ISOZYME ALLELIC FREQUENCIES RELATED TO SELECTION AND GENE-FLOW HYPOTHESES[1]

HENRY E. SCHAFFER AND F. M. JOHNSON

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27607

Manuscript received September 24, 1973

Revised copy received January 29, 1974

ABSTRACT

Significant correlations between allelic frequencies and environmental variables in a number of insect species have been demonstrated by multivariate techniques. Since many environmental variables show a strong relationship to geographic location and since gene flow between populations can also produce patterns of gene frequencies which are related to the geographic location, both selection and gene-flow hypotheses are consistent with the observed correlations. The genetic variables can be corrected for geographic location and so for linear gene-flow patterns. If, after correction, the genetic variables still show significant correlations with similarly corrected environmental variables, then these correlations are consistent with hypotheses of selection but not of gene flow. The data of JOHNSON and SCHAFFER (1973) have been reanalyzed using the method of canonical correlation after correction for geographical location by means of multiple regression. Five of the nine loci studied exhibit significant canonical correlations. These results, under the assumption of linear gene flow, support hypotheses of selective action of environmental variables in the genotype-environment relationships observed.

THE relationships between isozyme (allozyme) allelic frequencies in natural populations and various measurements of the environments in which these populations exist have been studied in a number of insect species. JOHNSON *et al.* (1969) found statistically significant relationships in the harvester ant, *Pogonmyrmex barbatus*. KOJIMA *et al.* (1972) studied *Drosophila pavani*, ROCKWOOD-SLUSS, JOHNSTON and HEED (1973) studied *D. pachea*, TOMASZEWSKI, SCHAFFER and JOHNSON (1973) studied *P. badius*, and JOHNSON and SCHAFFER (1973) studied *D. melanogaster*. In all of these cases, statistically significant relationships were demonstrated between allelic frequency variation and some characteristics of the environment. The environmental characteristics include the temperature, precipitation and humidity records maintained by the Weather Bureau. ROCKWOOD-SLUSS, JOHNSTON and HEED (1973) also included the concentration of several chemical compounds in the senita cactus which is the food source for *D. pachea*.

[1] Paper Number 4173 of the Journal Series of the North Carolina State Agricultural Experiment Station, Raleigh, N.C. This investigation was supported by NIH Research Grant Number GM 11546 from the National Institute of General Medical Sciences and by contract number AT-(40-1)3980 from the U.S. Atomic Energy Commission.

The statistical methodology used in these studies, except for in that of KOJIMA *et al.* (1972), has primarily involved multivariate methods such as principal components and canonical correlation; TAYLOR and MITTON (1972, 1974) have used factor analysis. All of these methods are used to synthesize new variables from the original variables. The new variables can be considered to be patterns which are identified in the original data. These patterns are sought in the data to satisfy such criteria as maximum variance or correlation, depending on which multivariate method is used.

The end result in each study has been the identification of allelic frequency (genetic) patterns and environmental patterns which exhibit a statistically significant correlation. Most of the environmental patterns identified in the studies mentioned above have shown a relationship with the geographical locations at which the populations were collected and the environmental measurements taken. It is natural that such environmental characteristics as elevation, temperature and precipitation show strong relationships with geographical location. Thus when patterns of these variables show correlations with genetic patterns, hypotheses based on selective effects of the environment become confounded with those based on non-selective location-dependent effects, e.g., gene flow as a result of drift and migration. This confounding of the two categories of hypotheses has been pointed by SCHAFFER and JOHNSON (1973). However, when the environmental and genetic patterns are not completely correlated with geographical location, it is possible to partially differentiate between effects of the two hypotheses. The results of JOHNSON and SCHAFFER (1973) will be reanalyzed in such a manner.

MODEL AND ANALYSIS

A cline for an allelic frequency may arise as the result of a wave of migration, in which the frequency of migrants carrying a new allele gradually diminishes over distance, or from the meeting of two populations with different frequencies of an allele. Considering a two-dimensional geographical area instead of a transect and letting the frequency of the allele represent the third dimension, such a cline will appear as a surface. In the simplest case, the allelic frequency will be proportional to the distance involved and so the allelic frequency surface will be a plane. Subsequent instances of gene flow for the same allele may alter the orientation of the plane but will leave it as a plane. If more than two alleles are present, then additional allele will add one more dimension and the graph can no longer be directly visualized, as it requires more than three dimensions for its representation. However, the surface described in the higher dimensional space, termed a flat, is analogous to the plane which was visualized in three dimensions (KENDALL 1961).

Clines in gene frequencies can also be produced as the result of the intergradation of two populations. More than two populations may also be involved. Such intergradation can result in a cline which tends to be a linear combination of the genes of the two populations with coefficients proportional to the distance from the ends of the overlapping region. Such a cline should be, to a first approximation, a "flat".

If the observed allelic frequencies at a locus are plotted as additional dimensions versus the geographical locations at which the collections were made, a higher dimensional representation will result. To the extent that the points representing the allelic frequencies fit a flat, the differences in allelic frequencies between locations could have resulted from the type of gene flow described above. Many environmental factors also approximate a flat when plotted against the geographical coordinates. Thus selective adaptation of allelic frequencies to these factors would also tend to produce allelic frequency points which are consistent with a flat. In this way, two possible causative agencies are confounded.

Environmental variables are also observed to be in patterns related to the geographical location. The strongest such relationship in the environmental data presented by JOHNSON and SCHAFFER (1973) was for average annual temperature, which has 98.5% of its variation linearly related to geographical location. The weakest was average noon relative humidity with 1.5% of its variation related to location. In each case this percentage was the multiple correlation coefficient (R^2) resulting from a multiple regression of the environmental variable on the independent variables of latitude and longitude. In all of these analyses the geographical area is being treated as a two-dimensional area with latitude and longitude comprising a rectangular coordinate system.

To the extent that both the environmental variables and the possible results of gene migration show similar patterns, either of them could explain equally well the appearance of such patterns in the gene frequencies. The presence of these patterns can be seen by the size of the R^2 values given in Table 1. The re-

TABLE 1

Analysis of genetic and environmental data with respect to geographical location

Variable	No. of sites	R^2	First canonical correlation
EstC	30	.65**	.66
Est6	42	.25**	.72**
Adh†	42	.88**	.43
α-gpdh	42	.23**	.72**
Acph	42	.18*	.62
Odh	41	.64**	.56*
Mdh	42	.12	.65**
Pgm	25	.36**	.82**
Aph‡	14	.84**	.70
Elevation	42	.78**	
Avg. ppt.	42	.90**	
Avg. temp.	42	.99**	
Avg. midnight rel. humid.	42	.39**	
Avg. noon rel. humid.	42	.01	
Avg. wind speed	42	.31**	

For loci with multiple alleles the largest R^2 for regression on geographical location is given.
* $p < .05$; ** $p < .01$.
The canonical correlations are computed using the corrected data as described in the text.
† Multiple correlation coefficient (R) for regression of corrected genetic data on environment is given instead of canonical correlation since only two alleles were found at this locus.
‡ Not all environmental measurements included because of multiple collinearities.

moval of the geographical "flat" pattern from both gene frequency data and the environmental measurements will yield corrected sample data (deviations) which are uncorrelated with "flat" migrational patterns. It will also remove from the genetic data any clines which are caused by environmental variation which is linearly related to the geographical coordinates. Thus this correction for gene migration may also remove certain effects of selection, if such are present, and so produce a minimum estimate of the influence of the environment.

The correction for geographical location can be expressed geometrically in terms of a dimensional space where each variable (genetic or environmental) is plotted against the two dimensions representing the geographical coordinates. The best fitting plane (in the sense of minimizing the sum of squared distances of points from the plane) is passed through the points and the new corrected variables are the deviations (signed distances) of the points from the plane. The new variables are orthogonal to the geographical coordinates, and so are uncorrelated with any possible flat gene migration pattern. Both the genetic and environmental data are re-expressed as deviations in this manner. The arithmetic can be done using multiple regression analysis. Each gene frequency and each environmental variable is regressed on the "independent" variables of latitude and longitude and the deviations calculated. Using these deviations as the new variables, they can be analyzed to detect relationships between the patterns of genetic and environmental variation.

The closest such relationship is a correlation which can be calculated using the multivariate statistical method of canonical correlation. This method (KENDALL and STUART 1966; MORRISON 1967) finds the genetic pattern and the environmental pattern which show the greatest possible correlation in a set of data. These patterns are linear combinations of the individual variables in exactly the same sense as in principal components analysis, except here they are chosen as a pair, one genetic pattern and one environmental pattern, to maximize the correlation between them. Additional pairs of patterns can be chosen which are uncorrelated with previous pairs and which have lower correlation. Only the first (largest) correlation will be discussed here. The first canonical correlation between each locus and the environmental variables is presented for each locus in Table 1. The significance is calculated by means of a chi-square test. All calculations were done using the Statistical Analysis System (SERVICE 1972).

RESULTS AND DISCUSSION

Of the nine loci analyzed, five show significant or highly significant canonical correlations, one (*Acph*) is almost significant ($p=.09$) and three show obviously statistically nonsignificant results. The number of observations for the *Aph* locus is insufficient for this type of analysis. Even though the power of the test for the *Aph* locus is undoubtedly low, it is included for comparison with the previous analysis given by JOHNSON and SCHAFFER (1973).

These results strongly indicate that there is an actual correlation between genetic and environmental patterns which are not correlated with geographical

location. This is not consistent with the neutral hypothesis of gene migration but is predicted by the hypothesis that the environment selectively influences the genetic polymorphisms studied. Thus these results support the selection hypothesis.

While the data have been corrected for linear effects of gene migration, there are patterns of gene migration which may not be linearly related to the geographical location—for example, migration along a winding river or through a mountain pass. Considering the distances between the sites and the total geographical area involved, it does not appear that the data should exhibit genetic patterns depending on such fine detail of the migration patterns. However, this possibility does keep the above tests from conclusively demonstrating the effect of environment completely corrected for gene migration.

No assumption is made concerning linearity of the observed gene frequencies, or environmental factors which may be acting selectively. In fact, the analysis is a search for geographical non-linearities in the genetic patterns which are correlated with corresponding non-linearities in the environmental patterns. In this manner, selective influences which are not linearly related to the geographic locations can be detected, and only selective influences which are not linearly related to the geographic coordinates can be detected.

The statistical analyses used on these data are computationally involved and complex, involving multiple regression analysis of each variable and a subsequent canonical correlation analysis on the array of deviations. The chi-square test for the significance of the canonical correlation is an approximate test of significance, and the data which were analyzed are not distributed according to the multivariate normal distribution. In consideration of the foregoing, it is of interest to determine whether the test being used is of correct size or whether the sequence of analyses would show correlations between genetic and environmental patterns, even if there actually were none. This question was answered by a sampling study in which the data were simulated by random numbers and for which the same analyses were conducted.

The same geographical coordinates as in the real data were used, and random gene frequencies and environmental readings were generated. There were four alleles at the locus. Since gene frequencies are constrained to add to unity, one gene frequency was omitted from the analyses, as was done with the data from the collections. Three sets of simulations were run in which the environmental and genetic variables were chosen from different distributions. In two sets the gene frequencies were distributed multinomially with equal frequencies. In both of these sets the environmental variables were distributed identically and independently. In one set the uniform distribution over the unit interval was used and in the other set the normal distribution with zero mean and unit variance was used. The third set was similar to the second, except that the variables had unequal variances and correlations were introduced within each group of genetic and environmental variables.

In each set the probabilities, or significance values, of the canonical correlation were plotted and nicely approximated the uniform distribution which should

be exhibited with random data if the analysis does not introduce spurious correlations. The fraction of significant results, at the .05 level, in the three sets was 1/22, 1/25 and 2/25, which are all close to the expected .05 level. These results indicate that the method of analysis used here does not introduce significant correlations when they are not actually present.

LITERATURE CITED

JOHNSON, F. M. and H. E. SCHAFFER, 1973 Isozyme variability in species of the genus *Drosophila*. VII. Genotype-environment relationships in populations of *D. melanogaster* from the eastern U. S. Biochem. Genet. **10**: 149–163.

JOHNSON, F. M., H. E. SCHAFFER, J. E. GILLASPY and E. S. ROCKWOOD, 1969 Isozyme-environment relationships in natural populations of the harvester ant, *Pogonomyrmex barbatus*, from Texas. Biochem. Genet. **3**: 429–450.

KENDALL, M. G., 1961 *A Course in the Geometry of n Dimensions*. Hafner Publishing Company, New York.

KENDALL, M. G. and A. STUART, 1966 *The Advanced Theory of Statistics*. Vol. 3. Hafner Publishing Company, New York.

KOJIMA, K., P. SMOUSE, S. YANG, D. S. NAIR and D. BRNCIC, 1972 Isozyme frequency patterns in *Drosophila pavani* associated with geographical and seasonal variables. Genetics **72**: 721–731.

MORRISON, D. F., 1967 *Multivariate Statistical Methods*. McGraw-Hill Book Company, New York.

ROCKWOOD-SLUSS, E. S., J. S. JOHNSTON and W. B. HEED, 1973 Allozyme genotype-environment relationships. I. Variation in natural populations of *Drosophila pachea*. Genetics **73**: 135–146.

SCHAFFER, H. E. and F. M. JOHNSON, 1973 Alternative hypotheses explaining the correlation of geographical patterns of isozyme gene frequencies and environmental factors. Genetics **74**: s242 (abstract).

SERVICE, J., 1972 *A User's Guide to the Statistical Analysis System*. Student Supply Stores, N. C. State University, Raleigh, N. C.

TAYLOR, C. and J. B. MITTON, 1972 Multivariate analysis of genetic variation. Genetics **71**: s63–s64 (abstract). ——, 1974 Multivariate analysis of genetic variation. (manuscript in preparation).

TOMASZEWSKI, E. K., H. E. SCHAFFER and F. M. JOHNSON, 1973 Isozyme genotype-environment associations in natural populations of the harvester ant, *Pogonomyrmex badius*. Genetics **75**: 405–421.

Corresponding editor: R. W. ALLARD