Detecting Small Amounts of Gene Flow From Phylogenies of Alleles

Montgomery Slatkin

Department of Zoology, University of California, Berkeley, California 94720

Manuscript received April 25, 1988
Accepted for publication November 12, 1988

ABSTRACT

The method of coalescents is used to find the probability that none of the ancestors of alleles sampled from a population are immigrants. If that is the case for samples from two or more populations, then there would be concordance between the phylogenies of those alleles and the geographic locations from which they are drawn. This type of concordance has been found in several studies of mitochondrial DNA from natural populations. It is shown that if the number of sequences sampled from each population is reasonably large (10 or more), then this type of concordance suggests that the average number of individuals migrating between populations is likely to be relatively small ($Nm < 1$) but the possibility of occasional migrants cannot be excluded. The method is applied to the data of E. Bermingham and J. C. Avise on mtDNA from the bowfin, *Amia calva*.

NEW methods for examining DNA from natural populations provide new opportunities for understanding processes governing the evolution of those populations. In this paper, I will suggest that data on the geographic distribution of segments of DNA (alleles) can provide a test for small amounts of gene flow among populations. To use this test, there must be sufficient information about the sequences of the alleles that it is possible to infer their phylogeny. At the present time, there is abundant data of this type for mitochondrial DNA, for which the lack of recombination in animals makes it relatively easy to reconstruct phylogenies. AVISE *et al.* (1987) review studies of the geographic distribution of mtDNAs within species.

The particular geographic pattern of alleles that I will be concerned with here is one in which there is concordance between the phylogenies of alleles sampled from a species and the geographic locations from which those alleles were sampled. An example of this kind of data is from the bowfin (*Amia calva*), a freshwater fish living in streams on the Atlantic and Gulf of Mexico coasts of the United States [see AVISE *et al.* (1987) Figure 3]. BERMINGHAM and AVISE (1986) sampled individuals from several streams and found that an unrooted parsimony tree of the mtDNAs showed two distinct clusters, one containing mtDNAs from western Florida, Alabama, Mississippi and Louisiana, and the other containing mtDNAs from eastern Florida, Georgia and South Carolina. From this observation, BERMINGHAM and AVISE concluded that there is no or effectively no gene flow between these two geographic areas. They found the same pattern in three other species.

This pattern of concordance suggests that the entire ancestry of the mtDNAs sampled from each of the two regions were present only in those regions. That is, the ancestors of the samples for the Atlantic states and eastern Florida were all in that region and the ancestors of the Gulf of Mexico samples were all in that region. In what follows, I will compute the probability of "nonimmigrant" ancestry in a single population. I will also show that, if the concordance between the phylogenies of alleles and their geographic locations is taken as evidence of nonimmigrant ancestry, then this result can be used to place an approximate upper bound on the level of gene flow between populations. A subsequent paper (M. SLATKIN and W. MADDISON, unpublished data) will address the problem of inferring the number of migration events in the ancestry of samples of alleles from a subdivided population.

To carry out the analysis, I will use the method of coalescents, which was introduced by KINGMAN (1982a, b) and which depends on the assumption that all sequences sampled are neutral. TAVARÉ (1984) and WATTERSON (1985) review the literature in this area. The method I will use is closely related to the "lines-of-descent" approach of GRIFFITHS (1979). Recently, TAKAHATA (1988) has developed the theory of coalescents in two populations with gene flow between them. He showed that the general results cannot be found in closed form except for relatively small numbers of alleles sampled from each population. The following analysis is in effect a special case of TAKAHATA's more general model but one for which relatively simple answers can be obtained.

EQUILIBRIUM MODEL

We will consider populations that have been separated from one another for indefinitely long. For convenience, each segment of DNA will be referred to as an allele.

One population: The main result can be derived from a model of a single population containing N haploid individuals. Assume that all N individuals contribute equally to an infinite pool of gametes and then a fraction m of the gamete pool is replaced by immigrants every generation. The results for a diploid population are the same if N is replaced by $2N$, but because the data available now are for mitochondrial DNA the haploid model is more relevant. The problem is to find the probability that all the ancestors of a sample of n alleles from this population were residents of this population. This probability will be referred to as the probability of "nonimmigrant ancestry." The method for solving this problem is the usual one in analyzing coalescent processes.

Consider the n alleles in a sample and their ancestry in the previous generation. The number of alleles that are descended from immigrants in the previous generation is a binomial random variable with parameter m and sample size n. Let M_i be the probability that i of the n alleles are descended from immigrants in the previous generation. If $nm \ll 1$, which is the case we will be concerned with here, then $M_0 \approx 1 - nm$, $M_1 \approx nm$, and $M_i = O((nm)^2)$ for $i > 1$, where $O((nm)^2)$ indicates that the quantity is of the same order of magnitude as $(nm)^2$ or smaller.

Let $G_{n,k}$ be the probability that n alleles are descended from k nonimmigrant alleles in the preceding generation. I will assume that $N \gg 1$ and $n \ll N$, in which case

$$G_{n,n} = (1 - m)^n\left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right) \dots \left(1 - \frac{n-1}{N}\right) \approx 1 - nm - \frac{n(n-1)}{2N} \quad (1)$$

$$G_{n,n-1} \approx \frac{n(n-1)}{2N}$$

and $G_{n,k} = O(n^2/N^2)$ for $k < n - 1$ (Tavaré 1984).

If we consider the ancestry of the n alleles going backward in time, there are, under our assumption that $m \ll 1$, $1/N \ll 1$ and $n \ll N$, only three things that can happen in any generation: one of the alleles is an immigrant, which occurs with probability nm; there is a *coalescence* (meaning that two of the alleles are descended from a common ancestor in the previous generation), which occurs with probability $n(n - 1)/(2N)$; or there are n alleles in the previous generation. In last case, nothing has happened so we can disregard that possibility and continue to look one more generation in the past. In tracing back the ancestry of the n alleles, we will either come to a coalescence or to the detection of an immigrant. From (1), the probability that the coalescence occurs first is

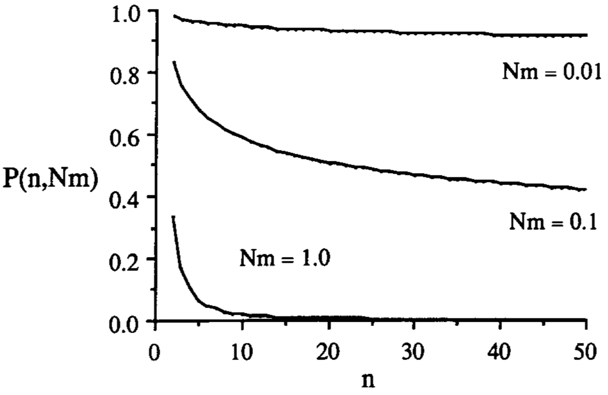

FIGURE 1.—A graph of the probability that all n alleles drawn from a population are descended from an ancestral allele that was in that population. These curves are obtained from Equation 1 in the text.

$$\pi_n = \frac{\dfrac{n(n-1)}{2N}}{\dfrac{n(n-1)}{2N} + nm} = \frac{(n-1)}{(n-1) + 2Nm}$$

and the probability that an immigrant is detected first is $1 - \pi_n$. Note that for a given value of $2Nm$, π_n increases with n. The reason is that as n increases the chance that each allele is descended from another allele in the sample is proportional to $n - 1$ but the chance that each allele is an immigrant is m and is independent of n.

After there is a coalescence, there are $n - 1$ alleles and we can consider the same model with n replaced by $n - 1$. By induction, the probability that all n alleles are descended from nonimmigrants is

$$P(n,Nm) = \pi_2\pi_3 \dots \pi_n = \frac{(n-1)!}{(2Nm)_{(n)}} \quad (2)$$

where $x_{(n)} = x(x + 1) \dots (x + n - 1)$.

The derivation of (2) is similar to the theory of Ewens (1972), Griffiths (1979) and others who were concerned with the number of distinct alleles in a sample of n alleles from a randomly mating population. The only difference is that mutation in those theories is replaced by immigration here. Equation 2 is equivalent to that of Ewens (1972, Eq. 23) with $i = 1$ and $\theta = 2Nm$) for the probability of only a single type of allele in a sample of n alleles. There is a factor of 2 difference because the present model is of a haploid population.

It is easy to evaluate $P(n, Nm)$ for different values of Nm. Some values are shown in Figure 1 which illustrate the main result of this analysis: for values of n that are typical of samples of mtDNA taken from natural populations, $P(n, Nm)$ is very small unless $Nm < 1$. With even moderate levels of gene flow ($Nm =$

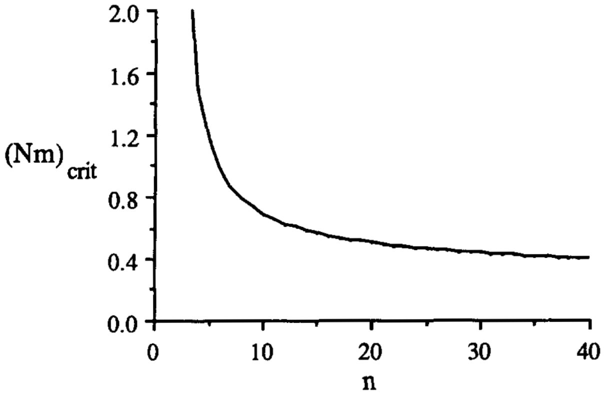

FIGURE 2.—A graph of the solution to the equation $P(n, Nm) = 0.05$ for different values of n. If $Nm > (Nm)_{\text{crit}}$, then the probability that all n alleles in a sample from a population are descended from an allele that was also in the population is less than 0.05.

1.0), there is a very small probability of nonimmigrant ancestry. For lower levels of gene flow ($Nm = 0.1$), however, $P(n, Nm)$ is still not 1 and there is some chance that one or more of the ancestors of the n alleles are immigrants.

Two-populations: For two populations, the joint probability of nonimmigrant ancestry in both is just the product of the probabilities for each population because the coalescent events in each populations are independent. If n_1 alleles are sampled from a population of size N_1 with immigration rate m_1 and n_2 alleles are sampled from a second population of size N_2 with immigration rate m_2, then the probability of nonimmigrant ancestry in both populations is $P(n_1, N_1m_1)P(n_2,N_2m_2)$.

APPLICATION

We can illustrate how the above results are used by applying them to the data of BERMINGHAM and AVISE (1986) on mtDNA in *Amia calva*. They analyzed mtDNA from 75 individuals and found 13 distinct types by using 13 restriction enzymes that cut the mtDNA at an average of 54 sites. A parsimony analysis of the 13 distinct types placed them in two groups, one (with 4 distinct types in 16 sampled) from western Florida, Alabama, Mississippi and Louisiana and the other (with 9 distinct types in 59 sampled) from South Carolina, Georgia and eastern Florida. In the present notation, $n_1 = 16$ and $n_2 = 59$ and we want to use Equations 1 and 2 to find upper bounds on N_1m_1 and N_2m_2 that are consistent with these observations.

There are two ways to proceed. One is to assume N_1m_1 and N_2m_2 differ and use Equation 1 for each population separately. For either population considered separately we can find the value of Nm, $(Nm)_{\text{crit}}$, for which $P(\text{n}, (Nm)_{\text{crit}}) = 0.05$ (or some other threshold probability). If $Nm > (Nm)_{\text{crit}}$, the probability of the finding concordance is less than 5%. Figure 2 shows $(Nm)_{\text{crit}}$ for different values of n. Surprisingly, $(Nm)_{\text{crit}}$ does not decrease very rapidly with n. Increasing n from 20 to 40 decreases $(Nm)_{\text{crit}}$ from 0.50 to only 0.40. For the particular example we are concerned with here, the critical value of N_1m_1 (with $n_1 = 16$) is 0.547 and the critical value of N_2m_2 (with $n_2 = 59$) is 0.356.

The other way to proceed is to assume $N_1m_1 = N_2m_2 = Nm$ (or some other fixed relationship) and then use Equation 2 to place an upper bound on Nm by solving for $P(n_1, Nm)P(n_2, Nm) = 0.05$. Using this method, the upper bound on Nm is 0.202. Both approaches show that it is likely that gene flow between these two populations is relatively rare, as BERMINGHAM and AVISE (1986) concluded.

Note that in this analysis, it is number of mtDNAs sampled from each population that determines the bounds on Nm, not the number of distinct mtDNA types found. That is true as long as the individuals are randomly sampled from each population. Any tendency to chose close relatives would reduce the number of independent alleles sampled.

DISCUSSION AND CONCLUSIONS

The above results show that if the concordance between the geographic distribution of neutral alleles and their phylogeny is assumed to indicate non-immigrant ancestry of those alleles then it is possible to place an upper bound on the amount of gene flow between populations. The application of this method confirms the conclusion of Avise *et al.* (1987) that the finding of such a concordance indicates that there is little or no gene flow between populations. However, the upper bound on Nm is not as small as intuition might suggest and some gene flow might have been occurring and not be detected by this kind of analysis.

The method described here is based on the assumption that the populations sampled have been separated for indefinitely long. The results of GRIFFITHS (1979) and WATTERSON (1985) show that this assumption is valid if the time of separation is much larger than four times the largest of the population sizes after separation. In that case, alleles in each population would have their most recent common ancestor after the separation, so any similarity of alleles would be due to gene flow and not to common descent from an ancestral allele present before the separation. If the time of separation is much less than that time, it is unlikely that concordance of allelic phylogenies and geographic locations would be observed.

This approach to analyzing DNA sequence data does not make full use of the information in the data because the phylogeny of a sample of alleles alone does not indicate the extent of divergence among those alleles. It may be possible to develop methods that take the extent of divergence into account and produce a more accurate estimate of levels of gene flow.

This research has been supported in part by National Science Foundation grant BSR-8500258 and National Institutes of Health grant GM40282. I thank N. TAKAHATA for sending me a copy of his paper on the *n*-coalescent in a subdivided population and J. C. AVISE, R. R. HUDSON, C. MORITZ, N. TAKAHATA and an anonymous referee for helpful comments on earlier versions of this paper.

LITERATURE CITED

AVISE, J. C., J. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, J. E. NEIGEL, C. A. REEB and N. C. SAUNDERS, 1987 Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. **18:** 489–522.

BERMINGHAM, E., and J. C. AVISE, 1986 Molecular zoogeography of freshwater fishes in the southeastern United States. Genetics **113:** 939–965.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

GRIFFITHS, R. C., 1979 Exact sampling distributions from the infinite neutral alleles model. Adv. Appl. Prob. **11:** 326–354.

KINGMAN, J. F. C., 1982a The coalescent. Stochast. Proc. Appl. **13:** 235–248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

TAKAHATA, N., 1988 The *n* coalescent in two partially isolated diffusion populations. Genet. Res. (in press).

TAVARÉ, S. 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Popul. Biol. **26:** 119–164.

WATTERSON, G. A., 1985 The genetic divergence of two populations. Theor. Popul. Biol. **27:** 298–317.

Communicating editor: R. R. HUDSON