Hierarchical Analysis of Nucleotide Diversity in Geographically Structured Populations

Kent E. Holsinger and Roberta J. Mason-Gamer[1]

Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043

Manuscript received August 1, 1994
Accepted for publication November 4, 1995

ABSTRACT

Existing methods for analyzing nucleotide diversity require investigators to identify relevant hierarchical levels before beginning the analysis. We describe a method that partitions diversity into hierarchical components while allowing any structure present in the data to emerge naturally. We present an unbiased version of NEI's nucleotide diversity statistics and show that our modification has the same properties as WRIGHT's F_{ST}. We compare its statistical properties with several other F_{ST} estimators, and we describe how to use these statistics to produce a rooted tree of relationships among the sampled populations in which the mean time to coalescence of haplotypes drawn from populations belonging to the same node is smaller than the mean time to coalescence of haplotypes drawn from populations belonging to different nodes. We illustrate the method by applying it to data from a recent survey of restriction site variation in the chloroplast genome of *Coreopsis grandiflora*.

POPULATION geneticists have long recognized that the genetic diversity present in a species is hierarchically structured. In addition to differences among individuals within any one population, there may be differences among populations within a given geographical region, differences among populations from different geographical regions, and differences among entire geographical regions. SEWALL WRIGHT (1951, 1965) introduced *F*-statistics as a way of assessing genetic differentiation at each of these levels, and the intervening 40 years have repeatedly demonstrated how useful this approach can be (see, for example, the reviews on patterns of electrophoretic diversity by NEVO *et al.* 1984, NEI 1987, or HAMRICK and GODT 1990).

In the past 10 years, population geneticists have turned to increasingly sensitive techniques for detecting genetic variation, with a view toward uncovering fine-scale population structure that may be undetectable in routine allozyme surveys. In particular, population surveys of restriction site and nucleotide sequence variation are now becoming routine. Unfortunately, *F*-statistics, whether those originally proposed by WRIGHT or their more recent modifications (COCKERHAM 1969, 1973; NEI 1973; WEIR and COCKERHAM 1984; LONG 1986), are not appropriate for analysis of the variation revealed by these new techniques. *F* statistics are calculated from allele frequencies at many, independently inherited loci. The variation revealed in restriction site and sequence studies, however, almost always results from differences at sites that are not independently inherited.

Corresponding author: Kent E. Holsinger, Department of Ecology and Evolutionary Biology, U-43, University of Connecticut, Storrs, CT 06269-3043. E-mail: kent@darwin.eeb.uconn.edu

[1] *Current address:* Harvard University Herbaria, 22 Divinity Ave., Cambridge, MA 02138.

EXCOFFIER *et al.* (1992) recently described a method appropriate for the analysis of restriction site and sequence data that is related to WRIGHT's *F*-statistics. It uses a distance matrix to describe the relationship among haplotypes in the sample and partitions the variation present in the sample into three components: variation within populations, variation among populations within a particular geographical region, and variation among regions. It is formally equivalent to a nested analysis of variance of the nucleotide diversity in the sample, and it is a simple generalization to several hierarchical levels of the analysis of variance approach WEIR and BASTEN (1990) originally suggested. NEI (1982) proposed a method related to his G_{ST} statistics for allozymes (NEI 1973), while TAKAHATA and PALUMBI (1985), LYNCH and CREASE (1990), and NEI and MILLER (1990) proposed similar methods for estimating the average number of nucleotide substitutions between populations. Valuable as these methods are, they all suffer from one important limitation. As with any nested analysis of variance, the investigator must specify the hierarchical structure of the data before beginning the analysis. In other words, each method requires that a predetermined hierarchical structure be imposed on the data, not inferred from it. Thus, they are not well suited to identifying the hierarchical structure that best reflects the pattern of genetic differentiation among populations.

In this paper, we describe a method for describing the hierarchical structure of nucleotide diversity in a sample that builds on these approaches. To develop this method we provide a bias correction to NEI's (1982) nucleotide diversity statistics. This statistic provides a direct analog of F_{ST} that is appropriate for haplotype

frequency data, and we compare its statistical behavior with that of other F_{ST} measures based on nucleotide sequence data. We call this measure $\hat{g}_{st}$, to emphasize its close relationship to NEI's (1982) g_{st}. Using the fact that F_{ST} is equal to the ratio of the average coalescence times for different pairs of genes (SLATKIN 1991), we show how $\hat{g}_{st}$ can be used to group populations based on the average time to coalescence for pairs of haplotypes. The results of this analysis can be displayed as a tree diagram depicting the pattern of genetic differentiation among populations. The mean time to coalescence for two haplotypes drawn from the same node of the tree is less than that for two haplotypes drawn from different nodes. In addition to providing a more detailed description of the pattern of nucleotide diversity than methods that require pre-specified hierarchical categories, the structure of the tree can be interpreted as reflecting either patterns of gene exchange or phylogenetic relationship and the statistical significance of the structure that is revealed can be assessed. We illustrate the method by applying it to data derived from a recent survey of restriction site variation in the chloroplast genome of *Coreopsis grandiflora* in the southern United States (MASON-GAMER *et al.* 1995).

THE METHOD

Measuring nucleotide sequence diversity: Let n_{ik} be the number of individuals with haplotype i collected from population k and $n_k = \Sigma_i n_{ik}$. Then $\hat{x}_{ik} = n_{ik}/n_k$ is the maximum-likelihood estimate of x_{ik}, the frequency of haplotype i in population k. Similarly, $\hat{x}_i = \Sigma_k \hat{x}_{ik}/n$ is the maximum-likelihood estimate of its average frequency, $x_i = \Sigma_k x_{ik}/n$, in the set of populations being considered, where n is the number of populations included in the sample. Let δ_{ij} be the number of differences found between the ith and jth haplotypes in the sample. For restriction site surveys, δ_{ij} is the number of restriction site differences found between two haplotypes i and j. For nucleotide sequence surveys, δ_{ij} is the number of nucleotide differences found between two sequences i and j.

Let ν_k be the average number of differences between two haplotypes drawn at random from population k. NEI and TAJIMA (1981) showed that

$$\hat{\nu}_k = \frac{n_k}{n_k - 1} \sum_{ij} \hat{x}_{ik}\hat{x}_{jk}\delta_{ij} \tag{1}$$

is an unbiased estimator for ν_k. Similarly, we define ν as the average number of differences between two haplotypes drawn at random without respect to the population from which they were drawn: $\nu = \Sigma_{ij} x_i x_j \delta_{ij}$.

Because the sample is drawn from several populations, a correction is necessary to obtain an unbiased estimate of ν from sample data (*cf.* NEI and CHESSER 1983). In making this correction, we consider only the stochastic variation that arises from the sampling process involving the actual populations chosen, not the stochastic variation arising from the choice of populations (*cf.* WEIR and COCKERHAM 1984). Let $\nu^* = \Sigma_{ij} \hat{x}_i\hat{x}_j\delta_{ij}$. Then

$$E(\nu^*) = \sum_{ij} E(\hat{x}_i\hat{x}_j)\delta_{ij}.$$

Assuming there is no covariance in the sampling of haplotypes from different populations, $E(\hat{x}_{ik}\hat{x}_{jl}) = x_{ik}x_{jl}$ for $l \neq k$ (*cf.* TAKAHATA and PALUMBI 1985). Recalling this, we find that

$$E(\hat{x}_i\hat{x}_j) = E\left(\frac{1}{n}\sum_k \hat{x}_{ik}\right)\left(\frac{1}{n}\sum_l \hat{x}_{jl}\right) = \frac{1}{n^2}\sum_{kl} E(\hat{x}_{ik}\hat{x}_{jl})$$

$$= \frac{1}{n^2}\sum_k \sum_{l\neq k} x_{ik}x_{jl} + \frac{1}{n^2}\sum_k E(\hat{x}_{ik}\hat{x}_{jk}) = \frac{1}{n^2}\sum_k \sum_{l\neq k} x_{ik}x_{jl}$$

$$+ \frac{1}{n^2}\sum_k \left(\frac{n_k - 1}{n_k}\right)x_{ik}x_{jk} = x_i x_j - \frac{1}{n^2}\sum_k \frac{x_{ik}x_{jk}}{n_k}.$$

Thus,

$$\hat{\nu} = \sum_{ij} \hat{x}_i\hat{x}_j\delta_{ij} + \frac{1}{n^2}\sum_k \frac{\hat{\nu}_k}{n_k}, \tag{2}$$

is an unbiased estimator for ν. Let π_{ij} be the probability that haplotypes i and j differ at any nucleotide. Then the probability that two randomly chosen haplotypes differ at any nucleotide is $\pi = \Sigma_{ij} x_i x_j \pi_{ij}$ (NEI and TAJIMA 1981). If the differences in the sample are nucleotide sequence differences, then an unbiased estimator for π is

$$\hat{\pi} = \frac{\nu}{N}, \tag{3}$$

where N is the length of the nucleotide sequences. If the differences in the sample are restriction site differences, then an unbiased estimator for the nucleotide diversity is

$$\hat{\pi} = \frac{\nu}{2\,\Sigma_i\, m_i r_i}, \tag{4}$$

where m_i is the number of restriction sites detected with the ith restriction enzyme and r_i is the number of nucleotides in its recognition sequence (NEI and TAJIMA 1981; NEI 1987). The nucleotide diversity within each population, $\hat{\pi}_k$, is estimated in the same way.

We now have unbiased estimates for the average nucleotide diversity within populations, namely the mean of the $\hat{\pi}_k$, and for the nucleotide diversity in a set of populations. Following NEI (1973), let $\hat{g}_{st}$ be the proportion of diversity in the sample due to differences among populations. Then

$$\hat{g}_{st} = \frac{\hat{\pi} - \bar{\hat{\pi}}}{\hat{\pi}}, \tag{5}$$

where $\bar{\hat{\pi}} = \Sigma\hat{\pi}_k/n$ is the mean of the $\hat{\pi}_k$ (*cf.* NEI 1982;

TAKAHATA and PALUMBI 1985). $\hat{g}_{st}$ differs from NEI's (1982) g_{st} in two ways: it includes a bias correction both for average nucleotide diversity within populations and for total nucleotide diversity in the sample and average nucleotide diversity within populations is not a weighted average based on subpopulation size.

If f_{ij} is the probability that haplotypes i and j are identical at any nucleotide, and f is the probability that two haplotypes chosen at random match at any nucleotide, then

$$f = \sum_{ij} x_i x_j f_{ij} = \sum_{ij} x_i x_j (1 - \pi_{ij}) = 1 - \pi. \quad (6)$$

Similarly, if $\bar{f}$ is the mean across populations of the probability that two haplotypes chosen at random from the same population match at any nucleotide, then

$$\bar{f} = \frac{1}{n} \sum_{k} \sum_{ij} x_{ik} x_{jk} f_{ij} = \frac{1}{n} \sum_{k} \sum_{ij} x_{ik} x_{jk} (1 - \pi_{ij}) = 1 - \bar{\pi}. \quad (7)$$

Using (6) and (7) it is easy to see that $\hat{g}_{st}$ is a good estimator for F_{ST}, because

$$g_{st} = \frac{(1 - f) - (1 - \bar{f})}{1 - f} = \frac{\bar{f} - f}{1 - f},$$

which is precisely WRIGHT's definition for F_{ST}. Although $\hat{g}_{st}$ is not an unbiased estimator of g_{st}—because $E(\hat{\bar{\pi}}/\hat{\pi}) \neq E(\hat{\bar{\pi}})/E(\hat{\pi})$—the bias will be small for moderate to large sample sizes (STUART and ORD 1987).

Haplotype diversity and coalescence times: SLATKIN (1991) pointed out that if mutation is rare, then

$$F_{ST} = \frac{\bar{t} - \bar{t}_0}{\bar{t}}, \quad (8)$$

where $\bar{t}$ is the average time to coalescence of two genes drawn at random without respect to population and $\bar{t}_0$ is the average time to coalescence of two genes drawn at random from the same population. If f_i is $\bar{f}$ in one set of populations, f_j is $\bar{f}$ in a second set of populations, and f_{ij} is F_{ST} in the combined set of populations, then WRIGHT (1951) showed that

$$1 - f_{ij} = (1 - g_{ij})\{1 - (pf_i + qf_j)\}, \quad (9)$$

where g_{ij} is the proportion of genetic variation in the whole sample owing to differences between populations in the first and second set, p is the fraction of populations in the sample from the first group, and q is the fraction of populations in the sample from the second group. For notational convenience we use g_{ij} instead of g_{st} when discussing the degree of genetic differentiation between two specified sets of populations, where i and j are the population indices.

Let $\bar{t}_{ij}$ be the average time to coalescence of two genes drawn at random without respect to which set they are drawn from, $\bar{t}_i$ be the average time to coalescence of two genes drawn at random from the first set of populations, and $\bar{t}_j$ be the average time to coalescence of two genes drawn at random from the second set of populations. Then it follows from (8) that

$$f_{ij} = \frac{\bar{t}_{ij} - \bar{t}_0}{\bar{t}_{ij}} \quad (10)$$

and

$$pf_i + qf_j = p[1 - (t_0/t_i)] + q[1 - (t_0/t_j)] = 1 - \frac{t_0}{\tilde{t}}. \quad (11)$$

where $\tilde{t} = \{p(1/\bar{t}_i) + q(1/\bar{t}_j)\}^{-1}$ is the harmonic mean of $\bar{t}_i$ and $\bar{t}_j$ (*cf.* SLATKIN 1993). Thus,

$$g_{ij} = \frac{t_{ij} - \tilde{t}}{t_{ij}}. \quad (12)$$

Just as F_{ST} measures the increase in coalescence time resulting from the grouping of genes into populations, hierarchically expanded F-statistics measure the increase in coalescence time resulting from the grouping of populations into more inclusive sets of populations. Thus, pairwise F-statistics provide a convenient measure of the evolutionary divergence between two sets of populations (PIAZZA and MENOZZI 1983; NEI 1987; SLATKIN 1991). The increased time to coalescence for genes from different populations or from different groups of populations may mean either that the rate of gene exchange between the two sets of populations is smaller than that within each set or that the two sets of populations stopped exchanging genes entirely some time ago. The patterns produced by these processes cannot be distinguished from F-statistics alone (FELSENSTEIN 1982).

Assessing methods of estimating F_{ST}: The relationship between F_{ST} and coalescence times allows us to compare the accuracy of several different estimators F_{ST}. We used the algorithm described in APPENDIX A to build many replicate coalescent trees for sampled haplotypes under a wide variety of population sizes and migration structures. For each of the coalescent trees in our sample, we constructed a random nucleotide sequence to serve as the common ancestor of all alleles in that coalescent sample. All bases occurred at each position in this sequence with equal probability. Mutations were assigned randomly along each branch of the genealogy following a JUKES and CANTOR (1969) substitution model. Specifically, the number of mutations assigned was a Poisson random variate with mean equal to the product of branch length and mutation rate. The position of each mutation was chosen at random, and each of the three nucleotides that could replace the current nucleotide at that position was chosen with equal probability. We then compared estimates of F_{ST} calculated from the simulated sequence data using NEI's (1982) g_{st}, our $\hat{g}_{st}$, LYNCH and CREASE's (1990) N_{ST}, and WEIR and BASTEN's β with the known F_{ST} as calculated from the coalescent history of the sample.

TABLE 1

Performance of alternative F_{ST} measures

N_em	Population size	Sample configuration[a]	Mean squared error of estimate			
			$\hat{g}_{st}$	g_{st}	N_{ST}	β
0.5	2500	1	0.000499	0.000993	0.002031	0.001966
		2	0.000700	0.004511	0.002292	0.002201
		3	0.000644	0.002561	0.002156	0.01498
	500	4	0.002152	0.004388	0.003688	0.00366
		5	0.003771	0.02002	0.005577	0.005544
		6	0.002521	0.007348	0.004104	0.01552
1.0	2500	1	0.000240	0.000971	0.001066	0.001031
		2	0.000459	0.005369	0.001295	0.001246
		3	0.000365	0.002896	0.001228	0.01054
	500	4	0.00124	0.00461	0.002325	0.002313
		5	0.002998	0.02345	0.004506	0.00448
		6	0.001747	0.008071	0.002896	0.01395
2.0	2500	1	0.000107	0.001003	0.000454	0.00440
		2	0.002378	0.006071	0.002378	0.000605
		3	0.000216	0.003192	0.000610	0.007807
	500	4	0.000749	0.004598	0.001315	0.001308
		5	0.002188	0.02626	0.00338	0.003361
		6	0.001113	0.008762	0.00187	0.01126

[a] See Table 2 for the list of sample configurations.

For each combination of parameters, we constructed a sample of 1000 coalescent trees. For each coalescent tree, we constructed 10 sequence data sets using different randomly generated starting sequences and independent realizations of the mutation process.

Results for one set of simulations involving sequences of 500 nucleotides drawn from five populations with 500 or 2500 individuals in each population are shown in Table 1. For each of the two population sizes, several different sample configurations were also considered (Table 2). Remarkably, the mean squared error of $\hat{g}_{st}$ is two or more times smaller than that of any of the remaining estimators for every population size and sample configuration presented. Simulations involving sequences of 50 nucleotides or 1500 nucleotides showed the same pattern. With longer sequences, of course, the mean squared error of all estimators is reduced, but $\hat{g}_{st}$ benefitted more from this effect than the other estimators. Increasing the sequence length from 500 to 1500 nucleotides reduced the mean squared error of $\hat{g}_{st}$ by an average of 49% while the mean squared error of the remaining estimates was reduced by less than 24%.

TABLE 2

Sample configurations for the F_{ST} analysis

Number	Sample configuration
1	25,25,25,25,25
2	10,10,10,10,10
3	5,15,25,35,45
4	12,12,12,12,12
5	5,5,5,5,5,5
6	4,8,12,16,20

The results presented in Table 1 ignore two aspects of haplotype sampling that may be critically important in interpreting F_{ST} statistics: haplotypes samples are typically gathered from a small subset of the populations that may actually be exchanging migrants and effective population sizes may differ substantially from one sample population to another. To examine the first of these effects, we performed a series of simulations in which haplotypes were sampled from five randomly chosen populations out of a total of 21 populations. The mean squared errors reported in Table 3 are quite similar to those for simulations in which the five populations sampled represent the entire set of populations exchanging migrants. These results suggest that F_{ST} estimates may be relatively insensitive to the proportion of populations sampled, although much more extensive simulations would be necessary to confirm this conjecture. To examine the second effect, we performed a series of simulations in which population sizes differed dramatically from one another. In the most extreme case, the five populations had sizes of 100, 250, 2500, 25000, and 100,000 and resulted in a mean squared error for $\hat{g}_{st}$ of 2.23×10^{-4}. In the least extreme case, the five populations had sizes of 500, 1000, 2500, 4000, and 4500 and resulted in a mean squared error for $\hat{g}_{st}$ of 2.50×10^{-4}. These errors are comparable with those reported in Table 1 and Table 3. Thus, it appears that variation in population sizes has little impact on the accuracy of F_{ST} estimates.

Our simulations included both multiple coalescent trees for each set of parameter values and multiple realizations of the mutation process for each coalescent tree. We took advantage of this structure to partition

TABLE 3

Population sampling and F_{ST} estimates

N_em	Population size	Sample configuration[a]	Mean squared error of estimate			
			$\hat{g}_{st}$	g_{st}	N_{ST}	β
0.25	2500	1	0.000469	0.000974	0.003109	0.003145
		2	0.000515	0.003354	0.003495	0.003291
		3	0.000417	0.002016	0.002689	0.020080
0.5	2500	1	0.000318	0.000870	0.002241	0.002095
		2	0.000482	0.004608	0.002294	0.002196
		3	0.000344	0.002563	0.002642	0.015600
2.5	2500	1	0.000036	0.000921	0.000394	0.000346
		2	0.000110	0.006372	0.000576	0.000526
		3	0.000087	0.003265	0.000386	0.006239
5.0	2500	1	0.000022	0.000995	0.000168	0.000146
		2	0.000062	0.006259	0.000321	0.000243
		3	0.000052	0.003381	0.000272	0.006965
10.0	2500	1	0.000014	0.001009	0.000014	0.000061
		2	0.000039	0.006543	0.000118	0.000086
		3	0.000029	0.003513	0.000139	0.004209

[a] See Table 2 for the list of sample configurations.

the overall variance in F_{ST} estimates into two components. The first of these reflects the inaccuracies associated with estimating the actual coalescent history of the haplotypes in our sample. It is a result of estimation error. The small mean squared error associated with each of the F_{ST} estimates, and especially with $\hat{g}_{st}$, shows that estimation error is small and that $\hat{g}_{st}$ and the other estimators provide a good estimate of the average coalescent history of a sample of haplotypes. The second variance component reflects the differences among coalescent histories that can be produced by the same population parameters. It reflects intrinsic variance associated with the drift-migration process. For the simulations reported in Table 1, the intrinsic drift variance is between five and 25 times greater than the estimation variance. In fact, the drift variance appears to be more than 10 times greater than the estimation variance whenever expected sequence divergence is greater than ~1%. In short, $\hat{g}_{st}$ and other F_{ST} estimators can provide a good estimate of the coalescent history of a particular sample of haplotypes, but inferences about migration rates based on those F_{ST} estimates may be problematic.

Uncovering hierarchical structure: The results in the preceding sections show that $\hat{g}_{st}$ is a useful measure of genetic differentiation between pairs of populations (or between paired sets of populations). How can we use this information to uncover hierarchical structure in the data? One way would be to treat the pairwise $\hat{g}_{st}$s as distances and analyze them using one of the existing methods for phylogenetic analysis of distance data (*cf.* PIAZZA and MENOZZI 1983; MERCURE *et al.* 1993). This approach is unsatisfactory in two ways. First, methods for phylogenetic analysis of distance data generally assume that branch lengths are additive, but inspection of (9) shows that the $\hat{g}_{st}$s are not additive. Second, none of the existing methods for phylogenetic inference from distance data take advantage of the mathematical partitioning of diversity components among hierarchical levels implicit in (9). The following algorithm corrects both of these deficiencies:

(1) Construct a list of nodes, treating each population as a separate node.
(2) Compute $\hat{g}_{ij}$ between all pairs of nodes in the node list, using (5) to calculate f_{ij}, f_i, and f_j. ($f_k = 0$ for any k in which the node consists of a single population.)
(3) Combine the two nodes with the smallest $\hat{g}_{ij}$ into a single node.
(4) Return to (2) until the node list is empty.

This algorithm is similar to the widely-used unweighted pair-group method with arithmetic mean (UPGMA) for cluster analysis (SNEATH and SOKAL 1973). In both that method and this one, distances from each new node to every other node are recomputed at each step. Our method differs from UPGMA in the way in which the distances are recomputed. In UPGMA, if i and j are combined into a new node, k, then $\hat{g}_{kn} = (\hat{g}_{in} + \hat{g}_{jn})/2$. In our method, $\hat{g}_{kn}$ is computed from (9). As a result, our method guarantees that the average time to coalescence for haplotypes drawn from populations belonging to the same node is less than the average time to coalescence for a haplotype drawn from any of those populations and from a population belonging to a different node.

In addition to producing a tree that describes the hierarchical structure, it is possible to assess the statistical significance of the genetic differentiation detected at each hierarchical level. We show in APPENDIX B that if population haplotype frequencies are known, not esti-

TABLE 4

Divergence between cpDNA haplotypes in *Coreopsis grandiflora*

	A	*A4*	*A15*	*A19*	*B*	*B2*	*B13a*	*B13b*	*B15*	*B16*	*B17a*	*B17b*	*B18*
A		3	1	1	13	19	15	14	14	14	15	16	14
A4	0.090		4	4	16	22	18	17	17	17	18	19	17
A15	0.030	0.123		2	12	18	14	13	13	13	14	15	13
A19	0.030	0.123	0.061		14	20	16	15	15	15	16	17	15
B	0.395	0.487	0.365	0.426		8	2	1	3	1	2	3	1
B2	0.538	0.674	0.552	0.614	0.243		10	9	9	9	10	11	9
B13a	0.456	0.548	0.426	0.488	0.060	0.304		3	5	3	5	5	3
B13b	0.425	0.517	0.395	0.457	0.030	0.273	0.090		4	2	3	4	2
B15a	0.428	0.519	0.397	0.459	0.091	0.275	0.151	0.121		4	5	6	4
B16	0.425	0.518	0.396	0.458	0.030	0.274	0.091	0.060	0.121		1	2	2
B17a	0.456	0.548	0.426	0.488	0.060	0.304	0.121	0.090	0.151	0.030		1	3
B17b	0.489	0.579	0.547	0.519	0.091	0.335	0.151	0.121	0.182	0.060	0.030		4
B18	0.425	0.518	0.396	0.458	0.030	0.274	0.091	0.060	0.121	0.060	0.091	0.121	

Number of restriction site differences (above the diagonal) and percent nucleotide divergence (below the diagonal, calculated according to NEI and TAJIMA 1981). From MASON-GAMER *et al.* (1995).

mated, $g_{ij} \geq 0$ with equality if and only if the haplotype frequencies in i and j are identical, as would be expected from its close relationship with F_{ST} and G_{ST}. This provides a simple test for the presence of genetic differentiation between two nodes. Let

$$\hat{\Omega} = \sum_k (\hat{x}_k^{(i)} - \hat{x}_k^{(j)})^2, \qquad (13)$$

where $\hat{x}_k^{(i)}$ is the sample frequency of haplotype k in node i and $\hat{x}_k^{(j)}$ is the sample frequency of haplotype k in node j. Then $g_{ij} = 0$, if and only if $\Omega = 0$. To determine the distribution of Ω under the null hypothesis we construct a large number of random samples in which haplotypes are assigned randomly to populations in each node in proportion to their average frequency in the combined sample. We calculate Ω from each sample, producing a null distribution to which we compare our sample $\hat{\Omega}$ (*cf.* LONG 1986; WEIR and COCKERHAM 1984; ZHIVOTOVSKY 1988).

AN EXAMPLE WITH CPDNA HAPLOTYPES

While restriction site and nucleotide sequence variation in chloroplast DNA (cpDNA) has been used extensively at the interspecific level and above, (*e.g.*, PALMER *et al.* 1988), it has been less widely used for intraspecific studies because of its low rate of sequence evolution (PALMER 1985, 1987; WOLFE *et al.* 1987; BIRKY 1988; CLEGG *et al.* 1990). Several recent studies have shown, however, that the amount of intraspecific cpDNA variation is greater than previously thought (reviewed in HARRIS and INGRAM 1991; SOLTIS *et al.* 1992). We illustrate here the techniques described above can be used to provide new insight into patterns of population differentiation by applying them to data derived from an extensive restriction site survey of *C. grandiflora*, a morphologically variable member of the sunflower family found through much of the southern United States.

The data: MASON-GAMER *et al.* (1995) provide complete details on the sampling and laboratory procedures. Briefly, the sample consists of ~20 individuals from each of 14 populations in Georgia and Arkansas, representing both varieties found in Georgia (var. *grandiflora* and var. *saxicola*) and all three varieties found in Arkansas (var. *grandiflora*, var. *saxicola*, and var. *harveyana*). After extracting total DNA with a simple modification of standard procedures (SAGHAI-MAROOF *et al.* 1984; DOYLE and DOYLE 1987), samples were digested with eight restriction enzymes, each of which cuts cpDNA frequently: *Alu*I (AC/GT), *Hae*III (GG/CC), *Hha*I (GCG/C), *Hinf*I (G/ANTC), *Mbo*I (/GATC), *Msp*I (C/CGG), *Rsa*I (GT/AC), and *Taq*I (T/CGA). The resulting fragments were separated on 1.25–1.5% agarose gels and bidirectionally blotted (SMITH and SUMMERS 1980) to reusable nylon membranes. Membrane-bound DNA fragments were hybridized with eight ^{32}P-labeled, cloned fragments of the lettuce chloroplast genome (JANSEN and PALMER 1987) that an earlier study of *Krigia* (another member of the sunflower family) had suggested are the most variable regions of the genome (KIM *et al.* 1992).

Among the 273 cpDNAs assayed from these 14 populations, 33 of the 427 restriction sites detected were polymorphic, and we detected 13 distinct haplotypes (Table 4). The haplotypes may differ at only one site (0.030% sequence divergence) or at as many as 22 restriction sites (0.674% sequence divergence). The 14 sampled populations show a wide range of population structures (Table 5), from those in which only a single haplotype was detected to those in which as many as four haplotypes were detected. Not only do populations appear to be quite different from one another, haplotypes are unevenly distributed among populations. The *A* haplotype, for example, is found in eight populations while several others are found in only one.

TABLE 5

Haplotypic composition of the population samples

Sample no.	Sample size	Haplotypes present	Haplotype frequencies
Georgia			
1	21	*A*	1.000
2	18	*A*	0.500
		*B*2	0.500
3	17	*A*	1.000
4	20	*A*	0.850
		*A*4	0.150
8	23	*A*	1.000
9/10	20	*A*	1.000
Arkansas			
13	20	*B*	0.700
		*B*13*a*	0.250
		*B*13*b*	0.050
14	20	*B*13*a*	1.000
15	31	*A*	0.097
		*A*15	0.194
		*B*13*a*	0.258
		*B*15	0.452
16	21	*B*	0.049
		*B*16	0.290
17	19	*B*	0.895
		*B*17*a*	0.053
		*B*17*b*	0.053
18		*B*	0.500
		*B*18	0.500
19	19	*A*	0.632
		*A*19	0.368
20	20	*A*	0.400
		*A*19	0.600

From MASON-GAMER *et al.* (1995).

Results: The results of our hierarchical analysis of haplotype diversity in this sample are presented in Figure 1. Three major groups of populations are evident: those in which only *A*-type genomes are present, those in which only *B*-type genomes are present, and those in which both *A*-type and *B*-type genomes are present. Within the *A*-type genome group, the populations are further divided into western (Arkansas) and eastern (Georgia) populations. The *B*-type genome group is found only in Arkansas. The only *B*-type genome found in Georgia, *B*2, is found in a population that also contains haplotype *A*. The *B*2 genome is also unusual in that it is highly divergent from other *B*-type genomes and is found nowhere else in the sample (Tables 4 and 5; MASON-GAMER *et al.* 1995). MASON-GAMER *et al.* (1995) noted that *A*-type genomes are found both in other members of Coreopsis sect. Coreopsis, the section to which *C. grandiflora* belongs, and in species belonging to other sections of Coreopsis. Similarly, *B*-type genomes are found both in other members of sect. Coreopsis and in members of other sections. The pattern of genome distribution among taxa suggests that the divergence between the two genome groups predates the origin of *C. grandiflora*. If so, then each population having both *A*-type and *B*-type genomes actually represents two populations with respect to cpDNA haplotype, because there is no evidence of cpDNA recombination in natural populations.

In our sample, only two populations (population 2 from Georgia and population 15 from Arkansas) have both *A*-type and *B*-type genomes present. Figure 2 presents the results of a hierarchical analysis of haplotype diversity in which population 2 and population 15 are each treated as two populations: once composed entirely of *A*-type haplotypes and one composed entirely of *B*-type haplotypes (populations 2A/2B and 15A/15B, respectively). The major difference between the relationships as depicted here and in Figure 1 is, not surprisingly, that there are two major groups of populations: those in which only *A*-type genomes are present and those in which only *B*-type genomes are present. In fact, >90% of the total nucleotide diversity present in the sample is a result of the divergence between *A*-type and *B*-type genomes. With respect to maternal phylogeny, therefore, *A*-type genome populations in Arkansas are more closely related to *A*-type genome populations in Georgia than they are to *B*-type genome populations in Arkansas. Similarly, *B*-type genome populations in Arkansas share a more recent common maternal ancestor with population 2B in Georgia than they do with *A*-type genome populations in Arkansas.

In spite of the predominant role of genome type in structuring nucleotide diversity in *C. grandiflora*, Figure 2 also makes it apparent that both the *A*-type and *B*-type genomes found in population 15 (Arkansas) are more similar to other genomes of their type in Arkansas than to other genomes of their type in Georgia. Similarly, population 2A is more similar to other *A*-type genome populations from Georgia than it is to any *A*-type genome populations from Arkansas. In short, this analysis suggests that the geographical separation of populations in Georgia and Arkansas has been accompanied by significant genetic differentiation, a pattern that was not apparent in earlier analyses of electrophoretic variation (CRAWFORD and SMITH 1984; COSNER and CRAWFORD 1990). Comparable restriction site or nucleotide sequence data on nuclear encoded genes will be required before we can determine if the differences observed between these sets of data reflect the higher level of population subdivision expected for cpDNA markers than for nuclear markers because of its predominantly maternal transmission (BIRKY *et al.* 1983, 1989) or the influence of hybridization and introgression events (*e.g.*, RIESEBERG 1991; WHITTEMORE and SCHAAL 1991; reviewed in RIESEBERG and BRUNSFELD 1992). It appears, however, that *A*-type genome populations are more closely connected with others in that genome group than with any in the *B*-genome group, even though both *A*-type and *B*-type populations occur in both geographic regions and there is significant genetic differentiation between them.

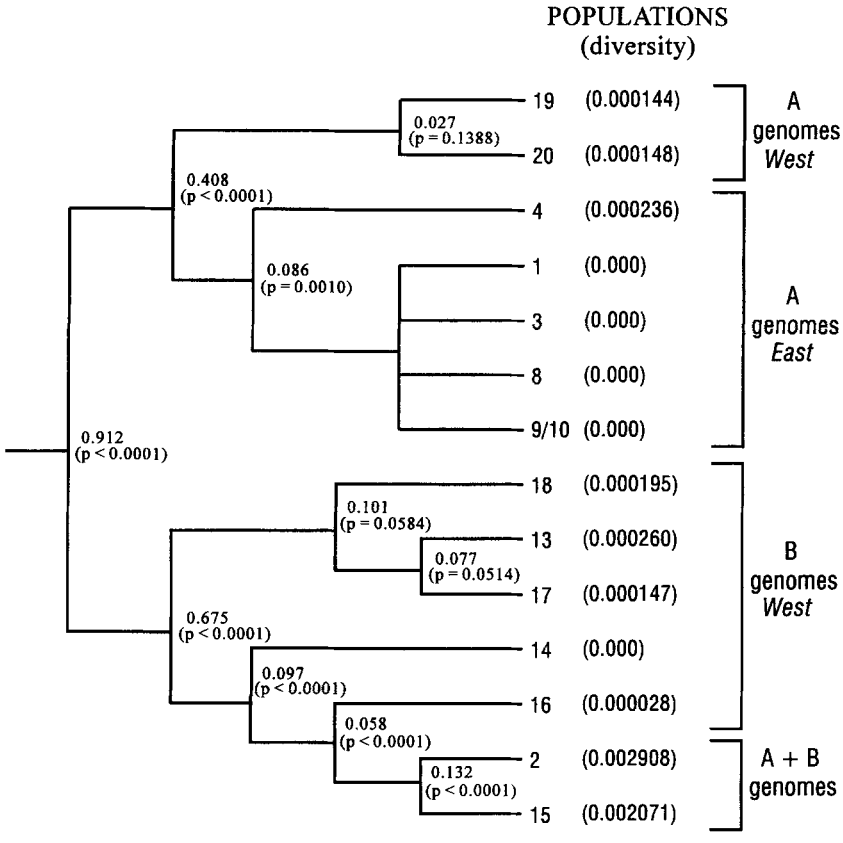

FIGURE 1.—Hierarchical analysis of haplotype diversity in *Coreopsis grandiflora*. Population numbers correspond to those in Table 5, and the numbers in parentheses following them are the nucleotide sequence diversity within each population, estimated following NEI and TAJIMA (1981). The number at each node is the distance (g_{ij}) between its two daughter nodes, and the *P*-value reported is the probability of obtaining a distance that great or greater under the null hypothesis of no differentiation between the daughter nodes (based on 10,000 random resamplings).

DISCUSSION

The method we introduced here for studying patterns of haplotype diversity within and among populations is similar to several existing procedures. Like Φ statistics calculated from a distance matrix in which the distance between two haplotypes is a simple count of the differences between them (EXCOFFIER *et al.* 1992), the measure of among population differentiation we propose, $\hat{g}_{st}$, does not correct for multiple substitutions. It differs in this respect from the similar measures proposed by TAKAHATA and PALUMBI (1985), LYNCH and CREASE (1990), and NEI and MILLER (1990). At least when levels of divergence are less than a few percent, however, $\hat{g}_{st}$ provides more accurate estimates of F_{ST} than other statistics that have been proposed for this purpose. Like all of these measures, $\hat{g}_{st}$ allows us to partition the observed diversity into within and among population components (*cf.* Equation 9). Unlike other methods that have been proposed, however, any hierarchical structure present in the data emerges naturally from the algorithm we suggest for pairwise calculations of $\hat{g}_{st}$. The hierarchical structure need not be imposed prior to the analysis. We regard this as the most novel and useful feature of our method.

In circumstances where multiple substitutions are more frequent, our method may still be employed as a way to summarize the hierarchical structure of the data. Any hierarchical structure revealed, however, should not be interpreted in terms of mean coalescence times. In fact, when dealing with data of this type, it may be useful to consider other methods of calculating within and among population diversity. By interpreting δ_{ij} as the number of steps between haplotypes i and j on a minimum-spanning tree connecting all haplotypes, for example, the resulting tree will connect populations whose haplotypes are, on average, separated by fewer mutational events than those belonging to different nodes (*cf.* EXCOFFIER *et al.* 1992). Alternatively, the evolutionary distance between haplotypes i and j could be calculated using formulas for restriction sites or nucleotide sequences that correct for multiple substitutions, base composition biases, among-site rate variation, or mutational biases. In either of these cases, the hierarchical analysis we propose would be done directly on the ν and ν_k, using $\hat{g}'_{st} = (\hat{\nu} - \bar{\hat{\nu}})/\hat{\nu}$ as the measure of among population differentiation. Because $\hat{g}_{st}$ depends on the ratio of $\bar{\pi}$ to π and $\bar{\nu}$ and ν differ from these only by a constant for low levels of sequence divergence, the results of an analysis on $\hat{g}'_{st}$ will differ substantially from those of an analysis on $\hat{g}_{st}$ only when some of the haplotypes are very divergent from one another. Similarly, data from electrophoretic surveys could be used to uncover hierarchical structure simply by using NEI's G_{ST} statistics (NEI 1973; NEI and CHESSER 1983) in pairwise comparisons instead of $\hat{g}_{st}$. In short, the method we describe here provides a flexible framework for the analysis of genetic variation in spatially structured populations, a framework that is particularly appropriate

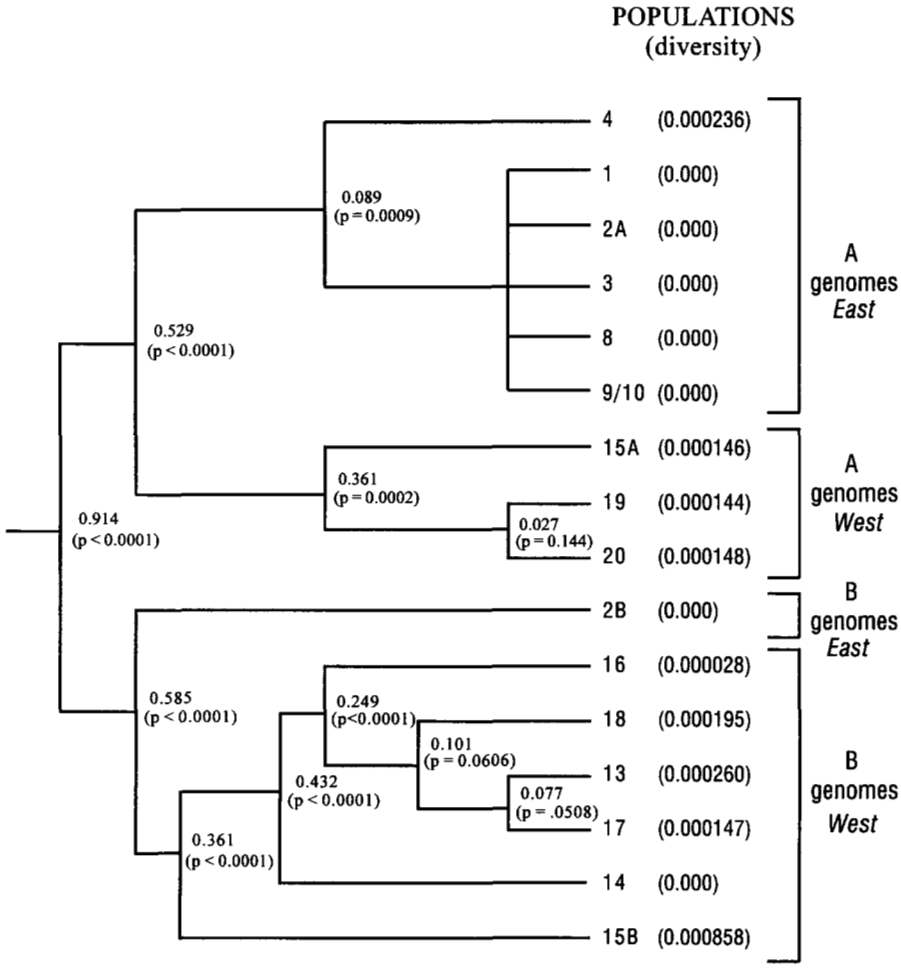

FIGURE 2.—Hierarchical analysis of haplotype diversity in *Coreopsis grandiflora*, treating the *A* haplotypes and *B* haplotypes in populations 2 and 15 as separate populations (2a and 2b, 15a and 15b). See the caption in Figure 1 for an explanation of the statistics reported on the tree.

when interest is centered on discovering hierarchical structure that is present in the data rather than on determining whether the data fits a preconceived notion of what that structure should be.

This work was supported by a grant from the University of Connecticut Research Foundation to K.E.H. and by a Grant-in-Aid of Research from Sigma Xi and a National Science Foundation Doctoral Dissertation Improvement grant (BSR-9105167) to R.J.M.

Note: The computer programs used to evaluate the performance of F_{ST} measures and to analyze restriction site diversity in *C. grandiflora* can be obtained from the senior author on request. They are available either as source code (ANSI C) or as a precompiled executable (386 MS-DOS).

LITERATURE CITED

BIRKY, C. W., 1988 Evolution and variation in plant chloroplast and mitochondrial genomes, pp. 23–53, in *Plant Evolutionary Biology*, edited by L. D. GOTTLIEB and S. K. JAIN. Chapman and Hall, New York.

BIRKY, C. W., P. FUERST and T. MARUYAMA, 1989 Organelle gene diversity under migration and drift: equilibrium expectations, approach to equilibrium, effects of heteroplasmic cells, and comparison to nuclear genes. Genetics **121:** 613–627.

BIRKY, C. W., T. MARUYAMA and P. FUERST, 1983 An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics **103:** 513–527.

CLEGG, M. T., G. H. LEARN and E. M. GOLENBERG, 1990 Evolution of chloroplast DNA. pp. 135–149, in *Evolution at the Molecular Level*, edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer, Sunderland, MA.

COCKERHAM, C. C., 1969 Variance of gene frequencies. Evolution **23:** 72–84.

COCKERHAM, C. C., 1973 Analyses of gene frequencies. Genetics **74:** 679–700.

COSNER, M. B., and D. J. CRAWFORD, 1990 Allozyme variation in *Coreopsis* sect. *Coreopsis* (Asteraceae). Syst. Bot. **15:** 256–265.

CRAWFORD, D. J., and E. B. SMITH, 1984 Allozyme divergence and interspecific variation in *Coreopsis grandiflora* (Compositae). Syst. Bot. **9:** 219–225.

DOYLE, J. J., and J. L. DOYLE, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19:** 11–15.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

FELSENSTEIN, J., 1982 How can we infer geography and history from gene frequencies? J. Theor. Biol. **96:** 9–20.

HAMRICK, J. L., and M. J. W. GODT, 1990 Allozyme diversity in plant species. pp. 43–63, in *Plant Population Genetics, Breeding, and Genetic Resources*, edited by A. H. D. BROWN, M. T. CLEGG, A. T. KAHLER and B. S. WEIR. Sinauer Associates, Sunderland, MA.

HARRIS, S. A., and R. INGRAM, 1991 Chloroplast DNA and biosystematics: the effects of intraspecific diversity and plastid transmission. Taxon **40:** 393–412.

JANSEN, R. K., and J. D. PALMER, 1987 Chloroplast DNA from lettuce and *Barnadesia* (Asteraceae): structure, gene localization, and characterization of a large inversion. Curr. Genet. **11:** 553–564.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein mole-

cules. pp. 21–132, in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KIM, K.-J., R. K. JANSEN and B. L. TURNER, 1992 Phylogenetic and evolutionary implications of interspecific chloroplast DNA variation in dwarf dandelions (*Krigia;* Asteraceae). Amer. J. Bot. **79:** 708–715.

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

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

LONG, J. C., 1986 The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's *F*-statistics. Genetics **112:** 629–647.

LYNCH, M., and T. J. CREASE, 1990 The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7:** 377–394.

MASON-GAMER, R. J., K. E. HOLSINGER and R. K. JANSEN, 1995 Chloroplast DNA haplotype variation within and among populations of *Coreopsis grandiflora* (Asteraceae). Mol. Biol. Evol. **12:** 371–381.

MERCURE, A., K. RALLS, K. P. KOEPFLI and R. K. WAYNE, 1993 Genetic subdivisions among small canids: mitochondrial DNA differentiation of swift, kit, and arctic foxes. Evolution **47:** 1313–1328.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3321.

NEI, M., 1982 Evolution of human races at the gene level. pp. 167–181, in *Human Genetics, Part A: The Unfolding Genome, Proceedings of the Sixth International Congress of Human Genetics,* edited by B. BONNÉ-TAMIR, T. COHEN and R. M. GOODMAN. Alan R. Liss, Inc., New York.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia Univ. Press, New York.

NEI, M., and R. K. CHESSER, 1983 Estimation of fixation indices and gene diversities. Ann. Hum. Genet. **47:** 253–259.

NEI, M., and J. C. MILLER, 1990 A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics **125:** 873–879.

NEI, M., and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

NEVO, E., A. BEILES and R. BEN-SHLOMO, 1984 The evolutionary significance of genetic diversity: ecological, demographic, and life history correlates. pp. 13–213, in *Evolutionary Dynamics of Genetic Diversity,* edited by G. S. MANI. Springer-Verlag, Berlin.

NOTOHARA, M., 1990 The coalescent and the genealogical process in geographically structured population. J. Math. Biol. **29:** 59–75.

PALMER, J. D., 1985 Evolution of chloroplast and mitochondrial DNA in plants and algae. pp. 131–140, in *Monographs in Evolutionary Biology: Molecular Evolutionary Genetics,* edited by R. J. MACINTYRE. Plenum, New York.

PALMER, J. D., 1987 Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. Am. Nat. **130:** S6–S29.

PALMER, J. D., R. K. JANSEN, H. J. MICHAELS, M. W. CHASE and J. W. MANHART, 1988 Chloroplast DNA variation and plant phylogeny. Ann. MO Bot. Gard. **75:** 1180–1206.

PIAZZA, A., and P. MENOZZI, 1983 Geographic variation in gene frequencies. pp. 444–450, in *Numerical Taxonomy,* NATO ASI series, Vol. 61, edited by J. FELSENSTEIN. Springer-Verlag, Heidelberg.

RIESEBERG, L. H., 1991 Homoploid reticulate evolution in *Helianthus* (Asteraceae): evidence from ribosomal genes. Am. J. Bot. **78:** 1218–1237.

RIESEBERG, L. H., and S. J. BRUNSFELD, 1992 Molecular evidence and plant introgression. pp. 151–176, in *Molecular Systematics of Plants,* edited by P. S. SOLTIS, D. E. SOLTIS and J. J. DOYLE. Chapman and Hall, New York.

SAGHAI-MAROOF, M. A., K. M. SOLIMAN, R. A. JORGENSEN and R. W. ALLARD, 1984 Ribosomal DNA spacer length polymorphism in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. USA **81:** 8014–8018.

SLATKIN, M., 1991 Inbreeding coefficients and coalescence times. Genet. Res. **58:** 167–175.

SLATKIN, M., 1993 Isolation by distance in equilibrium and nonequilibrium populations. Evolution **47:** 264–279.

SMITH, G. E., and M. D. SUMMERS, 1980 The bidirectional transfer of DNA and RNA to nitrocellulose or diazobenzyloxymethyl paper. Anal. Biochem. **109:** 123–129.

SNEATH, P. H. A., and R. R. SOKAL, 1973 *Numerical Taxonomy.* W. H. Freeman, San Francisco, CA.

SOLTIS, D. E., P. S. SOLTIS and B. G. MILLIGAN, 1992 Intraspecific chloroplast DNA variation: systematic and phylogenetic implications. pp. 117–150, in *Molecular Systematics of Plants,* edited by P. S. SOLTIS, D. E. SOLTIS and J. J. DOYLE. Chapman and Hall, New York.

STUART, A., and J. K. ORD, 1987 *Kendall's Advanced Theory of Statistics,* Volume 1. Oxford Univ. Press, New York.

TAKAHATA, N., and S. R. PALUMBI, 1985 Extranuclear differentiation and gene flow in the finite island model. Genetics **109:** 441–457.

WEIR, B. S., and C. J. BASTEN, 1990 Sampling strategies for distances between dna sequences. Biometrics **46:** 551–582.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WHITTEMORE, A. T., and B. A. SCHAAL, 1991 Interspecific gene flow in sympatric oaks. Proc. Natl. Acad. Sci. USA **88:** 2540–2544.

WOLFE, K. H., W.-H. LI and P. M. SHARPE, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84:** 9054–9058.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **15:** 323–354.

WRIGHT, S., 1965 The interpretation of population structure by *F*-statistics with special regards to systems of mating. Evolution **19:** 395–420.

ZHIVOTOVSKY, L. A., 1988 Some methods of analysis of correlated characters. pp. 423–432, in *Proceedings of the II International Conference on Quantitative Genetics,* edited by B. S. WEIR, G. EISEN, M. M. GOODMAN and G. NAMKOONG. Sinauer Assoc., Sunderland, MA.

Communicating editor: B. S. WEIR

APPENDIX A

NOTOHARA (1990) extends KINGMAN's coalescent (1982a,b) to geographically structured populations with arbitrary population sizes and arbitrary migration rates among populations. Let m_{ij} be the proportion of individuals in population i that are immigrants from population j. Let q_{ij} be the proportion of individuals in population i that move to population j. m_{ij} is the backward migration rate, and q_{ij} is the forward migration rate. They are related as

$$m_{ij} = \begin{cases} \dfrac{N_j}{N_i} q_{ji} & \text{if} \quad k \neq j \\ -\dfrac{1}{N_i}\left(\sum k \neq i N_k q_{ki}\right) & \text{if} \quad k = j \end{cases} \tag{A1}$$

where N_k is the size of population k. In this formulation, the effective population size is assumed to equal the census population size. Let n_k be then number of haplotypes in the sample from population k.

Looking backward over the genealogy of the sampled alleles, at each generation one of two events may occur: coalescence or migration. Let α_k be the probability of a coalescent event in population k, and let β_{ij} be the probability that an allele in population i was in population j in the preceding generation. If we assume that populations are large enough and migration rates are small enough that the probability of two events happening simultaneously is negligible, then

$$\alpha_k = \frac{n_k(n_k - 1)}{4N_k} \tag{A2}$$

and

$$\beta_{ij} = n_i m_{ij} \quad (i \neq j). \tag{A3}$$

Furthermore, the probability that neither a coalescence nor a migration event occurs is

$$\phi = 1 - \sum_k \left(\alpha_k - \sum_{i \neq k} \beta_{ik} \right). \quad \text{(A4)}$$

Thus, the time back to the first event is exponentially distributed with mean $1/\phi$. Let $\alpha = \Sigma_k \alpha_k$ and $\beta = \Sigma_k \Sigma_{i \neq k} \beta_{ik}$. Then the probability that the first event is a coalescent event is

$$p_c = \frac{\alpha}{\alpha + \beta}, \quad \text{(A5)}$$

and the probability that the first event is a migration event is

$$p_m = \frac{\beta}{\alpha + \beta}. \quad \text{(A6)}$$

To construct the coalescent structure of a sample we use the following algorithm:

(1) Calculate all α_k and β_{ij}, ϕ, α, β, p_c, and p_m for the current sample configuration.
(2) Select the time for the next event at random from an exponential distribution with mean $1/\phi$.
(3) Select a random number, u, from a uniform distribution on the interval (0, 1). If $u < p_c$ as given by (A.5), then go to (4), otherwise go to (5).
(4) The next event is a coalescent event. Select a random number, u, on the interval (0, 1). Let K be the largest integer k such that the inequality

$$\frac{\Sigma_i^k \alpha_k}{\alpha} < u$$

is satisfied.

(a) The coalescent event occurred in population K. Select two haplotypes at random from this population and do the coalescence.
(b) Count total number of haplotypes remaining after the coalescence event. If there is only one, the coalescence structure is complete. If there is more than one, return to (1).

(5) The next event is a migration event. Select a random number, u, on the interval (0, 1). Let M be the largest integer m such that the inequality

$$\frac{\Sigma_i^m \Sigma_{i \neq m} \beta_{im}}{\beta} < u$$

is satisfied.

(a) The haplotype migrated into population M. Select a random number, u, on the interval (0, 1). Let N be the largest integer n such that the inequality

$$\frac{\Sigma_i^n \Sigma_{i \neq M} \beta_{iM}}{\Sigma_{i \neq M} \beta_{iM}}$$

is satisfied.

(b) The haplotype migrated from population N. Select a haplotype at random from population M and move it to population N. Return to (1).

APPENDIX B

It is apparent from (A.9) that $g_{ij} = 0$ if and only if $f_{ij} - (pf_i + qf_j) = 0$. Letting π_{ij} represent the nucleotide diversity in the combined sample, $\bar{\pi}$ the average within population diversity in the combined sample, and π_i the diversity in node i, it is a matter of algebra to show that

$$f_{ij} - (pf_i + qf_j) = \frac{\bar{\nu}}{\nu_i \nu_j \nu_{ij}} [\nu_{ij} - (p\nu_i + q\nu_j)]$$

$$= \frac{\bar{\nu}}{\nu_i \nu_j \nu_{ij}} \left[-pq \sum_{kl} (x_k^{(i)} - x_k^{(j)})(x_l^{(i)} - x_l^{(j)}) \delta_{kl} \right]. \quad \text{(B1)}$$

The partial derivative of the term in square brackets with respect to $x_l^{(j)}$ is

$$2pq \sum_k (x_k^{(i)} - x_k^{(j)}) \delta_{kl}. \quad \text{(B2)}$$

If $\mathbf{y}$ is the vector whose components are $x_k^{(i)} - x_k^{(j)}$ and Δ is the matrix of δ_{kl}, then the condition for $x_k^{(j)}$ to be a critical point of $f_{ij} - (pf_i + qf_j)$ is

$$\Delta \mathbf{y} = 0,$$

which requires either $|\Delta| = 0$ or $\mathbf{y} = 0$. But $\delta_{kl} > 0$ for $i \neq j$ and $\delta_{kk} = 0$ whenever each haplotype is distinct, which guarantees $|\Delta| \neq 0$. Because $\mathbf{y} = 0$ if and only if $x_k^{(i)} = x_k^{(j)}$ for all k, $x_k^{(i)} = x_k^{(j)}$ is a unique critical point for $f_{ij} - (pf_i + qf_j)$.

It is evident from (B.2) that $x_k^{(i)} = x_k^{(j)}$ is a minimum for each k individually. Because the x_k form a complete basis for the space, $x_k^{(i)} = x_k^{(j)}$ for all k is also a minimum for (B.1). Thus,

$$g_{ij} \geq 0 \quad \text{(B3)}$$

with equality only when $x_k^{(i)} = x_k^{(j)}$ for all k.