*Indian J. Genet.*, **49**(2): 227–230 (1989)

# ON THE METHOD OF CLASSIFICATION USING $D^2$ STATISTIC AND THE IMPORTANCE OF CHARACTERS CONTRIBUTING TO GENETIC DIVERGENCE

V. ARUNACHALAM AND A. BANDYOPADHYAY

*Indian Agricultural Research Institute*
*New Delhi 110012*

**(Received: March 23, 1987; accepted: October 6, 1988)**

## ABSTRACT

**The method of classification using $D^2$ statistic as given by Singh and Chaudhary [7] is examined in depth and shown to be basically erroneous. A relook on the paper that identifies the character contributing most to genetic divergence by Singh [8] brings to light the complexities of the problem which remain essentially unsolved.**

**Key words:** Genetic divergence, $D^2$ statistic, classification.

The method of classifying genetic stocks on the basis of genetic divergence measured by Mahalanobis' distance statistic ($D^2$) is now well established in plant breeding [1–4]. The theoretical basis of classification using $D^2$ is explained by Rao [5] with a number of examples from various fields. An algorithm for computing $D^2$ is available [6] and included in the subroutine library of various computers in India. But the method of grouping suggested by Tocher [5] uses a norm to decide whether a genetic stock can be included in a cluster. No set formula to determine this norm can be devised, since the norm is not constant for all clusters but would depend on the range and magnitude of variation in $D^2$ values. However, Singh and Chaudhary [7] chose to define this norm without logic; grouping based on their norm results in the first one or two groups containing a large proportion of genetic stocks and the rest ending up in single-stock clusters. The problem needs to be looked in proper perspective.

In addition to grouping, attempts were made earlier to identify characters that largely contribute to $D^2$ values with a view to saving time and cost in measuring characters that do not contribute substantially to genetic differentiation. In a recent paper, Singh [8] pointed out some conceptual errors in the methods adopted earlier and proposed a method amending the earlier defects. But several gaps persist in his amended method too, which needs a careful scrutiny.

This paper addresses itself to these two problems.

## MATERIALS

The practical example given by Singh and Chaudhary [7] in chapter 12 titled Classificatory Analysis (pp. 204–214) was reworked starting from the raw data of

Tables 9–12 on pp. 49–50. The 8 varieties were grouped using the basic principles enunciated in [5] and also using the method of [7]*.

## DISCUSSION

### GROUPING BASED ON $D^2$ STATISTIC

According to Rao [5], "No formal rules can be laid down for finding the clusters because a cluster is not a well defined term. The only criterion appears to be that any two groups belonging to the same cluster should at least on the average show a smaller $D^2$ than those belonging to two different clusters." The device by Tocher helps essentially in achieving this criterion.

In the Tocher's method, it is usual to make preliminary groups such that the various $D^2$ values between varieties within any group are comparable in magnitude. Thus, it is possible that $D^2$ values in one group may be in the range of 1–10, and in some other in the range of 150–250. A final grouping is obtained from the preliminary groups as a next step using the essential principle that the values of "average $D^2$" (A) and "increase in $D^2$ – increase in number of $D^2$" (B) should be comparable when any variety is added to a group. It is, however, not possible to specify how much the difference (A – B) should be. This would depend on the magnitude of $D^2$ and on the range of variation.

Singh and Chaudhary [7], on the other hand, proposed a maximum value for the quantity B referred to above. The two methods were applied to the example cited in ' Materials' above. The Tocher's method led to the following groups: I: 2; II: 4; III: 1, 5, 8; IV: 3, 6; and the method of Singh and Chaudhary gave the following groups in the same material: I: 1, 4, 6, 7; II: 5, 8; III: 2; and IV: 3. The two methods led to completely different grouping, the common group being only the one containing variety 2. This is natural, as variety 2 is unique in having high $D^2$ values with every other variety.

The method of Singh and Chaudhary is defective since it ignores the pattern of $D^2$ variation and the basic definition of a group, and uses a single illogical value for the norm.

### IMPORTANCE OF CHARACTERS CONTRIBUTING TO GENETIC DIVERGENCE

Earlier workers (for example [8]) have interpreted the direction and magnitude of the elements of various characters in the first two canonical vectors that are the best linear combinations.

Singh [8] argued that these elements cannot have any logical interpretation on the importance of individual characters as they do not correspond in their extent

*A number of errors is found in the computation of $D^2$ almost in every step starting from the computation of common dispersion matrix: The grouping given in the text of this paper is the correct one.

of contribution to differentiation. Hence he has derived the canonical vector corresponding to X values and suggested that the character corresponding to the element having maximum value (ignoring direction) in the first canonical vector is the one contributing maximum to $D^2$. Thus, his paper, in fact, aimed at no more than identifying the character with maximum contribution to $D^2$. It was also identified as character j for which the value of S.j ($= \sum_{i=1}^{p} w^{ij} d_i d_j$, where $w^{ij}$ was inverse of the common dispersion matrix) was maximum. The same argument would also imply that the character next in importance would be the one corresponding to the element of the canonical vector next best in value. Alternatively, it would be character r for which S. r would be the next best to S. j in value, and so on. A scrutiny of the example worked out in [8] would show that the order of importance given by the two criteria did not agree entirely, though the most important character was identified correctly by both of them. His criteria essentially stem from the following fact:

If $l_i$ (i = 1, p) are the elements of the first canonical vector and if $K = \sum_{i=1}^{p} l_i d_i$, where $d_i$ is the difference in the means of two populations with respect to each original character $X_i$, then $D^2$ with respect to K

$$= K^2 \text{ (} \oint \text{ 9c.2 [5])}$$
$$= \Sigma (l_i d_i)^2$$

We note that the value of $D^2$ depends on the values of $l_i$ and $d_i$ which determine the value of the linear combination K. Hence the individual values of $l_i$ (that too, ignoring their sign) alone cannot decide the final value of K, as both the direction and magnitude of $l_i$ and $d_i$ are equally important. The $l_i$ values corresponding to X values derived by Singh [8] and also those corrected using "standard weights" can be, at best, of limited help.

## REFERENCES

1. S. R. Chandrasekhariah, B. R. Murty and V. Arunachalam. 1969. Multivariate analysis of divergence in the genus *Eu-sorghum*. Proc. Nat. Inst. Sci., India, B, **35:** 172–195.

2. B. R. Murty and V. Arunachalam. 1966. The nature of divergence in relation to breeding systems in some crop plants. Indian J. Genet., **26A:** 188–198.

3. B. R. Murthy, V. Arunachalam and I. J. Anand. 1973. Effect of environment on the genetic divergence among some populations of linseed. Indian J. Genet., **33:** 304–313.

4. S. Vairavan, E. A. Siddiq, V. Arunachalam and M. S. Swaminathan. 1973. A study of the nature of divergence in rice from Assam and Northeast Himalayas. Theor. Appl. Genet., **43:** 213–221.

5. C. R. Rao. 1952. Advanced Statistical Methods in Biometrical Research. Wiley, New York, U.S.A.

6. B. R. Murty and V. Arunachalam. 1967. Computer programmes for some problems in biometrical genetics. I. Use of Mahalanobis' $D^2$ in classificatory problems. Indian J. Genet., **27**: 60–69.

7. R. K. Singh and B. D. Chaudhary. 1977. Biometrical Methods in Quantitative Genetic Analysis. Kalyani Publishers, New Delhi.

8. Daljit Singh. 1981. The relative importance of characters affecting divergence. Indian J. Genet., **41**: 237–245.