Aquilo Ser. Zool. 20: 111–117. 1979

Estimation of genetic variation by the components of gene diversity

RISTO H. ALATALO & RAUNO V. ALATALO (1979)

Department of Genetics, University of Oulu, Oulu, Finland

Received 15. VIII. 1978

Alatalo, R. H. & Alatalo, R. V. 1979: Estimation of genetic variation by the components of gene diversity. — Aquilo, Ser. Zool. 20: 111–117.

Genetic variation is apportioned to components due to variation within and between populations or samples (years etc.) using the antilogarithm of Shannon's entropy as index of gene diversity. Differences in allele frequencies between samples are tested by a likelihood-ratio test, G-test. This test requires pooling of samples instead of proportions for calculation of average diversities. Without testing it is possible to partition diversity by pooling proportions, which gives equal weight to each sample. G-test also requires that natural logarithms are used for calculation of Shannon's entropy. Formulae of gene diversity are given for a two-dimensional measurement, say annual and between-population variation. Diversity is particularly suitable for the measurement of association in contingency tables, but these associations are tested as a by-product. Several loci can be analyzed simultaneously, because likelihood-ratios, degrees of freedoms, and gene diversities are additive: they can be summed over many loci.

I. Introduction

Genetic variation in a population has been often measured by the proportion of polymorphic loci and the average heterozygosity per locus. However, the classification of polymorphic locus is based on an artificial limit (say the frequency of the commonest allele equal or less than 0.95). On the other hand, genetic variation can be measured unambiguously. Nei (1973, 1975) suggested average homozygosity and heterozygosity as measures of genetic variation within and between populations. Kimura and Crow (1964) proposed the effective number of alleles per locus, which is the reciprocal of homotsygosity in a population. Further, Lewontin (1972) used Shannon's entropy (i.e. diversity index) to apportion genetic variation within and between populations.

We introduce a method modified and extended from the Lewontin's (1972) method. We give also the connection between the Lewontin's method and the likelihood-ratio test, which uses chi-square distribution.

II. Shannon's entropy as an index of gene diversity

Lewontin (1972) partitioned Shannon's entropy without weighting samples (pooled proportions), but stated that weighted averaging (pooled samples) should decrease total and between-

population diversity. Allan (1975) gave formulae with weights and they are included in following notation.

Shannon's entropy is related to likelihood-ratio tests used for contingency tables (i.e. G-test, see Sokal and Rohlf 1969). G-test requires that natural logarithms are used for the calculation of diversity and averages are calculated weighting each sample by its relative size (cf. Colwell 1974). This paper gives a modified and extended partition of Shannon's entropy (cf. Lewontin 1972, Allan 1975, Alatalo & Alatalo 1977).

Suppose that there are s alleles at a locus, which is investigated from r populations during c years. Gene diversity for a sample from the ith population in the jth year is

$$H_{ij} = -\sum_k \frac{N_{ijk}}{N_{ij.}} ln \frac{N_{ijk}}{N_{ij.}}, \quad (1)$$

where N_{ijk} is the number of kth allele in jth year in ith population, and $N_{ij.}$ is the number of genes studied in jth year from ith population (in diploid case two times the number of individuals). Gene diversity for the total sample pooled over populations and years is

$$H_{...} = -\sum_k \frac{N_{...k}}{N_{...}} ln \frac{N_{...k}}{N_{...}}, \quad (2)$$

where $N_{...k}$ is the number of kth alleles in the pooled sample, and $N_{...}$ is the number of genes studied.

The average gene diversity within samples is

$$\bar{H}_{AB} = \sum_i \sum_j \frac{N_{ij.}}{N_{...}} (H_{ij}). \quad (3)$$

The average gene diversity within populations pooled over years is

$$\bar{H}_A = \sum_i \frac{N_{i..}}{N_{...}} (H_{i.}), \quad (4)$$

where $N_{i..}$ is the number of genes studied from the ith population pooled over years, and $H_{i.}$ is diversity in notation (1), which is calculated within the ith population pooling years. The average annual gene diversity pooling populations is

$$\bar{H}_B = \sum_j \frac{N_{.j.}}{N_{...}} (H_{.j}) \quad (5)$$

where $N_{.j.}$ is the number of genes studied in the jth year from pooled population and $H_{.j}$ i diversity in notation (1), which is calculatec within the jth year pooling populations.

III. Likelihood-ratio test for genetic variation

The likelihood-ratio test is based on the diversities defined in the former section. Test statistic G is distributed as chi-square, if the null hypothesis is true, and hence various associations can be tested. However, a test statistic like chi-square is not a measure of association (e.g. Poole 1974). On the other hand, test procedure derived from diversity is not always most powerful statistically speaking (see Margolin & Light 1974), but more elegant methods either require more complex calculation or do not measure association (e.g. Bishop *et al.* 1975). Diversity can be used particularly for the measurement of association (see section IV), and associations are tested as a by-product.

Null hypothesis, that annual populations have no genetic differences (i.e. no variability between years and populations), is tested by statistic

$$G_{AB} = 2N_{...}(H_{..} - \bar{H}_{AB}) \quad (6)$$

which under null hypothesis is distributed as chi-square with $(rc-1)(s-1)$ degrees of freedom.

Genetic variation between populations is tested by statistic

$$G_A = 2N_{...}(H_{..} - \bar{H}_A), \quad (7)$$

which under null hypothesis is distributed as chi-square with $(r-1)(s-1)$ degrees of freedom. Similarly, genetic variation between years is tested by statistic

$$G_B = 2N_{...}(H_{..} - \bar{H}_B), \quad (8)$$

which under null hypothesis is distributed as chi-square with $(c-1)(s-1)$ degrees of freedom.

Genetic variation between populations within years is tested by statistic

$$G_{A/B} = 2N_{...}(\bar{H}_B - \bar{H}_{AB}), \quad (9)$$

which under null hypothesis is distributed as chi-square with $(r-1)c(s-1)$ degrees of freedom. Analogously, genetic variation between years within populations is tested by statistic

$$G_{B/A} = 2N_{...}(\bar{H}_A - \bar{H}_{AB}), \quad (10)$$

which under null hypothesis is distributed as chi-square with $r(c-1)(s-1)$ degrees of freedom.

Several loci can be tested simultaneously by combining each locus adding corresponding test statistics and degrees of freedom. Degrees of freedom must be lowered, if there are any marginal zeroes (Bishop *et al.* 1975) arising from a missing allele in a population or a missing sample from some population in a year.

IV. Estimation of components of gene diversity

Standardized components are estimated by the antilogarithm of Shannon's entropy *exp (H)* (Alatalo & Alatalo 1977). However, the antilogarithmic transformation is made after averaging and not first before averaging as previously suggested. By this modification G-test is reached as a by-product and components calculated using alternative ways in multidimensional contingency tables are standardized more regularly (Alatalo & Alatalo MS).

The antilogarithm of Shannon's entropy *exp (H)* indicates how many equally common alleles are needed to produce the same diversity H as is present in the sample (cf. effective number of species, Peet 1974); hence its unit is "the number of equally common alleles". The number of alleles is easier to interpret than any information theoretical unit, say bit. This is one reason to use the antilogarithmic transformation (see Alatalo & Alatalo 1977).

Between-sample gene-diversity V_{AB} measures the extent to which samples (populations in each year) differ genetically from each other in the locus:

$$V_{AB} = \frac{exp(H_{..}) - exp(\bar{H}_{AB})}{exp(H_{..}) - 1}. \quad (11)$$

The complement to this is within-sample gene-diversity L_{AB}, which measures to what extent the genetic variation was expressed within samples relative to the total genetic variation observed in the locus:

$$L_{AB} = 1 - V_{AB} = \frac{exp(\bar{H}_{AB}) - 1}{exp(H_{..}) - 1}. \quad (12)$$

Between-population gene-diversity V_A estimates the degree of genetic differentiation in the locus between populations relative to the total variation:

$$V_A = \frac{exp(H_{..}) - exp(\bar{H}_A)}{exp(H_{..}) - 1}. \quad (13)$$

Exclusive-between-population gene-diversity $V_{A/B}$ measures the average extent to which populations have differentiated within each year:

$$V_{A/B} = \frac{exp(\bar{H}_B) - exp(\bar{H}_{AB})}{exp(H_{..}) - 1}. \quad (14)$$

Between-year gene-diversity V_B standardized to the total variation is

$$V_B = \frac{exp(H_{..}) - exp(\bar{H}_B)}{exp(H_{..}) - 1}. \quad (15)$$

Exclusive-between-year gene-diversity $V_{A/B}$ is a measure of annual genetic variation within populations on average:

$$V_{B/A} = \frac{exp(\bar{H}_A) - exp(\bar{H}_{AB})}{exp(H_{..}) - 1}. \quad (16)$$

Interaction component of gene diversity R_{AxB} measures the extent to which genetic differences between samples arise from a different variation between populations in each year:

$$R_{AxB} = \quad (17)$$

$$\frac{exp(\bar{H}_{AB}) - exp(\bar{H}_A) - exp(\bar{H}_B) + exp(H_{..})}{exp(H_{..}) - 1}.$$

R_{AxB} gives positive values, if years and populations indicate the same genetic differentiation between samples. A negative value arises from the annual variation in genetic differentiation between populations.

Components of diversity are given relative to the total diversity or standardized; hence components make an additive partition:

$$V_{A/B} + V_{B/A} + R_{AxB} + L_{AB} = V_A + V_B - R_{AxB} + L_{AB} = V_A + V_{B/A} + L_{AB} = 1. \quad (18)$$

The last partition is hierarchical, like those presented by Lewontin (1972) and Allan (1975). Consequently, the partition of diversity

in this paper is an extension of Lewontin's and Allan's method to multidimensional tables and at the same time a modification (owing to the antilogarithmic transformation). Further, we connect the partition of diversity with testing of components, and hence we call this method analysis of diversity (analogous with the analysis of variance).

V. Components of gene diversity in *Aph-3* locus of *Drosophila subobscura*

Partition of gene diversity is demonstrated by Saura's *et al.* (1973) data on *Aph-3* locus of *Drosophila subobscura* providing variation of allele frequencies both between-populations and annually. This two-dimensional measurement (population x year) is an example of multivariate analysis of diversity. Allele frequencies in each sample are given in Table 1.

In Kolmperä gene diversity of *Aph-3* locus is higher than in Helsinki and Tvärminne (Table 2). In 1969 gene diversity is higher than values observed in other years (Table 3) and then gene diversity is particularly high in the sample from Kolmperä (Table 1).

Table 1. Allele frequencies and gene diversities in *Aph-3* locus of *D. subobscura* (Saura *et al.* 1973).

Population	Sample	Alleles					Gene diversity	
Year	size	*90*	*97*	*100*	*107*	*111*	H	exp(H)
Helsinki								
1970	338	—	—	0.52	0.48	—	0.6923	1.9983
1971	18	—	—	0.50	0.44	0.06	0.8676	2.3811
1972	24	—	—	0.54	0.42	0.04	0.8293	2.2917
Kolmperä								
1969	20	0.40	—	0.55	—	0.05	0.8451	2.3282
1970	62	—	—	0.71	0.29	—	0.6024	1.8266
1971	56	—	0.02	0.48	0.50	—	0.7702	2.1602
1972	24	—	—	0.46	0.54	—	0.6897	1.9931
Tvärminne								
1969	26	—	—	0.65	0.35	—	0.6450	1.9061
1970	26	—	—	0.42	0.58	—	0.6813	1.9764
1971	38	—	—	0.37	0.63	—	0.6581	1.9311
1972	48	—	—	0.46	0.52	0.02	0.7780	2.1771

Table 2. Allele frequencies in *Aph-3* locus for each population pooling year.

Population	Sample	Alleles					Gene diversity	
	size	*90*	*97*	*100*	*107*	*111*	H	exp(H)
Helsinki	380	—	—	0.52	0.47	0.01	0.7212	2.0570
Kolmperä	162	0.05	0.01	0.57	0.36	0.01	0.8978	2.4543
Tvärminne	138	—	—	0.46	0.53	0.01	0.7289	2.0728

Table 3. Allele frequencies in *Aph-3* locus for each year pooling populations.

Year	Sample	Alleles					Gene diversity	
	size	*90*	*97*	*100*	*107*	*111*	H	exp(H)
1969	46	0.17	—	0.61	0.20	0.02	1.0088	2.7423
1970	426	—	—	0.54	0.46	—	0.6896	1.9929
1971	112	—	0.01	0.45	0.53	0.01	0.7787	2.1786
1972	96	—	—	0.48	0.50	0.02	0.7797	2.1809

Table 4. Analysis of diversity for allele frequencies in *Aph-3* locus of *D. subobscura*.

Component of gene diversity	exp(H)	%	Likelihood-ratio test		
			G	df	Probability
Total	2.2009	—	—	—	—
Within-population	2.1486	95.6	—	—	—
Within-year	2.0930	91.0	—	—	—
Within-sample	2.0265	85.5	—	—	—
Between-population	0.0523	4.4	32.78	8	5×10^{-6}
Between-year	0.1079	9.0	68.41	12	6×10^{-10}
Between-sample	0.1744	14.5	112.34	40	9×10^{-9}
Population Year	0.0665	5.5	43.93	15	0.0001
Year/Population	0.1221	10.2	79.56	22	2×10^{-8}
Population x year	−0.0142	−1.2	—	—	—

Analysis of diversity (Table 4) reveals that 14.5 % of gene diversity can be ascribed to samples collected in different years from Helsinki, Kolmperä and Tvärminne. This between-sample component arises from a statistically significant variation of allele frequencies between samples ($p=9\times10^{-9}$). 9.0 % of this variation is contributed by annual variation, which is also highly significant ($p=6\times10^{-10}$). Between-population variation is lower (4.4 %), although it is still significant ($p=5\times10^{-6}$). The average variation of allele frequencies annually in each population (10.2 %) is higher than the annual variation for pooled populations (9.0 %). This difference gives a negative interaction component (−1.2 %).

The annual variation of allele frequencies is highest in Kolmperä (Table 5). In Tvärminne annual variation is higher than in Helsinki, although these between-year components are not statistically significant. The between-population variation of allele frequencies was exceptionally high in 1969. There is one sample with exceptional allele frequencies in 1969 from Kolmperä (Table 1). In 1971 and 1972 allele frequencies had no statistically significant differences between populations (Table 6).

Table 5. Annual variation of allele frequencies in *Aph-3* locus of *D. subobscura* for each population.

Population	exp(H) %	Likelihood-ratio test		
		G	df	Probability
Helsinki	2.3	9.04	4	0.06
Kolmperä	29.8	63.02	12	6×10^{-9}
Tvärminne	5.2	7.51	6	0.28
Year/Population	10.2	79.56	22	2×10^{-8}

Table 6. Between-population variation of allele frequencies in *Aph-3* locus of *D. subobscura* for each year.

Year	exp(H) %	Likelihood-ratio test		
		G	df	Probability
1969	38.1	25.47	3	0.00001
1970	2.2	9.46	2	0.009
1971	5.6	6.92	6	0.33
1972	2.0	2.09	4	0.72
Population/Year	5.5	43.93	15	0.0001

VI. Discussion

Statisticians have not reached a generally accepted standard method for the measurement of association in a multidimensional contingency table. Different methods are needed for each type of measurement problem (Bishop *et al.* 1975). For example, chi-square test can be used to determine whether a certain association is probable or not, but the test never measures the intensity of association (Poole 1974). Consequently, analyses for discrete data are divided into two groups: they measure associations or they test them (see Bishop *et al.* 1975).

Mitton (1977) suggested the simultaneous use of many loci to estimate the degree of differentiation between populations. However, this may sometimes be unpractical, if several polymorphic loci are studied and many multilocus genotypes are observed only once. Then

it is more practical to analyze each locus separately; hence likelihood-ratios, degrees of freedom and gene diversities are additive. The sums of these can be used for testing and estimation the variation of all gene loci of a species.

Averaging of gene diversities is presented using weight for each sample. On the contrary, if samples are considered equivalent, weights are omitted. In this case total and between-population diversities are usually higher than corresponding values with weights (Lewontin 1972). Weighting is essential for the likelihood-ratio test, because it requires pooling of the number of alleles instead of allele frequencies. The estimation of components gives more unambiguous results, if samples are relative to population size, and weighting does not arise from an arbitrary sample size.

To measure differentiation between populations for their gene pools the loci studied should be a random sample of all loci (Nei 1975). In this case monomorphic loci are also present and they have no between-population variation; hence within-population component is 100 % for a monomorphic locus (Järvinen *et al.* 1976). Monomorphic loci are not useful for the partition of genetic variation, but if the number of substitutions per locus is to be estimated they should be included (Mitton 1977). On the other hand, partition of variation to within-population and between-population components is not the best approach, if only the differences in allele frequencies between populations are studied. For example, between-population component can be equal to within-population component, although populations have no alleles in common (Mitton 1977). In this case the genetic difference between populations is maximal, but between-population variation is only half of its maximum.

In general, diversity indices are more or less dependent on sample size (e.g. Caswell 1976). There are corrected formulae for Shannon's entropy, which increase diversity for small samples and usually lower between-sample diversity (see Järvinen & Sammalisto 1976, Atkinson & Schorrocks 1977). If the estimation of components is preferred, a corrected formula can be used, although it makes the likelihood-ratio test conservative. Further, this test is approximate in any case for small samples (Margolin & Light 1974). For large samples the corrected formula does not affect results markedly.

Assumptions required for a likelihood-ratio test are usually met with in a measurement of genetic variation. The most likely bias is dependence between successive samples, which in many case is overlook in a testing of biological data (Eberhardt 1976). Components of diversity require no assumptions about the population structure (Järvinen *et al.* 1976). However, the use of diversity is generally based on the assumption that random samples are used (Eberhardt 1976).

The study of genetic variation needs more sophisticated statistical methods, although progress in collection of samples and identification of alleles is a key problem when better results are desired. Statistical methods for the analysis of multidimensional contingency tables are under continuous development. The need to study partition of diversity and alternative methods seems to be urgent.

References

Alatalo, R. H. & Alatalo, R. V. 1979: Analysis of diversity: Pielou's, Nei's and Lewontin's methods revised. Manuscript, Dept. Gen., Univ. Oulu.

Alatalo, R. V. & Alatalo, R. H. 1977: Components of diversity: Multivariate analysis with interaction. — Ecology 58: 900—906.

Allan, J. D. 1975: Components of diversity. — Oecologia 18: 359—367.

Atkinson, W. & Shorrocks, B. 1977: Breeding site specificity in the domestic species of Drosophila. — Oecologia 29: 223—232.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. 1975: Discrete multivariate analysis. — 558 pp. Cambridge, Mass.

Caswell, H. 1976: Community structure: A neutral model analysis. — Ecol. Monogr. 46: 327—354.

Colwell, R. K. 1974: Predictability, constancy, and contingency of periodic phenomena. — Ecology 55: 1148—1153.

Eberhardt, L. L. 1976: Quantitative ecology and impact assessment. — J. environ. Mgmt 4: 27—70.

Järvinen, O. & Sammalisto, L. 1976: Regional trends in the avifauna of Finnish peatland bogs. — Ann. Zool. Fennici 13: 31—43.

Järvinen, O., Sisula, H., Varvio-Aho, S.-L. & Salminen, P. 1976: Genic variation in isolated marginal populations of the Roman Snail, Helix pomatia L. — Hereditas 82: 101—110.

Kimura, M. & Crow, J. F. 1964: The number of alleles that can be maintained in a finite population. — Genetics 49: 725—738.

Lewontin, R. C. 1972: The apportionment of human diversity. — Evol. Biol. 6: 381—398.

Margolin, B. H. & Light, R. J. 1974: An analysis of variance for categorical data. II: Small sample comparison with chi-square and other competitiors. — J. Amer. Statist. Assoc. 69: 755—764.

Mitton, J. B. 1977: Genetic differentiation of races of man as judged by single-locus and multilocus analyses. — Amer. Natur. 111: 203—212.

Nei, M. 1973: Analysis of gene diversity in subdivided populations. — Proc. Nat. Acad. Sci. 70: 3321—3323.

Nei, M. 1975: Molecular population genetics and evolution. — 288 p. Amsterdam.

Peet, R. K. 1974: The measurement of species diversity. Ann. Rev. Ecol. Syst. 5: 285—307.

Poole, R. W. 1974: An introduction to quantitative ecology. — 532 pp. New York.

Saura, A., Lakovaara, S., Lokki, J. & Lankinen, P. 1973: Genic variation in central and marginal populations of Drosophila subobscura. — Hereditas 75: 33—46.

Sokal, R. R. & Rohlf, F. J. 1969: Biometry. 368 pp. San Francisco.