Am. J. Hum. Genet. 43:709–725, 1988

Population Amalgamation and Genetic Variation: Observations on Artificially Agglomerated Tribal Populations of Central and South America

Ranajit Chakraborty,* Peter E. Smouse,† and James V. Neel†

*Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston; and †Department of Human Genetics, University of Michigan Medical School, Ann Arbor

Summary

The interpretation of data on genetic variation with regard to the relative roles of different evolutionary factors that produce and maintain genetic variation depends critically on our assumptions concerning effective population size and the level of migration between neighboring populations. In humans, recent population growth and movements of specific ethnic groups across wide geographic areas mean that any theory based on assumptions of constant population size and absence of substructure is generally untenable. We examine the effects of population subdivision on the pattern of protein genetic variation in a total sample drawn from an artificial agglomerate of 12 tribal populations of Central and South America, analyzing the pooled sample as though it were a single population. Several striking findings emerge. (1) Mean heterozygosity is not sensitive to agglomeration, but the number of different alleles (allele count) is inflated, relative to neutral mutation/drift/equilibrium expectation. (2) The inflation is most serious for rare alleles, especially those which originally occurred as tribally restricted "private" polymorphisms. (3) The degree of inflation is an increasing function of both the number of populations encompassed by the sample and of the genetic divergence among them. (4) Treating an agglomerated population as though it were a panmictic unit of long standing can lead to serious biases in estimates of mutation rates, selection pressures, and effective population sizes. Current DNA studies indicate the presence of numerous genetic variants in human populations. The findings and conclusions of this paper are all fully applicable to the study of genetic variation at the DNA level as well.

Introduction

The pattern of genetic variation within a population is best described by the relative frequencies of different genetic variants at a series of loci unselected with respect to variation. The interpretation of such data with regard to the relative roles of the different evolutionary factors that produce and maintain genetic variation depends critically on our assumptions concerning effective population size and the level of migration between neighboring populations. The lack of historical demographic records for most organisms makes it difficult to validate any particular assumptions. In humans, where detailed demographic data are obtainable, commonly made assumptions are demonstrably false. Recent expansions of population size and movements of specific ethnic groups across wide geographic areas can often be documented historically or adduced from archaeological/linguistic studies. Other animal species are undoubtedly subject to similar perturbations. Despite these obvious complications, most manipulations of population data have been based on the assumption that the basic sampling unit is a single, panmictic breeding population whose effective size (N_e) is either known or can be estimated without error.

Our intent here is to examine the effects of population subdivision on the pattern of human genetic variation and its interpretation by creating an *artificial* popu-

Received January 13, 1988; revision received June 27, 1988.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225.

0002-9297/88/4305-0018$02.00

lation from a precisely specified agglomerate of 12 tribal populations of Central and South America. We will analyze the agglomerate as though it were a single population, as is the normal practice when analyzing large cosmopolitan populations. We will address four questions: (1) How is genetic variation in an agglomerated sample affected by the number of subpopulations encompassed by that sample? (2) How is the pattern of variation altered as the degree of genetic heterogeneity of those subpopulations increases? (3) Are the effects of population amalgamation spread across all allelic classes, or are the effects more pronounced for rare alleles? (4) What erroneous conclusions might result from treating an agglomerate population as a single, internally panmictic breeding unit of size N_e, however N_e is defined? We show that the amalgamation history of a population is a critical determinant of the pattern of genetic variation within that population, extending a very preliminary treatment of this point by Neel (1978).

Populations Sampled and Genetic Loci Examined

To identify the factors that influence the pattern of genetic variation within an agglomerate population, we have conducted an experiment involving the manipulation of data from our own studies of genetic variation in 12 Central and South American Indian tribes examined for electrophoretic variation in the protein products of the same 27 genetic loci. Considerable efforts have been expended to supplement the genetic data with the detailed ethnographic and demographic information necessary to establish these tribes as essentially discrete breeding units. During the 20-year period of these studies, an effort has been made to hold the basic sampling and assay techniques constant. These tribes were all originally selected for study because they were relatively undisturbed and unadmixed, but all have been influenced to one degree or another by post-Columbian developments. De facto evidence for restricted gene flow across tribal boundaries for a considerable period of time is provided by the observation of several tribally restricted "private" polymorphisms with allele frequencies as high as .10 that have *not* spread into nearby tribes (Neel 1980). We have employed one-dimensional protein electrophoresis to identify genetic variation in these studies. Although this technique only detects mutations involving charge and/or conformational changes in the coding region of DNA sequences, we will show later that the problems discussed in the present paper also affect the interpretation of population data on genetic variants detected by modern DNA techniques.

Tribal Groups

The 12 tribes we have used are the Ayoreo (AYO), Baniwa (BAN), Cayapo (CAY), Guaymi (GUA), Kraho (KRA), Machusi (MAC), Makiritare (MAK), Panoa (PAN), Piaroa (PIA), Ticuna (TIC), Wapashina (WAP), and Yanomama (YAN). The approximate map positions of the centroids of these tribal samples are shown in figure 1. Although some of these tribes (e.g., GUA, PAN, and YAN) have very wide geographic distributions and exhibit a well-defined infrastructure, the extent of subdivision within most of them is somewhat smaller than the extent of divergence among them (see Smouse 1982; Smouse et al. 1982). Where examples of recent intertribal gene flow have been documented (Chagnon et al. 1970; Neel et al. 1977*a*, 1977*b*; Long and Smouse 1983), samples from individuals of known mixed tribal origin have been deleted from consideration in order to minimize problems created by tribal admixture. These particular tribes also have relatively little admixture with non-Indians (Neel 1978); where such admixture can be identified, we have deleted the non-Indians and their children from the sample. This careful "genetic editing" results in a tabulation of relatively pure tribal allele counts not presented in any of the earlier accounts of these populations.

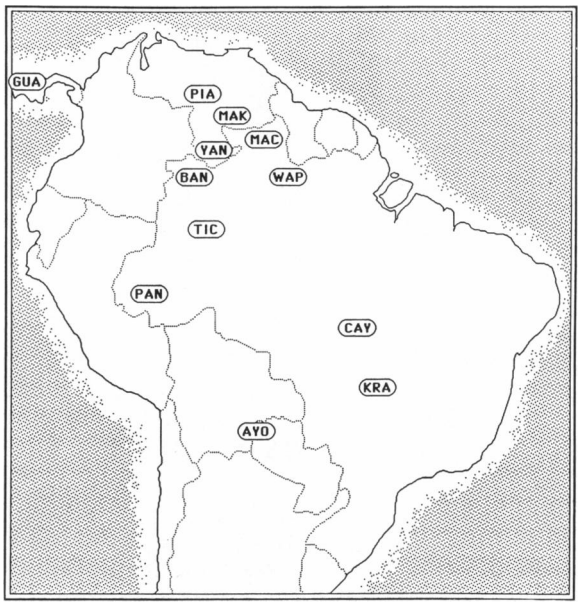

Figure 1. Map locations of the 12 Amerindian tribes.

Genetic Loci

The 27 electrophoretic loci included in this analysis are acid phosphatase-1 (ACP1), adenosine deaminase (ADA), adenalyte kinase-1 (AK1), albumin (ALB), carbonic anhydrase-1 (CA1), carbonic anhydrase-2 (CA2), ceruloplasmin (CRPL), esterase A (ESA), esterase D (ESD), galatose-1-phosphate uridyl transferase (GALT), phosphoglucomutase 1 and 2 (PGM1 and PGM2), haptoglobin (HP), hemoglobin-α (HBα), hemoglobin-β (HBβ), hemoglobin-δ (HBδ), isocitrate dehydrogenase (ICD), lactate dehydrogenase A and B (LDHA and LDHB), malate dehydrogenase (MDH), nucleoside phosphorylase (NP), peptidase A and B (PEPA and PEPB), 6-phosphogluconate dehydrogenase (6PGD), phosphohexose isomerase (PHI), transferrin (TF), and triosephosphate isomerase (TPI). The raw data (more than 575,000 allele counts) are deferred to Appendix A. The genetic data are summarized by Neel et al. (1977*a*, 1977*b*), Neel (1978), Salzano et al. (1978, 1984), and Mohrenweiser et al. (1979).

The Amount and Pattern of Genetic Variation in Tribal and Agglomerated Samples

To address the questions posed earlier, we utilize two different treatments: (a) an analysis of gene diversity and (b) an analysis of the numbers of alleles in various frequency classes. The two treatments yield different results, and the contrast is particularly revealing with respect to the impact of hidden agglomeration on the amount and pattern of genetic diversity.

Genic Diversity (Heterozygosity)

We will use heterozygosity (h) as a measure of allelic diversity. For our purposes, h is defined as the probability that two alleles drawn from a population are not identical, and it is computed as

$$h = 1 - \sum_{j=1}^{K} p_j^2 , \qquad (1)$$

where $p_1, p_2, \ldots, p_K$ represent the true allele frequencies of K alleles at a particular locus. When averaged over loci, this quantity reflects the per-locus genic diversity in the population. In a random mating population, h is mathematically equivalent to the proportion of heterozygotes, so that gene diversity is also a reflection of average heterozygosity per locus. For the present analysis we have used an unbiased estimate of h, specifically

$$\hat{h} = \frac{n}{n-1}\left[1 - \sum_{j=1}^{K} (n_j/n)^2\right], \qquad (2)$$

where $n_1, n_2, \ldots, n_K$ are the allele counts of K alleles at a locus in a sample of n genes drawn from the population (Nei and Roychoudhury 1974). Ideally, equation (2) should be used when the sampled genes are independent, i.e., when no related individuals have been included in the sample. This condition was not met in the studies from which the allele count data are obtained, because sets of relatives were included in the samples. Note, however, that even when individuals are related, $\hat{h}$ remains an *unbiased* estimate of h (Chakraborty 1978), although its nominal sampling precision is inflated. Since, in the present context, the sampling variance of $\hat{h}$ does not enter into our analysis (either explicitly or implicitly), the presence of related individuals in the samples does not affect our conclusions. Furthermore, we have no reason to believe that the degrees of relatedness among sampled individuals are different for different tribes, so the estimates should be consistent across tribes.

Under the assumptions of selective neutrality and mutation-drift equilibrium, the *expected* h in the population is given by

$$E(h) = M/(M + 1) , \qquad (3)$$

where $M = 4N_e v$, in which N_e is the effective population size and v is the mutation rate per locus per generation (Kimura and Crow 1964). We may estimate M from observed h by

$$\hat{M} = \hat{h}/(1 - \hat{h}) , \qquad (4)$$

where $\hat{h}$ is obtained from equation (2), averaged over all loci. It might be argued that this estimate is not a sufficient statistic for M, and hence not of maximum efficiency, but we will show later that $\hat{h}$ is robust in the face of agglomeration, while the more usual (and sufficient) statistic described below is profoundly sensitive to agglomeration. $\hat{M}$ is thus the estimate of choice in the face of possible agglomeration.

Numbers of Alleles

We present the allele frequency spectrum within each tribe and that for the total sample of 12 tribes in table 1; this summary was obtained by pooling data on all 27 loci surveyed. Note that even though these distributions could have been computed for each locus separately, single-locus analysis of the allele frequency spectrum is known to have a large stochastic error, in-

Table 1

Numbers of Alleles in Different Frequency Intervals for 12 Central and South American Indian Populations Assayed for 27 Protein Loci

	Allele Frequency Class (p_1, p_2)									
Tribe	<.005	.005–.01	.01–.05	.05–.10	.10–.30	.30–.70	.70–.90	.90–.95	>.95	Total
AYO	0	0	0	0	1	4	1	0	24	30
BAN	1	1	1	2	2	2	2	2	22	35
CAY	1	1	1	1	2	4	2	1	22	35
GUA	5	0	2	5	1	2	1	5	20	41
KRA	0	1	1	0	2	4	2	0	23	33
MAC	5	0	3	1	1	4	1	1	23	39
MAK	2	1	1	1	2	2	2	1	23	35
PAN	1	0	2	1	2	2	2	1	23	34
PIA	0	1	1	1	5	0	5	1	21	35
TIC	3	0	2	1	2	4	2	0	23	37
WAP	6	1	3	1	3	2	3	1	22	42
YAN	3	1	2	1	2	0	2	1	24	36
Pooled	27	4	3	1	2	2	2	1	23	65

terfering with any fine-tuned statistical analysis (Nei 1975; Ewens 1979). When data from 12 tribes are pooled, the total number of alleles per locus (65/27 = 2.407) becomes roughly one more than the average found within any one tribe (1.333).

In addition to the total number of alleles per locus, the allele frequency spectrum also provides estimates of the numbers of alleles in particular allele frequency classes. The additional alleles that result from pooling are not uniformly distributed over all allele frequency classes, as shown in table 1 and in figure 2, where panel (*a*) represents the allele frequency profile within a single tribe (averaged over all 12 tribes) and panel (*b*) represents the corresponding profile for the agglomerated sample for all 12 tribes. Note that except for the first allele frequency class (allele frequency <.005), these two histograms are almost identical, suggesting that the effect of agglomeration is almost entirely reflected in an apparent excess of rare alleles; the numbers of polymorphic alleles are not appreciably changed in an agglomerated sample. Although we will explicate these data further in a sequel, we note that this does not necessarily imply that the effects of agglomeration are simply described by the number of subpopulations hidden within an agglomerate. This is so because "private" alleles may have appreciable frequencies within a single tribe but be "rare" with respect to their regional occurrence. Since many of the private alleles are locally polymorphic, their combined effect on genetic dissimilarities among tribes is not negligible.

The expected number of different alleles in a sample is readily demonstrated to be an increasing function of sample size (Ewens 1972), and the results of table 1 and figure 2 are not unexpected. The question is whether the excess total number of alleles (or the excess rare alleles) can be accounted for entirely by the increase in sample size that accompanies pooling. There are two ways to examine this question.

The Ewens Expectation.—The first approach relies on the assumptions of selective neutrality and mutation/drift/equilibrium. It would be naive to assert that all of the variants under discussion in the present paper are strictly neutral, but the stochastic factor is so prominent in the evolution of small tribal groups that these variants can be treated as effectively neutral, in the sense that even modest selection pressures have negligible impact in small tribal gene pools (Neel and Thompson 1978; Thompson and Neel 1978). Given the assumption of selective neutrality, the expected total number of alleles that one should encounter in a sample of n genes is given by

$$E(k) = \sum_{m=0}^{n-1} \frac{M}{M + m}, \qquad (5)$$

where M is defined as before (Ewens 1972). Inserting the value of $\hat{M}$ into equation (5), we can obtain the expected number of alleles per locus for the specific sample sizes encountered in this survey. This provides a contrast of the observed number of alleles per locus

with that expected for given sample sizes for each tribe separately, as well as for the pooled sample.

We have used this theory to contrast, in figure 3, the observed total numbers of alleles with their expectations. The vertical bars for each data point in figure 3 represent a 2-standard-error interval for the total number of alleles per locus (computed from interlocus variations of this statistic). The statistical congruence of the observed and expected numbers of alleles within each individual tribe is obvious (allowing for variation in sample sizes), as in the *large excess* number of alleles in the agglomerated sample (the right-most data point of fig. 3).

The expected number of alleles in a specific allele frequency class has to be obtained somewhat differently. Chakraborty et al. (1980), Chakraborty (1981), and Chakraborty and Griffiths (1982) extended Ewens's sampling theory to obtain the expected number of alleles in the frequency interval (p_1, p_2) by the following expression:

$$k(p_1, p_2) = \sum_{m = [np_1]+1}^{[np_2]} \frac{M}{m} \cdot \frac{n!}{(n - m)!} \cdot \frac{\Gamma(n + M - m)}{\Gamma(n + M)}, \quad (6)$$

where Γ is a gamma function, $[np_1]$ is the largest integer contained in np_1, and $[np_2]$ is analogously defined.

The estimate of M given by equation (4) can be inserted into equation (6) to obtain the expected number of alleles in a specific frequency interval (p_1, p_2). The expected frequencies of variants occurring as singletons has an even simpler formula,

$$k(\text{singletons}) = Mn/(n + M - 1). \quad (7)$$

Chakraborty and Griffiths (1982) have shown that the numbers of rare alleles and singletons are Poisson distributed, and the variances of these statistics are the same as the expectations shown in equations (6) and (7).

We present the results of a formal analysis of the frequencies of rare alleles ($p_2 < .005$) and singletons from our tribal samples in table 2. The observed numbers of such alleles are counted from the data reported in table 1 and Appendix A, while the expectations are based on equations (6) and (7), respectively. Since the distributions of rare and singleton alleles are Poisson (Chakraborty and Griffiths 1982), the 95% confidence interval (CI) of such Poisson observations can be calculated for 27 trials (the number of independent loci), and these values are also shown in table 2. Note that for each of the 12 tribes the observed numbers of rare and singleton alleles are always within the respective

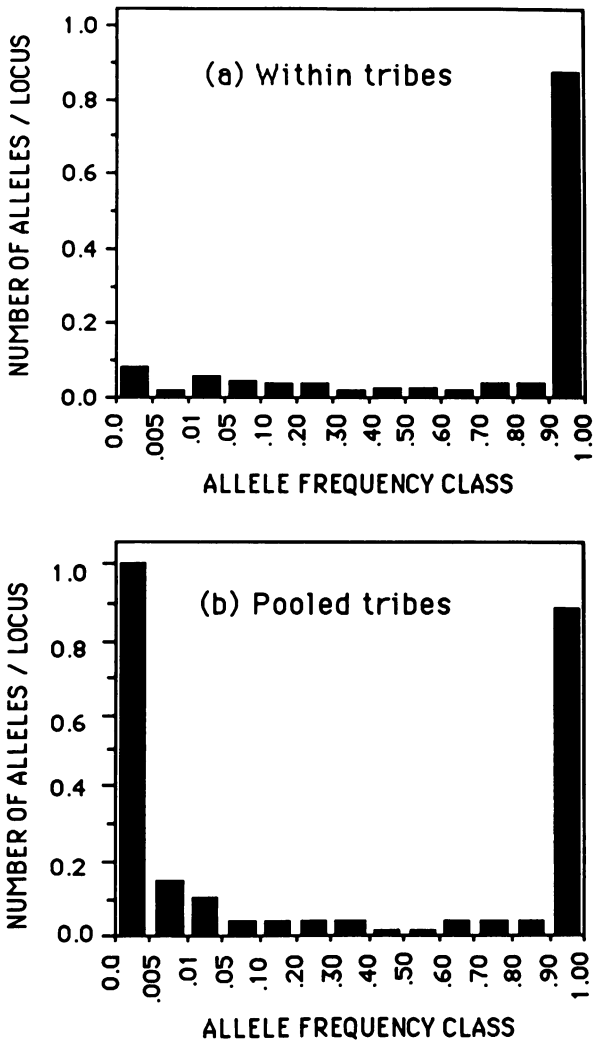

Figure 2. Observed allele frequency profile in 12 Amerindian populations. Panel (*a*) represents the allele frequency profile in an individual population, averaged over 12 tribes; panel (*b*) represents the allele frequency profile in the pooled sample of 12 tribes.

95% CIs, whereas for the agglomerated sample the observed numbers are in significant excess. It is particularly important to note that since both the expectations and confidence limits are based on the respective sample sizes, the excess in the total population *cannot* be ascribed to the larger sample size of the pooled data. A different visualization of the same point has been presented by Neel (1978). By using Poisson theory, we have avoided any difficulties that might arise from using our observed sample sizes (slightly inflated because of relatedness in the samples) in the determination of CIs.

Since the effect of agglomeration on the allele frequency profile is most evident among the rare alleles,

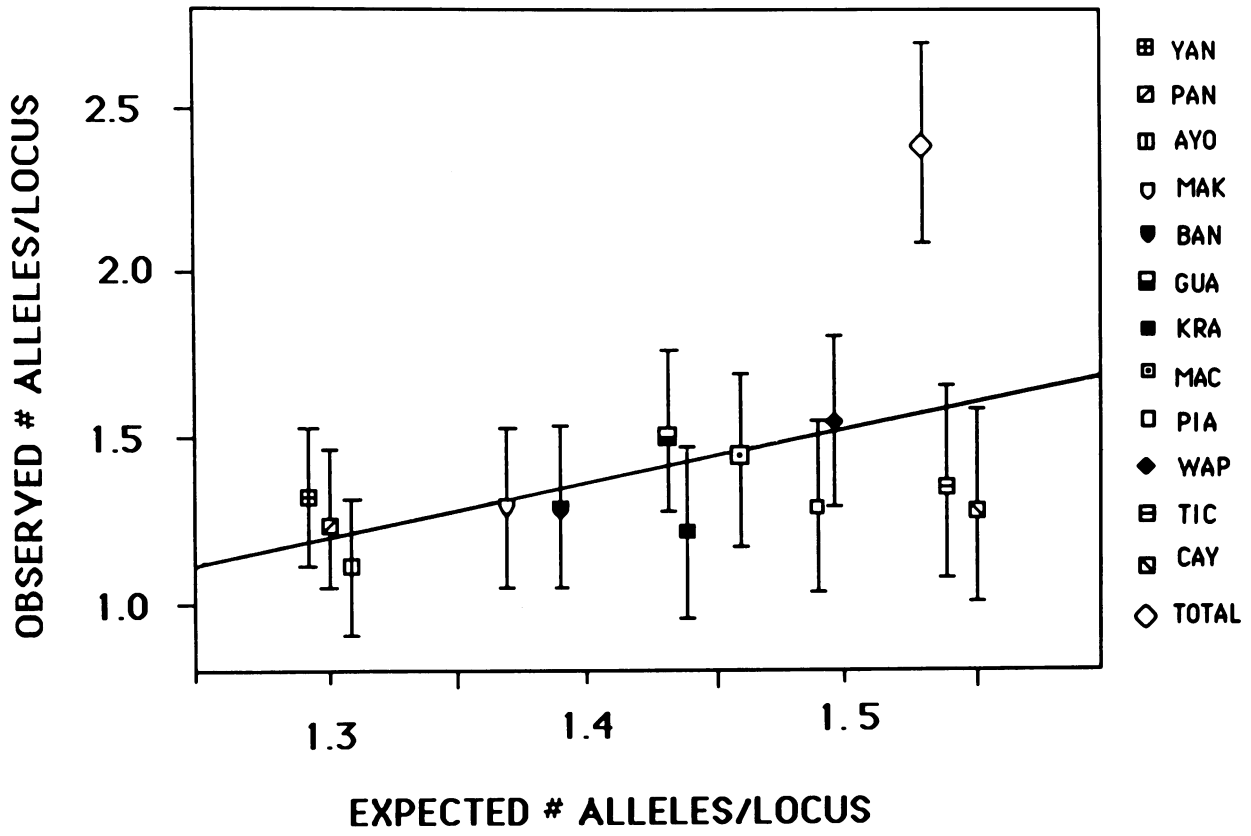

Figure 3. Observed vs. expected number of alleles per locus in 12 Central and South American Indian tribes and in the pooled sample. The expected numbers of alleles are computed using the equilibrium sampling theory of neutral alleles. The vertical bars represent two standard error intervals, while the solid dots represent the observed values.

genic diversity ($\hat{h}$), not being very sensitive to rare alleles, should not be much affected by agglomeration. In table 3 we show that this expectation is empirically borne out. This table also presents the average sample size per locus in these tribes, as well as the average allele counts in several low-frequency classes. We conclude from this table that if h is used as the summary measure of genic variation in the sample, amalgamation will usually go undetected. This, in essence, also justifies the use of equation (4) to estimate M, because the most efficient estimator (k) is exquisitely model dependent and would have given erroneous inference regarding M in the presence of agglomeration.

For purposes of analysis, we have assumed that each of the tribes has maintained a constant size during its evolution, so that $M = 4N_e v$ is estimated from their respective h values. Like the neutrality assumption, this assumption is not strictly true; modest departures from constancy, however, will not vitiate the results. We do not find any *significant* departures of the allele frequency profiles from expectations within tribes, arguing for virtual panmixia within tribes. Indeed, most of the observed numbers of alleles are smaller than expected (see table 2); however, the observed values for the MAC, WAP, YAN, and GUA are slightly higher than the expectations for them. There is evidence for some admixture between the MAC and WAP (Neel et al. 1977*a*, 1977*b*) and for tribal substructure within each of the YAN (Spielman et al. 1974) and GUA (Barrantes et al. 1982) samples. The genetic differences between the hidden components of these four groups are small enough, however, that the departures from expectation are all nonsignificant. We show later that the size of the departure increases with the genetic distance among the hidden components of an agglomerate. Thus, in spite of some genetic heterogeneity within our basic (tribal) sampling units, the equilibrium sampling theory of alleles may be reasonably adequate within single tribes. In order to avoid any potential difficulties from this source, however, we find it useful to introduce an alternative

Table 2

Observed and Expected Number of Singleton and Rare ($P < .005$) Alleles Found in 27 loci for 12 Amerindian Tribes and in The Total Sample

Tribe	No. of Singleton and Rare Alleles at 27 Loci					
	Singleton			Rare		
	Observed	Expected[a]	95% CI[b]	Observed	Expected[a]	95% CI[b]
AYO	1	1.30	(0, 3)	0	1.91	(0, 4)
BAN	1	1.50	(0, 4)	1	2.73	(0, 5)
CAY	0	2.01	(0, 5)	1	4.56	(1, 8)
GUA	3	1.52	(0, 4)	5	3.90	(1, 7)
KRA	0	1.85	(0, 4)	0	1.85	(0, 4)
MAC	0	1.70	(0, 4)	5	3.96	(1, 7)
MAK	0	1.36	(0, 3)	2	3.08	(1, 6)
PAN	1	1.15	(0, 3)	1	2.15	(0, 5)
PIA	0	2.14	(0, 5)	0	2.14	(0, 5)
TIC	1	1.70	(0, 4)	3	5.71	(2, 10)
WAP	2	1.84	(0, 4)	6	4.26	(1, 8)
YAN	2	.90	(0, 3)	3	3.33	(1, 7)
Pooled	8	1.55	(0, 4)	27	7.61	(3, 12)

[a] Computed by using eq. (7b) for rare alleles (with $p_2 = .005$) and eq. (9a) for singletons, employing the sample sizes presented in Appendix 1. The numbers are multiplied by 27, to reflect the expectations for 27 scored loci.

[b] Computed by using the Poisson expectation for the distribution of rare and singleton alleles (Chakraborty and Griffiths 1982).

approach which circumvents the panmictic assumptions.

Sample size adjustment.—The alternative approach is based solely on sampling considerations. We begin with the supposition that the observed allele frequencies for each tribe are consistent and unbiased estimates of their population (parametric) values. This claim is supported by the fact that for each of these tribes our samples represent an appreciable fraction of the total population (Neel 1973; Neel and Rothman 1978; Neel et al. 1986*a*). Assuming that the observed allele frequencies are unbiased estimates of their tribe-specific distributions, we can then ask, if a sample of alleles is drawn from these allele frequency distributions, how many al-

Table 3

Average Heterozygosity and the Number of Alleles/Locus in 12 Indian Tribes of Central and South America

Tribe	Average No. of Genes Sampled/Locus	Average h (%)	No. of Alleles/Locus		
			Total	Singles	Rare[a]
AYO	404	4.51	1.111	.000	.000
BAN	754	5.22	1.296	.037	.037
CAY	1,104	6.89	1.296	.000	.037
GUA	1,410	5.28	1.519	.111	.185
KRA	382	6.43	1.222	.000	.000
MAC	1,390	5.65	1.444	.074	.185
MAK	1,033	4.75	1.296	.000	.074
PAN	669	4.17	1.259	.037	.037
PIA	292	7.33	1.296	.000	.000
TIC	3,234	5.89	1.370	.037	.111
WAP	1,231	6.05	1.556	.074	.222
YAN	4,135	3.32	1.333	.074	.111
Pooled	16,036	5.37	2.407	.296	1.000

[a] $P < .005$.

leles should be observed in each allele frequency class? For a particular locus, let $p_1, \ldots, p_K$ represent the frequencies of K alleles in a population. The expected number of alleles in a random sample of size n is given by

$$E(k_n) = K - \sum_{j=1}^{K} (1 - p_j)^n \approx K - \sum_{j=1}^{K} \exp\{-np_j\}, \quad (8)$$

the proof for which is provided in Appendix B. This appendix also shows that when the observed allele frequencies are taken as unbiased estimates of their respective population frequencies, the number of alleles at a locus, and the allele counts in specific allele frequency classes, can be computed for any arbitrary sample size, thus avoiding the effect of sample size differences among the populations studied. The results of this computation are presented in table 4, using $n = 300$ for each tribe. Two sample size values have been chosen for the total population: $n = 3{,}600$, representing a partition of equal sample sizes from each of the 12 populations, and $n = 300$, to scale down the size of the sample from the agglomerated population to that of each tribe. The effect of sampling from an agglomerated population is found mainly in the rare alleles. Since the increase of allele numbers is seen for both agglomerated population sample sizes, we conclude that the excess rare alleles of tables 1–3 are *not* an artifact of larger sample size in the amalgamated sample. We also note that this result implies that the extensively sampled tribes (YAN and GUA) do not disproportionately affect our inferences.

The Magnitude of Heterogeneity

It is clear that sampling from an agglomerated population results in a deviation from the allele frequency spectrum predicted by traditional neutral theory. Our next step is to examine just how these deviations change with the number of populations and their degree of genetic divergence. In particular, we ask (as mentioned earlier) whether both of these factors are equally important. In view of our earlier result that the effect of agglomeration is most conspicuous in the rare allele class, one might suspect that the degree of genetic divergence among the constituent subpopulations hidden within an agglomerated sample may not be a relevant factor. We now demonstrate that this suspicion is *not* correct.

Such a demonstration requires an examination of the genetic diversity among the 12 tribes. Of the 27 loci examined here for electrophoretic variants, only five (ACP1, ESD, GALT, HP, and PGM1) are polymorphic (at least two alleles with $p > .01$). Fortunately, data are also available on polymorphisms at 10 additional loci whose products were studied (for the most part) by serological methods: MNSs, Rh, Duffy, Kidd, Diego, P.

Table 4

Expected Number of Alleles in Each Tribe and in the Pooled Sample When Sample Size is Reduced to 300 or 3,600

Tribe	Average No. of Genes Sampled/Locus	No. of Alleles/Locus		
		Total	Singles	Rare[a]
AYO	300	1.108	.000	.000
BAN	300	1.271	.036	.036
CAY	300	1.270	.000	.035
GUA	300	1.493	.106	.106
KRA	300	1.220	.000	.000
MAC	300	1.428	.069	.069
MAK	300	1.278	.000	.069
PAN	300	1.247	.035	.035
PIA	300	1.296	.000	.000
TIC	300	1.316	.033	.034
WAP	300	1.540	.069	.069
YAN	300	1.289	.060	.060
Pooled:				
300		2.368	.891	.891
3,600		2.371	.979	.979

[a] $P < .005$.

Lewis, Gm, Inv, and Gc. These additional genetic data can be found in the work of Salzano et al. (1972, 1977, 1986), Ward et al. (1975), Neel et al. (1977*a*), Gershowitz and Neel (1978), Mestriner et al. (1980), and Neel et al. (1980). We have not used such loci to study the allelic spectra of these tribes, because the standard serological approach seldom uncovers "novel" markers for these systems. Such data are, however, useful in taxonomic studies. A total of 195,688 allele product determinations have been performed for these 10 loci. Combining the two data sets, we have 15 polymorphisms for the computation of genetic distances.

We have used Nei's (1972) standardized genetic distance and the unweighted-pair-group mean algorithm (UPGMA) of Sneath and Sokal (1973) to construct a genetic dendrogram for the 12 tribes (fig. 4). This method produces a rooted dendrogram, and the lengths of the internodes and branchs reflect times of divergence for the various tribes, under the assumptions of selective neutrality and a constant rate of evolution. We resort to the use of a dendrogram at this juncture as a convenient way to define a set of agglomerated samples representing different degrees of genetic divergence, on the premise that the process of natural agglomeration involves increasingly unrelated populations as time goes on. Figure 4 shows the result of this analysis, which suggests that the 12 tribes have diverged in a hierarchical fashion, with the branching pattern generally following known linguistic affiliations. The linguistically anomalous grouping of the WAP (Arawak speakers) with the MAC (Carib speakers), rather than with the BAN (Arawak speakers), is probably due to recent admixture between the WAP and the MAC, alluded to earlier and described by Neel et al. (1977*a*, 1977*b*), but not subject to the precise documentation that would permit the removal of admixture from the sample. We have no ready explanation for the intrusion of the TIC (who speak an "isolated" language) into the Arawak-Carib grouping. Any minor differences between this dendrogramic depiction of the genetic relationships among the tribes and that previously published (Ward et al. 1975) are most probably the result of our choice of populations and loci; both factors are known to affect dendrogram topology and estimated branch lengths (Felsenstein 1973; Nei et al. 1983).

The present dendrogram, whether or not it represents absolute truth, will serve our objective of determining how the allele frequency spectrum in a sample depends on the degree of genetic divergence among subpopulations subsumed by that sample. The initial stages of agglomeration will usually involve closely related, neighboring groups; successive stages will involve progressively more distantly related groups. By agglomerating in dendrogram-compatible fashion, i.e., from right to left, we maintain some semblance of realism. All 12 tribes diverge from one another at node 11, and we can construct only ${}^{11}C_1 = 11$ possible pairs of tribes diverging at this level (AYO and any one of the other 11 tribes). There are only 55 trios compatible with this same node (${}^{11}C_2$ combinations of the 11 tribes, each amalgamated with AYO), 165 quartets, and so on. Similarly, when we consider node 4 (where the MAC, WAP, BAN, and TIC meet), we consider only three pairs of tribes (MAC-TIC, WAP-TIC, and BAN-TIC), since all other pairwise combinations of these four tribes will refer to a different level of genetic divergence (e.g., the BAN-MAC combination refers to the evolutionary distance across node 3).

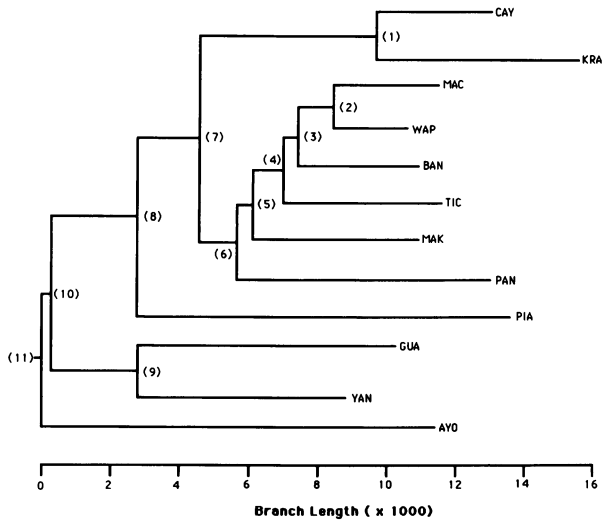

Figure 4. Evolutionary relationships among 12 Central and South American Indian tribes, based on data from 39 blood group and protein loci. Nei's (1972) standardized genetic distance is used, along with the UPGMA network. The branch length estimates are computed by the minimum-branch-length method. The numbers in parentheses are the internal nodes used in fig. 5.

The results of these evaluations are shown in figure 5. Panel (*a*) is based on the total number of alleles per locus, panel (*b*) on the number of rare alleles ($p < .005$). The sample size for the agglomerated sample is taken as 300 for each computation. It is clear that the excess number of alleles in an amalgamated sample is an increasing function of both the number of tribes encompassed in the amalgamated sample and the average degree of genetic diversity among them. The function approaches an asymptote with respect to both variables,

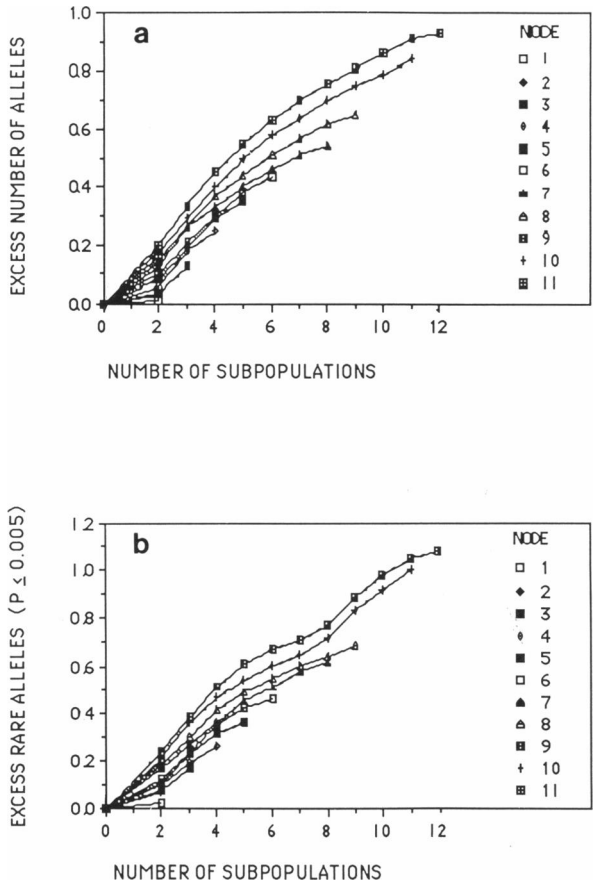

Figure 5. Number of alleles in an amalgamated sample, as a function of number of tribes encompassed in the sample and level of genetic divergence among them: panel (*a*) shows the total number of alleles, and panel (*b*) shows the rare alleles ($P < .005$); the numbers indicate the dendrogram node employed in fig. 4.

but the approach is slower with respect to the number of tribes. This conclusion is not dependent on the order of agglomeration; any other sequence would yield comparable results.

Discussion and Conclusion

The artificial amalgamation experiment conducted here indicates that even if each sampling unit exhibits an allele frequency spectrum in *apparent* agreement with expectations of the mutation/drift/equilibrium theory, the allele frequency spectrum of the agglomerated sample can deviate substantially from theoretical expectation. The fact that the deviation is most conspicuous for rare alleles has a number of interrelated implications.

First, on the basis of Ewens's (1972) neutral sampling theory, several authors have argued that the total number of alleles (k) is a sufficient statistic for the parameter $M = 4N_e v$ under the assumption of neutral mutation-drift equilibrium (Ewens 1974, 1979; Ewens and Gillespie 1974; Watterson 1978). Nei (1977) has argued, however, that selection will drive disadvantageous alleles toward extinction, thus reducing the *total* number of segregating alleles, relative to neutral expectation. On the other hand, positive and negative selection have nearly counterbalancing effects on the number of *rare* alleles (Nei 1977; Kimura 1983). Nei (1977) proposed that the number of rare alleles in a sample yields a better indirect estimate of the mutation rate than is the total number. Neel and Rothman (1978) reached the same conclusion, using a somewhat different formulation.

While that strategy is attractive for a single random mating population, it is not tenable for agglomerated populations, because the rare allele count is substantially inflated in such populations. Any indirect estimation of mutation rates profits from large sample sizes, leading to the temptation to pool data from several different populations. This will inevitably lead to upwardly biased mutation-rate estimates. Even when carefully defined tribal populations are used to estimate mutation rates, there may be problems. We have commented elsewhere (Spielman et al. 1974) on the amount of subdivision within tribes. Our own indirect mutation-rate estimates based on tribal totals (Neel et al. 1986*a*) yield higher estimates than the direct approach based on the same markers (Neel et al. 1986*b*). To some unknown extent, this may result from the ineluctable subdivision within larger tribes.

Second, criteria need to be established for the choice and/or the pooling of populations for analytic purposes. For instance, Chakraborty et al. (1980) analyzed samples from 138 different species for departures from the predictions of the neutral mutation model. A frequent observation was an apparent excess of rare alleles. To enlarge the sample sizes for that analysis, however, populations that showed Nei's (1972) genetic distances of 0.05 or less were pooled. In view of our results here, that was probably too generous a pooling criterion. The maximum genetic distance between any two tribes in our amalgamation experiment is only 0.034 (between AYO and PIA); yet, the effect of amalgamation on the increase of rare allele frequencies is quite dramatic. We conclude that pooling over subpopulations exhibiting even much smaller genetic distances would also produce a nontrivial deviation of the allele frequency spectrum from neutral expectation. Interpretation of the evolutionary causes of that variation may therefore be

problematic. In particular, we question the appropriateness of including data on civilized human populations for inferences regarding either past bottlenecks (although these may well have occurred) or the persistence of slightly deleterious mutations (see Ohta 1976; Nei 1987).

Third, the use of frequencies of rare alleles to estimate the effective migration rate between populations must be approached with caution. Slatkin (1985) has proposed a computer simulation method to estimate the average number of migrants exchanged between populations (N_em) from the spatial distribution of rare alleles. Slatkin and Takahata (1985) have provided analytical justification for these predictions. Singh and Rhomberg (1987) have recently applied Slatkin's method to a study of dispersal in *Drosophila melanogaster*. In applying such theory, it will be important to establish that each locality is unitary (e.g., see Strobeck 1987), so that the local frequency of private alleles is not affected by undetected amalgamation of the form described here.

Finally, recognition of recent population agglomeration could be important for the interpretation of human DNA polymorphism. DNA data collected from ethnically heterogeneous populations have frequently been pooled in order to increase sample sizes (e.g., see Chakravarti et al. 1984, 1986; Murray et al. 1984). We suspect that the sampling frame may make a major contribution to the apparently large number of low-frequency variants, since that number is an increasing function of the genetic heterogeneity encompassed by the sample, which is substantial in most of these studies. As large-scale surveys of less internally heterogeneous populations are undertaken, we can anticipate a (perhaps considerable) reduction in the numbers of rare variants recovered/1,000 individuals screened, although it must be remembered that even single cosmopolitan populations are regional agglomerates. We anticipate DNA results that are entirely compatible with those for the protein studies already done (e.g., see Harris et al. 1973; Neel et al. 1978, 1988), given a comparable sampling frame. Moreover, and less comforting, we may be confident that the picture of DNA polymorphism that emerges from agglomerated samples—and all cosmopolitan populations are agglomerates—will provide as poor a basis for rigorous inference about the causes and meaning of the variants as have the available protein data from agglomerated population samples.

Acknowledgments

This work was supported by grants GM-R01–22394 (R. C.), GM-R01–32589 (P.E.S.), and DE-FGO2–87ER 60533 (J.V.N.). We thank J. Towrek and R. Bush for computational and graphics help, respectively, and both of them for extensive comments on the manuscript. We also thank two anonymous reviewers for helpful critiques.

Appendix A

Allele Count Data for 12 Amerindian Populations from Central and South America for 27 Genetic Loci

	Population											
Locus/Allele	AYO	BAN	CAY	GUA	KRA	MAC	MAK	PAN	PIA	TIC	WAP	YAN
						A. ALB						
Normal	716	754	1,870	1,434	386	1,388	1,417	670	290	1,522	1,222	6,482
MAK-1	0	0	0	0	0	0	13	0	2	0	0	0
MAK-3	0	0	0	0	0	0	2	0	0	0	0	0
YAN-2	0	0	0	0	0	0	0	0	0	0	0	521
MAKU	0	0	0	0	0	0	0	0	0	0	24	4
YAN-1	0	0	0	0	0	0	0	0	0	0	0	1
						B. CRPL						
B	516	754	1,269	1,434	384	1,384	1,290	670	290	1,516	1,232	6,815
A CAY-1	0	0	67	0	0	1	0	0	0	0	4	0
A MAC-1	0	0	0	0	0	3	0	0	0	0	0	0
C WAP-1	0	0	0	0	0	0	0	0	0	0	4	0
C WAP-2	0	0	0	0	0	0	0	0	0	0	2	0
A YAN-1	0	0	0	0	0	0	0	0	0	0	0	1

Appendix A (continued)

Locus/Allele	Population											
	AYO	BAN	CAY	GUA	KRA	MAC	MAK	PAN	PIA	TIC	WAP	YAN
C. HP												
HPT-1	321	391	887	774	268	724	613	417	208	2,337	557	5,308
HPT-2	405	363	629	656	116	529	797	213	82	1,193	637	1,108
D. TF												
C	518	742	1,450	1,357	384	1,386	1,436	670	265	3,487	1,246	6,744
D CHI	0	12	0	74	0	0	0	0	27	43	0	0
D GUA-1	0	0	0	1	0	0	0	0	0	0	0	0
E. ACP1												
A	73	55	209	80	111	39	77	41	67	220	75	86
B	169	699	671	1,292	271	1,328	1,349	629	225	2,916	1,155	6,516
TIC-1	0	0	0	0	0	0	0	0	0	390	0	0
C	0	0	0	4	0	1	0	0	0	0	0	0
GUA	0	0	0	42	0	0	0	0	0	0	0	0
F. ADA												
ADA-1	362	754	934	1,373	382	1,372	654	670	292	3,524	1,224	3,326
ADA-2	0	0	0	1	0	0	0	0	0	2	0	0
G. AK1												
AK-1	364	754	1,172	1,404	382	1,366	1,316	670	292	3,523	1,230	5,212
AK-2	0	0	0	0	0	0	0	0	0	1	0	0
H. CA2												
1	282	713	786	1,398	380	1,484	780	669	292	2,523	1,225	612
2	0	0	0	2	0	0	0	1	0	3	3	0
BAN-1	0	41	0	0	0	0	0	0	0	0	0	0
I. ESA												
A	282	752	788	1,415	380	1,428	776	670	292	2,526	1,199	746
D MAC-1	0	0	0	0	0	56	0	0	0	0	29	0
D GUA-1	0	0	0	1	0	0	0	0	0	0	0	0
J. ESD												
ESD-1	282	592	392	1,370	165	1,012	622	558	222	1,712	978	479
ESD-2	0	158	394	52	215	472	158	112	70	874	248	133
K. GALT												
1	230	749	781	1,244	382	1,251	749	662	254	3,368	1,102	1,927
D	0	5	11	156	0	117	27	8	38	154	127	19
WAP-1	0	0	0	0	0	0	0	0	0	0	1	0
L. PGM1												
1	302	621	905	1,327	296	1,136	1,203	590	215	2,925	949	6,380
2	68	133	287	77	86	242	229	80	77	605	281	304
10 MAC-1	0	0	0	0	0	14	0	0	0	0	0	0

Appendix A (continued)

Locus/Allele	Population											
	AYO	BAN	CAY	GUA	KRA	MAC	MAK	PAN	PIA	TIC	WAP	YAN
M. ICD												
Normal	370	754	1,042	1,416	382	1,366	806	670	284	3,530	1,230	3,982
PIA-1	0	0	0	0	0	0	0	0	8	0	0	0
N. LDHB												
Normal	370	754	934	1,319	382	1,368	1,014	670	292	3,530	1,230	4,084
GUA-1	0	0	0	103	0	0	0	0	0	0	0	0
O. MDH												
1	372	754	936	1,416	382	1,366	1,014	670	292	3,530	1,230	4,006
2 MAC-1	0	0	0	0	0	2	0	0	0	0	0	0
P. PEPA												
1	276	754	1,042	1,406	374	1,482	814	670	292	3,530	1,208	3,898
WAP-1	0	0	0	0	0	0	0	0	0	0	20	0
KRA-1	0	0	0	0	8	0	0	0	0	0	0	0
Q. PEPB												
1	372	753	1,054	1,416	382	1,486	818	654	292	3,530	1,230	4,010
BAN-1	0	1	0	0	0	0	0	0	0	0	0	0
PAN-1	0	0	0	0	0	0	0	16	0	0	0	0
R. 6PGD												
A.	356	754	1,169	1,283	382	1,389	1,425	670	292	3,528	1,218	6,416
C	0	0	3	115	0	3	5	0	0	0	12	0
S. PHI												
1	372	754	1,041	1,414	379	1,368	818	670	292	3,530	1,229	4,010
2 CAY-1	0	0	9	0	3	0	0	0	0	0	0	0
3 WAP-1	0	0	0	0	0	0	0	0	0	0	1	0
T. CA1												
Normal	282	754	788	1,404	380	1,366	780	670	292	2,526	1,228	748
U. PGM2												
Normal	366	754	1,200	1,390	382	1,394	1,432	670	292	3,530	1,230	6,688
V. HBα												
Normal	684	754	1,788	1,414	382	1,370	960	670	292	3,530	1,230	5,206
W. HBβ												
Normal	684	754	1,788	1,414	382	1,370	960	670	292	3,530	1,230	5,206
X. HBδ												
Normal	684	754	1,054	1,402	382	1,368	960	670	292	3,530	1,230	5,206

Appendix A (continued)

Locus/Allele	Population											
	AYO	BAN	CAY	GUA	KRA	MAC	MAK	PAN	PIA	TIC	WAP	YAN
						Y. LDHA						
Normal	370	754	934	1,422	382	1,368	1,014	670	292	3,530	1,230	4,084
						Z. NP						
Normal	228	754	738	1,374	382	1,364	780	670	292	3,528	1,228	684
						AA. TPI						
Normal	228	754	790	1,400	380	1,368	780	670	292	3,530	1,230	686

Appendix B

Sampling Theory for Agglomerated Populations

I. The Number of Different Alleles Recovered from a Sample of n Genes Collected from an Agglomerate of r Populations

Let p_{ij} be the frequency of the jth allele in the ith population, $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, K$. The total sample size has a typical partition $(n_1, n_2, \ldots, n_r)$, where n_i is the sample size from the ith population, such that $0 < n_i < n$ and $\Sigma_i n_i = n$. For a given partition, let us define (for the jth allele)

$$X_j = \begin{cases} 1 \text{ if the } j\text{th allele is included in the total sample} \\ 0 \text{ otherwise} \end{cases}$$

with $Y = \Sigma_j X_j$ defined as the total number of alleles in the pooled sample. The probability that the jth allele is absent from the total sample is

$$\begin{aligned} \Pr(X_j = 0) &= \text{Prob } (j\text{th allele is absent from all } r \text{ population samples}) \\ &= \prod_{i=1}^{r} (1 - p_{ij})^{n_i}, \end{aligned} \tag{A1}$$

since the populations are independently sampled. It follows that

$$\Pr(X_j = 1) = 1 - \prod_{i=1}^{r} (1 - p_{ij})^{n_i}. \tag{A2}$$

The expected number of alleles for the partition $(n_1, n_2, \ldots, n_r)$ is given by

$$\begin{aligned} E(Y \mid n_1, \ldots, n_r) &= \sum_{j=1}^{K} E(X_j) \\ &= \sum_{j=1}^{K} \left[1 - \prod_{i=1}^{r} (1 - p_{ij})^{n_i} \right] \\ &= K - \sum_{j=1}^{K} \left[\prod_{i=1}^{r} (1 - p_{ij})^{n_i} \right]. \end{aligned} \tag{A3}$$

Since $0 < p_{ij} < 1$, and n_i's are generally large, we may use the approximation

$$(1 - p_{ij})^{n_i} \approx \exp\{-n_i p_{ij}\}, \tag{A4}$$

so that the expected number of alleles in the total sample of n genes with the observed partition of sample is given by

$$E(Y \mid n_1, \ldots, n_r) \approx K - \sum_{j=1}^{K} \exp\{-n\bar{p}_j\}, \tag{A5a}$$

where $\bar{p}_j$ is the average frequency of the jth allele in the total sample.

Note that the above derivation is based on the assumption that the partition of n, giving the sample sizes drawn from each subpopulation $(n_1, n_2, n_3, \ldots, n_r)$ is fixed. If $w_1, w_2, \ldots, w_r$ represent the relative sizes of the r populations, with $0 < w_i < 1$, and if $\Sigma_i w_i = 1$, the probability of observing the partition itself is given by the multinomial probability function

$$\Pr(n_1, \ldots, n_r \mid n) = \frac{n!}{\prod_{i=1}^{r} n_i!} \prod_{i=1}^{r} w_i^{n_i}, \tag{A6}$$

and hence the expected number of alleles in the total sample of size n genes, permuted over all possible partitions of n, is given by

$$E(Y) = \sum_{(n1, \ldots, nr)} \frac{n! \prod_{i=1}^{r} w_i^{n_i} \left[\sum_{j=1}^{r} \{1 - \prod_{i=1}^{r}(1 - p_{ij})^{n_i}\} \right]}{\prod_{i=1}^{r} n_i!}$$

$$= \sum_{j=1}^{K} \left[\sum_{(n1, \ldots, nr)} \frac{n! \prod_{i=1}^{r} w_i^{n_i}}{\prod_{i=1}^{r} n_i!} \right]$$

$$- \sum_{j=1}^{K} \left[\sum_{(n1, \ldots, nr)} \frac{\left(n! \prod_{i=1}^{r} [w_i(1 - p_{ij})]^{n_i} \right)}{\prod_{i=1}^{r} n_i!} \right]$$

$$= K - \sum_{j=1}^{K} \left[1 - \sum_{i=1}^{r} w_i (1 - p_{ij}) \right]^n$$

$$= K - \sum_{j=1}^{K} (1 - \bar{p}_j)^n \approx K - \sum_{j=1}^{K} \exp \{-n\bar{p}_j\} , \quad \text{(A5b)}$$

the same as in equation (A5a).

Hence, as long as the approximation (A4) is valid (which is true whenever each population is represented by a sample size larger than, say, 100), the specific partition of n into $n_1, n_2, \ldots, n_r$ is enough to compute the expected number of alleles in the sample, and we need not be concerned with all possible partitions of the total sample size in a conglomerate sample.

The variance of Y (the number of alleles in the pooled sample) can be written as

$$\text{Var } (Y) = \text{Var} \left[\sum_{j=1}^{K} X_j \right] = \sum_{j=1}^{K} \text{Var } [X_j] + \sum_{j \neq j'=1}^{K} \sum \text{Cov } [X_j, X_{j'}] . \quad \text{(A7)}$$

Since the X_j's are Bernoulli variates,

$$\text{Var } [X_j] = \prod_{i=1}^{r} (1 - p_{ij})^{n_i} \left[1 - \prod_{i=1}^{r} (1 - p_{ij})^{n_i} \right] , \quad \text{(A8)}$$

and, furthermore, noting that

$$\Pr (X_j = 0, X_{j'} = 0) = \prod_{i=1}^{r} (1 - p_{ij} - p_{ij'})^{n_i} , \quad \text{(A9)}$$

we have

$$\text{Cov } [X_j, X_{j'}] = \prod_{i=1}^{r} (1 - p_{ij} - p_{ij'})^{n_i} - \prod_{i=1}^{r} [(1 - p_{ij}) (1 - p_{ij'})]^{n_i} . \quad \text{(A10)}$$

Inserting equations (A8) and (A10) into equation (A7), and using the approximation (A4), we then have

$$\text{Var } (Y) = \sum_{j=1}^{K} \exp \{- n\bar{p}_j\} [1 - exp \{- n\bar{p}_j\})]. \quad \text{(A11)}$$

2. *Number of Alleles of a Specific Allele Frequency Class in a Sample of n Genes Chosen from an Ensemble of r Populations*

This technique can be used to compute the number of alleles in a specific frequency class, say (p_1, p_2). Let us redefine the indicator variables, X_j, as

$$X_j = \begin{cases} 1 \text{ if the } j\text{th allele has } \ell \text{ copies in the sample of } n \text{ genes.} \\ 0 \text{ otherwise.} \end{cases} \quad \text{(A12)}$$

Let $(\ell_1, \ell_2, \ldots, \ell_r)$ be a partition of ℓ, such that there are ℓ_i copies of the jth allele from the ith population in the total sample, with $0 < \ell_i < \ell$, and $\sum \ell_i = \ell$. We can thus write

$$\Pr (X_j = 1) = \Sigma_\ell \left[\frac{\ell!}{\prod_{i=1}^{r} \ell_i!} \prod_{i=1}^{r} \binom{n_i}{\ell_i} p_{ij}^{\ell_i} (1 - p_{ij})^{n_i - \ell_i} \right] , \quad \text{(A13)}$$

where Σ_ℓ is the summation over all partitions of ℓ. The expected number of alleles in the allele frequency class (p_1, p_2) is

$$E[k(p_1, p_2)] = \sum_{\ell = [np_1]+1}^{[np_2]} \sum_{j=1}^{K} \Pr (X_j = 1) , \quad \text{(A14)}$$

where $[n]$ represents the largest integer contained in n. There are only r possible partitions of ℓ for singleton alleles ($\ell = 1$). Hence

$$\Pr(X_j = 1) = \sum_{i=1}^{r} n_i p_{ij} (1 - p_{ij})^{n_i - 1} \cdot \prod_{i \neq i' = 1}^{r} (1 - p_{i'j})^{n_{i'}}$$

$$= \prod_{i=1}^{r} (1 - p_{ij})^{n_i} \cdot \sum_{i=1}^{r} \frac{n_i p_{ij}}{1 - p_{ij}} . \quad \text{(A15)}$$

Therefore, the expected number of singleton alleles in the total sample of n genes can be computed as

$$\text{E(singletons)}$$

$$= \sum_{j=1}^{K} \left[\prod_{i=1}^{r} (1 - p_{ij})^{n_i} \left(\sum_{i=1}^{r} \frac{n_i p_{ij}}{(1 - p_{ij})} \right) \right] . \quad \text{(A16)}$$

References

Barrantes, R., P. E. Smouse, J. V. Neel, H. W. Mohrenweiser, and H. Gershowitz. 1982. Migration and genetic infrastructure of the Central American Guaymi and their affinities with other tribal groups. Am. J. Phys. Anthropol. **58:** 201–214.

Chagnon, N. A., J. V. Neel, L. R. Weitkamp, H. Gershowitz, and M. Ayres. 1970. The influence of cultural factors on the demography and pattern of gene flow from the Makiritare to the Yanomama Indians. Am. J. Phys. Anthropol. **32:**339–349.

Chakraborty, R. 1978. Number of independent genes examined in family surveys and its effect on gene frequency estimation. Am. J. Hum. Genet. **30:**550–552.

———. 1981. Expected number of rare alleles per locus in a sample and estimation of mutation rates. Am. J. Hum. Genet. **33:**481–483.

Chakraborty, R., P. A. Fuerst, and M. Nei. 1980. Statistical studies on protein polymorphism in natural populations. III. Distribution of allele frequencies and the number of alleles per locus. Genetics **94:**1039–1063.

Chakraborty, R., and R. C. Griffiths. 1982. Correlation of heterozygosity and the number of alleles in different frequency classes. Theor. Popul. Biol. **21:**205–218.

Chakravarti, A., K. H. Beutow, S. E. Antonarakis, P. G. Waber, C. D. Boehm, and H. H. Kazazian. 1984. Nonuniform recombination within the human β-globin gene cluster. Am. J. Hum. Genet. **36:**1239–1258.

Chakravarti, A., S. C. Elbein, and M. A. Permutt. 1986. Evidence for increased recombination near the human insulin gene: implications for disease association studies. Proc. Natl. Acad. Sci. USA **83:**1045–1049.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:**87–112.

———. 1974. A note on the sampling theory for infinite alleles and infinite sites models. Theor. Popul. Biol. **6:** 143–148.

———. 1979. Mathematical population genetics. Springer, Berlin.

Ewens, W. J., and J. H. Gillespie. 1974. Some simulation results for the neutral allele model, with interpretations. Theor. Popul. Biol. **6:**35–57.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. **22:**240–249.

Gershowitz, H., and J. V. Neel. 1978. The immunoglobulin allotypes (Gm and Km) of twelve Indian tribes of Central and South America. Am. J. Phys. Anthropol. **49:**289–302.

Harris, H., D. A. Hopkinson, and E. B. Robson. 1973. The incidence of rare alleles determining electrophoretic variants: data on 43 enzyme loci in man. Ann. Hum. Genet. **37:**237–253.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics **49:**725–738.

Long, J. C., and P. E. Smouse. 1983. Intertribal gene flow between the Ye'cuana and Yanomama: genetic analysis of an admixed village. Am. J. Phys. Anthropol. **61:**411–422.

Mestriner, M. A., A. L. Simões, and F. M. Salzano. 1980. New studies on the esterase D polymorphism in South American Indians. Am. J. Phys. Anthropol. **52:**95–101.

Mohrenweiser, H., J. V. Neel, M. A. Mestriner, F. M. Salzano, E. Migliazza, A. L. Simões, and C. M. Yoshihara. 1979. Electrophoretic variants in three Amerindian tribes: the Baniwa, Kanamari, and Central Pano of western Brazil. Am. J. Phys. Anthropol. **50:**237–246.

Murray, J. C., K. A. Mills, C. M. Demopulos, S. Hornung, and A. G. Motulsky. 1984. Linkage disequilibrium and evolutionary relationships of DNA variants (restriction enzyme fragment length polymorphisms) at the serum albumin locus. Proc. Natl. Acad. Sci. USA **81:**3486–3490.

Neel, J. V. 1973. "Private" genetic variants and the frequency of mutation among South American Indians. Proc. Natl. Acad. Sci. USA **70:**3311–3315.

———. 1978. Rare variants, private polymorphisms, and locus heterozygosity in Amerindian populations. Am. J. Hum. Genet. **30:**465–490.

———. 1980. Isolates and private polymorphisms. Pp. 173–193 *in* A. Ericksson, ed. Population structure and genetic disorders. Academic Press, London.

Neel, J. V., H. Gershowitz, H. W. Mohrenweiser, B. Amos, D. D. Kostyu, F. M. Salzano, M. A. Mestriner, D. Lawrence, A. L. Simões, P. E. Smouse, W. J. Oliver, R. S. Spielman, and J. V. Neel, Jr. 1980. Genetic studies on the Ticuna, an enigmatic tribe of Central Amazonas. Ann. Hum. Genet. **44:**37–54.

Neel, J. V., H. Gershowitz, R. S. Spielman, E. C. Migliazza, F. M. Salzano, and W. J. Oliver. 1977*a*. Genetic studies of the Macushi and Wapishana Indians. II. Data on 12 genetic polymorphisms of the red cell and serum proteins: gene flow between the tribes. Hum. Genet. **37**:207–219.

Neel, J. V., H. W. Mohrenweiser, E. D. Rothman, and J. M. Naidu. 1986*a*. A revised indirect estimate of mutation rates in Amerindians. Am. J. Hum. Genet. **38**:649–666.

Neel, J. V., and E. D. Rothman. 1978. Indirect estimates of mutation rates in tribal Amerindians. Proc. Natl. Acad. Sci. USA **75**:5585–5588.

Neel, J. V., C. Satoh, K. Gorike, M. Fujita, N. Takahashi, J. Asakawa, and R. Hazama. 1986*b*. The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc. Natl. Acad. Sci. USA **83**:389–393.

Neel, J. V., C. Satoh, P. E. Smouse, J. Asakawa, N. Takahashi, K. Goriki, M. Fujita, T. Kageoka, and R. Hazama. 1988. Protein variants in Hiroshima and Nagasaki: tales of two cities. Am. J. Hum. Genet. (submitted).

Neel, J. V., R. J. Tanis, E. C. Migliazza, R. S. Spielman, F. Salzano, W. J. Oliver, M. Morrow, and S. Bachofer. 1977*b*. Genetic studies of the Macushi and Wapishana Indians. I. Rare genetic variants and a "private polymorphism" of esterase A. Hum. Genet. **36**:81–107.

Neel, J. V., and E. A. Thompson. 1978. Founder effect and the number of private polymorphisms observed in Amerindian tribes. Proc. Natl. Acad. Sci. USA **75**:1904–1908.

Neel, J. V., N. Veda, C. Satoh, R. E. Ferrell, R. J. Tanis, and H. B. Hamilton. 1978. The frequency in Japanese of genetic variants of 22 proteins. V. Summary and comparison with data on Caucasians from the British Isles. Ann. Hum. Genet. **41**:429–441.

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

———. 1975. Molecular population genetics and evolution. North Holland–American Elsevier, New York.

———. 1977. Estimation of mutation rate from rare protein variants. Am. J. Hum. Genet. **29**:225–232.

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

Nei, M., and A. K. Roychoudhury. 1974. Sampling variance of heterozygosity and genetic distance. Genetics **76**: 379–390.

Nei, M., F. Tajima, and Y. Tateno. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. **19**:153–170.

Ohta, T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphisms. Theor. Popul. Biol. **10**:254–275.

Salzano, F. M., H. Gershowitz, P. C. Junqueira, J. P. Wodall, F. L. Black, and W. Hierholzer. 1972. Blood groups and H-Le^a salivary secretion of Brazilian Cayapo Indians. Am. J. Phys. Anthropol. **36**:417–426.

Salzano, F. M., H. Gershowitz, H. Mohrenweiser, J. V. Neel, P. E. Smouse, M. A. Mestriner, T. A. Weimer, M. H. L. P. Franco, A. L. Simões, J. Constans, A. E. Oliveira, and M. J. de Freitas e Melo. 1986. Gene flow across tribal barriers and its effect among the Amazonian Icana River Indians. Am. J. Phys. Anthropol. **69**:3–14.

Salzano, F. M., H. Mohrenweiser, H. Gershowitz, J. V. Neel, M. A. Mestriner, A. L. Simões, J. Constans, and M. J. de Melo e Freitas. 1984. New studies on the Macushi Indians of northern Brazil. Ann. Hum. Biol. **11**:337–350.

Salzano, F. M., J. V. Neel, H. Gershowitz, and E. C. Migliazza. 1977. Intra and intertribal genetic variation within a linguistic group: the Ge-speaking Indians of Brazil. Am. J. Phys. Anthropol. **47**:337–348.

Salzano, F. M., F. Pages, J. V. Neel, H. Gershowitz, R. J. Tanis, R. Moreno, and M. H. L. P. Franco. 1978. Unusual blood genetic characteristics among the Ayoreo Indians of Bolivia and Paraguay. Hum. Biol. **50**:121–136.

Singh, R. S., and L. R. Rhomberg. 1987. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. I. Estimates of gene flow from rare alleles. Genetics **115**:313–322.

Slatkin, M. 1985. Rare alleles as indicators of gene flow. Evolution **39**:53–65.

Slatkin, M., and N. Takahata. 1985. The average frequency of private alleles in a partially isolated population. Theor. Popul. Biol. **28**:314–331.

Smouse, P. E. 1982. Genetic architecture of swidden agricultural tribes from the lowland rain forests of South America. Pp. 139–178 *in* M. Crawford and J. Mielke, eds. Current developments in anthropological genetics. Vol. **2**: Ecology and population structure. Plenum. New York.

Smouse, P. E., R. S. Spielman, and M. H. Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am. Nat. **110**:445–463.

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman, San Francisco.

Spielman, R. S., E. C. Migliazza, and J. V. Neel. 1974. Regional linguistic and genetic differences among Yanomama Indians. Science **184**:637–644.

Strobeck, C. 1987. Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117**:149–153.

Thompson, E. A., and J. V. Neel. 1978. The probability of founder effect in a tribal population. Proc. Natl. Acad. Sci. USA **75**:1442–1445.

Ward, R. H., H. Gershowitz, M. Layrisse, and J. V. Neel. 1975. The genetic structure of a tribal population, the Yanomama Indians. XI. Gene frequencies for 10 blood groups and the ABH-Se traits in the Yanomama and their neighbors; the uniqueness of the tribe. Am. J. Hum. Genet. **27**:1–30.

Watterson, G. A. 1978. An analysis of multi-allelic data. Genetics **88**:171–179.