

Experimental design in intraspecific organelle DNA sequence studies III: statistical measures of sampling success

Jason A. Holt¹, Sierra D. Stoneberg Holt^{1,2}, Petr Bureš²

¹ Box 37, Hinsdale, Montana 59241, U.S.A. jholt@seznam.cz (author for correspondence)

² Institute of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, Brno 611 37, Czech Republic

Statistical methods are proposed for analyzing the experimental design, preliminary results, and final results of phylogenetic studies of organelle DNA sequence at low taxonomic levels. Such studies require sampling numerous individuals, many of which share identical haplotypes. The proportions of the haplotypes sampled can help answer the following questions: (1) Is one haplotype so dominant that the particular DNA region is without meaningful variation within the scope of the study? (2) Were all prevalent haplotypes found? (3) What are the proportions of each haplotype within the studied group? (4) What percentage of the studied group can be confidently asserted to belong to the haplotypes that were found? Examples are given in which the statistics techniques are applied to data drawn from the botanical literature. Tables are included as a quick reference for the researcher who wishes to circumvent calculation. A Microsoft® Excel 2000 spreadsheet (titled “HaploPro.xls”) for performing some of the more complicated calculations is offered online. Finally, the limitations of these methods and their applicability to nuclear DNA and other characters studies are discussed.

KEYWORDS: bulking, haplotype frequency confidence intervals, inclusion-exclusion principle, intraspecific sequence studies, phylogeography, proportion estimates

INTRODUCTION

Intraspecific DNA sequence studies, such as phylogeographic studies, are characterized by large sample sizes and low diversity (Whittemore & Schaal, 1991; Lessa & Applebaum, 1993; Demesure & al., 1996; Strand & al., 1996; Dumolin-Lapègue & al., 1997, 1998; Schaal & Olsen, 2000; Mátyás & Sperisen, 2001; Posada & Crandall, 2001). Such studies pose specific challenges for statistical analysis. This paper proposes statistical methods that are appropriate for use in evaluating DNA sequence data at the intraspecific level.

Intraspecific organelle DNA sequences. — The raw data obtained in an organelle DNA sequence study at low taxonomic levels consists of a large number of samples assigned to a relatively small number of haplotypes. Ideally, DNA sequences from multiple regions of the genome will be used (Lessa & Applebaum, 1993; Nordborg & Innan, 2002). For example, Demesure & al. (1996: 2516) found 11 haplotypes in a survey of 2 regions of plastid DNA for 399 individuals of *Fagus sylvatica* L. in Europe. Strand & al. (1996: Table 2) found 5 haplotypes in a survey of one region of plastid DNA for 251 individuals from two *Aquilegia* spp. in the southwestern United States and Mexico. Dumolin-Lapègue & al. (1998: 1324) found 12 haplotypes in a survey of 2 regions of plastid DNA and 2 regions of mitochondrial DNA for 1,749 individuals from four *Quercus* spp. from southern France.

In this paper, “haplotypes” are classes of DNA sequences distinguishable by the methods employed to analyze the sampled individuals. Sequencing each specimen is not necessarily a feasible method of determining haplotypes in intraspecific phylogenetic studies (Lessa & Applebaum, 1993; Cotton, 1997; Nataraj & al., 1999; Taylor, 1999; Stoneberg Holt & Bureš, 2007), and specimens are often assigned to haplotype based on, e.g., banding patterns rather than actual DNA sequence. Thus, screening methods may not uncover all unique sequences within the sample. Methods used to screen for haplotypes in previously published studies include restriction fragment length polymorphisms of PCR products (PCR-RFLP) (Demesure & al., 1996; Dumolin-Lapègue & al., 1997, 1998; Mátyás & Sperisen, 2001), denaturing gradient gel electrophoresis (DGGE) (Strand & al., 1996), single-strand conformation polymorphism (SSCP) (Fujii & al., 1997), and allele-specific amplification (Dumolin-Lapègue & al., 1998).

Sample bulking. — Because so many specimens share the same haplotype, it can be efficient to mix PCR products from a number of individuals and screen them all simultaneously—a technique known as bulking. Effective bulking requires a reasonably high probability that all individuals in the bulk will share an identical haplotype and a screening method sensitive enough to recognize differing haplotypes despite the dilution factor. When these conditions are met, many of the bulks will be found to be homogeneous, thus saving the time and money that would have been spent

analyzing each individual in the bulk. Heterogeneous bulks must be analyzed further—unless the researcher decides, based on the preliminary bulking study, that one haplotype is so dominant that no meaningful information can be obtained by further analysis. A technique for calculating the percentage dominance from the results of a bulking study is described in the Methods section.

Bulking experiments have been carried out to determine the limits of sensitivity of some screening methods. For example, chemical cleavage of mismatch (CCM) with fluorescent visualization was sensitive enough to use with bulks the equivalent of 9 sampled individuals plus standard (Verpy & al., 1994) and enzyme mismatch cleavage (EMC) was effective with bulks the equivalent of 15 sampled individuals plus standard using the enzyme CEL I (Colbert & al., 2001) and up to 20 sampled individuals plus standard using the enzyme *Tma* Endonuclease V with ligase (Huang & al., 2002). These may be the upper limits of bulking sensitivity, or future advances may extend them. All these bulking-tested methods require the formation of heteroduplexes between a known sequence and the tested sample (Stoneberg Holt & Bureš, 2007). Methods based on different approaches may also prove amenable to bulking.

Statistical evaluation of variation. — This paper addresses several of the statistical questions about the relationship between the actual haplotype distributions and those observed in a phylogenetic study. They are:

- (1) If only one haplotype is found, what is the statistical support for the claim that the given combination of DNA region, organism, geographic area, and screening method is without meaningful variation?
- (2) If one haplotype is dominant, what is its proportion within the studied group?
- (3) What percentage of homogeneous bulks from a preliminary bulking survey indicates lack of meaningful variation?
- (4) How likely is it that all common haplotypes were found by the study?
- (5) What are the proportions of each haplotype within the studied group?
- (6) What percentage of the studied group can be confidently asserted to be made up of the haplotypes that were found?

METHODS

After analyzing a few of the specimens collected, the researcher may already have some idea of the relative frequencies of the haplotypes in the study. If most of the analyzed specimens have the same haplotype, the researcher may suspect that the variation in the study will

be uninformative. The first three techniques discussed below are those which help the researcher determine the proportion of the dominant haplotype. The exact percentage of dominance required to justify modifying the parameters of the study is left to the researcher.

The fourth technique gives the researcher an a priori target sample size which provides a certain confidence level that all prevalent haplotypes will be found by the sample.

Once all the data have been collected, the final two techniques can be applied to estimate the proportions of the haplotypes within the studied group and to calculate, a posteriori, the minimum proportion of the population that can be confidently assigned to the found haplotypes.

Estimating percentage dominance when only one haplotype is found. — If all sampled individuals have a single haplotype, it may be that the haplotype strongly dominates the studied group. The researcher cannot be certain that there are no other haplotypes to be found, but it is possible to estimate the single haplotype's proportion within the population (p). The probability of sampling this haplotype exclusively in a sample of size n is p^n (assuming that the studied group is large compared with the sample size). If the observed data correspond to an event with probability greater than a given significance level α , then $p^n \geq \alpha$. This relationship establishes a lower bound for p with confidence $1 - \alpha$:

$$p \geq \sqrt[n]{\alpha}. \quad (\text{Eq. 1})$$

For example, if the researcher tests 44 individuals and observes only a single haplotype, the researcher can be 99% confident that this haplotype represents at least $\sqrt[44]{1-0.99} \approx 90\%$ of the studied group; i.e., if the haplotype is found in less than 90% of the studied group, the probability of obtaining the observed results is less than 1%.

This relationship can also be used to determine a sample size n that guarantees a probability of at least $1 - \alpha$ of finding multiple haplotypes, provided the most prevalent haplotype represents less than a certain proportion p of the population. Table 1 provides minimal sample sizes for selected levels of dominance and standard significance

Table 1. When a sample contains only one haplotype, this table shows the sample size necessary to support claims that that haplotype dominates the studied group. The presented values are sample sizes (n) at selected levels of dominance (p) and significance (α).

Probability of sampling at least one additional haplotype ($1 - \alpha$)	Percentage of study group represented by a single haplotype (p)		
	90%	95%	99%
90%	22	45	230
95%	29	59	299
99%	44	90	459

levels. (For a discussion of using this technique to demonstrate uniparental inheritance of organelle DNA, see Milligan [1992].)

This statistic can be applied to results already found in the literature. In the *Fagus sylvatica* phylogeographic study by Demesure & al. (1996: 2516), a pilot test of 10 plastid and 4 mitochondrial PCR products was made in 12 individuals. For 4 plastid and 4 mitochondrial products, no variation was determined, and these primers were abandoned. These researchers could be 90% confident that for these primers, the one found haplotype would be present in at least 82% of *F. sylvatica* (but see “Caveats” in the Discussion section). Mátyás & Sperisen (2001: Table 2) found only Haplotype 1 in 50 *Quercus petraea* (Matt.) Liebl. individuals from population CH145. These researchers could be 90% confident that Haplotype 1 accounts for at least 95% of this population, 95% confident that it accounts for at least 94%, and 99.9% confident that it accounts for at least 87%.

Calculating a confidence interval for the proportion of a single haplotype. — Even if a few of the specimens analyzed have a different haplotype, the researcher may feel that one haplotype dominates the sample. Does it dominate the group studied? To answer this question, the researcher can calculate a confidence interval for the proportion.

Calculating a confidence interval for a single proportion is a procedure well known to biologists in general, and it has already been used in phylogenetics in particular: Mátyás & Sperisen (2001) calculated confidence intervals for the haplotype proportions they found in populations of *Quercus* spp. in the Swiss Alps using (Eq. 3) below. Single proportion confidence interval techniques are presented here (1) for completeness, (2) to discuss techniques which may be more appropriate than the standard technique, and (3) to introduce equations which will be needed for discussions in later sections.

The confidence interval presented by “most introductory statistics textbooks” (according to Agresti & Coull [1998: 119]) is:

$$\hat{p}_{\pm} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \quad (\text{Eq. 2})$$

where \hat{p} is the proportion of the haplotype in the sample, \hat{p}_{-} and \hat{p}_{+} are the lower and upper bounds of the confidence interval, n is the number of specimens in the sample, $z_{\alpha/2}$ is the critical value of the standard normal distribution corresponding to a tail with probability $\alpha/2$, and α is the level of significance chosen by the researcher (i.e., $1-\alpha$ is the level of confidence desired). Agresti & Coull (1998) refer to this as the “Wald confidence interval.”

This approximation is less reliable for small n or for values of \hat{p} near 1, making it inappropriate for this application. Newcombe (1998: 868) goes so far as to say

“... it is strongly recommended that intervals calculated by these methods ([Eq. 2] and its counterpart with continuity correction) should no longer be acceptable for the scientific literature ...”

Agresti & Coull (1998) make a good case for using the score confidence interval (also reviewed in Newcombe, 1998):

$$\hat{p}_{\pm} = \frac{2n\hat{p} + A \pm \sqrt{A(A + 4n\hat{p}(1-\hat{p}))}}{2(n+A)}, \quad (\text{Eq. 3})$$

where A is $z_{\alpha/2}^2$ and the other symbols are as in (Eq. 2) above. (Note that, unlike the interval in [Eq. 2], \hat{p} is not in the center of the interval.) HaploPro (Electronic suppl.) can be used to calculate score confidence intervals for a single proportion. See Appendix 1.

Agresti & Coull (1998) demonstrated that for 95% confidence intervals, the Wald interval (Eq. 2) can be adjusted to perform similarly to the score interval (Eq. 3) when one artificially adds two successes and two failures. If k of the n specimens analyzed are of the dominant haplotype, then $\hat{p} = k/n$. To add two successes and two failures, replace \hat{p} by $(k+2)/(n+4)$ and n by $n+4$ in (Eq. 2).

The score interval performs well even for small values of n , but for certain values of the true proportion p near 1 the probability of obtaining a confidence interval that contains p is much less than the nominal confidence level (Agresti & Coull, 1998; Newcombe, 1998)—e.g., a probability of only 83.1% that a random sample will yield a \hat{p} giving a 95% confidence interval containing p (Newcombe, 1998: 868). The problem, usually (Agresti & Coull, 1998: 122; Newcombe, 1998: Table II), is that the actual p is closer to 1 than the upper bound of the confidence interval. The researcher requiring a precise upper bound on p for small n and only one individual with a different haplotype ($k = n-1$) should use an exact method to calculate the confidence interval. In the context of this paper, however, the researcher already knows from the data that p is near 1, but not exactly 1, and is not concerned with discovering how near. The researcher is interested primarily in the lower bound—desiring to know if there is, e.g., a 10% chance that the actual level of dominance is low enough to warrant further study—and the score interval should be suitable for this application.

See Mátyás & Sperisen (2001) for an example of using the score confidence interval in the literature. An example is also given at the end of the following section.

Estimating dominance of a haplotype from bulked data. — A researcher may begin by testing each sampled organism individually, expecting to find a variety of haplotypes. If nearly all of the first individuals tested have the same haplotype, the researcher may wish to consider bulking, if the haplotype detection method permits. (See Introduction.)

When bulking with a heteroduplex-based method (see Introduction), several individuals are tested against the same standard, yielding a binary result: either all match the standard, or at least one differs from the standard. When one haplotype (the standard) is dominant, most of the bulks will match the standard. Individuals from the bulks that do not match can be tested further, but if rare haplotypes are uninformative for the study, the researcher may need only an estimate of the proportion of the dominant haplotype. This estimate can be obtained from the bulked data, even without knowing the exact contents of the non-matching bulks.

If a sample of size n is divided into m bulks of size b ($n = mb$) and k of those bulks match the standard, then the quantity k/m provides an estimate of p^b , where p is the proportion of the dominant in the population sampled. (The principle is the same as that discussed in the context of [Eq. 1] above.) Thus p can be estimated by: $\hat{p} = \sqrt[b]{k/m}$. Note that this proportion calculated from the bulked data is less reliable than the proportion that could be obtained if each specimen's haplotype were known because $m < n$.

Table 2 gives some values of \hat{p} for selected combinations of b and k/m . Note that even if only half of the bulks match the standard, it can indicate that the haplotype is strongly dominant, if the bulks are large enough.

The precision of this estimate depends on the number of bulks (m). A confidence interval for p^b , yielding bounds \hat{p}_- and \hat{p}_+ , can be calculated using one of the methods discussed in the previous section. (Note that k/m should be substituted for \hat{p} , and m for n when using [Eq. 3] for this purpose.) Taking the b -th root of these bounds yields bounds \hat{p}_- and \hat{p}_+ . The level of confidence remains the same.

This technique is based on the assumption that the individuals to be bulked are grouped independently, i.e., grouped according to an arbitrary or random criterion having no relation to haplotype. If, for example, individuals are grouped according to collection site, then the analysis above does not apply. Thus, this design for a bulking experiment is recommended only when the researcher

has reason to believe that there is little variation or if the researcher is only interested in a rough estimate of the proportion of the dominant haplotype. If variation is expected, the researcher should group together specimens expected to have the same haplotype, and proportions should be calculated from the complete data once every individual has been classified.

Strand & al. (1996: 1823) studied *Aquilegia* spp. "found in rather small, isolated populations in mesic, high-altitude canyons separated by intervening desert." In such a situation, researchers may suspect that each population will probably be homogeneous, possibly for different haplotypes, and such was the case for 14 of 18 tested populations (Strand & al., 1996: Table 2). The paper implies (p. 1824) that at least 526 samples plus standards were run to test 251 individuals. Therefore, an experimental design that preliminarily bulked members of the same population might significantly reduce the time and cost of the study and/or allow more individuals to be sampled, even if all individuals were re-tested separately to confirm the preliminary results. However, because such bulks would not be grouped independently but based on predicted structure within the data, the technique presented in this section would not be applicable.

This technique relies on having all bulks be of the same size (b). In practice, this is rarely the case. The researcher may have decided to bulk after screening several specimens individually. These individuals can all be considered as one bulk of size b , either a matching bulk if they all have the same haplotype or a non-matching bulk if at least one of them has a different haplotype. If the number of these individuals is greatly in excess of b , the researcher may wish to treat them as several bulks of size b . Making groups after the fact makes the statistics less valid, but the researcher can guarantee a conservative estimate by distributing the non-dominant haplotypes evenly. Even if the researcher starts bulking from the beginning, it is unlikely that the number of samples will work out perfectly so that $n = mb$. Some will be left over. The simplest thing to do is to ignore this irregularly sized bulk when computing the statistics. The especially conservative researcher can ignore this bulk if it matches and include it in the analysis if it does not match the standard.

A hypothetical example of the use of this statistic can be based on the data of Dumolin-Lapègue & al. (1998: Table 2). In studying the association between plastid and mitochondrial lineages in 4 *Quercus* spp. from southern France, they found 518 studied members of plastid Haplotype 10 to have one phase of a stem-loop induced micro-inversion and 4 to have the opposite phase. Had they wanted a preliminary estimate of the predominance of the common haplotype and been using a method conducive to bulking, they could have made 32 bulks of 16 individuals each (with 10 individuals left over). The 4 individuals with

Table 2. This table gives the proportion of a dominant within a studied group, based upon the results of a bulking study. The presented values are estimates of the proportion of the dominant (\hat{p}) at given combinations of individuals per bulk (b) and proportions of bulks found to be homogenous (k/m) expressed as a percentage. The value m is the number of bulks, and k is the number of those bulks matching the standard.

Percentage of bulks matching standard (k/m)	Number of individuals in each bulk (b)			
	2	4	8	16
50%	0.707	0.841	0.917	0.958
80%	0.894	0.946	0.972	0.986
90%	0.949	0.974	0.987	0.993

the rare haplotype would have been found in from one to four bulks. If four bulks were heterogeneous, then 87.5% of the bulks would be homogeneous. Taking the 16th root gives an estimated proportion of $\hat{p} = \sqrt[16]{0.875} = 99.2\%$. Calculating the 95% score confidence interval (Eq. 3) gives $\hat{p}_-^{16} = 0.719317$ and $\hat{p}_+^{16} = 0.950299$. Taking the 16th root of these bounds gives a 95% confidence interval of 97.9%–99.7%, where the lower bound is rounded down and the upper bound is rounded up. If all 4 fell into one bulk (the least likely scenario), the estimated proportion would be 99.8% with a 95% score confidence interval of 98.9%–100%. This compares well with the values found by screening all 522 specimens individually: an estimated proportion of 99.2% with a 95% score confidence interval of 98.0%–99.8%.

Designing a study to sample all prevalent haplotypes. — Ideally, more than one haplotype will be well represented within a study. Unless every individual in the group of interest is tested, it is impossible to be certain that all haplotypes have been observed; however, one can calculate the probability of observing all haplotypes with substantial representation in the population. For a given sample size (n), the probability (p) of observing all haplotypes which have a proportion in the population of at least $1/k$ is bounded by

$$p \geq 1 + \sum_{i=1}^{k-1} (-1)^i \binom{k}{i} \left(1 - \frac{i}{k}\right)^n \quad (\text{Eq. 4})$$

(Note that the meanings of p and k here differ from the usage in other sections.) A discussion of this inequality can be found in Appendix 2. A Microsoft® Excel 2000 spreadsheet, HaploPro.xls (Appendix 1; Electronic suppl.), has been implemented that performs this calculation.

Table 3, calculated using HaploPro, shows sample sizes necessary to guarantee selected probabilities of sampling all haplotypes that represent a given minimum percentage of the population. It can also be used to get a rough estimate of the coverage of a particular sample. For example, if the researcher tests 100 individuals, that is enough to be 99.9% certain that all haplotypes with a proportion of 10% or more of the studied group have been found, but not quite enough to be 90% certain that all haplotypes with a proportion of 5% or more have been found. Table 3 indicates that it is easier to achieve a high probability of sampling all haplotypes of a moderate size than a moderate probability of sampling all haplotypes of a small size.

HaploPro can be used to evaluate the sampling rigor achieved by Dumolin-Lapègue & al. (1998: 1322). Sampling 1,749 *Quercus* spp. individuals gives over 99.999% certainty that every haplotype that makes up more than 1% of the population of southern France has been sampled. By contrast, in a survey of many traits of *Dupontia fisheri* R. Br., Brysting & al. (2004) sequenced the *trnL-F* region of 19 specimens. These researchers can be over 99.999%

certain that they have sampled every haplotype that makes up at least 50% of the population and 90% certain that they have sampled every haplotype that makes up at least 20% of the population.

Confidence intervals for multiple proportions. — The proportion of a haplotype within the sample (\hat{p}_i) provides an unbiased estimator of its proportion within the studied group. The precision of the estimators can be measured by calculating simultaneous confidence intervals for multinomial proportions. The intervals are simultaneous in the sense that the researcher can be confident that each true proportion is in the given interval, i.e., the confidence level applies to all the intervals simultaneously.

The standard method for evaluating simultaneous confidence intervals for multinomial proportions is Goodman's (1965) method, based on the work of Quesenberry & Hurst (1964). The upper (\hat{p}_i^+) and lower (\hat{p}_i^-) bounds of the interval are given by:

$$\hat{p}_i^\pm = \frac{2n\hat{p}_i + A \pm \sqrt{A(A + 4n\hat{p}_i(1 - \hat{p}_i))}}{2(n + A)},$$

where n is the sample size. This is identical to (Eq. 3) above, except that here the parameter A is chosen differently. Given the number of haplotypes found (h) and the desired confidence level ($1 - \alpha$), let $\alpha' = \alpha/(2h)$. In the normal distribution table, find the critical z -value corresponding to α' , i.e., $z_{\alpha'}$ is the value such that $P(X \geq z_{\alpha'}) = \alpha'$ for a random variable X with standard normal distribution. Then $A = z_{\alpha'}^2$. The spreadsheet HaploPro (Electronic suppl.) uses this method to estimate confidence intervals for haplotype proportions.

Note that when z is chosen to correspond to $\alpha/2$ instead of $\alpha/(2h)$, the formula gives the score confidence interval for a single proportion (Eq. 3) discussed above. Mátyás & Sperisen (2001) used (Eq. 3) (with $z = 1.96$, corresponding to $\alpha/2 = 0.025$) to calculate non-simultaneous 95% confidence intervals for the frequencies of the 3 haplotypes they found in a population (CH147) of *Quercus robur* L. In this case, the confidence statement applies to each of the three proportions individually, but is not applicable to

Table 3. This table gives the sample sizes necessary to claim that all prevalent haplotypes have been sampled. The presented values are sample sizes (n) at given minimum proportions ($1/k$) and probabilities (p), both expressed as percentages.

Probability of sampling all prevalent haplotypes (p)	Minimum proportion of prevalent haplotypes ($1/k$)		
	10%	5%	1%
90.0%	44	103	683
95.0%	51	117	754
99.0%	66	149	916
99.9%	88	194	1,146

all intervals simultaneously. For any proportion considered individually, one can be 95% confident that it lies in its interval, but one can only be 85% confident that all three lie in their intervals simultaneously. (Mátyás & Sperisen [2001] used $z = 1.96$, corresponding to a tail probability of 0.025. For Goodman's [1965] method, $0.025 = \alpha/(2h)$. In this case $h = 3$. So $\alpha = 6 \times 0.025 = 0.15$. Thus the confidence level for Goodman [1965] confidence intervals is $1 - \alpha = 0.85$.)

Goodman's (1965) method applies to a multinomial distribution where every individual is in one of h categories. In practice, there may be an unknown number of haplotypes which have not been found. However, because these haplotypes are probably rare (otherwise they would have been found by the study), it is reasonable to approximate the underlying multinomial distribution with one having only h categories.

There are several alternative methods for calculating simultaneous confidence intervals for multinomial proportions. Fitzpatrick & Scott (1987) observe that opinion polls frequently use $\hat{p} \pm 1/\sqrt{n}$ where n is the sample size and \hat{p} is the sample proportion. They demonstrate that this gives 90% confidence intervals and recommend $\hat{p} \pm 1.13/\sqrt{n}$ for the 95% level and $\hat{p} \pm 1.40/\sqrt{n}$ for the 99% level. May & Johnson (2000) give a macro for the SAS computer statistics package that constructs intervals using the method of Sison & Glaz (1995) and observe that Goodman's (1965) method performs well when the number of cells (haplotypes) is 10 or less, while Sison & Glaz's (1995) method is good for cases where there are many cells, each with roughly the same number of observations. It is expected that intraspecific organelle DNA studies will generally fall into the former category.

Simultaneous confidence intervals are to be preferred over single-proportion intervals for several reasons. First, they can be used to estimate the proportion of the population that can be confidently assigned to a haplotype, as described below. Second, they are more conservative (though larger) than single-proportion intervals. Third, they apply to the studied group as a whole. There may be cases, however, where the researcher is focusing only on one particular haplotype and the single-proportion interval is more appropriate.

Determining confidence intervals with HaploPro (Appendix 1; Electronic suppl.) can be demonstrated on plastid data from three *Quercus* spp. (Dumolin-Lapègue & al., 1998; Mátyás & Sperisen, 2001). The data used in these calculations are shown in Table 4. Dumolin-Lapègue & al. (1998) can be 99% confident that in southern France the true proportion of Haplotype 1 lies between 6.4% and 10.9% while simultaneously the true proportion of Haplotype 7 lies between 37.0% and 44.8%. Mátyás & Sperisen (2001) can be 99% confident that in the Swiss Alps the true proportion of Haplotype 1 lies between 31.1% and 41.0% and the true proportion of Haplotype 7 lies between 56.8% and 66.7%. Thus, there is strong statistical evidence that Haplotype 7 is more prevalent than Haplotype 1 in both regions, and both haplotypes make up a larger proportion of the population in the Swiss Alps than in southern France. This result appears to be supported by the tables and maps from the European-wide study of Dumolin-Lapègue & al. (1997), which show that Haplotypes 10, 11, and 12 are also common in southern France.

Measuring sampling thoroughness a posteriori.

— A technique has been presented above which ensures a given probability of sampling all haplotypes that are sufficiently prevalent in the studied group. This is not the same as guaranteeing that the proportion of unrepresented haplotypes is small, because there may be a number of rare haplotypes that add up to a large percentage of the population. For example, Table 3 shows that a sample of size 44 has a 90% probability of including every haplotype that represents at least 10% of the population. However, if the studied group contains 100 haplotypes, each with a proportion of only 1%, no sample of 44 can represent more than 44% of the population.

If the sample consists of only one haplotype, then (Eq. 1) can be used to estimate the proportion $(1-p)$ of all haplotypes not sampled.

If the sample consists of several haplotypes, simultaneous confidence intervals can be used to establish a lower bound for the proportion of each haplotype. The researcher is, e.g., 90% confident that every true proportion (p_i) is greater than its estimated lower bound (\hat{p}_i). The sum of these lower bounds, $\sum \hat{p}_i$, represents the proportion of the population that can be confidently classified into

Table 4. This table summarizes the data taken from the botanical literature and used to demonstrate the use of HaploPro for determining confidence intervals. The studied species were *Quercus robur*, *Q. pubescens* Willd., and *Q. petraea*.

Study	Number of sampled individuals	Number of haplotypes found	Individuals belonging to Haplotype 1	Individuals belonging to Haplotype 7
Dumolin-Lapègue & al. (1998: Table 2), southern France	1,710	9	143	699
Mátyás & Sperisen (2001: Table 1), Swiss Alps	1,036	10	372	641

Note: Mátyás & Sperisen (2001) did not include *Q. pyrenaica* Willd. For this table, 39 *Q. pyrenaica* individuals were excluded from Dumolin-Lapègue & al. (1998) using the assumptions (supported by Dumolin-Lapègue & al., 1997) that they had no unique haplotypes and no members belonging to Haplotype 1 or Haplotype 7.

one of the found haplotypes. The remainder consists of found and unfound haplotypes.

Goodman’s (1965) method can be used to establish these lower bounds, but a few refinements are recommended. As noted above, Goodman’s interval is simply a score confidence interval with a different choice of α' . For a single proportion, $\alpha' = \alpha/2$; for simultaneous confidence intervals $\alpha' = \alpha/(2h)$, where h is the number of observed haplotypes. The 2 is in the denominator because the interval is two-sided, corresponding to a confidence $1-\alpha/2$ in each end of the interval. When we are only interested in lower bounds, we may take $\alpha' = \alpha/h$. (In such a case, the upper bound of the interval is 1.) However, we wish to allow for the existence of unfound haplotypes. This implies that the actual number of haplotype categories is one more than was found in the study. (The extra category consists of all unsampled haplotypes.) Therefore, for this application, we recommend using $\alpha' = \alpha/(h + 1)$. The implementation of Goodman’s (1965) method in HaploPro (Appendix 1; Electronic suppl.) has two columns, one corresponding to $\alpha' = \alpha/(2h)$, the other corresponding to $\alpha' = \alpha/(h + 1)$ for calculating only lower bounds.

Because we are assuming some haplotypes were not found, $h + 1$ is a better estimate of the number of categories than h in this application; however combining unsampled haplotypes into a single category is of questionable validity because the categories should be defined a priori. The confidence in the lower bounds will be most valid in

cases where one missed haplotype has a proportion much larger than all the other missed haplotypes combined, or where the missed haplotypes can be combined into a single category that could have been defined a priori— e.g., “all haplotypes with proportion less than 0.05%.”

The sum of lower bounds estimates the proportion of the population with haplotypes that were found, but the estimate is very conservative. It is true that if all true proportions are greater than their lower bounds, then their sum must be greater than the sum of the lower bounds; however the converse is not necessarily true. Even when some true proportions are below their lower bounds, other proportions may be far enough above to make up the difference. Thus one’s confidence that the true sum is greater than the lower bound sum is actually much better than the level stated. For this reason, the statistic can be used to show that a study was thorough, but it is less useful for indicating that a study is incomplete. It can, however, be used to compare different studies to say which is more likely to have unobserved variation. It can also be used to compare the samples of different populations within the same study, indicating where further sampling will yield the most information. That information will be in the form of new haplotypes and/or increased confidence in the true proportions of already found haplotypes.

Examples of the application of this statistic are given in Table 5. Clearly, larger sample sizes enable the researcher to more confidently predict haplotype proportions in the

Table 5. This table demonstrates calculating the sum of 90% confidence lower bounds in data from the literature.

Group ^a	Total individuals	Number of haplotypes found	Individuals per haplotype	Lower bound (%)	Sum of lower bounds
<i>Quercus</i> spp. Mátyás & Sperisen (2001: Table 1)	1,036	10	641	58.25	91.49
			372	32.47	
			10	0.46	
			6	0.22	
			2	0.04	
			5 types with 1	0.01	
<i>Quercus</i> spp. Whittemore & Schaal (1991: Table 1)	129	7	67	42.19	68.67
			2 types with 19	9.06	
			12	5.00	
			6	1.93	
			4	1.07	
			2	0.36	
CH148 Mátyás & Sperisen (2001: Table 2)	50	2	31	49.03	75.57
			19	26.54	
CHA Strand & al. (1996: Table 2)	10	3	6	31.26	43.81
			3	10.77	
			1	1.78	
NUL Strand & al. (1996: Table 2)	10	2	9	61.82	63.78
			1	1.96	

^a“Group” gives the name of the studied taxon or population and the reference.

studied group. This statistic indicates that by sampling 1,036 members of three *Quercus* spp. from the Swiss Alps, Mátyás & Sperisen (2001) may have 90% confidence in proportions accounting for over 91% of the studied group. Contrastingly, sampling 10 members of *Aquilegia longissima* A. Gray ex S. Watson population CHA in Mexico gives Strand & al. (1996) 90% confidence in proportions accounting for merely 43% of the population.

This statistic is affected by the presence of clear dominants. Strand & al. (1996) also sampled ten individuals from population NUL. Unlike population CHA, population NUL had only two haplotypes and a strong dominant, so this statistic indicates 90% confidence in proportions accounting for over 63% of the population.

Finding many haplotypes leaves more room for uncertainty. Whittemore & Schaal (1991) sampled over twice as many individuals from five *Quercus* spp. as Mátyás & Sperisen (2001) did from *Q. petraea* population CH148. However, because Whittemore & Schaal (1991) found seven haplotypes, they could have 90% confidence in proportions representing only 68% of the population. Mátyás & Sperisen (2001) found just two haplotypes and could have 90% confidence in proportions representing over 75% of the population.

This statistic can be used to get a conservative upper bound on the size of the group of unfound haplotypes. In the study of Mátyás & Sperisen (2001), 75% of population CH148 is accounted for with 90% confidence. Unfound haplotypes can only occur in the remaining 25%, so they can claim 90% confidence that they missed less than 25% of the population. This does not mean, however, that there is a 10% chance that they missed more than 25% of the population. That 25% accounts for a mixture of found and unfound haplotypes in unknown proportions.

DISCUSSION

Have all haplotypes been found? — This question is possible to answer in the affirmative only in very special situations—e.g., when the entire population is included in the study or when the number of possible haplotypes can be calculated a priori. If the population is large relative to the sample size and the number of possible haplotypes is large, then the researcher must accept that there may be some rare mutations that will not be included in the study. Therefore, we have concerned ourselves with analyzing haplotypes in terms of their proportions.

Instead of trying to find all haplotypes, we recommend trying to find all prevalent haplotypes. The researcher can define “prevalent” a priori and use (Eq. 4) to estimate the probability that the sample included at least one of each, or the researcher can choose a standard probability (such as 95%) and vary k in (Eq. 4) to find the smallest

proportion $1/k$ that is guaranteed to have been found (with 95% percent probability).

Another alternative to trying to find all haplotypes is to try to find enough of the haplotypes so that the unfound haplotypes account for only a small percentage of the population. The sum of lower bounds statistic discussed in the previous section can be used to get an estimate of the percentage of the population that has haplotypes missed by the sample. This estimate is conservative, however, and it does not make very strong statements even for very thorough samples.

Both (Eq. 4) and the sum of lower bounds statistic are measures of thoroughness, useful for comparing the relative coverages of samples in different studies or samples of different populations within the same study. The researcher may wish to calculate both because the two statistics approach the idea of thoroughness differently. The sum of lower bounds statistic is particularly useful as a measure because it takes into account the structure of the sample, whereas the results of (Eq. 4) depend only on the number of individuals sampled. On the other hand, unlike the sum of lower bounds statistic, (Eq. 4) can be used a priori when designing a study.

Caveats. — The statistical methods presented here assume that a true random sample of the professed group under study was conducted. This group may include part of a population or one or several populations, depending on the problem being addressed. With ideal sampling techniques, these methods are appropriate for the case where the group is spread over a large geographic area but is not easy to divide into several distinct populations. If the group consists of several well-defined populations, these techniques should be applied to each population separately. However, the researcher can make valid statements about proportions throughout the entire group, provided each population is represented proportionally in the study. In the *Fagus sylvatica* phylogeographic study by Demesure & al. (1996: 2516), a pilot test was made with 12 individuals from 12 populations “selected to represent as much as possible the whole range of the species.” This justifies applying the statistic (Eq. 1) to it, as was done above. It should be noted, however, that simple representation is not sufficient. Populations should be represented in the sample in proportion to the number of individuals they contain if the statistics are to be used to make valid statements about proportions within the taxon as a whole.

These models assume that specimen collection has no discernable impact on the proportions within the population, i.e., that sampling one individual does not affect the probability of sampling another individual with the same haplotype. This will generally be the case as long as the sample is only a miniscule fraction of the individuals in the group being studied. If this is not the case, these tech-

niques are less appropriate. Note that if every individual in the studied group is included in the sample, most of the questions considered above are irrelevant.

Special considerations for nuclear DNA sequences. — The statistics techniques presented here can, in some cases, be applied to *nr*DNA. It must be noted that one sampled individual actually corresponds to two or more statistical samples of haplotypes. Thus, the sample size n is not the number of sampled individuals, but the sum of the ploidy levels of the sampled individuals, e.g., a hexaploid is effectively a bulk of six potentially different alleles (comparable to organelle haplotypes). Furthermore, the statistics will only be valid to the extent that alleles within individuals occur independently (which is the case for a single, randomly mating population).

If this independence assumption does not hold, the techniques presented may still be applied, except that the unit of study is not an allele but an allele combination. In a group of diploid individuals, for example, instead of calculating confidence intervals for the proportions of alleles A and B, the researcher could calculate confidence intervals for the proportions of genotypes AA, AB, and BB. Note that the sample size n in this case is the same as the number of sampled individuals because each individual is a sample of only one genotype.

In particular, if a study of *nr*DNA reveals only one allele, each sampled individual falls into the category “homozygous for allele A” (regardless of ploidy level), and (Eq. 1) can be used to estimate the size of this category.

Application to other characters. — These methods can be applied to characters other than sequence data, if certain conditions are met.

The population must be sampled randomly with respect to the character. If the character is visible at the population level (petal color for large flowers, ploidy level in some cases), the researcher must be certain that the sample did not favor a particular type.

The character must have discrete values. For non-discrete characters such as pollen diameter, culm height, relative DNA content, or GC/AT proportions, the researcher can create groups, such as 1–2 mm, 2–5 mm, greater than 5 mm. Grouping can also be used to apply these statistics to combinations of characters, such as Haplotype A/less than 1 mm, Haplotype A/greater than 1 mm, Haplotype B/less than 1 mm, etc.

In some cases, the researcher can determine a priori the number of haplotypes. (Here we use “haplotype” as a generic term for character value, character group, or combination of characters.) If this number is small, some of the methods may need to be modified. The techniques of (Eq. 1) and (Eq. 3) for measuring dominance still apply, but when calculating Goodman (1965) confidence intervals (using HaploPro, for example), the researcher may wish to consider using the number of possible haplotypes (h')

in place of the number of haplotypes that were found (h). The researcher is warned, however, that the confidence intervals for the haplotypes that were found 0 times are less reliable than those calculated for haplotypes that were actually found. Goodman’s (1965) method performs well when the number of haplotypes is ten or less, but if there are many values, each with roughly the same number of observations, the researcher may want to use the method of Sison & Glaz (1995) (according to May & Johnson [2000] as stated above).

The two measures of sampling thoroughness do not apply if the researcher finds all the haplotypes, but they can be used if several of the haplotypes are missing. When calculating the sum of lower bounds statistic, the researcher should use $\alpha' = a/h'$ instead of $\alpha' = a/(h+1)$. (To use HaploPro for this calculation, enter one less than the number of possible haplotypes instead of the number of haplotypes found.) The inequality (Eq. 4) still applies, but when the number of possible haplotypes is known, it may be possible to get a tighter estimate. For example, when calculating the probability of sampling every haplotype that makes up at least 5% of the population, (Eq. 4) takes into account the possibility that there could be as many as 20 of them. (See Appendix 2.) If it is known that there are only 4 possible haplotypes, the probability of missing a prevalent one is less than it would be if there were 20 of them. On the other hand, if it is known that there are 30 possible haplotypes, the bound given by (Eq. 4) cannot be improved upon because the possibility exists that the population contains 20 of them, each making up 5% of the population.

CONCLUSIONS

Studies that investigate organelle DNA sequence data at low taxonomic levels present several statistical challenges. Measures for addressing six different questions concerning sampling effectiveness have been proposed in this paper. They include predicting the extent of a single dominant both with and without bulking methods, measuring the thoroughness of a sample, and calculating confidence intervals for the proportions of found haplotypes, with or without an overwhelming dominant. These measures can help researchers evaluate their sampling design, plan future studies, and objectively present their results.

ACKNOWLEDGEMENTS

The authors wish to thank two anonymous reviewers for helpful suggestions on earlier versions of the manuscript. This research was undertaken during the graduate studies of SDSH,

supported by a U.S. National Science Foundation Graduate Research Fellowship, a U.S. Student Fulbright Grant, and an Honor Society of Phi Kappa Phi Fellowship, and was funded by the Ministry of Education of the Czech Republic Research Projects MSM 0021622416 and LC 06073.

LITERATURE CITED

- Agresti, A. & Coull, B.A.** 1998. Approximate is better than “exact” for interval estimation of binomial proportions. *Amer. Statist.* 52: 119–126.
- Brysting, A.K., Fay, M.F., Leitch, I.J. & Aiken, S.G.** 2004. One or more species in the arctic grass genus *Duportia*?—a contribution to the Panarctic Flora project. *Taxon* 53: 365–382.
- Colbert, T., Till, B.J., Tompa, R., Reynolds, S., Steine, M.N., Yeung, A.T., McCallum, C.M., Comai, L. & Henikoff, S.** 2001. High-throughput screening for induced point mutations. *Pl. Physiol.* 126: 480–484.
- Cotton, R.G.H.** 1997. Slowly but surely towards better scanning for mutations. *Trends Genet.* 13: 43–46.
- Demesure, B., Comps, B. & Petit, R.J.** 1996. Chloroplast DNA phylogeography of the common beech (*Fagus sylvatica* L.) in Europe. *Evolution* 50: 2515–2520.
- Dixon, C.J.** 2006. A means of estimating the completeness of haplotype sampling using the Stirling probability distribution. *Molec. Ecol. Notes* 6: 650–652.
- Dumolin-Lapègue, S., Demesure, B., Fineschi, S., Le Corre, V. & Petit, R.J.** 1997. Phylogeographic structure of white oaks throughout the European continent. *Genetics* 146: 1475–1487.
- Dumolin-Lapègue, S., Pemonge, M.-H. & Petit, R.J.** 1998. Association between chloroplast and mitochondrial lineages in oaks. *Molec. Biol. Evol.* 15: 1321–1331.
- Fitzpatrick, S. & Scott, A.** 1987. Quick simultaneous confidence intervals for multinomial proportions. *J. Amer. Statist. Assoc.* 82: 875–878.
- Fujii, N., Ueda, K., Watano, Y. & Shimizu, T.** 1997. Intraspecific sequence variation of chloroplast DNA in *Pedicularis chamissonis* Steven (Scrophulariaceae) and geographic structuring of the Japanese “alpine” plants. *J. Pl. Res.* 110: 195–207.
- Goodman, L.A.** 1965. On simultaneous confidence intervals for multinomial proportions. *Technometrics* 7: 247–254.
- Huang, J., Kirk, B., Favis, R., Soussi, T., Paty, P., Cao, W. & Barany, F.** 2002. An endonuclease/ligase based mutation scanning method especially suited for analysis of neoplastic tissue. *Oncogene* 21: 1909–1921.
- Lessa, E.P. & Applebaum, G.** 1993. Screening techniques for detecting allelic variation in DNA sequences. *Molec. Ecol.* 2: 119–129.
- Mátyás, G. & Sperisen, C.** 2001. Chloroplast DNA polymorphisms provide evidence for postglacial re-colonisation of oaks (*Quercus* spp.) across the Swiss Alps. *Theor. Appl. Genet.* 102: 12–20.
- May, W.L. & Johnson, W.D.** 2000. Constructing two-sided simultaneous confidence intervals for multinomial proportions for small counts in a large number of cells. *J. Statist. Software* 5: 6.
- Milligan, B.G.** 1992. Is organelle DNA strictly maternally inherited? Power analysis of a binomial distribution. *Amer. J. Bot.* 79: 1325–1328.
- Nataraj, A.J., Olivos-Glander, I., Kusukawa, N. & Highsmith, W.E., Jr.** 1999. Single-strand conformation polymorphism and heteroduplex analysis for gel-based mutation detection. *Electrophoresis* 20: 1177–1185.
- Newcombe, R.G.** 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 17: 857–872.
- Nordborg, M. & Innan, H.** 2002. Molecular population genetics. *Curr. Opin. Pl. Biol.* 5: 69–73.
- Posada, D. & Crandall, K.A.** 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16: 37–45.
- Quesenberry, C.P. & Hurst, D.C.** 1964. Large sample simultaneous confidence intervals for multinomial proportions. *Technometrics* 6: 191–195.
- Schaal, B.A. & Olsen, K.M.** 2000. Gene genealogies and population variation in plants. *Proc. Natl. Acad. Sci. U.S.A.* 97: 7024–7029.
- Sison, C.P. & Glaz, J.** 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. *J. Amer. Statist. Assoc.* 90: 366–369.
- Sokal, R.R. & Rohlf, F.J.** 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. 3rd ed. W.H. Freeman and Company, New York.
- Stoneberg Holt, S.D. & Bureš, P.** 2007. Experimental design in intraspecific organelle DNA sequence studies I: haplotype detection methods. *Taxon* 56: 137–144.
- Strand, A.E., Milligan, B.G. & Pruitt, C.M.** 1996. Are populations islands? Analysis of chloroplast DNA variation in *Aquilegia*. *Evolution* 50: 1822–1829.
- Sveshnikov, A.A. (ed.)** 1965. *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions*. English translation by Scripta Technica, Inc.: Gelbaum, B. R. (ed.). 1978. Dover Publications, Inc., New York, New York.
- Taylor, G.R.** 1999. Enzymatic and chemical cleavage methods. *Electrophoresis* 20: 1125–1130.
- Verpy, E., Biasotto, M., Meo, T. & Tosi, M.** 1994. Efficient detection of point mutations on color-coded strands of target DNA. *Proc. Natl. Acad. Sci. U.S.A.* 91: 1873–1877.
- Weisstein, E. W.** 1999a. Inclusion-exclusion principle [Online, last modified Sept. 2, 2003]. Available at <http://mathworld.wolfram.com/Inclusion-ExclusionPrinciple.html> (last accessed July 9, 2007) in: Weisstein, E. W. (ed.). 1995. *MathWorld—A Wolfram Web Resource*. Wolfram Research, Inc., Champaign, IL.
- Weisstein, E. W.** 1999b. Stirling Number of the Second Kind [Online, last modified Dec. 29, 2006]. Available at <http://mathworld.wolfram.com/StirlingNumberoftheSecondKind.html> (last accessed July 9, 2007) in: Weisstein, E. W. (ed.). 1995. *MathWorld—A Wolfram Web Resource*. Wolfram Research, Inc., Champaign, IL.
- Whittemore, A.T. & Schaal, B.A.** 1991. Interspecific gene flow in sympatric oaks. *Proc. Natl. Acad. Sci. U.S.A.* 88: 2540–2544.

APPENDIX 1

HaploPro, available as an Electronic supplement and at <http://www.sci.muni.cz/botany/e/zdroje/haplo.pro> for download, is a simple Microsoft® Excel 2000 spreadsheet of 59 kb that calculates (Eq. 4) and Goodman (1965) confidence intervals.

The first worksheet of HaploPro, “Found Haplotypes”, calculates (Eq. 4). The user enters the minimum proportion a haplotype must have to be considered prevalent. (This minimum proportion must be at least 1%.) Because (Eq. 4) is designed for proportions of the form $1/k$, the spreadsheet calculates the smallest integer k such that $1/k$ is less than or equal to the specified proportion. (If the specified proportion cannot be expressed as $1/k$ or if it is less than 1%, a message appears indicating that the calculation actually applies to a proportion different from that which was entered.) When sample size is entered, the probability that all haplotypes of proportion $1/k$ and greater are represented in the sample (and inversely the probability that haplotypes of that proportion have been

missed) is displayed. Only proportion and sample size can be entered, but specific significance levels can be quickly found by varying sample size.

Confidence interval calculations are performed on the second sheet of HaploPro, “Confidence Intervals”. Here the user enters the number of haplotypes found, the number of individuals in the sample, the number of individuals with the haplotype of interest, and the desired confidence level. The spreadsheet calculates the proportion of the sample with the given haplotype and upper and lower bounds for the proportion of the studied group with the given haplotype. Confidence intervals must be determined for each haplotype individually. To use this worksheet to calculate a score confidence interval for a single proportion (Eq. 3), set the number of haplotypes to 1. When calculating the sum of the lower bounds statistic, the number in the Lower Bound box should be used. The calculations are identical to those in the Confidence Interval box except that $\alpha' = \alpha/(h + 1)$ instead of $\alpha/2h$ is used when calculating $z_{\alpha'}$.

APPENDIX 2

In this appendix, we discuss the inequality (Eq. 4) and explain why it is the correct solution to the problem of estimating the probability of sampling every sufficiently prevalent haplotype.

Suppose we are interested in every haplotype that makes up at least 5% of the population. If we sample $n = 45$ individuals, the probability that we fail to sample a particular haplotype of size 5% is $(1 - 0.05)^{45} = 9.94\%$. For any one particular haplotype of size 5%, we can be 90% certain that we have sampled it. However, there may be several haplotypes with this proportion. Suppose there are two such haplotypes (and many others, each with proportion less than 5%). We can calculate the probability of failing to sample one of them using the Addition Principle: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. (See, for example Sokal & Rohlf, 1995: 65.)

The probability of missing at least one of them is given by $P(\text{missing haplotype } A \text{ or } B) = P(\text{missing haplotype } A) + P(\text{missing haplotype } B) - P(\text{missing both}) = 9.94\% + 9.94\% - (1 - 0.10)^{45} = 19.01\%$. The third term must be subtracted because it is counted twice: the events “missing haplotype A” and “missing haplotype B” both include the event “missing both A and B.”

If there are three such haplotypes, the probability of missing one of them is given by $P(\text{missing } A, B, \text{ or } C) = P(\text{missing } A) + P(\text{missing } B) + P(\text{missing } C) - P(\text{missing } A \text{ and } B) - P(\text{missing } B \text{ and } C) - P(\text{missing } C \text{ and } A) + P(\text{missing } A, B, \text{ and } C)$. Note that the event “missing A, B, and C” is added in each of the first three terms (e.g.,

“missing A” includes “missing A, B, and C”) and is subtracted in each of the second three terms (e.g., “missing A and B” includes “missing A, B, and C”), thus the $P(\text{missing } A, B, \text{ and } C)$ term must be added at the end. When there are three haplotypes, each with proportion 5%, the probability of missing one of them is $3 \times (1 - 0.05)^{45} - 3 \times (1 - 0.10)^{45} + (1 - 0.15)^{45} = 27.28\%$.

Note that if the proportion of one of the prevalent haplotypes is greater than 5%, then it is less likely to be missed. Thus, if there is only one haplotype with proportion *at least* 5%, the probability of missing it is at most 9.94%. If there are only two haplotypes with proportion *at least* 5%, the probability of missing one of them is at most 19.01%. If there are only three haplotypes with proportion *at least* 5%, the probability of missing one of them is at most 27.28%.

As the number of prevalent haplotypes increases, the probability of finding all of them decreases. This is not just a property of this example. It is true in general: the more haplotypes there are of this proportion, the greater the probability that one of them will be missed. Thus the worst case is when there are as many of them as possible. In this example, there cannot be more than 20 haplotypes that have a proportion of at least 5%.

To calculate the probability of missing one of k prevalent haplotypes we use a generalization of the Addition Principle, the Inclusion-Exclusion Principle:

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i) - \sum_{i,j:1 \leq i < j \leq k} P(A_i \cap A_j) + \sum_{i,j,l:1 \leq i < j < l \leq k} P(A_i \cap A_j \cap A_l) - \dots + (-1)^{k+1} P\left(\bigcap_{i=1}^k A_i\right).$$

(See, e.g., Sveshnikov, 1965: 16; Weisstein, 1999a.) This equation says that the probability of one of k events occurring can be calculated by summing the probabilities of each event, then subtracting the cases in which two events occur simultaneously (because they have been counted twice) then adding the cases in which three events occur simultaneously (because they were counted multiple times in the previous subtraction) and so on.

When calculating the probability of missing a haplotype of proportion at least $1/k$, the worst case is when there are k haplotypes and each is present in the studied group with proportion exactly $1/k$. If we sample n individuals, what is the probability that we will find a representative of each haplotype?

The probability of missing a given haplotype is $(1 - 1/k)^n$. If we group i haplotypes together, the probability of missing all of them is $(1 - i/k)^n$. For each i , the number of such groups is given by the binomial coefficient:

$$\binom{k}{i} = \frac{k!}{i!(k-i)!}$$

(See, e.g., Sokal & Rohlf, 1995: 72.) Thus, each of the summands in the Inclusion-Exclusion Principle can be replaced by the expression

$$\binom{k}{i} \left(1 - \frac{i}{k}\right)^n$$

Because the summands are alternately added and subtracted, we multiply each term by $(-1)^{i+1}$ and we have: $P(\text{missing at least one of } k \text{ haplotypes})$

$$= \sum_{i=1}^{k-1} (-1)^{i+1} \binom{k}{i} \left(1 - \frac{i}{k}\right)^n$$

(Note that the sum could go all the way to k , but the last term would include the factor $(1 - k/k)$: the probability of not sampling any of the k haplotypes is zero.) We conclude that

$$\begin{aligned} &P(\text{sampling each haplotype at least once}) \\ &= 1 - P(\text{missing at least one haplotype}) \\ &= 1 + \sum_{i=1}^{k-1} (-1)^i \binom{k}{i} \left(1 - \frac{i}{k}\right)^n \end{aligned} \quad (\text{Eq. 5})$$

(This problem and its solution can also be found in Sveshnikov [1965, Problem 5.34, pp. 22, 382], where it is formulated as a group of passengers boarding train cars.)

The inequality (Eq. 4) can be used a priori with no assumptions made about the true proportions within the population because the assumptions used to derive (Eq. 5) are the worst case. Haplotypes with proportions greater than $1/k$ are more likely to be sampled than those with proportions equal to $1/k$, so the probability of sampling all prevalent haplotypes is lowest when every prevalent haplotype has proportion $1/k$. There cannot be more than k haplotypes with proportion $1/k$ (or greater), and if there are fewer, then the probability of sampling them all is greater. Therefore, the inequality (Eq. 4) gives a lower bound on the probability of sampling at least one of every haplotype that has a proportion of at least $1/k$.

The inequality (Eq. 4) only gives information about prevalent haplotypes. There may be many haplotypes with proportion less than $1/k$ and the sum of their proportions may exceed $1/k$.

It should be noted that

$$1 + \sum_{i=1}^{k-1} (-1)^i \binom{k}{i} \left(1 - \frac{i}{k}\right)^n = \frac{k!}{k^n} S_{n,k}$$

where $S_{n,k}$ is a Stirling number of the second kind. (See, e.g., Weisstein, 1999b.) In Dixon (2006),

$$\frac{k!}{(k-x)!k^n} S_{n,x}$$

is used to estimate the number of haplotypes (k) in a population, based on the number of haplotypes found (x) and the sample size (n), under the assumption that all haplotypes in the population are of roughly equal proportion.

A version of this manuscript, before review and revision, appeared in Stoneberg Holt, S.D. 2004. *The trnL-F plastid DNA region and its application to phylogeographic analysis in Poa pratensis agg.* PhD thesis. Department of Botany, Masaryk University, Brno.