F-statistics and analysis of gene diversity in subdivided populations

BY MASATOSHI NEI

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

INTRODUCTION

In a study of genetic differentiation of populations Wright (1943, 1951) introduced the following formula:

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

where \( F_{IT} \) and \( F_{IS} \) are the correlations between the two uniting gametes to produce the individuals relative to the total population and relative to the subpopulations, respectively, whereas \( F_{ST} \) is the correlation between two gametes drawn at random from each subpopulation and measures the degree of genetic differentiation of subpopulations. \( F_{IS} \) and \( F_{IT} \) are often called fixation index and may become negative, while \( F_{ST} \) is always non-negative. Cockerham (1969, 1973) studied detailed properties of these F-statistics in terms of the intraclass correlation of alleles for various subdivisions of populations.

Wright’s or Cockerham’s definition of F-statistics has been made in terms of neutral genes, and it is not clear whether the same definition applies in the presence of selection and migration particularly when there are multiple alleles. Wright has also implicitly assumed that the number of subpopulations is infinitely large. In practice, this number is often very small, so that the relationship between Wright’s theory and real data is not very clear. In recent years there have been a number of attempts to reformulate the F-statistics in terms of probability treatment (Crow & Kimura, 1970, p. 108; Jacquard, 1974, p. 169), but they have not been very successful. This is because (1) the probability treatment introduced by Malécot (1948) is based on the same assumptions as those in Wright’s theory of inbreeding and (2) for defining Malécot’s identity of genes by descent we must know the pedigrees of individuals within and between subpopulations in the presence of migration. The difficulty of relating the probabilities of gene identity to Bernstein’s (1930) \( \alpha \) (equivalent to Wright’s \( F_{IS} \)) or Wahlund’s (1928) variance of gene frequency (related to Wright’s \( F_{ST} \)) in real-world populations was recently emphasized by Smith (1974).

Nei (1973) introduced a somewhat different approach to the study of genetic differentiation of populations. He showed that the gene diversity of the total population can be partitioned into its components, i.e. intrasubpopulational and intersubpopulational gene diversities, when gene diversity is defined as the heterozygosity (frequency of heterozygotes) expected under Hardy–Weinberg equilibrium. This gene diversity analysis is primarily designed to be applied to the average gene diversity for a large number of loci among a finite number of subpopulations. However, it can also be applied to a single locus. In Nei’s theory gene diversity is defined by using the gene frequencies at the present generation, so that no assumption is required about the pedigrees of individuals, selection, and migration in the past.

In this paper I shall extend this idea and show that the F-statistics need not be defined as the correlation of uniting gametes but can be defined as a function of observed and expected
heterozygosities. The $F$-statistics thus defined can be applied to any situation whether there is selection or not and no matter how many alleles are segregating at a locus. The relationship between the $F$-statistics analysis and gene diversity analysis will also be discussed.

### $F$-statistics for a pair of alleles

Let us first study how $F$-statistics can be expressed in terms of gene and genotype frequencies for the case of two alleles at a locus. Note that most theoretical studies on $F$-statistics have been done with respect to this case.

Consider a population divided into $s$ subpopulations in each of which Hardy–Weinberg equilibrium does not necessarily hold. We designate the two alleles by $A_1$ and $A_2$. Let $p_i$ and $F_{Is}$ be the gene frequency of $A_1$ and the fixation index in the $i$th subpopulation, respectively. The frequency of homozygote $A_1A_1$ in the $i$th subpopulation is then given by

$$P_i = p_i^2 + F_{Is} p_i (1 - p_i).$$  \hspace{1cm} (2)

Therefore,

$$F_{Is} = \frac{p_i - p_i^2}{p_i (1 - p_i)}.$$  \hspace{1cm} (3)

Note that the above formulae hold true whether there is selection or not (Wright, 1965).

Following Wright (1965) and Kirby (1975), we define $F_{IS}$ as a weighted average of $F_{Is}$. Namely,

$$F_{IS} = \frac{\sum_{i=1}^{s} w_i P_i (1 - p_i) F_{Is}}{\sum_{i=1}^{s} w_i P_i (1 - p_i)},$$  \hspace{1cm} (4)

where $w_i$ is the relative size of the $i$th subpopulation with $\Sigma_i w_i = 1$. (In most instances $w_i = 1/s$ may be assumed, since population size is quite transitory and geneticists are generally interested in gene frequency differences between populations, disregarding the effect of population size.)

A number of authors (e.g. Neel & Ward, 1972; MacCluer, 1974) defined $F_{IS}$ as $\Sigma_i w_i F_{Is}$. However, if we use this definition, Wright’s formula (1) cannot be obtained. In fact, to get formula (1) we must use an average weighted with heterozygosities for all $F$’s, as will be seen in the following. At any rate, (4) reduces to

$$F_{IS} = \frac{\Sigma_i w_i (P_i - p_i^2)}{\Sigma w_i P_i (1 - p_i)} = \frac{P - \overline{p}^2}{\overline{p} - \overline{p}^2} = 1 - \frac{\overline{p} - P}{\overline{p} - \overline{p}^2},$$  \hspace{1cm} (5)

where $P = \Sigma_i w_i P_i$ is the relative frequency of $A_1A_1$ in the total population, $\overline{p} = \Sigma_i w_i p_i$ is the mean gene frequency and $\overline{p}^2 = \Sigma_i w_i p_i^2$. On the other hand, $P$ may be written as

$$P = (1 - F_{IT}) \overline{p}^2 + F_{IT} \overline{p}.$$  \hspace{1cm} (6)

Therefore,

$$F_{IT} = \frac{P - \overline{p}^2}{\overline{p} (1 - \overline{p})} = 1 - \frac{\overline{p} - P}{\overline{p} - \overline{p}^2}.$$  \hspace{1cm} (7)

$F_{ST}$ can be obtained by inserting (5) and (7) into equation (1). It becomes

$$F_{ST} = \frac{\overline{p}^2 - \overline{p}^2}{\overline{p} (1 - \overline{p})} = 1 - \frac{\overline{p} - \overline{p}^2}{\overline{p} - \overline{p}^2}.$$  \hspace{1cm} (8)
It is clear that (8) agrees with the usual definition of $F_{ST}$, i.e. $\sigma_p^2/\overline{\sigma}(1-\overline{\sigma})$, where $\sigma_p^2$ is the variance of gene frequency among subpopulations.

In the above formulation we used the gene and genotype frequencies for $A_1$. Obviously, similar formulae can be obtained for $A_2$. The formula for $F_{ST}$ for $A_2$ is exactly the same as that for $A_1$, since $\overline{\sigma}-\overline{\sigma}^2$ is equal to $1-\overline{\sigma}-(1-\overline{\sigma})^2$. It can also be shown that $F_{IS}$ and $F_{IT}$ are the same for $A_1$ and $A_2$. Therefore, $F_{IS}$ and $F_{IT}$ can be written as

$$F_{IS} = 1 - \frac{\overline{\sigma} + (1-\overline{\sigma}) - P_1 - P_2}{2(\overline{\sigma}-\overline{\sigma}^2)} = 1 - \frac{1 - J_0}{2(\overline{\sigma}-\overline{\sigma}^2)},$$

$$F_{IT} = 1 - \frac{1 - J_0}{2\overline{\sigma}(1-\overline{\sigma})},$$

where $P_1$ and $P_2$ are the observed genotype frequencies of $A_1A_1$ and $A_2A_2$ in the entire population, respectively, and $J_0 = P_1 + P_2$ is the observed homozygosity. If we note $2(\overline{\sigma}-\overline{\sigma}^2) = 2\Sigma w_i p_i(1-p_i)$, it is clear that $F_{IS}$, $F_{IT}$ and $F_{ST}$ are all a function of observed and expected heterozygosities. This property will be used explicitly in the next section.

**Extension of $F$-statistics**

If there are more than two alleles at a locus, the genotype frequencies in a population cannot be specified completely by the gene frequencies and a single $F_{IS}$ parameter except for the special case of random differentiation with infinite subpopulations (Nei, 1985). When there are $n$ alleles, $n(n-1)/2$ $F_{IS}$ parameters are required for a complete specification of genotype frequencies (Weir, 1970; Rao & Chakraborty, 1974). If we consider only homozygotes, however, $n$ $F_{IS}$ parameters are sufficient. Thus, the frequency of homozygotes for the $k$th allele, i.e. $A_kA_k$, in the $i$th subpopulation may be written as

$$P_{ik} = p_{ik}^2 + F_{ISik} p_{ik}(1-p_{ik}),$$

where the subscript $k$ refers to the $k$th allele. Therefore,

$$F_{ISik} = \frac{P_{ik} - p_{ik}^2}{p_{ik}(1-p_{ik})}.$$  

It is clear that the $F_{IS}$, $F_{IT}$, and $F_{ST}$ for the $k$th allele can be defined as in the case of the foregoing section. Hence

$$F_{ISk} = \frac{P_k - \overline{p}_k^2}{\overline{p}_k - \overline{p}_k^2},$$

$$F_{ITk} = \frac{P_k - \overline{p}_k^2}{\overline{p}_k - \overline{p}_k^2},$$

$$F_{STk} = \frac{\overline{p}_k^2 - \overline{p}_k^2}{\overline{p}_k - \overline{p}_k^2},$$

where $P_k = \Sigma w_i P_{ik} \overline{p}_k$, $\overline{p}_k = \Sigma w_i p_{ik}$, and $\overline{p}_k^2 = \Sigma w_i p_{ik}^2$. Needless to say, these quantities satisfy formula (1).
As noted earlier, $F_{ISi}^k$ may vary from allele to allele. It is, however, possible to define a unified fixation index for all alleles in the $i$th sub-population. Namely,

$$F_{ISi} = \frac{\sum_{i=1}^{n} p_{ik}(1-p_{ik}) F_{ISi}^k}{\sum_{k=1}^{n} p_{ik}(1-p_{ik})} = \frac{\Sigma_k (P_{ik} - p_{ik}^2)}{\Sigma_k p_{ik}(1-p_{ik})} = (H_{Si} - H_{Oi})/H_{Si},$$

where $H_{Si} = 1 - \Sigma_k P_{ik}^2$ and $H_{Oi} = 1 - \Sigma_k P_{ik}$ are the expected and observed heterozygosities, respectively. Namely, $F_{ISi}$ is defined as the ratio of the difference between the expected and observed heterozygosities to the expected heterozygosity, whether there is selection or not. Note that this definition is different from that used by Jain & Workman (1967) and Rao & Chakraborty (1974).

Following (4), $F_{IS}$ is now defined as $\Sigma_i w_i H_{Si} F_{ISi}/\Sigma w_i H_{Si}$, and becomes

$$F_{IS} = (H_S - H_O)/H_S,$$

where $H_S$ and $H_O$ are the weighted averages of $H_{Si}$ and $H_{Oi}$ over all subpopulations, respectively. The above formula can also be obtained from (13a) in the following way:

$$F_{IS} = \frac{\Sigma_k (P_k - P_{ik}) F_{ISi}^k}{\Sigma_k (P_k - P_{ik})^2}.$$

It is clear that the fixation index in the total population ($F_{IT}$) can be defined in the same way as (15) by using the gene and genotype frequencies in the total population. It becomes

$$F_{IT} = (H_T - H_O)/H_T,$$

where $H_T = 1 - \Sigma P_k^2$. This again can be obtained from (13b) by using a definition similar to (16). Similarly, we define $F_{ST}$ as

$$F_{ST} = (H_T - H_S)/H_T.$$

Obviously, $F_{IS}$, $F_{IT}$ and $F_{ST}$ thus defined satisfy formula (1). It is clear that in the case of two alleles ($n = 2$) formulae (15), (17) and (18) reduce to (9), (10) and (8), respectively.

In the above formulation we first defined $F_{IS}$, $F_{IT}$ and $F_{ST}$ for each of the $n$ alleles (or homozygotes) at a locus and then derived the weighted averages of these quantities for the locus. However, it is also possible to define $F_{ISi}$, $F_{IT}$ and $F_{ST}$ for each of the $n(n-1)/2$ pairs of alleles (or heterozygotes) and then derive the weighted averages. Namely, the observed frequency of heterozygote $A_k A_l$ in the $i$th subpopulation may be written as $P_{ikl} = 2P_{ik} P_{li}(1-F_{ISi}^k)$. Therefore,

$$F_{ISi} = \frac{2P_{ik} P_{li} - P_{ikl}}{2P_{ik} P_{li}}.$$

Note that this definition is different from that of Rao and Chakraborty (1974). The weighted average of $F_{ISi}$ over all subpopulations is then given by

$$F_{IS} = \frac{2P_{ik} P_{li} - P_{ikl}}{2P_{ik} P_{li}}.$$
F-statistics and gene diversity analysis

where \( \bar{p}_k \bar{p}_1 = \sum \psi_i p_{ik} \bar{p}_i \) and \( p_{k1} = \sum \psi_i p_{ik} \bar{p}_1 \). Similarly, \( F_{IT} \) and \( F_{ST} \) for heterozygote \( A_k A_1 \) are given by

\[
F_{IT} = \frac{2 \bar{p}_k \bar{p}_1 - p_{MT}}{2 \bar{p}_k \bar{p}_1},
\]

\[
F_{ST} = \frac{2 \bar{p}_k \bar{p}_1 - 2 \bar{p}_k \bar{p}_i}{2 \bar{p}_k \bar{p}_1}.
\]

(20b)

(20c)

Note that \( F_{ST} \) is identical with Nei’s (1965) earlier definition of this quantity, i.e.

\(-\text{Cov}(p_k, p_1)/(p_k \bar{p}_1)\),

where Cov \((p_k, p_1)\) is the covariance of \( p_k \) and \( p_1 \) over subpopulations. It is now easy to see that the weighted averages of (20a), (20b), and (20c) over all heterozygotes become identical with (15), (17) and (18), respectively.

Therefore, at a locus with \( n \) codominant alleles we can compute \( n(n+1)/2 \) sets of \( F_{IS}, F_{IT} \) and \( F_{ST} \), if we include those obtained by (13a)–(13c). Of course, they are not all independent of each other.

It is interesting to see that all \( F_{IS}, F_{IT} \) and \( F_{ST} \) can be defined in terms of expected and observed heterozygosities and take a similar mathematical form (15, 17, 18). Since these are defined in terms of the present gene and genotype frequencies, they can be applied to any situation, whether there is selection or not. However, the F-statistics defined in this way still have a property similar to that of Wright’s definition. That is, \( F_{IS} \) and \( F_{IT} \) both can be negative, but \( F_{ST} \) is a non-negative quantity, since \( H_F \geq H_S \). As in the case of Wright’s definition, \( F_{IS} \) and \( F_{IT} \) measure the deviations of genotype frequencies from Hardy–Weinberg proportions in the subpopulations and in the total population, respectively, whereas \( F_{ST} \) measures the degree of genetic differentiation of subpopulations. Conceptually, however, there is one important difference between the two definitions. Namely, in Wright’s theory, \( F_{IS}, F_{IT} \) and \( F_{ST} \) are defined relative to a hypothetical population that is composed of an infinite number of subpopulations, whereas in the present definition they are defined relative to the existing total population in which gene frequency survey is made.

In the above definition of F-statistics we used population gene frequencies. Therefore, if all individuals of a population are examined for gene frequency survey, the F-statistics computed have no sampling variance. However, the F-statistics defined above are still subject to the error due to random genetic drift (sampling error at the time of reproduction), since the population is finite by definition. If population size is relatively small, the variance due to drift is considerably large and the F-statistics may fluctuate from generation to generation. To evaluate the drift variance, we must know the breeding structure of the population. For some special types of breeding structure, the drift variance of \( F_{ST} \) has already been evaluated (Nei & Chakravarti, 1977; Nei, Chakravarti & Tateno, 1977).

Numerical example

Gershowitz et al. (1967) studied the genotype frequencies for the MNSs blood group system in three populations of the Xavante Indians in South America. In this study virtually all individuals of the populations are sampled, so that the sampling error at the time of survey is practically nil. Treating MS, Ms, NS and Ms chromosome segments as multiple alleles, they estimated the gene frequencies for the three populations under the assumption of Hardy–Weinberg equilibrium. The gene frequencies as well as the four homozygote frequencies are presented in Table 1. From these gene and genotype frequencies the \( F_{IS} \) values for each allele
Table 1. Gene ($p_i$) and homozygote ($P_i$) frequencies for the MNSs locus in three populations of the Xavante American Indians (from Gershonowitz et al. 1967) and estimates of $F_{ISk}$

<table>
<thead>
<tr>
<th>Allele</th>
<th>São Domingos (79)*</th>
<th>São Marcos (287)</th>
<th>Simões Lopes (171)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p$</td>
<td>$P$</td>
<td>$F_{ISk}$</td>
</tr>
<tr>
<td>$MS$</td>
<td>0.33</td>
<td>0.1392</td>
<td>0.1370</td>
</tr>
<tr>
<td>$Ms$</td>
<td>0.38</td>
<td>0.1646</td>
<td>0.0857</td>
</tr>
<tr>
<td>$NS$</td>
<td>0.09</td>
<td>0.0000</td>
<td>0.0989</td>
</tr>
<tr>
<td>$Ns$</td>
<td>0.20</td>
<td>0.0380</td>
<td>0.0125</td>
</tr>
<tr>
<td>Weighted average</td>
<td>—</td>
<td>—</td>
<td>0.0575</td>
</tr>
</tbody>
</table>

* Number of individuals examined.

Table 2. $F$-statistics for the MNSs locus in the Xavante population of American Indians

<table>
<thead>
<tr>
<th>Allele</th>
<th>Average gene frequency ($\bar{p}$)</th>
<th>$P^*$</th>
<th>$\bar{p}^2$</th>
<th>$\bar{p}^3$</th>
<th>$F_{ISk}$</th>
<th>$F_{ITk}$</th>
<th>$F_{STk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MS$</td>
<td>0.3633</td>
<td>0.1613</td>
<td>0.1320</td>
<td>0.1326</td>
<td>0.1244</td>
<td>0.1266</td>
<td>0.0026</td>
</tr>
<tr>
<td>$Ms$</td>
<td>0.4067</td>
<td>0.1684</td>
<td>0.1654</td>
<td>0.1681</td>
<td>0.0013</td>
<td>0.0012</td>
<td>0.0013</td>
</tr>
<tr>
<td>$NS$</td>
<td>0.0833</td>
<td>0.0058</td>
<td>0.0069</td>
<td>0.0102</td>
<td>0.0602</td>
<td>0.0144</td>
<td>0.0432</td>
</tr>
<tr>
<td>$Ns$</td>
<td>0.1407</td>
<td>0.0197</td>
<td>0.0215</td>
<td>0.0229</td>
<td>0.0258</td>
<td>0.0144</td>
<td>0.0112</td>
</tr>
<tr>
<td>Weighted average</td>
<td>—</td>
<td>—</td>
<td>0.0321</td>
<td>—</td>
<td>—</td>
<td>0.0436</td>
<td>—</td>
</tr>
</tbody>
</table>

* Average frequency of homozygotes over the three populations with an equal weight.
† Average of $p^3$ over the three populations with an equal weight.

can be computed by using (12). The results obtained are given in Table 1. It is seen that $F_{ISk}$ is high for allele $MS$ but low for $NS$ in all three populations. This suggests the possibility that either selection or non-random mating or both caused an excess of homozygote $MS/MS$ and a deficiency of $NS/NS$. Therefore, I tested the significance of the excess and the deficiency by using the formula $\chi^2 = NF_{ISk}$ (Li, 1955), where $N$ is the number of individuals studied. This test showed that the excess of $MS/MS$ in São Marcos was significant at the 1% level, while all other excesses and deficiencies were nonsignificant. As indicated earlier, $F_{ISk}$ can also be computed for each heterozygote. However, this quantity does not seem to be very informative unless the number of individuals in each subpopulation is very large. Note also that in the MNSs system $MS/Ns$ and $Ms/NS$ cannot be distinguished, so that $F_{ISk}$ is not obtainable for these genotypes.

The $F$-statistics for the entire population of Xavante Indians are presented in Table 2. In the computation of these values $\omega_i = 1/3$ was assumed. The $F$-statistics, particularly $F_{ISk}$ and $F_{ITk}$, vary with allele considerably. The $\chi^2$ test mentioned above shows that the $F_{ISk}$ for allele $MS$ is again significantly different from 0. If the number of subpopulations is large, the deviation of $F_{STk}$ from that expected under random differentiation may be tested by the method of Nei & Imaizumi (1966). In the present case, however, this method is not efficient since there are only three subpopulations. Neel & Ward (1972) have computed the values of $F_{IS}$, $F_{IT}$ and $F_{ST}$ for the $MN$ and $Ss$ loci separately in the same population by using a somewhat different method. The averages of their values for the two loci are close to our values of $F_{IS}$, $F_{IT}$ and $F_{ST}$ (averages in Table 2).
F-statistics and gene diversity analysis

Nei (1973) defined gene diversity for a single locus as the heterozygosity expected under Hardy–Weinberg equilibrium, disregarding the actual genotype frequencies in the population. Namely, gene diversity is a measure of the degree of genic variation in a population and has nothing to do with the frequency of heterozygotes except in a Hardy–Weinberg population. With this definition, I have shown that $H_T = H_S + D_{ST}$, where $D_{ST}$ is the interpopulational gene diversity. I have called the ratio ($G_{ST}$) of $D_{ST}$ to $H_T$ the coefficient of gene differentiation. This coefficient is identical with the $F_{ST}$ defined in this paper. If we note that $G_{ST} = 1 - H_S/H_T$, $H_S = 1 - J_S$, and $H_T = 1 - J_T$, where $J_S$ and $J_T$ are the average expected homozygosity within subpopulations and the expected homozygosity in the total population, respectively, we obtain

$$1 - J_S = (1 - J_T) (1 - G_{ST})$$

(Nei, 1973). This formula is similar in form to (1). To some readers this similarity seems to have been a source of confusion, particularly because some authors have used the notation $F$ or $f$ to represent $J$. However, $J_S$ and $J_T$ are entirely different parameters from $F_{IS}$ and $F_{IT}$. Note that $J_S$ and $J_T$ never become negative, unlike $F_{IS}$ and $F_{IT}$. S. Wright (personal communication) has suggested that the notation $F$ should not be used to represent homozygosity.

Gene diversity analysis was proposed primarily to be applied to the average gene diversity for a large number of loci. In this case the average gene diversities, $H_T$, $H_S$ and $D_{ST}$, are obtained simply by taking the averages of $H_T$, $H_S$ and $D_{ST}$ over all loci. With these average gene diversities, formula (21) again holds if we replace $J_S$, $J_T$ and $G_{ST}$ by $J_S = 1 - H_S$, $J_T = 1 - H_T$ and $G_{ST} = D_{ST}/H_T$, respectively. One may wish to obtain average $F$-statistics for many loci corresponding to average gene diversities. They are given by

$$\bar{F}_{IS} = \Sigma_1 H_S(i) F_{IS(i)} / \Sigma_1 H_S(i) = (\bar{H}_S - \bar{H}_O)/\bar{H}_S,$$

$$\bar{F}_{IT} = \Sigma_1 H_T(i) F_{IT(i)} / \Sigma_1 H_T(i) = (\bar{H}_T - \bar{H}_O)/\bar{H}_T,$$

$$\bar{F}_{ST} = \Sigma_1 H_T(i) F_{ST(i)} / \Sigma_1 H_T(i) = (\bar{H}_T - \bar{H}_S)/\bar{H}_T,$$

where $i$ refers to the $i$th locus. These average $F$-statistics again satisfy formula (1).

So far we have considered only one degree of hierarchical subdivision of population. However, the above theory can be extended to any number of degrees of hierarchical subdivisions. For example, if each subpopulation is subdivided into a number of colonies, the total gene diversity ($H_T$) may be partitioned as follows: $H_T = H_C + D_{CS} + D_{ST}$, where $H_C$ and $D_{CS} = H_S - H_C$ are the gene diversities within and between colonies within subpopulations (Nei, 1973). On the other hand, Wright’s $F$-statistics for this case can be written as $1 - F_{IT} = (1 - F_{IC})(1 - F_{CS})(1 - F_{ST})$, where $F_{IC}$, $F_{CS}$ and $F_{ST}$ are $(H_C - H_O)/H_C$, $(H_S - H_O)/H_S$ and $(H_T - H_S)/H_T$, respectively. Here $H_O$ is the observed heterozygosity within colonies.

It is now clear that gene diversity analysis and $F$-statistics analysis are closely related to each other. However, there are some differences between them. In the former the gene diversity of the total population is directly decomposed into its components, and the component gene diversities are often expressed as a proportion of the total diversity (e.g. $H_C/H_T$, $D_{CS}/H_T$, $D_{ST}/H_T$ in the above example). In the latter analysis, however, the quantities concerned are always the ratios of two different types of gene diversities. Note also that these ratios are not necessarily the same as the corresponding component gene diversities expressed as proportions of the total.
gene diversity. The motivation of gene diversity analysis is also different from that of $F$-statistics analysis. The former was developed primarily to estimate the inter- and intra-populational genic variations with respect to the entire genome of the organism concerned (see Nei (1975, p. 152) for the application of this method to various organisms). Thus, in this analysis it is important to use a large number of loci which are ideally a random sample from the total genome. On the other hand, $F$-statistics analysis is primarily concerned with the relationship between the genotype frequencies in the total population and in the subpopulations for a single locus.

Conceptually, gene diversity analysis is similar to Lewontin’s (1972) apportionment of genetic variability in terms of the Shannon information measure $I = -\sum p_i \log_2 p_i$. Both methods are concerned with partitioning of the genetic variation of the total population into its components. Biologically, however, the Shannon information measure does not seem to have much meaning, whereas gene diversity is the most important measure of genetic variability of a population and can be related to the number of codon differences per locus. The theoretical properties of gene diversities under evolutionary forces have been studied extensively (see Nei, 1975). Therefore, gene diversity analysis seems to permit a more meaningful interpretation of the results obtained than the analysis by the Shannon information measure.

Wright’s (1943, 1965) theory of evolutionary change of $F$-statistics depends on the assumption of infinite number of subpopulations. Recently, Nei & Chakravarti (1977) and Nei et al. (1977) removed this assumption and studied the evolutionary change of $F_{ST}$ (or $G_{ST}$) in a subdivided population of finite size with and without migration. This theory will be useful for making the biological interpretation of the results obtained by $F$-statistics analysis. The evolutionary changes of gene diversity under mutation pressure in a finite number of subpopulations have been studied by Nei (1975, p. 121).

**Summary**

It is shown that Wright’s $F$-statistics can be defined as ratios of gene diversities of heterozygote frequencies rather than as the correlations of uniting gametes. This definition is applicable irrespective of the number of alleles involved or whether there is selection or not. The relationship between $F$-statistics and Nei’s gene diversity analysis is discussed.

This study was supported by PHS research grant GE 20293 and NSF research grant DEB 76-06069.

**References**


F-statistics and gene diversity analysis


NEEL, J. V. & WARD, R. H. (1972). The genetic structure of a tribal population, the Yanomama Indians. VI. Analysis by F-statistics (including a comparison with the Makiritare and Xavante). Genetics 72, 639–66.


