Human Genetic Diversity and the Nonexistence of Biological Races

Author(s): Jeffrey C. Long and Rick A. Kittles
Source: Human Biology, 81(5/6):777-798. 2009.
Published By: Wayne State University Press
DOI: http://dx.doi.org/10.3378/027.081.0621
URL: http://www.bioone.org/doi/full/10.3378/027.081.0621

Human Genetic Diversity and the Nonexistence of Biological Races

JEFFREY C. LONG[1] AND RICK A. KITTLES[2]

Abstract Sewall Wright's population structure statistic, F_{ST}, measured among samples of world populations is often 15% or less. This would indicate that 85% of genetic variation occurs within groups while only 15% can be attributed to allele frequency differences among groups. In this paper, we show that this low value reflects strong biases that result from violating hidden assumptions that define F_{ST}. These limitations on F_{ST} are demonstrated algebraically and in the context of analyzing dinucleotide repeat allele frequencies for a set of eight loci genotyped in eight human groups and in chimpanzees. In our analyses, estimates of F_{ST} fail to identify important variation. For example, when the analysis includes only humans, $F_{ST} = 0.119$, but adding the chimpanzees increases it only a little, $F_{ST} = 0.183$. By relaxing the underlying statistical assumptions, the results for chimpanzees become consistent with common knowledge, and we see a richer pattern of human genetic diversity. Some human groups are far more diverged than would be implied by standard computations of F_{ST}, while other groups are much less diverged. We discuss the relevance of these findings to the application of biological race concepts to humans. Four different race concepts are considered: typological, population, taxonomic, and lineage. Surprisingly, a great deal of genetic variation within groups is consistent with each of these concepts. However, none of the race concepts is compatible with the patterns of variation revealed by our analyses.

Human racial classifications have vexed anthropologists and geneticists for the last half-century, since biologists began to question the validity of taxonomy below the species level (Wilson and Brown 1953). The following flaws were repeatedly pointed out (Lieberman and Jackson 1995; Keita and Kittles 1997; Templeton 1998). (1) There are no agreed-upon criteria for when to assign formal names to groups that might more appropriately be considered aggregates of local populations. (2) Race classifications fail for phenotypically intermediate populations. (3) They fail for individuals who trace their ancestry to two or more named races. (4) They are defied by sets of characters that show independent geographic trends. (5) It has been difficult to relate many human populations as distinct evolutionary lineages.

[1]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109-0618.
[2]National Human Genome Center, Howard University, Washington, D.C. 20060.

Human Biology, August 2003, v. 75, no. 4, pp. 449–471.

KEY WORDS: F_{ST}, GENE IDENTITY, RACE.

Another objection to racial classifications has been advanced for the past 30 years. This view holds that race classifications are unjustifiable when the variation within groups exceeds the variation between groups. Lewontin (1972) showed that 85.4% of the human species diversity could be attributed to allelic variation within populations, 8.3% to variation between populations within races, and only 6.3% to interracial variation. Many other studies have reiterated this finding (Barbujani et al. 1997). Most have used the statistic F_{ST} (Wright 1951, 1965, 1969, 1978) or very closely related statistics (Cockerham 1969; Nei 1987; Excoffier et al. 1992). F_{ST} measures the extent of subpopulation differentiation as the decrease in heterozygosity relative to that which would be expected if mating were at random throughout the entire population. F_{ST} can be interpreted equivalently as measure of gain of homozygosity. Wright (1969, 1978) envisioned this increase in homozygosity as the progress towards fixation. F_{ST} measured among samples of world populations is often 15% or less when computed as an average over many alleles or loci. There are alleles and loci that individually show higher and lower F_{ST} values, but these exceptions are greatly overwhelmed by the collective weight of many alleles analyzed simultaneously (Bowcock et al. 1991, 1994).

This paper questions the validity of this ubiquitous finding. We show that estimates of F_{ST} will fail dramatically to identify important differentiation among groups, because the outcome of analyses is strongly biased by violating two hidden assumptions: that expected gene identity (Nei 1987) is the same in every population and that divergence between all pairs of populations is equal and independent. We also show that the value F_{ST} depends on allele frequencies, and so it is not free to vary from 0.0 to 1.0. The precise nature of these limitations on F_{ST} is demonstrated algebraically and in the context of data analysis. The data consist of a large set of dinucleotide repeat allele frequencies at eight loci estimated for eight human populations and a sample of chimpanzees (Deka et al. 1995). When the analysis includes only human populations, $F_{ST} = 0.119$. When the chimpanzees are added, $F_{ST} = 0.183$. Although adding the chimpanzees increases the value of F_{ST}, neither estimate differs substantially from 0.15.

To demonstrate the impact of unequal levels of gene identity and nonindependence of divergence of all pairs of subpopulations, we estimate F_{ST} using a sequential model fitting approach (Urbanek et al. 1996). We also use a formal measure of lack of fit to evaluate the basic F_{ST} variance decomposition (Cavalli-Sforza and Piazza 1975) and to assess the impact of relaxing the violated assumptions. We conclude that the ubiquitous finding $0.10 \leq F_{ST} \leq 0.15$ is due primarily to statistical artifact. There is little meaning to simple partitions of human genetic variation on a worldwide scale, and the broad acceptance of F_{ST} as a valid measure has prevented a deeper understanding of human variation.

Microsatellite Repeat Data Set

The data consist of published dinucleotide repeat allele frequencies at eight loci for eight human populations sampled from five different continents, and a

Table 1. Sample Descriptions

Population	*Geographic Origin*	*Individuals*
Sokoto	Africa (Nigeria)	117
German	Europe (Germany)	95
CEPH	Europe (Utah)	76
Samoa	Pacific (Polynesia)	110
Kalam	Pacific (New Guinea)	149
Kachari	Asia (India)	50
Dogrib	North America	66
Pehuenche	South America	103
Chimp	–	83

sample of chimpanzees (Deka et al. 1995). The sample sizes are relatively large, $n_k \geq 50$, for all groups, including chimpanzees (Table 1). Each locus is highly polymorphic; these eight loci segregate a total of 136 alleles. The human samples mostly represent local populations, although the European-derived groups are more cosmopolitan. The German participants were recruited in northern Germany and the CEPH sample is composed mostly of individuals with northern European heritage living in Utah. We will generally refer to populations by their local names in order to maintain precision; however, their geographic origin will usually be indicated as well. It is theoretically possible that analyzing the chimpanzees does not dramatically increase F_{ST} because high mutation rates have produced similar allele frequency distributions in the two species. This possibility is easily discounted by visual inspections of the frequency distributions at all eight loci (Figure 1). Clearly, there is overlap in the actual alleles observed at all eight loci; however, each locus presents large frequency differences between species.

Theoretical Background

The parameter F_{ST} has a precise and unambiguous definition in the context of a population model where each subpopulation is considered to be an independent replicate of the evolutionary process (Cockerham 1969; Weir and Cockerham 1984; Weir 1996; Wright 1951, 1969). In the absence of such a model, F_{ST} can be defined only for the exact set of subpopulations being analyzed, and there cannot be valid generalizations from the analysis of a specific data set to the species as a whole (Cockerham and Weir 1986). We will review this model as Weir and Cockerham (1984) presented it. By doing so, we will be able to demonstrate that F_{ST} is indeed dependent on allele frequencies and that estimates of F_{ST} are biased by violating the basic assumptions of equal gene identity within subpopulations and independence of divergence among subpopulations.

Basic Model and Definitions. Consider the alleles at a single selectively neutral locus in a population that is composed of a very large number s of independently

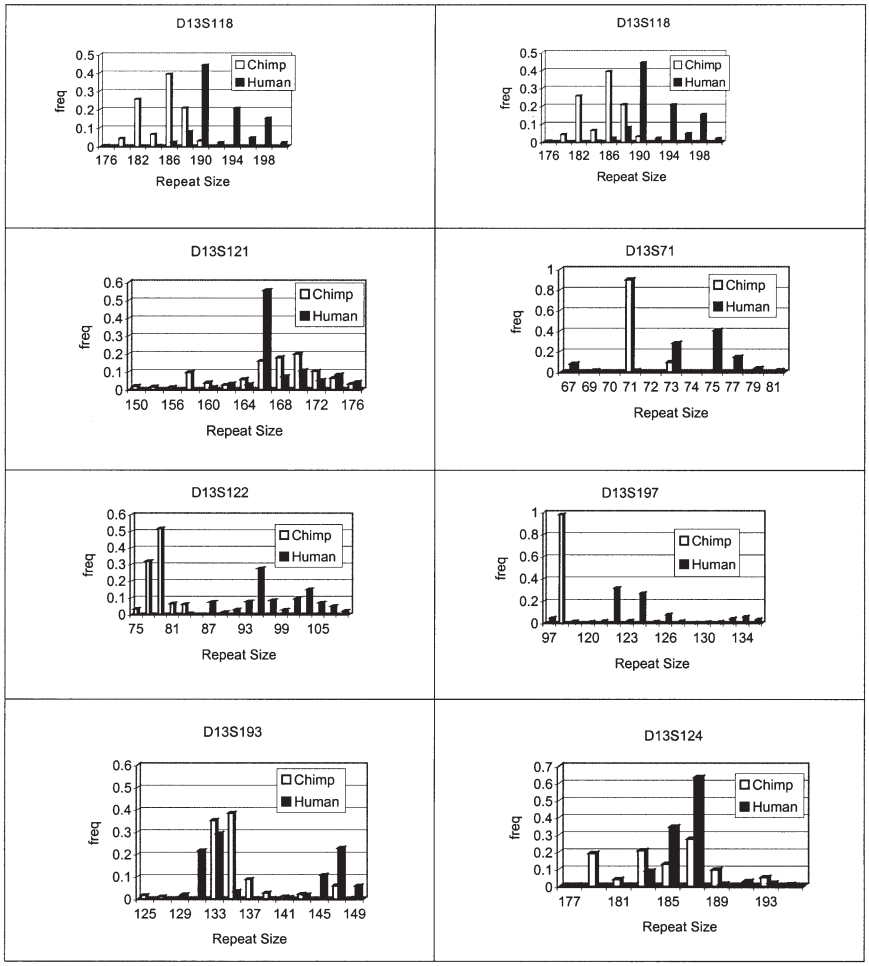


Figure 1. Dinucleotide repeat allele size distributions for eight loci in humans and chimpanzees. The frequencies for humans are unweighted averages from eight groups representing diverse worldwide localities (see Deka et al. 1995). Notice that there are clear distinctions between humans and chimpanzees at all loci despite the presence of overlap between the distributions.

evolving subpopulations, i.e., $s \rightarrow \infty$. For simplicity, let all subpopulations be panmictic and of equal effective size. Let p_{ki} be the frequency of the *i*th allele in the *k*th subpopulation and $p_i = (1/s) \sum_{k=1}^{s} p_{ki}$ be the frequency of the *i*th allele in the total population. The differences in allele frequencies among subpopulations are assumed to be the result of genetic sampling, i.e., the process whereby the gametes forming each generation are a random sample from the gametes that formed the previous generation. Random populations models embody two important properties

(Weir 1996; Weir and Hill 2002). First, they envision a single base population that serves as a reference for evaluating probabilities and expectations about the genetic sampling process. Since allele frequencies within the subpopulations vary at random, each subpopulation is equally representative of the total population and no single subpopulation is of particular importance. The expected allele frequency is the same for all subdivisions, $E[p_{ki}] = p_i$. Second, the process of genetic sampling causes some alleles within a subpopulation to be related because they are recent copies of an ancestral allele that existed as a single copy.

Consider the following probabilities J that a random pair of alleles will be identical in state: For alleles drawn from the same subpopulation,

$$J_k = \sum_i p_{ki}^2; \tag{1}$$

for alleles drawn from different subpopulations,

$$J_{kl} = \sum_i p_{ki} p_{li}; \tag{2}$$

and for alleles drawn from the total population,

$$J_T = \sum_i p_i^2. \tag{3}$$

Nei (1987) has referred to these probabilities as gene identities in the kth subpopulation, between the kth and lth subpopulations, and in the total population, respectively. Gene identity has a clear biological interpretation; it is the expected proportion of homozygous genotypes assuming random mating. We may write the gene identity in the total population in terms of allele frequencies and the gene identities within and between subpopulations

$$\begin{aligned} J_T &= \left(\sum_k \sum_i p_{ki}^2 + \sum_{k \neq l} \sum_i p_{ki} p_{li} \right) \Big/ s^2 \\ &= \left(\sum_k J_k + \sum_{k \neq l} J_{kl} \right) \Big/ s^2 \\ &= \frac{J_S}{s} + \frac{\sum_{k \neq l} J_{kl}}{s^2}. \end{aligned} \tag{4}$$

Here, J_S is the average gene identity within subpopulations, $J_S = \Sigma_k \, J_k / s$. Additional algebra shows that

$$J_T = J_S - D_{ST}, \tag{5}$$

where

$$D_{ST} = \sum_{k \neq l} \frac{\left[(J_k + J_l)/2 - J_{kl} \right]}{s^2} = \frac{1}{2} \sum_{k \neq l} \sum_i \frac{(p_{ki} - p_{li})^2}{s^2}.$$

Since $D_{ST} \geq 0$, $J_T \leq J_S$, with equality only when $p_{ki} = p_{li}$, for all alleles i and all pairs of subpopulations k and l. The quantities above were defined for a single locus in the interest of simplicity, but the approach can be extended to multiple loci by averaging over independent loci.

The parameter F_{ST} is defined as the relative increase in gene identity that is attributable to population subdivision:

$$F_{ST} = \frac{D_{ST}}{1 - J_T} = \frac{J_S - J_T}{1 - J_T}. \tag{6}$$

F_{ST} can be interpreted as a composite measure of allelic differentiation among subpopulations, relative to the limiting amount under complete fixation (Wright 1978).

Following Cockerham (1969, 1973) and Weir and Cockerham (1984), the expected gene identity over evolutionary replications is the same within each subdivision k:

$$E[J_k] = J_S = J_T + (1 - J_T) F_{ST}. \tag{7}$$

Similarly, the expected gene identity between pairs of subdivisions is the same for all pairs k and l:

$$E[J_{kl}] = J_T. \tag{8}$$

These expectations provide the basis for estimating F_{ST} from genetic marker data (Weir and Cockerham 1984). An important nuance of Equation (8) is that the gene identity between every pair k and l of subpopulations is informative about the parameter J_T.

Maximum F_{ST}. It is sometimes claimed that F_{ST} is a measure of population differentiation that is independent of allele frequencies (e.g., Harpending and Rogers 2000). This claim is based on the fact that in the definition of F_{ST} the absolute divergence among populations D_{ST} is standardized by the total gene diversity $1 - J_T$. However, the independence of F_{ST} and allele frequency is a misperception. The maximum value that F_{ST} can take is easily derived to be

$$J_S = \frac{1}{s} \sum_k \sum_i p_{ki}^2,$$

which is clearly a function of allele frequencies. Consider the state of maximum differentiation among subpopulations where no subpopulation shares an allele with any other subpopulation, i.e., $J_{kl} = 0$ for all pairs of subpopulations, k and l.

Substituting these values into Equation (4) yields $J_T = J_S/s$. Then, by substituting this value for J_T into Equation (6),

$$\max(F_{ST}) = \frac{J_S - (J_S/s)}{1-(J_S/s)}. \tag{9}$$

With many subpopulations, $J_S/s \to 0$ and $\max(F_{ST}) \to J_S$. Thus, $F_{ST} = 1.0$ only if every subpopulation is fixed, and F_{ST} can never be very high for genetic loci with high heterozygosity.

Critical Assumption No. 1. The first critical assumption is that gene identities and, hence, effective population sizes are equal for all subpopulations. If there is variation in subpopulation sizes, then the expected relatedness among alleles will vary across subpopulations. F_{ST} as defined by Equation (6) then relates to some kind of average of population specific F_{ST}'s (Weir 1996) and Equation (7) no longer holds. There are two ways around this problem. First, as Nei (1987) shows, Equation (3) for F_{ST} is valid if p_i is defined as a weighted average, $p_i = \Sigma_{k=1}^{s} w_k p_{ki}$ with $w_k = N_{e_k} \big/ \Sigma_{k=1}^{s} N_{e_k}$. This is not a very useful solution to the problem because effective sizes are abstract quantities that are not easily measured for natural populations including humans. Applying it would also require that the distribution of subpopulation effective sizes in the total sample is an unbiased representation of subpopulation effective sizes for the species as a whole. In fact, data analyses have almost never used averages weighted by effective population sizes. Second, more recently Urbanek et al. (1996) and Weir and Hill (2002) have shown that it is possible to estimate subpopulation specific values of F_{ST}. This is the strategy that will be pursued in the data analysis of the next section.

Critical Assumption No. 2. The second critical assumption is that every subpopulation is evolving independently. The consequences of violating this assumption are best illustrated by extending the model to account for an additional level of population structure. Following Wright (1951, 1965, 1969, 1978), Nei (1987), and others, let the subpopulations be nested within different geographic regions. Suppose further that the number of different regions, r, is large, and that each region contains a large number, s, of subpopulations. The frequency of the ith allele in the kth subpopulation of the gth region is now denoted p_{gki}, each region has an average allele frequency

$$p_{g\cdot i} = \frac{1}{s}\sum_{k=1}^{s} p_{gki},$$

and the frequency for the total population is the average over regions,

$$p_i = \frac{1}{r}\sum_{g=1}^{r} p_{g\cdot i}.$$

As long as all other aspects of the random populations model are maintained, the expected allele frequency for all subpopulations and all regions is the same and equal to the total population frequency, $E[p_{gki}] = E[p_{g.i}] = p_i$.

We can now define gene identity for pairs of genes from the same subpopulation

$$J_S = \frac{1}{rs} \sum_{g=1}^{r} \sum_{k=1}^{s} \sum_{i} p_{gki}^2,$$

from the same region

$$J_R = \frac{1}{r} \sum_{g=1}^{r} \sum_{i} p_{g\cdot i}^2,$$

and from the total population $J_T = \Sigma_i \, p_i^2$. This formulation leads to two new F statistics,

$$F_{SR} = \frac{J_S - J_R}{1 - J_R} \tag{10a}$$

and

$$F_{RT} = \frac{J_R - J_T}{1 - J_T}, \tag{10b}$$

that are related to F_{ST} as defined in Equation (6) by the well-known formula

$$(1 - F_{ST}) = (1 - F_{SR})(1 - F_{RT}). \tag{11}$$

According to this version of the random populations model (Cockerham and Weir 1986), the expectations for gene identity are

$$E[J_{gk}] = J_S = J_T + (1 - J_T) F_{ST} \tag{12}$$

for pairs of alleles from the same subpopulation k,

$$E[J_{gk,gl}] = J_T + (1 - J_T) F_{RT} \tag{13}$$

for pairs of alleles randomly drawn from different subpopulations k and l in the same region g, and

$$E[J_{gk,hl}] = J_T \tag{14}$$

for pairs of alleles randomly drawn from subpopulations k and l in different regions g and h. Notice that there is an important change from the situation where all subpopulations are evolving independently. Now the expected gene identity

between two subpopulations depends on whether or not they are in the same geographic region. An unbiased estimate of J_T is obtained only by averaging gene identities between pairs of subpopulations in different regions. If subpopulations within the same region are included in the average, then the estimate of J_T would be biased upwards, and concomitantly, estimates of F_{ST} are biased downwards.

Natural populations, of course, have a more complex structure than the simple models of subdivision described above. We would expect that the usual consequence of failing to recognize levels of subdivision is to bias estimates of J_T upwards and estimates of F_{ST} downwards. We also note that there will not be a single value F_{RT} as defined by Equation (10b) if effective population sizes differ between regions. The consequences of variable effective population size and evolutionary nonindependence compound each other. Evolutionary independence cannot be achieved in a hierarchically structured population unless every level is completely balanced. For example, each subpopulation must have the same number of individuals and each continent must have the same number of subpopulations.

Methods

Gene identity coefficients representing various levels of population hierarchy were estimated using the method of Urbanek et al. (1996). Estimates of F_{ST} were obtained from the estimated gene identity coefficients by taking the appropriate ratios, e.g., Equations (6) and (10).

Models, Estimation Procedures, and Hypothesis Tests. The method of Urbanek et al. requires allele frequencies sampled from $t \leq s$ subpopulations, from which a $t \times t$ matrix $\boldsymbol{J}$ is constructed with estimates of gene identities within populations $\hat{J}_k$ in the diagonal positions and estimates of gene identities between populations $\hat{J}_{kl}$ in the off-diagonal positions. The symbol $\hat{}$ is used here and elsewhere to designate an estimate rather than a parameter. Estimates of gene diversity at different levels of population hierarchy (e.g., J_S, J_R, J_T) are obtained by fitting hierarchical models to the matrix $\boldsymbol{J}$. Each hierarchical arrangement predicts a different theoretical form for the matrix $\boldsymbol{J}$ that will be denoted by a matrix Σ. As shown by Cavalli-Sforza and Piazza (1975), the theoretical matrix Σ specified by a particular hierarchy can be expressed as a linear combination

$$\Sigma = \sum_{g=0}^{m} \sigma_g \mathbf{G}_g, \tag{15}$$

where $\mathbf{G}_0, \mathbf{G}_1, \ldots, \mathbf{G}_m$ are fixed symmetric linear independent $t \times t$ matrices composed of zeros and ones, and the coefficients (σ_g) are theoretical values for gene identities implied by the internal and external nodes of the hierarchy. The number of parameters that are estimated depends on the levels of population subdivision

in the hierarchy and assumptions about whether or not gene identities J_k are equal. In some simple cases maximum likelihood estimators based on this model take explicit algebraic forms (see Urbanek et al. 1996), but in most situations, estimation requires numerical analysis. Anderson (1973) developed a system of expectation-maximization (EM) equations that can be applied for maximum likelihood estimation in more complex situations:

$$\sum_{f=0}^{m} tr\left(\hat{\Sigma}^{-1}\mathbf{G}_g\hat{\Sigma}^{-1}\mathbf{G}_f\right)\hat{\sigma}_f = tr\left(\hat{\Sigma}^{-1}\mathbf{G}_g\hat{\Sigma}^{-1}\boldsymbol{J}\right), \qquad g = 0,1,...,m. \tag{16}$$

The advantage of the maximum likelihood approach to estimation is that it opens the possibility for testing hypotheses using likelihood ratio statistics. As a starting point, we assume that the parameter estimates are multivariate normally distributed (e.g., Weir and Hill 2002). Then tests can be developed for two kinds of null hypotheses. The first sort of hypothesis is that the observed matrix $\boldsymbol{J}$ deviates from a proposed theoretical matrix Σ no more than would be expected from random genetic drift alone. This uses the standard likelihood ratio statistic

$$\Lambda = \nu \cdot [\ln(\det\hat{\Sigma} - \ln(\det\boldsymbol{J}) + tr(\boldsymbol{J}\hat{\Sigma}^{-1}) - t], \tag{17}$$

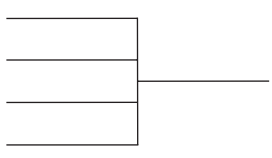
where ν is the number of independent observations in the analysis (Morrison 1976). For the present purposes, an observation is an allele at a locus and the number of independent observations is the (total alleles − loci − 1). In the limit of large sample size (in this case independent alleles), Λ is distributed as a chi-square random variable with degrees of freedom equal to $t(t+1)/2$ minus the number of parameters required to fit the specified tree (Cavalli-Sforza and Piazza 1975). This form of the likelihood ratio statistic differs slightly from the one provided by Cavalli-Sforza and Piazza (1975) by correcting a printed mistake and by adding the usually negligible term $[tr(\boldsymbol{J}\hat{\Sigma}^{-1}) - t]$ for completeness. The second sort of hypothesis involves a constraint on a specific component, say $\sigma_g = \sigma_{g-1}$. Comparing the log likelihood maximized under no constraint with the log likelihood maximized under the constraint tests this hypothesis. The likelihood ratio test statistic for this hypothesis is computed as

$$K = \Lambda_1 - \Lambda_0, \tag{18}$$

where Λ_1 and Λ_0 are computed by Equation (17) under the unconstrained and constrained hypotheses, respectively.

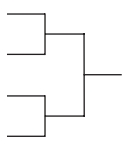
Although examples of how to use this estimation and testing procedure have been presented elsewhere (e.g., Cavalli-Sforza and Piazza 1975, Urbanek et al. 1996), a brief review is helpful here. Assume that allele frequencies have been sampled from $t = 4$ subpopulations, and a matrix $\boldsymbol{J}$ has been computed. Under the supposition that each subpopulation has evolved independently and that they have equal effective sizes, Equation (15) takes the form shown below. A tree diagram representing this model is shown on the right:

$$\Sigma = J_S \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + J_T \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$



By contrast, suppose that the four subpopulations were sampled such that two subpopulations were drawn from each of two regions. This requires expanding Equation (15) to include a component that corresponds to the regional structure, as shown below. Once again, a tree diagram representing this model is shown on the right.

$$\Sigma = J_S \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + J_R \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} + J_T \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$



Three hypotheses are now of interest. First, can tree No. 1 be rejected by its lack of fit to the data (i.e., the full matrix $\boldsymbol{J}$)? This hypothesis is tested by computing the test statistic in Equation (17) and comparing it to a chi-square distribution with $4 \times 3/2 - 2 = 4$ degrees of freedom. Second, can tree No. 2 be rejected by its lack of fit to the data? This hypothesis is also tested by computing the test statistic in Equation (17) and comparing it to a chi-square distribution; however, the degrees of freedom are now $4 \times 3/2 - 3 = 3$ because the model estimates one more parameter (J_R). Third, does the second model fit the observations ($\boldsymbol{J}$) significantly better than the first model? For this, the test statistic defined by Equation (18) is computed and compared to a chi-square distribution with $3 - 2 = 1$ degree of freedom. The null hypothesis implicit in this test is H_0: $J_R = J_T$.

Caveats. The test statistics Λ and K may reject their respective null hypothesis too readily because of violation of the multivariate normality assumption. Nevertheless, they remain useful qualitative measures. Notice that $\Lambda = 0$ when the theoretical matrix Σ fits the observed matrix $\boldsymbol{J}$ perfectly. Similarly, $K = 0$ when the estimate of Σ under the unconstrained hypothesis fits the observed matrix $\boldsymbol{J}$ no better than the estimate of Σ under constrained hypothesis.

It is worth noting that the formulas above have not dealt with all of the complicating factors in analysis. For example, we have reduced the complexity of equations by ignoring the small bias factor related to sample size. Fortunately, the sample size effect is modest for the data to be analyzed here because a relatively large number of individuals were genotyped for each locus and population (Deka et al. 1995). Nei and colleagues (Nei 1987) have developed useful corrections for sample size bias that can be applied when necessary. However, there are less appreciated biases, such as those created by data gaps caused by routine polymerase

chain reaction failures or limited reagents and DNA samples. There are not simple formulas for removing these biases but their first order effects can be removed by jackknifing (Quenouille 1956; Urbanek et al. 1996).

Sequential Model Fitting. Our strategy involves sequentially fitting four models; each model relaxes an important constraint from the previous model. Model I proposes a single level of nesting: genes within subpopulations. Model II maintains the structure of model I but relaxes the assumption of evolutionary independence by fitting two levels of nesting: genes within subpopulations and subpopulations within regions. Model III creates a third level of nesting that maintains all of the aspects of the previous model, but it nests the non-African subpopulations as a group distinct from the African subpopulation. Model IV maintains the structure of model III, but gene identity is fit individually for each subpopulation. The residual lack of fit Λ of each successive model is compared to that of its predecessor. The important components of population structure are therefore identified by a large reduction in Λ accompanied by a modest decrease in degrees of freedom.

Results

Figure 2 portrays the hierarchies defined by the four models described above. The scale on each diagram is gene identity. The total length, from the origin to the terminal nodes, represents the gene identity within subpopulations J_S. The length from the origin to the first division represents the gene identity in the total population J_T. The difference between these quantities, $J_S - J_T$, is D_{ST}, the gene diversity between subpopulations. For purposes of comparison, this difference is indicated by the shaded bar beneath the scale.

The simple partition underlying model I (Figure 2A) assumes that the gene identity between populations, J_{kl}, is the same between all pairs of populations ($E[J_{kl}] = J_S - J_T = 0.083$) and that the gene identity within populations J_k is the same for all populations ($E[J_k] = J_S = 0.383$). However, the actual values of gene identity provided in Table 2 show large departures from these expectations, and it is unsurprising that the residual lack of fit $\Lambda = 402.9$ is very large relative to its degrees of freedom, $df = 34$. While $\hat{F}_{ST} = 0.119$ is low, the model fits the data very poorly.

Figure 2B portrays the two-level hierarchy in model II. Again, the total length of the tree portrays J_S, and the length from the origin to the first bifurcation represents J_T. However, in the geographic regions represented by more than one subpopulation, we estimate a component of gene identity specific to subpopulations in the same region. The total gene identity between populations is again indicated by the shaded bar. It is nearly unchanged by this procedure, and the estimate $\hat{F}_{ST}$ is hardly changed. However, this relaxes the assumption of complete evolutionary independence between subpopulations and by doing so a substantial improvement in the fit is achieved; the residual lack of fit $\Lambda = 322.7$ with $df = 33$. Nevertheless, the residual lack of fit is still overwhelming.

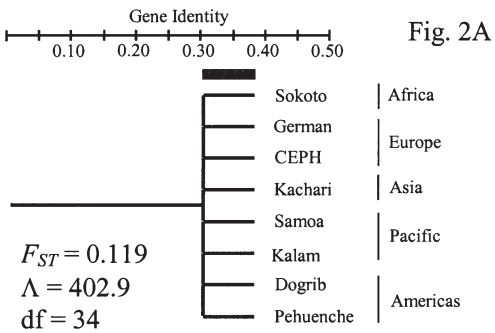


Fig. 2A

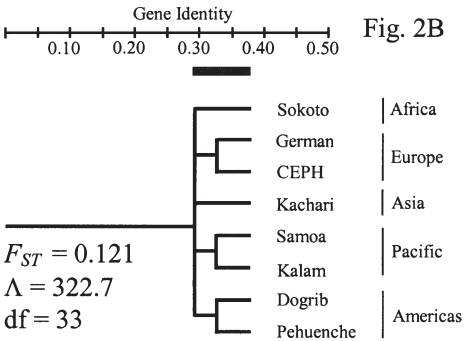


Fig. 2B

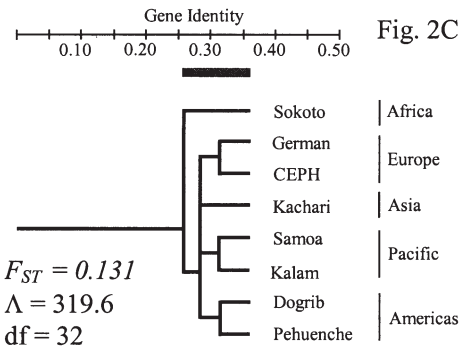


Fig. 2C

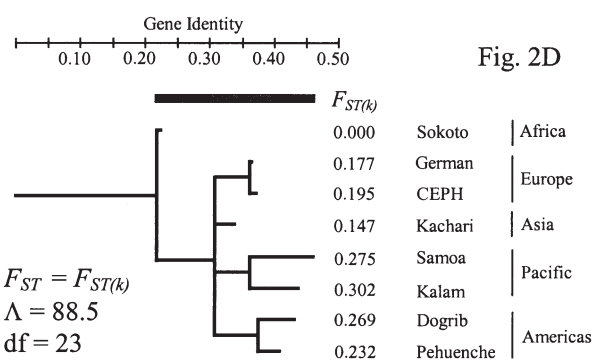


Fig. 2D

Figure 2. Tree diagrams for sequential *F*-statistic models. The length of each branch represents the increase in gene identity (homozygosity). The distance from the origin to the first node represents the quantity J_T. The distance from the origin to a terminal node represents the quantity J_S in 2**A**–2**C** and J_k in 2**D**. The shaded bar above each tree measures the level of gene diversity among populations D_{ST}. The statistic Λ is ideally distributed as a chi-square random variable with the indicated degrees of freedom. Each tree indicates a statistically significant lack of fit.

Table 2. Gene Identity Within and Between Samples[a]

Sample	*Sokoto*	*German*	*CEPH*	*Karachi*	*Samoa*	*Kalam*	*Dogrib*	*Pehuenche*	*Chimp*
Sokoto	**0.213**	0.205	0.196	0.212	0.247	0.220	0.231	0.249	0.077
German	0.129	**0.325**	0.312	0.276	0.339	0.262	0.314	0.361	0.056
CEPH	0.133	0.011	**0.311**	0.262	0.314	0.250	0.303	0.344	0.055
Kachari	0.097	0.079	0.093	**0.307**	0.349	0.384	0.305	0.355	0.076
Samoa	0.260	0.187	0.225	0.151	**0.541**	0.384	0.403	0.453	0.093
Kalam	0.178	0.204	0.214	0.184	0.177	**0.404**	0.301	0.334	0.086
Dogrib	0.214	0.160	0.167	0.160	0.197	0.265	**0.463**	0.443	0.076
Pehuenche	0.237	0.125	0.144	0.119	0.158	0.258	0.099	**0.522**	0.069
Chimp	0.472	0.625	0.613	0.568	0.769	0.645	0.724	0.797	**0.413**

a. Gene identity within populations is given along the diagonal in boldface. Gene identity between pairs of subpopulations is given in the upper triangle. Nei's (1987) minimum genetic distance between pairs of subpopulations is given in the lower triangle.

Model III (Figure 2C) expands the hierarchy to three levels by nesting the non-African subpopulations together independently from the African subpopulation. This configuration is consistent with a large body of molecular data (Bowcock et al. 1991, 1994; Cavalli-Sforza et al. 1988, 1994; Nei and Roychoudhury 1993; Jorde et al. 1995) that indicates a primary division between subpopulations south of the Sahara and the remainder of the world. Interestingly, $\hat{F}_{ST}$ for the total sample increases modestly, but the increase in fit resulting from this is paltry, $\Lambda = 319.6$. In fact, it does not attain statistical significance because the cost of adding an additional parameter to the model is the loss of a degree of freedom.

Model IV (Figure 2D) retains the nesting structure of model III. However, it relaxes the assumption that gene identity is the same in all subpopulations. This requires estimating eight new parameters. There is a dramatic improvement in fit, $\Lambda = 88.5$, despite the loss of these degrees of freedom ($df = 23$). F_{ST} is rendered meaningless as a single quantity in the context of this model because the individual values $\hat{J}_k$ are not averaged to obtain $\hat{J}_S$. A population-specific F_{ST} can be defined:

$$F_{ST(k)} = \frac{J_k - J_T}{1 - J_T}.$$

Clearly, $F_{ST(k)}$ is interpreted as the relative gain in gene identity in the kth subpopulation relative to the gene identity in the population as a whole. Estimates of $\hat{F}_{ST(k)}$ vary broadly among the eight subpopulations. The basic estimation (model I) yielded $\hat{F}_{ST} = 0.119$, which is a small fraction of the greatest value $\hat{F}_{ST(\text{Kalam})} = 0.302$ and greatly exceeds the smallest value $\hat{F}_{ST(\text{Sokoto})} = 0.000$. This later value implies that the Sokoto possesses more gene diversity than that harbored by all of the non-African populations combined together. Interestingly, some of the non-African populations exhibit a reduction in gene diversity that approaches F_{ST} values typical of those in sibling species (Wright 1978). An unexpected result

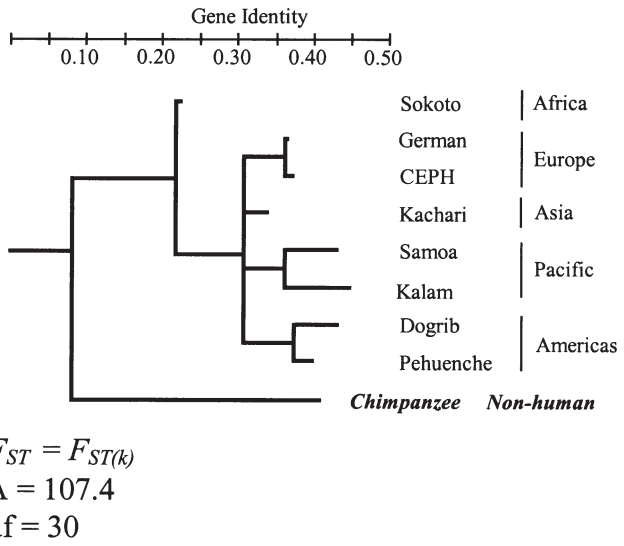


Figure 3. Tree diagram for *F*-statistic analysis of both humans and chimpanzees allowing for a complex human population structure.

of fitting this model is evident on the branch dividing the Sokoto from the non-African populations. Notice that this branch was small and not statistically significant in the former model, which constrained gene identity within subpopulations. Here it is the largest (and most statistically significant) branch of the tree. While the relative increase in gene identity differs by subpopulation, in some subpopulations it is almost twice that observed in the earlier models.

The estimates of F_{ST} for the non-African subpopulations indicate greater population differentiation than those typically encountered in population genetic studies, but this is even more stunning when we consider that in all cases $\max(F_{ST}) \ll 1.0$. For the New Guinea subpopulation, $\max(F_{ST}) = \hat{J}_{\text{Kalam}} = 0.404$ and $\hat{F}_{ST(\text{Kalam})} = 0.302$. In other words, this subpopulation exhibits only 75% of the genetic divergence possible without further increasing gene identity. For the African Sokoto, $\max(F_{ST}) = \hat{F}_{ST(\text{Sokoto})} = 0.213$ and $\hat{F}_{ST(\text{Sokoto})} = 0.000$. The maximum permissible level of divergence is low and more importantly this population has not lost gene diversity relative to the situation that would be envisioned if all eight human subpopulations formed one panmictic unit.

It is now worth asking whether or not the result for chimpanzees will be clarified by taking into account the statistical issues raised above. For this purpose, we added the chimpanzee sample to the complete structure of model IV. As shown in Figure 3, the chimpanzees stand well apart from the human samples. $F_{ST(\text{Chimpanzee})} = 0.364$, double the value for humans and chimpanzees when it is assumed that all subpopulations, regardless of species, are evolving independently. $F_{ST(\text{Chimpanzee})}$ is also well beyond the arbitrary one quarter threshold recommended

by Wright (1978) for very great differentiation. Moreover, $\max(F_{ST}) = \hat{J}_{\text{Chimpanzee}} = 0.413$ so that chimpanzees exhibit 88% of possible divergence from the baseline that would be formed by drawing genes at random from all subpopulations (chimpanzees and humans). The remaining 12% possibly reflects back mutations and constraints on allele sizes with repeat polymorphisms. This is consistent with Garza et al. (1995), who found that allele sizes at several microsatellite loci in chimpanzees and humans are too similar to be evolving without constraint.

Discussion

This study succeeds in demonstrating the shortcomings of F_{ST} for two principal reasons. First, the chimpanzee sample provided an example where F_{ST} fails to show obviously significant differentiation. This is an attention getter. Second, we recognized that F_{ST} implicitly assumes a specific model tree. This recognition led us to the test for lack of fit, and to a structured approach to exploring models with greater complexity. The need to test the fit of trees has been recognized and developed by others (Cavalli-Sforza and Piazza 1975; Templeton 1998), but tests have not been applied to analyses using F_{ST}. Ironically, the tree that represents the usual model for estimating F_{ST} is the simplest and worst fitting tree, yet results based on such simple models have been widely accepted as portraying a fundamental property of human diversity (e.g., Brown and Armelagos 2001; Templeton 1998).

We show that the initial assumptions of partitioning gene diversity can heavily bias the outcome of analyses. It is particularly disturbing that assumptions made while estimating parameters at a low level of hierarchy (i.e., within subpopulations) can have large effects on estimates of parameters at higher levels (i.e., among subpopulations). The seemingly innocuous assumption of equal gene identity within groups greatly affects the estimated relationship between the African Sokoto population and the seven non-African populations (Figures 2C, 2D). While each model in our sequential analysis fits the data better than its predecessor, our best model still has a significant lack of fit. This questions the stability of conclusions from this analysis. It must be acknowledged that a different model, perhaps one that incorporates gene flow or other mechanisms for nonindependence across major branches, would provide a better fit and portray the patterns of relationship differently. In fact, a recent analysis of a sample of geographically diverse individuals (Wilson et al. 2001) revealed an imperfect correspondence between ethnicity and empirical clusters identified by a numerical algorithm. This finding was important because the genetic clusters were better predictors of allele frequencies at drug metabolizing loci than were the a priori ethnic labels.

A recent study examined the distribution of F_{ST} estimates obtained individually for 26,530 single nucleotide polymorphisms (Akey et al. 2002). It was postulated that some loci affected by natural selection could be identified because they would provide either unusually high or unusually low estimates of F_{ST}. This intriguing idea follows the work of Lewontin and Krakauer (1973), who developed

a formal statistical test for selection nearly three decades ago. The Lewontin-Krakauer test assumes a simple island model of population structure, the same model that was simulated by Akey et al. (2002). Unfortunately, Robertson (1975) showed that the expected variation of F_{ST} estimates over loci is greatly exaggerated by complex hierarchical structure. Robertson (1975) also showed that human populations appeared to have complex hierarchical structure. Our analyses of microsatellites in this paper confirm Robertson's finding; human population structure is indeed complex and the simple island model is a very poor approximation. The evidence provided by F_{ST} to detect natural selection should be interpreted with great caution.

The widespread belief that F_{ST} measures relative diversity independently of absolute diversity has been contradicted in several recent publications (Nagylaki 1998; Hedrick 1999; Balloux and Lugon-Moulin 2002). As we show, the maximum for F_{ST} is the within subpopulation gene identity and therefore quite dependent on allele frequency (also see Hedrick 1999). This is a likely explanation for the strong negative correlation between estimated F_{ST} and microsatellite heterozygosity reported by Bowcock et al. (1994). It is also a likely explanation for the more recent finding that estimates of F_{ST} on a worldwide basis are lower for microsatellite polymorphisms than they are for biallelic *Alu* insertion polymorphisms (Bamshad et al. 2003).

Although it was instructive here to compare $F_{ST(k)}$ to its theoretical maximum in several cases, this is unlikely to remove all scale effects. In fact, scale-free statistical measures do not exist (Lewontin 1988). It is important to be aware of scale effects and their impact on the analysis being conducted. Sewall Wright (1978: 82) wrote, "The fixation index is thus not a measure of degree of differentiation in the sense implied in the extreme case by absence of any common allele. It measures differentiation within the total array in the sense of the extent to which the process of fixation has gone toward completion." Wright (1978) provides a hypothetical example for two subpopulations where F_{ST} is only one-third despite the fact that they have no alleles in common. This is the reason that Sewall Wright (1978), a founder of population genetics and the principal architect of F_{ST}, believed that $F_{ST} = 0.15$ represents a moderate to great level of differentiation.

The belief that human genetic diversity on a global scale can be reduced to simple statistical partitions has limited our understanding of diversity and thwarted training in biological anthropology. For example, a current textbook (Boyd and Silk 2000) states, "Geneticists computed the amount of variation in these characters within each local group, among groups within each race, and among the races. They found that there is much more genetic variation *within* local groups than there is *among* local groups or among races themselves. Differences within local groups account for about 85% of all the variation in the human species. To put this another way, suppose a malevolent extraterrestrial wiped out the entire human species except for one local group, which it preserved in an extraterrestrial zoo. The alien could pick any local group at random—the Efe, the Inuit, the citizens of Ames, Iowa, or the people of Patagonia—and then wipe

Table 3. Biological Definitions of Race

Concept	*Reference*	*Definition*
Essentialist	Hooton (1926)	"A great division of mankind, characterized as a group by the sharing of a certain combination of features, which have been derived from their common descent, and constitute a vague physical background, usually more or less obscured by individual variations, and realized best in a composite picture."
Population	Dobzhansky (1970)	"Races are genetically distinct Mendelian populations. They are neither individuals nor particular genotypes, they consist of individuals who differ genetically among themselves."
Taxonomic	Mayr (1969)	"An aggregate of phenotypically similar populations of a species, inhabiting a geographic subdivision of the range of a species, and differing taxonomically from other populations of the species."
Lineage	Templeton (1998)	"A subspecies (race) is a distinct evolutionary lineage within a species. This definition requires that a subspecies be genetically differentiated due to barriers to genetic exchange that have persisted for long periods of time; that is, the subspecies must have historical continuity in addition to current genetic differentiation."

out the rest of the humans on the planet. This group would still contain on average 85% of the genetic variation that exists in the entire human species." However, our analysis indicates that it would make a great difference which group is chosen. For example, no gene diversity would be lost if the Sokoto were chosen while nearly one-third would be lost by choosing the subpopulation from Papua New Guinea. It is important to point out here that the rich genetic diversity within Africans is a robust finding that is not peculiar to the loci or specific samples analyzed here. Recently, Yu et al. (2002) assayed nucleotide substitutions in 50 randomly chosen noncoding DNA segments (~500 base pairs) in 30 individuals: 10 Africans, 10 Europeans, and 10 Asians. The subjects within each continent were chosen widely from dispersed geographic locations. Interestingly, nucleotide diversity was greater within the Africans than within either Asians or Europeans. More importantly, the nucleotide diversity was greater within Africans than *between* Europeans and Asians.

We must also ask about the relevance of our findings to the existence of human races. It is difficult to answer this question because race has a spectrum of biological definitions, despite the fact that two prominent biologists have recently argued that the definition of race is clear to most people (Crow 2002; Mayr 2002). Table 3 reviews four different biological race concepts. It begins with the typological views of early-20th-century biological anthropology and proceeds to more modern population-based definitions. In most circumstances, biologists equate races with subspecies and these terms are used interchangeably. Surprisingly, a

good deal of genetic diversity within groups is consistent with the entire spectrum represented by these concepts. This is clear from the wording of both Hooton's typological concept and Dobzhansky's population concept.

It is less apparent a good deal of genetic variation within groups is consistent with Mayr's taxonomic concept. Nevertheless, complete taxonomic resolution is possible even with a great deal of within group variation. This is exemplified in this study by the locus *D13S122* (Figure 1) where both humans and chimpanzees are highly polymorphic but their divergence is near complete. Moreover, when Rosenberg and colleagues (2001) applied a numerical clustering algorithm (Pritchard et al. 2000) to 377 microsatellite loci genotyped on a large sample of individuals from diverse geographic regions, they found strong evidence for the existence of six geographically associated clusters despite the fact that the within-population component of genetic variation accounted for between 93% and 95% of the total genetic variation.

High within-group variance is also consistent with the population lineage concept, because mutation will introduce novel variation within a divergent lineage over a long time. Therefore, gene diversity within old lineages is expected to be near its mutation/drift equilibrium. The relative proportion of variation within and among groups therefore appears to be meaningless as a criterion for judging the validity of races or subspecies as defined by biologists.

The biological concepts of race identified in the preceding paragraph are distinct from common lay conceptualizations of race. One such lay concept postulates the existence of near-uniform groups of individuals that can be identified by a few externally visible traits such as skin color (Keita and Kitties 1997). The AAPA statement on race (American Association of Physical Anthropologists 1996) articulates a counterargument to this popular view. In fact, our findings are consistent with the key features of the AAPA view: that all human populations derive from a common ancestral group, that there is great genetic diversity within all human populations, and that the geographic pattern of variation is complex and presents no major discontinuity. There are many other lay concepts of race but these are inconsistent with genetic inheritance and are unrelated to levels of variation or percentages of ancestry. For example, in the Untied States the pernicious *one drop rule* (Fredrickson 2002) is based on a social inheritance pattern—any traceable ancestry to a subordinate group—that is inconsistent with all five modes of genetic transmission: autosomal, X-linked, Y-linked, mitochondrial, and pseudoautosomal. In Brazil, the equivalent of race is *Color*, which is judged for an individual along a continuum determined by a composite of pigmentation, hair form, and facial morphology (Parra et al. 2003). Ancestry is not a direct criterion for *Color* and even full siblings can be classified differently for *Color*. A recent study in Brazil (Parra et al. 2003) has shown that *Color* is a poor predictor of genomic ancestry as determined by the assay of marker alleles.

In conclusion, the value of F_{ST} computed from worldwide samples reflects misleading statistical biases. The patterns of variation within and between groups are too intricate to be reduced to a single summary measure. Surprisingly, a great

deal of variation within groups is compatible with biological race concepts and therefore partitions of genetic variation such as those achieved by simple statistics such as F_{ST} do not provide critical tests for the existence of races as defined by biologists. Four decades ago, Frank Livingstone declared the nonexistence of human races (Livingstone 1963). It is now time for geneticists and anthropologists to stop worrying about what does not exist and to discover what does exist.

Acknowledgments We thank Mike Boehnke, Keith Hunley, Ripan Malhi, Connie Mulligan, Sumaiya Olatunde, Carrie Pfaff, and Ken Weiss for their constructive comments on earlier versions of this paper. The authors accept sole responsibility for the contents.

Received 22 July 2002; revision received 9 May 2003.

Literature Cited

Akey, J. M., G. Zhang, K. Shang et al. 2002. Interrogating a high-density SNP map for signatures of natural selection. *Genome Research* 12:1805–1814.

American Association of Physical Anthropologists. 1996. AAPA statement on biological aspects of race. *Am. J. Phys. Anthropol.* 101:569–570.

Anderson, T. W. 1973. Asymptotically efficient estimation of covariance matrices with linear structure. *Ann. Statistics* 1:135–141.

Balloux, F., and N. Lugon-Moulin. 2002. The estimation of population differentiation with microsatellite markers. *Molecular Ecology* 11:155–165.

Bamshad, M. J., S. Wooding, W. S. Watkins et al. 2003. Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* 72:578–589.

Barbujani, G., A. Magani, E. Minch et al. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. (USA)* 94:4516–4519.

Bowcock, A. M., J. R. Kidd, J. Mountain et al. 1991. Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proc. Natl. Acad. Sci. (USA)* 88:839–843.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde et al. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455–457.

Boyd, R., and J. B. Silk. 2000. *How Humans Evolved*, 2nd ed. New York: W. W. Norton.

Brown, R. A., and G. J. Armelagos. 2001. Apportionment of racial diversity: A review. *Evol. Anthropol.* 10:34–40.

Cavalli-Sforza, L. L., and A. Piazza. 1975. Analysis of evolution: Evolutionary rates, independence, and treeness. *Theor. Popul. Biol.* 37:3–25.

Cavalli-Sforza, L. L., A. Piazza, P. Menozzi et al. 1988. Reconstruction of human evolution: Bringing together genetic, archeological, and linguistic data. *Proc. Natl. Acad. Sci. (USA)* 85:6002–6006.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes* (abridged paperback edition). Princeton, NJ: Princeton University Press.

Cockerham, C. C. 1969. The variance of gene frequencies. *Evolution* 23:72–84.

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

Cockerham, C. C., and B. S. Weir. 1986. Estimation of inbreeding parameters in stratified populations. *Ann. Hum. Genet.* 50:271–281.

Count, G. W. 1950. *This is Race*. New York: Henry Schumann.

Crow, J. F. 2002. Unequal by nature: A geneticist's perspective on human differences. *Daedalus* 131:81.

Deka, R., L. Jin, M. D. Shriver et al. 1995. Population genetics of dinucleotide $(dC–dA)_n$.$[(dG–dT)_n$ polymorphisms in world populations. *Am. J. Hum. Genet.* 56:461–474.

Dobzhansky, T. 1970. *Genetics of the Evolutionary Process*. New York: Columbia University Press.
Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131:479–491.
Fredrickson, G. M. 2002. *Racism: A Short History*. Princeton, NJ: Princeton University Press.
Garza, J. C., M. Slatkin, and N. B. Freimer. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* 12:594–603.
Harpending, H., and A. Rogers. 2000. Genetic perspectives on human origins and differentiation. *Ann. Rev. Genomics Hum. Genet.* 1:361–385.
Hedrick, P. W. 1999. Highly variable loci and their interpretation in evolution and conservation. *Evolution* 53:313–318.
Hooton, E. A. 1926. Methods of racial analysis. *Science* 63:75–81.
Jorde, L. B., M. J. Bamshad, W. S. Watkins et al. 1995. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57:523–538.
Keita, S. O. Y., and R. A. Kitties. 1997. The persistence of racial thinking and the myth of racial divergence. *Am. Anthropol.* 99:534–544.
Lewontin, R. C. 1972. The apportionment of human diversity. *Evol. Biol.* 6:381–398.
Lewontin, R. C. 1988. On measures of gametic disequilibrium. *Genetics* 120:849–852.
Lewontin, R. C., and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of selective neutrality of polymorphism. *Genetics* 74:175–195.
Lieberman, L., and F. L. C. Jackson. 1995. Race and three models of human origin. *Am. Anthropol.* 97:231–242.
Livingstone, F. B. 1963. On the non-existence of human races. *Curr. Anthropol.* 3:279–281.
Mayr, E. 1969. *Principles of Systematic Zoology*. New York: McGraw-Hill.
Mayr, E. 2002. The biology of race and the concept of equality. *Daedalus* 131:89.
Morrison, D. F. 1976. *Multivariate Statistical Methods*. New York: McGraw-Hill.
Nagylaki. T. 1998. Fixation indices in subdivided populations. *Genetics* 148:1325–1332.
Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.
Nei, M., and A. K. Roychoudhury. 1993. Evolutionary relationships of human populations on a global scale. *Mol. Biol. Evol.* 10:927–943.
Parra, F. C., R. C. Amado, J. R. Lambertucci et al. 2003. Color and genomic ancestry in Brazilians. *Proc. Natl. Acad. Sci. USA* 100:177–182.
Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
Quenouille, M. 1956. Notes on bias in estimation. *Biometrika* 43:253–260.
Robertson, A. 1975. Gene frequency distributions as a test of selective neutrality. *Genetics* 81:775–785.
Rosenberg, N. A., J. K. Pritchard, J. L. Weber et al. 2001. Genetic structure of human populations. *Science* 298:2381–2385.
Templeton, A. R. 1998. Human races: A genetic and evolutionary perspective. *Am. Anthropol.* 100: 632–650.
Urbanek, M., D. Goldman, and J. C. Long. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: A new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol. Biol. Evol.* 13:943–953.
Weir, B. S. 1996. *Genetic Data Analysis*. Sunderland, MA: Sinauer Associates, Inc.
Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1358–1370.
Weir, B. S., and W. G. Hill. 2002. Estimating *F*-statistics. *Annu. Rev. Genet.* 36:721–750.
Wilson, E. O., and W. L. Brown. 1953. The subspecies concept and its taxonomic application. *Syst. Zool.* 2:214–220.
Wilson, J. G., M. E. Weale, A. C. Smith et al. 2001. Population genetic structure of variable drug response. *Nature Genetics* 29:265–269.
Wright, S. 1951. The genetical structure of populations. *Annals of Eugenics* 15:323–354.

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

Wright, S. 1969. *Evolution and the Genetics of Populations*. Vol. 2, *The Theory of Gene Frequencies*. Chicago: University of Chicago Press.

Wright, S. 1978. *Evolution and the Genetics of Populations*. Vol. 4, *Variability Within and Among Natural Populations*. Chicago: University of Chicago Press.

Yu, N., F.-C. Chen, S. Ota et al. 2002. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* 161:269–274.