Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human Mitochondrial DNA Restriction Data

Laurent Excoffier,*,† Peter E. Smouse* and Joseph M. Quattro*,‡

**Center for Theoretical and Applied Genetics (CTAG), Cook College, Rutgers University, New Brunswick, New Jersey 08903-0231, †Department of Anthropology and Ecology, University of Geneva, 1227 Carouge, Switzerland, and ‡Department of Biological Sciences, Hopkins Marine Station, Stanford University, Pacific Grove, California 93950*

Manuscript received November 1, 1991
Accepted for publication February 10, 1992

ABSTRACT

We present here a framework for the study of molecular variation within a single species. Information on DNA haplotype divergence is incorporated into an analysis of variance format, derived from a matrix of squared-distances among all pairs of haplotypes. This analysis of molecular variance (AMOVA) produces estimates of variance components and *F*-statistic analogs, designated here as Φ-statistics, reflecting the correlation of haplotypic diversity at different levels of hierarchical subdivision. The method is flexible enough to accommodate several alternative input matrices, corresponding to different types of molecular data, as well as different types of evolutionary assumptions, without modifying the basic structure of the analysis. The significance of the variance components and Φ-statistics is tested using a permutational approach, eliminating the normality assumption that is conventional for analysis of variance but inappropriate for molecular data. Application of AMOVA to human mitochondrial DNA haplotype data shows that population subdivisions are better resolved when some measure of molecular differences among haplotypes is introduced into the analysis. At the intraspecific level, however, the additional information provided by knowing the exact phylogenetic relations among haplotypes or by a nonlinear translation of restriction-site change into nucleotide diversity does not significantly modify the inferred population genetic structure. Monte Carlo studies show that site sampling does not fundamentally affect the significance of the molecular variance components. The AMOVA treatment is easily extended in several different directions and it constitutes a coherent and flexible framework for the statistical analysis of molecular data.

OUR knowledge of population genetic diversity has improved considerably over the last decade, with the application of molecular techniques to evolutionary studies. Quantitative resolution has improved as larger numbers of haplotypic markers are defined within each sample. Moreover, information on the degree of divergence between alleles/restriction haplotypes/DNA sequences has become available. Whenever we can make mutational or recombinational assumptions about the relationships among haplotypes, special phylogenetic reconstruction algorithms are available to characterize evolutionary relationships more precisely (see reviews by FELSENSTEIN 1988; SWOFFORD and OLSEN 1990).

Although no precise analytic model for the full population distribution of molecular differences among a set of interconnected haplotypes is known, the expected mean number of site differences between sets of panmictic (WATTERSON 1975) and subdivided (SLATKIN 1987) populations has been derived under simple assumptions. When a species exhibits subdivision, we expect both increased haplotypic diversity and a larger number of segregating sites for genomes sampled from different demes (SLATKIN 1987). The use of information on the molecular connection of DNA haplotypes should be valuable in population genetic analyses.

Population genetic structure within a species has traditionally been studied using departures of allele frequencies from panmictic expectations. Several estimation procedures related to WRIGHT's (1951, 1965) *F*-statistics have been proposed for the treatment of polymorphic systems (COCKERHAM 1969, 1973; NEI 1977; WEIR and COCKERHAM 1984; LONG 1986). A few studies have tried to translate information on DNA restriction endonuclease haplotypes into estimates of the magnitude of intraspecific subdivision. LYNCH and CREASE (1990), using a phylogeny of haplotypes, provide estimates of the variance of nucleotide diversity for different sampling processes. TAKAHATA and PALUMBI (1985) compute the fraction of nucleotide diversity due to interpopulation genetic differences and provide an analogue of NEI's (1973) coefficient of gene differentiation (G_{ST}). Both methods involve nonlinear transformation of the original data set into estimates of genetic diversity. Several

assumptions on the underlying evolution of the molecule are required, assumptions that are neither always met nor generally verifiable. We need a more general methodology that does not depend so critically on the specific assumptions.

Our purpose here is to design an alternative methodology that makes use of the available molecular information gathered in population surveys, while remaining flexible enough to accommodate different types of assumptions about the evolution of the genetic system. We propose to extend the work of COCKERHAM (1973), LONG (1986) and LONG, SMOUSE and WOOD (1987) on allelic correlations among demes to a comparable analysis of haplotypic diversity. Using the fact that a conventional sum of squares (SS) may be written as the sum of squared differences between all pairs of observations (LI 1976), we construct a hierarchical analysis of molecular variance directly from the matrix of squared-distances between all pairs of haplotypes. Beyond its clear relation to an analysis of variance, the method has the additional advantage that several different assumptions can be imposed on the haplotype differentiation process, each of which translates into a different distance matrix, with no change in the structure of the subsequent analysis. When all interhaplotypic distances are presumed equal, the analysis is tantamount to a multiallelic (multivariate) analysis of variance (see WEIR and COCKERHAM 1984; LONG 1986; LONG, SMOUSE and WOOD 1987). Alternatively, we can use the mean number of restriction site differences, patristic distances along a given network, or nucleotide diversity as measures of interhaplotypic distances.

We illustrate with an analysis of human mitochondrial DNA (mtDNA) restriction site data, performing a nested analysis of molecular variance on five regional collections, each represented by two different populations. The hierarchical model employs "Within Populations" (WP), "Among Populations/Within Groups" (AP/WG), and "Among Groups" (AG) components of diversity. To illustrate the impact of different sets of assumptions concerning the origins of the haplotypic variants, we employ alternative distance metrics to examine the amount and pattern of genetic subdivision. We use permutational procedures on the original interindividual squared distance matrix to provide significance tests for each of the hierarchical variance components and related F-statistic analogues. We also study the importance of site choice on the significance of the different statistics, using resampling techniques (EFRON 1982).

METHODOLOGICAL DEVELOPMENTS

Phenetic distances between restriction haplotypes: We assume that restriction analysis has been performed on a non-recombining DNA segment. For

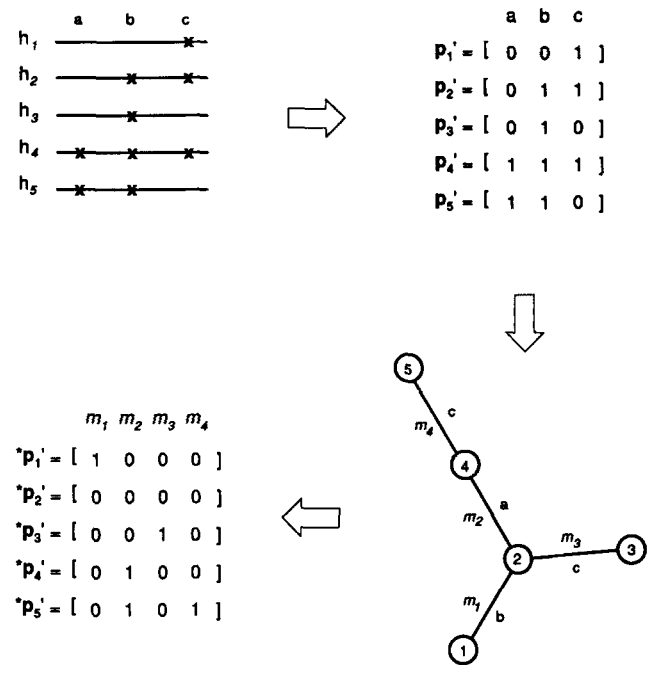

FIGURE 1.—Example of the steps involved in the computations of the boolean vectors $*\mathbf{p}_j$ of mutational events. Each haplotype (h_j) is first translated into a boolean vector ($\mathbf{p}_j$) of presence or absence of restriction sites. The second step involves the construction of a phylogenetic network, where each haplotype is linked by a single or a series of mutational events to all the other haplotypes through a unique pathway. The final step is the coding of each haplotype (h_j) as a boolean vector ($*\mathbf{p}_j$) of occurrences of mutational events (m_i) from a given haplotype chosen as a reference to h_j. In this example, haplotype 2 has been chosen as the reference.

N individuals assayed with a standard set of restriction enzymes, S polymorphic restriction sites are identified. A restriction haplotype (h), defined as the combination of presence or absence of the various restriction sites, may be considered as an S-dimensional boolean vector of the form

$$\mathbf{p}' = [p_1 p_2 p_3 p_4 \ldots p_S], \tag{1}$$

where $p_s = 1$ if h is cut at site s, and $p_s = 0$ if it is not (upper right, Figure 1). The difference between two haplotypes h_j and h_k is then defined as $(\mathbf{p}_j - \mathbf{p}_k)$

$$(\mathbf{p}_j - \mathbf{p}_k)' = [(p_{1j} - p_{1k})(p_{2j} - p_{2k}) \ldots (p_{Sj} - p_{Sk})]. \tag{2}$$

Each polymorphic site contributes additional information, without necessarily being evolutionarily independent. We define a Euclidean distance metric (δ^2_{jk}) between haplotypes h_j and h_k as

$$\delta^2_{jk} = (\mathbf{p}_j - \mathbf{p}_k)'\mathbf{W}(\mathbf{p}_j - \mathbf{p}_k), \tag{3a}$$

where $\mathbf{W}$ is a matrix of differential weights for the various sites. The weight matrix $\mathbf{W}$ takes any of several forms, depending upon how we wish to use ancillary information. If all sites are assumed independent and equally informative, $\mathbf{W} = \mathbf{I}$, the identity matrix, and the distance metric is equal to the number of restric-

tion-site differences. This Euclidean metric is commonly employed for population differences (NEI and TAJIMA 1981), but it may be used just as easily for differences between single haplotypes. In the case where $\mathbf{W}$ is diagonal, $\mathbf{W} = \text{diag}\{w_s^2\}$, weighting sites differentially but treating them as independent, Equation 3a can be rewritten as

$$\delta_{jk}^2 = \sum_{s=1}^{S} w_s^2(p_{sj} - p_{sk})^2. \tag{3b}$$

The rest of the analysis does not depend on which particular form of $\mathbf{W}$ has been chosen; we will assume that the weight matrix has been set in advance, returning to the definition of metrics and the choice of $\mathbf{W}$ for the human illustration.

Evolutionary distances between haplotypes: The DNA haplotypes can sometimes be related mutationally and arranged into a network (see Figure 1). We may then use the number of mutations along the network as a measure of evolutionary divergence between any two haplotypes. Network distance does not generally equal phenetic distance, either because of homoplasy (convergent site changes or reverse mutations) or because the translation from the changes we see to those we infer are nonlinear (*e.g.*, TAKAHATA and PALUMBI 1985; LYNCH and CREASE 1990). We can always modify the definition of the $\mathbf{p}$'s and $\mathbf{W}$ in such a way that we can apply Equation 3a to provide an evolutionary distance. We define a given haplotype as a vector ($*\mathbf{p}$) of independent and unique mutational events, described sequentially from any fixed position in the network (lower left, Figure 1), rather than as a vector of restriction-site presence or absence indicators (upper right, Figure 1). If M mutational events are recognized, each haplotype is defined as a vector of dimension $M \geqq S$. Our evolutionary distance metric becomes

$$*\delta_{jk}^2 = (*\mathbf{p}_j - *\mathrm{p}_k)'\mathbf{W}(*\mathbf{p}_j - *\mathbf{p}_k). \tag{3c}$$

In the absence of homoplasy and keeping $\mathbf{W}$ constant, $*\delta_{jk}^2$ is the same as δ_{jk}^2. The important point is that once a metric has been set, the following is general.

Partitioning a distance matrix into hierarchical components: Our application will concern mtDNA data. Consider a haploid genetic system where interhaplotypic distances are identical to distances between individuals. We can arrange a set of N individuals from I populations into a distance matrix, $\mathbf{D}^2$, partitioned into a series of submatrices corresponding to particular subdivisions:

$$\mathbf{D}^2 = \begin{bmatrix} \begin{bmatrix} [\mathbf{D}_{11}^2] & [\mathbf{D}_{12}^2] \\ [\mathbf{D}_{21}^2] & [\mathbf{D}_{22}^2] \end{bmatrix} & \cdots & [\mathbf{D}_{1I}^2] \\ \cdots & & \cdots \\ [\mathbf{D}_{I1}^2] & \cdots & [\mathbf{D}_{II}^2] \end{bmatrix}, \tag{4}$$

where the elements of the block-diagonal submatrices $\mathbf{D}_{ii}^2$ contain pairwise squared-distances (δ_{jk}^2) between individuals of the same (ith) population, and those of the off-diagonal matrix blocks $\mathbf{D}_{ii'}^2$ contain pairwise squared-distances between individuals, one from the ith and the other from the i'th population. Individuals may also be grouped at higher levels, according to such non-genetic criteria as geography, ecological environment, or language.

A conventional sum of squares [$\text{SS}_{(\text{Total})}$] may be written, barring a constant ($2N$), as the sum of squared differences between all pairs of N items (LI 1976). In the multidimensional case, using vectors instead of scalars, the conventional sum of squares becomes a sum of squared deviations (SSD) from the centroid of a multidimensional space. Thus,

$$\begin{aligned} \text{SSD}_{(\text{Total})} &= \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{W}(\mathbf{x}_j - \bar{\mathbf{x}}) \\ &= \frac{1}{N}\sum_{j=1}^{N-1}\sum_{k>j}^{N} (\mathbf{x}_j - \mathbf{x}_k)'\mathbf{W}(\mathbf{x}_j - \mathbf{x}_k), \end{aligned} \tag{5a}$$

or

$$\begin{aligned} \text{SSD}_{(\text{Total})} &= \frac{1}{2N}\sum_{j=1}^{N}\sum_{k=1}^{N} (\mathbf{x}_j - \mathbf{x}_k)'\mathbf{W}(\mathbf{x}_j - \mathbf{x}_k) \\ &= \frac{1}{2N}\sum_{j=1}^{N}\sum_{k=1}^{N} \delta_{jk}^2, \end{aligned} \tag{5b}$$

because $\delta_{jj}^2 = 0$ for all haplotypes h_j. This transformation applies equally to the total array of individuals in the data set, to those within each population separately (within the diagonal blocks, $\mathbf{D}_{ii}^2$), and to those belonging to a particular subdivision (within the diagonal blocks, $\mathbf{D}_{11}^2$, $\mathbf{D}_{12}^2$, $\mathbf{D}_{21}^2$, and $\mathbf{D}_{22}^2$).

Where individuals are arranged into populations and populations nested within groups defined *a priori* on nongenetic criteria, we employed a linear model on the pattern first described by COCKERHAM (1969, 1973) and refined upon by others (WEIR and COCKERHAM 1984; LONG 1986)

$$\mathbf{p}_{jig} = \mathbf{p} + \mathbf{a}_g + \mathbf{b}_{ig} + \mathbf{c}_{jig}, \tag{6}$$

where $\mathbf{p}_{jig}$ indexes the jth chromosome, here equivalent to the jth individual ($j = 1, \ldots, N_{ig}$) in the ith population ($i = 1, \ldots, I_g$) in the gth group ($g = 1, \ldots, G$), and $\mathbf{p}$ is the unknown expectation of $\mathbf{p}_{jig}$, averaged over the whole study. The effects are a for group, b for populations and c for individuals within populations. The effects are assumed to be additive, random, uncorrelated, and to have the associated variance components (expected squared deviations) σ_a^2, σ_b^2, and σ_c^2, respectively.

Relying on the standard decomposition, we note that for any choice of hierarchical partition of the N individuals into strata, we can write

$$\begin{aligned} \text{SSD(Total)} = \text{SSD(Among Strata)} \\ + \text{SSD(Within Strata)}, \end{aligned} \tag{7}$$

placing us in traditional analysis of variance framework, designated here as Analysis of Molecular Variance, AMOVA (Table 1). For illustration, we shall partition the total sum of squared deviations, SSD(Total), into components for variation within populations, SSD(WP), variation among populations within regional groups, SSD(AP/WG), and variation among regional groups, SSD(AG)

$$\text{SSD(WP)} = \sum_{g=1}^{G}\sum_{i=1}^{I_g}\frac{\sum_{j=1}^{N_{ig}}\sum_{k=1}^{N_{ig}}\delta_{jk}^2}{2N_{ig}} \tag{8a}$$

SSD(AP/WG)

$$= \sum_{g=1}^{G}\left(\frac{\sum_{i=1}^{I_g}\sum_{j=1}^{N_{ig}}\sum_{i'=1}^{I_g}\sum_{k=1}^{N_{i'g}}\delta_{jk}^2}{\sum_{i=1}^{I_g}2N_{ig}} - \sum_{i=1}^{I_g}\frac{\sum_{j=1}^{N_{ig}}\sum_{k=1}^{N_{ig}}\delta_{jk}^2}{2N_{ig}}\right) \tag{8b}$$

$$\text{SSD(AG)} = \left(\frac{\sum_{j=1}^{N}\sum_{k=1}^{N}\delta_{jk}^2}{2N} - \sum_{g=1}^{G}\frac{\sum_{i=1}^{I_g}\sum_{j=1}^{N_{ig}}\sum_{i'=1}^{I_g}\sum_{k=1}^{N_{ig}}\delta_{jk}^2}{\sum_{i=1}^{I_g}2N_{ig}}\right) \tag{8c}$$

The corresponding mean squared deviations (MSD) are obtained by dividing each SSD by the appropriate degrees of freedom, as reported in Table 1. The n-coefficients in Table 1 represent the average sample sizes of particular hierarchical levels, allowing for unequal sample sizes,

$$n = \frac{\sum_{g=1}^{G}\sum_{i=1}^{I_g}N_{ig} - \sum_{g=1}^{G}\left(\frac{\sum_{i=1}^{I_g}N_{ig}^2}{\sum_{i=1}^{I_g}N_{ig}}\right)}{\sum_{g=1}^{G}I_g}, \tag{9a}$$

$$n' = \frac{\sum_{g=1}^{G}\left(\frac{\sum_{i=1}^{I_g}N_{ig}^2}{\sum_{i=1}^{I_g}N_{ig}}\right) - \frac{\sum_{g=1}^{G}\sum_{i=1}^{I_g}N_{ig}^2}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}N_{ig}}}{G-1}, \tag{9b}$$

$$n'' = \frac{\sum_{g=1}^{G}\sum_{i=1}^{I_g}N_{ig} - \frac{\sum_{g=1}^{G}\left(\sum_{i=1}^{I_g}N_{ig}\right)^2}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}N_{ig}}}{G-1}. \tag{9c}$$

The variance components (σ^2's) of each hierarchical level are extracted by equating the mean squares (MSDs) to their expectations. The structure of the analysis is that described for F-statistics (COCKERHAM 1969, 1973), but it allows for the haploid transmission of mitochondrial genomes. It may also be useful to employ an analogous array of haplotypic correlation measures, which we shall term Φ-statistics to avoid confusion. Following COCKERHAM's lead, we have

$$\begin{aligned}\sigma_c^2 &= (1 - \Phi_{ST})\sigma^2,\\ \sigma_b^2 &= (\Phi_{ST} - \Phi_{CT})\sigma^2,\\ \sigma_a^2 &= \Phi_{CT}\sigma^2,\end{aligned} \tag{10a}$$

where $\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$; Φ_{ST} is viewed as the correlation of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the whole species; Φ_{CT} as the correlation of random haplotypes within a group of populations, relative to that of random pairs of haplotypes drawn from the whole species, and Φ_{SC} as the correlation of the molecular diversity of random haplotypes within populations, relative to that of random pairs of haplotypes drawn from the region. Still following the analogy, we rewrite the equations (10a) in terms of the Φ-statistics

$$\Phi_{ST} = \frac{\sigma_a^2 + \sigma_b^2}{\sigma^2}, \quad \Phi_{CT} = \frac{\sigma_a^2}{\sigma^2}, \quad \Phi_{SC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_c^2}. \tag{10b}$$

We shall not require it for mtDNA, but for the case of diploid genetic systems, the procedure employs within-individual haplotypic diversity as an additional level, following COCKERHAM (1973) and LONG (1986) exactly. The only difficulty is that DNA haplotype diversity within nuclear genes is often assayed from homozygous individuals, to avoid confusion over linkage phase in multisite heterozygotes. If one cannot avoid the resulting sampling biases, one should probably avoid the within-individual level of the hierarchy. The limitations arising from the precise assumptions of the F-statistics treatment (random sampling to create the initial subdivisions at each level, pure drift and no migration) are almost never met in natural populations. The same comments apply to the Φ-statistics. Proper caution is necessary when interpreting these coefficients, but we may nevertheless view them as convenient summarizations of the packaging of genetic information within and among populations, being one for one with the variance components.

Testing significance of the variance components and Φ-statistics: Considerable discussion has emerged over which method to use for testing the significance of the variance-components (WEIR and COCKERHAM 1984; LONG 1986; ZHIVOTOVSKY 1988). The method requiring the fewest assumptions is permutational analysis of the null distribution for each variance-component. Under the null hypothesis, samples are considered as drawn from a global population, with variation due to random sampling in the construction of populations. To obtain a null distribution, we allo-

TABLE 1

General design for hierarchical analysis of molecular variance (AMOVA)

Source of variation	d.f.	MSD	Expected MSD
Among regions	$G - 1$	MSD/(AG)	$\sigma_c^2 + n' \sigma_b^2 + n'' \sigma_a^2$
Among populations within regions	$\sum_{g=1}^{G} I_g - G$	MSD/(AP/WG)	$\sigma_c^2 + n \sigma_b^2$
Among individuals within populations	$N - \sum_{g=1}^{G} I_g$	MSD/(WP)	σ_c^2
Total	$N - 1$		

cate each individual to a randomly chosen population, while holding sample sizes constant at the realized values. This amounts to random permutation of the rows (and corresponding columns) of the squared-distance matrix (MANTEL 1967). The variance-components are estimated from each of a large number (say 1000) of permuted matrices. We use this procedure to obtain the null distribution and to test for the significance of Φ_{ST} and σ_c^2.

Two other permutation schemes are useful. The first assumes that the regions are real but that the populations within them are not, permuting individuals within regional groups without regard to population, a procedure used to obtain the null distributions of Φ_{SC} and σ_b^2. The second assumes that while the populations are real, the regional groupings are artificial, permuting whole populations across groups. In this case, the sizes of the groups (but not those of the populations) vary with each permutational run. This randomization scheme is used to obtain the null distribution of Φ_{CT} and σ_a^2.

Restriction site sampling: The sampling of nucleotides has been shown to be a major source of variability for estimates of molecular diversity (LYNCH and CREASE 1990). One can legitimally ask whether the results of our study depend on the particular array of restriction sites employed. We examine the influence of site sampling on the genetic structure of the populations, using a site resampling plan similar to the bootstrap used by EFRON (1982). Under the assumption that the observed 62 sites are representative of all potential mtDNA sites, we obtain the distribution of the variance components and associated Φ-statistics by Monte Carlo simulation, using 500 random collections of sites. For each collection, the procedure is as follows: (a) Draw a given number of sites from the observed array of 62 sites, at random and with replacement. Given the choice of sites, the haplotype of each individual is then taken as the combination of the original states of those randomly chosen sites; (b) compute interhaplotypic distances on the basis of the newly defined haplotypes and perform an AMOVA analysis. The distances are simply computed from Equation 3b, with all w_s^2 equal to 1; and (c) permute the matrix 500 times, and test the significance of the different statistics with the previously described procedures.

ILLUSTRATION WITH HUMAN mtDNA HAPLOTYPES

Due to its high relative mutation rate (BROWN, GEORGE and WILSON 1979; BROWN *et al.* 1982), mtDNA presents many distinct haplotypes in different demes. Prevailing maternal transmission in mammals (GILES *et al.* 1980; GYLLENSTEN *et al.* 1991) favors higher levels of population subdivision than is true for nuclear DNA markers (BIRKY, MARUYAMA, and FUERST 1983; BIRKY, FUERST and MARUYAMA 1989). Barring migration, these two effects should produce increasingly non-overlapping sets of restriction haplotypes as divergence time between populations increases (WATTERSON 1985). Both of these features are evident in human mtDNA, which is small (16,569 bp, ANDERSON *et al.* 1981), rapidly evolving, and apparently free of recombination.

Restriction haplotypes of human mtDNA have been sampled from a substantial number of populations (for a review of the two main data set, see EXCOFFIER 1990; STONEKING *et al.* 1990). Our purpose is to illustrate the methodology described above, rather than to reopen the question of human origins raised elsewhere (CANN, STONEKING and WILSON 1987; EXCOFFIER and LANGANEY 1989; EXCOFFIER 1990; STONEKING *et al.* 1990). We consider here ten populations for which ample data are available in the literature (Table 2). These particular populations were chosen to represent five "regional groups" of two populations each (Figure 2). The samples have also been analyzed for polymorphism with the five restriction enzymes most commonly used in human studies, *Bam*HI: GGATCC, *Hpa*I: GTTAAC, *Hae*II: (A/G)GCGC(T/C), *Ava*II: GG(T/A)CC, and *Msp*I: CCGG. Among the 672 mtDNAs assayed from these ten populations, 34 of 62 recognizable sites were found to be polymorphic.

In a sample of 672, we cannot expect to see all 2^{34} possible haplotypes, but sample size considerations aside, the absence of recombination practically guarantees large amounts of disequilibrium among the 34

TABLE 2

Haplotypic composition of the population samples by region

Sample No.	Sample name	Reference	Sample size	Haplotype frequencies[a]
Asia				
				1 8 9 13 28 47 48 49 50 51 52 53 54
1	Tharu	BREGA *et al.* (1986)	91	*48 2 5 23 2 2 2 1 1 1 2 1 1*
				1 6 8 9 12 13 27 28 29
2	Oriental	JOHNSON *et al.* (1983)	46	*32 1 2 4 2 2 1 1 1*
West Africa				
				1 2 7 10 27 39 52 64 65 66 67 68 71
3	Wolof	SCOZZARI *et al.* (1988)	110	*23 39 29 2 2 5 2 2 1 1 1 2 1*
				1 2 6 8 34 39 69
4	Peul	SCOZZARI *et al.* (1988)	47	*11 19 12 2 1 1 1*
America				
				1 6 39 46
5	Pima	WALLACE, GARRISON and KNOWLER (1985)	63	*59 2 1 1*
				1 47 95
6	Maya	SCHURR *et al.* (1990)	37	*30 4 3*
Europe				
				1 6 11 18 21 38 47 82 83
7	Finnish	VILKKI, SAVONTAUS and NIKOSKELAINEN (1988)	110	*87 2 4 3 8 2 2 1 1*
				1 2 6 18 21 23 34 42 47 56 57 72 73 75 76 77
8	Sicilian	SEMINO *et al.* (1989)	90	*50 3 9 11 1 1 1 1 5 1 2 1 1 1 1 1*
Middle-East				
				1 6 11 17 22 36 37 38 39
9	Israeli Jews	BONNÉ-TAMIR *et al.* (1986)	39	*15 14 1 1 4 1 1 1 1*
				1 2 6 7 22 31 40 41 42 43 44 45
10	Israeli Arabs	BONNÉ-TAMIR *et al.* (1986)	39	*22 1 1 1 6 2 1 1 1 1 1 1*
			672	

[a] For each population, haplotype numbers are reported on the first line and their absolute frequencies are shown in italic on the second line.

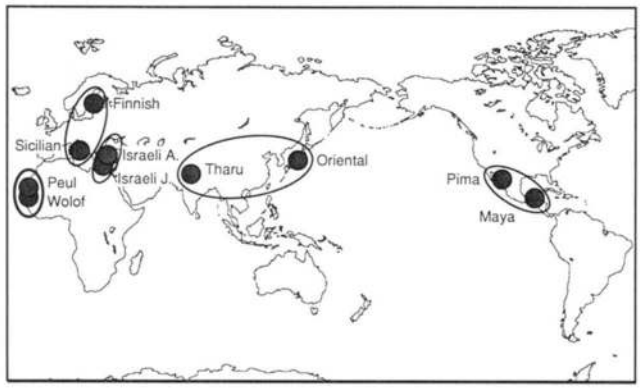

FIGURE 2.—Geographic location of the population samples.

sites. The treatment we have developed above does not require independence of the restriction sites. Only 35 haplotypes would be observed if each site had been the subject of a single mutational event; there is a high level of homoplasy. Nevertheless, all 56 haplotypes may be linked by single mutational events in a parsimonious network (Figure 3), with only two missing intermediates. Neither of these missing haplotypes (probably representing extinct intermediates, rather than sampling holes) has been found in human studies to date. These 56 haplotypes are a subset of a much larger world-wide collection reviewed in EXCOFFIER (1990). The network presented in Figure 3 is a minimum spanning tree (PRIM 1957), obtained by the algorithm found in the NTSYS package (ROHLF 1990). The procedure is similar to that producing Wagner trees (FARRIS 1970), but differs by using the observed haplotypes as the nodes of the network, rather than as branch tips of the tree. Wagner trees and Prim networks are alternative ways of viewing the same data, but the network better conveys the connections between the haplotypes.

Haplotypic diversity among samples is pictured in Figure 4, where the darkened circles indicate presence of a given haplotype in a particular population sample. A common feature of each sample is the presence of type 1 (the large central circle) in substantial frequencies. Other, less common haplotypes (2, 6, 7, 11, 39), are found in samples from different geographic regions. Each sample also possesses a series of private haplotypes, restricted to a single sample and not found elsewhere. Populations within a region tend to occupy similar portions of the network, sharing more than one haplotype, and differing by small mutational

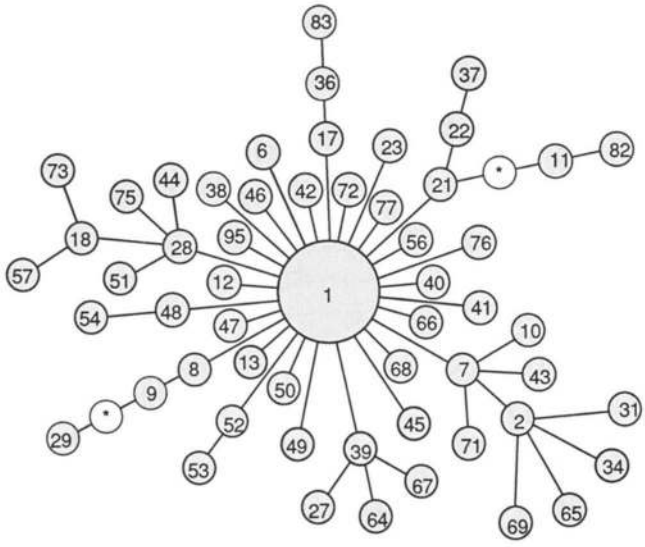

FIGURE 3.—Minimum spanning network of 56 haplotypes found among 10 populations. Each link between haplotype represents a unique mutational event. Two haplotypes marked with asterisks have not been found among sampled human populations. The designation of each haplotype follows that of the publication where they have been first described (listed in Table 2). The universal haplotype 1 has been enlarged for easy recognition.

steps. Populations in different regions tend to occupy different (although partially overlapping) parts of the network. Regional diversification represents both haplotype frequency changes and some degree of phyletic radiation, probably smoothed by gene flow.

Alternative definitions of the distance metric: We have performed hierarchical analysis of variance on four different matrices of inter-haplotypic squared distances, computed from different assumptions about the evolutionary process that produced the mtDNA haplotypes. The four matrics are: $\mathbf{D}_1$ (a standard Euclidean metric counting the differences among haplotypes), $\mathbf{D}_2$ (an equidistant metric based on the idea that haplotypes are merely distinguishable), $\mathbf{D}_3$ (a distance measured along the network, but also incorporating additional geographic and probabilistic information), and $\mathbf{D}_4$ (a matrix allowing for nonlinearity of changes along the network).

$\mathbf{D}_1$: This first input matrix is based on a phenetic distance metric, amounting to a simple count of the number of restriction-site differences between two haplotypes. One would choose this type of metric when the identities of restriction-sites are well defined and some haplotypes are clearly more different than others but where no network connecting the haplotypes is available. The results of our hierarchical partition are reported in Table 3 under $\mathbf{D}_1$. The proportion of the "among regions" variance component is large (21.12%), but the "among populations/within region" percentage is low (3.49%), relative to the "within populations" variance component. All three variance components are highly significant. We present the null distributions of σ_a^2, σ_b^2 and σ_c^2 in Figure 5, obtained by the three different permutation procedures described above. The null distributions of Φ-statistics are highly correlated with those of the associated variance components [$\mathrm{Corr}(\sigma_a^2,\Phi_{CT}) > 0.99$; $\mathrm{Corr}(\sigma_b^2,\Phi_{SC}) > 0.99$; $\mathrm{Corr}(\sigma_c^2,\Phi_{ST}) < -0.99$] and would thus have virtually identical shapes. For the permutation of whole populations across regions, testing σ_a^2 and Φ_{CT}, there are 945 possible ways of allocating 10 populations to five groups of two populations each ($10!/(5!\,2^5)$). Only one combination of populations was found to give a slightly larger value than the observed σ_a^2. As shown in Figure 5a, the null distribution is clearly bimodal. A certain number of other combinations also give σ_a^2 values that are almost as large as our observed value. Interestingly, all combinations of this higher peak show the two African populations grouped together in a single region. On the contrary, each time Peul and Wolof populations are separated in different regions, σ_a^2 values are small and found in the lower peak around zero. Large regional diversity may then be attributed to differences between the African group and all other regions, the composition of which is of no real importance. This fact would not have emerged from a standard F-ratio test. In the combination giving a maximum σ_a^2 value, the Asiatic, Middle-Eastern, and Western African groups are preserved, but the Pima are grouped with Finns and the Maya with Sicilians. Although clearly significant, our arbitrarily chosen geographic groupings are not optimum for maximizing the "among region" diversity.

The AMOVA treatment on the input distance matrix $\mathbf{D}_1$ has close connections with TAKAHATA and PALUMBI's (1985) technique, which leads to a G_{ST} analog, after a nonlinear transformation of restriction-site changes into nucleotide diversity estimates. Without entering into much detail, we would merely point out that TAKAHATA and PALUMBI's equation (17), defining an affinity measure within populations ($\hat{I}$), may be modified as an affinity measure between any two haplotypes j and k ($\hat{I}_{jk}$) by letting TAKAHATA and PALUMBI's variable l be the total number of restriction-sites present in the whole collection of haplotypes, rather than that for the specific pair of haplotypes j and k. Following the analogy, we also need to define an affinity measure between an "individual and itself." The most convenient definition is the number of restriction sites for that individual, the definition most in keeping with the spirit of TAKAHATA and PALUMBI (1985). These simple changes preserve the Euclidian closure of the inter-haplotypic distance measure if we use $d_{jk}^2 = \hat{I}_{jj} + \hat{I}_{kk} - 2\,\hat{I}_{jk}$, which turns out to be identical to our phenetic distance δ_{jk}^2, defined in (3a).

$\mathbf{D}_2$: This second input matrix assumes that all haplotypes are equidistant. The evolutionary relations

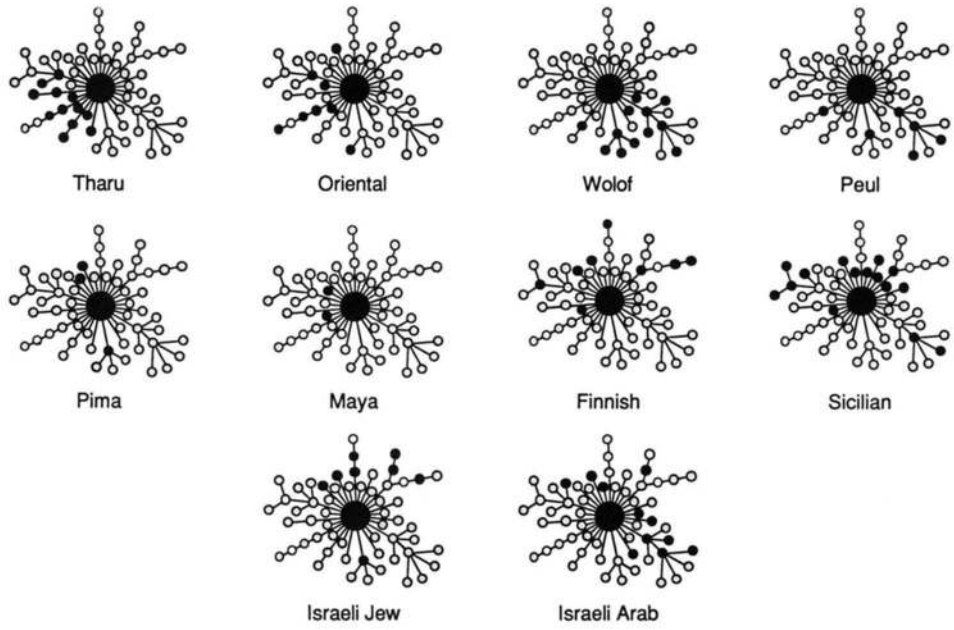

FIGURE 4.—Haplotypic diversity of each of the 10 population samples. The position of the haplotypes are identical for each population and are homologous to those of Figure 3. The haplotypes found within each population sample are shown as black circles.

TABLE 3

Hierarchical analysis of variance on four different square matrices of distances between haplotypes

		$\mathbf{D}_1$ (haplotypic)				$\mathbf{D}_2$ (multiallelic)			
		Observed partition				Observed partition			
Variance component		Variance	% total	P^a	Φ-statistics	Variance	% total	P^a	Φ-statistics
Among regions	σ_a^2	0.134	21.12	0.002	$\Phi_{CT} = 0.211$	0.055	15.73	0.008	$\Phi_{CT} = 0.157$
Among populations/regions	σ_b^2	0.022	3.49	<0.0001	$\Phi_{SC} = 0.044$	0.013	3.59	<0.0001	$\Phi_{SC} = 0.043$
Within populations	σ_c^2	0.478	75.39	<0.0001	$\Phi_{ST} = 0.246$	0.281	80.68	<0.0001	$\Phi_{ST} = 0.193$
		$\mathbf{D}_3$ (Prim network)				$\mathbf{D}_4$ (nonlinear)			
Among regions	σ_a^2	0.142	21.99	0.002	$\Phi_{CT} = 0.220$	$0.127\ 10^{-5}$	21.30	0.002	$\Phi_{CT} = 0.213$
Among populations/regions	σ_b^2	0.021	3.29	<0.0001	$\Phi_{SC} = 0.042$	$0.020\ 10^{-5}$	3.31	<0.0001	$\Phi_{SC} = 0.042$
Within populations	σ_c^2	0.484	74.72	<0.0001	$\Phi_{ST} = 0.253$	$0.449\ 10^{-5}$	75.39	<0.0001	$\Phi_{ST} = 0.246$

[a] Probability of having a more extreme variance component and Φ-statistic than the observed values by chance alone. Φ_{CT} and σ_a^2 are tested under random permutations of whole populations across regions. Φ_{SC} and σ_b^2 are tested under random permutation of individuals across populations but within the same region. Φ_{ST} and σ_c^2 are tested under random permutation of individuals across populations without regard to either their original populations or regions.

between distinguishable haplotypes are assumed to be unknown, a standard treatment for allozymes or other protein systems (see, however, RICHARDSON and SMOUSE 1976; RICHARDSON, SMOUSE and RICHARDSON 1977). This treatment is also applicable to antigenic systems, or even to molecular fingerprint analysis, where the banding pattern of two individuals either matches or does not. The Φ-statistics become the usual multiallelic *F*-statistics (LONG 1986). The results of our hierarchical analysis are presented in Table 3 under $\mathbf{D}_2$. Most of the haplotype diversity (80.68%) is found within each population, but an appreciable amount still (15.73%) separates regions. The differences among populations within regions are small (3.59%). For the two procedures that involve permutation of individuals across populations, testing σ_b^2, σ_c^2, Φ_{ST} and Φ_{SC}, our observed variance components showed extreme values in all cases. Seven permutations of whole populations across regions were found to yield greater σ_a^2 (and Φ_{CT}) than our observed value. Although the result is still significant, we clearly lose geographic resolution with this metric.

$\mathbf{D}_3$: Our third matrix is based on a distance metric computed along the evolutionarily parsimonious network shown in Figure 3. When several connections of equal length are possible for a particular haplotype, two additional rules are used to make a choice (EXCOFFIER and LANGANEY 1989). The first is a probability criterion; a link between two rare (<5%) haplotypes is less likely than a link between rare and frequent (>5%) haplotypes. The second criterion is geographic; links between haplotypes that are found

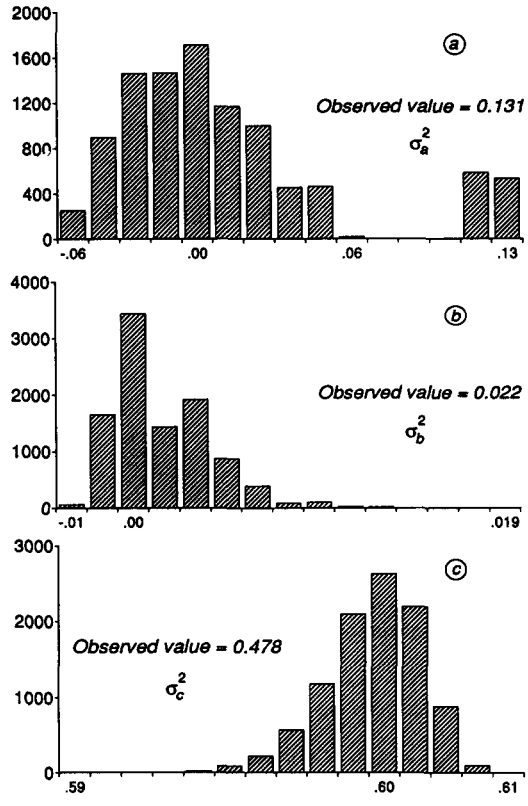

FIGURE 5.—Null distributions of the molecular variance components obtained through different random permutations of the large matrix of squared interindividual distances $\mathbf{D}_1$ of dimension 672 (see text). (a) Distribution of σ_a^2; (b) distribution of σ_b^2; (c) distribution of σ_c^2.

within the same population or within the same region are favored over links between types from different regions. These distances differ from those of $\mathbf{D}_1$ whenever we have homoplasic mutations along the network (23 cases out of 57). As exemplified in Figure 1 for the differences between haplotypes 3 and 5, single-site changes from (+) to (−) to (+) or from (−) to (+) to (−) along the network, scored as a distance of zero for $\mathbf{D}_1$, are scored as a distance of 2 for $\mathbf{D}_3$. The results are presented in Table 3, labeled as $\mathbf{D}_3$. The observed σ_a^2 value accounts for a slightly larger fraction of the total genetic variability (21.99%) than is the case for $\mathbf{D}_1$, but there is again one regional combination of populations which produces a larger σ_a^2 than that observed, and it is the same one described before (Pima + Finns and Maya + Sicilians). The handling of homoplasies does not modify the outcome.

$\mathbf{D}_4$: This fourth input matrix is made up of weighted evolutionary distances, measured along the PRIM network shown in Figure 3. The weighting matrix ($\mathbf{W}$) is now a diagonal matrix of dimension $M = 55$ (total number of haplotypes − 1), where each $\mathbf{w}_{mm}$ is equal to the nucleotide diversity (d_{xy}) between adjacent types x and y in the network, so that $\mathbf{W} = \text{diag}\{d_{xy}^2\}$. Different methods have been used to estimate nucleotide diversity (d_{xy}^2) from restriction-site data (ENGELS 1981; EWENS, SPIELMAN and HARRIS 1981; NEI and TAJIMA 1981, 1983; KAPLAN 1983; NEI and MILLER 1990). For simplicity, we have used Equation 4 from NEI and MILLER (1990), which yields results very close to the maximum-likelihood estimates of NEI and TAJIMA (1983). For each adjacent pair of haplotypes x and y on the network, we estimate the nucleotide diversity d_{xy} by

$$d_{xy} = \frac{\sum_{e=1}^{E} S_e r_e d_{xy(e)}}{\sum_{e=1}^{E} S_e r_e}, \qquad (11)$$

where E is the number of enzyme classes examined, S_e is the mean number of restriction sites present in haplotypes x and y for the enzyme class e, r_e is the length of the recognition sequence of the e-th enzyme class (for our enzymes $r_e = 4$, 14/3, 16/3 or 6), and $d_{xy(e)}$ is the fraction of nucleotide substitutions per site between sequences x and y, estimated for the enzyme class e. The computation of (11) is quite simple in our case, because adjacent haplotypes are separated by single mutation changes in most cases, so the numerator involves only one term. Substituting (11) in (3), we have

$$\begin{aligned} {}^*\delta_{jk}^2 &= \{({}^*\mathbf{p}_j - {}^*\mathbf{p}_k)'\mathbf{W}^{1/2}\}\{({}^*\mathbf{p}_j - {}^*\mathbf{p}_k)'\ \mathbf{W}^{1/2}\}' \\ &= ({}^*\mathbf{p}_j - {}^*\mathbf{p}_k)'\ \mathbf{W}\ ({}^*\mathbf{p}_j - {}^*\mathbf{p}_k) \qquad (12) \\ {}^*\delta_{jk}^2 &= \sum_{m=1}^{M} d_{mm}^2({}^*p_{mj} - {}^*p_{mk})^2, \end{aligned}$$

where M is the total number mutational events or links between haplotypes in the minimum spanning network, as defined above. This analysis is analogous to that done for $\mathbf{D}_3$, but here the branches linking each adjacent haplotypes are of length d_{xy}^2 instead of 1. This weighting scheme enables us to incorporate the nucleotide diversity in an Euclidian framework, and to perform an analysis very similar to that developed in LYNCH and CREASE (1990), but with considerably less computation. Using this strategy, we only need to compute nucleotide diversity with (11) between the M adjacent pairs of haplotypes on the network. The nucleotide diversity between a pair of nonadjacent haplotypes is the sum of the stepwise nucleotide diversities along the path joining these two haplotypes. The results of the AMOVA are again reported in Table 3, now labeled as $\mathbf{D}_4$. The figures for both variance component fractions and Φ-statistics are essentially similar to those obtained for $\mathbf{D}_1$ and $\mathbf{D}_3$, with an important fraction of molecular diversity separating regions (21.3%). Careful examination of the input distance matrix (not shown) generated by (12) shows that the amount of nucleotide diversity emerging from single restriction-site changes is very

similar for different enzymes. Branch-lengths between adjacent haplotypes on the network are virtually identical, except for the two cases where more than one restriction-site change is involved.

Genetic structure and DNA site sampling: We evaluated the sensitivity to site sampling by examining the $\mathbf{D}_1$ partition for a random sample of sites, with the number of sites ranging from 5 to 62. We report the percentages of significant values ($\alpha < 0.05$) for the variance components in Figure 6. These three power curves are indistinguishable from those for the Φ-statistics, which are suppressed. As anticipated, the percentage of significant results increases with the number of sampled sites for all statistics; σ_c^2 and Φ_{ST} approach 100% significant outcomes when as few as 40 sites are taken into account. When 62 sites are randomly sampled, σ_b^2 and Φ_{SC} are significant in 99.8% of all replicates, whereas σ_a^2 and Φ_{CT} are significant 94.8% of the time. The component of molecular variance among regions exhibits least power and requires the largest number of restriction sites, suggesting that differences among regions are due to specific sites and mutations. On the whole, however, these high levels of significance show that the inferred genetic structure of our sampled populations is not a sampling artifact and that reliable inference does not require an inordinately large number of sites. We have not carried the analysis to more than 62 sites, because an increase in the number of sampled sites would mean the occurrence of new haplotypes, the distribution of which among populations is unknown from our data.

That conclusion is subject, however, to the assumption that the 62 sites observed are representative of all sites of the mtDNA molecule. Our sites, sampled from an empiric set, are, however, not entirely random. As a practical matter, restriction enzymes that do not generate restriction site variation are usually discarded from the assay battery. The enzymes used here are used routinely in human work precisely because they do exhibit substantial polymorphism. They almost surely do not provide a random representation of the human mtDNA genome, and our collection of sites is certainly biased towards excess polymorphism. The fact that the variation encountered is also geographically structured was not used as a criterion of choice. Indeed, a recent work (STONEKING *et al.* 1990) using additional enzymes revealing even greater polymorphism shows as much geographic structure as we have demonstrated here. It seems probable that a truly random sample of sites (or nucleotides), a larger fraction of which would be monomorphic, would be required to demonstrate the same level of infra-specific structure we have described here. The question of whether our chosen genetic markers are representative set is one more often dealt with by assumption

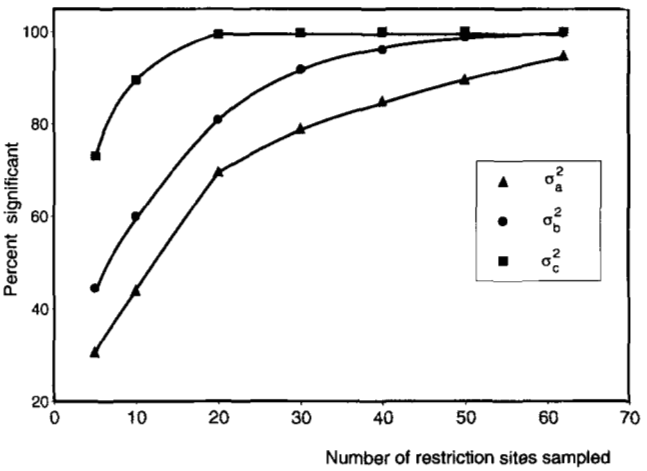

FIGURE 6.—Percentage of significant variance components as a function of haplotype size (in number of restriction-sites). A given number of sites is drawn at random with replacement from the original 62 restriction sites and variance component significance is tested by 500 permutations of the original matrix of squared interindividual distances (see text). This process is repeated 500 times to find the percentage of significant outcomes at the level $\alpha = 0.05$. Φ-statistics curves are almost identical to corresponding variance components and are not reported on the graph.

than proof. Empirically, we see no alternative but testing of the data we have.

LYNCH and CREASE (1990) studied nucleotide sampling analytically, showing that it constituted a major source of variance in estimating diversity at the nucleotide level. Our results are somewhat at odds with theirs. In our case however, the unit studied for its diversity is not the nucleotide but the haplotype, which is itself a collection of sites. The variance of haplotypic diversity due to site sampling appears to be lower than the variance of nucleotide diversity due to the same sampling process. When the number of sites per haplotype is reduced, site sampling becomes increasingly important as shown in Figure 6. For a haplotype with only 5 sites σ_c^2 is significant in 73% of all replicates, σ_b^2 in 44.4%, and σ_a^2 in only 30.8%, showing the importance of site sampling in this case.

DISCUSSION

Human population radiation: Hierarchical analysis of human mtDNA variability shows substantial subdivision among human populations, but with a large fraction of the variation found within populations (>74%). A similar value (69%) has been derived using a G_{ST} approach on another human mtDNA data set (STONEKING *et al.* 1990). Our rather contrived regional groups exhibit a high level of divergence. Populations within regions were shown to be significantly (but minimally) differentiated. Our results suggest that extensive studies within each of the regions are needed to determine whether the much greater divergence observed "among regions" than "among populations/within regions" is an artifact of our arbi-

trary choice of populations, a sampling consequence of isolation-by-distance, or whether there are steep boundary zones of limited genetic exchange between regions. Such zones have come under increasing scrutiny of late (BARBUJANI, ODEN and SOKAL 1989; BARBUJANI and SOKAL 1990, 1991), and a generic answer to the "boundary question" will only be available from a study of more evenly spaced samples.

Regional differentiation is more apparent when the degree of difference between haplotypes is taken into account, in keeping with the observation that molecular distances are larger for pairs of haplotypes drawn from different regions than from the same region (Figure 4). This suggests that a substantial fraction of the mtDNA variability among regions is due to divergent arrays of haplotypes, ultimately attributable to the occurrence of new mutations along the path to regional radiation. It is initially surprising that computing distances along the network only slightly enhances the regional differences in our data set. On further reflection, however, the results make sense. Homoplasies due to recurrent mutations mainly affect low frequency haplotypes that are located at the tips of the network. Both the low frequency of such haplotypes and their network placement will minimally affect the hierarchical partition of variation. The computation of evolutionary distances along a network should yield greater additional resolution for taxonomic assemblages of greater internal radiation, where extinction of intermediates would lead to homoplasic mutations of higher frequency and of more central position.

Nonlinear transformation of restriction-site differences into estimates of nucleotide diversity between haplotypes also does not substantially affect the haplotypic variance partition. We attribute this result to the low divergence between adjacent haplotypes on the network. As most of the links between adjacent haplotypes involve unique restriction-site changes, taking into account the fact that a particular site involves four-, five- or six-base recognition sequences does not matter much here. Thus, the additional assumptions involved in the nonlinear translation, such as uniform substitution rates at different sites and identical substitution probabilities for the four nucleotides, may not be necessary in delineating the internal genetic structure of a single species. However, such nonlinear transformations could be useful if the analysis included individuals from different species with larger interhaplotypic differences.

These conclusions may depend on the choice of the network presented in Figure 3, which was built before the AMOVA analyses were performed. Its basic structure had already been determined in previous publications (JOHNSON *et al.* 1983; EXCOFFIER and LANGANEY 1989). When a high level of homoplasy is present in the data, as it is here, the parsimony criterion does not lead to a unique network, as is also the case for most phylogeny reconstruction algorithms, and a large number of equally parsimonious networks could have been imposed. The question of how to choose among equally parsimonious networks (or trees) is a problem that cannot be settled here. Our contention is merely that given a minimum spanning (parsimonious) network, buttressed by frequency and geographic criteria, an eminently "sensible" network, one can use the methods developed here for a useful partition of the variation. For the example at hand, the additional wrinkle of measuring distance along the network does not provide any additional resolution. Whether we could do better with a different network, and how to choose such a network, we will leave for a later paper.

Our analysis of regional differences shows that the geographic criterion used to define regional groups is quite reasonable as a first approximation. Slightly greater regional divergence was found with an alternative partition of the populations. The European region contains the most internal diversity, whereas the Amerindian region contains the least. The two "alternative" regions Sicily + Maya and Pima + Finland present intermediate "within region" diversities, which slightly lower the total "within region" variability and increase the "among region" variance component. One might consider that σ_a^2 could itself be useful as a criterion for defining supra-population groups. This situation also shows that we need to examine more closely the extent to which each region or each population contributes to the total molecular diversity, as the variance components or Φ-statistics do not bring us much detail of the patterning of the species variability. As has already been done for the multiallelic case (LONG, SMOUSE and WOOD 1987), our analysis framework could be extended to a partitioning of the among-population variability into pairwise population distance components.

Methodological considerations: We have introduced an analytical method for studying the genetic structure of populations that permits use of as much (or as little) of the available information on the molecular nature of DNA haplotypes as is desired. It extends procedures that explicitly use an analysis of variance format (COCKERHAM 1969, 1973; WEIR and COCKERHAM 1984; LONG 1986; LONG, SMOUSE and WOOD 1987) to estimate the degree of intra-specific genetic subdivision. If we can legitimately assume that populations become differentiated by drift alone, then we can expect a linear relation between divergence time and allelic correlation for short periods (REYNOLDS, WEIR and COCKERHAM 1983). In our case, population differences in restriction pattern have clearly arisen from genetic drift of existing variants, from the intro-

duction of new mutations, and from some degree of gene flow, so we will not extrapolate our results as far as a divergence-time interpretation.

The point of the current exercise is neither to estimate unknown population parameters from our variance components nor to define exactly how or at what rate these population differences have developed. Our purpose here is to demonstrate how to delineate the extent of genetic differentiation within and among populations. The approach is general enough to deal with any organism and to study any type of structure (hierarchical or otherwise) that one might wish to consider. The underlying (distance matrix) structure of the analysis permits flexible exploration of a given data set. Several different distance matrices, one for each particular set of assumptions, may be taken as alternate inputs and their influence on the outcome evaluated. The relation to *F*-statistics is straightforward, though subject to the usual limitations. More important is the realization that the whole array of least-squares methods (analysis of variance, analysis of covariance, regression, correlation, principal coordinates analysis, factor analysis, etc.) is accessible from this same distance matrix. We have tapped only a small portion of the available repertoire here.

Significance testing with permutation procedures is both easy and essentially assumption free; in particular, we are freed from the testing limitations of normal theory, so useful in analysis of variance but so inappropriate here. We can address several questions with the same data set. We might even wish to test the difference between outcomes formally, based on different squared-distance matrices. As the computation of the variance components involves only manipulation of the original input distance metrics, the outcome will only be as different as the inputs. Squared-distance matrices may be compared using a normalized Mantel test (SMOUSE, LONG and SOKAL 1986).

If one wishes to translate restriction site differences into estimates of the fraction of nucleotide differences between pairs of haplotypes (π_{jk}), several procedures are available (ENGELS 1981; EWENS, SPIELMAN and HARRIS 1981; NEI and TAJIMA 1981, 1983; KAPLAN 1983; NEI and MILLER 1990), any one of which can be used to modify the interhaplotypic squared distances in our technique. Additional translation may permit linearization of these estimates with divergence time. Such transformations have the additional advantage of being independent of the number of restriction sites surveyed. We have seen, however, that this process does not fundamentally alter either our estimates of the variance components. Extension of this methodology to DNA sequence data is straightforward and can be achieved through a redefinition of the interchromosomal distance metric. As several methods are already available for this purpose in the literature (SWOFFORD and OLSEN 1990), one is free to choose. We will content ourselves here with the observation that the use of a Euclidean metric has some natural advantages, not the least of which is that a matrix of such distances can be used for other purposes than phylogenetic analysis. The considerable variety of data types made available by molecular biology needs a statistical analysis framework that is coherent but also sufficiently flexible to accommodate the different types of questions inherent in each particular situation. The AMOVA treatment presented here is intended to serve as the beginning of just such a framework.

The authors thank OSCAR GAGGIOTTI and ANDRÉ LANGANEY for their comments on the manuscript, as well as MICHAEL LYNCH and another (anonymous) reviewer for their suggestions. L.E. was funded by FNRS Switzerland 32-28784.90 and 32-27845.89, and INSERM France 900 814, P.E.S. by NJAES/USDA-32102, JMQ by the Roosevelt Fund, American Museum of Natural History and by the Leathem-Steinetz-Stauber Fund, Rutgers University. An analysis of molecular variance program (AMOVA), including the permutational testing procedures, is available on request from L.E.

LITERATURE CITED

ANDERSON, S., A. T. BANKIER, B. G. BARREL, M. H. L. DE BRUIJN, A. R. COULSON, J. DROUIN, I. C. EPERON, D. P. NIERLICH, B. A. ROE, F. SANGER, P. H. SCHREIER, A. J. H. SMITH, R. STADEN and I. G. YOUNG, 1981 Sequence and organization of the human mitochondrial genome. Nature **290:** 457–465.

BARBUJANI, G., N. L. ODEN and R. R. SOKAL, 1989 Detecting areas of abrupt change in maps of biological variables. Syst. Zool. **38:** 376–389.

BARBUJANI, G., and R. R. SOKAL, 1990 The zones of sharp genetic change in Europe are also language boundaries. Proc. Natl. Acad. Sci. USA **87:** 1816–1819.

BARBUJANI, G., and R. R. SOKAL, 1991 Genetic population structure of Italy. II. Physical and cultural barriers to gene flow. Am J. Hum. Genet. **48:** 398–411.

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.

BONNÉ-TAMIR, B., M. J. JOHNSON, A. NATALI, D. C. WALLACE and L. L. CAVALLI-SFORZA, 1986 Human mitochondrial DNA types in two Israeli populations—a comparative study at the DNA level. Am. J. Hum. Genet. **38:** 341–351.

BREGA, A., R. GARDELLA, O. SEMINO, G. MORPURGO, G. B. ASTALDI RICOTTI, D. C. WALLACE and A. S. SANTACHIARA-BERENECETTI, 1986 Genetic studies on the Tharu population of Nepal: restriction endonuclease polymorphisms of mitochondrial DNA. Am. J. Hum. Genet. **39:** 502–512.

BROWN, W. M., M. GEORGE, JR. and A. C. WILSON, 1979 Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA **76:** 1967–1971.

BROWN, W. M., E. M. PRAGER, A. WANG and A. C. WILSON, 1982 Mitochondrial DNA sequences of primates: tempo and mode of evolution. J. Mol. Evol. **18:** 225–239.

CANN, R. L., M. STONEKING and A. C. WILSON, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

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

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

EFRON, B., 1982 The Jacknife, the Bootstrap and Other Resampling Plans. Regional Conference Series in Applied Mathematics, Vol 38. Society for Industrial and Applied Mathematics, Philadelphia.

ENGELS, W. R., 1981 Estimating genetic divergence and genetic variation with restriction endonucleases. Proc. Natl. Acad. Sci. USA **78:** 6329–6333.

EWENS, W. J., R. S. SPIELMAN and H. HARRIS, 1981 Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc. Natl. Acad. Sci. USA **78:** 3748–3750.

EXCOFFIER, L., 1990 Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. **30:** 125–139.

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

FARRIS, J. S., 1970 Methods for computing Wagner trees. Syst. Zool. **19:** 83–92.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:** 521–565.

GILES, R. E., H. BLANC, H. M. CANN and D. C. WALLACE, 1980 Maternal inheritance of human mitochondrial DNA. Proc. Natl. Acad. Sci. USA **77:** 6715–6719.

GYLLENSTEN, U., D. WHARTON, A. JOSEFSSON and A. C. WILSON, 1991 Paternal inheritance of mitochondrial DNA in mice. Nature **352:** 255–257.

JOHNSON, M. J., D. C. WALLACE, S. D. FERRIS, M. C. RATTAZZI and L. L. CAVALLI-SFORZA, 1983 Radiation of human mitochondrial DNA types analyzed by restriction endonuclease cleavage patterns. J. Mol. Evol. **19:** 255–271.

KAPLAN, N., 1983 Statistical analysis of restriction enzyme map data and nucleotide sequence data, pp. 75–106 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

LI, C. C., 1976 *Population Genetics.* Boxwood, Pacific Grove, Calif.

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.

LONG, J. C., P. E. SMOUSE and J. W. WOOD, 1987 The allelic correlation structure of Gainj- and Kalam-speaking people. II. The genetic distance between population subdivisions. Genetics **117:** 273–283.

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

MANTEL, N., 1967 The detection of disease clustering and a generalized regression approach. Cancer Res. **27:** 209–220.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70:** 3321–3323.

NEI, M., 1977 *F*-statistics and the analysis of gene diversity in subdivided populations. Ann. Hum. Genet. **41:** 225–233.

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.

NEI, M., and F. TAJIMA, 1983 Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105:** 207–217.

PRIM, R. C., 1957 Shortest connection networks and some generalizations. Bell Syst. Tech. J. **36:** 1389–1401.

REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: Basis for a short term genetic distance. Genetics **105:** 767–779.

RICHARDSON, R. R., and P. E. SMOUSE, 1976 Patterns of electrophoretic mobility. I. Interspecific comparisons in the *Drosophila mulleri* complex. Biochem. Genet. **14:** 447–466.

RICHARDSON, R. R., P. E. SMOUSE and M. E. RICHARDSON, 1977 Patterns of molecular variation. II. Associations of electrophoretic mobility and larval substrate within species of the *Drosophila mulleri* complex. Genetics **85:** 141–154.

ROHLF, F. J., 1990 NTSYS. Numerical Taxonomy and Multivariate Analysis System. Ver. 1.60. Exeter Publ. Ltd., Setauket, N.Y.

SCHURR, T. G., S. W. BALLINGER, Y.-Y. GAN, J. A. HODGE, D. A. MERRIWEATHER, D. N. LAWRENCE, W. C. KNOWLER, K. M. WEISS and D. C. WALLACE, 1990 Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am. J. Hum. Genet. **46:** 613–623.

SCOZZARI, R., A. TORRONI, O. SEMINO, G. SIRUGO, A. BREGA and A. S. SANTACHIARA-BERENECETTI, 1988 Genetic studies on the Senegal population. I. Mitochondrial DNA polymorphisms. Am. J. Hum. Genet. **43:** 534–544.

SEMINO, O., A. TORRONI, R. SCOZZARI, A. BREGA, G. DE BENEDICTIS and A. S. SANTACHIARA BENERECETTI, 1989 Mitochondrial DNA polymorphisms in Italy. III. Population data from Sicily: a possible quantitation of African ancestry. Ann. Hum. Biol. **53:** 193–202.

SLATKIN, M., 1987 The average number of sites separating DNA sequences drawn from a subdivided population. Theor. Popul. Biol. **32:** 42–49.

SMOUSE, P. E., J. C. LONG and R. R. SOKAL, 1986 Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. **35:** 627–632.

STONEKING, M., L. B. JORDE, K. BHATIA and A. C. WILSON, 1990 Geographic variation in human mitochondrial DNA from Papua New Guinea. Genetics **124:** 717–733.

SWOFFORD, D. L., and G. J. OLSEN, 1990 Phylogeny reconstruction, pp. 411–501 in *Molecular Systematics*, edited by D. M. HILLIS and C. MORITZ. Sinauer Associates, New York.

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

VILKKI, J., M.-L. SAVONTAUS and E. V. NIKOSKELAINEN, 1988 Human mitochondrial types in Finland. Hum. Genet. **80:** 317–321.

WALLACE, D. C., K. GARRISON and W. C. KNOWLER, 1985 Dramatic founder effect in Amerindian mitochondrial DNAs. Am. J. Phys. Anthrop. **68:** 149–155.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

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

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

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugen. **1:** 323–334.

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 Associates, Sunderland, Mass.

Communicating editor: E. THOMPSON