

Human genetic diversity: Lewontin's fallacy

A.W.F. Edwards

Summary

In popular articles that play down the genetical differences among human populations, it is often stated that about 85% of the total genetical variation is due to individual differences within populations and only 15% to differences between populations or ethnic groups. It has therefore been proposed that the division of *Homo sapiens* into these groups is not justified by the genetic data. This conclusion, due to R.C. Lewontin in 1972, is unwarranted because the argument ignores the fact that most of the information that distinguishes populations is hidden in the correlation structure of the data and not simply in the variation of the individual factors. The underlying logic, which was discussed in the early years of the last century, is here discussed using a simple genetical example. *BioEssays* 25:798–801, 2003.
© 2003 Wiley Periodicals, Inc.

“When a large number of individuals [of any kind of organism] are measured in respect of physical dimensions, weight, colour, density, etc., it is possible to describe with some accuracy the population of which our experience may be regarded as a sample. By this means it may be possible to distinguish it from other populations differing in their genetic origin, or in environmental circumstances. Thus local races may be very different as populations, although individuals may overlap in all characters;...” R.A. Fisher (1925). See Ref. 1.

“It is clear that our perception of relatively large differences between human races and subgroups, as compared to the variation within these groups, is indeed a biased perception and that, based on randomly chosen genetic differences, human races and populations are remarkably similar to each other, with the largest part by far of human variation being accounted for by the differences between individuals. Human racial classification is of no social value and is positively destructive of social and human relations. Since such racial classification is now seen to be of virtually no genetic or taxonomic significance either, no justification can be offered for its continuance”. R.C. Lewontin (1972). See Ref. 2.

“The study of genetic variations in *Homo sapiens* shows that there is more genetic variation within populations than between populations. This means that two random individuals from any one group are almost as different as any two random individuals from the entire world. Although it may be easy to observe distinct external differences between groups of people, it is more difficult to distinguish such groups genetically, since most genetic variation is found within all groups.” *Nature* (2001). See Ref. 3.

Introduction

In popular articles that play down the genetical differences among human populations it is often stated, usually without any reference, that about 85% of the total genetical variation is due to individual differences within populations and only 15% to differences between populations or ethnic groups. It has therefore been suggested that the division of *Homo sapiens* into these groups is not justified by the genetic data. People the world over are much more similar genetically than appearances might suggest.

Thus an article in *New Scientist*⁽⁴⁾ reported that in 1972 Richard Lewontin of Harvard University “found that nearly 85 per cent of humanity’s genetic diversity occurs among individuals within a single population.” “In other words, two individuals are different because they are individuals, not because they belong to different races.” In 2001, the *Human Genome* edition of *Nature*⁽³⁾ came with a compact disc containing a similar statement, quoted above.

Such statements seem all to trace back to a 1972 paper by Lewontin in the annual review *Evolutionary Biology*.⁽²⁾ Lewontin analysed data from 17 polymorphic loci, including the major blood-groups, and 7 ‘races’ (Caucasian, African, Mongoloid, S. Asian Aborigines, Amerinds, Oceanians, Australian Aborigines). The gene frequencies were given for the 7 races but not for the individual populations comprising them, although the final analysis did quote the within-population variability. “The results are quite remarkable. The mean proportion of the total species diversity that is contained within populations is 85.4%.... Less than 15% of all human genetic diversity is accounted for by differences between human groups! Moreover, the difference between populations within a race accounts for an additional 8.3%, so that only 6.3% is accounted for by racial classification.”

Gonville and Caius College, Cambridge, CB2 1TA, UK.
E-mail: awfe@cam.ac.uk
DOI 10.1002/bies.10315
Published online in Wiley InterScience (www.interscience.wiley.com).

Lewontin concluded "Since . . . racial classification is now seen to be of virtually no genetic or taxonomic significance . . . , no justification can be offered for its continuance" (full quotation given above).

Lewontin included similar remarks in his 1974 book *The Genetic Basis of Evolutionary Change*⁽⁵⁾ "The taxonomic division of the human species into races places a completely disproportionate emphasis on a very small fraction of the total of human diversity. That scientists as well as nonscientists nevertheless continue to emphasize these genetically minor differences and find new 'scientific' justifications for doing so is an indication of the power of socioeconomically based ideology over the supposed objectivity of knowledge."

The fallacy

These conclusions are based on the old statistical fallacy of analysing data on the assumption that it contains no information beyond that revealed on a locus-by-locus analysis, and then drawing conclusions solely on the results of such an analysis. The 'taxonomic significance' of genetic data in fact often arises from correlations amongst the different loci, for it is these that may contain the information which enables a stable classification to be uncovered.

Cavalli-Sforza and Piazza⁽⁶⁾ coined the word 'treeness' to describe the extent to which a tree-like structure was hidden amongst the correlations in gene-frequency data. Lewontin's superficial analysis ignores this aspect of the structure of the data and leads inevitably to the conclusion that the data do not possess such structure. The argument is circular. A contrasting analysis to Lewontin's, using very similar data, was presented by Cavalli-Sforza and Edwards at the 1963 International Congress of Genetics.⁽⁷⁾ Making no prior assumptions about the form of the tree, they derived a convincing evolutionary tree for the 15 populations that they studied. Lewontin,^(2,5) though he participated in the Congress, did not refer to this analysis.

The statistical problem has been understood at least since the discussions surrounding Pearson's 'coefficient of racial likeness'⁽⁸⁾ in the 1920s. It is mentioned in all editions of Fisher's *Statistical Methods for Research Workers*⁽¹⁾ from 1925 (quoted above). A useful review is that by Gower⁽⁹⁾ in a 1972 conference volume *The Assessment of Population Affinities in Man*. As he pointed out, "... the human mind distinguishes between different groups *because* there are correlated characters within the postulated groups."

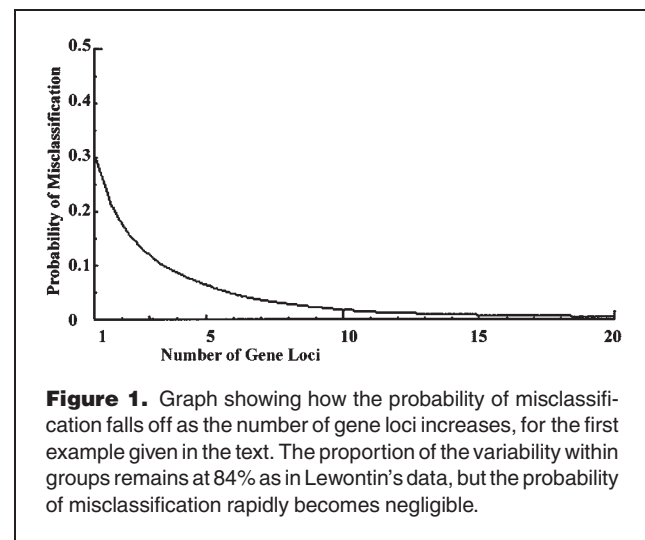
The original discussions involved anthropometric data, but the fallacy may equally be exposed using modern genetic terminology. Consider two haploid populations each of size n . In population 1 the frequency of a gene, say '+' as opposed to '-', at a single diallelic locus is p and in population 2 it is q , where $p + q = 1$. (The symmetry is deliberate.) Each population manifests simple binomial variability, and the overall variability is augmented by the difference in the means. The

natural way to analyse this variability is the analysis of variance, from which it will be found that the ratio of the within-population sum of squares to the total sum of squares is simply $4pq$. Taking $p = 0.3$ and $q = 0.7$, this ratio is 0.84; 84% of the variability is within-population, corresponding closely to Lewontin's figure. The probability of misclassifying an individual based on his gene is p , in this case 0.3. The genes at a single locus are hardly informative about the population to which their bearer belongs.

Now suppose there are k similar loci, all with gene frequency p in population 1 and q in population 2. The ratio of the within-to-total variability is still 84% at each locus. The total number of '+' genes in an individual will be binomial with mean kp in population 1 and kq in population 2, with variance kpq in both cases. Continuing with the former gene frequencies and taking $k = 100$ loci (say), the mean numbers are 30 and 70 respectively, with variances 21 and thus standard deviations of 4.58. With a difference between the means of 40 and a common standard deviation of less than 4.6, there is virtually no overlap between the distributions, and the probability of misclassification is infinitesimal, simply on the basis of counting the number of '+' genes. Fig. 1 shows how the probability falls off for up to 20 loci.

One way of looking at this result is to appreciate that the total number of '+' genes is like the first principal component in a principal component analysis (Box 1). For this component the between-population sum of squares is very much greater than the within-population sum of squares. For the other components the reverse will hold, so that overall the between-population sum of squares is only a small proportion (in this example 16%) of the total. But this must not beguile one into thinking that the two populations are not separable, which they clearly are.

Each additional locus contributes equally to the within-population and between-population sums of squares, whose



Box 1. Principal component analysis

Principal Component Analysis (PCA) is a way of teasing out the more important information in multivariate data, where the high dimensionality renders simple graphical presentation impossible. The procedure can easily be understood even with just two variates, though its use might then be unnecessary. Taking an example from anthropometry where PCA originated, we might have data on the lengths and breadths of a number of human skulls. Each skull can be represented by a point in a diagram whose two axes are length and breadth. Since length and breadth will almost certainly be associated to some extent, the points will tend to be spread out preferentially in a certain direction, stretching from short length and breadth (a small skull) to long length and breadth (a large skull).

PCA defines this direction precisely as that of the line for which the sum of the squares of the perpendicular distances from the points to the line is a minimum. This line passes through the centre of gravity of the points, and a simple application of Pythagoras's Theorem shows that the one-dimensional array of the points defined by the feet of the perpendiculars from the points to the line then has the maximum possible sum of squares. In other words, the variability of the data has been partitioned into two components, one of which, along this line, is known as the (First) Principal Component because it encapsulates as much of the variability as can be represented in one dimension. The Second Component, at right angles to the First, encapsulates the remainder, which is, of course, a minimum.

These two components can be used as replacement axes on the graph. Sometimes the First Component will have an obvious meaning, as would be the case with the skulls, where it is clear that it corresponds in a general way to 'size'. Similarly the Second Component corresponds in some sense to 'shape', because a skull whose data-point is far from the line of the First Component will either be longer and narrower than the norm, or shorter and broader.

The procedure generalises to any number of variates, and the successive First, Second, Third, ... Components are then mutually-orthogonal directions partitioning the total variability into ever-decreasing amounts. A graph of the first two components will represent as much of the information as is possible using only two dimensions.

proportions therefore remain unchanged but, at the same time, it contributes information about classification which is cumulative over loci because their gene frequencies are correlated.

Classification

It might be supposed, though it would be wrong, that this example is prejudiced by the assumptions that membership of the two populations is known in advance and that, at each locus, it is the same population that has the higher frequency of the '+' gene. In fact the only advantage of the latter simplifying assumption was that it made it obvious that the total number of '+' genes is the best discriminant between the two populations.

To dispel these concerns, consider the same example but with '+' and '-' interchanged at each locus with probability 1/2, and suppose that there is no prior information as to which population each individual belongs. Clearly, the total number of '+' genes an individual contains is no longer a discriminant, for the expected number is now the same in each group. A cluster analysis will be necessary in order to uncover the groups, and a convenient criterion is again based on the analysis of variance as in the method introduced by Edwards and Cavalli-Sforza.⁽¹⁰⁾ Here the preferred division into two clusters maximises the between-clusters sum of squares or, what is the same thing, minimises the sum of the within-clusters sums of squares.

As pointed out by these authors, it is extremely easy to compute these sums for binary data, for all the information is contained in the half-matrix of pairwise distances between the individuals, and at each locus this distance is simply 0 for a match and 1 for a mismatch of the genes. Since interchanging '+' and '-' makes no difference to the numbers of matches and mismatches, it is clear that the random changes introduced above are irrelevant. Continuing the symmetrical example, the probability of a match is $p^2 + q^2$ if the two individuals are from the same population and $2pq$ if they are from different populations. With k loci, therefore, the distance between two individuals from the same population will be binomial with mean $k(p^2 + q^2)$ and variance $k(p^2 + q^2)(1 - p^2 - q^2)$ and if from different populations binomial with mean $2kpq$ and variance $2kpq(1 - 2pq)$. These variances are, of course, the same.

Taking $p = 0.3$, $q = 0.7$ and $k = 100$ as before, the means are 58 and 42 respectively, a difference of 16, the variances are 24.36 and the standard deviations both 4.936. The means are thus more than 3 standard deviations apart (3.2415). The entries of the half-matrix of pairwise distances will therefore divide into two groups with very little overlap, and it will be possible to identify the two clusters with a risk of misclassification which tends to zero as the number of loci increases.

By analogy with the above example, it is likely that a count of the four DNA base frequencies in homologous tracts of a genome would prove quite a powerful statistical discriminant for classifying people into population groups.

Conclusion

There is nothing wrong with Lewontin's statistical analysis of variation, only with the belief that it is relevant to classification.

It is not true that “racial classification is...of virtually no genetic or taxonomic significance”. It is not true, as *Nature* claimed, that “two random individuals from any one group are almost as different as any two random individuals from the entire world”, and it is not true, as the *New Scientist* claimed, that “two individuals are different because they are individuals, not because they belong to different races” and that “you can’t predict someone’s race by their genes”. Such statements might only be true if all the characters studied were independent, which they are not.

Lewontin used his analysis of variation to mount an unjustified assault on classification, which he deplored for social reasons. It was he who wrote “Indeed the whole history of the problem of genetic variation is a vivid illustration of the role that deeply embedded ideological assumptions play in determining scientific ‘truth’ and the direction of scientific inquiry”.⁽⁵⁾ In a 1970 article *Race and intelligence*⁽¹¹⁾ he had earlier written “I shall try, in this article, to display Professor Jensen’s argument, to show how the structure of his argument is designed to make his point and to reveal what appear to be deeply embedded assumptions derived from a particular world view, leading him to erroneous conclusions.”

A proper analysis of human data reveals a substantial amount of information about genetic differences. What use, if any, one makes of it is quite another matter. But it is a dangerous mistake to premise the moral equality of human beings on biological similarity because dissimilarity, once revealed, then becomes an argument for moral inequality. One is reminded of Fisher’s remark in *Statistical Methods and Scientific Inference*⁽¹²⁾ “that the best causes tend to attract to their support the worst arguments, which seems to be equally true in the intellectual and in the moral sense.”

Epilogue

This article could, and perhaps should, have been written soon after 1974. Since then many advances have been made in both gene technology and statistical computing that have facilitated the study of population differences from genetic

data. The magisterial book of Cavalli-Sforza, Menozzi and Piazza⁽¹³⁾ took the human story up to 1994, and since then many studies have amply confirmed the validity of the approach. Very recent studies^(14,15) have treated *individuals* in the same way that Cavalli-Sforza and Edwards treated *populations* in 1963, namely by subjecting their genetic information to a cluster analysis thus revealing genetic affinities that have unsurprising geographic, linguistic and cultural parallels. As the authors of the most extensive of these⁽¹⁵⁾ comment, “it was only in the accumulation of small allele-frequency differences across many loci that population structure was identified.”

References

1. Fisher RA. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd. 1925.
2. Lewontin RC. The apportionment of human diversity. In: Dobzhansky T, Hecht MK, Steere WC, editors. *Evolutionary Biology* 6. New York: Appleton-Century-Crofts. 1972. p 381–398.
3. The Human Genome. *Nature* 2001;409:following p 812.
4. Ananthaswamy A. Under the skin. *New Scientist* 2002;174:34–37.
5. Lewontin RC. *The Genetic Basis of Evolutionary Change*. New York: Columbia University Press. 1974.
6. Cavalli-Sforza LL, Piazza A. Analysis of evolution: evolutionary rates, independence and treeness. *Theor Pop Biol* 1975;8:127–165.
7. Cavalli-Sforza LL, Edwards AWF. Analysis of human evolution. *Proc. 11th Internat. Congr. Genetics, The Hague 1963, Genetics Today* 3. Oxford: Pergamon. 1965. p 923–933.
8. Pearson K. On the coefficient of racial likeness. *Biometrika* 1926;18: 105–117.
9. Gower JC. Measures of taxonomic distance and their analysis. In: Weiner JS, Huizinga J, editors. *The Assessment of Population Affinities in Man*. Oxford: Clarendon. 1972. p 1–24.
10. Edwards AWF, Cavalli-Sforza LL. A method for cluster analysis. *Biometrics* 1965;21:362–375.
11. Lewontin RC. Race and intelligence. *Bulletin of the Atomic Scientists*. March 1970;2–8.
12. Fisher RA. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd. 1956.
13. Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton University Press. 1994.
14. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;155:945–959.
15. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovskiy LA, Feldman MW. Genetic structure of human populations. *Science* 2002;298:2381–2385.