Variance-Partitioning and Classification in Human Population Genetics

NOAH A. ROSENBERG

In his 2003 paper on human genetic diversity ([192]), Edwards develops a simple model of allele frequencies in two populations, using it to examine the potential for classifying an individual genome into one of the two populations on the basis of its alleles. He employs the model to critique interpretations of the magnitude and significance of human genetic diversity tracing to the celebrated paper of Lewontin (1972).

Edwards draws a distinction between two questions about human genetic variation. The first is the variance-partitioning question initially posed by Lewontin (1972): in a partition of worldwide human genetic variation into components representing differences among individuals within populations, differences among populations within major geographic groups, and differences among geographic regions, what are the numerical values of the variance components? The second is the classification question: can an individual be accurately classified into one of a series of specified populations on the basis of his or her multilocus genotypes? The notable contribution of Edwards's article is the lucidity with which it uses a simple two-population model to demonstrate that the result for the former question need not imply the result for the latter; even if the within-population variance is the largest component, classification can still be possible.

As Edwards's treatment shows, the statistical distinction between the two questions lies in the way in which information is combined across multiple loci. In a variance-partitioning analysis, different loci amount to separate trials, each of which produces its own estimate of the desired variance components. The variance partition for a multilocus dataset is computed as an average across loci; additional loci merely refine the precision of the estimate and ensure that it approaches the value that would be obtained if data were available genome-wide. In a classification analysis, however, information is *accumulated* across loci, rather than simply averaged; while any particular allelic type might occur at only slightly different frequencies across groups, multilocus *combinations* of allelic types differ

dramatically in their probability of originating in different source populations. Classification is possible not because any one genetic character is population-specific or even substantially different in frequency across groups, but because multilocus combinations of characters are much more likely to arise in one group than in the others. Variance-partitioning does not consider multilocus combinations; in fact, in many implementations, including Lewontin's, it is performed directly from allele frequencies – without any knowledge of individual multilocus genotypes.

Edwards writes that his article "could, and perhaps should, have been written soon after 1974" (p. 801). In a sense, it was. Shortly after the appearance of Lewontin's (1972) variance partition and the related work of Nei and Roychoudhury (1972, 1974), Spielman and Smouse introduced the classification problem through a multivariate assignment of individual members of geographically dispersed Amazonian natives by village, village cluster, or tribe (Spielman and Smouse 1976, Smouse and Spielman 1977), reporting a classification success "far in excess of random expectation" (Spielman and Smouse 1976, p. 330). Mitton (1978) then used multilocus combinations to implement a form of variance-partitioning, finding substantially greater difference among populations than in the averaging work of Lewontin (1972). This paper drew three sharp critiques, framed primarily as statistical objections (Chakraborty 1978, Lewontin 1978, Powell and Taylor 1978), to which Mitton (1978), employing principal-component analysis, responded by arguing that "multivariate analyses may provide resolution of groups that is not apparent in a sequence of single-locus analyses" (p. 1143).

A commentary by Neel (1981) characterized the distinct questions appearing in the exchange. These were first, the problem of Lewontin, Nei, and Roychoudhury (Neel 1981, p. 83): "What proportion of all the genetic variation within some large group can be apportioned to differences among subgroups and among individuals, on the average, over all known loci?" And second, the problem of Spielman, Smouse, and Mitton (Neel 1981, p. 83): "Are the levels of allelic frequency variation found between human populations sufficient to generate a useful taxonomy?" Neel wrote that "Chakraborty, Lewontin, Powell, and Taylor appear to me to be chastising Mitton for asking his question rather than the one already posed by Lewontin, Nei, and Roychoudhury." Thus, in [192], Edwards's attempt to separate a similar pair of questions occupies a parallel position in the literature to Neel's commentary (and another by Chakraborty [1982], analyzing the latest work of Smouse *et al*. [1982]), except that in Edwards's critique, it is Lewontin who is chastised.

The efforts represented in Edwards's paper, in the early literature, and in further recent attempts – including one co-authored by Lewontin – to clearly distinguish

questions at issue in the study of human genetic variation (Feldman and Lewontin 2008, Barbujani and Colonna 2010, Rosenberg 2011) illustrate the ongoing challenge of providing a genetic portrayal that simultaneously represents both similarity and difference in an accurate manner. As Edwards remarks, the variance-partitioning result has been taken as the basis for statements that it does not directly imply. For example, the frequently reported summary of this result, in the form of a statement that 85% of human genetic variation is within populations, might deceptively appear to respond to a question not about variance-partitioning, but rather about the presence and absence of alleles: of the genetic variants that exist in the human genome, how many are present within a given population? This latter question has a result smaller than 85%, with quite different values occurring in different parts of the world (Rosenberg 2011).

However, a connection between formulations of the classic population-genetic statistic F_{st} in a type of variance-partitioning calculation and in an analysis that examines times of coalescence of genealogical lineages, in the same and different populations, does mean that variance components have a surprising equivalence with results on pairwise similarity and difference (Rosenberg 2011). In this equivalence, the within-population variance component for worldwide human genetic variation – computed using a heterozygosity statistic instead of the entropy statistic in Lewontin's original work – is approximately equal to the ratio of the number of pairwise genetic differences for two individuals, chosen from the same population, to the number of pairwise differences for two individuals who are chosen arbitrarily from anywhere in the world. Thus, the *Nature* statement quoted by Edwards (p. 250, this volume), "two random individuals from any one group are almost as different as any two random individuals from the entire world," if considered in terms of pairwise differences, and assuming 85% is taken to be a large number, is a legitimate – though non-obvious – consequence of Lewontin's result.

The fallacy to which Edwards refers in the title of [192] concerns the conflation of variance-partitioning and classification, but he also critiques Lewontin's view that a high value of the variance component for differences among individuals within populations argues against human racial classification. The destructive typological race theories of the past, requiring small numbers of genetic characters with dramatic biologically significant differences across groups to be typical of the genome-wide pattern, are clearly belied by the high within-population variance component. In the disagreement between Edwards and Lewontin, however, much about their senses of the meaning of "race" is left unstated in both positions. The interpretation of the magnitude of human genetic similarities and differences in relation to "race" demands more explicit context on what properties a theory of the existence or non-existence of "races" entails for human variation. It also requires an analysis of whether or not the necessary criteria are satisfied by the

actual patterns in human populations. Nevertheless, Edwards's critique helps to illuminate that in assessing the applicability of any such theory of race, the answers to many separate questions about human genetic similarity and difference must be simultaneously explained.

A phenotypic perspective that extends beyond an exclusively genotypic analysis is surely relevant for such inquiries. Edge and Rosenberg (2015) extended the model in [192] to incorporate a selectively neutral quantitative phenotype, examining three additional questions related to variance-partitioning and classification. (1) How does the mean difference between two populations in the value of a phenotype change as the number of loci that influence the trait increases? (2) What proportion of variance in a phenotype is expected for the between-population variance component? (3) Does a phenotype become increasingly useful for identifying the source population of an individual as the number of loci grows? In each case, the model predicts that a single neutral quantitative trait has similar properties in terms of variance-partitioning and classification to those of a single genetic locus. This conclusion, building on Edwards's framework, solidifies the link that the variance-partitioning and classification results in the genetic studies of Lewontin and Edwards extend analogously to a class of phenotypes. Thus, [192] has contributed not only to characterizing the multiple distinct questions that are of interest in the study of human genetic similarities and differences, but also to providing a model for their further investigation.

References

Barbujani, G. and Colonna, V. 2010. Human genome diversity: frequently asked questions. *Trends in Genetics* 26, 285–295.

Chakraborty, R. 1978. Single-locus and multilocus analysis of genetic differentiation of the races of man: a critique. *The American Naturalist* 112, 1134–1138.

Chakraborty, R. 1982. Allocation versus variation: the issue of genetic differences between human racial groups. *The American Naturalist* 120, 403–404.

Edge, M. D. and Rosenberg, N. A. 2015. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Studies in History and Philosophy of Science, Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 52, 32–45.

Feldman, M. W. and Lewontin, R. C. 2008. Race, ancestry, and medicine. In *Revisiting Race in a Genomic Age*, eds. Koenig, B. A., Lee, S. S. J., Richardson, S.S., Piscataway, NJ: Rutgers University Press, 89–101.

Lewontin, R. C. 1972. The apportionment of human diversity. In *Evolutionary Biology*, eds. Dobzhansky, T., Hecht, M. K., Steere, W. C., New York: Appleton-Century-Crofts, vol. 6, 381–398.

Lewontin, R. C. 1978. Single- and multiple-locus measures of genetic distance between groups. *The American Naturalist* 112, 1138–1139.

Mitton, J. B. 1977. Genetic differentiation of races of man as judged by single-locus and multilocus analyses. *The American Naturalist* 111, 203–212.

Mitton, J. B. 1978. Measurement of differentiation: reply to Lewontin, Powell, and Taylor. *The American Naturalist* 112, 1142–1144.

Neel, J. V. 1981. The major ethnic groups: diversity in the midst of similarity. *The American Naturalist* 117, 83–87.

Nei, M. and Roychoudhury, A. K. 1972. Gene differences between Caucasian, Negro, and Japanese populations. *Science* 177, 434–436.

Nei, M. and Roychoudhury, A. K. 1974. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *American Journal of Human Genetics* 26, 421–443.

Powell, J. R. and Taylor, C. E. 1978. Are human races "substantially" different genetically? *The American Naturalist* 112, 1139–1142.

Rosenberg, N. A. 2011. A population-genetic perspective on the similarities and differences among worldwide human populations. *Human Biology* 83, 659–684.

Smouse, P. E. and Spielman, R. S. 1977. How allocation of individuals depends on genetic differences among populations. In *Human Genetics: Proceedings of the Fifth International Congress of Human Genetics, Mexico City, 10–15 October 1976*, eds. Armendares, S., Lisker, R., Amsterdam: Excerpta Medica, 255–260.

Smouse, P. E., Spielman, R. S., and Park, M. H. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *The American Naturalist* 119, 445–463.

Spielman, R. S. and Smouse, P. E. 1976. Multivariate classification of human populations I. Allocation of Yanomama Indians to villages. *American Journal of Human Genetics* 28, 317–331.