Studies in History and Philosophy of Biological and Biomedical Sciences 52 (2015) 32–45

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity

Michael D. Edge*, Noah A. Rosenberg

Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA, 94305-5020, USA

ARTICLE INFO

Article history:
Available online 9 February 2015

Keywords:
Genetic differentiation
Population genetics
Quantitative genetics
Race

ABSTRACT

Researchers in many fields have considered the meaning of two results about genetic variation for concepts of "race." First, at most genetic loci, apportionments of human genetic diversity find that worldwide populations are genetically similar. Second, when multiple genetic loci are examined, it is possible to distinguish people with ancestry from different geographical regions. These two results raise an important question about human phenotypic diversity: To what extent do populations typically differ on phenotypes determined by multiple genetic loci? It might be expected that such phenotypes follow the pattern of similarity observed at individual loci. Alternatively, because they have a multilocus genetic architecture, they might follow the pattern of greater differentiation suggested by multilocus ancestry inference. To address the question, we extend a well-known classification model of Edwards (2003) by adding a selectively neutral quantitative trait. Using the extended model, we show, in line with previous work in quantitative genetics, that regardless of how many genetic loci influence the trait, one neutral trait is approximately as informative about ancestry as a single genetic locus. The results support the relevance of single-locus genetic-diversity partitioning for predictions about phenotypic diversity.

1. Introduction

Going back to Lewontin's 1972 study of human genetic diversity, many investigators have reported that at typical genetic loci, most of the allelic variation in statistical partitions of human genetic variation is "within," rather than "between," populations (e.g. Barbujani, Magagni, Minch, & Cavalli-Sforza, 1997; Brown & Armelagos, 2001; Li et al., 2008; Rosenberg et al., 2002; Rosenberg, Pritchard, et al., 2003). Many of these studies presented their results as estimates of F_{ST}, a quantity that can be interpreted as the proportion of allelic variance—that is, variance in a binary random variable representing the presence or absence of a specific allele—attributable to differences in allele frequencies between populations (Holsinger & Weir, 2009). Estimates of worldwide human F_{ST} and F_{ST}-like quantities have ranged from ~0.05 (e.g. Rosenberg et al., 2002) to ~0.15 (e.g. Barbujani et al., 1997), meaning that 5–15% of allelic variance at a representative locus is due to between-population differences in allele frequencies—or, equivalently, that 85–95% lies in the within-population variance component.

In spite of this result, which shows that human groups have similar allele frequencies at most variable loci, it is possible to infer the continental ancestry of individual people using genetic data alone (e.g. Bamshad et al., 2003; Bowcock et al., 1994; Mountain & Cavalli-Sforza, 1997; Rosenberg et al., 2002; Tang et al., 2005). Ancestry inference is performed by pooling information from many loci. Each locus provides only a small amount of information about population membership, but when many loci are used, their information can be combined to distinguish among potential ancestries.

In 2003, A. W. F. Edwards provided a particularly clear explanation of the way in which multiple loci can be used to classify accurately even when each individual locus is only weakly informative (Edwards, 2003). Edwards' point was not new—it appeared in earlier arguments about allelic-variance partitioning and

* Corresponding author. Tel.: +1 650 724 5122.
E-mail address: medge@stanford.edu (M.D. Edge).

http://dx.doi.org/10.1016/j.shpsc.2014.12.005
1369-8486/

classification (e.g., Mitton, 1977; Neel, 1981; Smouse, Spielman, & Park, 1982)—but he used an accessible model that clarified the result.

What do single-locus variance partitioning and multilocus classification studies lead us to expect about phenotypic differences between human populations? The finding that human groups have similar allele frequencies at most genetic loci has been used to support arguments that most large, genetically-based phenotypic differences between groups are exceptions to the genomic rule (e.g., Brown & Armelagos, 2001; Feldman & Lewontin, 2008; Goodman, 2000). Indeed, single-locus partitioning studies do suggest that human populations will not differ widely on most traits controlled by a single genetic locus. But the fact that classification is possible using many loci seems to suggest that human groups might differ more substantially on traits influenced by many loci. If populations can be distinguished with multilocus genotypes, then it is possible that phenotypes controlled by multilocus genotypes could differ markedly between populations. Should we expect to see larger differences between human populations for traits influenced by many loci than for traits influenced by a single locus?

Here, we extend Edwards' model—which has already proven to be an effective framework for describing results about allelic-variance partitioning and classification—to study the expected level of between-population difference for a selectively neutral quantitative trait. Other researchers have studied this question in other contexts (e.g. Berg & Coop, 2014; Chakraborty & Nei, 1982; Felsenstein, 1973; Lande, 1976, 1992; Rogers & Harpending, 1983; Whitlock, 1999), but by basing our analysis in Edwards' model, we explicitly connect questions about trait differences to questions about multilocus ancestry inference. We show that for a random quantitative trait under the extended Edwards model, two groups are not unduly likely to differ on traits that are determined by many loci, even when the loci influencing the trait would provide a sufficient basis for accurate classification. In particular, the expected level of difference between the populations' mean trait values is, in two senses made more precise below, approximately equal to the magnitude of single-locus genetic difference between the populations. Similarly, a typical multilocus trait contributes approximately as much information for classification as does a single genetic locus.

2. The Edwards model

Risch, Burchard, Ziv, and Tang (2002, box 1), Edwards (2003), and Tal (2012) have examined related classification models involving accumulations of information across loci; here, we consider the simplest of these models, that of Edwards (2003). We first describe key features of the model, and we then introduce a quantitative trait.

Suppose we have two haploid populations of equal size, labeled A and B. At one genetic locus, the probability that an individual from population A has an allele we label "1" is p, with $p \in (0,1/2)$, and the probability of allele "0" is $q = 1 - p$. In population B, the allele frequencies are switched: The probability of "1" is q and the probability of "0" is p. Table 1 shows the allele frequencies by population.

Table 1
The frequencies of the "0" and "1" alleles in each population, with $p + q = 1$.

Population	Allele	
	"0"	"1"
A	q	p
B	p	q

We can represent the genotype of an individual at the locus as a random variable L that takes values of 0 and 1, and we can represent population membership of an individual as a random variable M that takes values A and B. Within each population, the allelic variance at the locus—that is, the variance of L—is $Var(L|M = A) = Var(L|M = B) = pq$. This result follows from the status of $L|M$ as a Bernoulli random variable with probability either p or q.

When not conditioning on population membership, the genotype at the locus is still a Bernoulli random variable, but now, because populations A and B are equal in size, the probability of observing a "1" is 1/2:

$$\begin{aligned} P(L = 1) &= P(M = A)P(L = 1|M = A) \\ &\quad + P(M = B)P(L = 1|M = B) \\ &= \frac{1}{2}p + \frac{1}{2}q = \frac{1}{2}[p + (1 - p)] = \frac{1}{2}. \end{aligned}$$

The total unconditional variance of L is therefore $Var(L) = P(L = 0)P(L = 1) = 1/4$.

The proportion of the total allelic variance that is "within populations"—that is, the proportion of the total variance that remains after conditioning on an individual's population membership—is the conditional variance of L given M divided by the total variance of L:

$$Var(L|M)/Var(L) = pq/(1/4) = 4pq.$$

Because the total allelic variance is the sum of within- and between-population components, the proportion of the total variance in allelic types that is "between populations," or F_{ST}, is

$$F_{ST} = (1/4 - pq)/(1/4) = 1 - 4pq. \quad (1)$$

Mimicking estimates for the between-region and between-population proportion of genetic diversity from Lewontin (1972) and subsequent studies, if we assume $p < q$, then we might take p between 0.3 and 0.4—an interval that produces within-population variance proportions from 0.84 to 0.96—as approximately reflecting differences between human groups at a typical locus.

Suppose we want to classify individuals into populations using the genotype at the locus. That is, we wish to predict population membership M after observing an individual's allele. If $p < q$, then the decision rule with the greatest prediction accuracy is to assign individuals with a "0" allele to population A and individuals with allele "1" to population B (Rosenberg, Li, Ward, & Pritchard, 2003). That is, we assign an individual to the population in which its allele is most common. Misclassification occurs for individuals from population A with a "1" allele and individuals from population B with a "0" allele. The total misclassification probability is

$$\begin{aligned} &P(L = 1|M = A)P(M = A) + P(L = 0|M = B)P(M = B) \\ &= \frac{1}{2}p + \frac{1}{2}p = p. \end{aligned}$$

Thus, if we use a single locus for classification, then the misclassification rate is p.

Suppose now that instead of being limited to one locus, we use k loci to classify. We represent the genotypes of a random individual at the k loci as random variables $L_1,\ldots,L_k$, denoting the total number of "1" alleles at the loci by the random variable $S = \sum_{i=1}^{k} L_i$. Assume that for all loci, allele frequencies in each population are the same as at the single locus described above, and that conditional on population membership, alleles at separate loci are independent. In other words, conditional on population membership, the sum S of