Update to Long and Kittles's "Human Genetic Diversity and the Nonexistence of Biological Races" (2003): Fixation on an Index

Jeffrey C. Long

Human Biology, Volume 81, Numbers 5-6, October-December 2009, pp. 799-803 (Article)

Published by Wayne State University Press

Update to Long and Kittles's "Human Genetic Diversity and the Nonexistence of Biological Races" (2003): Fixation on an Index

JEFFREY C. LONG[1]

Sewall Wright's fixation index F_{ST} measured among samples of world populations is often 0.15 or less when computed as an average over many alleles or loci. To many, this result indicates that the genetic similarities among human populations far outweigh the differences. For example, a finding like this led Richard Lewontin to claim that human races have no genetic or taxonomic significance (Lewontin 1972). Despite the far-reaching proclamations that researchers make from F_{ST}, few have questioned the validity of how it is applied or interpreted.

Earlier in this decade, Rick Kittles and I took an unusually critical look at F_{ST} (Long and Kittles 2003). We analyzed a unique data set composed of short tandem repeat (STR) allele frequencies for eight loci genotyped in both humans and chimpanzees (Deka et al. 1995). These data made it possible to see how F_{ST} played out when no one could dispute taxonomic and genetic significance. The answer surprised us. F_{ST} was pretty close to the canonical 0.15 shown so many times for human populations. In our analysis, F_{ST} was 0.12 for humans, but for humans and chimpanzees together, F_{ST} rose only to 0.18. Indeed, we found one locus, *D13S122*, where the size range of human and chimpanzee alleles hardly overlapped, yet F_{ST} equaled 0.15 (Figure 1). We ultimately found that the genetic and statistical model underlying F_{ST} does not fit well to human populations. Specifically, human population structure strongly biases the outcome of analyses by violating two assumptions: first, that expected genetic diversity is the same in every population; and second, that divergence between all pairs of populations is equal and independent. These assumptions are explicit and clear in the major statistical papers on estimating F_{ST} (Cockerham 1969; Weir and Cockerham 1984; Weir and Hill 2002), but most researchers ignore them. More important, Kittles and I introduced a way to relax these assumptions by using generalized hierarchical models that nest smaller units, such as genes, into larger units, such as individuals, populations, and geographic regions. In our approach, it is possible to restate many hierarchical models as either expansions or reductions of each other, and by comparing a sequence of nested models, we are able to identify those

[1]Department of Anthropology, University of New Mexico, Albuquerque, NM 87131.

Human Biology, October–December 2009, v. 81, nos. 5–6, pp. 799–803.

KEY WORDS: F_{ST}, SHORT TANDEM REPEATS (STRS), SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS), CHIMPANZEES, HIERARCHICAL MODELS, NESTED MODELS, GENETIC DIVERSITY, RACIAL TYPING.

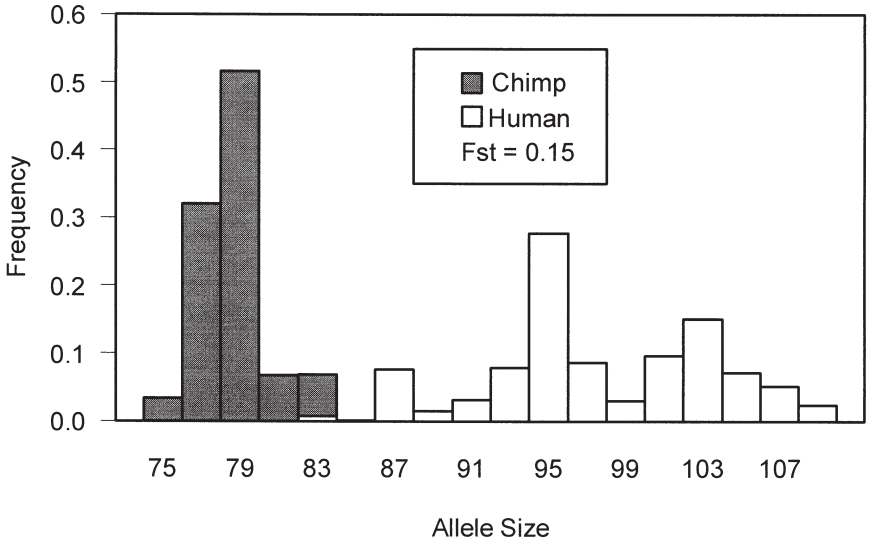

Figure 1. Dinucleotide repeat allele size distributions for the *D13S122* locus in chimpanzees and humans. The frequencies for humans are unweighted averages from eight groups representing diverse worldwide localities (Deka et al. 1995). Notice the clear distinction between humans and chimpanzees, but $F_{ST} = 0.15$.

human demographic features that have the greatest effect on the distribution of genetic variation.

Continuing Lapses in Using F_{ST} Critically

A recent review notes that different kinds of genetic markers give different estimates of F_{ST} (Holsinger and Weir 2009). For example, F_{ST} estimated from STRs is 0.05, but F_{ST} estimated from single nucleotide polymorphisms (SNPs) is 0.09 (Li et al. 2008; Rosenberg et al. 2002). This discrepancy should be no surprise because F_{ST} depends on allele frequencies; it is inversely proportional to the variation within populations, and STRs are more variable than SNPs. Kittles and I demonstrated this with algebra, but the finding was not novel. Wright had pointed to it in his major synthesis (Wright 1978), and Phillip Hedrick had shown the same result for a slightly different statistic (Hedrick 1999). In 2009, my colleagues and I showed that patterns of variation in STRs and segregating sites in DNA sequences are concordant after standardizing variance components and fixation indexes (such as F_{ST}) relative to their theoretical maxima (Long et al. 2009).

Holsinger and Weir (2009) say correctly that the differences between data types make it hard to estimate population genetic parameters such as effective migration rates ($N_e m$) from F_{ST} (Holsinger and Weir 2009). However, they fail to warn readers about another looming problem—namely, estimates of scaled migration

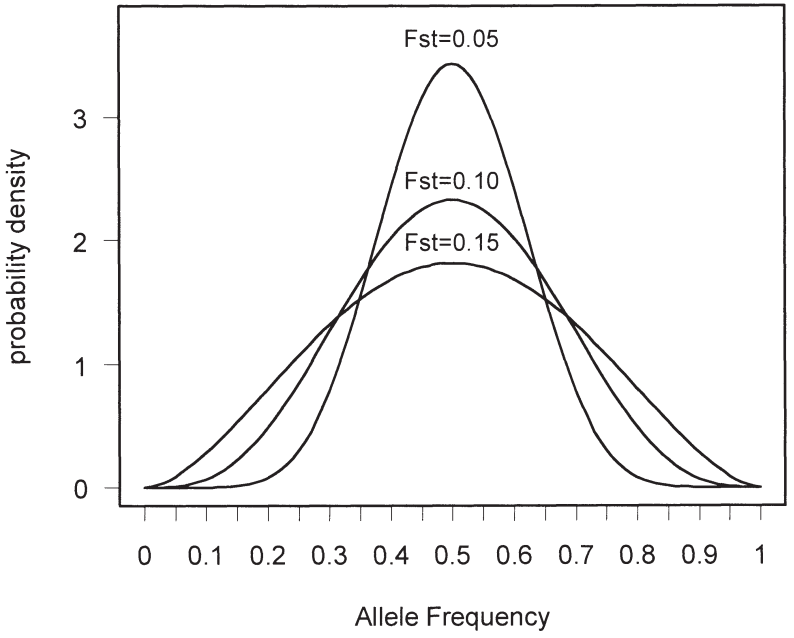

Figure 2. The relation between variation of gene frequencies among island populations and the value of F_{ST}. The mean allele frequency is assumed to be $q_T = 0.5$. Drawn according to Wright (1951).

rates from F_{ST} require a sample of genetic polymorphisms that represent variation in the genome as a whole. SNPs do not represent segregating nucleotide sites in the genome (Clark et al. 2005); compared to completely resequenced neutral loci, SNPs have higher variation within groups and mask differences in the extent of rare variants between populations. The effect of ascertainment bias is apparent from the fact that F_{ST} estimated from DNA sequences is nearly 0.16, which is higher than the F_{ST} estimated from either SNPs or STRs (Long et al. 2009; Wall et al. 2008).

F_{ST} and Race

Richard Lewontin's dismissal of race may not have led to the wide popularity of F_{ST} in population biology, but it did galvanize anthropology. Lewontin confronted race by trying to show that classical racial groupings accounted for too little of the total diversity to be of any value. In retrospect, it is odd that Lewontin felt that 15% of variation among groups is small and even odder that others have concurred. Sewall Wright, the inventor of F_{ST}, believed the opposite. To Wright, $F_{ST} = 0.05$ or even less indicates considerable differences, and $F_{ST} = 0.15$ reflects moderately great differences (Wright 1951, 1978). Low values of F_{ST} reflect large gene frequency differences in replicate populations (Figure 2). In

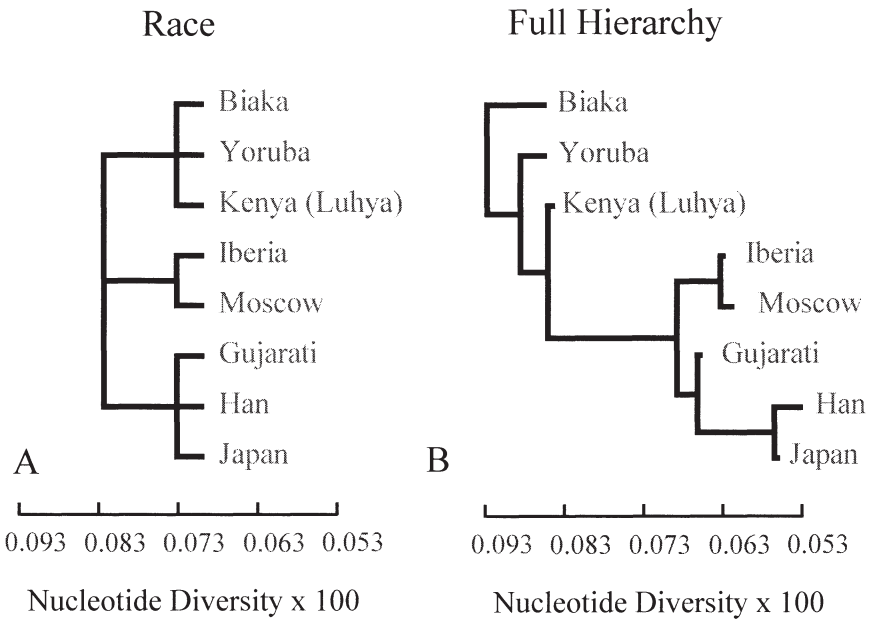

Figure 3. Diagrams of two population structure models fitted to DNA sequences. Each graph is calibrated to the nucleotide diversity scale below it. Statistical tests show that the full hierarchy fits the data significantly better than the race classification (Long et al. 2009).

other words, these seemingly small values of F_{ST} permit allele frequencies to drift widely among populations. Unfortunately, Lewontin did not contest the larger issue, which is whether or not races are a good way to portray the pattern of gene frequency differences between populations.

Now, with more genetic data and more populations sampled, we are able to revisit the race problem with greater accuracy. Recently, my colleagues and I have tested the usefulness of race as a way to describe genetic differences among populations by contrasting the results of racial classification with those from generalized hierarchical models (Long et al. 2009). Race fails! Figure 3 diagrams the contrast for a data set consisting of complete DNA sequences for 64 autosomal loci (38,000 bp total). Four resequenced individuals represent each population. A summary of the major problems with using race are as follows. First, imposing the classically defined race structure on populations causes us to estimate less diversity for the species as a whole than does allowing all populations to link back to a common base population in an unrestricted hierarchy. Second, using the race pattern causes us to estimate excess diversity within non-sub-Saharan African populations, but it estimates a deficit of diversity within sub-Saharan African populations. Third, the supposition of races forces all continental populations to diverge equally from a single ancestral node, whereas an unrestricted hierarchy

places the basal split within Africa. Fourth, in the classical race framework, European and Asian populations diverge from African populations independently, but the unrestricted hierarchy shows that European and East Asian populations link together before either links to sub-Saharan Africans.

Concluding Remarks

In the end, we must contend with the fact that a single index such as F_{ST} cannot tell us everything we might like to know. Even if many population structure models accurately predict F_{ST}, we cannot be so naïve as to assume that we can turn the process around and trust that we can translate that index back to the structure of a natural population. The mass of genetic information available to us today allows us to go beyond estimating one index. We can now evaluate patterns, and this should be our pursuit.

Received and accepted for publication 19 November 2009.

Literature Cited

Clark, A. G., M. J. Hubisz, C. D. Bustamante et al. 2005. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15:1496–1502.

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

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.

Hedrick, P. W. 1999. Highly variable loci and their interpretation in evolution and conservation. *Evolution* 53:313–318.

Holsinger, K. E., and B. S. Weir. 2009. Genetics in geographically structured populations: Defining, estimating, and interpreting *F*(*ST*). *Nat. Rev. Genet.* 10:639–650.

Lewontin, R. C. 1972. The apportionment of human diversity. *Evol. Biol.* 6:381–398.

Li, J. Z., D. M. Absher, H. Tang et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Long, J. C., and R. A. Kittles. 2003. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* 75(4):449–471.

Long, J. C., J. Li, and M. E. Healy. 2009. Human DNA sequences: More variation and less race. *Am. J. Phys. Anthropol.* 139:23–34.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber et al. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Wall, J. D., M. P. Cox, F. L. Mendez et al. 2008. A novel DNA sequence database for analyzing human demographic history. *Genome Res.* 18:1354–1361.

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.

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

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