Theor Appl Genet (1994) 89:198–200


R. Finkeldey

# A simple derivation of the partitioning of genetic differentiation within subdivided populations

Received: 24 November 1993 / Accepted: 21 December 1993

**Abstract** The genetic differentiation $\delta_T = 1 - \sum_i p_i^2$ within a subdivided population can be partitioned into two proportions, one due to differentiation within subpopulations and the other due to differentiation among subpopulations. A simple mathematical derivation of this procedure, known as Nei's $G_{ST}$-statistics, is presented. The significance of considering the differing relative sizes of subpopulations is stressed. Possible fields of applications for breeders and conservationists who are concerned with the establishment of genetically diverse populations are mentioned.

**Key words** Genetic differentiation · Population structure · Gene diversity · Differentiation measures

## Introduction

One of the most frequently computed measures of allelic variation at single gene loci within populations is the term $1 - \sum_i p_i^2$, which equals the probability that two alleles chosen at random from an infinite population differ in type. Nei 1973 suggested the term "gene diversity" for this probability of non-identity. As the expression equals the proportion of heterozygotes within a population at Hardy-Weinberg-equilibrium, it is often referred to as "expected heterozygosity", $H_e$ (Nei 1977). Gregorius (1987) proposed the interpretation of $\delta_T = 1 - \sum_i p_i^2$ as a measure of the total gene differentiation within an effectively infinite population. His notion shall be adopted here since the "diversity" of a population should directly reflect its effective number of types (Gregorius 1987) and since the term "heterozygosity" is inappropriate in the following context.

The problem to be addressed is the computation of the overall gene differentiation, $\delta_T(g)$, of a population $g$ that is divided into several subpopulations $j$ ($j = 1 \ldots n$). The gene differentiation within each subpopulation is $\delta_T(j)$, and each subpopulation is represented in a proportion $c_j$ of the population $g$. This problem is of relevance to a breeder who wants to compose a genetically diversified working collection from a set of accessions of a base collection. Another field of application is in gene conservation, if a genetic resource is to be composed of several previously distinct populations. Maximization of gene differentiation within an unstructured genetic resource may be accompanied by a large genetic load; hence, it will only rarely be the objective of a pooling of subpopulations. However, a knowledge of the factors that influence $\delta_T(g)$ is of interest to anyone concerned with the pooling of previously distinct populations.

## Computation of the total gene differentiation

Let a population $g$ be composed of $n$ subpopulations, each of relative size $c_j[c_l]$ ($\sum_j c_j = \sum_l c_l = 1$). Let $p_i(j)[p_i(l)]$ denote the relative frequency of the $i$th allele in the $j$th [$l$th] subpopulation and $p_i(g)$ denote the relative frequency of the $i$th allele in the pooled population ($\sum_i p_i(j) = \sum_i p_i(l) = \sum_i p_i(g) = 1$). For each subpopulation $j$ the total gene differentiation $\delta_T(j)$ equals $1 - \sum_i p_i(j)^2$, and the relative frequency of the $i$th allele in the pooled population $p_i(g)$ equals $\sum_j c_j p_i(j)$.

$\delta_T(g)$ may be computed as follows:

$$\delta_T(g) = 1 - \left[\sum_i p_i(g)^2\right]$$

$$= 1 - \left[\sum_i \left(\sum_j c_j \cdot p_i(j)\right)^2\right]$$

Communicated by P. M. A. Tigerstedt

R. Finkeldey[1]
Abteilung für Forstgenetik und Forstpflanzenzüchtung, Georg-August-Universität Göttingen, Büsgenweg 2, 37077 Göttingen, Germany

*Present address:*
FAO/UNDP Project RAS/91/004 ERDB Building, P. O. Box 157, College, Laguna 4031, Philippines

$$= 1 - \left[ \sum_i \left( \sum_j \sum_l c_j \cdot c_l \cdot p_i(j) \cdot p_i(l) \right) \right]$$

$$= 1 - \left[ \sum_j \sum_l c_j \cdot c_l \sum_i p_i(j) \cdot p_i(l) \right]$$

$$= 1 - \left[ \sum_j \sum_l c_j \cdot c_l \sum_i \left( \frac{p_i(j)^2 + p_i(l)^2}{2} - \frac{(p_i(j) - p_i(l))^2}{2} \right) \right]$$

$$= 1 - \left[ \frac{1}{2} \left( \sum_j \sum_l c_j \cdot c_l \sum_i p_i(j)^2 + \sum_j \sum_l c_j \cdot c_l \sum_i p_i(l)^2 \right) \right]$$

$$+ \sum_j \sum_l c_j \cdot c_l \sum_i \frac{(p_i(j) - p_i(l))^2}{2}$$

$$= 1 - \sum_j c_j \cdot \sum_i p_i(j)^2 + \sum_j \sum_l c_j \cdot c_l \cdot d(j,l) \quad (1)$$

where $d(j,l) = \frac{1}{2} \sum_i (p_i(j) - p_i(l))^2$

$$= \sum_j c_j \cdot \delta_T(j) + \sum_j \sum_l c_j \cdot c_l \cdot d(j,l) \quad (2)$$

$$= \bar{\delta}_T + \bar{D} \quad (3)$$

where $\bar{\delta}_T = \sum_j c_j \cdot \delta_T(j)$ and $\bar{D} = \sum_j \sum_l c_j \cdot c_l \cdot d(j,l)$

The quantity $d(j,l)$ was termed the "minimum genetic distance" (Nei 1987, p 219), since it can supposedly measure a minimum number of codon differences between two populations. $\bar{D}$ is a weighted average of minimum genetic distances between population for all $n^2$ pairwise combinations of subpopulations, where each distance is weighted by the product of the relative proportions of the two subpopulations within the pooled population ($\sum_i \sum_l c_j c_l = 1$). Hence, $\bar{D}$ may be interpreted as a measure of genetic differentiation among subpopulations. As $\bar{\delta}_T$ is a weighted average of genetic differentiation within each subpopulation, the calculation proves the possibility to partition $\delta_T(g)$ into a component of genetic differentiation within subpopulations and a component of genetic differentiation among subpopulations.

Since $d(j,l)$ is symmetrical [$d(j,l) = d(l,j)$] and since $d(j,j) = 0$, Eq. 2 may be rewritten for ease of computation as

$$\delta_T(g) = \sum_{j=1}^{n} c_j \cdot \delta_T(j) + 2 \cdot \sum_{j=1}^{n} \sum_{l=j+1}^{n} c_j \cdot c_l \cdot d(j,l).$$

## Discussion

Obviously, it holds that $d(j,l)$ cannot be negative and $d(j,l) = 0$ if and only if the allelic structures of the supopulations $j$ and $l$ are identical at the respective gene locus. Hence, $\bar{D}$ is nonnegative. The genetic differentiation within the pooled population, $\delta_T(g)$, is larger than the weighted average of genetic differentiation within subpopulations, $\bar{\delta}_T$, if at least two subpopulations are genetically distinct. The genetic differentiation of the pooled populations equals the differentiation within supopulations only if there is no differentiation among subpopulations.

Numerical values of Nei's minimum genetic distance $d_m = d(j,l)$ are for the same set of frequency data consistently smaller than most other frequently computed distance measures such as Nei's standard and maximum genetic distances (Nei 1987, p 220), Roger's distance (Wright 1978, p 91), which is the square root of $d(j,l)$, and the distance $d_0$ (Gregorius 1984), which is identical to the "Prevosti"-distance (Nei 1987, p 210). $\bar{D}$, the contribution of the genetic differentiation within the pooled population that is due to differentiation among subpopulations, is a weighted average of minimum genetic distances. Hence, the relative weights of genetic differentiation within subpopulations versus genetic differentiation among subpopulations for the total genetic differentiation $\delta_T(g)$ should be assessed by comparisons of $\bar{\delta}_T$ and $\bar{D}$, but not rely upon other differentiation measures.

The partitioning of total gene differentiation into a component due to differentiation within subpopulations and one due to differentiation among subpopulations is the basis of the $G_{ST}$-statistics of Nei (1973), which is equivalent to Wright's $F_{ST}$-statistics in computation (Wright 1969, p 295). The strict correspondence of the above computation to Nei's $G_{ST}$ (Nei 1973) is easily shown, if all subpopulations are equally weighted ($v_{j=1}^{n} c_j = 1/n$). For this special case holds $H_T = \delta_T(g)$, $H_S = \bar{\delta}_T$, and $D_{ST} = \bar{D}$.

Nei (1977, 1987, p 160, p 188, respectively) claims that equal weights of all subpopulations are a resonable presumption in most instances as population sizes are transitory and geneticists frequently disregard the effect of population size. However, the disregard of population sizes in most population genetic studies on genetic differentiation among subpopulations is presumbly due to difficulties in their estimation. As mentioned before, the consideration of relative population sizes is of interest to breeders and conservationists, who are concerned with the creation of pooled populations from previously disjoint populations. The significance of considering population sizes for the analysis of genetic differentiation among natural populations is stressed by the concluding numerical example.

## Numerical example

Genetic structures of the five natural populations of *Pinus radiata*, a North-American pine, were determined at 31 isozyme gene loci by Moran et al. (1988). Allelic frequency data and estimations of population sizes were

taken from their publication; measures of genetic differentiation were recomputed from Table 2 of their paper. The population sizes of all populations are sufficient so as to be able to disregard the correction for small populations recommended by Gregorius (1987), i.e., to compute the gene differentiation as $1-\sum_i p_i^2$ for single gene loci.

The natural distribution of *P. radiata* is confined to three comparatively large mainland populations (Año Nuevo, Monterey, Cambria) in California and two much smaller Mexican island populations (Cedros and Guadalupe; see Table 1). Allelic differentation within each population was computed as a mean over 13 isozyme gene loci, for which the frequency of the most frequent allele was $< 95\%$ in at least one population, and over all investigated 31 gene loci, i.e., including 14 polymorphic gene loci, for which the frequency of the same dominant type was $\geq 95\%$ in all populations, and for 4 monomorphic gene loci (Table 1). Mean allelic differentiation within populations was computed as the arithmetic mean over single locus differentiations (Gregorius 1987; Nei 1987, p 179). Results of the partitioning of genetic differentiation are presented in Table 2. Columns 1–4 show the results of the analysis of gene differentiation following Eq. 2 i.e., considering the different relative population sizes. Columns 5–8 show the results giving each population the same weight. This procedure is equivalent of the conventional analysis of "gene diversity" following Nei (1973). Column 8 ($G_{ST}=D_{ST}/H_T$) is interpreted as the proportion of genetic differentiation due to differentiation among populations. The same quantity was computed in column 4 ($\bar{G}=\bar{D}/\delta_T(g)$) taking into account the differing population sizes.

Genetic differentiation at isozyme gene loci among conifer populations is usually low; mean $G_{ST}$-values above 0.10 have been exceptions (El-Kassaby 1991). Hence, the $G_{ST}$ values (column 8) are extraordinarily high. However, "a high proportion of this interpopulation diversity is due to differences between the island and mainland populations", as noted by Moran et al. (1988). Since the relative population sizes of the island populations are small, the large differentiation among populations vanishes if the differing population sizes are considered (compare columns 4 and 8 of Table 2).

A breeder or conservationist who wants to establish a genetically diversified population may wish to know the effect of an overrepresentation of the small and peripheral island populations on the genetic differentiation within an artificially established population. Since the mean $H_T$ is clearly larger than the mean $\delta_T(g)$, and equal representation of all five populations within a pooled population would clearly increase its genetic differentiation, if compared to a representation of each population according to its population size. The partitioning of genetic differentiation shows that this effect is not caused by a higher average differentiation within populations due to the change in the relative weights of the populations, since the mean $\bar{\delta}_T$ is nearly the same as the mean $H_S$. However, the comparison of columns 3 and 7 reveals that the differentiation among populations is considerably increased if all populations are equally weighted.

**Table 1** The five natural populations of *Pinus radiata*: estimated population sizes ($N$), relative population sizes ($c_j=N_j/N_{\text{total}}$), and differentiation $\delta_T(j)$ within populations [mean over 13 polymorphic isozyme gene loci (95% criterion) and mean over 31 gene loci]. Data were taken from Moran et al. (1988)

| Population | $N$ | $c_j$ | $\delta_T(j)$ | |
|---|---|---|---|---|
| | | | 31 loci | 13 loci |
| Año Nuevo | 1 200 000 | 0.2150396 | 0.0905 | 0.1948 |
| Monterey | 3 000 000 | 0.5375990 | 0.0973 | 0.1916 |
| Cambria | 1 300 000 | 0.2329595 | 0.1121 | 0.2511 |
| Cedros | 80 000 | 0.0143360 | 0.0976 | 0.2177 |
| Guadalupe | 368 | 0.0000659 | 0.0939 | 0.2118 |

**Table 2** Analysis of gene differentiation within and among the five natural populations of *Pinus radiata*: Means over 13 polymorphic gene loci (95% criterion) and 31 gene loci

| | 1 $\delta_T(g)$ | 2 $\bar{\delta}_T$ | 3 $\bar{D}$ | 4 $\bar{G}$ | 5 $H_T$ | 6 $H_S$ | 7 $D_{ST}$ | 8 $G_{ST}$ |
|---|---|---|---|---|---|---|---|---|
| 13 loci | 0.217 | 0.207 | 0.011 | 0.051 | 0.258 | 0.213 | 0.045 | 0.173 |
| 31 loci | 0.104 | 0.099 | 0.005 | 0.048 | 0.117 | 0.098 | 0.019 | 0.162 |

## References


El-Kassaby YA (1991) Genetic variation within and among conifer populations: review and evaluation of methods. In: Fineschi S, Cannata F, Hattemer HH (eds) Biochemical markers in the population genetics of forest tress. SPB Academic Publ, The Hague, pp 61–76

Gregorius H-R (1984) A unique genetic distance. Biom J 26:13–18

Gregorius H-R (1987) The relationship between the concepts of genetic diversity and differentiation. Theor Appl Genet 74: 397–401

Moran GF, Bell JC, Eldridge KG (1988) The genetic structure and the conservation of the five natural populations of *Pinus radiata*. Can J For Res 18:506–514

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. Ann Hum Genet 41:225–233

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Wright S (1969) Evolution and the genetics of populations, vol 2: the theory of gene frequencies. University of Chicago Press, Chicago London

Wright S (1978) Evolution and the genetics of populations, vol 4: variability within and among natural populations. University of Chicago Press, Chicago London