

Estimating the number of founder lineages from haplotypes of closely linked SNPs

RAPHAËL LEBLOIS* and MONTGOMERY SLATKIN

Department of Integrative Biology, University of California at Berkeley, Berkeley, CA 9472-3140, USA

Abstract

We consider an isolated population founded by a small number of individuals randomly chosen from a source population of known genetic composition at a known time in the past. We develop a Monte-Carlo maximum-likelihood method for estimating the number of founding individuals from the haplotype frequencies at several SNP (single nucleotide polymorphism) loci in a sample. We assume the isolated population was founded recently enough that that mutation can be ignored and that haplotype frequencies in the source population have not changed. We apply the method to simulated data and show that it is unbiased. With a reasonable number of individuals sampled, it is possible to estimate the number of founders within a factor of 2. We show that the performance of the method is not degraded substantially if the frequencies of the rare haplotypes in the source are not known precisely and if there is some recombination. We illustrate the use of our method by applying it to a previously published data set from a recently founded population of wolves (*Canis lupus*) in Scandinavia.

Keywords: coalescent theory, founder effect, gene genealogy, linkage disequilibrium, maximum-likelihood estimation

Received 9 December 2006; revision received 13 December 2006; accepted 8 January 2007

Introduction

The establishment of a new population by a small number of founding individuals can result in substantial differences in allele and haplotype frequencies between the newly founded population and the source of founders. This episode of genetic evolution is called a founder effect. In evolutionary biology, founder effects have been suggested to cause rapid morphological evolution (Mayr 1954), evolution across 'adaptive valleys' (Wright 1932), and evolution of reproductive isolation (Carson & Templeton 1984). Furthermore, founder effects are thought to be responsible for the unusually high incidence of some Mendelian genetic diseases in isolated human populations (Vogel & Motulsky 1996) and to have created conditions suitable for linkage disequilibrium mapping (Hästbacka *et al.* 1992).

Although founder effects are frequently called upon as agents of rapid genetic change, most evidence for founder effects has been circumstantial, except in rare instances when historical records indicate the number of founders. In some cases, the circumstantial evidence is very strong. Populations of plants and animals on remote islands and archipelagos such as the Galapagos were almost certainly founded by small numbers of individuals; there is no plausible way for large numbers of founders to have gotten there. In other cases, there is less agreement. Although the cichlid fishes of the African Rift Valley lakes have undergone rapid morphological divergence and speciation, founder effects may not have played an important role (Meyer *et al.* 1990). Individuals of Ashkenazi Jewish ancestry have numerous monogenic diseases in relatively high frequency, but there is disagreement over whether those high frequencies are the result of selection or founder effects (Slatkin 2004).

Because of the potential importance of founder effects, it would be desirable to have a method for estimating the number of founders of an isolated population. That would allow quantitative assessment of the intensity of a founder event. In this study, we introduce a method that is applicable

Correspondence: Montgomery Slatkin, Fax: 510-643-6264; E-mail: slatkin@berkeley.edu

*Present address: Unité Origine, Structure et Evolution de la Biodiversité, Département Systématique et Evolution, Muséum National d'Histoire Naturelle, 16 Rue Buffon, 75005 Paris, France

to data at closely linked single nucleotide polymorphism (SNP) loci. We assume that an isolated population was founded at a known time in the past by individuals from a source population that is still extant. For convenience, we refer to the isolated population as the island population. We assume that individuals from both the island and source populations have been typed at several closely linked SNP loci and that haplotype frequencies in both samples have been estimated. The problem is to estimate the number of founder individuals from the genetic data.

Our method is similar to one developed by Anderson & Slatkin (2007). The difference is that Anderson and Slatkin assumed that data are available from a set of unlinked multi-allelic loci. Their method is appropriate for the analysis of data from unlinked allozyme and microsatellite loci, for which it is reasonable to assume that all or nearly all alleles in the source population have been identified. The method presented here focuses on haplotypes at linked SNP loci, for which the number of haplotypes may be so large that many of the rarer haplotypes will not be found in a sample from the source population. Furthermore, low levels of recombination can create haplotypes found in the isolated population that are not present in the source population. To account for both of these possibilities, we group low frequency haplotypes into a single category and show that making this approximation in the presence of realistic amounts of recombination does not degrade the performance of our method.

Other theoretical studies have modelled the effects of bottlenecks in population size and founder events. Vincek *et al.* (1997) estimated the number of founders of the Darwin's finches on the Galapagos Islands by analysing variation at a major histocompatibility (*Mhc*) locus. Vincek *et al.* simulated the effect of a founder on the number of overdominant alleles in an isolated population and compared their results with observations of the number of *Mhc* alleles currently found in Darwin's finches. Because the rate of loss of overdominant alleles is very small, their analysis is appropriate for populations that were founded in the distant past. Vincek *et al.* (1997) did not have samples from the potential source population so instead assumed that the pattern of variation in the source was similar to patterns of variation in class II *Mhc* loci in humans. They assumed, as we will, that no new mutations occurred since the founder event. They inferred that the founder population for the Galapagos finches contained at least 30 individuals.

Gaggiotti *et al.* (2004) considered a slightly different problem than we do. They used multilocus genetic data to estimate the relative contributions of different source populations to an island population and to evaluate the relative importance of other factors, such as geographical distance and size of the source population, in affecting the contributions

of each source population. Gaggiotti *et al.* (2004) used simulated data to show that their method performed well under a wide variety of assumptions. Foll & Gaggiotti (2005) described an improved program, COLONISE, that implements this method.

A closely related problem is finding genetic evidence for a past bottleneck in population size. This problem is more difficult because the population before the bottleneck, which corresponds to the source population, is not available for study. The pattern of pairwise differences in the DNA sequences of nonrecombining genes such as mitochondrial DNA can indicate multiple episodes of population growth, which can be interpreted as evidence of one or more bottlenecks (Rogers & Harpending 1992). Several studies have shown that allele frequencies following a bottleneck will tend to be more even than expected in a population of constant size (Nei *et al.* 1975; Watterson 1984; Maruyama & Fuerst 1985; Cornuet & Luikart 1996; Luikart & Cornuet 1998). Cornuet & Luikart (1996) and Luikart & Cornuet (1998) evaluated the power of two tests for past bottlenecks based on difference between the observed heterozygosity and the heterozygosity expected for the observed number of alleles. They found that these tests have some power but only if the bottleneck was quite recent. These tests are unable to estimate the population size during the bottleneck, which corresponds to the number of founders in our analysis.

Approximate Bayesian computation (ABC) methods (Beaumont *et al.* 2002; Chan *et al.* 2006) have been used to estimate the numbers of founders. Estoup *et al.* (2001) estimated founder population sizes and other parameters of a complex demographic model of several island populations of cane toads, and Estoup & Clegg (2003) estimated founder and current population sizes of the bird species *Zosterops lateralis* in New Zealand, Tasmania and nearby islands. In both cases, as well as in the recent study by Chan *et al.* (2006), prior information was needed for the analysis.

Hey (2005) adapted the program IM (Nielsen & Wakeley 2001) to estimate of the number of founders of the Native American populations. IM implements a Bayesian method that simultaneously estimates the time of divergence of populations, the sizes at the time of divergence, the extent of gene flow since divergence, and mutation rates. Hey (2005) estimated the founder size of Native Americans to be approximately 80 individuals. He performed extensive simulations showing that the method performed well on data simulated under a model embodying essential features of the history.

Our analysis is similar to that of Estoup *et al.* (2001), Estoup & Clegg (2003), and Hey (2005) but is based on a much simpler model. We assume that the history of the island population and its extent of isolation are known and that only the founder size is to be inferred. By focusing

on a more narrowly defined problem, we can examine the ability to infer a population genetic parameter under the best possible conditions. The method was tested on an SNP data set from a recently founded population of wolves in Scandinavia.

Method

The problem is to estimate the number of founding chromosomes for an island population established T generations in the past. The island population was established by a group of individuals randomly sampled from a single source population which still exists and is available for study. The time T is assumed to be small enough that mutation is not important, and the source population is assumed to be large enough that the allele and haplotype frequencies observed today represent the frequencies when the island population was established. In other words, we allow for genetic drift acting between foundation event and current time on the island but not in the source.

Exact method

We assume that haplotypes at L loci can be determined in a sample from the isolated population. Our focus will be on single nucleotide polymorphisms (SNPs), which are diallelic; the total number of possible haplotypes is $k = 2^L$. We begin by assuming no recombination, in which case our model is equivalent to a model with k alleles at a single locus. With that restriction, the theory is equivalent to that of Anderson & Slatkin (2007). The data consists of the numbers of each of the k distinct haplotypes in a sample from the island population today ($t = 0$), represented as a vector, $\mathbf{n} = [n_1, \dots, n_k]$. The number of chromosomes sampled is $n = \sum_{i=1}^k n_i$. The haplotype frequencies in the source population are $\mathbf{p} = [p_1, \dots, p_k]$. Although those frequencies have to be estimated from other data, we assume initially they are known without error; we discuss later the effects of error in estimates of \mathbf{p} . Part of the goal of our analysis is to establish whether it is possible in principle to estimate the number of founders in the highly idealized model we examine. Then we can consider how good an estimate can be obtained when various complicating factors are taken into account.

In the exact method, we assume all elements of \mathbf{n} and \mathbf{p} are known. An isolated population is founded by $m/2$ diploid individuals T generations in the past and has grown according to a function $N(t)$ between $t = 0$ (the present) and $t = T[N(T) = m]$. We assume exponential growth at rate r : $N(t) = me^{r(T-t)}/2$. Given r and m , the current population size is determined, $N(0) = me^{rT}/2$. The sample of n chromosomes is drawn from the island population at $t = 0$. At a locus, those n chromosomes are descended from j ancestors at time T in the past when the population was founded;

j is a random variable which has a known distribution, given by equation 6.1 of Tavaré (1984) with t replaced by time rescaled to account for population growth

$$\tau = \int_0^T \frac{dt'}{2N(t')} \tag{eqn 1}$$

(Griffiths & Tavaré 1994):

$$\Pr(j|n,r) = \sum_{i=j}^n \frac{(2i-1)(-1)^{i-j} j_{(i-1)} n_{[i]}}{j!(i-j)!n_{(i)}} e^{-i(i-1)\tau/2} \tag{eqn 2}$$

where $a_{(i)} = a(a+1) \dots (a+i-1)$ and $a_{[i]} = a(a-1) \dots (a-i+1)$. The parameter r is added to the left hand side to emphasize the dependence on it. All the dependence on r is through its effect on τ , which can itself be regarded as the free parameter of the model. Equation 2 assumes that only one coalescent event occurs at a time, which is equivalent to assuming a diffusion approximation for allele frequency evolution on the island. This equation represents the coalescent process with drift driven by the exponential growth rate in the island population between foundation event and time of sampling.

Given j , the allelic composition at the locus, the configuration of the sample, $\mathbf{j} = \{j_1, \dots, j_k\}$, has a multinomial distribution with frequencies $\mathbf{p} = \{p_1, \dots, p_k\}$ and sample size j :

$$\Pr(\mathbf{j} | \mathbf{p}) = \left(\frac{j!}{j_1! \dots j_k!} \right) p_1^{j_1} \dots p_k^{j_k}$$

To find the probability of data, $\mathbf{n} = \{n_1, \dots, n_k\}$ given \mathbf{j} , we use the fact that under the assumptions of the neutral coalescent (Kingman 1982), all configurations of the descendants are equally probable, meaning

$$\Pr(\mathbf{n} | \mathbf{j}) = \frac{\prod_{l=1}^k \binom{n_l - 1}{j_k - 1}}{\binom{n - 1}{j - 1}}$$

To find the overall probability of the data, we sum over all j and all configurations of alleles in the founders:

$$\Pr(\mathbf{n} | r) = \sum_{j=2}^m \Pr(j | r) \sum_{j_1 + \dots + j_k = j} \Pr(\mathbf{j} | \mathbf{p}) \Pr(\mathbf{n} | \mathbf{j})$$

This probability is the likelihood of r , as a function of the data.

For small j , this sum can be evaluated exactly. For larger j , the exact calculation is impractical, but the sum can be approximated by using the Monte Carlo procedure. To approximate the inner sum for each j , we generated replicate sets of $\{j_1, \dots, j_k\}$ from the appropriate multinomial distribution, and then took the numerical average over

replicates. We then summed over j . If data are available for independent loci, the likelihoods computed for each locus are multiplied. From the likelihood, the maximum likelihood estimate (MLE) of r is found. In the analysis of simulated data, we present results in terms of m under the assumption that the current population size is known.

Approximate method

The exact method breaks down if there are a large number of possible haplotypes because the elements of the vector $\mathbf{p} = \{p_1, \dots, p_k\}$ cannot be estimated accurately. It is quite likely that when a sample from the source population is taken for the purpose of estimating $\{p_1, \dots, p_k\}$, some rare haplotypes that are present in the island population will not be found in the sample from the source population. In that case, the exact method fails because the likelihood of r is 0 if a haplotype found on the island has a 0 frequency in the source.

To avoid this problem, the approximate method uses only those haplotypes that are relatively common in the source population. We choose a threshold value of haplotype frequency, Q , and keep track of only those haplotypes for which $p_i = Q$. We assume there are K such haplotypes and that the haplotypes are numbered so that $p_i \geq Q$ for $1 = i = K$ and $p_i < Q$ for $i > K$. The value of Q is arbitrary and can be adjusted to suit the available data. All other haplotypes are put in a single category, for which the net frequency is $p_0 = \sum_{i>K} p_i$. The approximate method is the same as the exact method described above with the difference that the low frequency haplotypes are treated as a single haplotype regardless of their identity. In the equations for exact method, k is set to $K + 1$ and p_k is set to p_0 .

Simulation tests

We carried out a series of simulation tests of the performance of the two methods described above. We assumed that source population was of constant size and adjusted the parameters so that a specified number of haplotypes were present in source population. Haplotype frequencies in the source population were simulated using the program *ms* (Hudson 2002). We used a very large sample (10 000 chromosomes) in order to have a good indication of the haplotypes frequencies in the entire population. For each simulated data set, we computed the number of SNPs, the number of haplotypes and the frequencies of each haplotypes in the sample. The different haplotypes of each simulated data set and their frequencies were then assigned to a given 'class' depending on the number of SNPs and the number of haplotypes (for example, eight SNPs and nine haplotypes). By running *ms* for a large number of independent loci, we could thus obtain a specified number of replicates with the right number of SNPs and haplotypes. The parameter

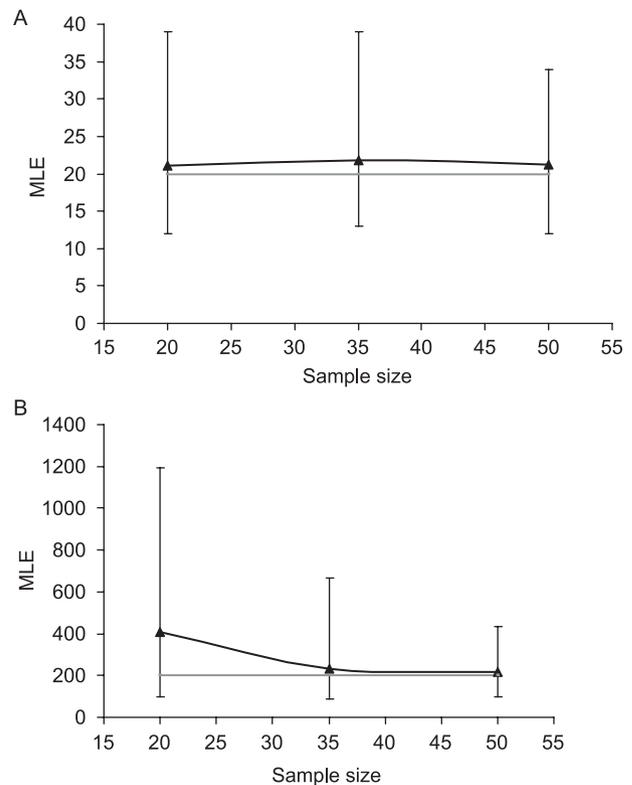


Fig. 1 Effects of varying the sample size. Results were obtained by combining information from 10 loci in each of 100 replicate data sets. In part A, $m = 20$; in part B, $m = 200$.

$\theta (= 4N\mu)$ was set to 0.2, 0.8 and 1.0 for 4, 8 and 12 SNPs, respectively, in order to maximize the number of samples in the target class. For each set of parameter values, we chose randomly 100 samples in a given class and used them to estimate the frequencies of the source population. For each set of parameter values, we estimated m for each of 100 replicate data sets. The figures show the mean estimate and approximate 95% confidence intervals obtained by dropping the lowest and highest two estimates (of 100).

Figures 1, 2 and 3 show the performance of the exact method on simulated data. In all cases, the final population size is 2×10^6 . In all figures the true value is indicated a grey line and the results from analysing the simulations are connected by black lines. Figure 1 shows the effect of varying the sample size (n). For smaller m (20), there is little effect of increasing sample size. For larger m the confidence interval becomes smaller for larger sample size. In both cases, the estimator appears to be nearly unbiased and the extent of bias is only weakly dependent on sample size. Figure 2 shows the effect of varying the number of independent loci. In these cases and others not shown, increasing sample size reduces the range of estimates obtained. Figure 3 shows the effect of varying the number of haplotypes in the source population. The performance improves as the number of haplotypes increases.

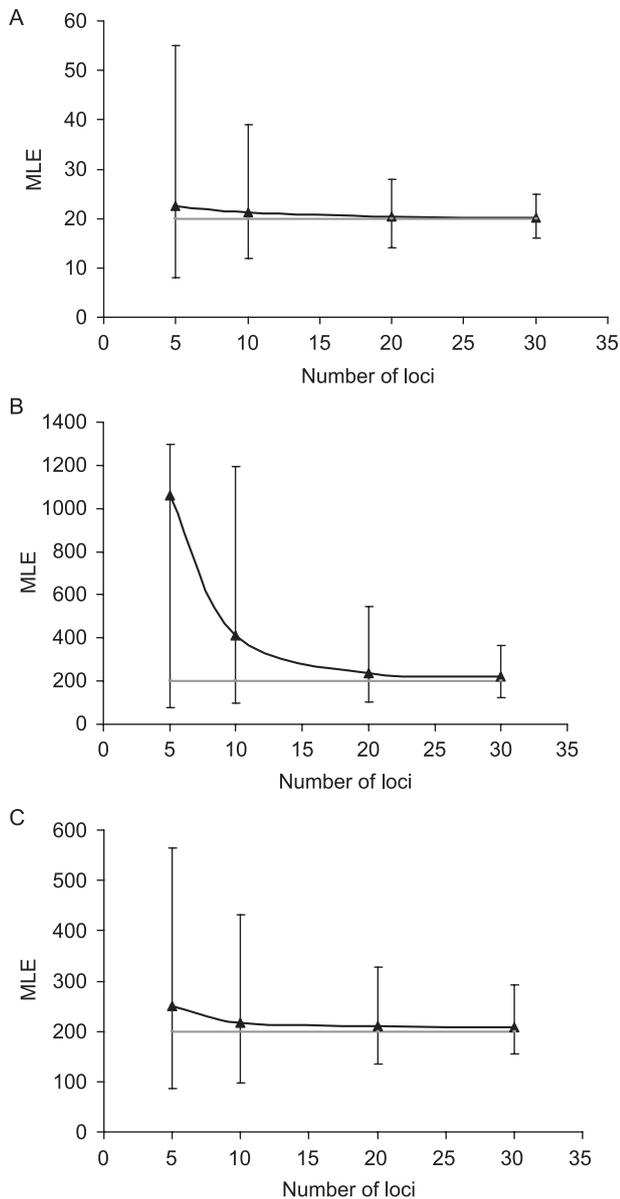


Fig. 2 Effects of varying the number of loci. In part A, $m = 20$; in part B, $m = 200$ and $n = 20$; in part C, $m = 20$ and $n = 50$.

Figures 4 and 5 show the performance of the approximate method. Figure 4 shows the effect of varying Q , the threshold values of haplotype frequencies. The case with $Q = 0$ corresponds to the exact method. For the smaller value of m , estimates of m are surprisingly insensitive to variation in Q . The approximate method remains nearly unbiased. For larger m , however, the bias increases substantially with Q . For $Q = 0.2$, the average estimate of m is approximately half the true value.

Figure 5 shows the effect of applying the approximate method (with $Q = 0.05$) on data simulated with recombination. The eight loci were assumed to be evenly spaced on the chromosome; the recombination rate c is frequency of

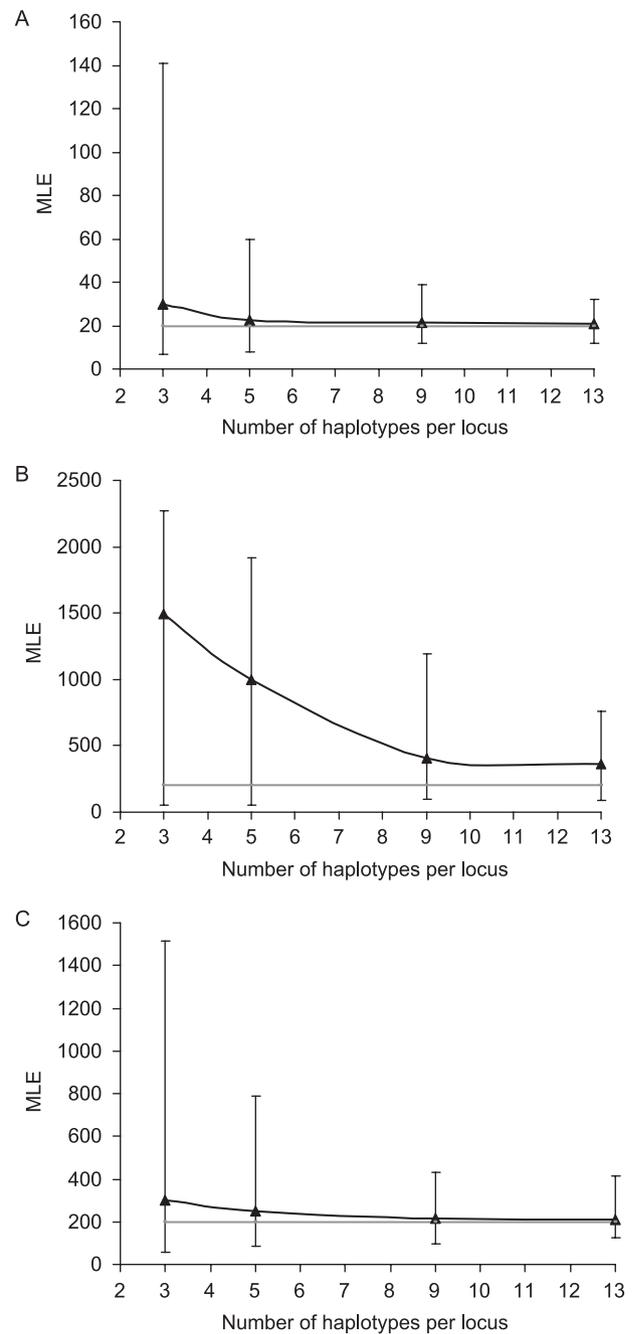


Fig. 3 Effects of varying the number of haplotypes. In part A, $m = 20$; in part B, $m = 200$ and $n = 20$; in part C, $m = 20$ and $n = 50$.

recombination between the SNPs at each end. Recombination has a relatively small effect on the performance of the approximate method under the conditions we considered. The reason is that, because the isolated population was assumed to have been founded recently, $T = 100$, recombination did not have time to erode the linkage disequilibrium created at the time of the founder event and alter substantially the frequencies of the common haplotypes.

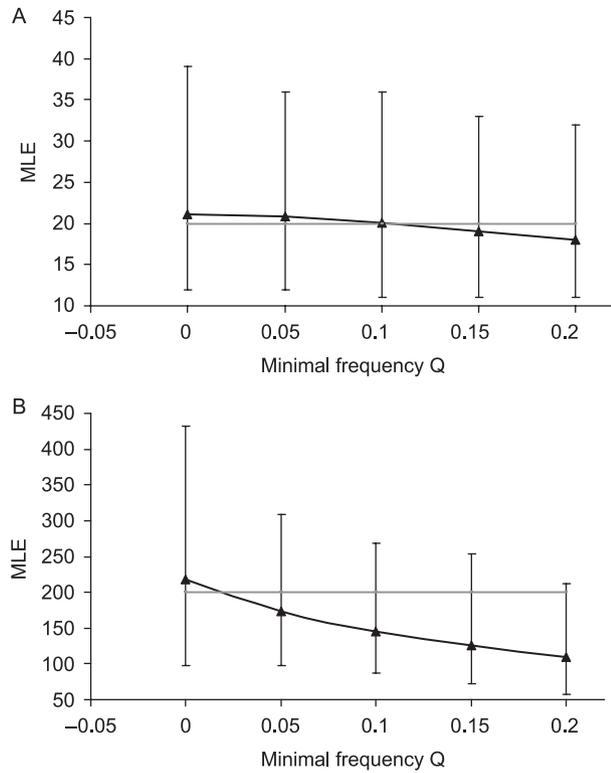


Fig. 4 Approximate method in the absence of recombination. In part A, $m = 20$ and $n = 20$; in part B, $m = 200$ and $n = 50$.

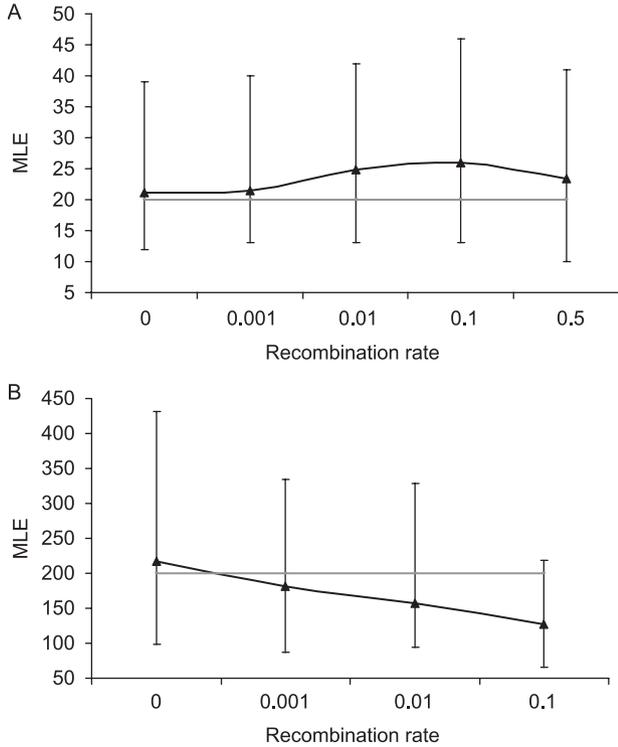


Fig. 5 Approximate method in the presence of recombination. In both parts, $Q = 0.05$. In part A, $m = 20$ and $n = 20$; in part B, $m = 200$ and $n = 50$.

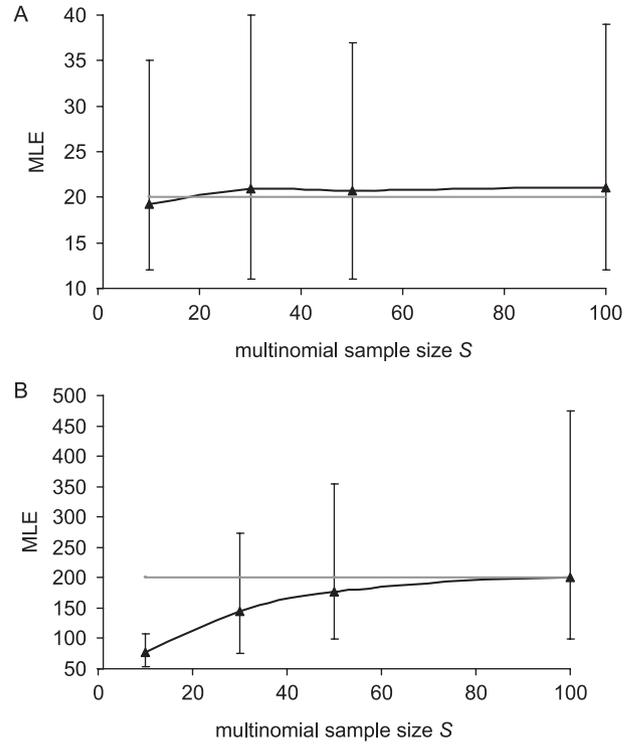


Fig. 6 Effects of estimating haplotypes frequencies using a multinomial sample of size S . In both parts, $Q = 0.05$. In part A, $m = 20$ and $n = 20$; in part B, $m = 200$ and $n = 50$.

The maximum recombination rate, 0.5, corresponds to a rate of $0.5/7 \sim 0.071$ between adjacent sites. Recombination will of course create new rare haplotypes but those are not distinguished when implementing the approximate method.

Errors in estimating haplotype frequencies

Haplotype frequencies in the source population, the elements of \mathbf{p} , are not known exactly; they are estimated from samples from the source population. To assess the effects of errors in the estimates of \mathbf{p} and also to indicate the effects of changes in \mathbf{p} after the population was founded, we modified our method slightly. For a given data set, \mathbf{n} , and set of source frequencies, \mathbf{p} , we drew a multinomial sample of size S . In the sample, the haplotype frequencies are indicated by the vector \mathbf{p}' . In implementing our exact method, we used \mathbf{p}' instead of \mathbf{p} . Figure 6 shows the effect of varying S .

Application

To illustrate the use of our method, we analysed a data set published by Seddon *et al.* (2005) who studied populations of Scandinavian and Finnish wolves (*Canis lupus*). After a dramatic decline, the Scandinavian wolf population was

considered extinct by the 1960s (Wabakken *et al.* 2001). However, in 1983, a breeding wolf pack, for which a micro-satellite study indicated a Finnish/Russian origin, was unexpectedly discovered in southern Scandinavia. This newly established population started to grow exponentially in the early 1990s after the arrival of a male immigrant from the same Finnish population (Wabakken *et al.* 2001; Vila *et al.* 2003).

Seddon *et al.* (2005) typed 24 unlinked SNPs from the Scandinavian and Finnish populations. We reanalysed their data set to determine whether our results are consistent with the known history of the Scandinavian population. We used the following parameter values. Wabakken & Rønning (2003) estimated the total population size to be 76–96 individuals in the winter 2002/2003. Assuming that the ratio of effective to census population size (N_e/N_c) is between 0.26 and 0.42, we assumed that the current effective size is estimated 20–40 individuals (Vila *et al.* 2003; Aspi *et al.* 2006; Liberg 2006). Finally, based on a generation time of 3–6 years (Aspi *et al.* 2006; Mech & Seal 1987; Liberg 2006), we assumed the Scandinavian population was founded four to six generations ago.

Seddon *et al.* (2005) genotyped 108 wolves that had been sampled between 1983 and 2003. However, as our method assumes only a single sample, we reanalysed data from each sampling year between 1998 and 2003 separately. The results are presented in Table 1. Our method indicates that the number of founders for this wolf population was small (four to eight individuals), as is consistent with the historical record. The approximate 95% confidence intervals, based on the log-likelihood curve, are large enough that there may indeed have been only two founders of the Scandinavian population.

This SNP data set is not optimal because each locus is composed of only one SNP. As a consequence, variability at each locus is low and the full power of our method cannot be utilized. Also, the sample sizes are small and the model on which our method is based does not account for immigration after the founding event. Nevertheless, our results are approximately consistent with the known historical record.

Discussion and conclusions

Our results show that it is possible to estimate the number of founders of an isolated population by examining variation in haplotype frequencies in that population and the source of founders. Although the conditions we have assumed are restrictive, namely that both the source of founders and the time at which the population was founded are known, those conditions are met in some situations. Many new populations have been established or re-established both accidentally and intentionally. Even when the historical record indicates the number of founding individuals, as in

Table 1 Application of the method described in the text to the data set of Seddon *et al.* (2005). MLE indicates the maximum likelihood estimate of the number of founders; 95% CI is 95% support interval of the MLE, defined to be the largest and smallest values of the number of founders for which the log-likelihood of the number of founders decreases by 2

Sampling year	Sample size	Current population size	Generations since establishment	MLE	95% CI
1998	8	20	4	4	[2–9]
		40	4	5	[3–8]
		20	6	6	[2–17]
1999	6	20	4	7	[2–19]
		40	4	5	[2–16]
		20	6	15	[4–20]
2000	18	20	4	6	[3–11]
		40	4	7	[5–10]
		20	6	10	[3–17]
2001	26	30	4	7	[5–10]
		40	4	8	[5–10]
		30	6	7	[3–15]
2002	21	30	4	7	[4–11]
		40	4	7	[5–11]
		30	6	12	[4–17]
2003	9	20	4	6	[3–15]
		40	4	5	[3–11]
		20	6	13	[4–18]

the example of the Scandinavian wolf population, the analysis of genetic variation in those populations can either confirm what is known or imply that additional, previously unknown founder individuals have contributed to the population of interest.

Although our method estimates the number of founding lineages when explicit assumptions are made about the time of founding and subsequent population growth, it actually estimates the net extent of genetic drift since the population was founded, which is summarized by the derived quantity τ , defined in equation 1. For some purposes, it is better to estimate τ directly, thus dispensing with the need to make explicit assumptions about the history of the isolated population. That would be appropriate, for example, when testing for selection on one locus in the isolated population.

Our results show that under the conditions we considered, it is possible to estimate the number of founders of an isolated population with some confidence. We recognize, however, that the assumptions underlying our analysis are quite restrictive. The time of founding, the subsequent history of growth, and the identity and composition of the source population are assumed known. Under these conditions, it is possible to isolate and quantify the net effect of genetic drift because genetic drift is the principal source of randomness. That is not likely to be true under

all conditions. In other cases, genetic drift may have effects that are of the same order of magnitude as the uncertainty in other components of the model, thus making it very difficult to separately quantify the effect of genetic drift and accurately estimate the number of founders. Part of the purpose of a model like ours is to make explicit the various sources of randomness and their role in affecting estimates of parameter values.

Electronic resources

The source code of a C program that implements the method described in this study and the data files that were analysed to produce the results in Table 1 are available at <http://ib.berkeley.edu/labs/slatkin/software.html>.

Acknowledgements

We thank E. C. Anderson and B. Rannala for helpful comments on this project. This research was supported by a research grant from the US National Institutes of Health (R01-GM40282) to M. S. and a Lavoisier fellowship from the French Ministry of Foreign Affairs to R. L.

References

- Anderson EA, and Slatkin M (2007) Estimation of the number of individuals founding colonized populations. *Evolution*, in press.
- Aspi J, Roininen E, Ruokonen M, Kojola I, Vila C (2006) Genetic diversity, population structure, effective population size and demographic history of the Finnish wolf population. *Molecular Ecology*, **15**, 1561–1576.
- Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.
- Carson HL, Templeton AR (1984) Genetic revolutions in relation to speciation phenomena: the founding of new populations. *Annual Review of Ecology and Systematics*, **15**, 97–131.
- Chan YL, Anderson CNK, Hadly EA (2006) Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA. *Public Library of Science Genetics* **2**, e59.
- Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics*, **144**, 2001–2014.
- Estoup A, Clegg SM (2003) Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. *Molecular Ecology*, **12**, 657–674.
- Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C (2001) Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics*, **159**, 1671–1687.
- Foll M, Gaggiotti OE (2005) COLONISE: a computer program to study colonization processes in metapopulations. *Molecular Ecology Notes*, **5**, 705–707.
- Gaggiotti OE, Brooks SP, Amos W, Harwood J (2004) Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology*, **13**, 811–825.
- Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **344**, 403–410.
- Hästbacka J, De La Chapelle A, Kaitila I *et al.* (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genetics*, **2**, 204–211.
- Hey J (2005) On the number of new world founders: a population genetic portrait of the peopling of the Americas. *Public Library of Science Biology* **3**, e193.
- Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics (Oxford)*, **18**, 337–338.
- Kingman JFC (1982) The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.
- Liberg O (2006) Genetic aspects of viability in small wolf populations – with special emphasis on the Scandinavian wolf population. *Report from an international expert workshop at Färna Herrgård, Sweden 1st–3rd May 2002. Report 5436*. Swedish Environmental Protection Agency.
- Luikart G, Cornuet JM (1998) Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, **12**, 228–237.
- Maruyama T, Fuerst PA (1985) Population bottlenecks and non-equilibrium models in population-genetics. 2. Number of alleles in a small population that was formed by a recent bottleneck. *Genetics*, **111**, 675–689.
- Mayr E (1954) Change of genetic environment and evolution. In: *Evolution as a Process* (eds Huxley J, Hardy AC, Ford EB), pp. 157–180. Allen & Unwin, London.
- Mech LD, Seal US (1987) Premature reproductive activity in wild wolves. *Journal of Mammalogy*, **68**, 871–873.
- Meyer A, Kocher TD, Basasibwaki P, Wilson AC (1990) Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature*, **347**, 550–553.
- Nei M, Maruyama T, Chakraborty R (1975) The bottleneck effect and genetic variability in populations. *Evolution*, **29**, 1–10.
- Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.
- Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Molecular Biology and Evolution*, **9**, 552–569.
- Seddon JM, Parker HG, Ostrander EA, Ellegren H (2005) SNPs in ecological and conservation studies: a test in the Scandinavian wolf population. *Molecular Ecology*, **14**, 503–511.
- Slatkin M (2004) A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. *American Journal of Human Genetics*, **75**, 282–293.
- Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, **26**, 119–165.
- Vila C, Sundqvist AK, Flagstad O *et al.* (2003) Rescue of a severely bottlenecked wolf (*Canis lupus*) population by a single immigrant. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **270**, 91–97.
- Vincek V, O’Huigin C, Satta Y *et al.* (1997) How large was the founding population of Darwin’s finches? *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **264**, 111–118.
- Vogel F, Motulsky AG (1996) *Human Genetics: Problems and Approaches*. Springer-Verlag, New York.

- Wabakken P, Rønning H (2003) *Ulv I Skandinavia Vinteren 2002–03: Foreløpig Statusrapport*. Høgskolen i Hedmark, Evenstad, Norway.
- Wabakken P, Sand H, Liberg O, Bjarvall A (2001) The recovery, distribution, and population dynamics of wolves on the Scandinavian peninsula, 1978–98. *Canadian Journal of Zoology-Revue Canadienne de Zoologie*, **79**, 710–725.
- Watterson GA (1984) Allele frequencies after a bottleneck. *Theoretical Population Biology*, **26**, 387–407.
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: *Proceedings of the 6th International Congress of Genetics*, **1**, 356–366.

Raphaël Leblois is currently working on statistical methods estimating demographic parameters from genetic data using coalescent theory and maximum likelihood inference methods. He is particularly interested in natural population history and dispersal inference under various realistic models such as isolation by distance or divergence with migration models. Montgomery Slatkin is a theoretical population geneticist who works on models of evolutionary and human genetics.
