AMERICAN JOURNAL OF HUMAN BIOLOGY 2:165–170 (1990)

Dispersal, Effective Population Size, and the Genetic Structure of the Contemporary United States

WALTER D. KOENIG
Hastings Reservation and Museum of Vertebrate Zoology, University of California, Carmel Valley, California 93924

ABSTRACT I estimate effective population size (N_e) and the inbreeding coefficient (F_{ST}) for contemporary United States using Wright's isolation by distance model (Wright: *Genetics* 28:114–138, 1943) and parent-offspring dispersal distances obtained from individuals surveyed as part of a study of modern dispersal patterns. N_e is estimated to be minimally 3.61×10^7 and more likely closer to 8.05×10^7; based on these values, F_{ST} is between 1.59×10^{-7} and 9.28×10^{-9}, depending on whether it is measured relative to the United States population or the world at large. Not all the assumptions of the isolation by distance model are met by modern populations, and thus the results must be interpreted with caution. They suggest, however, that both mobility within and immigration into contemporary United States are great enough to make the probability of inbreeding and random genetic drift negligible factors in producing future evolutionary change. In contrast, gene flow, acting as both a constraint against geographic differentiation within the United States and by introducing new genes via international immigration, is likely to be a dominant evolutionary force in this population.

Studies of contemporary non-industrialized human societies indicate that current within-population genetic diversity and future evolutionary change may be largely determined by random genetic drift (Cavalli-Sforza, 1969; Cavalli-Sforza and Bodmer, 1971). Although the apparent increase in mobility since the development of mechanized transport (Boyce et al., 1971) makes this less likely to be true in industrialized societies, this assumption has not been tested. Such a test can be made by calculating effective population size (N_e), proportional to the rate at which genetic drift leads to fixation of alleles and loss of genetic heterzygosity (Crow and Kimura, 1970), and the inbreeding coefficient (F_{ST}), measuring the expected degree of genetic differentiation among subdivisions of a population (Wright, 1969).

Here I calculate N_e and F_{ST} for the contemporary United States. From the several migration models which can be applied to human populations, I chose Wright's (1943) isolation by distance model. This model specifically focuses on the situation in which, because of limited dispersal, individuals within a more-or-less continuously distributed population are isolated by distance rather than by geographic barriers or gaps in the habitat. It is thus particularly appropriate to human populations in industrialized societies where the advent of mechanized transport has greatly reduced the formerly prominent role of geographic barriers in constraining dispersal patterns. Individual isolates are rarely identifiable, thereby making the application of alternatives such as island models and matrix methods unsuitable. A review of the application of these approaches and their derivatives to studies of the genetic structure of human population can be found in Jorde (1980).

Although the isolation by distance model is appropriate for some of the most salient features of modern industrialized societies, it entails at least two assumptions that are not likely to be met: 1) a uniform distribution of individuals throughout their range and 2) panmixia within neighborhoods. The sensitivity of the analysis to violation of these assumptions may be considerable, at least with respect to some local populations. Thus, the results obtained from these analyses must be viewed with caution.

Received August 17, 1989; accepted November 19, 1989.

METHODS

Neighborhood size, N, is defined as the population of a region in which the parents of individuals born near the center may be treated as if drawn at random. Following Wright (1946),

$$N = 4\pi\rho\sigma^2 \quad (1)$$

where ρ is population density and σ^2 is the mean-square dispersal distance. From N, effective population size, N_e, can be estimated as

$$N_e = N\,C_K\,C_{RS}\,C_{GT} \quad (2)$$

where C_K, C_{RS}, and C_{GT} are correction factors compensating for non-normality of the dispersal distribution, non-random variation in lifetime progeny production, and overlapping generations, respectively (Barrowclough and Coats, 1985). These correction factors are discussed in greater detail below.

Parent-offspring dispersal distance was estimated from individuals sampled from eight secondary school reunion booklets obtained as part of a larger study of migration patterns in contemporary United States (Koenig, 1988). Booklets were from eight schools located in Alabama, Ohio, Wisconsin, and California, and included both urban and rural areas. Schools were chosen whose reunions had occurred recently, therefore maximizing the probability that printed addresses would be current. In autumn 1984, a survey was sent to 1,095 individuals arbitrarily selected from the booklets; 607 (55%) replies were received. Individuals were asked their sex, birthplace, birthplace of their (first) spouse, and the birthplace of their first child, if any.

From these data, I determined the latitude and longitude for all localities and the great-circle distances between 1) birthplace of respondents and the birthplace of their first child and 2) birthplace of respondent's spouses and the birthplace of their first child. In order to confine the analyses to a geographically contiguous and homogeneous political unit, only individuals both born and whose first child was born within the continental United States or adjacent Canada (excluding Alaska and Hawaii) were included. Root-mean-square (RMS) distances were estimated from the formula

$$\mathrm{RMS} = \sqrt{[1/2n)\sum_{i=1}^{n} x_i^2]} \quad (3)$$

(Rockwell and Barrowclough, 1987), where x_i are parent-offspring distances and n is the sample size.

The degree of genetic differentiation expected among subdivisions is measured by the inbreeding coefficient F_{ST} (Wright, 1969). F_{ST} is hierarchical, measured relative to a specific larger population. At the level of the United States, F_{ST} can be estimated (Wright, 1951; Rockwell and Barrowclough, 1987) as

$$F_{ST} = (1 - Kt_k)/(1 + Kt_k), \quad (4)$$

where K is the number of demes in the species' range and

$$\begin{aligned} Kt_k = \exp - \{&(1/N_e) \times [\ln(K - 0.5) + 0.5772]\} \\ &+ (1/(2N_e^2)) \times [1.645 - 2/(2K - 1)] \\ &+ (1/(3N_e^3)) \times [1.202 - 2/(2K - 1)^2] + \cdots] \end{aligned} \quad (5)$$

F_{ST} can also be estimated at the global level from N_e and the rate of international immigration. Several models have been described, including the island model of Wright (1943), an elaboration of this model by Nei et al. (1977), and Malécot's two-dimensional continuous model (Cavalli-Sforza and Bodmer, 1971). All three of these yielded comparable results when applied to the data gathered here, and thus I present values based only on Wright's (1943) island model. From this model,

$$F_{ST} = 1/(1 + 4N_e m), \quad (6)$$

where m is the migration rate per generation.

RESULTS

Dispersal data for respondents alone and for the combined sample of respondents and their spouses, divided by sex, are presented in Table 1. As previously reported by Koenig (1989), values for females tend to be slightly greater than for males. The distributions of parent-offspring dispersal distances for respondents were not significantly different from those of respondent's spouses of the

TABLE 1. Estimation of parent-offspring dispersal distances in contemporary United States

	Mean (km)	Root-mean-square (km)	Standard deviation	Kurtosis	N
Respondents					
Males	678.2	896.6	1,074.2	2.330	189
Females	849.8	1060.2	1,237.9	0.758	234
Combined	773.2	990.5	1,169.4	1.344	423
Respondents and spouses					
Males	711.1	917.2	1,086.0	1.814	420
Females	760.4	983.8	1,166.5	1.349	410
Combined	734.1	950.7	1,125.1	1.581	832

same sex (Kolmogorov-Smirnov $Z = 0.78$ for males, 0.93 for females; both $P > 0.50$ [2-tailed]). Thus, for subsequent analyses I combined respondents and their spouses. Root-mean-square distances for the two sexes combined ranged from 532.8 km (Oshkosh, Wisconsin) to 1,291.9 km (San Jose, California) among the eight localities. Overall mean RMS distance was 950.7 km, several times greater than that reported for any other human population to date (Wijsman and Cavalli-Sforza, 1984; Koenig, 1988).

The 1980 population of the continental United States (excluding Alaska and Hawaii) was 2.250×10^8 individuals (US Dept. of Health, Education and Welfare, 1980); population density, given area of 7.984×10^6 km^2, was 28.2 km^{-2}. Substituting these values into Equation 1, $N = 4\pi(28.2)\ (950.7)^2 = 3.20 \times 10^8$ individuals for the complete sample; for the individual localities N ranged from 1.01×10^8 to 5.91×10^8 individuals. All but the lowest of these estimates is greater than the census size of the United States, indicating that there is little or no isolation by distance within this population and that parents of individuals born near the center of the country may be treated as if drawn at random. For subsequent analyses, I assume that the best estimate for neighbourhood size is the census size (2.25×10^8). However, as a conservative minimum I also calculated values based on the low estimate for $N(1.01 \times 10^8)$.

The three correction factors C_K, C_{RS}, and C_{GT} correct for demographic features altering N_e relative to N (Barrowclough and Coats, 1985). C_K depends on the precise parent-offspring dispersal distribution and is slightly >1 when the dispersal distribution is mildly leptokurtic and <1 when the distribution is either platykurtic or strongly leptokurtic (Wright, 1969). The parent-offspring distribution for the combined sample was slightly leptokurtic (Table 1) yielding $C_K = 1.013$.

C_{RS} depends on the variance in lifetime progeny production. If variance is less than binomial, $N_e > N$; otherwise, $N_e < N$. Assuming a stable population,

$$C_{RS} = \bar{k}/[\bar{k} - 1 + (V_k/\bar{k})] \qquad (7)$$

where $\bar{k}$ and V_k are the mean and variance of lifetime population progeny production (Crow and Kimura, 1970). For estimates of these parameters, I used the total number of live offspring born to women in the United States who were between 40 and 54 years of age in 1980, an age cutoff which includes >99% of the offspring born during the lifetime of this cohort (Demographic Yearbook, 1986). From these data, $\bar{k} = 3.00$ and $V_k = 4.09$. Substituting into equation (7), $C_{RS} = 0.892$. Unfortunately, comparable data are not available for men, who might be expected to have greater variance in lifetime progeny than would women.

C_{GT} adjusts for the effects of age structure and delayed breeding in populations with overlapping generations, factors which decrease N_e relative to N. I calculated C_{GT} by the algebraic method of Emigh and Pollak (1979). Using 1979 census data from the United States, $C_{GT} = 0.396$. From these estimates and Equation 2, $N_e = 8.05 \times 10^7$ when N equals the census size. Using the minimum estimate for N, $N_e = 3.61 \times 10^7$.

Using $N_e = 8.05 \times 10^7$ and Equation 4, $F_{ST} = 9.28 \times 10^{-9}$ relative to the 1980 United States potential for nearly three subdivisions of this size. According to the minimum estimate for N_e, $F_{ST} = 3.16 \times 10^{-8}$ relative to the potential for six such subdivi-

sions. These small values indicate that, given the current levels of mobility, virtually no stochastic genetic differentiation can be expected to arise within the United States.

At the global level, estimates of F_{ST} depend on m, the rate of international immigration. During the period 1951–1979, a 29-year time period roughly corresponding to generation length, there were approximately 9.80×10^6 immigrants to the United States (Houstoun et al., 1984), amounting to 4.35% of the 1980 population. Using this value for m and Equation 6, $F_{ST} = 7.12 \times 10^{-8}$ (1.59×10^{-7} using the minimum estimate for N_e).

DISCUSSION

Limitations and potential bias

These results are dependent on a variety of assumptions, some of which are violated to an unknown extent. First, they are potentially biased in that the survey was not a random sample of the study population. Potential sources of bias include 1) nonrandom geographic distribution of schools from which the samples were obtained, 2) nonrandom sampling of individuals as a consequence of choosing only those graduating or nearly graduating from secondary school, 3) bias among graduating individuals that were located at the time of the reunion and thus listed in the booklets, and 4) bias among those responding to the survey. Although the specific effects of these problems are unknown, they are unlikely to have inflated the observed mean dispersal distances (Koenig, 1988).

Second, as mentioned in the introduction, neither the assumptions of a uniform distribution or of panmixia at a local level are met by modern United States society, in violation of the isolation by distance model. Problems associated with the non-uniform distribution of the population are in most cases minimized by the fluidity of the population. For example, all samples analyzed here yielded large dispersal values, even though RMS dispersal distances varied slightly over two-fold in part because of the non-uniform distribution of individuals among localities. However, non-uniform distribution of individuals is in some cases extreme. Genetic structure of the resulting geographically isolated communities may at least in some cases be quite different than that described here for the population at large.

A similar, even more important difficulty is non-panmixia as a consequence of social, economic, and ethnic stratification. Moderate examples of such stratification include any number of ethnic groups in large urban areas between which intermarriage is relatively uncommon. Extreme examples include numerous subgroups existing in varying degrees of isolation from the main United States population such as rural Hispanics (Devor, 1980), the Ramah Navajo in New Mexico (Spuhler and Kluckhohn, 1953), Dunkers in Pennsylvania (Glass et al., 1952), and Hutterites in South Dakota and Minnesota (Mange, 1964). Because of their small census size and isolation, these communities have relatively high inbreeding coefficients; for example, Mange (1964) estimated that $F_{ST} = 0.022 \pm 0.014$ for the Hutterites.

As a consequence of these problems, the results obtained here must be interpreted with caution and cannot be considered representative of all local populations within the United States. However, if isolation is not complete and gene flow exceeds, on average, one individual per generation, substantial genetic differentiation due to drift will be prevented (Wright, 1951; Slatkin, 1987). It is likely that few minority groups, even among those attempting to retain a high degree of isolation, succeed in reducing gene flow below this level (Cavalli-Sforza and Bodmer, 1971). Thus, although important for the genetic structure of particular local subgroups, the overall effect of stratification on the measures of population structure calculated here for the United States as a whole is probably small.

An additional potential bias in the estimation of F_{ST} relative to the world at large is that immigrants are not chosen randomly from the common gene pool (that is, the world population). However, the results are so striking that correction for this bias would be unlikely to substantially alter the conclusions.

Consequences of large effective population size

The F_{ST} estimates obtained here are two orders of magnitude lower than those previously reported for any other human population (Jorde, 1980, 1984) and are much smaller than that necessary for mutation to maintain significant genetic variation at equilibrium (Lande and Barrowclough, 1987). For example, genetic heterozygosity

declines due to drift by a factor of $1 - (1/2N_e)$ per generation (Wright, 1931) and is therefore lost by a factor of only 0.999,999,994 per generation when $N_e = 8.05 \times 10^7$. Hence, at equilibrium, expected heterozygosity for selectively neutral polymorphisms is very high. Heterozygosity per locus $\overline{H} = 4N_e\mu/(1 + 4N_e\mu)$ where μ = the mutation rate (Crow and Kimura, 1970). At the relatively low μ of 10^{-7} per locus per generation (Dobzhansky, 1970), estimated $\overline{H}$ at equilibrium equals 0.9699 (0.9352 when $N_e = 3.61 \times 10^7$).

The estimated values for N_e and F_{ST} are sufficient to make it unlikely that selection can produce any appreciable geographic variation within the United States population. The distance over which adaptation to local conditions can occur, ℓ_c, is equal to $\sigma/\sqrt{s}$, where σ equals the root-mean-square dispersal distance and s is the selection coefficient against a homozygote in a 2-allele, additive model (Slatkin, 1973; May et al., 1975). No pocket of selective differences whose length is less than ℓ_c can produce a cline (Wijsman and Cavalli-Sforza, 1984). Given a relatively high selection coefficient of $s = 0.1$ and $\sigma = 950.7$ km, $\ell_c = 3{,}006$ km (1,685 km for the minimum estimate of σ). It is unlikely that there exists a selective factor that could maintain a cline of this magnitude for very long.

These results demonstrate that both mobility within and immigration into contemporary United States are great enough to make the probability of inbreeding and random genetic drift negligible factors in producing future evolutionary change. In contrast, gene flow, acting as both a constraint against geographic differentiation within the United States and by introducing new genes via international immigration, is likely to be a dominant evolutionary force in this population.

ACKNOWLEDGMENTS

I thank those who allowed me to use their high school reunion booklets, those who responded to the survey, and K. Heck for help in organizing the data. J.L. Dickinson, J.L. Patton, M. Ridley, and two reviewers provided valuable comments on the manuscript.

LITERATURE CITED

Barrowclough GF, and Coats SL (1985) The demography and population genetics of owls, with special reference to the conservation of the spotted owl (*Strix occidentalis*). In Gutiérrez RJ and Carey AB (eds.): Ecology and Management of the Spotted Owl in the Pacific Northwest. Portland OR: USDA, Forest Service Tech. Rep. PNW–185, pp. 74–85.

Boyce AJ, Küchemann CF, and Harrison GA (1971) Population structure and movement patterns. In Brass W (ed.): Biological Aspects of Demography. London: Taylor & Francis, pp. 1–9.

Cavalli-Sforza LL (1969) Genetic drift in an Italian population. Sci. Amer. *221*(2):30–37.

Cavalli-Sforza LL, and Bodmer WF (1971) The Genetics of Human Populations. San Francisco: W. H. Freeman.

Crow JF, and Kimura M (1970) An Introduction to Population Genetics Theory. New York: Harper & Row.

Demographic Yearbook (1986) New York: United Nations.

Devor EJ (1980) Marital structure and genetic isolation in a rural Hispanic population in northern New Mexico. Am. J. Phys. Anthropol. *53*:257–265.

Dobzhansky T (1970) Genetics of the Evolutionary Process. New York: Columbia University Press.

Emigh TH, and Pollak E (1979) Fixation probabilities and effective population numbers in diploid populations with overlapping generations. Theor. Pop. Biol. *15*:86–107.

Glass B, Sacks MS, Jahn EF, and Hess C (1952) Genetic drift in a religious isolate: an analysis of the causes of variation in blood group and other gene frequencies in a small population. Am. Nat. *86*:145–159.

Houstoun MF, Kramer RG, and Barret JM (1984) Female predominance in immigration to the United States since 1930: a first look. Inter. Migration Rev. *18*:908–963.

Jorde LB (1980) The genetic structure of subdivided human populations: a review. In Mielke JH and Crawford MH (eds.): Current Developments in Anthropological Genetics, vol. 1, Theory and Methods. New York: Plenum Press, pp. 135–208.

Jorde LB (1984) A comparison of parent-offspring and marital migration data as measures of gene flow. In Boyce AJ (ed.): Migration and Mobility. London: Taylor & Francis, pp. 83–96.

Koenig WD (1988) Internal migration in the contemporary United States: comparison of measures and partitioning of stages. Hum. Biol. *60*:927–944.

Koenig WD (1989) Sex-biased dispersal in the contemporary United States. Ethol. & Sociobiol. *10*:263–278.

Lande R, and Barrowclough GF (1987) Effective population size, genetic variation, and their use in population management. In Soulé ME (ed.): Viable Populations for Conservation. Cambridge: Cambridge University Press, pp. 87–123.

Mange AP (1964) Growth and inbreeding of a human isolate. Hum. Biol. *36*:104–133.

May R, Endler JA, and McMurtrie RE (1975) Gene frequency clines in the presence of selection opposed by gene flow. Am. Nat. *109*:659–675.

Nei M, Chakravarti A, and Tateno Y (1977) Mean and variance of F_{ST} in a finite number of incompletely isolated populations. Theor. Pop. Biol. *11*:291–306.

Rockwell RF, and Barrowclough GF (1987) Gene flow and the genetic structure of populations. In Cooke F and Buckley PA (eds.): Avian Genetics. London: Academic Press, pp. 223–255.

Slatkin M (1973) Gene flow and selection in a cline. Genetics *75*:733–756.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. Science *236*:787–792.

Spuhler JN, and Kluckhohn C (1953) Inbreeding coefficients of the Ramah Navaho population. Hum. Biol. *25*:295–317.

Wijsman EM, and Cavalli-Sforza LL (1984) Migration and genetic population structure with special reference to humans. Ann. Rev. Ecol. Syst. *15*:279–301.

Wright S (1931) Evolution in Mendelian populations. Genetics *16*:97–159.

Wright S (1943) Isolation by distance. Genetics *28*: 114–138.

Wright S (1946)Isolation by distance under diverse systems of mating. Genetics *31*:39–59.

Wright S (1951) The genetical structure of populations. Ann. Eugenics *15*:323–354.

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

US Dept Health, Education and Welfare (1980) Vital Statistics of the United States. Washington, DC: US Gov. Printing Office.