case the phylogeographic signal would be high, the noise low and the researcher may find highly significant differences using sequence statistics even with low sample size. This is just another way of saying that the researcher might get lucky and examine a case with a very large effect size.

Another alternative for investigating population structure when many or most individuals have unique haplotypes is to use K^*_{ST} instead of ϕ_{ST}. We are not suggesting that researchers try a battery of statistics and use whatever yields a significant result. Rather, we suggest that some simple diagnostics, like regressing genetic distances on geographic distances and seeing what proportion of the population have common haplotypes, should suggest which statistics will be both most appropriate and most powerful. If researchers are interested in population structure at levels with demographically trivial dispersal but evolutionarily high dispersal then sequence statistics are neither appropriate nor powerful.

References

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Hudson, R. R., D. D. Boos, and N. L. Kaplan. 1992. A statistical test for detecting geographic subdivision. Mol. Biol. Evol. 9:138-151.

Roff, D. A. and P. Bentzen. 1989. The statistical analysis of mitochondrial DNA polymorphisms: χ^2 and the problem of small samples. Mol. Biol. Evol. 6:539-545.

Taylor, B. L., S. J. Chivers, and A. E. Dizon. 1997. Using statistical power to interpret genetic data to define management units for marine mammals. In Molecular Genetics of Marine Mammals. eds. A. E. Dizon, S. J. Chivers, and W. F. Perrin. Special Publication 3:347-364 Allen Press, Inc., Lawrence, Kansas, U.S.A.

Taylor, B. L., S. J. Chivers, and A. E. Dizon. SC/F2K/J4. Estimating the statistical power to detect population subdivision using mitochondrial DNA.

Taylor, B. L., S. J. Chivers, S. Sexton and A. E. Dizon. In press. Using simulation models that incorporate uncertainty to estimate dispersal rates from mitochondrial DNA data. Conservation Biology.

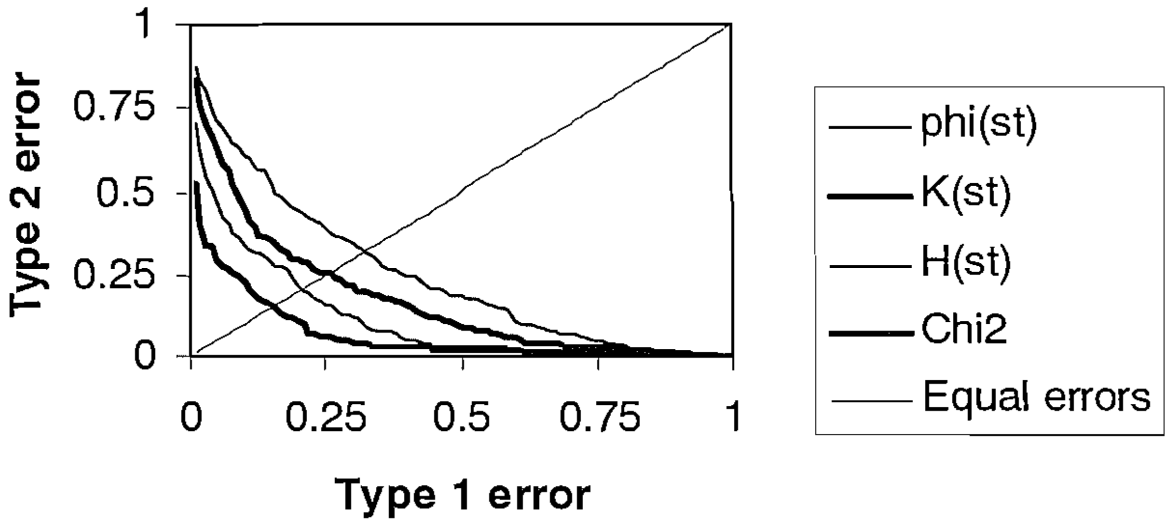


Figure 1. A comparison of statistics of population subdivision differentiation using tradeoff curves. The case shown is when both populations had an abundance of 1,000 and 40 samples were taken from each. Statistics are listed from least powerful (highest line,ϕ_{ST}) to most powerful (lowest line χ^2). The line of equal Type 1 to Type 2 errors is also shown for visual clarity.

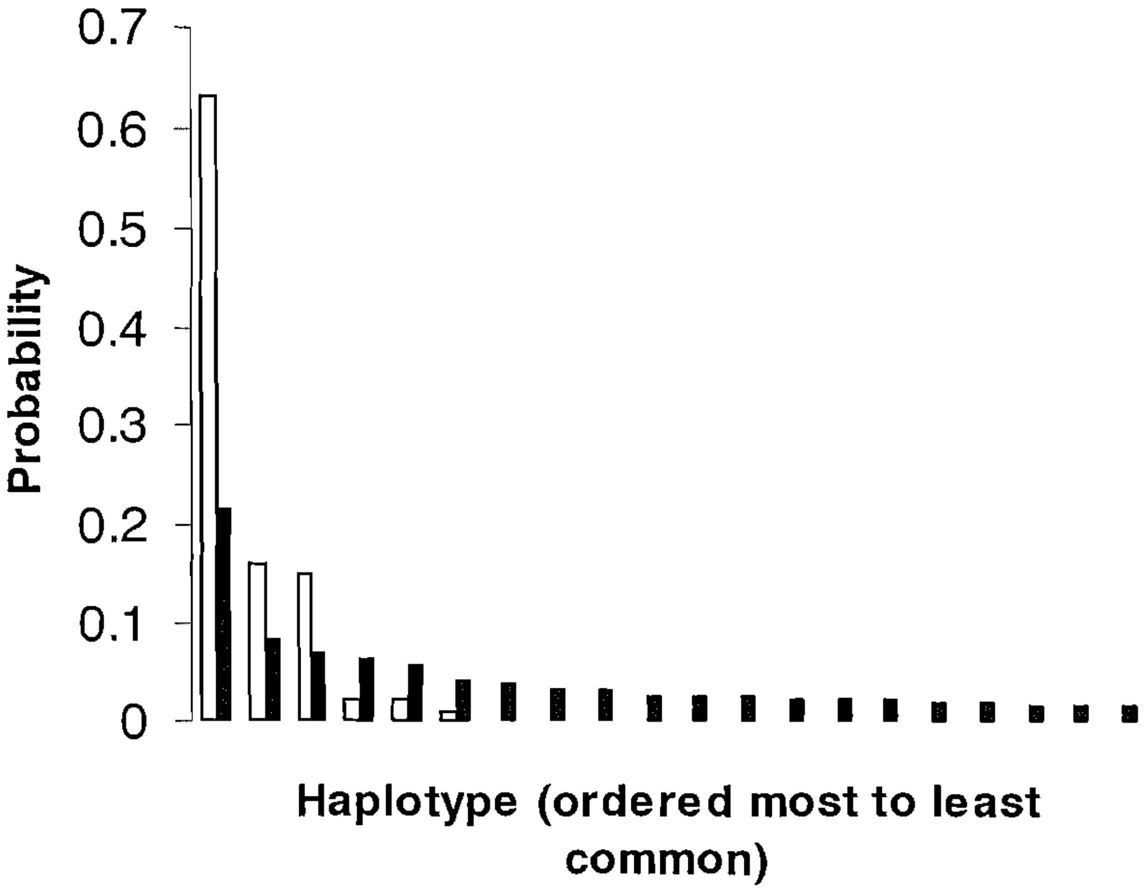
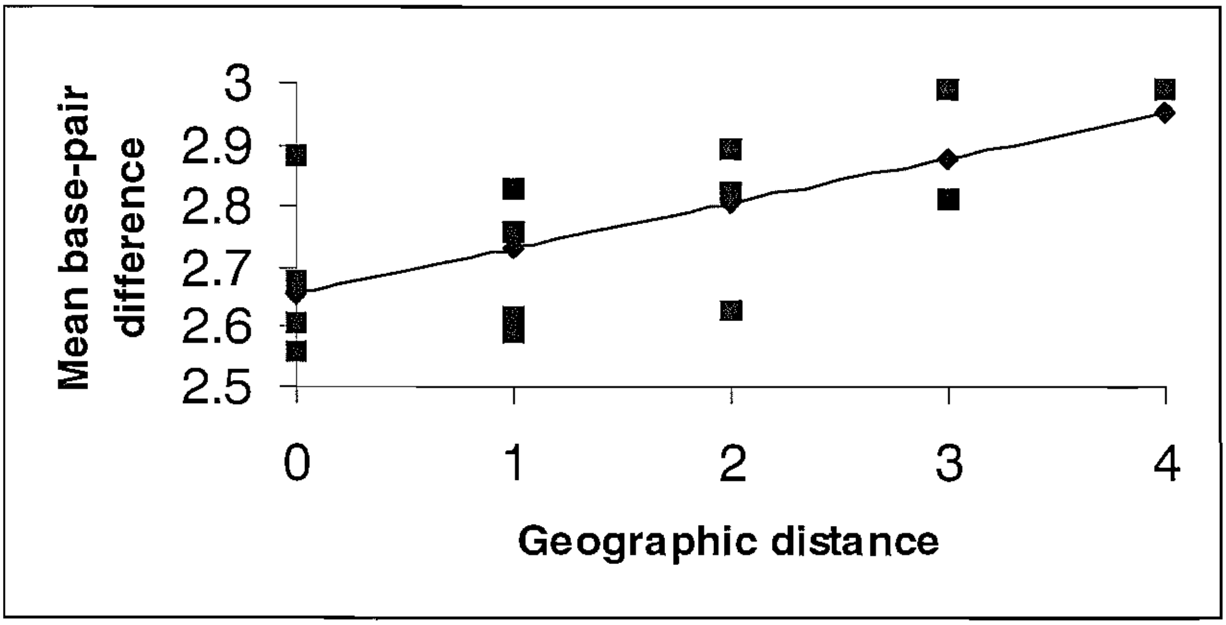


Figure 2. The probability of individuals belonging to different haplotypes for sample population pairs with abundances of 100 (white bars) and 1,000 (black bars). For visual clarity we show only the 20 most common haplotypes which would include 100% of the individuals for the N = 100 case and 88% of the N = 1,000 case.



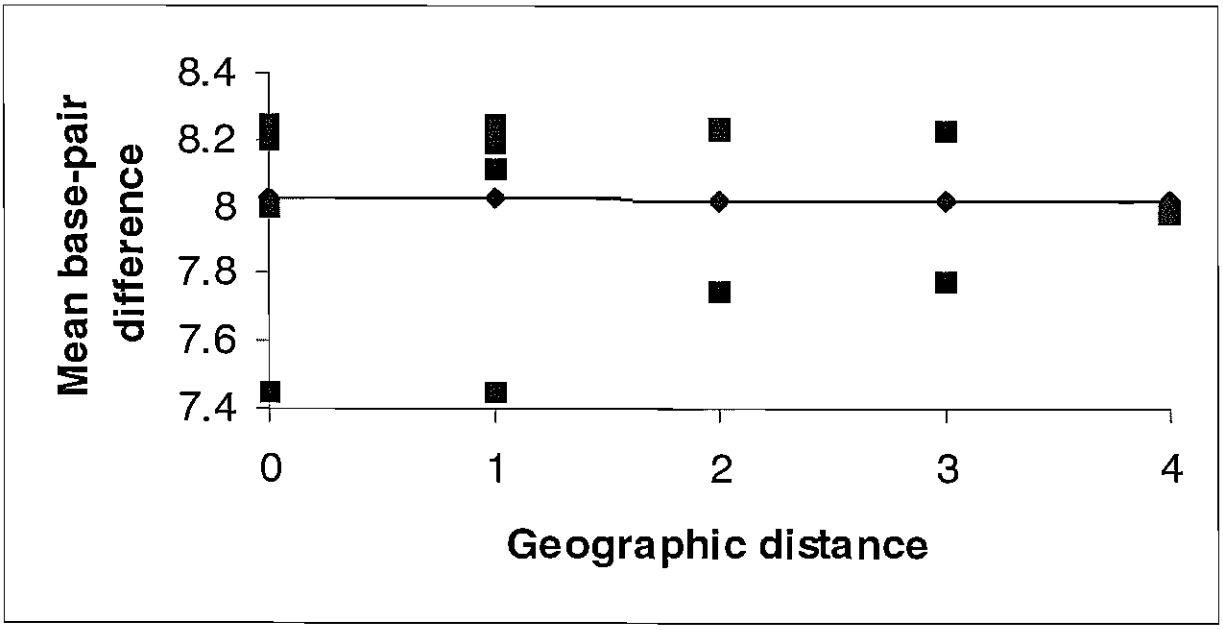


Figure 3. Mean base-pair difference by geographic distance for observed population comparisons (black squares) and as predicted by regressing base-pair differences on geographic distance (diamonds connected by line for visual clarity) for N = 100 (3a) and N = 1,000 (3b).