Humangenetik 21, 85—88 (1974)

A Note on Nei's Measure of Gene Diversity in a Substructured Population*

Ranajit Chakraborty

Center for Demographic and Population Genetics,
University of Texas Health Science Center at Houston

Received October 22, 1973

Summary. The decomposition of the total gene diversity in a substructured population into its components is presented when the sizes of the subpopulations are not necessarily equal. The effect of unequal subpopulation sizes on the coefficient of gene diversity is also discussed. The sampling variance of this index of subdivision, derived here, can be used to measure the extent of the effect of subdivision more objectively.

Zusammenfassung. Die Untergliederung der gesamten genetischen Verschiedenheit in einer substrukturierten Population in ihre Komponenten wird dargestellt für den Fall, daß die Subpopulationen nicht notwendig gleich groß sind. Die Auswirkung ungleicher Größen der Subpopulationen auf den Koeffizienten der genetischen Verschiedenheit wird diskutiert. Die Stichproben-Varianz dieses Index der Unterteilung wird abgeleitet; sie kann helfen, das Ausmaß des Effektes der Unterteilung objektiver zu messen.

In a substructured population, Wright (1943) studied the variation of gene frequencies among subpopulations by the fixation indices or F-statistics. Nei, recently, studied this problem again and showed that the use of these F-statistics is not as general as it was thought to be. Since in reality the number of subpopulations in a substructured population is never infinite and often very small, and mutations are known to occur in any generation, resulting in entirely new alleles (and thus producing a large number of loci with more than two alleles at each locus), Nei (1973a, b) reformulated the problem covering the above situations. The new descriptive parameter, called as the coefficient of gene differentiation, depends on the decomposition of gene diversity (heterozygosity) of the total population (H_T) into its components, i.e., the gene diversity within (H_S) and between (D_{ST}) populations. The coefficient of gene differentiation is then defined as the proportion of the interpopulational gene diversity ($G_{ST} = D_{ST}/H_T$).

Our aim here is to present the decomposition of H_T into its components when the sizes of the subpopulations are unequal. Furthermore, we shall also present an expression for the sampling variance of estimate of G_{ST} which can possibly be used to interpret the relative role of substructuring at different levels of hierarchy.

Decomposition of H_T into Its Components

Following Nei's notation (1973a), let us denote the number of subpopulations. Let x_{ik} be the frequency of the k^{th} allele in the i^{th} subpopulation which is of size N_i.

* Work supported by grant 5P01 GM 19513 SSS from U.S. National Institutes of Health.

The gene identity (probability of identity of two randomly chosen genes) in this subpopulation is given by

$$J_i = \sum_k x_{ik}^2$$

while the same in the total population is

$$J_T = \sum_k x_{.k}^2 \qquad (1)$$

when $x_{.k} = \sum_i w_i x_{ik}$, in which w_i is the relative size of the i^{th} subpopulation ($w_i = N_i/N$, $n = \sum_i N_i$).

It is easy to show that

$$J_T = \sum_k \left(\sum_i w_i^2 x_{ik}^2 + \sum_{i \neq j}\sum w_i w_j x_{ik} x_{jk}\right) = \sum_i w_i^2 J_i + \sum_{i \neq j}\sum w_i w_j J_{ij} \qquad (2)$$

where $J_{ij} = \sum_k x_{ik} x_{jk}$ denote the gene identity between the i^{th} and j^{th} subpopulation.

The gene diversity between two different subpopulations i and j as defined by the minimum number of codon differences per locus is given by

$$D_{ij} = (J_i + J_j)/2 - J_{ij}$$

and thus

$$\begin{aligned} J_T &= \sum_i w_i^2 J_i + \sum_{i \neq j}\sum w_i w_j [(J_i + J_j)/2 - D_{ij}] = \sum_i w_i J_i - \sum_{i \neq j}\sum w_i w_j D_{ij} \\ &= J_S - D_{ST} \end{aligned} \qquad (3)$$

where $J_S = \sum_i w_i J_i$ and $D_{ST} = \sum_{i \neq j}\sum w_i w_j D_{ij}$. Note that J_S and D_{ST} has the same interpretation as in Nei (1973a). Transforming J_T and J_S into diversity measures ($H_T = 1 - J_T$, $H_S = 1 - J_S$) we get the same decomposition as in Nei (1973a) given by

$$H_T = H_S + D_{ST}. \qquad (4)$$

It may be noted that when $w_i = 1/s$ for all $i = 1, 2, \ldots s$ (i.e., when N_i's are equal), H_S and D_{ST} are identical to the expressions given in Nei (1973a).

It may be worthwhile to see the effect of the approximations of assuming equality of subpopulation sizes. During the analysis of gene frequency variations at village/tribal levels of several human populations (Roychoudhury and Chakraborty, unpublished) we noticed that when the gene diversity (H_T) in the total population is very small (i.e., when the populations, as a whole, has a high genic identity), the relative approximation involved due to the assumption of equal population size becomes quite appreciable. However, in the cases where interpopulational gene diversity is very small as compared to intrapopulational one, the effect of the unequal population sizes is not appreciable. The situation is best described by considering the case of Makiritare Indians. Gershowitz *et al.* (1970), Arends *et al.* (1970) and Weitkamp and Neel (1970) studied 15 blood group loci and 17 protein loci in 7 villages of Makiritare Indians from southern Venezuela. The population sizes of those villages varied between 70 to 176. In Table 1 we present the total gene diversities (as measured by heterozygosities), net codon differences and the relative amount of approximation involved in the assumption of equal population sizes for each of these loci separately. The last row presents the same parameters averaged over all the serological and biochemical markers. Though there is a wide variation in the effect of approximation from locus to

Table 1. Gene diversity parameters (H_T and D_{ST}) and effect of unequal sizes of the subpopulations on G_{ST}-estimates for different gene markers in seven Makiritare villages

Locus	Gene diversity in the total population (H_T)	Interpopulational gene diversity (D_{ST})	Effect of approximation on G_{ST}[a] (%)
Serological			
MN	0.4053	0.0114	1.22
Ss	0.4985	0.0296	29.48
P	0.4943	0.0092	25.54
Rh(C)	0.4897	0.0024	16.24
Rh(E)	0.4946	0.0029	36.16
Duffy	0.3775	0.0054	5.51
Kidd	0.4331	0.0120	14.18
Diego	0.3271	0.0425	30.92
Lewis	0.4987	0.0452	15.15
Average[b]	0.2679	0.0107	20.38
Biochemical			
Hp	0.4808	0.0350	19.17
Gc	0.2825	0.0078	57.00
Lp	0.2193	0.0013	19.64
Alb.	0.0244	0.0004	2.61
AP	0.1056	0.0016	59.27
PGM_1	0.2683	0.0123	3.40
6PGD	0.0207	0.0004	14.76
Average[b]	0.0824	0.0035	22.31
Average[b] (overall loci)	0.1694	0.0069	20.90

[a] Effect of approx. = (difference in G_{ST}-estimates with unequal wt. and equal wt.)/G_{ST}-estimate with unequal wt.

[b] Including the monomorphic loci (ABO, Kell, Lu, Wr, ABH-Secretor and Rh(D) among the serological and Tf, Cp, Ps, PGM_2, AK, LDH-A, LDH-B, G-6-PD, Oxidase and ADA among the biochemical markers).

locus, from the average values we see that the effect is more or less similar for the serologic and biochemical markers even though heterozygosity figures differ by a factor of nearly three. This may be ascribed to the small value of interpopulational gene diversity (D_{ST}).

Sampling Variance of Coefficient of Gene Differentiation

The relative measure of gene differentiation in a substructured population (G_{ST}) is defined by

$$G_{ST} = D_{ST}/H_T.$$

Nei and Roychoudhury (1974) studied the sampling variances of heterozygosity and genetic distances. From their study it can be seen that if $H_T^{(1)}, H_T^{(2)}, \ldots, H_T^{(n)}$ denotes the gene diversities (as measured by heterozygosities) at n randomly chosen loci, the sampling variance of average of these quantities is given by

$$V(H_T) = \frac{1}{n(n-1)} \sum_{k=1}^{n} (H_T^{(k)} - H_T)^2 \tag{5}$$

where $H_T = \frac{1}{n} \sum_{k=1}^{n} H_T^{(k)}$.

Similarly, for each locus we may compute D_{ST} and thus the sampling variance of D_{ST}, $V(D_{ST})$ is given by a similar expression. The covariance between D_{ST} and H_T at each locus is also computed similarly.

Thus, we have

$$V(G_{ST}) \cong G_{ST}^2 \left\{ \frac{V(D_{ST})}{D_{ST}^2} + \frac{V(H_T)}{H_T^2} - \frac{2\,\mathrm{Cov.}\,(H_T, D_{ST})}{D_{ST} \cdot H_T} \right\}.$$

It may be noted that in (5) n also includes the number of monomorphic loci selected in the sample.

This sampling variance can be used to study the significance of the effect of subdivision. In the case with Makiritare Indians we get a G_{ST} estimate of 0.0405 with a standard error (given by the square root of the expression in $V(G_{ST})$) of 0.0099 which shows a significant effect of subdivision in the present sample of Makiritare Indians. The investigators of the project also came to the same conclusion though their treatment largely depends upon the study of individual gene markers and justifications being provided by demographic parameters. Thus, the present analysis seems to have solved, at least partially the problem raised by Ward and Neel (1970): "One way to quantitate the degree of microdifferentiation is by using genetic distance functions, although assigning variances to them is complicated by the same factors that raise difficulties in using χ^2."

Acknowledgements. I am grateful to Dr. M. Nei for many helpful comments during the preparation of this note. The work is supported by the Grant 5P01 GM 19513 SSS from U.S. National Institutes of Health.

References

Arends, T., Weitkamp, L. R., Gallango, M. L., Neel, J. V., Schultz, J.: Gene frequencies and microdifferentiation among the Makiritare Indians. II. Seven serum protein systems. Amer. J. hum. Genet. **22**, 536 (1970)

Gershowitz, H., Layprisse, M., Layprisse, Z., Neel, J. V., Brewer, C., Chagnon, N., Ayres, M.: Gene frequencies and microdifferentiations among the Makiritare Indians. I. Eleven blood group systems and the ABH-Le secretor traits: A note on Rh gene frequency determinations. Amer. J. hum. Genet. **22**, 515—525 (1970)

Nei, M.: Analysis of gene diversity in subdivided populations. Proc. nat. Acad. Sci. (Wash.) (in press, 1973a)

Nei, M.: Dynamics of gene differentiation among a finite number of populations. Amer. J. hum. Genet. (submitted, 1973b)

Nei, M., Roychoudhury, A. K.: Sampling variances of heterozygosity and genetic distance. Genetics (in press, 1974)

Ward, R. H., Neel, J. V.: Gene frequencies and microdifferentiation among the Makiritare Indians. IV. A comparison of a genetic network with ethnohistry and migration matrices; a new index of genetic isolation. Amer. J. hum. Genet. **22**, 538—561 (1970)

Weitkamp, L., Neel, J. V.: Gene frequencies and microdifferentiation among the Makiritare Indians. III. Nine erythrocyte enzyme systems. Amer. J. hum. Genet. **22**, 533—537 (1970)

Wright, S.: Isolation by distance. Genetics **28**, 116—138 (1943)

Ranajit Chakraborty
Center for Demographic and Population Genetics
University of Texas Health Science Center
Houston, Texas 77025, USA