Evolution, 59(8), 2005, pp. 1633–1638

A STANDARDIZED GENETIC DIFFERENTIATION MEASURE

PHILIP W. HEDRICK
School of Life Sciences, Arizona State University, Tempe, Arizona 85287-4501
E-mail: philip.hedrick@asu.edu

Abstract.—Interpretation of genetic differentiation values is often problematic because of their dependence on the level of genetic variation. For example, the maximum level of G_{ST} is less than the average within population homozygosity so that for highly variable loci, even when no alleles are shared between subpopulations, G_{ST} may be low. To remedy this difficulty, a standardized measure of genetic differentiation is introduced here, one which has the same range, 0–1, for all levels of genetic variation. With this measure, the magnitude is the proportion of the maximum differentiation possible for the level of subpopulation homozygosity observed. This is particularly important for situations in which the mutation rate is of the same magnitude or higher than the rate of gene flow. The standardized measure allows comparison between loci with different levels of genetic variation, such as allozymes and microsatellite loci, or mtDNA and Y-chromosome genes, and for genetic differentiation for organisms with different effective population sizes.

Key words.—Gene flow, heterozygosity, linkage disequilibrium, microsatellite loci, mutation.

Received February 9, 2005. Accepted May 31, 2005.

Determination of the extent of genetic differentiation between groups is of great interest in a number of biological fields including evolution, conservation, plant and animal breeding, and anthropology. There are a number of approaches used to quantify and visualize genetic differentiation (Excoffier 2001; Charlesworth et al. 2003) but it is often difficult to compare the amount of differentiation between different studies, loci, or organisms (Charlesworth 1998; Nagylaki 1998; Hedrick 1999). Here I propose a standardized measure of genetic differentiation, related to the widely used genetic differentiation measure G_{ST}, which has a range from 0 to 1 for all loci, independent of the extent of within subpopulation genetic variation.

GENERAL BACKGROUND

Wright (1951, 1965) developed an approach to partition the genetic variation in a subdivided population that is commonly used and provides an obvious description of differentiation. He denoted F_{ST} as a measure of the genetic differentiation, where S and T indicate the subpopulation and total population levels, as

$$F_{ST} = \frac{V(q)}{\bar{q}(1 - \bar{q})} \tag{1}$$

where q is the frequency of allele A_2 at a biallelic locus and $V(q)$ is the variance over subpopulations. If the variance is 0, that is, the same allele frequencies are present in all the subpopulations, then $F_{ST} = 0$. If the variance is at its binomial maximum, $q(1 - q)$, that is, the subpopulations are fixed for different alleles, then $F_{ST} = 1$. In other words, the range of F_{ST} is from 0 to 1 for a biallelic locus.

Nei (1973) provided an estimate of F_{ST} for a locus with multiple alleles, assuming Hardy–Weinberg proportions, as

$$G_{ST} = \frac{H_T - H_S}{H_T} \tag{2}$$

where H_S is the average subpopulation Hardy–Weinberg heterozygosity and the total population heterozygosity is $H_T = 1 - \Sigma\, \bar{p}_i^2$ for any number of alleles, where $\bar{p}_i$ is the average frequency of allele i over subpopulations.

Unlike F_{ST} for two alleles, the magnitude of G_{ST} is particularly dependent on the amount of genetic variation for highly variable genes, such as microsatellite loci, where both H_S and H_T can approach unity. As a result, G_{ST} may not range from 0 to 1 and can be very small even if the subpopulations have nonoverlapping sets of alleles (Hedrick 1999). This at first seems counterintuitive because, as we discussed above in the two-allele case, when the subpopulations are monomorphic for different alleles, $F_{ST} = 1$. However, G_{ST} measures the amount of variation between subpopulations, relative to the total population variation, and does not specify the identity of the alleles involved.

The magnitude of G_{ST} can also be written as

$$G_{ST} = 1 - \frac{H_S}{H_T} < 1 - H_S \tag{3}$$

where $1 - H_S$ is the average within subpopulation homozygosity. From this, it is obvious that the differentiation measure cannot exceed the level of within-subpopulation homozygosity, no matter what evolutionary factor is influencing the amount and pattern of variation. Obviously, when using highly polymorphic markers that make the level of homozygosity low, then the maximum G_{ST} value must also be greatly reduced.

A NEW STANDARDIZED MEASURE

Theory

One approach to circumvent the restriction on the range of G_{ST} is to standardize the observed value of G_{ST} by the maximum level that it can obtain for the observed amount of genetic variation. This is analogous to the measure of linkage disequilibrium, D', of Lewontin (1964), which is D/D_{max}, where D is the traditional measure of linkage disequilibrium and D_{max} is the maximum value it can take, given the observed allele frequencies.

Assume that there are k equal-sized subpopulations and that each allele in each subpopulation is unique. That is, all

TABLE 1. Examples illustrating the effect of heterozygosity on measures of genetic differentiation. (a), (b), and (c) have $H_S = 0.25$ and $G_{ST(max)} = 0.6$ whereas (d), (e), and (f) have $H_S = 0.58$ and $G_{ST(max)} = 0.266$.

	(a) Subpopulation		(b) Subpopulation		(c) Subpopulation	
Allele	1	2	1	2	1	2
1	0.1	—	0.1	—	0.9	—
2	0.9	—	0.9	0.8	0.1	0.2
3	—	0.8	—	0.2	—	0.8
4	—	0.2	—	—	—	—
H_S	0.25		0.25		0.25	
$G_{ST(max)}$	0.6		0.6		0.6	
G_{ST}	0.6		0.057		0.593	
G'_{ST}	1.0		0.095		0.988	
	(d) Subpopulation		(e) Subpopulation		(f) Subpopulation	
Allele	1	2	1	2	1	2
1	0.5	—	0.1	—	0.6	—
2	0.5	—	0.6	0.5	0.3	—
3	—	0.3	0.3	0.3	0.1	0.2
4	—	0.3	—	0.2	—	0.3
5	—	0.4	—	—	—	0.5
H_S	0.58		0.58		0.58	
$G_{ST(max)}$	0.266		0.266		0.266	
G_{ST}	0.266		0.025		0.256	
G'_{ST}	1.0		0.094		0.964	

subpopulations have nonoverlapping sets of alleles, or the maximum difference possible for the observed amount of genetic variation within subpopulations. If we let p_{ij} be the frequency of allele A_i in subpopulation j, the maximum total heterozygosity for a set of allele frequencies is then

$$H_{T(max)} = 1 - \sum_i \left(\frac{p_{i1}}{k}\right) - \sum_i \left(\frac{p_{i2}}{k}\right)^2 - \cdots - \sum_i \left(\frac{p_{ik}}{k}\right)^2$$
$$= 1 - \frac{1}{k^2} \sum_i \sum_j p_{ij}^2.$$

In other words, $H_{T(max)}$ is the maximum heterozygosity possible in the total population, given the observed heterozygosity within subpopulations. The maximum G_{ST} is then

$$G_{ST(Max)} = \frac{H_{T(max)} - H_S}{H_{T(max)}}.$$

We can write H_S as

$$H_S = \frac{1}{k}\left(1 - \sum_i p_{i1}^2 - \sum_i p_{i2}^2 - \cdots - \sum_i p_{ik}^2\right)$$
$$= 1 - \frac{1}{k} \sum_i \sum_j p_{ij}^2$$

therefore,

$$H_{T(max)} = (k - 1 + H_S)/k \quad \text{and}$$
$$G_{ST(max)} = \frac{(k-1)(1-H_S)}{k - 1 + H_S}. \tag{4a}$$

Note that as k becomes large, $G_{ST(max)}$ approaches $1 - H_S$ because $H_{T(max)}$ approaches 1.

At this limit, all populations have nonoverlapping sets of alleles and the genetic distance is maximized; for example, the standard genetic distance of Nei (1972) is ∞. Jin and Chakraborty (1995) also derived this expression as the asymptotic level of G_{ST} over k subpopulations when they are all descended from a common ancestral population and became completely isolated over time (see also Long and Kittles 2003).

We can then define the standardized G_{ST}, G'_{ST}, as

$$G'_{ST} = \frac{G_{ST}}{G_{ST(max)}} = \frac{G_{ST}(k - 1 + H_S)}{(k-1)(1-H_S)}. \tag{4b}$$

This is the proportion of $G_{ST(max)}$ that the observed G_{ST} represents, given the level of heterozygosity within the subpopulations. (Note that Nei (1987:191) used G'_{ST} differently to indicate a redefined G_{ST} value that is independent of k.) For two equal-sized populations, these expressions become

$$G_{ST(max)} = \frac{1 - H_S}{1 + H_S} \quad \text{and} \tag{5a}$$
$$G'_{ST} = \frac{G_{ST}(1 + H_S)}{(1 - H_S)}. \tag{5b}$$

Numerical Examples

Let us give some specific numerical examples to illustrate the calculation of $G_{ST(max)}$ and G'_{ST}. First, let us examine a situation where there is relatively low heterozygosity, $H_S = 0.25$ and $G_{ST(max)} = 0.6$. Assume that each of two populations has different alleles and they each contribute half to the total population (Table 1a). In this first example, $G_{ST} = 0.6$ so that $G'_{ST} = 1$. In other words, the level of differentiation is the maximum possible, given the genetic variation in the subpopulations, and is consistent with F_{ST} for two alleles, that is, when different alleles are fixed in different populations, $F_{ST} = 1$. Now let us assume that some alleles are shared between populations. In Table 1b, where there is relatively high overlap in allele frequencies, $G_{ST} = 0.057$ and $G'_{ST} =$

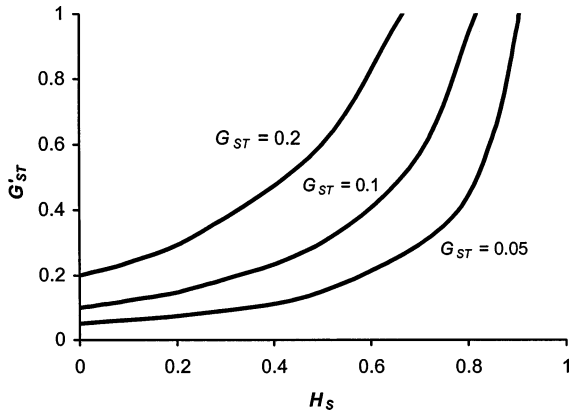

FIG. 1. The magnitude of the standardized genetic differentiation G'_{ST} for three levels of G_{ST}, as a function of the within population heterozygosity H_S when there are two subpopulations.

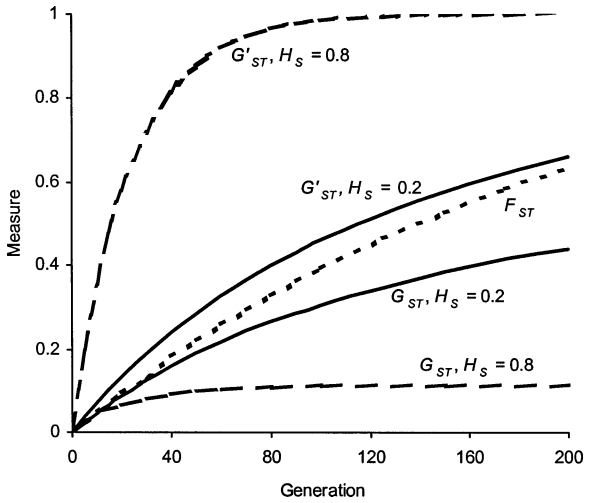

FIG. 2. The expected change over time in F_{ST} (short, broken line) and in G_{ST} and G'_{ST} for two levels of heterozygosity (H_S = 0.8, broken lines, and H_S = 0.2, solid lines) when there is complete isolation of the subpopulations and N_e = 100.

0.095, or 9.5% of the maximum possible. In Table 1c, where there is much less overlap in allele frequencies, $G_{ST} = 0.593$ and $G'_{ST} = 0.988$, or 98.8% of the maximum possible.

Now let us examine the situation where there is higher heterozygosity, $H_S = 0.58$ and $G_{ST(max)} = 0.266$. Again when there are no shared alleles, Table 1d, $G'_{ST} = 1$. When some alleles are shared, similar in magnitude to that in the example with a lower level of heterozygosity, then G'_{ST} values are very similar to those for the lower level of heterozygosity. However, notice that the levels of G_{ST} (d), (e), and (f) are only about 44% that in (a), (b), and (c).

Another way to see how the extent of differentiation is influenced by heterozygosity is to determine G'_{ST} for given a G_{ST} value over the range of possible heterozygosities using expression 5b (Fig. 1). For a range of lower H_S values, G'_{ST} remains not much larger than G_{ST} values. For example, for $G_{ST} = 0.05$, G'_{ST} is 0.12 for $H_S = 0.4$. As heterozygosity increases, $G_{ST(max)}$ decreases so the difference between G'_{ST} and G_{ST} increases rapidly. To determine the maximum heterozygosity for a given G_{ST} value, we can set G'_{ST} in expression (5b) to 1 and solve for H_S as

$$H_S = \frac{1 - G_{ST}}{1 + G_{ST}}.$$

Therefore, for G_{ST} values of 0.2, 0.1, and 0.05 to be obtainable, H_S must be less than 0.667, 0.818, and 0.905, respectively.

DISCUSSION

The standardized measure of genetic differentiation proposed here has properties that make it useful for evaluating the extent of divergence among subpopulations in a number of different situations. Below, in order to provide further context for this measure, I elaborate on two aspects of this measure. That is, I examine the expected change in this measure with time and the relative impact of gene flow versus mutation on genetic differentiation. In addition, I also discuss when application of this standardized measure is most appropriate and situations in which it may be useful.

Change in Genetic Differentiation over Time

Wright (1943) showed that when there is no gene flow between subpopulations, then

$$F_{ST(t)} = 1 - e^{-t/2N_e} \qquad 6a$$

when the subpopulations diverged from a common ancestral population t generations ago and N_e is the effective population size. Note that F_{ST} approaches unity because genetic drift within each subpopulation results in H_S approaching zero. Chakraborty and Jin (1992) showed that under the infinite allele model (IAM) for k subpopulations diverged from a common ancestral population t generations ago, and which have remained in isolation since,

$$G_{ST(t)} = \frac{\left(1 - \frac{1}{k}\right)(1 - H_S)[1 - e^{-tH_S/N_e(1-H_S)}]}{H_S + \left(1 - \frac{1}{k}\right)(1 - H_S)[1 - e^{-tH_S/N_e(1-H_S)}]}. \qquad (6b)$$

They assumed that each of the subpopulations remained at mutation-genetic drift balance so that H_S remained constant over time. As t increases, then the terms in brackets approach unity and the expression approaches $G_{ST(max)}$. Similarly, using the above expression for $G_{ST(t)}$

$$G'_{ST(t)} = \frac{G_{ST(t)}(1 + H_S)}{(1 - H_S)} \qquad (6c)$$

which, from a starting point of zero, approaches its maximum of unity. In other words, under this model, the value of $G'_{ST(t)}$ between 0–1 reflects the time since the subpopulations diverged.

To compare these predictions, Figure 2 gives the expected increase in $F_{ST(t)}$ over time (this change is independent of the heterozygosity) and the expected increase in $G_{ST(t)}$ and $G'_{ST(t)}$ for two different H_S levels (this assumes mutation-genetic drift equilibrium). Note that for the lower H_S, the

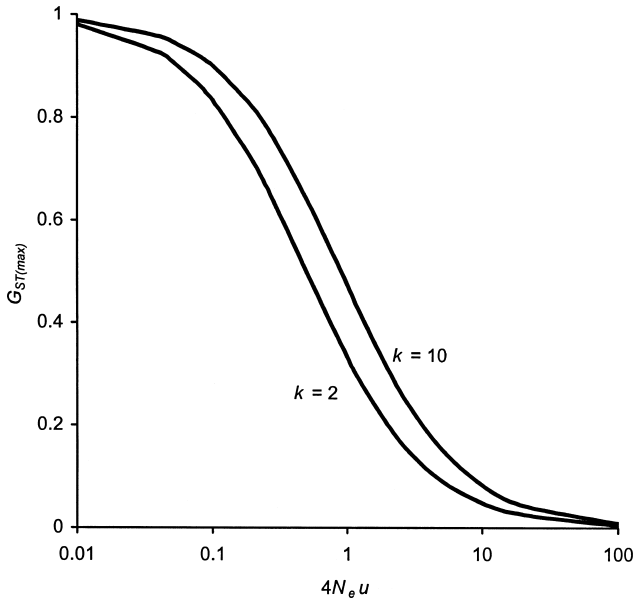

FIG. 3. The influence of the mutation rate, as indicted by $4N_e u$, on $G_{ST(max)}$ for two and ten subpopulations.

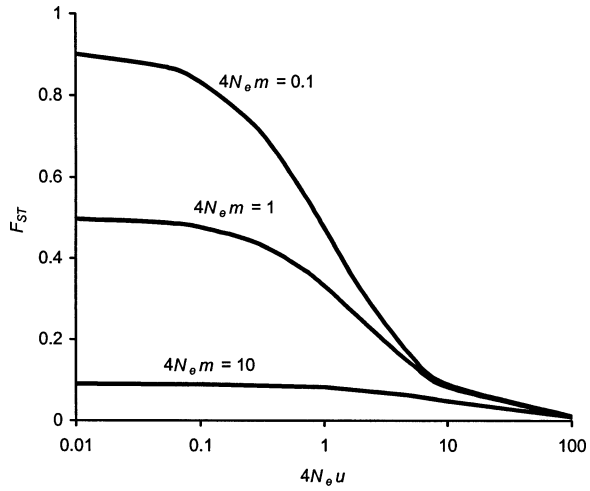

FIG. 4. The approximate value of F_{ST} in a finite population under the island model for different levels of gene flow ($4N_e m$) and mutation ($4N_e u$).

increase in G'_{ST} is only slightly larger than for F_{ST}. However, with a higher H_S, then G'_{ST} approaches the maximum of 1 faster. In other words, the rate of increase of F_{ST} is dependent only on genetic drift while the rate of increase of G'_{ST} depends both on genetic drift and mutation (resulting in a constant H_S over time). When two populations have recently diverged and a new small population size has a major impact of differentiation (such as in endangered species), then F_{ST} sometimes may be more useful than G'_{ST} because F_{ST} is independent of H_S (N. Ryman, pers. comm.). In other words, using F_{ST} loci with different levels of heterozygosity, such as allozymes and microsatellites, should be impacted similarly under this short-term scenario.

Impact of Gene Flow and Mutation on Genetic Differentiation

When the rate of gene flow m is higher than the mutation rate u, that is, $N_e m > N_e u$, then theory predicts that gene flow will dominate the level of genetic differentiation between groups (Balloux and Lugon-Moulin 2002). However, when the mutation rate is at the same level or higher than the level of gene flow, then mutation has a significant impact on genetic differentiation. For example, assume that there is an equilibrium level of heterozygosity within subpopulations for the IAM

$$H_S = \frac{4N_e u}{4N_e u + 1}.$$

Substituting this value of H_S in equation (4a), the maximum differentiation becomes

$$G_{ST(max)} = \frac{k - 1}{k(4N_e u + 1) - 1}. \quad (7)$$

To illustrate the maximum values of G_{ST} for different levels of mutation, Figure 3 plots $G_{ST(max)}$ for two and ten subpopulations. For high mutation rate loci, such as microsatellites, $4N_e u > 1$ and G_{ST} is relatively small. For low mutation rate loci, such as allozymes or SNPs, $4N_e u < 0.1$ and G_{ST} approaches unity. In other words, the size of G_{ST} for isolated populations is strongly influenced by the amount of variation (determined entirely by mutation here) at a locus. Of course, for all the values in Figure 3, $G'_{ST} = 1$.

Let us now examine the joint effects of gene flow and mutation in a finite population under the island model. The probability that alleles are identical by descent (e.g., see Hedrick 2005; p. 309) is modified by the probability that both alleles are not migrants $(1 - m)^2$ or mutants $(1 - u)^2$ so that

$$f_t = \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right) f_{t-1}\right](1 - m^2(1 - u)^2. \quad (8a)$$

Assuming equilibrium and that $f = F_{ST}$, then

$$F_{ST} = \frac{(1 - m)^2(1 - u)^2}{2N - (2N - 1)(1 - m)^2(1 - u)^2} \approx \frac{1}{4Nm + 4Nu + 1}. \quad (8b)$$

Using this approximation, we can get some guidance about when it is important to use G'_{ST}. In general, $4N_e u$ needs to be of the same magnitude or larger than $4N_e m$ to have an impact on F_{ST} (Fig. 4). For example, when $4N_e m = 1$, then for $4N_e u = 0$, $F_{ST} = 0.5$ and for $4N_e u = 1$, $F_{ST} = 0.333$. Or if we use $4N_e u$ as a general surrogate for H_S, then F_{ST} is generally independent of H_S when $4N_e u$ is less than $4N_e m$.

Researchers are often interested in what level of gene flow, that is, the number of migrants Nm that will result in a given level of genetic differentiation. When only gene flow influences differentiation, assuming $4Nu = 0$, equation 8b can be solved for an estimate of the number of migrants as

$$Nm = \frac{1 - F_{ST}}{4F_{ST}}. \quad (9a)$$

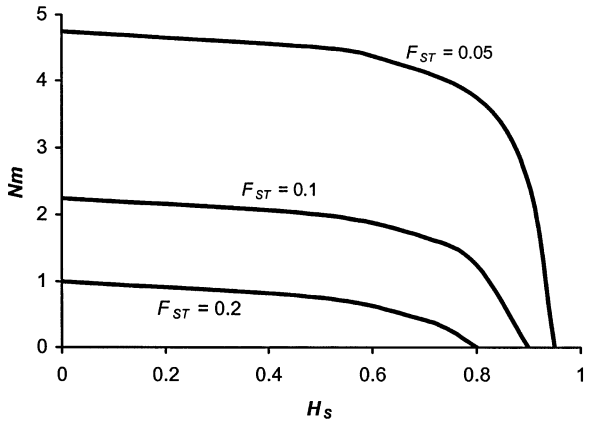

FIG. 5. The expected number of migrants *Nm* for different levels of F_{ST} and within population heterozygosity H_S.

If both gene flow and mutation occur, and assuming $4Nu = H_S/(1 - H_S)$, expression (8b) becomes

$$Nm = \frac{1 - F_{ST}[1 + H_S/(1 - H_S)]}{4F_{ST}}. \quad (9b)$$

In other words, the expected value of *Nm* is reduced as a function of the term in brackets in the above equation. To illustrate, Figure 5 shows the level of *Nm* for different F_{ST} values. Obviously only when H_S becomes large for a given F_{ST} value is the expected number of migrants greatly reduced.

Situations Where the Standardized Measure Is Useful

The standardized measure G'_{ST} should allow more appropriate comparisons between loci with different mutation rates, such as allozyme and microsatellite loci or mtDNA and Y-chromosome genes, and help resolve problems interpreting levels of genetic differentiation. A situation in which G'_{ST} would be particularly useful is in comparing differentiation of genes that may have different amounts of variation, such as mtDNA and Y chromosome genes. For example, Seielstad et al. (1998) found that F_{ST} was less for Y chromosome than for mtDNA genes in humans and concluded that this was evidence for a higher female than male gene flow rate. However, the Y chromosome markers that they used were much less variable than the mtDNA markers, apparently contributing to the F_{ST} patterns they observed. For example, when Wilder et al. (2004) used Y and mtDNA markers with similar levels of variation, they observed similar estimates of gene flow for females and males.

In addition, the standardized measure allows comparison of levels of genetic differentiation in different organisms that may have very different effective population sizes. A biological example of the application of G'_{ST} in this case is one discussed by Nei (1987, p. 191) where he compares levels of allozyme differentiation in the bacterium, *Escherichia coli*, and the kangaroo rat, *Dipodomys ordii*. *Escherichia coli* has an effective population size many orders of magnitude larger than that of kangaroo rats, with consequently a much larger level of genetic variation. Nei (1987) concluded that in *E. coli* ''the extent of gene differentiation is as great as that of kangaroo rats'' based on his measure of the absolute gene differentiation, $k(H_T - H_S)/(k - 1)$, which was 0.028 for both species. However, G'_{ST} values are 0.090 and 0.683 in *E. coli* and the kangaroo rats, respectively, clearly indicating much higher differentiation in the kangaroo rats than in *E. coli*.

Difficulty in the interpretation of statistics measuring genetic differentiation has long been recognized, even by Wright (1978, p. 82). With the standardized measure of genetic differentiation proposed here, the magnitude is the proportion of the maximum differentiation possible for the level of population homozygosity observed, a value that can be used universally to compare levels of genetic differentiation from many different organisms and loci.

ACKNOWLEDGMENTS

I thank L. Excoffier, O. Gaggiotti, C. Goodnight, A. Jones, S. Kalinowski, I. Olivieri, P. Pamilo, N. Ryman, M. Slatkin, R. Waples, and an anonymous reviewer for comments on an earlier version of the manuscript and the Ullman Professorship for support.

LITERATURE CITED

Balloux, F., and N. Lugon-Moulin. 2002. The estimation of population differentiation with microsatellite markers. Mol. Ecol. 11:155–168.

Chakraborty, R., and L. Jin. 1992. Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum. Genet. 88:267–272.

Charlesworth, B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. 15:538–543.

Charlesworth, B., D. Charlesworth, and N. H. Barton. 2003. The effects of genetic and geographic structure on neutral variation. Annu. Rev. Ecol. Evol. Syst. 34:99–125.

Excoffier, L. 2001. Analysis of population subdivision. Pp. 271–307 *in* D. J. Balding, M. Bishop, and C. Cannings, eds. Handbook of statistical genetics. John Wiley, New York.

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

———. 2005. Genetics of populations, 3rd ed. Jones and Bartlett, Boston, MA.

Jin, L., and R. Chakraborty. 1995. Population structure, stepwise mutation, heterozygote deficiency, and their implications in DNA forensics. Heredity 74:274–285.

Lewontin, R. C. 1964. The interaction of selection and linkage. 1. General considerations; heterotic models. Genetics 49:49–67.

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

Nagylaki, T. 1998. Fixation indices in subdivided populations. Genetics 148:1325–1332.

Nei, M. 1972. Genetic distance between populations. Am. Nat. 106: 283–292.

———. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

Seielstad, M. T., E. Minch, and L. L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. Nat. Genet. 20:278–280.

Wilder, J. A., S. B. Kingan, A. Mobasher, M. M. Pilington, and M. F. Hammer. 2004. Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat. Genet. 36:1122–1125.

Wright, S. 1943. Isolation by distance. Genetics 28:114–128.

———. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. 70:3321–3323.

———. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.
———. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19: 395–420.
———. 1978. Evolution and the genetics of populations: variability within and among natural populations. Univ. Chicago Press, Chicago, IL.

Corresponding Editor: C. Goodnight