Theor Appl Genet (1988) 76:947–951

The meaning of genetic variation within and between subpopulations

H.-R. Gregorius

Abteilung für Forstgenetik und Forstpflanzenzüchtung, Georg-August-Universität, Büsgenweg 2, 3400 Göttingen, F. R. Germany

Received June 15, 1988; Accepted July 5, 1988
Communicated by P. M. A. Tigerstedt

Summary. Argumentation is presented which indicates that the additive decomposition of the total genetic variation of a population into variation within and between (among) its subpopulations suffers from conceptual inconsistency. While the measurement of variation between subpopulations can be shown to be identical to the measurement of subpopulation differentiation, the notion of variation within subpopulations, when viewed as a single measurement, cannot be derived as an independent and cogent concept. Rather, it appears to be merely technically defined as the arithmetic difference between the total variation and the variation between subpopulations, and this difference happens to be non-negative for concave measures of variation such as the (statistical) variance or certain measures of genetic diversity. In order to overcome the conceptual inconsistency, "variation between subpopulations" could be regarded as subpopulation differentiation and the notion of "variation within subpopulations" could be replaced by measurements of the degree to which the variation in the total population is represented within the subpopulations. A complementary situation with respect to total variation is thus realized once more, and appropriate measures can be directly derived from existing ones.

Key words: Genetic variation – Within subpopulations – Between subpopulations – Genetic differentiation – Genetic diversity

Introduction

"Variation within subpopulations" and "variation between subpopulations" are two concepts generally seen as being mutually complementary with respect to total variation in a subdivided population. This view is manifested in the statistical Eq.

$$V(X) = E(V(X \mid Y)) + V(E(X \mid Y)), \quad (1)$$

where X and Y are random variables and X is real-valued; E and V denote the expectation and the variance, respectively; $X \mid Y$ symbolizes X under the condition of Y. Thus, the total variance equals the sum of the expectation of the conditional variances and the variance of the conditional expectations. When X refers to the trait whose variation in a population is to be studied and Y represents the subpopulations, then $V(X)$ reflects the total variation, and $E(V(X \mid Y))$ the variation within and $V(E(X \mid Y))$ the variation between subpopulations. Clearly, the applicability of this equation depends on the assumption that the statistical variance is an appropriate measure of the extent of variation, and it does not hold if other measures of variation, such as that of the standard deviation ($\sqrt{V}$), are considered. This is worth emphasizing since the standard deviation rather than the variance is usually considered to be the appropriate measure of the variation of metric traits. Priority has apparently been given to the mathematical convenience associated with computing second moments over the rigour of conceptual reasoning.

In population genetics, the standard methods of measuring genetic variation in subdivided populations are also based on Eq. (1). Since frequencies of genetic types (alleles) serve as a means for characterizing variation, the problem arises of how to regard these measurements as a real-valued trait X. The solution to this problem consists in fixing a single genetic type and assigning to each individual a value of 1 or 0 according to whether it does or does not show this type (such variables are usually called "indicator variables"). The thus resulting trait, X_i (i represents the genetic type under consideration), has an ex-

pectation $E(X_i) = p_i$ and a variance $V(X_i) = p_i(1-p_i)$, where p_i is the relative frequency of the i-th genetic type. Hence, when

$p_i(j) :=$ relative frequency of the i-th genetic type in the j-th subpopulation ($\sum_i p_i(j) = 1$), and

$c_j \quad :=$ relative size of the j-th subpopulation ($\sum_j c_j = 1$),

then $E(V(X_i|Y)) = \sum_j c_j p_i(j)(1-p_i(j)) = p_i - \sum_j c_j \cdot p_i(j)^2$ and $V(E(X_i|Y)) = \sum_j (p_i(j) - p_i)^2 \cdot c_j$. Here Y refers to the subpopulations and thus to the indices j, and $p_i = \sum_j p_i(j) \cdot c_j$ is the relative frequency of the i-th type in the total population.

In order to incorporate all genetic types, the sums

$$V_T := \sum_i V(X_i) = 1 - \sum_i p_i^2 ,$$

$$V_{WS} := \sum_i E(V(X_i|Y)) = \sum_j c_j \cdot (1 - \sum_i p_i(j)^2) ,$$

$$V_{BS} := \sum_i V(E(X_i|Y)) = \sum_i \sum_j (p_i(j) - p_i)^2 \cdot c_j$$

are taken, for which analogously to Eq. (1) the identity

$$V_T = V_{WS} + V_{BS} \tag{2}$$

holds. Again, the measure of the genetic variation in the total population (V_T) is additively composed of the measures of genetic variation V_{WS} within and V_{BS} between subpopulations. In Wright's (1978) terminology $V_T = y_T$, $V_{WS} = y_{ST}$ and $V_{BS} = y_{DT}$, while in Nei's (1973) terminology $V_T = H_T$, $V_{WS} = H_S$ and $V_{BS} = D_{ST}$. When expressed as a fraction of the total variation, the variation between subpopulations becomes $V_{BS}/V_T = F_{ST} = G_{ST}$, where F_{ST} is Wright's notation and G_{ST} is from Nei. The different viewpoints from which the quantities in Eq. (2) were derived by the two authors should not obscure the fact that they are mathematically identical. Consequently, these viewpoints could be considered alternative interpretations of the same formal result.

Yet, despite the general acceptance and application of Eq. (2) in the analysis of population genetic data, fundamental objections can be raised to the conceptual and statistical reasoning of this Eq. First of all, V_T, V_{WS} and V_{BS} cannot be assumed to be variances of a single real-valued trait, as would be required in order to employ the concept represented by Eq. (1). Moreover, the use of the indicator variables X_i, which is necessary to derive Eq. (2), is arbitrary. In view of this, it is almost superfluous to add that to use variance to measure variation as mentioned above is basically questionable.

In other approaches, variance as a measure of variation is replaced by diversity, although the concept suggested by Eq. (1) or (2) is maintained. Diversity is a non-negative function v of the relative frequencies of the types, i.e. $v = v(p_1, p_2, \ldots)$, and it meets a number of conditions among which concavity guarantees that $v_T \geq \sum_j c_j \cdot v_j$, where $v_T := v(p_1, p_2, \ldots)$ *and* $v_j := v(p_1(j), p_2(j), \ldots)$. v_T corresponds to the total variation V_T, and since v_j is the diversity in the j-th subpopulation, the average $\sum_j c_j \cdot v_j$ is addressed as the measure V_{WS} of variation within subpopulations. In this approach, variation between subpopulations can not be derived independently and is therefore defined as the difference $V_T - V_{WS}$. Note that the fulfilment of the necessary requirement $V_T - V_{WS} \geq 0$ is solely guaranteed by the assumption of concavity. For the particular form $v = -\sum_i p_i \cdot \log p_i$ this was recommended by Lewontin (1972) to be a method describing the decomposition of variation. The greatest weakness of this method is that it provides no reasoning for why the difference $V_T - V_{WS}$ should be an appropriate measure of variation between subpopulations. The tacit assumption implied here is that the two kinds of variation are always complementary.

In view of the arbitrariness that appears in these examples, it is desirable to restate the intuitive basis as well as the purposes of the concept independently of the lines of sight suggested by Eq. (1). Having done this, the statements implicit in Eq. (2) could be related in a possibly more direct manner to the underlying concept.

Conceptual considerations

Whenever variation among the members of a population is considered for qualitative characteristics, such as genes or genotypes, the frequencies of the types involved are of primary concern. The amount of variation is, consequently, directly associated with the number of types and their frequencies in the population. Highly variable populations contain many different types at approximately equal frequencies. The computation of variances or standard deviations, however, is not an immediate issue here because the individual measurements are not of a metric nature.

If in a population all subpopulations (for convenience the term "deme" will be used instead of the term "subpopulation") under study show the same pattern of variation, then variation between demes is generally said to be absent. This is maintained for demes of different sizes as long as there are no differences in their relative composition. In this sense, each deme is representative of the total population, so that variation within demes is equated to the variation in the total population. So far, the concept works, since all of the variation is due to variation within demes and none is due to variation between them.

At the other extreme, if all demes are unique in the sense that they do not share individuals of the same type, the pattern of variation in the total population can be obtained only by the union of all demes. Hence, total variation is completely dissociated between the demes, and one might be inclined to conclude that total variation is entirely due to variation between demes. However, the unique contribution of each deme to the total variation need not consist in a single type of individual, since uniqueness does not imply fixation of a deme if the number of types in the population exceeds the number of demes. Thus, even though the maximum possible difference between demes is reached, variation may exist within demes to arbitrary degrees, which contradicts the above conception and makes it very difficult to unambiguously distinguish the contributions of the sources of variation (for more detailed reasoning, see Gregorius and Roberds 1986).

This dilemma becomes even more apparent when the extreme values of the fraction $V_{BS}/V_T = F_{ST}$ are considered. Provided $V_T > 0$ (i.e. the total population is not fixed to a single type), F_{ST} attains its maximum value of 1 if, and only if, all demes are fixed. This property is in complete accordance with the purpose for which Wright originally derived F_{ST}, namely to measure the extent of fixation. However, its validity is questionable for the present purpose, since $F_{ST} = 1$ independent of whether all demes are fixed for different types or, with only one exception, for the same type. Thus, the situation where all demes make unique contributions to the total variation is equivalent to that where only a vanishing minority does.

In fact, this criticism applies to a much broader class of measures of variation, including the above-mentioned concave measures of diversity. To see this, suppose that v now denotes any non-negative measure of variation based on relative or absolute frequencies, for which $v = 0$ only if the population is fixed for a single type. Furthermore, suppose that the variation V_{WS} within demes is defined as any type of average over the amounts v of variation within each of the demes, and that this average becomes 0 only if all of the single deme amounts of variation are equal to 0. Then, if Eq. (2) is assumed to apply, $V_{BS}/V_T = (V_T - V_{BS})/V_T = 1 - V_{WS}/V_T$, so that $V_{BS}/V_T = 1$ if and only if $V_{WS} = 0$. However, by assumption, $V_{WS} = 0$ only if all demes are fixed. It thus turns out that any method of measuring variation within and between demes suffers from the "fixation dilemma", if it falls into the category described by Eq. (2) and if V_{WS} is an average of the single deme measures of variation.

The last explanations strongly support the idea that the two types of variation, i.e. "within" and "between (among) demes", should not be considered as complementary with respect to the total variation; instead, they deserve to be treated separately and independently. For example, individuals that were previously classified as representing the same type may become distinguishable due to the availability of better techniques of identification. In such a case, a situation of fixation of all demes could be changed drastically without affecting the quality of the demes to be unique. In other words, differences between demes need not be associated with differences within demes. This is an aspect of great importance in population genetics, where improvements in biochemical techniques enable the identification of ever increasing amounts of genetic variation.

Ideally, an individual belonging to a particular deme contributes to variation between demes if its type does not occur in any of the other demes; that is, if it is unique for its deme. However, a given type is usually represented by more than one individual in several demes. Hence, among the individuals of a given type in a particular deme, only that number can be considered as unique for this deme, which remains after subtraction of all individuals of the same type belonging to the other demes. If a type is evenly represented in all demes, then the whole type does not contribute to variation between demes, since none of the pertinent individuals is unique for any of the demes. Consequently, the sum over all individuals that are unique for any of the demes directly measures the amount of total variation that is due to only differences between demes (variation between demes). The mathematical formulation taking into account different deme sizes will be given in the next section.

This principle of uniqueness can also be applied to arrive at a direct measure of variation within each single deme or within the total population. Each individual in the population (or deme) simply has to be conceived of as a deme of its own. There remains the problem of how to combine the single deme measures to give one measure for all demes simultaneously, i.e. to give the variation within demes. Taking averages is a popular solution, however, as long as no conceptually cogent reasons are given for why one and not another type of average is applied, a suspicion of arbitrariness lingers. One could even argue that disregarding the differences between the amounts of variation within the demes, as is the case with taking averages, ignores an essential characteristic of the pattern of variation in a subdivided population. For example, using averages, a situation where the amount of variation within demes is negatively correlated with their sizes could not be properly distinguishable from one where all demes show about the same amounts of variation. In both situations, the amount of variation between demes could be identical, so that the above phenomenon would never have been detected. In fact, what is involved here is a third type of variation, namely variation between demes of variation within demes. Therefore, the concept of variation within demes, when viewed in terms of a single measurement, is probably not very useful and will thus not be further pursued.

A measure of variation between demes

If in two populations $\boldsymbol{n}$ and $\boldsymbol{n}'$ of equal size n, individuals of type i have (absolute) frequency n_i and n_i', respectively, then the number of i-type individuals by which $\boldsymbol{n}$ differs from $\boldsymbol{n}'$ is equal to

$$\alpha_i(\boldsymbol{n}-\boldsymbol{n}') := \max\{n_i - n_i', 0\} = \tfrac{1}{2}\cdot(n_i - n_i') + \tfrac{1}{2}\cdot|n_i - n_i'|\,.$$

This follows immediately from the fact that if $n_i > n_i'$, then $\boldsymbol{n}$ has an excess of $n_i - n_i'$ i-type individuals when compared to $\boldsymbol{n}'$. Otherwise, if $n_i \leq n_i'$, then n has no i-type individuals by which it differs from $\boldsymbol{n}'$.

Conversely, $\boldsymbol{n}'$ differs from $\boldsymbol{n}$ in

$$\alpha_i(\boldsymbol{n}'-\boldsymbol{n}) = \tfrac{1}{2}\cdot(n_i' - n_i) + \tfrac{1}{2}\cdot|n_i' - n_i|$$

individuals of type i. Hence, since $\sum_i n_i = \sum_i n_i'$ by assumption, the total number of individuals by which $\boldsymbol{n}$ differs from $\boldsymbol{n}'$ in type is equal to

$$\alpha(\boldsymbol{n}-\boldsymbol{n}') = \sum_i \alpha_i(\boldsymbol{n}-\boldsymbol{n}') = \tfrac{1}{2}\sum_i |n_i - n_i'|$$
$$= n\cdot\tfrac{1}{2}\sum_i\left|\frac{n_i}{n} - \frac{n_i'}{n}\right|.$$

When these considerations are extended to populations $\boldsymbol{n}$ and $\boldsymbol{m}$ of unequal size n and m, respectively, it must be taken into account that the differences in the number of individuals of a given type may be due merely to the difference in population size, which would be an undesired effect. Hence, one population has to be compared with another on the basis of the relative composition of the latter. This, in turn, requires that the latter be transformed in such away that it has the size of the former, but without changing its relative composition. Consequently, if the number of individuals by which $\boldsymbol{n}$, say, differs from $\boldsymbol{m}$ is to be determined, the frequencies m_i of the types in $\boldsymbol{m}$ have to be replaced by $m_i\cdot n/m$. Thus, $\alpha_i(\boldsymbol{n}-\boldsymbol{m})$ becomes

$$\alpha_i(\boldsymbol{n}-\boldsymbol{m}) = \max\left\{n_i - m_i\cdot\frac{n}{m}, 0\right\}$$
$$= n\cdot\frac{1}{2}\left[\frac{n_i}{n} - \frac{m_i}{m} + \left|\frac{n_i}{n} - \frac{m_i}{m}\right|\right].$$

Therefore, in the general case

$$\alpha(\boldsymbol{n}-\boldsymbol{m}) = n\cdot\frac{1}{2}\sum_i\left|\frac{n_i}{n} - \frac{m_i}{m}\right|,$$

and

$$\frac{1}{n}\cdot\alpha(\boldsymbol{n}-\boldsymbol{m}) = \frac{1}{m}\cdot\alpha(\boldsymbol{m}-\boldsymbol{n}) =: d(\boldsymbol{n},\boldsymbol{m})\,.$$

$d(\boldsymbol{n},\boldsymbol{m}) = d(\boldsymbol{m},\boldsymbol{n})$ is a measure of the distance between $\boldsymbol{n}$ and $\boldsymbol{m}$ (Gregorius 1974), and it specifies the proportion of individuals by which two populations differ from each other in type.

The above method can now be applied to a subdivided population to measure the amount by which total variation is due to differences between demes. The key idea is to characterize each deme by the number of individuals by which it differs in types from the remainder of the population. This number corresponds directly to those individuals in a deme whose types are not repeated in individuals belonging to any of the other demes. Hence, the sum of these individuals taken over all demes represents exactly that part of the total variation which is attributable exclusively to uniqueness of the demes and thus to differences between demes. In order to arrive at a formal representation, consider the following quantities:

$n_i(j) :=$ the number of individuals of type i in deme j;
$n_i := \sum_j n_i(j)$ or the number of individuals of type i in the total population;
$n(j) := \sum_i n_i(j)$ or the number of individuals in deme j (deme size);
$n := \sum_j n(j) = \sum_i n_i$ or size of the total population.

The number of individuals of type i by which deme j differs from the remainder of the population will now be denoted as $\alpha_i(j)$, and considering that this remainder has size $n - n(j)$ and contains $n_i - n_i(j)$ individuals of type i, it follows from the foregoing derivations that

$$\alpha_i(j) = \max\left\{n_i(j) - (n_i - n_i(j))\cdot\frac{n(j)}{n-n(j)}, 0\right\}$$
$$= n(j)\cdot\frac{1}{2}\left[\frac{n_i(j)}{n(j)} - \frac{n_i - n_i(j)}{n-n(j)} + \left|\frac{n_i(j)}{n(j)} - \frac{n_i - n_i(j)}{n-n(j)}\right|\right].$$

Consequently, the number $\alpha(j)$ of individuals by which deme j differs in types from the remainder of the population is equal to

$$\alpha(j) = \sum_i \alpha_i(j) = n(j)\cdot\frac{1}{2}\sum_i\left|\frac{n_i(j)}{n(j)} - \frac{n_i - n_i(j)}{n-n(j)}\right|,$$

and, therefore, the total number of individuals α by which the demes differ from each other in types is

$$\alpha = \sum_j \alpha(j)\,.$$

Thus, one arrives at the result that α divided by the total population size, which will be denoted as $\delta = \alpha/n$, is the desired measure of the proportion of the total variation that is due to differences between demes. In more precise wording: δ equals the proportion of individuals in the total population by which the demes differ in types from each other. Setting $c_j := n(j)/n$ and $D_j := \alpha(j)/n(j)$ (which is the distance between the j-th deme and the remainder of the population), it follows that $\delta = \sum_j c_j\cdot D_j$,

which is then seen to be identical to the measure previously derived by Gregorius and Roberds (1986) for quantification of the amount of deme differentiation. In other words, based on the present approach, the concepts of 'variation between demes' and 'deme differentiation' are identical. In particular, δ does not suffer from the "fixation dilemma", since it assumes its maximum value of 1 only when all demes are unique; that is, when they do not share individuals of the same type.

Concluding remarks

The measure δ of variation between demes can be consistently extended to a measure of variation within a non-subdivided population by considering each individual as a deme of its own, as was mentioned before. This results in the measure δ_T of differentiation of the total population given by

$$\delta_T = \sum_i \frac{n_i \cdot (n - n_i)}{n \cdot (n-1)},$$

which is identical to Simpson's measure of diversity (Gregorius 1987). Again, n_i is the number of i-type individuals in the total population; $n = \sum_i n_i$ is the population size; and $0 \leq \delta_T \leq 1$. In accordance with the general features of δ, $\delta_T = 0$ only if the total population is fixed, and $\delta_T = 1$ only if all members of the population differ in type (are distinguishable).

Hence, δ and its special form δ_T are measures of variation which, at least at first sight, appear to correspond to the components V_{BS} and V_T in Eq. (2), whereas V_{WS} remains as yet unspecified. If the concept underlying Eq. (2) should apply, the inequality $V_T \geq V_{BS}$ must hold in all cases, which, by the assumed correspondence, should translate into the inequality $\delta \leq \delta_T$. However, the last inequality obviously does not hold in those cases where all demes are unique in their types and at least two individuals (necessarily belonging to the same deme) are identical in type, since then $\delta = 1$ but $\delta_T < 1$.

Even if δ is rejected for any reason as a measure corresponding to V_{BS}, Simpson's widely applied measure δ_T still disturbs the picture evoked by Eq. (2). Consider the extreme situation where the demes are not differentiated, i.e. where the relative frequencies of the types within a deme are the same for all demes and thus for the total population ($\delta = 0$). Then the δ_T-value of the j-th deme is $(1 - \sum_i p_i^2) \cdot n(j)/(n(j) - 1)$, where the p_i's denote the relative frequencies and $n(j)$ is the size of the j-th deme ($\sum_j n(j) = n$). Similarly, the δ_T-value of the total population is $(1 - \sum_i p_i^2) \cdot n/(n-1)$. Consequently, since $n(j)/(n(j) - 1) > n/(n-1)$ for all j, the amount of variation within each deme exceeds that of the total population. Hence, by correspondence, $V_{WS} > V_T$ independent of the averaging process applied for the computation of V_{WS}. This again fundamentally contradicts the concept implied by Eq. (2).

In conclusion, it appears that the notion of "variation within demes", when specified by a single value, has no generally valid conceptual basis. Therefore, the idea that total variation in a subdivided population should be composed of one part reflecting variation within demes and a seond part reflecting variation between demes is probably useless. The measurement of variation between demes appears to be tantamount to the measurement of subpopulation differentiation, and this need not be strictly associated with measurements of variation in the total population. It is therefore recommended that the concepts of total variation and variation between demes be viewed as being largely mutually independent. The measures proposed support this view. Moreover, it is suggested that the notion of "variation within demes" be replaced by the measurement of the degree to which the variation in the total population is represented within the demes. δ and its components D_j serve this purpose. $D_j = 0$, for example, means that the j-th deme is completely representative of (undifferentiated with respect to) the total population, while $D_j = 1$ indicates the uniqueness of this deme. δ summarizes the single deme measures according to the weights of the demes. Hence, $1 - D_j$ and $1 - \delta$ can be addressed as measures of representation of the total variation by the demes, and they are thus exactly complementary to the respective measures of differentiation.

Acknowledgements. The author expresses grateful appreciation to J. H. Roberds for helpful discussions.

References

Gregorius H-R (1974) Genetischer Abstand zwischen Populationen. I. Zur Konzeption der genetischen Abstandsmessung. Silvae Genet 23:22–27

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

Lewontin RC (1972) The apportionment of human diversity. In: Dobzhansky Th, Hecht MK, Steere WMC (eds) Evolutionary biology, vol. 6. Appleton-Century-Crofts, New York, pp 381–398

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

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