Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3321–3323, December 1973

Analysis of Gene Diversity in Subdivided Populations

(population structure/ genetic variability/heterozygosity/gene differentiation)

MASATOSHI NEI

Center for Demographic and Population Genetics, University of Texas at Houston, Tex. 77025

Communicated by Sewall Wright, August 6, 1973

ABSTRACT **A method is presented by which the gene diversity (heterozygosity) of a subdivided population can be analyzed into its components, i.e., the gene diversities within and between subpopulations. This method is applicable to any population without regard to the number of alleles per locus, the pattern of evolutionary forces such as mutation, selection, and migration, and the reproductive method of the organism used. Measures of the absolute and relative magnitudes of gene differentiation among subpopulations are also proposed.**

In a genetic study of substructured populations, Wright (1–3) showed that the variation in gene frequency among subpopulations may be analyzed by the fixation indices or F-statistics. He derived the formula

$$1 - F_{IT} = (1 - F_{IS})\,(1 - F_{ST}), \qquad [1]$$

where F_{IT} and F_{IS} are the correlations between two uniting gametes to produce the individuals relative to the total population and relative to the subpopulations, respectively, while F_{ST} is the correlation between two gametes drawn at random from each subpopulation. F_{IT} and F_{IS} may become negative, but F_{ST} is nonnegative. The degree of gene differentiation among subpopulations may be measured by F_{ST}.

The F-statistics are applicable to any population if there are only two alleles at a locus. In the presence of multiple alleles, however, Eq. **1** no longer holds except for the special case of random differentiation with no selection (4). Recently, I (5, 6) proposed a new method of measuring the degree of gene differentiation between *a pair of populations*. This method is based on the identities of two randomly chosen genes within and between populations and independent of the number of alleles. In the following, I shall extend this method to the case of hierarchical structure of populations and show that the gene-frequency variation in a substructured population can be analyzed directly in terms of heterozygosity or of gene diversity, which will be defined later. This method can be applied to any population without regard to the number of alleles at a locus or to the pattern of evolutionary forces such as mutation, selection, and migration. It is also applicable to any organism, whether this is sexually or asexually reproducing or whether this is diploid or nondiploid, as far as gene frequencies can be determined. Such a method seems to be necessary to analyze rapidly increasing data on gene frequencies for protein loci.

Suppose that there are n alleles at a locus and the frequency of the kth allele is x_k in a population. The probabilities of identity and nonidentity of two randomly chosen genes are then given by $J = \sum_k x^2_k$ and $H = 1 - J$, respectively. The probability of nonidentity, H, is a measure of genic variation of a population and usually called *heterozygosity*. This word, however, is not appropriate for a nonrandom mating population. Therefore, I use the word *gene diversity* for this quantity. I also use the abbreviated word *gene identity* for J. Of course, if one is interested only in random mating populations, the words gene diversity and gene identity in the following may be replaced by heterozygosity and homozygosity, respectively.

Let us now consider a population that is subdivided into s subpopulations. Let x_{ik} be the frequency of the kth allele in the ith subpopulation. The gene identity in this subpopulation is given by

$$J_i = \sum_k x^2_{ik}, \qquad [2]$$

while the gene identity in the total population is

$$J_T = \sum_k x^2_{\cdot k}, \qquad [3]$$

where $x_{\cdot k} = \sum_i w_i x_{ik}$, in which w_i is the weight for the ith subpopulation ($\sum w_i = 1$). The quantity J_T may be written as

$$J_T = \sum_k \left(\sum_i w_i x_{ik}\right)^2$$

$$= \sum_k \left(\sum_i w^2_i x^2_{ik} + \sum_{i \neq j} w_i w_j x_{ik} x_{jk}\right).$$

If $w_i = 1/s$, then

$$J_T = \left(\sum_i \sum_k x^2_{ik} + \sum_{i \neq j} \sum_k x_{ik} x_{jk}\right)/s^2$$

$$= \left(\sum_i J_i + \sum_{i \neq j} J_{ij}\right)/s^2, \qquad [4]$$

where

$$J_{ij} = \sum_k x_{ik} x_{jk} \qquad [5]$$

is the gene identity between the ith and jth subpopulations.

Let us now define the gene diversity between the ith and jth populations as

$$D_{ij} = H_{ij} - (H_i + H_j)/2 = (J_i + J_j)/2 - J_{ij}, \quad [6]$$

where $H_i = 1 - J_i$ and $H_{ij} = 1 - J_{ij}$. I (6, 7) have called this parameter the minimum number of net codon differences per locus, but in the present context the word gene diversity seems to be better. Note that D_{ij} is $\sum_k (x_{ik} - x_{jk})^2/2$, so that it is nonnegative. If we use Eq. **6**, Eq. **4** reduces to

$$J_T = \{\sum_i J_i + \sum_{i \neq j} (J_i + J_j)/2 - \sum_{i \neq j} D_{ij}\}/s^2 = \{s\sum_i J_i - \sum_i \sum_j D_{ij}\}/s^2,$$

since $D_{ii} = 0$. Therefore,

$$J_T = (\sum J_i)/s - (\sum_i \sum_j D_{ij})/s^2 = J_S - D_{ST}, \quad [7]$$

where J_S is the average gene identity within subpopulations, and D_{ST} is the average gene diversity between subpopulations, including the comparisons of subpopulations with themselves. The gene diversity in the total population ($H_T = 1 - J_T$) is

$$H_T = H_S + D_{ST}, \quad [8]$$

where $H_S = 1 - J_S$. Thus, the gene diversity in the total population can be analyzed into the gene diversities within and between subpopulations.

The absolute magnitude of gene differentiation among subpopulations may be measured by D_{ST} or $\bar{D}_m$ given later, while the gene differentiation relative to the total population is given by

$$G_{ST} = D_{ST}/H_T. \quad [9]$$

The latter measure depends on the population used, and the estimate obtained in one population cannot be compared with that of another, unless the breeding system is similar for the two populations. If H_S is small, G_{ST} may be very large even if the absolute gene differentiation is small. G_{ST} is equivalent to Wright's F_{ST}, and we call it the coefficient of gene differentiation. If there are only two alleles at a locus, it can be shown that $H_T = 2\bar{x}(1 - \bar{x})$ and $D_{ST} = 2\sigma^2_x$, where $\bar{x}$ and σ^2_x are the mean and variance of the frequency of an allele among subpopulations, respectively. Therefore, G_{ST} becomes identical to F_{ST}, which is defined as $\sigma^2_x/\{\bar{x}(1 - \bar{x})\}$. This property was noted by H. Harpending (personal communication) in a numerical computation. S. Wright (personal communication) also pointed out that in the case of multiple alleles, G_{ST} is equal to the weighted average of F_{ST} for all alleles, i.e., $\bar{F}_{ST} = \sum_k \sigma^2_{x(k)}/\sum_k \bar{x}_k(1 - \bar{x}_k)$, where k refers to the kth allele.

From Eqs. **8** and **9** we obtain the equation $(1 - G_{ST})(1 - J_T) = 1 - J_S$. The difference between this equation and Eq. **1** occurs because F_{IS} and F_{IT} in Eq. **1** measure the deviations of genotype frequencies from Hardy–Weinberg proportions, while J_S and J_T are gene identities. Note that G_{ST}, J_T, and J_S are all nonnegative.

As mentioned earlier, D_{ST} includes the comparisons of subpopulations with themselves. If we exclude these comparisons, we have the interpopulational gene diversity defined as

$$\bar{D}_m \equiv \sum_{i \neq j} D_{ij}/\{s(s - 1)\} = sD_{ST}/(s - 1) \quad [10]$$

This absolute measure of gene differentiation is independent of the gene diversity within subpopulations, and thus it can be used for comparing the degrees of gene differentiation in different organisms. $\bar{D}_m$ may also be used to compute the interpopulational gene diversity relative to the intrapopulational gene diversity (7). That is,

$$R_{ST} = \bar{D}_m/H_S. \quad [11]$$

Formula **8** can easily be extended to the case where each subpopulation is further subdivided into a number of colonies. In this case, H_S may be analyzed into the gene diversities within and between colonies (H_C and D_{CS}, respectively). Therefore,

$$H_T = H_C + D_{CS} + D_{ST}. \quad [12]$$

This sort of analysis can be continued to any degree of hierarchical subdivision. The relative degree of gene differentiation attributable to colonies within subpopulations can be measured by $G_{CS(T)} = D_{CS}/H_T$. It can also be shown that $(1 - G_{CS})(1 - G_{ST})H_T = H_C$, where $G_{CS} = D_{CS}/H_S$. Expression **12** was derived on the basis of two levels of hierarchies. If we disregard the level of subpopulations, we have $H_T = H_C + D_{CT}$, where D_{CT} is the gene diversity between colonies within the total population. Therefore,

$$D_{CT} = D_{CS} + D_{ST}. \quad [13]$$

In his study of human diversity, Lewontin (8) made an analysis of gene-frequency variation analogous to Eq. **12**, by using the Shannon information measure. However, this measure is not directly related to any genetic entity, and it is difficult to make a genetic interpretation of the components corresponding to those in Eq. **12**.

Let us now consider the components of the gene diversity (D_{S12}) between two subpopulations that are composed of r and s colonies. Let x_{ik} and y_{jk} be the frequencies of the kth allele in the ith colony of the first subpopulation and the jth colony of the second, respectively. By definition,

$$D_{S12} = (J_{S1} + J_{S2})/2 - J_{S12},$$

where subscripts 1 and 2 refer to the first and second

populations, respectively. From Eq. **7**, $J_{Si} = J_{Ci} - D_{CSi}$ $(i = 1, 2)$. On the other hand,

$$J_{S12} = \sum_k x_{\cdot k} y_{\cdot k}$$

$$= \sum_i^r \sum_j^s \sum_k x_{ik} y_{jk}/(rs).$$

Let $D_{ij} = (J_i + J_j)/2 - J_{ij}$, where $J_i = \sum_k x^2_{ik}$, $J_j = \sum_k y^2_{jk}$, and $J_{ij} = \sum_k x_{ik} y_{jk}$. Then,

$$J_{S12} = \sum_i^r \sum_j^s \{(J_i + J_j)/2 - D_{ij}\}/(rs)$$

$$= (J_{C1} + J_{C2})/2 - D_{C12},$$

where $D_{C12} = \sum_{ij} D_{ij}/(rs)$. Therefore, we have

$$D_{S12} = D_{C12} - (D_{CS1} + D_{CS2})/2. \qquad [14]$$

Namely, the gene diversity between two subpopulations is equal to the average gene diversity between a pair of colonies, one from each of the two subpopulations, minus the average gene diversity between the colonies within subpopulations. Formula **14** may be used for estimating D_{C12} from D_{S12} and $(D_{C1} + D_{C2})/2$. It is noted that if we take the average of D_{S12} over all combinations of subpopulations, it reduces to D_{ST} in Eq. **13**, as expected.

So far we have considered only a single locus, but the present method is applicable to any number of loci, if we replace the gene diversity for a locus by the average gene diversity for all loci studied. In fact, in order to know a general picture of gene differentiation among subpopulations, a large number of loci that is a random sample of the genome should be used, including both polymorphic and monomorphic loci (7).

In the present paper, we were mainly concerned with the gene differentiation among closely related geographical populations. If the degree of gene differentiation is large, as is the case with a group of subspecies, and J_T is much smaller than J_S, D_{ST} in Eq. **7** (or $\bar{D}_m$ in Eq. **10**) is not a good measure of differentiation. In this case a better estimate may be obtained by $D_{ST} = -\log_e (J_T/J_S)$, in analogy with the genetic distance discussed in my earlier paper (6). Similarly, a better estimate of G_{ST} may be obtained by $-\log_e(J_T/J_S)/[-\log_e J_T]$.

I thank Dr. Sewall Wright for his valuable comments on the manuscript of this paper. This work was supported by U.S. Public Health Service Grant GM 20293.

1. Wright, S. (1943) *Genetics* **28**, 114–138.
2. Wright, S. (1951) *Ann. Eugenics*, **15**, 323–354.
3. Wright, S. (1965) *Evolution* **19**, 395–420.
4. Nei, M. (1965) *Evolution* **19**, 256–258.
5. Nei, M. (1972) *Amer. Naturalist* **106**, 283–292.
6. Nei, M. (1973) in *Genetics of Population Structure*, ed. Morton, N. E. (Univ. of Hawaii, Honolulu), in press.
7. Nei, M. & Roychoudhury, A. K. (1972) *Science* **177**, 434–436.
8. Lewontin, R. C. (1973) *Evol. Biol.* **6**, 381–398.