Evolution, 44(2), 1990, pp. 360–370

GENE GENEALOGIES WITHIN THE ORGANISMAL PEDIGREES OF RANDOM-MATING POPULATIONS

R. MARTIN BALL, JR.,[1] JOSEPH E. NEIGEL,[2] AND JOHN C. AVISE[1]
[1]*Department of Genetics, University of Georgia, Athens, GA 30602*
[2]*Department of Biology, University of Southwestern Louisiana, Lafayette, LA 70504*

Abstract.—Using computer simulations, we generated and analyzed genetic distances among selectively neutral haplotypes transmitted through gene genealogies with random-mating organismal pedigrees. Constraints and possible biases on haplotype distances due to correlated ancestry were evaluated by comparing observed distributions of distances to those predicted from an inbreeding theory that assumes independence among haplotype pairs. Results suggest that: 1) mean time to common ancestry of neutral haplotypes can be a reasonably good predictor of evolutionary effective population size; 2) the nonindependence of haplotype paths of descent within a given gene genealogy typically produces significant departures from the theoretical probability distributions of haplotype distances; 3) frequency distributions of distances between haplotypes drawn from "replicate" organismal pedigrees or from multiple unlinked loci within an organismal pedigree exhibit very close agreement with the theory for independent haplotypes. These results are relevant to interpretations of current molecular data on genetic distances among nonrecombining haplotypes at either nuclear or cytoplasmic loci.

Received October 21, 1988. Accepted November 13, 1989

With the advent of efficient laboratory methods for nucleotide sequencing and restriction-site mapping, it has become feasible to assay many DNA haplotypes at particular loci from a population or species. Such data can be analyzed phylogenetically to estimate the evolutionary relationships (gene genealogies or "gene trees" [Nei, 1987]) among the alleles of a gene (Avise, 1989). There is now a need for further development of a corresponding theory for haplotype genealogies at the within-species level. A suitable theory should include the expected effects on allelic relationships of various historical demographic factors such as population size and gene-flow pattern. Here we analyze one property of gene genealogies—the distribution of times to common ancestry for haplotypes—within random mating populations.

Suppose that, from each of a very large number of replicate, random-mating populations of effective size N_e, two haplotypes (alleles) were drawn at random from a nuclear gene locus. Suppose further that the times to common ancestry (the times to identity by descent) of these haplotype pairs were determined. In this idealized scenario, the probability $f(G)$ that two randomly chosen haplotypes are derived from the same ancestral haplotype that existed G generations prior is

$$f(G) = \left(\frac{1}{2N_e}\right)\left(1 - \frac{1}{2N_e}\right)^{G-1} \qquad (1)$$

(Tajima, 1983; Nei, 1987). For a maternally transmitted gene such as mitochondrial DNA (mtDNA) in higher animals, $2N_e$ in Equation (1) is replaced by $N_{e(f)}$, the effective size of the female population (Avise et al., 1988). Equation (1) is the probability distribution function of times to identity by descent among independent haplotypes. The form of this distribution is geometric, with mean $2N_e$ and variance $4N_e^2$.

In reality, such an empirical sampling design for haplotypes is not practical. Rather, an array of haplotypes from one "gene" (such as alcohol dehydrogenase or mtDNA) is normally assayed from one population or species (e.g., Aquadro et al., 1986; Avise et al., 1987; Kreitman, 1983). While genetic distances among haplotypes can be calculated and converted to a frequency distribution of times to common ancestry using a molecular-clock calibration (Avise et al., 1988), this distribution need not agree with the above theory because (among other possibilities) the haplotype distances are correlated due to coancestry through a particular gene genealogy within an organismal pedigree.

In this paper, we employ computer sim-

ulations to examine the possible constraints imposed by organismal pedigrees on ancestral relationships among selectively neutral haplotypes within random-mating populations. Results will be particularly relevant to current uses of empirical data on allelic distances to estimate evolutionary effective sizes of some populations and species in nature.

Materials and Methods

General Outline for the Models

Any current population of individuals is the product of demographically influenced matings and births that have occurred in preceding generations. The pedigree (record of matings and births) defines the familial relationships of individuals. The pedigree also defines a set of possible paths of descent, or gene genealogies, that could link the haplotypes in the current population to their progenitors. In principle, any two haplotypes chosen from the current generation can be traced back through the pedigree to the point where they both stem from the same ancestral haplotype. The time at which this most-recent common ancestor existed depends on the gene genealogy of the locus examined and the particular pair of haplotypes chosen.

Our model consists of three programs, LINELAND, TRICKLE, and NODEUP, which were designed, respectively, to 1) produce a random-mating population pedigree according to a given set of demographic parameters, 2) choose at random one possible gene genealogy through the pedigree, and 3) find the time to the most-recent common ancestor for random pairs of haplotypes traced through the gene tree. The programs were designed to compare the effects of drawing haplotypes of one gene from many independent population pedigrees with the effects of drawing haplotypes of many genes through a common population pedigree. Our simulation method, while much slower than those based on Hudson's (1983) algorithm, has the advantage that it allows us to look at many gene trees consistent with a single organismal pedigree. In addition, because our simulations involve direct observations of times to common ancestry, rather than approximations, we can utilize information from all of the individuals in a small population.

Space does not permit a complete listing of the programs outlined below, but such a listing is available upon request from R.M.B. The programs were written in Pascal and run on an IBM PS/2 Model 80 computer using the Turbo Pascal 4.0 compiler.

i) LINELAND.—The purpose of this program is to create a file representing the pedigree of a simulated population. The population is composed of nonoverlapping generations of dioecious individuals. Population size is density-regulated by assigning each female a random number of progeny from a Poisson distribution with parameter λ. In each generation, lambda is calculated as:

$$\lambda = \left(\frac{I}{F}\right)\exp\left(\frac{C - I}{C}\right)$$

where F and I are the numbers of females and total individuals, respectively, in the current generation, and C is the carrying capacity of the population (set at 100 individuals in most of our simulations).

The population size for each generation is stored in a data record, which also contains an array of records describing the sex and the parents of each individual in the population. An initial generation is created from 100 individuals whose sex is assigned at random. The first generation is then used to generate a second generation. Each female in the first generation is assigned a number of progeny as described above. Each offspring is assigned a father, drawn at random from the pool of males in the first generation, and a gender. The offspring record is then stored in a record for the second generation. After all of the females have been processed, the data record of the first generation is saved in the pedigree file. In the same manner, the second generation is used to create a third generation, and so on. This process is repeated until 2,000 generations have been recorded. The disk file thus contains the pedigree of a random-mating population.

ii) TRICKLE.—This program uses the population pedigree produced by LINELAND to produce a gene genealogy. For a given locus, each haplotype can be uniquely specified by identifying the individual in

which it exists and the parent from which it was inherited. Similarly, any haplotype in the parental generation is uniquely specified by the member of the parental generation in which it occurs and the grandparent from which it was inherited. We can therefore specify the link between a haplotype in the current generation and a haplotype in the previous generation by identifying the individual in which the current haplotype occurs, one of its parents, and one of its grandparents. Furthermore, haplotypes in the current generation that share the same parent and grandparent were identical by descent in the previous generation.

TRICKLE starts with the record of the last generation (P) produced by LINELAND. For each individual in the population, one of the parents is chosen at random and stored with the individual. The program then reads the record describing the preceding generation, $P - 1$, and chooses at random one of the two grandparents consistent with the parent already identified. A record containing the array of these combinations of individual, parent, and grandparent is referred to as a gene generation, to distinguish it from a generation of individual organisms. This array, which establishes the relationship between haplotypes in the current generation and their progenitors in the previous generation, is then stored to a disk file.

The program now finds the relationships between the haplotypes in generation $P - 1$ and those in generation $P - 2$, by considering the individuals in $P - 1$ as the "current" generation. The parents of generation P become the individuals of generation $P - 1$, and the grandparents of P become parents of $P - 1$. There could be duplicate pairs of individual and parent in $P - 1$, due to the common ancestry of two haplotypes in the previous generation. The program eliminates these duplicates to prevent the lineages from appearing to diverge once they have coalesced. The new generation is completed by choosing at random from the possible grandparents and storing the record. The program continues this process back through time until there is only one lineage in the "current" generation, all others having been combined as duplicates in the earlier iterations.

iii) NODEUP.—Nodeup is designed to reduce the information produced by TRICKLE to a more compact form. It also chooses random pairs of haplotypes in the final generation and finds the times to common ancestry (distances) between them.

NODEUP starts at the last generation of the gene genealogy and creates an array of linked lists. Each record in the lists refers to one of the haplotypes in the final generation, and each occupied element in the array represents a haplotype in the current generation. In the initial array, each element points to a list consisting of a single record which represents the same haplotype as the array element. As the program progresses, when two or more haplotypes coalesce in the file produced by TRICKLE, the lists connected to those haplotypes are combined into a single linked list which is then attached to the ancestor in the preceding generation. The times of linkage are recorded in a distance matrix. After all of the records have been linked into a single list, the program finishes by saving a file of distances between random pairs of haplotypes drawn from the final generation.

Note that the simulations deal directly with the times to common ancestry, as defined by the pedigree (i.e., no haplotypes are generated or monitored). Thus, in effect, the results summarize expectations for a perfect, metronomic molecular clock. In practice, genetic distances estimated from nucleotide sequences, site maps, or other molecular characters are not likely to accumulate at a constant rate. Any correlational pattern of coancestry for such empirical genetic distances might be even higher than those revealed by the simulations (see below).

Sampling Designs and Statistical Analyses

Using the programs described above, it is possible to create pedigrees and to simulate the descent of haplotypes (involving either nuclear or cytoplasmic genes) through them. The sampling design for the computer simulations is summarized in Figure 1. Through each of 50 independently generated organismal pedigrees (1, 2, 3, . . . , 50), 50 gene trees (A, B, C, . . . , Z, AA, . . . , XX) were followed. Within each of these 2,500 cells,

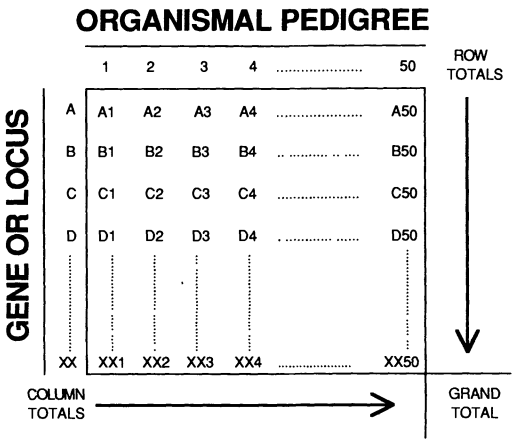


FIG. 1. Experimental design employed in the computer simulations of gene genealogies within and among organismal pedigrees.

a frequency distribution of times to common ancestry for randomly chosen haplotype pairs was constructed. So that no haplotype was used more than once, sampling of individuals was done without replacement, and each frequency distribution thus contained about 50 data points (one-half the number of individuals in the final generation of a population).

In Figure 1, a "column total" refers to a frequency distribution of haplotype distances for unlinked genes segregating within the same organismal pedigree, with one haplotype pair chosen at random from each of the 50 gene trees. Similarly, a "row total" refers to a frequency distribution of haplotype distances for any one gene sampled across independently generated pedigrees, with one haplotype pair chosen at random from each of 50 organismal pedigrees. Because the row and column totals are based on the same numbers of sampled data points (which is also very close to the number used within each gene-pedigree combination), statistical tests to detect departures from theory have the same power in all classes of comparison. The "grand total" in Figure 1 will refer to a frequency distribution of haplotype distances summed across the row or the column totals. Thus, for nuclear genes, the grand total consists of 2,500 data points, with one haplotype pair drawn from each of the 50 × 50 gene-pedigree combinations.

The theoretical frequency distributions of times to common ancestry among haplotypes were obtained from Equation (1), using as N_e one-half the mean time to common ancestry observed in the relevant simulation(s). These probability distributions of expected distances were then compared to the observed distributions by a) the Kolmogorov-Smirnov test to assess statistical significance of departures (Sokal and Rohlf, 1981, pp. 716–721) and b) a "pixel count" approach, which quantifies departures by measuring the area between a theoretical and an empirical curve (Lemke, 1985). Both the Kolmogorov-Smirnov and pixel-count methods evaluate cumulative distribution functions (Sokal and Rohlf, 1981, p. 716), which will be referred to throughout this paper.

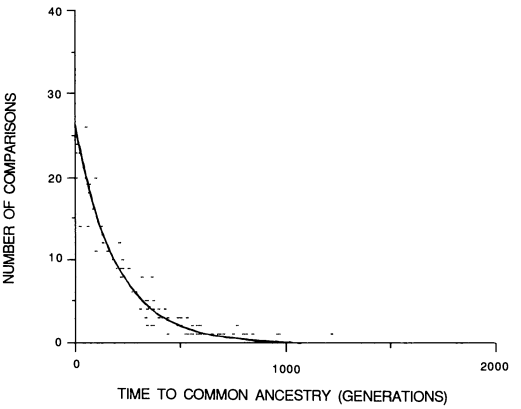
The evolutionary effective size (N_e) of each simulated population was also estimated by taking into account the sex ratio in each generation ($N_G = 4N_mN_f/[N_m + N_f]$, where N_m and N_f are the census numbers of males and females, respectively [Nei, 1987]) and then calculating the harmonic mean of the N_G values across generations.

RESULTS

"Grand Total" and "Within-Cell" Comparisons

The observed distribution of haplotype distances for nuclear genes is plotted in Figure 2, along with the corresponding cumulative distribution function for the 2,500 data points in the "grand total" summary from the simulations. Also shown in Figure 2 are the theoretical expectations for these haplotype distances, assuming that the draws were from populations of size $N_e = 99$. The agreement between theory and observations is very close, suggesting that the computer simulations are operating as intended.

The distribution of distances among mtDNA haplotypes from the "grand total" summation is plotted in Figure 3, in the same format. This comparison is necessarily based on 50 data points, because there is only one mtDNA gene genealogy in each organismal pedigree. Again, the agreement between theory and observation is very close. The observed mean time to common ancestry for mtDNA haplotypes (43.0 generations) is approximately one-fourth the



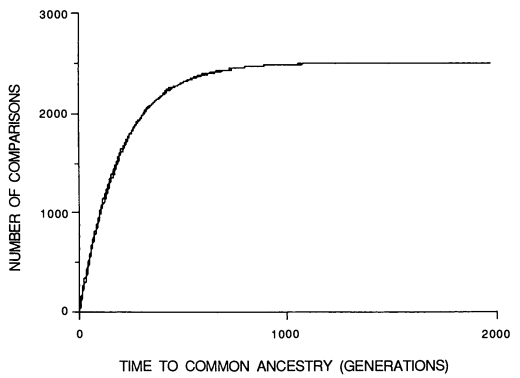
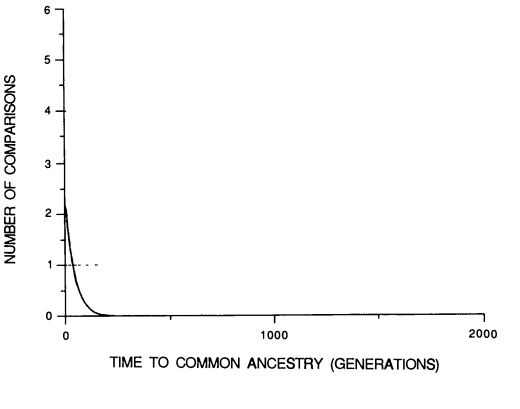


FIG. 2. Above) Frequency distributions of haplotype distances at nuclear loci in the 2,500 gene-pedigree comparisons involved in the "grand total" compilation. The curve is the theoretical expectation for independent haplotypes in a population of effective size $N_e = 99$. The points represent the numbers observed in the computer simulations. Below) Corresponding cumulative distribution functions for these theoretical and observed histograms (the two functions overlie one another and hence are virtually indistinguishable).



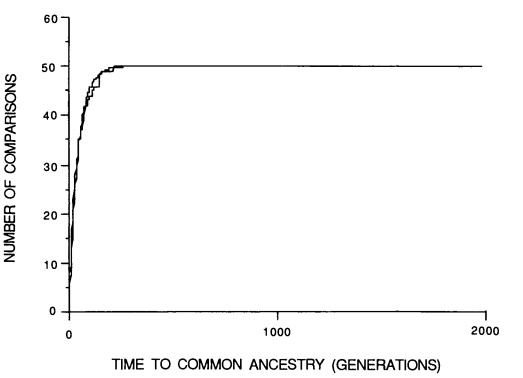


FIG. 3. Above) Frequency distributions of haplotype distances for maternally transmitted genes (such as mtDNA) in the 50 pedigrees involved in the "grand total" compilation. The curve is the theoretical expectation for independent haplotypes in a population of female effective size $N_{e(f)} = 43$. The points represent the numbers observed in the computer simulations. Below) Corresponding cumulative distribution functions representing the theoretical and observed values. The Kolmogorov-Smirnov test shows no significant difference between the cumulative distribution functions ($P = 0.57$).

time to common ancestry for nuclear haplotypes (198.0 generations), as expected. Results from the "grand total" comparisons provide confidence that, at this level of sampling, a good agreement can be expected between observed and theoretical distributions of haplotype distances, despite the fact that many of the gene genealogical tracings have been through the same organismal pedigree.

At the other end of our scale of analysis, frequency distributions of distances among pairs of haplotypes drawn from particular gene-pedigree combinations (the "within-cell" comparisons of Fig. 1) seldom agreed with the theoretical curves generated under the assumptions of independence among haplotypes. By the Kolmogorov-Smirnov test statistic, 49 of 50 cells departed significantly (at the $P < 0.05$ level) from theoretical expectations. As examples, three gene genealogies (representing three cells) are shown in Figure 4. The observed and expected cumulative distribution functions for these cells are shown in Figure 5.

Clearly, at some sampling level lying between the "within-cell" and the "grand total" comparisons of Figure 1, a transition

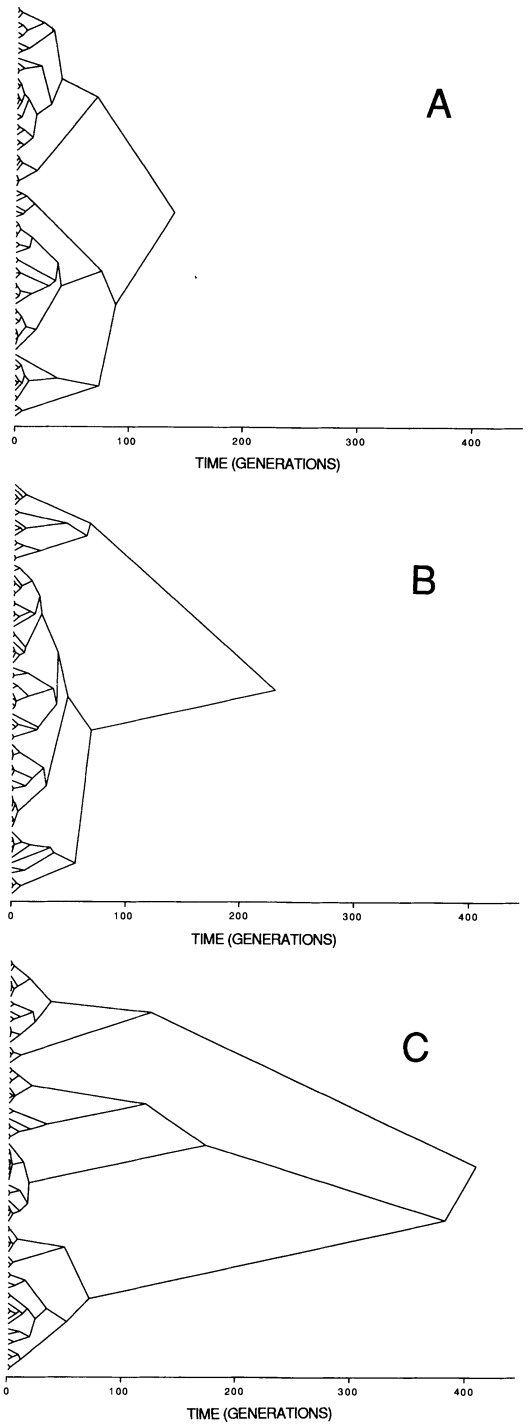


FIG. 4. Examples of gene genealogies representing within-cell comparisons. The three cases (A–C) produce the respective cumulative distribution functions shown in Figure 5.

to an acceptable level of independence among haplotypes required by theory [Equation (1)] has been made.

Comparisons Involving "Row Totals" and "Column Totals"

Frequency distributions of distances between pairs of haplotypes drawn from independent organismal pedigrees (the 50 "row totals" of Fig. 1) invariably showed good agreement with the appropriate theoretical distributions (no departures were statistically significant at the $P < 0.05$ level, as judged by Kolmogorov-Smirnov tests). Examples of the cumulative distribution functions are shown in Figure 6. Similarly, there were no significant departures from theoretical distributions of distances between haplotypes drawn from unlinked genes traced through the same organismal pedigree (the 50 "column totals" of Fig. 1). Examples are shown in Figure 7.

The pixel counts, which quantify the difference between theoretical and observed curves, were very similar in the "column total" and "row total" data summaries (Table 1). Results suggest that gene lineages transmitted through a single organismal pedigree show nearly as much independence as do gene lineages traced through separate organismal pedigrees generated under a common set of demographic conditions.

Effective Population Size

The mean times to common haplotype ancestry in our simulations were very close to the anticipated values of $2N_e$ (when N_e was calculated as the harmonic mean of the census population sizes after correction for sex ratio effects; see Materials and Methods). Thus, among the 2,500 randomly chosen gene-pedigree combinations, mean haplotype distance and monitored $2N_e$ were 198.0 and 194.6, respectively.

DISCUSSION

Our simulations have allowed an assessment of various sources of sampling influence on frequency distributions of haplotype distances in random-mating populations. An understanding of such influences is important, because some of the theoretical predictions derived from in-

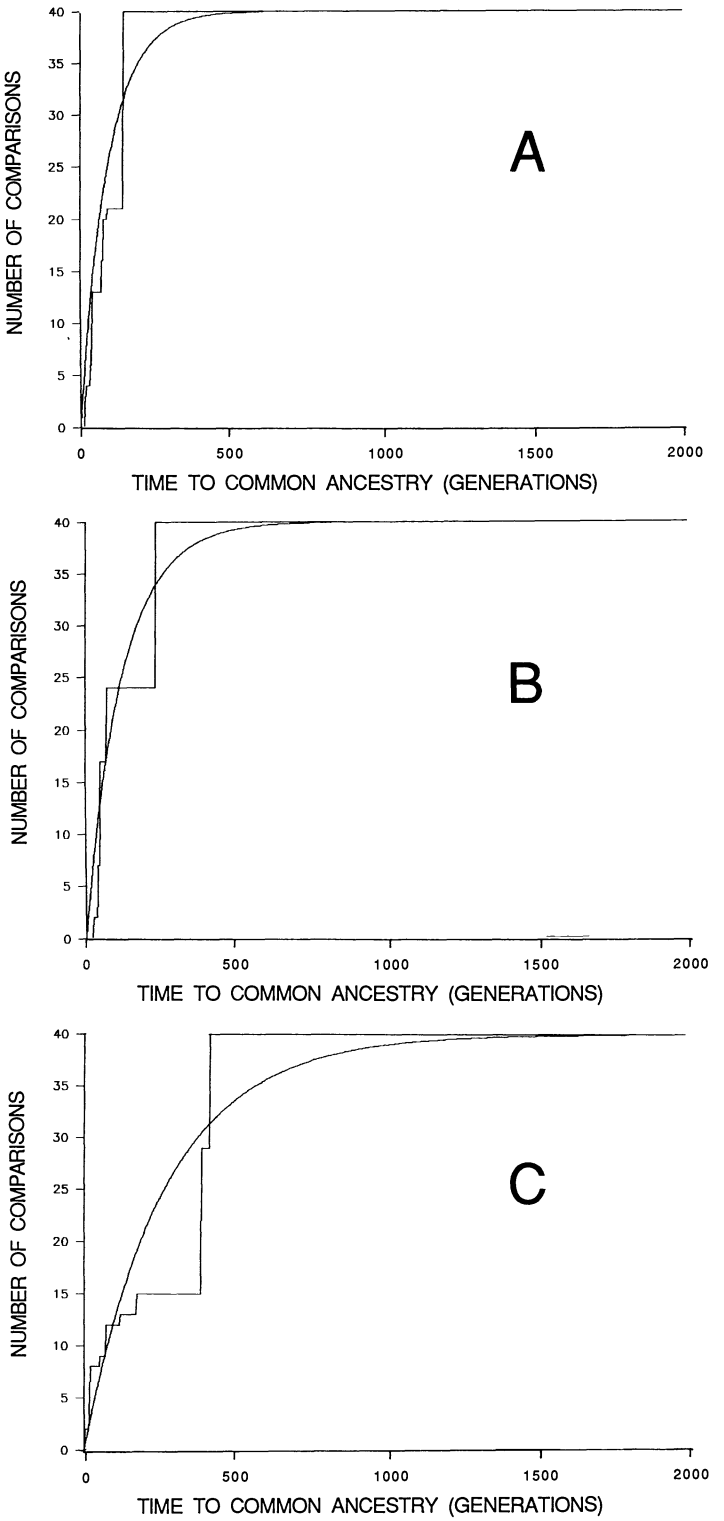


FIG. 5. Theoretical (smooth-appearing) and observed (stepped) cumulative distribution functions for three typical within-cell simulations (see Fig. 1 and text). The three cases (A–C) reflect the respective gene genealogies pictured in Figure 4. In each case, the parameter N_e for the theoretical curve was estimated from the observed data, in order to maximize the fit of the theoretical curve.

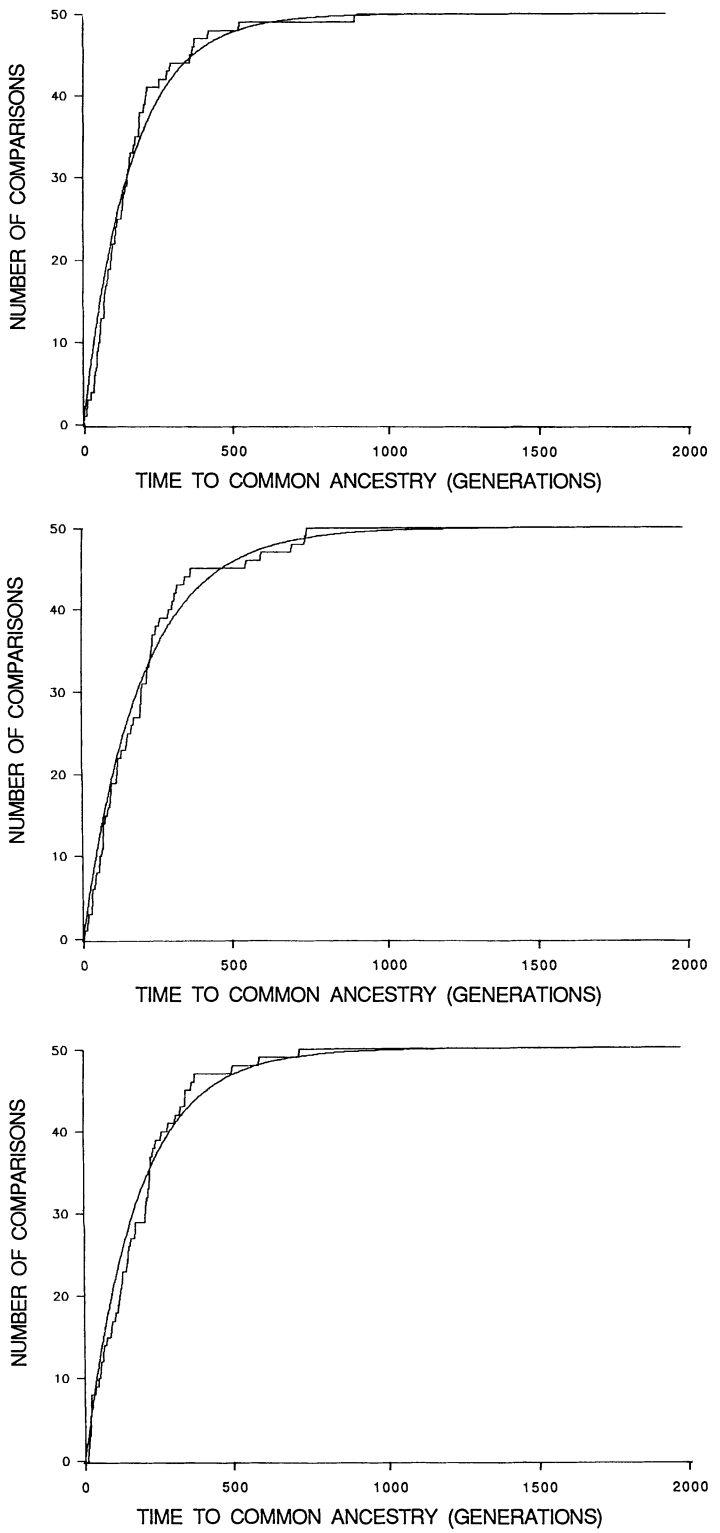


FIG. 6. Theoretical (smooth-appearing) and observed (stepped) cumulative distribution functions for three typical summaries of "row totals" (see Fig. 1 and text).

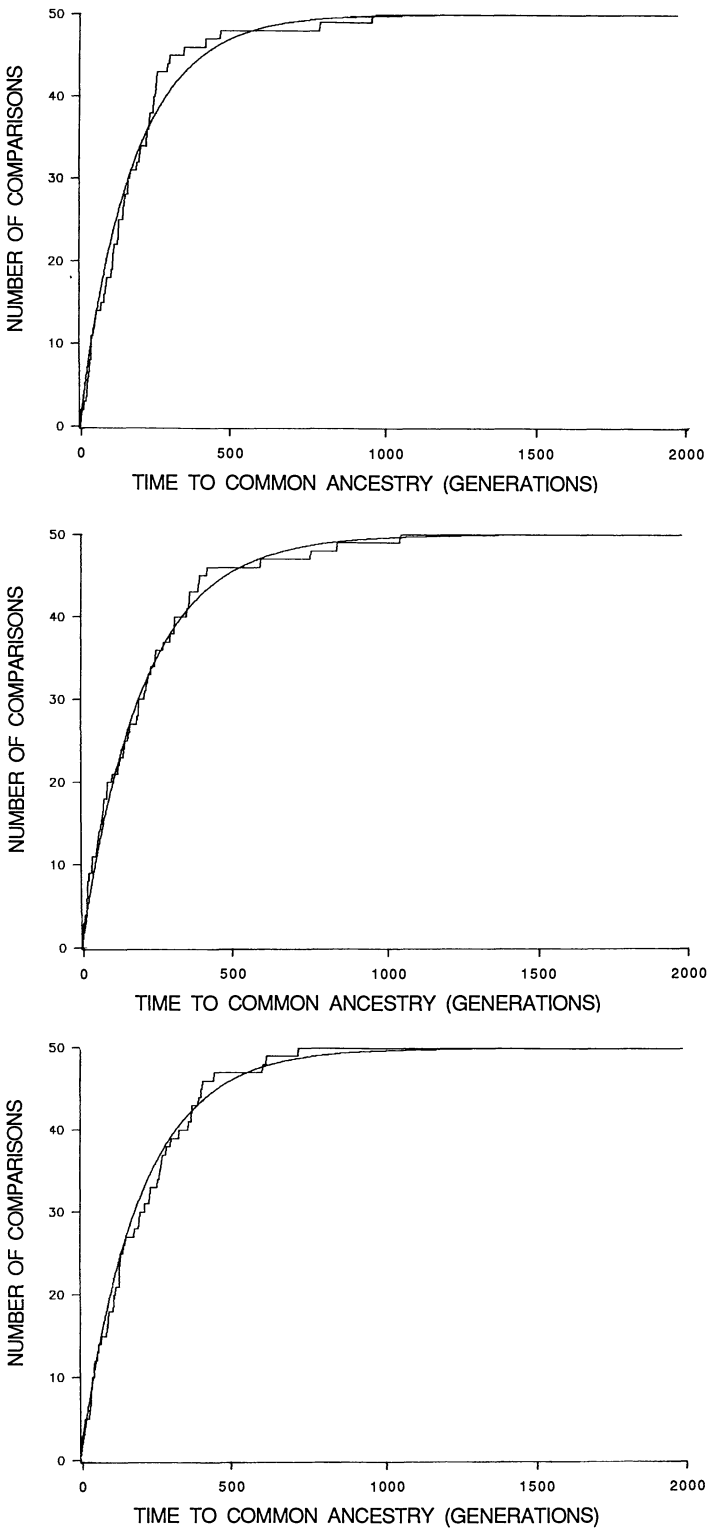


FIG. 7. Theoretical (smooth-appearing) and observed (stepped) cumulative distribution functions for three typical summaries of "column totals" (see Fig. 1 and text).

TABLE 1. Differences between theoretical and observed cumulative distribution functions, as quantified by the "pixel count" approach, for various sampling designs of haplotype distances within organismal pedigrees (see Fig. 1 and text).

Data summary	Pixel counts		Number of cases	
	Mean	SD	Examined	Significant[a]
Within-cell comparisons	1,812.2	891.9	50	49
Row totals	621.0	186.1	50	0
Column totals	617.2	187.1	50	0

[a] $P < 0.05$ (Kolmogorov-Smirnov test).

breeding theory [such as Equation (1)] apply to independent haplotypes, while most empirical data sets involve haplotypes whose relationships are partially correlated.

Lack of independence among haplotypes at a locus could arise from several sources, including 1) physical recombination or conversion acting to shuffle nucleotide sequences among alleles, 2) historical interconnectedness of haplotypes due to membership in a shared gene genealogy, and 3) transmission of haplotypes (even at unlinked loci) through a common organismal pedigree. In this study, we have neglected the first source listed above (by assuming a lack of recombination among the haplotypes at any one locus in our models) and have focused on the relative importance of gene genealogy and organismal pedigree on haplotype distances. The simulations were designed to compare the effects of sampling haplotypes at a locus through an organismal pedigree ("within-cell" effects), haplotypes at a locus through separate organismal pedigrees generated under the same set of demographic conditions ("row totals"), and haplotypes at unlinked loci within an organismal pedigree ("column totals").

The "within-cell" comparisons show that the frequency distributions of genetic distances among many pairs of haplotypes of one gene in a random-mating population will seldom conform precisely to the theory for independent alleles [Equation (1)]. These departures apparently arise because the haplotypes are connected in an historical gene genealogy which places constraints on the patterns of observable allelic relationship. For example, the time of first splitting of haplotypes in a gene genealogy introduces some constraint on the divergence times of haplotypes in later generations. Nonetheless, the constraints appear to be operating without strong directional bias, because mean distance among many haplotype pairs continues to show very good agreement with theoretical expectations.

With our sample sizes, we were unable to detect significant departures in either the "row total" or "column total" data summaries. Thus, to a first approximation, the theory for independent haplotypes would appear to apply quite well to any large empirical data set that entailed draws of random haplotypes from independent organismal pedigrees or draws of haplotype pairs from many unlinked loci within a population.

These results are relevant to recent interpretations of molecular data. For example, Avise et al. (1988) observed dramatically lower genetic distances among mtDNA haplotypes than were predicted from the current census population sizes in each of three species of vertebrates with high gene flow. They concluded that long-term effective population size was likely to have been much smaller (by two to three orders of magnitude) than present-day census size for these species. The restriction-site data of Avise et al. (1988) involved haplotypes at a single "gene" (mtDNA) within particular species and, hence, are analogous to the "within cell" comparisons in Figure 1. Our current simulations suggest that, notwithstanding violation of the assumptions of independence among haplotypes, N_e can indeed be estimated reasonably (certainly with respect to order of magnitude) from such within-cell comparisons. Avise et al. (1988) also noticed a significant difference between the observed and theoretical frequency distributions of mtDNA distances, even after appropriate modification of the theoretical curves to reflect the lower postulated N_e's. Our present simulations show that such departures are a likely and expected consequence of the inherent lack of independence among haplotype distances within a gene genealogy.

This study has dealt solely with distributions of haplotype distances in random-mating populations and represents only a

first step toward development of a broader theory for haplotype genealogies at the intraspecific level. Useful extensions of this approach will include comparisons of the shapes of gene genealogies within and among the population pedigrees of geographically structured species.

Acknowledgments

We thank M. Asmussen for generous donation of computer time. Work was supported by NSF grant BSR-8805360 to J.C.A.

Literature Cited

Aquadro, C. F., S. F. Desse, M. M. Bland, C. H. Langley, and C. Laurie-Ahlberg. 1986. Molecular population genetics of the alcohol dehydrogenase gene region of *Drosophila melanogaster*. Genetics 114:1165–1190.

Avise, J. C. 1989. Gene trees and organismal histories: A phylogenetic approach to population biology. Evolution 43:1192–1208.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. Ann. Rev. Ecol. Syst. 18:489–522.

Avise, J. C., R. M. Ball, and J. Arnold. 1988. Current versus historical population sizes in vertebrate species with high gene flow: A comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Molec. Biol. Evol. 5: 331–344.

Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217.

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

Lemke, K. 1985. Dispersal models for *Drosophila*. Ph.D. Diss. Univ. Georgia, Athens.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia Univ. Press, N.Y.

Sokal, R. R., and F. J. Rohlf. 1981. Biometry, 2nd Ed. Freeman, San Francisco, CA.

Tajima, F. 1983. Evolutionary relationships of DNA sequences in finite populations. Genetics 105:437–460.

Corresponding Editor: P. W. Hedrick