Empirical Tests of Some Predictions From Coalescent Theory With Applications to Intraspecific Phylogeny Reconstruction

Keith A. Crandall*,† and Alan R. Templeton*

**Department of Biology and †Department of Mathematics, Washington University, St. Louis, Missouri 63130-4899*

Manuscript received September 15, 1992
Accepted for publication March 18, 1993

ABSTRACT

Empirical data sets of intraspecific restriction site polymorphism in Drosophila have been gathered in order to test hypotheses derived from coalescent theory. Three main ideas are tested: (1) haplotype frequency in the sample contains information on the topological position of a given haplotype in a cladogram, (2) the frequency of a haplotype is related to the number of mutational connections to other haplotypes in the cladogram and (3) geographic location can be used to infer topological positioning of haplotypes in a cladogram. These relationships can then be used to better estimate intraspecific phylogenies in two ways: (1) rooting the phylogeny and (2) resolving ambiguities in a cladogram. This information will allow one to reduce the number of alternative phylogenies and incorporate the uncertainties involved in reconstructing intraspecific phylogenies into subsequent analyses that depend heavily on the topology of the tree.

RECENTLY, there has been interest in estimating phylogenies within-species for purposes as diverse as the calculation of migration rates (SLATKIN and MADDISON 1989) and the examination of associations between phenotype and biochemical/physiological functions (TEMPLETON, BOERWINKLE and SING 1987). Yet the theoretical tools for intraspecific phylogeny reconstruction are scarce. Most empirical analyses rely on methods designed for estimating between-species phylogenies. Subsequent analyses based on the phylogenetic relationships can be handicapped by the lack of resolving power of between-species methods. Here we test hypotheses relating to haplotype frequency and the topological relationship of haplotypes in order to refine estimates of intraspecific phylogenies. To do this, we utilize ideas from coalescent theory.

Coalescent theory was formally put forth in two seminal papers by KINGMAN (1982a,b). In these papers, KINGMAN derives a mathematical theory that describes the genealogical process of a sample of selectively neutral genes from a population, looking backward in time. This work has generated much subsequent discussion of the properties and applications of coalescent theory [for reviews, see TAVARÉ (1984), EWENS (1990) and HUDSON (1990)]. Coincident with the development of coalescent theory, there has been an increase in within-species molecular data sets resulting from the development and availability of recombinant DNA technologies. For example, KREITMAN's (1983) classic paper gave DNA sequence variants at the population level. While within-species sequence data are still scarce, restriction site data are abundant (*e.g.*, see Table 2). These studies and others have described levels of genetic variation in Drosophila at a number of loci. This study utilizes this wealth of empirical data to test some predictions of coalescent theory.

Three main questions will be addressed. (1) Can we use haplotype frequency information to refine the topological position of a given haplotype in a cladogram? (2) Is the frequency of a haplotype related to the number of mutational connections to other haplotypes within the cladogram? (3) Can geographic location be used to infer the topological position of a rare haplotype in a cladogram? Our results, in addition to other predictions from population genetic theory, have implications in allelic phylogeny reconstruction in two major areas: (1) in rooting the phylogeny and (2) in resolving ambiguities in the phylogeny. The predictions discussed above and the *a posteriori* probabilities that we establish can be used as *a priori* probabilities in these two areas of intraspecific phylogeny reconstruction. This will allow one to reduce the number of alternative phylogenies and incorporate the uncertainties involved in reconstructing intraspecific phylogenies into subsequent analyses which depend heavily on the topology of the tree.

THEORY

To test predictions of coalescent theory, we first describe the general theory of the n-coalescent with mutation. We then derive certain expectations relating to this process. Consider a Fisher-Wright model of evolution for a sample of n neutral genes from a population (by n genes we mean n samples from a

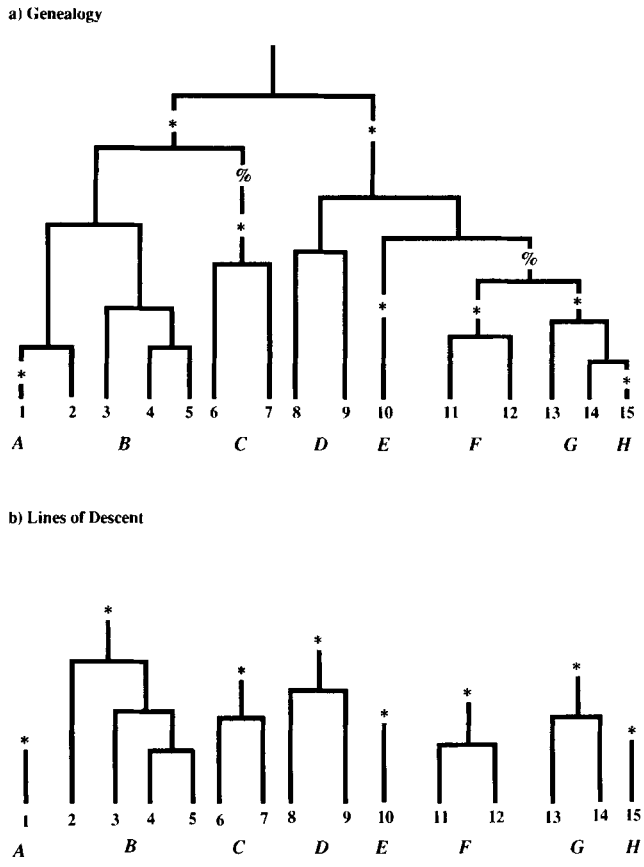


FIGURE 1.—The partitioning of coalescent events: (a) the genealogy and (b) the lines of descent. * represents a defining mutation and % represents previous mutations that are not defining events but are important in establishing the allelic phylogeny. Genes are represented by the numbers 1–15 and alleles (or haplotypes) are represented by the letters A–H.

particular gene region). Assume that $n \ll 2N$ where N is the inbreeding effective size of the population and $N \gg 1$. These assumptions imply that the probability $[1/(2N)]$ of any particular pair of genes coalescing in a given generation is very small. In particular, multiple coalescent events and multiple mutational events in the same generation in the ancestry of a sample of n genes can be ignored (KINGMAN 1982b).

Under this model, there are two possible defining events in the coalescent process: either a coalescence or a mutation. By a defining event, we mean either an event that results in a reduction in the number of genes looking back in time, *i.e.*, a coalescence, or a reduction in the number of allelic states looking back in time, *i.e.*, the first mutation going back in time. Throughout this paper we use the terms allele, allelic class, and haplotype synonymously to mean a set of $\leq n$ genes within which there is no observable variation. Figure 1a gives an example of genes (labeled 1–15) and the defining events in their ancestors. Figure 1b shows the lines of descent of the allelic classes (labeled A–H) and their originating mutations. Any mutation is assumed to produce a completely novel haplotype (*i.e.*, the infinite alleles model) as is reasonable in many biological situations. Under coalescent theory (Figure 2) (KINGMAN 1982a), the expectation of the number of mutations given n defining events is

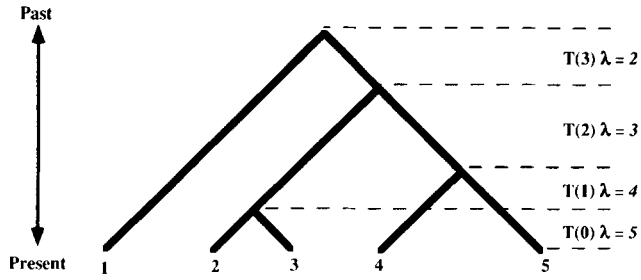


FIGURE 2.—Time segments as defined by the coalescent process. T represents the time interval looking back in time and λ gives the number of linages at that time interval (after HUDSON 1990).

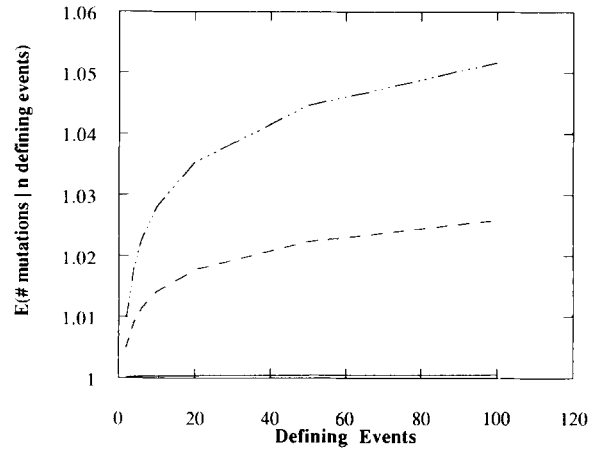


FIGURE 3.—The expected number of mutations given n defining events for three values of θ. See text, particularly Equation 5 for details. $\theta = 0.01$ (- - - -); $\theta = 0.005$ (- - -); $\theta = 0.0001$ (——).

$$E(\text{no. mutations} | n \text{ defining events}) = \sum_{\lambda=1}^{n} \frac{\theta}{\lambda + \theta - 1} = \theta \sum_{j=0}^{n-1} \frac{1}{\theta + j}. \quad (1)$$

Note that Equation 1 also gives the expected number of alleles in a sample of n genes.

From Equation 1 we see that as n increases the expected number of alleles increases; but as seen in Figure 3, this tends to plateau rapidly, especially with small values of θ. This result has important implications for sampling strategies for allelic phylogenies. It suggests that for extremely low values of θ, increasing sample size in a panmictic population will have little affect on the number of haplotypes found in the sample. For larger values of θ, the cost of sample size outruns the returns in unique haplotypes between 50 and 100 defining events (individuals) for biologically reasonable values of the parameters.

DONNELLY and TAVARÉ (1986) have developed a coalescent model which gives the distribution of class sizes for neutral alleles when ordered in increasing age, as well as a characterization of the ages of the

alleles. The results give specific relationships between ages of alleles and allele frequencies. Two main results follow from their discussion which are of particular interest. The first is the well established relationship (WATTERSON 1976; KELLY 1977; WATTERSON and GUESS 1977; DONNELLY and TAVARÉ 1986) that the probability that an allele represented n_i times in a sample of size n is the oldest allele in the sample is n_i/n,

$$\text{P (allele } i \text{ is the oldest allele)} = n_i/n. \quad (2)$$

A second result of DONNELLY and TAVARÉ (1986) is that the expected rank of alleles by age is equal to the rank of alleles by frequency; that is,

$$\text{E (rank in age of allele } i) = C(n_i/n), \quad (3)$$

where C is a normalizing constant depending on the number of alleles. An interesting note by TAVARÉ (1984; due to KELLY 1977) is that the absolute age of the oldest allele is independent of its frequency. We will now show how these results can lead to a more probabilistic framework from which one can refine estimates of intraspecific phylogenies.

MATERIALS AND METHODS

To examine some predictions from the coalescent theory described above, we will define our hypotheses within a phylogenetic framework. Let us first consider a genealogy (Figure 1a) of $n = 15$ (labeled 1–15) genes and $i = 8$ (labeled A–H) alleles. The first mutations looking backward in time in a line of descent are designated by an (*) and subsequent mutational events are shown as (%). This genealogy can be partitioned into two components. Figure 1b shows the first partition representing the lines of descent. This division gives information on the number of alleles, allele frequencies, and relative allele ages. What remains is the allelic phylogeny as defined by mutations (% in Figure 1a) which occur prior to defining events. The allelic phylogeny, Figure 4a, defines the evolutionary relationships between allelic classes. The data used to reconstruct this tree are given in Table 1. Using the methodologies of TEMPLETON, BOERWINKLE and SING (1987) and TEMPLETON, CRANDALL and SING (1992), we can construct a cladogram from the same data set (Figure 4b) to represent the allelic phylogeny in a more explicit way (*i.e.*, by showing explicitly the mutational changes associated with the given cladogram). These cladograms are networks of haplotypes that are interconnected using a Baysian procedure to evaluate the limits of parsimonious connections. Using this Baysian approach, TEMPLETON, CRANDALL and SING (1992) outline the conditions under which one can define a set of plausible cladograms for restriction site or DNA sequence data; that is, those cladograms that include all linkages among haplotypes that have a high cumulative probability (≥ 0.95) of being true. The TEMPLETON, CRANDALL and SING (1992) algorithm allows one to calculate the probability of multiple mutational events within a restriction site shared by two haplotypes that differ by that restriction site. Thus, connections are made only between haplotypes that have a probability of being unique, and which are therefore consistent with the infinite alleles assumption used in the theory given above.

This algorithm also detects haplotypes or gene regions

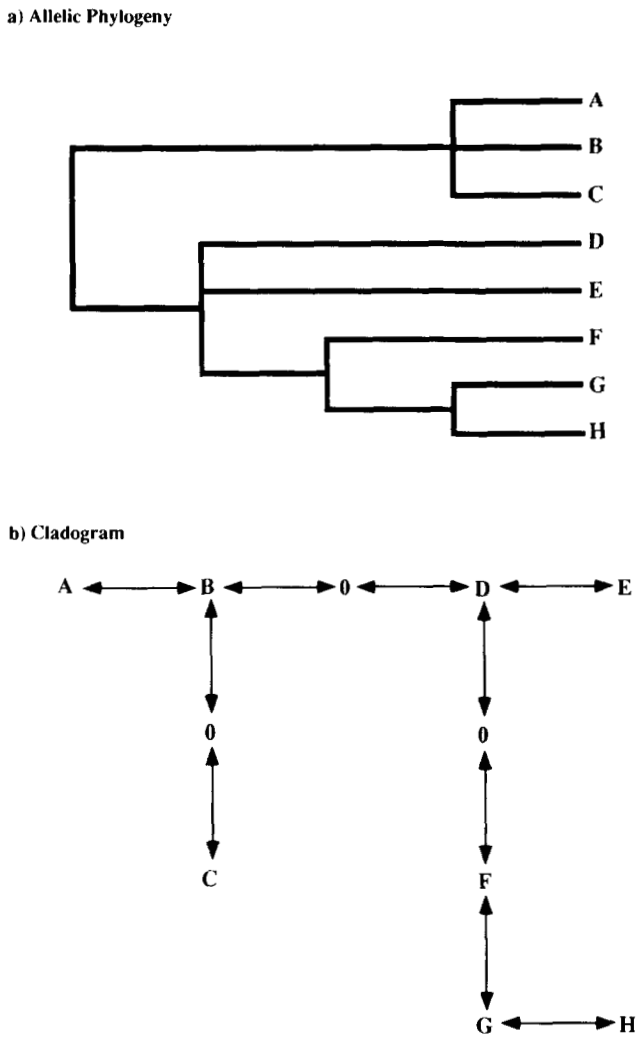


FIGURE 4.—The allelic phylogeny (a) produced by the EXACT SEARCH option in PAUP (SWOFFORD 1991) on the data matrix shown in Table 1. The cladogram (b) produced by the algorithm in TEMPLETON, CRANDALL and SING (1992) from the same data set. Haplotypes are labeled A–H and 0s represent missing intermediates. Both trees represent the genealogical process shown in Figure 1.

that are products of recombination. Recombination is inferred if a single recombinational event can resolve two or more homoplasies or if it can resolve a single homoplasy involving a mutation regarded as evolving in a completely parsimonious fashion (*e.g.*, an insertion/deletion). These criteria were established by AQUADRO *et al.* (1986) but have not been tested theoretically or empirically as to their ability to correctly infer recombination. TEMPLETON, CRANDALL, and SING (1992) have used these criteria with the algorithm of HEIN (1990) to estimate cladograms from gene regions with recombination. The inferred products of recombination are then either eliminated from subsequent analyses or the DNA region is subdivided into subregions in which little to no recombination has occurred. The subsequent tests of predictions from coalescent theory are defined in terms of cladograms estimated by this procedure. It is important to note that the Baysian procedure for evaluating the limits of parsimony and estimating the plausible set of cladograms does *not* incorporate the results from coalescent theory which we will subsequently test based on these cladograms.

TABLE 1

Data matrix for Figure 4

Haplotype	Restriction site profile									
A	1	1	0	0	0	0	0	0	0	0
B	0	1	0	0	0	0	0	0	0	0
C	0	1	1	1	0	0	0	0	0	0
D	0	0	0	0	1	0	0	0	0	0
E	0	0	0	0	1	1	0	0	0	0
F	0	0	0	0	1	0	1	0	0	1
G	0	0	0	0	1	0	1	1	0	1
H	0	0	0	0	1	0	1	1	1	1

Hypothetical data set used to calculate the trees in Figure 4. The data matrix describes the genealogy shown in Figure 1.

Also, to the extent that recombinational events are not detected through this procedure, the predictions of coalescent theory will be weakened because recombination destroys historical associations. Hence, our subsequent tests are conservative with respect to the possibility of undetected recombination.

These tests were performed on 29 data sets of Drosophila restriction site polymorphism collected from the literature. Table 2 gives a summary of the 29 data sets. We have concentrated on the restriction site data from Drosophila, as these data sets are the most abundant and extensive. An example of the resulting cladogram for the restriction site data of the *yellow-achaete-scute* region of *Drosophila melanogaster* (BEECH and BROWN 1989) is given in Figure 5. Often in reconstructing the cladograms for the various data sets, ambiguities exist within the network. In this example (Figure 5), haplotypes 9, 1, 5 and 2 are ambiguously connected. These ambiguous regions of the cladograms were not incorporated into tests of hypotheses used to resolve the ambiguities. Furthermore, information on haplotype frequencies was not incorporated into the cladogram estimation.

Throughout this study we use a variety of nonparametric statistical techniques. We choose to use nonparametric techniques because they make fewer assumptions about the underlying probability distributions from which data are assumed to come. In particular, nonparametric techniques do not assume that the underlying populations are normal, as do traditional statistical tests. Moreover, this collection of data sets is heterogeneous in nucleotide diversity, effective population sizes, migration rates, etc. We, therefore, believe nonparametric techniques are the most appropriate for this diverse compilation of data. While an argument can be made for the independence of these data sets, the assumption of data being identically distributed may be in question because of the heterogeneity of data sets being tested. Here again, nonparametric techniques tend to be more robust to violations of this assumption than standard techniques (HOLLANDER and WOLFE 1973). Where violations of the independent and identically distributed assumptions seem obvious to us, we perform resampling procedures in order to ameliorate problems associated with these violations.

TESTABLE HYPOTHESES AND RESULTS

First, we will explore the relationship between tip and interior haplotypes and haplotype frequency. We define "tip" haplotypes as those haplotypes that have only a single mutational connection to the other haplotypes within a cladogram or network. "Interior" haplotypes are those that have more than one mutational connection. Returning to Figure 4b, haplotypes A, C, E and H can then be classified as tips, while haplotypes B, D, F and G are interior. Nodes labeled "0" are missing intermediate haplotypes. DONNELLY and TAVARÉ (1986) have shown that the rank of alleles by age is proportional to their rank by frequency (*e.g.*, see our Equation 3). Additionally, GOLDING (1987) has pointed out that haplotypes of recent evolutionary origin occur preferentially at the tips of the clado-

TABLE 2

Summary of data sets

Species	Locus[a]	n^b	N^c	U^d	Ref[e]
D. melanogaster	*Adh*	49	30	24	1
D. melanogaster	*Adh*	87	50	36	2
D. melanogaster	*Adh*	18	10	8	3
D. melanogaster	*Adh*	88	32	20	24
D. melanogaster	*y-ac-sc*	105	25	16	4
D. melanogaster	*y-ac-sc*	43	15	13	5
D. melanogaster	*y-ac-sc*	64	24	21	6
D. melanogaster	*G6PD*	113	36	18	7
D. melanogaster	*Mtn*	88	18	15	8
D. melanogaster	*87Ahs*	29	9	7	9
D. melanogaster	*Est-6*	39	28	14	10
D. melanogaster	*rosy*	60	16	9	11
D. melanogaster	*mtDNA*	144	23	19	12
D. melanogaster	*notch*	37	21	15	13
D. melanogaster	*Zeste-tko*	64	27	20	14
D. melanogaster	*Amy*	43	20	16	15
D. simulans	*rosy*	30	23	11	11
D. ananassae	*Om(1D)*	58	41	11	17
D. ananassae	*vermillion*	60	17	17	18
D. ananassae	*forked*	59	41	18	18
D. pseudobscura	*Amy*	28	17	12	16
D. cyrtoloma	*mtDNA*	24	15	11	19
D. heteroneura	*mtDNA*	18	13	8	19
D. silvestris	*mtDNA*	30	23	19	19
D. melanogaster	*white*	38	27	17	20
D. subobscura	*mtDNA*	32	8	8	21
D. sulfurigasta bilimbata	*mtDNA*	56	11	11	22
D. sulfurigasta albostrigata	*mtDNA*	52	6	4	22
D. subobscura	*rp49*	54	23	19	23

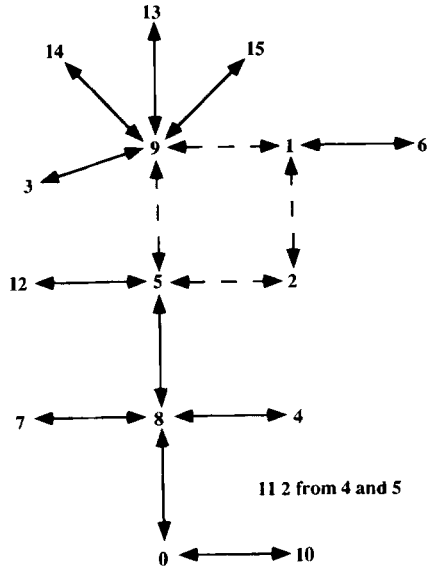
[a] Loci abbreviations: *Adh, alcohol dehydrogenase*; *y-ac-sc, yellow-achaete-scute*; *G6PD, glucose-6-phosphate dehydrogenase*; *Mtn, metallothionein*; *87Ahs, 87A heat shock*; *Est-6, esterase 6*; *mtDNA, mitochondrial DNA*; *Amy, amylase*; *rp49, ribosomal protein 49.*

[b] n = number of individuals.

[c] N = number of haplotypes.

[d] U = number of unambiguously connected haplotypes.

[e] References: 1, AQUADRO *et al.* (1986); 2, KREITMAN and AGUADÉ (1986); 3, LANGLEY, MONTGOMERY and QUATTLEBAUM (1982); 4, EANES, LABATE and AJIOKA (1989); 5, BEECH and BROWN (1989); 6, AGUADÉ, MIYASHITA and LANGLEY (1989a); 7, EANES *et al.* (1989); 8, LANGE, LANGLEY and STEPHAN (1990); 9, BROWN (1983); 10, GAME and OAKESHOTT (1990); 11, AQUADRO, LADO and NOON (1988); 12, HALE and SINGH (1987); 13, SCHAEFFER, AQUADRO and LANGLEY (1988); 14, AGUADÉ, MIYASHITA and LANGLEY (1989b); 15, LANGLEY *et al.* (1988); 16, AQUADRO *et al.* (1991); 17, STEPHAN (1989); 18, STEPHAN and LANGLEY (1989); 19, DESALLE (1984); 20, LANGLEY and AQUADRO (1987); 21, LATORRE, MOYA and AYALA (1986); 22, TAMURA, AOTSUKA and KITAGAWA (1991); 23, ROZAS and AGUADÉ (1991); 24, AGUADÉ (1988).



Haplotype	Frequency	Location
1	0.19	Spain, N. Carolina
2	0.02	N. Carolina
3	0.02	N. Carolina
4	0.12	N. Carolina
5	0.05	N. Carolina
6	0.05	N. Carolina
7	0.02	N. Carolina
8	0.07	Spain, N. Carolina
9	0.33	Spain, N. Carolina
10	0.02	N. Carolina
11	0.02	N. Carolina
12	0.02	N. Carolina
13	0.05	Spain
14	0.02	Spain
15	0.02	Spain

FIGURE 5.—A cladogram of the *yellow-achaete-scute* locus of *D. melanogaster* based on data from BEECH and BROWN (1989). The cladogram was constructed using the algorithm of TEMPLETON, CRANDALL and SING (1992). Haplotypes 9, 1, 5 and 2 are interconnected by dashed lines because their relationship is ambiguous. Haplotype 11 could not be connected without ambiguity to the main network.

gram. Similarly, EXCOFFIER and LANGANEY (1989) showed that haplotypes of low frequency usually occur at the tips of cladograms while haplotypes of high frequency occur in the interior. The logic behind this relationship comes from the relationship of age to frequency established by DONNELLY and TAVARÉ (1986). Older alleles (those of higher frequency in the population) have a greater probability of producing mutational derivatives, thereby becoming interior haplotypes than do younger haplotypes (those of lower frequency in the population). From these results, we will first test the following null hypothesis, that haplotype position (tip or interior) is unrelated to the frequency of that haplotype *vs.* the alternative that rarer haplotypes occur preferentially at the tips of cladograms and haplotypes with higher frequency occur in the interior of the cladogram. Note that the frequencies of haplotypes were not used in the cladogram estimation procedure; only the restriction fragment length polymorphisms associated with the particular haplotype.

We test this hypothesis in two ways in order to ameliorate effects of large data sets dominating the statistical analysis and to correct for the possibility of non-independence of haplotypes. First, we perform a sign test over all 29 data sets to test whether singletons (haplotypes represented by a single individual) occur preferentially at the tips of the cladograms. To control for differences in the number of tip haplotypes *vs.* interior haplotypes, we test the null hypothesis that the frequency of singletons at the tips relative to all tip haplotypes is equal to the frequency of singletons in the interior relative to all interior haplotypes against the alternative hypothesis that the frequency of singletons at the tips is greater than the frequency of singletons in the interior. For each data set, singletons unambiguously connected in a network were classified as tip or interior and the following statistic was calculated,

$$Z_n = \frac{\text{tip singletons}}{\text{total tips}} - \frac{\text{interior singletons}}{\text{total interiors}}.$$

A one-sided sign test was then performed on the Z_n's (p. 39, HOLLANDER and WOLFE 1973). With these data we reject the null hypothesis in favor of the alternative, that singletons occur preferentially at the tips, at the $P < 0.0002$ level of significance. The corresponding test for the absolute probability that a singleton is a tip was also highly significant ($P < 0.0002$).

To get a better idea of how tip or interior position is related to haplotype frequency, we carried out a second test of the hypothesis that rarer haplotypes occur preferentially at the tips of cladograms and haplotypes with higher frequency occur in the interior. To test this, we defined eight frequency classes based on within data set haplotype frequencies containing haplotypes in that frequency range. Only frequency classes that were represented in a minimum of 10 different data sets were used. We then performed a delete-one jackknife procedure (EFRON 1982) on the 29 data sets to estimate the proportion of tip haplotypes in each frequency class. Confidence intervals for these estimates were obtained by the method outlined in SOKAL and ROHLF (p. 797, 1981; but see EFRON 1982). The frequency classes and results are shown in Table 3. These results support the prediction, as rare haplotypes are found preferentially at the tips and haplotypes of high frequency are found in the interior. In fact, a graph of estimated tip probability *vs.* haplotype frequency (Figure 6) shows that the estimated tip probability can be well approximated as a linear function of haplotype frequency. The results shown in Table 3 or the linear relationship of Figure 6 can be used as point estimates for a prior probability that a haplotype is a tip or an interior

TABLE 3

Estimates of tip and interior probabilities

Haplotype frequency	Tip probability	Interior probability
0.01–0.02	0.842 (0.795, 0.889)	0.158 (0.111, 0.205)
0.03	0.811 (0.709, 0.914)	0.189 (0.086, 0.291)
0.04–0.05	0.646 (0.510, 0.781)	0.354 (0.219, 0.490)
0.06–0.07	0.567 (0.311, 0.823)	0.433 (0.177, 0.689)
0.08–0.09	0.524 (0.273, 0.774)	0.476 (0.226, 0.727)
0.10–0.11	0.380 (0.024, 0.737)	0.620 (0.263, 0.976)
0.12–0.15	0.257 (0.000, 0.568)	0.743 (0.432, 1.000)
≥0.16	0.096 (0.000, 0.205)	0.904 (0.795, 1.000)

Jackknife bias-corrected estimates of the probability of a haplotype being a tip or an interior given the haplotype has a particular frequency.

given it has a certain frequency. This probability is important in resolving cladogram ambiguities as will be discussed.

A second hypothesis that refines the above result is suggested by equation (3) and the observation by GOLDING (1987) that older alleles tend to have a greater number of mutational connections. Because the expected rank of the alleles by age is proportional to their frequency in the population (DONNELLY and TAVARÉ, 1986) (Equation 3) and frequency of an allele is the quantity easily estimated from a sampling of that population, we want to test the following hypotheses: the number of mutational connections is not correlated with the frequency of an allele *vs.* the alternative that the number of mutational connections increases with increasing frequency.

To perform this test we constructed a contingency table of within-data set frequency classes by the number of mutational connections. We then performed the Jonckheere-Terpstra distribution-free test for ordered alternatives (p. 120, HOLLANDER and WOLFE 1973). This allowed us to test the null hypothesis of no effect of the number of mutational connections on frequency *vs.* the alternative of an ordered effect (*i.e.*, increasing number of mutational connections with increasing frequency). Because of the lack of independence of frequency classes within the same data set, we chose to bootstrap the Jonckheere-Terpstra statistic (J^*). We performed two classes of bootstrap analysis. The first randomly sampled with replacement the number of mutational connections for the haplotype within each frequency class for the entire data set. For each of the 10,000 bootstrap replications the J^* statistic was calculated and saved. The bootstrapped values were then ordered and the J^* statistic from the original data set was compared to that distribution of J^*s generated by the bootstrapping procedure. An approximate 95% confidence interval was calculated by choosing the 250th and 9751st J^* statistics from the ordered distribution of 10,000 (EFRON 1982). The J^* statistic for the original data set was calculated to be 7.37 and the approximate 95% confidence interval for the bootstrapped values was (6.21, 8.55). Since $J^* = 0$ is not within the 95% confidence interval for the bootstrapped values, we conclude $J^* > 0$ and reject the null hypothesis of no order in favor of the ordered alternative, that is, haplotypes with greater frequency tend to have a greater number of mutational connections.

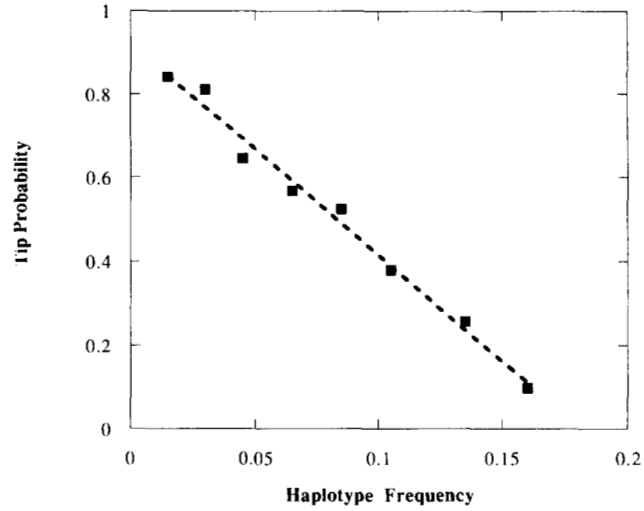


FIGURE 6.—Tip probability plotted against haplotype frequency with a linear approximation to the relationship. The function of the linear relationship is $y = 0.897 - 4.904x$ with a correlation coefficient of $R = 0.991$.

The second alternative bootstrap procedure was to bootstrap by data set, that is, calculate the J^* statistic for the entire data set, then randomly sample with replacement the 29 data sets and calculate new J^* statistics. This strategy controls for dependencies caused by both within-data set frequency dependencies and topological constraints on the number of connections. The resampling was performed 10,000 times resulting in an approximate 95% confidence interval of (5.66, 9.15) and a observed J^* statistic of 7.37. Once again, we can reject the null hypothesis in favor of the ordered alternative hypothesis, concluding that haplotypes with greater frequency tend to have a greater number of mutational connections.

An alternative approach is to limit the population of inference to only those haplotypes that are judged to be interior based on unambiguous connections; thereby eliminating possible bias due to the larger number of tip haplotypes. We restate the null hypothesis as no effect of the number of mutational connections on the frequency of interior haplotypes *vs.* the alternative of an ordered effect. Using the first bootstrapping scheme on this subset of the data, we reject the null hypothesis as the J^* for the original data set was 3.710 with an approximate 95% confidence interval of (2.855, 4.522). Therefore, even with the data restricted to interior haplotypes, we are able to reject the null hypothesis of no association between haplotype frequency and mutational connections.

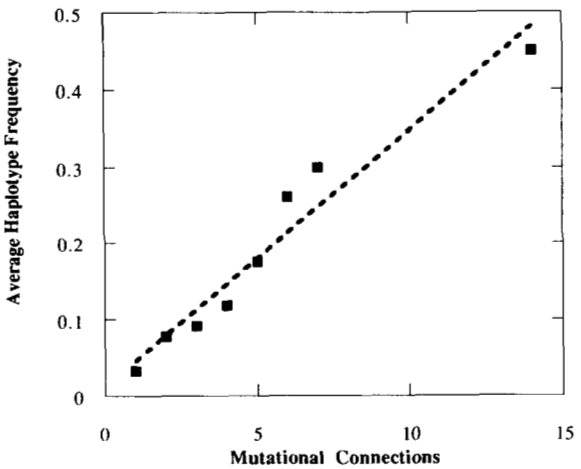


FIGURE 7.—Average haplotype frequency plotted against the number of mutational connections. Average haplotype frequency for mutational connections of 8–13 are missing as no haplotypes were observed in these classes. The function describing this linear relationship is $y = 0.015 + 0.033x$ with a correlation coefficient of $R = 0.976$.

Another line of support for mutational connections increasing with increasing frequency of haplotypes comes from a graphical representation of the average frequency for each class of mutational connections shown in Figure 7. Although this result is not independent from the tip to interior hypothesis, it does lend support to the idea that rare haplotypes occur at the tips (with one mutational connection) and more frequent haplotypes occur in the interior (with >1 mutational connections). Moreover, if one removes the tip class from Figure 7 (mutational connections equal to 1) the linear relationship between the number of mutational connections and haplotype frequency remains.

An additional prediction proposed by EXCOFFIER and LANGANEY (1989) also based on haplotype frequency is,

$$P \text{ (singleton is derived from allele } i) = n_i/n. \quad (4)$$

The reasoning behind this prediction is straightforward. Because most rare haplotypes are of recent evolutionary origin (DONNELLY and TAVARÉ 1986), they are recent mutational derivatives of some other haplotypes. Under a neutral model, the probability that they are derived from some preexisting haplotype is directly proportional to the frequencies of the preexisting haplotypes in the population. Since the mutation is most likely of recent origin (because it is rare), the current haplotype frequencies are good indicators of the haplotype frequencies at the time of its mutational origin. Therefore, the probability of a rare haplotype being mutationally derived from a non-rare haplotype should be approximately proportional to the frequency of the non-rare haplotype. We test the null hypothesis that singletons are equally likely to be connected to a common allele as to another singleton *vs.* the alternative that a singleton is more likely to be connected to a common allele than to another singleton. We control for the variation in numbers of singletons and non-singletons within and between data sets by calculating the frequencies of singleton to singleton and singleton to non-singleton connections for the 29 data sets using each data set as an independent sample. A sign test was performed with a large sample approximation (p. 39, HOLLANDER and WOLFE 1973). We rejected the null hypothesis that there is no frequency difference between singleton-singleton connections *vs.* a singleton-non-singleton connections at the $P = 0.0021$ level of significance. We conclude that there is a strong tendency for singletons to be connected to non-singletons rather than to other singletons in these data sets.

Geographically structured populations: WATTERSON (1985) examines the coalescent process in geographically subdivided populations using a Wright-Fisher model of random mating to derive an expectation for the number of alleles in common between the samples (Equation 3.5 in WATTERSON 1985). From this expectation, we get the prediction that in a geographically subdivided population with limited gene flow, a singleton is more likely to be connected to an allele in the same population than an allele in a different population. To test the null hypothesis of equal occurrences of singleton-same and singleton-different population connections, we tallied the singleton-same and the singleton-different population connections across loci, calculated the difference in frequencies of singleton-same population *vs.* singleton-different population connections, and performed the sign test. Only 19 of the 29 data sets were used in this case as only 19 were studies of more than one population. With these data, we rejected the null hypothesis of no difference between the frequency of singleton-same population connections *vs.* singleton-different population connections at the $P = 0.0154$ level of significance. We conclude that singletons are more likely to be connected to haplotypes from the same population than to haplotypes from different populations. This conclusion is supported without a test for significantly limited gene flow (see below) suggesting that this conclusion is robust to varying levels of gene flow (see Table 4).

TAKAHATA (1988) investigated the coalescent process in a partially isolated population and made predictions based on his results to the more than two subpopulation case. In particular, he predicted that the probability that allele i is the oldest is proportional to the number of subpopulations in which it is found,

$$P \text{ (allele } i \text{ is the oldest)} = Cs_i, \quad (5)$$

where s_i is the number of subpopulations containing allele i and $C = 1/\sum_{i=1}^{n} s_i$. The migration parameter, Nm, was estimated for the 19 data sets containing samples from more than one population using the

TABLE 4

Summary of geographic subdivision

Species	Locus[a]	r[b]	S[c]	Nm[d]	95% CI[e]	Ref[f]
D. melanogaster	*Adh*	4	25	5.4	2.0, 19.6	1
D. melanogaster	*Adh*	2	22	28.1	10.3, 1468.5	2
D. melanogaster	*Adh*	4	6	1.9	*	3
D. melanogaster	*y-ac-sc*	5	18	1.2	*	4
D. melanogaster	*y-ac-sc*	2	5	1.6	0.4, 4.1	5
D. melanogaster	*y-ac-sc*	3	11	1.8	0.8, 3.6	6
D. melanogaster	*G6PD*	5	26	2.8	*	7
D. melanogaster	*mtDNA*	2	7	1.8	0.5, 3.9	12
D. melanogaster	*notch*	3	13	16.2	4.7, 192.3	13
D. melanogaster	*Zeste-tko*	3	17	5.2	1.8, 11.8	14
D. melanogaster	*Amy*	4	19	15.3	*	15
D. ananassae	*Om(1D)*	3	9	1.5	0.5, 2.7	17
D. ananassae	*vermillion*	3	8	0.9	0.2, 2.1	18
D. ananassae	*forked*	3	15	3.7	1.8, 7.2	18
D. pseudobscura	*Amy*	3	8	6.2	*	16
D. melanogaster	*white*	3	13	64.5	*	20
D. subobscura	*mtDNA*	3	3	0.3	0.0, 1.7	21
D. sulfurigasta bilimbata	*mtDNA*	6	8	0.2	0.0, 0.4	22
D. sulfurigasta albostrigata	*mtDNA*	4	7	0.4	0.1, 0.8	22

[a] Loci abbreviations: *Adh, alcohol dehydrogenase*; *y-ac-sc, yellow-achaete-schute*; *G6PD, glucose-6-phosphate dehydrogenase*; *mtDNA, mitochondrial DNA*; *Amy, amylase.*

[b] r = number of subpopulations.

[c] S = minimum number of inferred migration events.

[d] Nm = estimated migration.

[e] 95% CI = 95% confidence interval for the Nm value. * indicates data sets for which no confidence interval could be calculated due to the highly unequal sample sizes between populations.

[f] References: 1, AQUADRO *et al.* (1986); 2, KREITMAN and AGUADÉ (1986); 3, LANGLEY, MONTGOMERY and QUATTLEBAUM (1982); 4, EANES, LABATE and AJIOKA (1989); 5, BEECH and BROWN (1989); 6, AGUADÉ, MIYASHITA and LANGLEY (1989a); 7, EANES *et al.* (1989); 12, HALE and SINGH (1987); 13, SCHAEFFER, AQUADRO and LANGLEY (1988); 14, AGUADÉ, MIYASHITA and LANGLEY (1989b); 15, LANGLEY *et al.* (1988); 16, AQUADRO *et al.* (1991); 17, STEPHAN (1989); 18, STEPHAN and LANGLEY (1989); 20, LANGLEY and AQUADRO (1987); 21, LATORRE, MOYA and AYALA (1986); 22, TAMURA, AOTSUKA and KITAGAWA (1991).

procedure of SLATKIN and MADDISON (1989). Limited gene flow was inferred if Nm and the associated 95% confidence interval was <1.0. Table 4 shows the estimated Nm values and their associated 95% confidence intervals. Only two data sets showed significantly limited gene flow based on this criterion, therefore, predictions from the relationship given by Equation 5 could not be tested using these data.

DISCUSSION: IMPLICATIONS FOR PHYLOGENY RECONSTRUCTION

Rooting: Rooting intraspecific phylogenies is especially difficult because outgroup comparison across species is often impractical due to the high similarity within species. Moreover, outgroup haplotypes are often distinct (separated from the ingroup by many mutational steps) resulting in a lack of resolving power between ingroup alternatives. Three important results from coalescent theory can be used to assign outgroup probabilities, thereby aiding in the rooting of a phylogeny. First, we consider the age of alleles. Recall from Equation 2 that the probability that allele i is the oldest is n_i/n. By definition, the oldest ancestral allele is the root of the phylogeny; therefore, allele i can serve as the outgroup with probability n_i/n, as it is the closest to the true root. Similarly, relationship (5) predicts that in a geographically structured population with limited gene flow, the best outgroup will be that allele which is found in the greatest number of subpopulations. If one uses relationship (5), one must first show that the population is geographically structured with limited gene flow (using, for example, SLATKIN and MADDISON 1989). Because the number of subpopulations within which a haplotype is found is not statistically independent from the frequency of the haplotype (*e.g.*, a singleton can only be found in one subpopulation), one possible way to use the subpopulation information is as a refinement of the frequency prior within a frequency class. So if, for example, a number of haplotypes have a frequency of 0.07, by Equation 2 they would have identical outgroup probabilities. But, if these haplotypes were distributed over various numbers of subpopulations shown to be geographically subdivided, then one could make a more refined statement on the outgroup probabilities within this frequency class based on the number of subpopulations in which each haplotype is found.

As an example of these outgroup criteria, we return to the data set of BEECH and BROWN (1989) in Figure 5. Considering just the relative frequency of alleles we conclude in this example that haplotype 9 is the most probable outgroup to the phylogeny, although with a probability of only 0.33. The next closest candidate is haplotype 1 with probability 0.19. No haplotypes within a frequency class are found in multiple populations; therefore, in this example the number of subpopulations in which a haplotype is found has no bearing on the outgroup probabilities. By these criteria, haplotype 9 is best supported as the most probable outgroup.

The second result from coalescent theory which can aid in rooting is the degree of connectedness; that is, the number of mutational connections of a haplotype in a cladogram. We have shown (see Figure 7) that more frequent haplotypes are expected to have a greater number of mutational connections and that this relationship is linear. By Equation 3, the expected rank in age of allele i is a linear function of its frequency in the population. It follows that the older allele i the greater the number of mutational connections it will have. Therefore,

$$\text{P (allele } i \text{ is the oldest)} = Cm_i, \qquad (6)$$

where m_i is the number of mutational connections of allele i and $= C = 1/\sum_{i=1}^{n} s_i$.

TEMPLETON and CASTELLOE (unpublished manu-

script) have supported this relationship by using a computer program by GRIFFITHS (1989) to show that the probability of allele i being the oldest allele in the population increases with increasing mutational connectedness. They also show that this probability is robust over a 30-fold range of biologically relevant θ values. Relationship (6) can be used as an additional criterion for refining outgroup probabilities for the allelic phylogeny. Returning to our example (Figure 5), we see that once again haplotype 9 is best supported by this criterion as it has four unambiguous connections and at least one more connection when the ambiguous block is resolved. Haplotype 8 has four unambiguous connections and haplotype 5 has two unambiguous connections plus at least one more upon resolution of the ambiguous block.

A third method of rooting is to use a genetic distance approach. From the theory described above, we know the expected time to total coalescence is equal to $4N(1 - n^{-1})$ or $4N$ for large n as in the infinite alleles case (KINGMAN 1982b), so

$$E \text{ (largest genetic distance)} = 2\mu\ 4N = 2\theta.$$

This result implies that mid-point rooting is yet another criterion with which to refine outgroup probabilities. One can simply calculate the distance of each haplotype from the mid-point of the cladogram and normalize these distances for all haplotypes to establish the probabilities of outgroup based on distance from the mid-point. Using mid-point rooting we found that once again haplotype 9 has the highest outgroup probability as it is closest to the mid-point. We point out that these rooting criteria are not evolutionarily independent of one another and are not expected to be. But they are statistically independent of one another as they sample independent sources of information (frequency, number of subpopulations, number of mutational connections, and distance from the mid-point) and can therefore be used to obtain overall outgroup probabilities for each haplotype. We are currently exploring possible ways of using this diverse set of information, as well as additional information, to assign statistically rigorous outgroup probabilities.

Resolving ambiguities: Our results also have important implications for resolving cladogram ambiguities. Ambiguities can arise in one of two ways. The first type, internal ambiguity, occurs when haplotypes are interconnected on the cladogram forming a closed loop which can be broken at any place. Figure 5 shows a cladogram from the *yellow-achaete-scute* locus of *D. melanogaster*. We can see that haplotypes 9, 1, 2 and 5 are interconnected, thereby forming an ambiguous portion of the cladogram. The second form of ambiguity, external ambiguity, arises with tip haplotypes that have alternative connections to a number of haplotypes. Haplotype 11 is an example of external

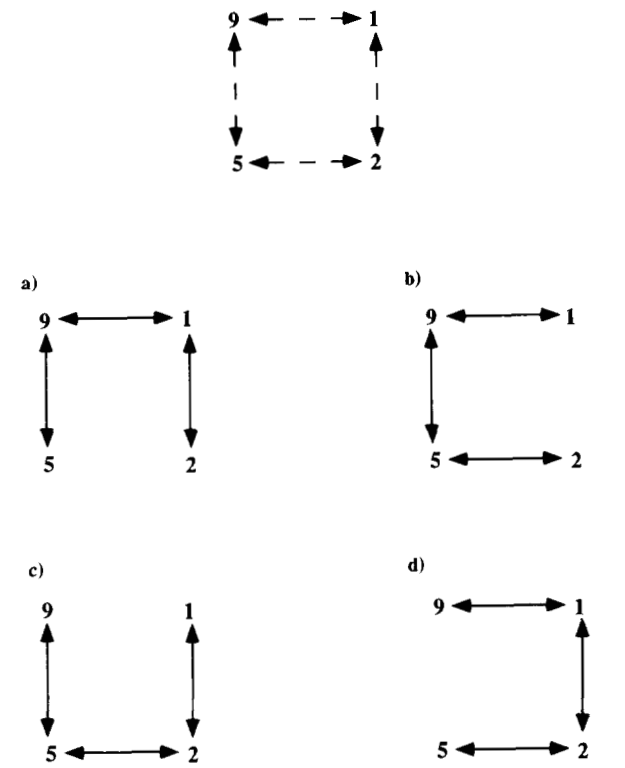


FIGURE 8.—The four possible resolutions of the ambiguous block in Figure 5.

ambiguity as it could be connected to either haplotype 4 or 5.

The first set of criteria helpful in resolving these ambiguities is the tip and interior relationship. We have shown that rare haplotypes occur preferentially at the tips of cladograms and non-rare haplotypes in the interior. The results from Table 3 can be used to assign probabilities to alternative cladograms. For example, of the four haplotypes in the ambiguous block shown in Figure 8, only haplotype 2 is ambiguous in its status as tip or interior. All other haplotypes (9, 5 and 1) are interior by virtue of their nonambiguous connections to other haplotypes. Therefore, of the four possible resolutions, **a** and **b** have haplotype 2 as a tip and resolutions **c** and **d** have it as an interior. From Table 3, we see haplotype 2 (with frequency 0.02) has a probability of 0.842 of being a tip and 0.158 of being an interior. Therefore, resolutions **a** and **b** are much more strongly supported by the tip/interior information than resolutions **c** and **d**. When utilizing these results based on haplotype frequencies, we point out that by equation (1) and Figure 3, one should have a sample of at least 50 individuals to be reasonably confident that one has a good representation of haplotypes in a sample.

Similarly, we can use the tip/interior results to resolve external ambiguities. In the above example, haplotype 11 is two steps (restriction site gains or losses) away from haplotype 4 and 5. Here, haplotype 4 is the ambiguous haplotype with respect to tip/

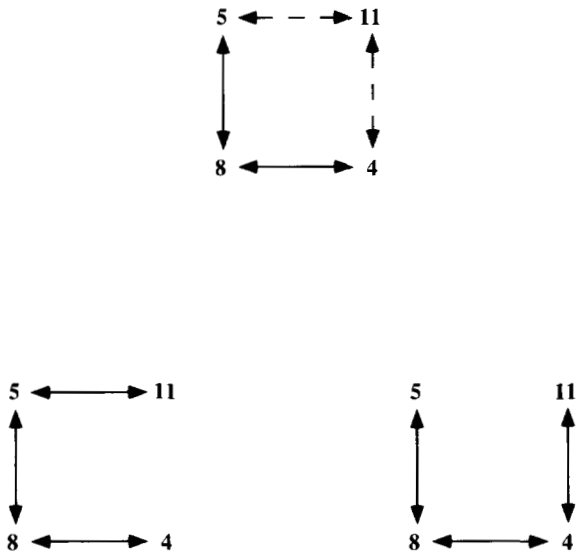


FIGURE 9.—The three possible resolutions of the ambiguous block resulting from the connection of haplotype 11 to the rest of the cladogram.

interior status. Figure 9 shows the ambiguous block with the connection of haplotype 11 and the two possible resolutions. Since haplotype 4 has a frequency of 0.12, Table 3 gives the tip probability as 0.257 and interior probability as 0.743. Based on this information, we would conclude that the second resolution is the more strongly supported.

The results obtained for geographically subdivided populations can also be used to resolve ambiguities in this data set as it samples two populations. We have shown that a singleton is more likely to be connected to haplotypes from the same population than to haplotypes from different populations. We can use these results to resolve internal ambiguities in the following way. Haplotypes 1 and 9 are found in both subpopulations, North Carolina and Spain, haplotypes 2 and 5 are found only in North Carolina. Therefore, using the criterion above, there is no information available based on geographic location. If haplotype 2, a singleton, was from Spain, connections to haplotypes 1 and 9 would be favored over the connection to haplotype 5. For resolution of the external ambiguity in this data set, geographic location has no information, as haplotypes 4, 5 and 11 are all from North Carolina. If haplotype 5 was from Spain and 4 from North Carolina, it would be more likely that haplotype 11 be connected to 4, a haplotype from the same population, than 5.

In addition to the information above, there are other forms of weighting one could use to achieve a better estimate of the phylogeny. For example, if one has strong support for a particular root or an appropriate outgroup so that polarity of restriction site change can be determined, models exist describing differential probabilities for gains and losses of restriction sites (TEMPLETON 1983). It is our intention in future work to develop a methodology of incorporating these diverse sources of information to construct a 95% confidence set of alternative phylogenies. GRIFFITHS (1987, 1989) has taken the first step in this direction by incorporating information on ages of alleles and mutational connections to calculate probabilities for alternative trees. We hope that this methodology will allow one to reduce the number of alternative phylogenies and incorporate all equally well supported phylogenies into subsequent analyses which depend so heavily on the topology of the tree.

We would like to thank JIM CHEVERUD, ERIC ROUTMAN and STAN SAWYER for helpful discussions on the statistical analyses and critical review of the manuscript. MONTY SLATKIN graciously provided a copy of his computer program to calculate Nm values and associated confidence intervals. This work was supported by the National Institutes of Health grants 1 R01 HL39107 and R01 GM31571 (to A.R.T.) and by a National Science Foundation Minority Graduate Fellowship (to K.A.C.).

LITERATURE CITED

AGUADÉ, M., 1988 Restriction map variation at the *Adh* locus of *Drosophila melanogaster* in inverted and noninverted chromosomes. Genetics **119:** 135–140.

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989a Restriction-map variation at the *Zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 123–130.

AGUADÉ, M., N. MIYASHITA and C. H. LANGLEY, 1989b Reduced variation in the *yellow-achaete-scute* region in natural populations of *Drosophila melanogaster*. Genetics **122:** 607–615.

AQUADRO, C. F., K. M. LADO and W. A. NOON, 1988 The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. Genetics **119:** 875–888.

AQUADRO, C. F., S. F. DESSE, M. M. BLAND, C. H. LANGLEY and C. C. LAURE-AHLBERG, 1986 Molecular population genetics of the *alcohol dehydrogenase* gene region of *Drosophila melanogaster*. Genetics **114:** 1165- 1190.

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*: The *amylase* gene region. Proc. Natl. Acad. Sci. USA **88:** 305–309.

BEECH, R. N., and A. J. L. BROWN, 1989 Insertion-deletion variation at the *yellow- achaete-scute* region in two natural populations of *Drosophila melanogaster*. Genet. Res. **53:** 7–15.

BROWN, A. J. L., 1983 Variation at the *87A* heat shock locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA. **80:** 5350–5354.

DESALLE, R., 1984 Mitochondrial DNA evolution and phylogeny in the Planitibia subgroup of Hawaiian Drosophila. Ph.D. Dissertation, Washington University, St. Louis.

DONNELLY, P., and S. TAVARÉ, 1986 The ages of alleles and a coalescent. Adv. Appl. Probab. **18:** 1–19.

EANES, W. F., J. LABATE and J. W. AJIOKA, 1989 Restriction-map variation with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 492–502.

EANES, W. F., J. W. AJIOKA, J. HEY and C. WESLEY, 1989 Restriction-map variation associated with the *G6PD* polymorphism in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 384–397.

EFRON, B., 1982 *The Jackknife, the Bootstrap, and Other Resampling Plans* (CBMS Regional Conference Series in Applied Mathe-

matics, Vol. 38). Society for Industrial & Applied Mathematics, Philadelphia.

EWENS, W., 1990 Population genetics theory–the past and the future, pp. 177–227 in *Mathematical and Statistical Developments of Evolutionary Theory*, edited by S. LESSARD. Kluwer Academic Publishers, New York.

EXCOFFIER, L., and A. LANGANEY, 1989 Origin and differentiation of human mitochondrial DNA. Am. J. Hum. Genet. **44:** 73–85.

GAME, A. Y., and J. G. OAKESHOTT, 1990 Associations between restriction site polymorphism and enzyme activity variation for *esterase 6* in *Drosophila melanogaster*. Genetics **126:** 1021–1031.

GOLDING, G. B., 1987 The detection of deleterious selection using ancestors inferred from a phylogenetic history. Genet. Res. **49:** 71–82.

GRIFFITHS, R. C., 1987 Counting genealogical trees. J. Math. Biol. **25:** 423–431.

GRIFFITHS, R. C., 1989 Genealogical-tree probabilities in the infinitely-many-site model. J. Math. Biol. **27:** 667–680.

HALE, L. R., and R. S. SINGH, 1987 Mitochondrial DNA variation and genetic structure in populations of *Drosophila melanogaster*. Mol. Biol. Evol. **4:** 622- 637.

HEIN, J., 1990 Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98:** 185–200.

HOLLANDER, M., and D. A. WOLFE, 1973 *Nonparametric Statistical Methods*. John Wiley & Sons, New York.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–42 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, Oxford.

KELLY, F. P., 1977 Exact results for the Moran neutral allele model. Adv. Appl. Prob. **9:** 197–201.

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

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

KREITMAN, M., 1983 Nucleotide polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. Nature **304:** 412–417.

KREITMAN, M., and M. AGUADÉ, 1986 Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

LANGE, B. W., C. H. LANGLEY and W. STEPHAN, 1990 Molecular evolution of Drosophila *metallothionein* genes. Genetics **126:** 921–932.

LANGLEY, C. H., and C. F. AQUADRO, 1987 Restriction-map variation in natural populations of *Drosophila melanogaster*: *White*-locus region. Mol. Biol. Evol. **4:** 651–663.

LANGLEY, C. H., E. MONTGOMERY and W. F. QUATTLEBAUM, 1982 Restriction map variation in the *Adh* region of Drosophila. Proc. Natl. Acad. Sci. USA **79:** 5631–5635.

LANGLEY, C. H., A. E. SHRIMPTON, T. YAMAZAKI, N. MIYASHITA, Y. MATSUO and C. F. AQUADRO, 1988 Naturally occurring variation in the restriction map of the *Amy* region of *Drosophila melanogaster*. Genetics **119:** 619–629.

LATORRE, A., A. MOYA and F. AYALA, 1986 Evolution of mitochondrial DNA in *Drosophila subobscura*. Proc. Natl. Acad. Sci. USA **83:** 8649–8653.

ROZAS, J., and M. AGUADÉ, 1991 Study of an isolated population at the nucleotide level: *rp49* region of a Canarian population of *Drosophila subobscura*. Mol. Biol. Evol. **8:** 202–211.

SCHAEFFER, S. W., C. F. AQUADRO and C. H. LANGLEY, 1988 Restriction-map variation in the *Notch* region of *Drosophila melanogaster*. Mol. Biol. Evol. **5:** 30–40.

SLATKIN, M., and W. MADDISON, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

SOKAL, R. R., and F. J. ROHLF, 1981 *Biometry*. W. H. Freeman, New York.

STEPHAN, W., 1989 Molecular genetic variation in the centromeric region of the X chromosome in three *Drosophila ananassae* populations. II. The *Om(1D)* locus. Mol. Biol. Evol. **6:** 624–635.

STEPHAN, W., and C. H. LANGLEY, 1989 Molecular genetic variation in the centromeric region of the *X* chromosome in three *Drosophila ananassae* populations. I. Contrasts between the *vermilion* and *forked* loci. Genetics **121:** 89–99.

SWOFFORD, D. L., 1991 *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.0r*. Computer program distributed by the Illinois Natural History Survey, Champaign, Ill.

TAKAHATA, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res. **52:** 213–222.

TAMURA, K., T. AOTSUKA and O. KITAGAWA, 1991 Mitochondrial DNA polymorphism in the two subspecies of *Drosophila sulfurigaster*: relationship between geographic structure of population and nucleotide diversity. Mol. Biol. Evol. **8:** 104–114.

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

TEMPLETON, A. R., 1983 Convergent evolution and nonparametric inferences from restriction data and DNA sequences, pp. 151–179 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

TEMPLETON, A. R., E. BOERWINKLE and C. F. SING, 1987 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. Genetics **117:** 343–351.

TEMPLETON, A. R., K. A. CRANDALL and C. F. SING, 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. III. Cladogram estimation. Genetics **132:** 619–633.

WATTERSON, G. A., 1976 Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. Theor. Popul. Biol. **10:** 239–253.

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

WATTERSON, G. A., and H. A. GUESS, 1977 Is the most frequent allele the oldest? Theor. Popul. Biol. **11:** 141–160.

Communicating editor: A. G. CLARK