Theor Appl Genet (1999) 98: 664–668

R. Finkeldey · O. Murillo

Contributions of subpopulations to total gene diversity

Received: 15 August 1998 / Accepted: 8 September 1998

Abstract The concept of the partitioning of genetic diversity into a component within and a component among populations (F_{ST}- or G_{ST}-statistics) can be easily expanded to compute the contribution of single subpopulations to total gene diversity. A subpopulation contributes to total gene diversity with its single-population gene diversity plus the (weighted) mean of Nei's minimum genetic distances to all subpopulations. The suggested method allows one to unambiguously rank subpopulations according to the amount they contribute to the total gene diversity. Genetic polymorphisms at four isozyme gene loci of *Alnus acuminata* in Costa Rica are used to illustrate the procedure and its biological interpretation.

Key words Gene diversity · Genetic differentiation · Population structure · *Alnus acuminata*

Introduction

Diversity at discretely varying traits refers to the existence of different types, e.g. trait expressions, in a set of data. The frequency distribution of types contains the complete information on the variation at a trait. A great variety of measures are in use to condense this information to a single value which characterises the diversity within the set. The following considerations mainly apply to the partitioning of allelic diversity within subdivided populations, but it is also possible to compute the contributions of single subsets to total diversity for any other genetic or non-genetic frequency distribution, e.g. frequencies of genotypes or species.

The term $1 - \sum_i p_i^2$ is one of the most simple, useful, and frequently applied diversity measures in genetics. It is intuitively appealing since it can be interpreted as the probability of non-identity of two randomly selected individuals from an infinite population. It equals the proportion of heterozygotes of a population in Hardy-Weinberg structure if the p_is represent allele frequencies, and thus has been defined as the "expected heterozygosity" (H_e). It was also defined as the "genetic differentiation within an effectively infinite population" (Gregorius 1987) or simply as "gene diversity" (Nei 1973).

If a population is divided into an arbitrary number of subpopulations it is possible to partition the total gene diversity into a component due to diversity within subpopulations and a component due to differentiation among subpopulations. Wright (1965) suggested a method for the partitioning of gene diversity based on fixation indices and correlations between uniting gametes (F-statistics). The procedure of Nei (1973) is conceptually different and does not take into consideration genotype frequencies or observed heterozygosities, but results in identical numerical values for the component of gene diversity due to differentiation among populations ($G_{ST} = F_{ST}$). A simple derivation of the partitioning of gene diversity has been provided by Finkeldey (1994).

F_{ST} or G_{ST} are the most frequently used measures to quantify genetic differentiation among (sub)populations. A single value is computed for each gene locus, which quantifies the amount of gene diversity due to differentiation among populations relative to the total gene diversity. We will show that the concept can be easily expanded to unambiguously determine the contribution of each single subpopulation to the total gene diversity.

Communicated by P. M. A. Tigerstedt

R. Finkeldey (✉)
Swiss Federal Institute for Forest, Snow and Landscape Research, Zürcherstrasse 111, CH-8093 Birmensdorf, Switzerland

O. Murillo
Escuela de Ingeniería Forestal, Instituto Tecnológico de Costa Rica, Apartado 159, 7050 Cartago, Costa Rica

Computation of subpopulation contributions

The following definitions are made following the notion of Finkeldey (1994): population G is 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 ith allele in the jth [lth] subpopulation and $p_i(t)$ denote the relative frequency of the ith allele in the pooled population G [$\sum_i p_i(j) = \sum_i p_i(l) = \sum_i p_i(t) = 1$]. The gene diversities of subpopulations are computed as $H(j) = 1 - [\sum_i p_i(j)^2]$. The total gene diversity $H_T = \delta_T(g)$ of the population may be partitioned according to Finkeldey (1994) as follows:

$$H_T = 1 - \left[\sum_i p_i(t)^2\right] = \sum_j H(j) + \sum_j \sum_l c_j \cdot c_l \cdot d(j,l), \quad (1)$$

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

Equation (1) may be rewritten as

$$H_T = \sum_j \left[c_j H(j) + c_j \sum_l c_l \cdot d(j,l) \right]. \quad (2)$$

Following the notion established by Nei (1973) the total gene diversity due to variation within populations is $H_s = \sum_j c_j H(j)$ and the diversity due to differentiation among populations is $D_{ST} = \sum_j c_j \sum_l c_l \cdot d(j,l)$. The term $d(j,l)$ has been defined as the minimum genetic distance between two populations (Nei 1987, p 219).

Equation (2) allows us to define the contribution of population j to the gene diversity within populations as $H_s(j) = c_j H(j)$, and the contribution of the same population to D_{ST} as $D_{ST}(j) = c_j \sum_l c_l d(j,l)$. The overall contribution of population j to the total gene diversity is computed by the summation of $H_S(j)$ and $D_{ST}(j)$, i.e. $H_T(j) = c_j[H(j) + \sum_l c_l \cdot d(j,l)]$.

The relative contributions of population j to the gene diversity within populations [$C_S(j)$], the genetic differentiation among populations [$C_{ST}(j)$], and the total gene diversity [$C_T(j)$], are computed as:

$$C_S(j) = H_S(j)/H_S, \quad (3)$$

$$C_{ST}(j) = D_{ST}(j)/D_{ST}, \quad (4)$$

$$C_T(j) = H_T(j)/H_T. \quad (5)$$

It obviously holds that $\sum_j C_S(j) = \sum_j C_{ST}(j) = \sum_j C_T(j) = 1$.

It is frequently biologically meaningful to disregard the effects of different population sizes, i.e. to weight populations equally ($v_{j=1}^n \; c_j = c_l = 1/n$). In this case equation (2) simplifies to

$$H_T = \sum_j \left[H(j)/n + \sum_l d(j,l)/n^2 \right]. \quad (6)$$

The computation of single population contributions to gene diversity is easily expanded to an arbitrary number of gene loci. Arithmetic means over all gene loci of all gene diversity and differentiation measures, i.e. of $H_S(j)$, $D_{ST}(j)$ and $H_T(j)$, replace the corresponding single-locus values for all populations, and the relative contributions to gene diversity and differentiation [$C_S(j)$, $C_{ST}(j)$ and $C_T(j)$] are computed based on these mean values.

Discussion

The contribution of a single subpopulation to the total gene diversity depends on its size relative to all other subpopulations, the gene diversity within the subpopulation, and a measure of genetic differentiation to all other subpopulations. The significance of within-subpopulation diversity for the contribution to total gene diversity is intuitively plausible and its computation is straightforward. Thus, only the effects of relative population sizes and gene differentiation are discussed.

$H(j)$ is a measure of the gene diversity only for effectively infinite populations. The corresponding measure of the "gene diversity" of a finite population is given by $(n_j/n_j - 1)\cdot(1 - \sum_i p_i^2)$ (n_j = size of population j) (Gregorius 1987). Thus, equation (2) refers to a set of subpopulations which all are large enough to disregard the effect of finite population sizes (e.g. larger than 50 individuals; Nei 1987, p 191), but which may still differ with regard to their relative sizes.

However, different population sizes are frequently disregarded either because they are difficult to estimate or because the addressed problem requires an equal representation of all subpopulations. Thus, it is frequently justified to disregard population sizes and to partition the total gene diversity according to equation (6).

The (weighted or arithmetic) average of Nei's minimum genetic distances to all other subpopulations is an appropriate measure for an assessment of the contribution of single populations to the total gene diversity due to population differentiation. It holds, obviously, that $d(j,j)$ is 0. Thus, Nei (1987, p 163 and p 190) suggested to disregard the $d(l,l)$s and to calculate the arithmetic mean over all minimum distances $d(j,l)$ with $j \neq l$ as a measure of absolute gene differentiation. However, this procedure is not appropriate if it is aimed to quantify the contribution of single populations to total gene diversity.

Gregorius and Roberds (1986) developed a measure of genetic differentiation based on the genetic distance $d_0 = 1/2 \sum_i [p_i(j) - p_i(l)]$ (Gregorius 1984) which quantifies the total genetic differentiation among subpopulations (δ) and the contributions of single subpopulations (D_j) to genetic differentiation. This procedure is conceptually related to the measurement of genetic diversity

(Gregorius 1987). However, D_j-values are absolute measures of genetic distances of populations to their respective complements and are thus computationally unrelated to the gene diversity within subpopulations. The method does not allow one to partition the total gene diversity into contributions of subpopulations which take into account and relate both gene diversity within subpopulations and gene differentiation among subpopulations.

The main advantage of our procedure is the possibility to unambiguously determine the contribution of single populations to total gene diversity. This allows us to rank populations according to their relative contributions to total gene diversity. Furthermore, it is possible to separate the contribution of each subpopulation into a component due to its gene diversity and a component due to gene differentiation with respect to all other subpopulations.

This concept is based solely on an analysis of frequency distributions. Thus, it may be used for any data set composed of at least two frequency distributions of the same type. Applications will usually be based on the allelic frequency distributions of populations. However, allelic frequency distributions may not only be inferred from populations or subpopulations, but also from other demes. For example, the method allows one to rank the genetically effective pollen clouds of single-seed trees (e.g. Finkeldey 1995) according to their contribution to the diversity of the total pollen cloud. Apart from allele frequencies the method is also applicable to other types of frequency distributions, e.g. frequencies of uniparentally inherited haplotypes or phenotypes controlled by dominant gene markers.

The F-statistic was originally developed to describe and analyse patterns of genetic variation in populations composed of an effectively infinite number of subpopulations (Nei 1977). The procedure outlined here is most informative if a taxonomic unit (e.g., a species, a subspecies or a variety) consists of only a few populations or subpopulations. Ideally, allelic frequency distributions at one or several marker loci should be known for all existing subpopulations. Data of this type become increasingly available due to a focus of population genetic and conservation genetic studies on rare or locally common species (see for example the numerical example that follows).

It is possible to rank populations according to their relative contributions to the total gene diversity. However, it is usually not advisable to use only this information for the selection of genetic resources, e.g. by selecting the populations which contribute most to total gene diversity. Selection criteria for genetic resources require a careful analysis of the observed traits and their properties. Furthermore, the diversity measure employed is rather insensitive to locally occurring rare alleles. Thus, genetic information may be lost even at the observed gene markers if the selection of populations is only based on their contributions to gene diversity.

Numerical example

Alnus acuminata spp. *arguta* (Schlectendal) Furlow is a broad-leaved tree species with a disjunct distribution range in Central America. A genetic inventory was performed in 17 populations covering the main distribution areas of the species in Central America (Murillo 1997). Eighteen out of twenty two investigated isozyme loci proved to be monomorphic. The allelic structures of the four polymorphic isozyme loci are listed in Table 1.

Table 1 Allelic structures at four isozyme gene loci in 17 populations of *A. acuminata* spp. *arguta* in Central America. N: sample size (no. of investigated trees)

No.	Population	N	*PGI-B*			*PGM-A*			*IDH-A*		*MNR-A*	
			B_1	B_2	B_3	A_1	A_2	A_3	A_1	A_2	A_1	A_2
1	Zarcero	62	0.09	0.75	0.16	0.20	0.00	0.80	0.83	0.17	1.00	0.00
2	Bajos Toro	61	0.03	0.93	0.04	0.11	0.06	0.83	0.84	0.16	1.00	0.00
3	Vara Blanca	61	0.16	0.72	0.12	0.21	0.00	0.79	0.57	0.43	0.99	0.01
4	Cartagos	61	0.02	0.75	0.23	0.06	0.00	0.94	0.58	0.42	0.98	0.02
5	Coronado A	43	0.00	0.82	0.18	0.13	0.00	0.87	0.51	0.49	0.95	0.05
6	Coronado B	60	0.00	0.85	0.15	0.01	0.00	0.99	0.60	0.40	0.97	0.03
7	Llano Grande	60	0.00	0.79	0.21	0.04	0.00	0.96	0.52	0.48	1.00	0.00
8	Irazú	48	0.00	0.99	0.01	0.02	0·00	0.98	0.60	0.40	1.00	0.00
9	Pacayas	62	0.00	0.88	0.12	0.03	0·00	0.97	0.65	0.35	1.00	0.00
10	Turrialba	50	0.00	0.85	0.15	0.02	0.00	0.98	0.25	0.75	1.00	0.00
11	El Empalme	63	0.00	0.82	0.18	0.06	0.00	0.94	0.82	0.18	0.98	0.02
12	Cañón	61	0.00	0.75	0.25	0.05	0.00	0.95	0.78	0.22	0.97	0.03
13	Copey	63	0.00	0.66	0.34	0.18	0.00	0.82	0.93	0.07	0.99	0.01
14	San Gerardo	48	0.00	0.63	0.37	0.14	0.00	0.86	0.74	0.26	0.99	0.01
15	Siberia	31	0.00	0.66	0.34	0.12	0.00	0.88	0.91	0.09	0.98	0.02
16	División	61	0.00	0.82	0.18	0.11	0.00	0.89	0.93	0.07	0.98	0.02
17	Boquete	35	0.00	0.54	0.46	0.34	0.04	0.62	0.76	0.24	1.00	0.00

Table 2 Relative contribution of subpopulations to the gene diversity within populations [$C_S(j)$], relative contribution of subpopulations to gene diversity among populations [$C_{ST}(j)$], and relative contribution of subpopulations to the total gene diversity [$C_T(j)$]), for four polymorphic isozyme systems and the gene pool in 17 populations of *A. acuminata* spp. *arguta*

Population	*PGI-B*			*PGM-A*			*IDH-A*			*MNR-A*			Gene pool		
	$C_S(j)$	$C_{ST}(j)$	$C_T(j)$	$C_S(j)$	$C_{ST}(j)$	$C_T(j)$	$C_S(j)$	$C_{ST}(j)$	$C_T(j)$	$C_S(j)$	$C_{ST}(j)$	$C_T(j)$	$C_S(j)$	$C_{ST}(j)$	$C_T(j)$
Zarcero	0.073	0.038	0.070	0.101	0.059	0.098	0.046	0.046	0.046	0.000	0.054	0.001	0.066	0.046	0.064
Bajos Toro	0.024	0.085	0.029	0.093	0.041	0.089	0.044	0.049	0.045	0.000	0.054	0.001	0.046	0.057	0.047
Vara Blanca	0.079	0.063	0.078	0.105	0.065	0.102	0.080	0.044	0.075	0.046	0.031	0.046	0.084	0.052	0.081
Cartagos	0.069	0.031	0.066	0.036	0.039	0.036	0.079	0.042	0.074	0.092	0.037	0.091	0.067	0.039	0.064
Coronado A	0.053	0.032	0.052	0.071	0.031	0.068	0.082	0.061	0.079	0.222	0.234	0.222	0.073	0.050	0.071
Coronado B	0.046	0.039	0.045	0.006	0.066	0.011	0.078	0.038	0.072	0.136	0.073	0.135	0.053	0.043	0.052
Llano Grande	0.060	0.030	0.057	0.024	0.048	0.026	0.081	0.058	0.078	0.000	0.054	0.001	0.059	0.049	0.058
Irazú	0.004	0.121	0.012	0.012	0.059	0.016	0.078	0.038	0.072	0.000	0.054	0.001	0.035	0.062	0.038
Pacayas	0.038	0.049	0.039	0.018	0.053	0.021	0.074	0.031	0.068	0.000	0.054	0.001	0.047	0.039	0.047
Turrialba	0.046	0.039	0.045	0.012	0.059	0.016	0.061	0.214	0.084	0.000	0.054	0.001	0.044	0.146	0.055
El Empalme	0.053	0.032	0.052	0.036	0.039	0.036	0.048	0.044	0.048	0.092	0.037	0.091	0.049	0.040	0.048
Cañón	0.067	0.033	0.065	0.030	0.043	0.031	0.056	0.036	0.053	0.136	0.073	0.135	0.057	0.036	0.055
Copey	0.081	0.065	0.080	0.093	0.047	0.089	0.021	0.081	0.030	0.046	0.031	0.046	0.059	0.071	0.060
San Gerardo	0.084	0.083	0.084	0.076	0.033	0.073	0.063	0.031	0.058	0.046	0.031	0.046	0.073	0.045	0.070
Siberia	0.081	0.065	0.080	0.067	0.030	0.064	0.027	0.072	0.034	0.092	0.037	0.091	0.056	0.064	0.057
División	0.053	0.032	0.052	0.062	0.030	0.059	0.021	0.081	0.030	0.092	0.037	0.091	0.043	0.061	0.045
Boquete	0.089	0.162	0.095	0.157	0.259	0.165	0.060	0.033	0.056	0.000	0.054	0.001	0.089	0.100	0.090

Conventional partitioning of gene diversity shows that the proportion of genetic variation due to differentiation among populations ($F_{ST} = G_{ST}$) is 0.076 for *PGI-B*, 0.080 for *PGM-A*, 0.149 for *IDH-A*, 0.015 for *MNR-A*, and 0.106 for the gene pool.

Contributions of single populations to the total gene diversity were computed for each of the polymorphic loci and the gene pool (four loci). The relative contributions of subpopulations to the gene diversity within populations, to the gene diversity among populations, and to the total gene diversity are presented in Table 2.

Population Boquete contributes most to the total gene diversity at the *PGI-B* gene locus. The frequency of allele B_3 is highest in Boquete which accounts for the highest contribution to within-population diversity and to gene differentiation. Population Irazú is almost fixed for the frequent allele B_2. As expected, the contribution of Irazú to within-population diversity and to the total diversity is lowest. However, Irazú is second highest with regard to its contribution to gene differentiation, due to the unusually low frequency of allele B_3. Population Bajos Toro ranks second lowest in its contribution to the total gene diversity due to the dominance of allele B_3, although it is one of the four populations which exhibits all three alleles at the *PGI-B* gene locus.

Boquete also contributes most to the diversity at the *PGM-A* locus. This population has the lowest frequency of the dominating allele A_3. The allele A_2, which was observed only in two populations, occurs at a low frequency in Boquete.

Population Turrialba contributes most to the total gene diversity at the *IDH-B* gene locus although its contribution to the diversity within populations is only average. The frequency of the allele A_1, which dominates in all other populations, is only 25% in Turrialba. Thus, this population clearly contributes most to the gene differentiation among populations.

Variation is low at the *MNR-A* gene locus. The frequency of the rare allele A_2 is highest in population Coronado A, which consequently contributes most to the diversity observed at this locus.

The gene pool values show that there are pronounced differences with regard to the contributions of single populations to the gene diversity at the four polymorphic gene loci, although the F_{ST}-value of 0.106 indicates only moderate differentiation. Population Boquete contributes 9.0% to the total diversity of the gene pool, but population Irazú contributes only 3.8% (cf. Table 2, last column), i.e. only 42.2% of that of Boquete. The contributions of single populations to gene diversity within populations [$H_S(j)$], gene diversity among populations [$D_{ST}(j)$], and the total gene diversity [$H_T(j)$] of the gene pool (four loci), are illustrated in Fig. 1.

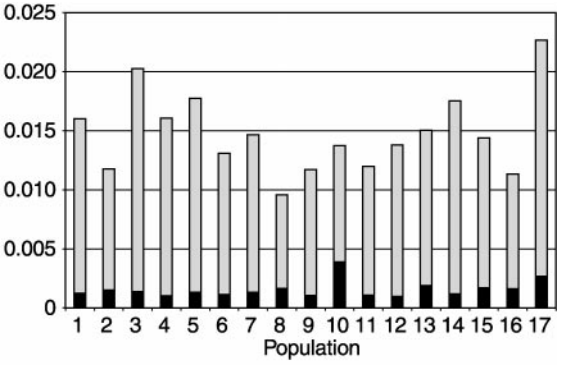

Fig. 1 Contributions of single populations to the gene diversity among populations ($D_{ST}(j)$; *black bars*) to the gene diversity within populations ($H_S(j)$; *grey bars*) and the the total gene diversity ($H_T(j) = H_S(j) + D_{ST}(j)$) for 17 populations of *A. acuminata*

References

Finkeldey R (1994) A simple derivation of the partitioning of genetic differentiation within subdivided populations. Theor Appl Genet 89 : 198–200

Finkeldey R (1995) Homogeneity of pollen allele frequencies of single-seed trees in *Picea abies* (L.) Karst. plantations. Heredity 74 : 451–463

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

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

Gregorius H-R, Roberds JH (1986) Measurement of genetical differentiation among subpopulations. Theor Appl Genet 71 : 826–834

Murillo O (1997) Genetische Untersuchungen an natürlichen Populationen von *Alnus acuminata* spp. *arguta* (Schlectendal) Furlow in Costa Rica und Panamá. Cuvillier Verlag, Göttingen, Germany

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 (1965) The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19 : 395–420