BioOne RESEARCH EVOLVED

A General Model of the Relationship between the Apportionment of Human Genetic Diversity and the Apportionment of Human Phenotypic Diversity

Author(s): Michael D. Edge and Noah A. Rosenberg
Source: Human Biology, 87(4):313-337.
Published By: Wayne State University Press
URL: http://www.bioone.org/doi/full/10.13110/humanbiology.87.4.0313

BioOne (www.bioone.org) is a nonprofit, online aggregation of core research in the biological, ecological, and environmental sciences. BioOne provides a sustainable online platform for over 170 journals and books published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Web site, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/page/terms_of_use.

Usage of BioOne content is strictly limited to personal, educational, and non-commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.

A General Model of the Relationship between the Apportionment of Human Genetic Diversity and the Apportionment of Human Phenotypic Diversity

Michael D. Edge[1]* *and Noah A. Rosenberg*[1]

ABSTRACT

Models that examine genetic differences between populations alongside a genotype–phenotype map can provide insight about phenotypic variation among groups. We generalize a simple model of a completely heritable, additive, selectively neutral quantitative trait to examine the relationship between single-locus genetic differentiation and phenotypic differentiation on quantitative traits. In agreement with similar efforts using different models, we show that the expected degree to which two groups differ on a neutral quantitative trait is not strongly affected by the number of genetic loci that influence the trait: neutral trait differences are expected to have a magnitude comparable to the genetic differences at a single neutral locus. We discuss this result with respect to population differences in disease phenotypes, arguing that although neutral genetic differences between populations can contribute to specific differences between populations in health outcomes, systematic patterns of difference that run in the same direction for many genetically independent health conditions are unlikely to be explained by neutral genetic differentiation.

1. Introduction

Since Lewontin's (1972) landmark partitioning of human genetic diversity, many studies have supported his claim that allele-frequency differences between geographically defined groups of people are relatively modest (Barbujani et al. 1997; Brown and Armelagos 2001; Rosenberg et al. 2002; Li et al. 2008). The findings of these previous studies have often been reported as estimates of F_{ST}, which can be interpreted as the proportion of variance in an allelic indicator variable attributable to allele-frequency differences between populations (Holsinger and 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).

Human F_{ST} estimates suggest that for phenotypes governed by a single typical genetic locus, population membership is likely to account for a relatively small proportion of the total variance of the trait. However, phenotypes are generally influenced by many loci, not just one. Large sets of loci can contain a great deal of information about population membership and can permit highly accurate ancestry inference, even if each locus has a small F_{ST} (Smouse et al. 1982; Bowcock et al. 1994; Mountain and Cavalli-Sforza 1997; Rosenberg

[1]Department of Biology, Stanford University, Stanford, California.

*Correspondence to: Michael D. Edge, Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305-5020. E-mail: medge@stanford.edu.

KEY WORDS: GENETIC DIFFERENTIATION, HEALTH DISPARITIES, POPULATION GENETICS, QUANTITATIVE GENETICS.

Human Biology, Fall 2015, v. 87, no. 4, pp. 313–337. Copyright © 2016 Wayne State University Press, Detroit, Michigan 48201

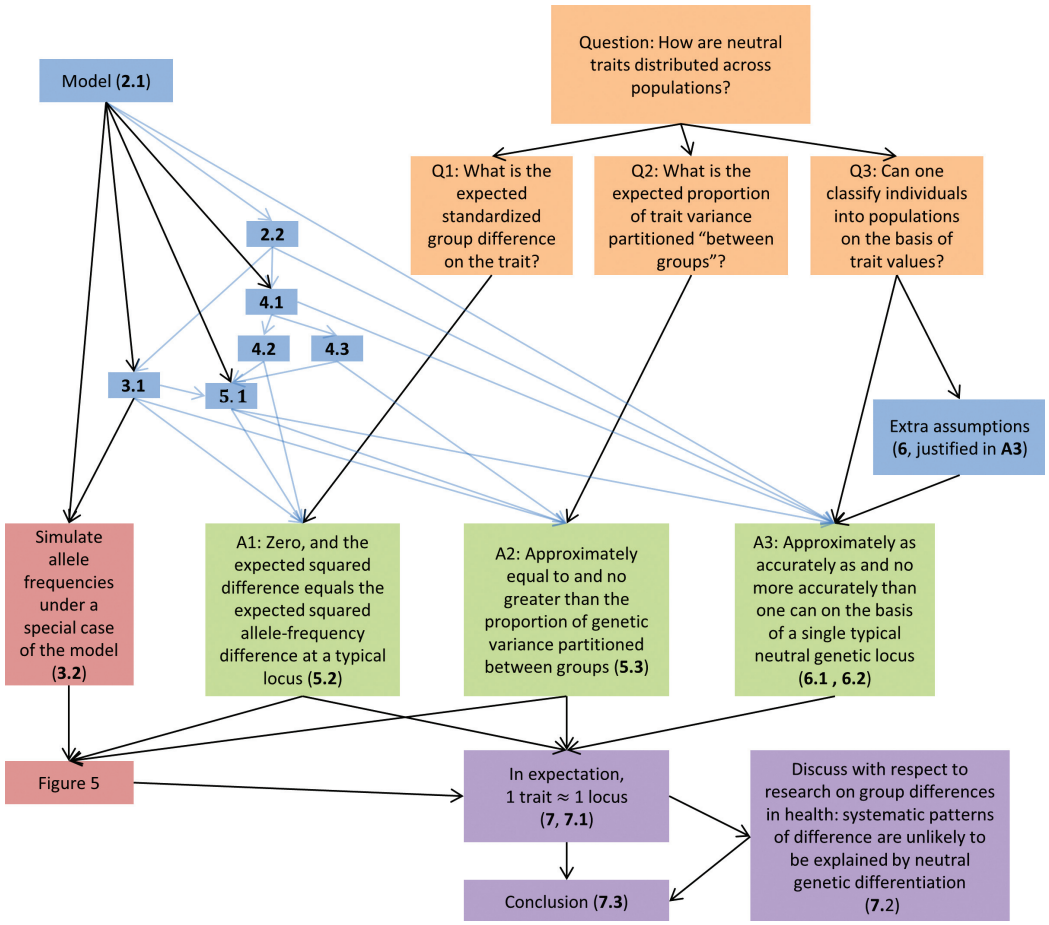

FIGURE 1. A conceptual map of this article. For boxes that correspond to specific subsections of the article, the subsection number is displayed in bold. Arrows indicate conceptual dependence, with blue arrows indicating that a subsection cites mathematical results obtained in another subsection. Motivating questions are in orange, mathematical machinery is in blue, simulations are in red, mathematical answers are in green, and interpretation of results is in purple.

et al. 2002; Bamshad et al. 2003; Edwards 2003; Li et al. 2008). Should we expect traits that are influenced by many loci to aggregate information about population membership across loci, leading to differences between populations that are more pronounced than those observed for traits influenced by fewer loci?

This question has been examined in many population-genetic and quantitative-genetic studies, both in theoretical models (Felsenstein 1973, 1986; Chakraborty and Nei 1982; Rogers and Harpending 1983; Lande 1992; Spitze 1993; Lynch and Spitze 1994; Whitlock 1999; Berg and Coop 2014; Edge and Rosenberg 2015) and in empirical applications in humans (Roseman 2004; Roseman and Weaver 2004; Weaver et al. 2007; Relethford 2010) and other organisms (for reviews, see Whitlock 2008; Leinonen et al. 2013), finding that, in the absence of selection, the expected degree to which groups differ on an additive, genetically determined trait does not depend on the number of loci that influence the trait. Put differently, a typical neutral trait conveys roughly the same degree of information about population membership as a single neutral locus, even if the trait is influenced by a large set of loci that would, if considered directly, permit accurate classification by population of origin.

Recently, to facilitate direct comparisons of multilocus genetic classification, single-locus genetic differentiation, and phenotypic differentiation, we developed a model that combines a simple model of multilocus genetic classification with a simple genotype–phenotype map. Our model enables genotype–phenotype comparisons to be performed in a statistical framework that permits exact computation and does not require detailed evolutionary assumptions (Edge and Rosenberg 2015). Our results agreed with those found with other models, highlighting the differences between polygenic phenotypic differentiation and information about population membership at multiple genetic loci.

In our past work (Edge and Rosenberg 2015), we applied strong assumptions about the allele-frequency distribution, and we examined only

Table 1. Summary of Notation

Symbol	Meaning
k	The number of loci that influence a quantitative trait
ℓ	The ploidy of the individuals being considered
M	An individual's population membership; takes values A and B
L_{ij}	An individual's allelic type at the jth allele at the ith locus; takes values 0 and 1
p_i	The frequency of the 1 allele at locus i in population A (Eq. 1)
q_i	The frequency of the 1 allele at locus i in population B (Eq. 1)
$\bar{p}$	The mean frequency of the 1 allele across loci in population A (Eq. 2)
$\bar{q}$	The mean frequency of the 1 allele across loci in population B (Eq. 2)
s_p^2	The variance across loci in the frequency of the 1 allele in population A (Eq. 3)
s_q^2	The variance across loci in the frequency of the 1 allele in population B (Eq. 3)
V_{ij}	An indicator for whether an individual's jth allele at the ith locus is a + allele (Eq. 4)
T	An individual's value for a quantitative trait (Eq. 4)
X_i	An indicator for whether the 0 or the 1 allele is also the + allele at locus i (Eq. 5)
S	The number of 1 alleles an individual carries (Eq. 6)
δ_i	The difference between populations in the frequency of the 1 allele at locus i (Eq. 8)
$\bar{\delta}$	The mean allele-frequency difference between populations, or $\bar{q} - \bar{p}$ (Eq. 9)
$\overline{\delta^2}$	The mean squared difference between populations in the frequency of the 1 allele (Eq. 10)
F_{ST}^k	The ratio of the mean (across k loci) within-population variance in an allelic indicator variable to the mean (across k loci) total variance in an allelic indicator variable (Eq. 14)
D_L^2	A function of F_{ST}^k that bears the same relationship to F_{ST}^k as the square of Cohen's d does to r^2 (Eq. 15)
$F_{ST(\ell)}^k$	A generalization of F_{ST}^k for the sum of ℓ independent allelic indicator variables (Eq. 16)
$D_{L(\ell)}^2$	A function of $F_{ST(\ell)}^k$ that bears the same relationship to $F_{ST(\ell)}^k$ as the square of Cohen's d does to r^2 (Eq. 17)
U_i	A transformation of the X_i: if $X_i = 1$, then $U_i = 1$; if $X_i = 0$, then $U_i = -1$ (Eq. 18)
D_T	The standardized difference between populations A and B on the trait (Eqs. 25, 34)
ρ_T^2	The proportion of the total variance in the trait attributable to between-population difference on the trait (Eqs. 26, 27, 41)
Q_{ST}	A quantitative-trait analogue of F_{ST}: if $\ell = 1$(haploid organisms), then $Q_{ST} = \rho_T^2$ (Eq. 28)
W_S	A quantity that equals 1 if an individual is classified into the wrong population on the basis of its value of S, and that equals 0 otherwise (Eq. 55)
W_T	A quantity that equals 1 if an individual is classified into the wrong population on the basis of its value of T, and that equals 0 otherwise (Eq. 56)

haploids. Here, we extend our earlier model to allow arbitrary allele-frequency distributions and arbitrary ploidy. Our results provide another way of establishing the result that between-group differentiation on a neutral trait mirrors between-group genetic differentiation at a neutral locus, one that makes minimal evolutionary assumptions. In Section 2, we describe our extended model. In Section 3, we define several measurements of between-group genetic differentiation. In Sections 4 and 5, we describe properties of two statistics that summarize the degree of difference between two populations on a quantitative trait. In Section 6, we introduce two simplifying assumptions that allow us to analyze the problem of inferring an individual's population of origin using either genetic or phenotypic information. Finally, we discuss the results with respect to the interpretation of population differences in disease phenotypes. Figure 1 provides a conceptual map of the structure of the article.

2. Preliminaries

2.1 Model

Our extended model is parallel to our previously reported model (Edge and Rosenberg 2015) and is similar to models used by Risch et al. (2002), Edwards (2003), and especially Tal (2012) to

investigate the problem of classifying individuals into populations using multilocus genetic data. For a summary of our notation, see Table 1.

We consider two populations of equal size, labeled A and B. In each individual, we consider k biallelic genetic loci. Each individual is ℓ-ploid ($\ell \geq 1$), carrying ℓ copies of each locus. At each locus, the allelic type more common in population B than in population A is labeled "1," and the other allelic type is labeled "0." Conditional on population membership, all of an individual's alleles are independent—both alleles at the same locus, as under Hardy-Weinberg equilibrium, and alleles at distinct loci, as under linkage equilibrium.

Let L_{ij} be an indicator random variable denoting whether the jth allele ($1 \leq j \leq \ell$) at locus i is the "1" allele, and let M be a random variable that represents an individual's population membership and that takes values A and B. The conditional probabilities that $L_{ij} = 1$ are

$$\begin{aligned} P(L_{ij} = 1 \mid M = A) &= p_i, \\ P(L_{ij} = 1 \mid M = B) &= q_i. \end{aligned} \tag{1}$$

The p_i and q_i obey $0 \leq p_i \leq q_i \leq 1$. The constraint $q_i \geq p_i$ holds because, by definition, the 1 allele is more common in population B than in population A. We also assume that, for at least one value of i, p_i or q_i does not equal 0 or 1; and for the limiting results in Section 6, we assume that, as the number of loci k approaches infinity, the number of loci at which $p_i \in (0, 1)$ and the number of loci at which $q_i \in (0, 1)$ both approach infinity.

Define $\bar{p}$ and $\bar{q}$ as the means across loci of the allele frequencies p_i and q_i:

$$\begin{aligned} \bar{p} &= \frac{1}{k}\sum_{i=1}^{k} p_i, \\ \bar{q} &= \frac{1}{k}\sum_{i=1}^{k} q_i. \end{aligned} \tag{2}$$

Define s_p^2 and s_q^2 as the variances across loci of the p_i and q_i, though they are not probabilistic variances because the p_i and q_i are nonrandom:

$$\begin{aligned} s_p^2 &= \frac{1}{k}\sum_{i=1}^{k} (p_i - \bar{p})^2, \\ s_q^2 &= \frac{1}{k}\sum_{i=1}^{k} (q_i - \bar{q})^2. \end{aligned} \tag{3}$$

We model a completely heritable, selectively neutral, additively determined trait as a function of the k loci described above. Specifically, an individual's value on the trait—represented by a random variable T—is a weighted sum of the individual's L_{ij} values. As in our previous work (Edge and Rosenberg 2015, Eq. 9), the weights are determined by labels at each locus, where for each trait we label one allele at each locus—either the 0 allele or the 1 allele—as the "+" allele and the other allele as the "–" allele. T is then equal to the number of + alleles carried by the individual. That is,

$$T = \sum_{i=1}^{k}\sum_{j=1}^{\ell} V_{ij}, \tag{4}$$

where $V_{ij} = 1$ if the jth allele at the ith locus is a + allele and $V_{ij} = 0$ otherwise.

Again following our previous work (Edge and Rosenberg 2015, Eq. 7), we assume that whether an allele is more common in population B than in population A (i.e., labeled "1") is independent of whether it is associated with larger trait values than the other allele (i.e., labeled "+"). This claim amounts to assuming that the alleles at k loci have not been under selection and have reached their current frequencies independently of their effect on the trait. We express this assumption with the random variable X_i, with $X_i = 0$ if the 0 allele and the + allele are identical at the ith locus and $X_i = 1$ if the 1 allele and the + allele are identical at the ith locus. Each trait is associated with a set of k values for the X_i, and for each of the k loci,

$$P(X_i = 0) = P(X_i = 1) = \tfrac{1}{2} \tag{5}$$

independently of the X_i for the other loci.

We introduce a statistic for comparison with the trait value T. We summarize the information about population membership available at an individual's k loci with the genotypic statistic S, the total number of 1 alleles—that is, alleles that are more common in population B than in population A—at k loci:

$$S = \sum_{i=1}^{k}\sum_{j=1}^{\ell} L_{ij}. \tag{6}$$

S is not generally an optimal basis for distinguishing members of population A and B (e.g., Tal 2012)—in principle, we could improve classification by more heavily weighting loci that have a greater allele frequency difference between populations—but we will show that classifications based on S approach perfect accuracy as k increases, as long as $\bar{p} \neq \bar{q}$.

Figure 2 shows a schematic of our model. The

model reduces to the one described in Edge and Rosenberg (2015) if one assumes (a) that $p_i = p$ and $q_i = q$ for all i, (b) that $q = 1 - p$, and (c) that the organisms being examined are haploid ($\ell = 1$).

2.2 The Poisson Binomial Distribution

Under our model, many relevant quantities have a Poisson binomial distribution, which arises when independent Bernoulli trials with possibly varying success probabilities are summed. By the central limit theorem, the Poisson binomial distribution converges to a normal distribution as the number of terms summed increases without bound, provided that the sum of the variances of the Bernoulli random variables approaches infinity (Deheuvels et al. 1989, Theorem 1.1). If Z is a Poisson binomial random variable with probabilities $p_1, \ldots, p_k$, then

$$\begin{aligned} E(Z) &= \sum_{i=1}^{k} p_i = k\overline{p}, \\ \mathrm{Var}(Z) &= \sum_{i=1}^{k} p_i(1-p_i) = k\overline{p}(1-\overline{p}) - ks_p^2, \end{aligned} \tag{7}$$

where $\overline{p}$ is as in Eq. 2 and s_p^2 is as in Eq. 3 (e.g., Edwards 1960).

3. Genetic Differentiation between Populations at a Single Locus

On the basis of our model, we define several statistics measuring the degree of genetic differentiation between populations at a single typical locus. In Section 5, we use these statistics to compare the degree of genetic differentiation at a typical locus to the expected degree of difference between populations in a neutral trait.

3.1 Single-Locus Differentiation Measures

One summary of the degree of single-locus genetic differentiation between populations is the difference between populations in the frequency of the 1 allele at the locus:

$$\delta_i = q_i - p_i. \tag{8}$$

We also define

$$\overline{\delta} = \frac{1}{k}\sum_{i=1}^{k} \delta_i = \overline{q} - \overline{p}, \tag{9}$$

$$\overline{\delta^2} = \frac{1}{k}\sum_{i=1}^{k} \delta_i^2 = \overline{\delta}^2 + s_\delta^2, \tag{10}$$

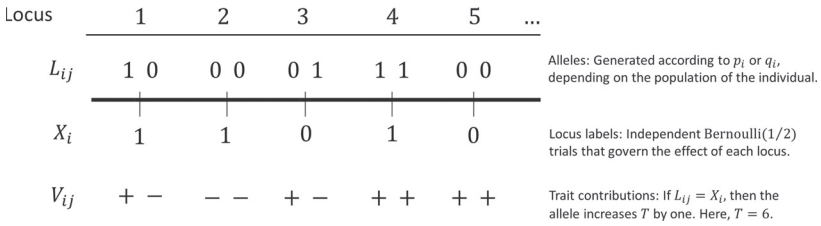

FIGURE 2. A schematic of our model for generating a quantitative trait. Five loci are shown for a diploid individual. The L_{ij} are the individual's alleles, which, conditional on population membership, are independent Bernoulli trials with probability at locus i equal either to p_i (if the individual is drawn from population A) or to q_i (if the individual is drawn from population B). At each locus, the frequency of the 1 allele is at least as large in population B as it is in population A. The X_i are labels indicating which allele at locus i leads to larger values of T; they are independent Bernoulli trials, each with probability ½. If an individual's jth allele at locus i (L_{ij}) matches the allele that leads to larger values of the trait for that locus (X_i), then V_{ij} takes the value "+"; otherwise, V_{ij} takes the value "–". T is equal to the sum of + alleles carried by the individual. In the case pictured, $T = 6$.

where $s_\delta^2 = \Sigma_{i=1}^{k}(\delta_i - \overline{\delta})^2/k$, and

$$\overline{\delta^4} = \frac{1}{k}\sum_{i=1}^{k} \delta_i^4 = \overline{\delta^2}^2 + s_{\delta^2}^2, \tag{11}$$

where $s_{\delta^2}^2 = \Sigma_{i=1}^{k}(\delta_i^2 - \overline{\delta^2})^2/k$.

The quantity δ (Eq. 8), the difference in population frequencies of one specific allele at a biallelic locus, is closely related to F_{ST} for a single locus. Specifically, with two populations and dropping the subscript i, single-locus F_{ST}, which we write as F_{ST}^1, is

$$\begin{aligned} F_{ST}^1 &= \frac{\delta^2}{4\left(\frac{p+q}{2}\right)\left(1-\frac{p+q}{2}\right)} \\ &= \frac{\delta^2}{2[p(1-p)+q(1-q)]+\delta^2} \end{aligned} \tag{12}$$

(e.g., Weir 1996; Rosenberg et al. 2003, Eq. 8; Holsinger and Weir 2009, Eq. 4). $F_{ST}^1 \in [\delta^2, \delta/(2-\delta)]$, and δ^2 never deviates from F_{ST}^1 by more than $(5\sqrt{5}-11)/2 \approx 0.0902$ (Rosenberg et al. 2003). F_{ST}^1 can be interpreted in terms of a ratio involving heterozygosity in the subpopulations and in the total population (Nei 1973), or as a ratio of variance components (Holsinger and Weir 2009); we emphasize the latter interpretation. Specifically, if L is an allelic indicator variable representing a single copy of a locus and M denotes population membership, then for a single locus,

$$F_{ST}^1 = \frac{\mathrm{Var}_M[E(L|M)]}{\mathrm{Var}(L)}. \tag{13}$$

The subscript M indicates that the variance in the numerator is taken with respect to group

membership. Equation 13 can be verified using the law of total variance, noting that $\mathrm{Var}_M[E(L|M)] = \delta^2/4$ and that $E_M[\mathrm{Var}(L|M)] = [p(1-p) + q(1-q)]/2$, and comparing with Eq. 12.

To summarize the overall degree of genetic differentiation at a group of k loci, we define an F_{ST} measure that summarizes the typical degree of differentiation at a locus chosen from a set of k loci, which we write as F_{ST}^k. For locus i, L_{i1} is an allelic indicator variable representing one copy of the locus. To compute F_{ST}^k, we sum the variance components that appear in Eq. 13 across all k loci and take their ratio:

$$F_{ST}^k = \frac{\sum_{i=1}^{k} \mathrm{Var}_M[E(L_{i1}\,|\,M)]}{\sum_{i=1}^{k} \mathrm{Var}(L_{i1})} = \frac{\overline{\delta^2}}{2[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2] + \overline{\delta^2}}. \quad (14)$$

This ratio is analogous to estimators of F_{ST} that involve a ratio of two variance estimates (e.g., Weir and Cockerham 1984, Eq. 10). In our model, however, F_{ST}^k is known and not estimated because the allele frequencies are known. Though F_{ST}^k is intended as an index of the degree of genetic differentiation at a single typical locus, F_{ST}^k is not equal to the mean of the F_{ST}^1 values computed for each locus separately. Rather, it is the ratio of the mean across loci of the between-group variance in allelic type to the mean across loci of the total variance in allelic type.

One interpretation of F_{ST} is as the proportion of the variance removed from an indicator variable for one copy of an allele by conditioning on population membership. F_{ST} is thus analogous to r^2, a measurement of effect size commonly used in meta-analysis, which can be interpreted as the proportion of variance in a dependent variable that is removed by conditioning on an independent variable (Fox 1997: 94). Another commonly used effect-size measurement applicable to differences between two groups is Cohen's d (Cohen 1988), the difference in group means on a dependent variable divided by the square root of the mean across groups of within-group variances of the independent variable. For equally sized groups, Cohen's d is related to r^2 by $d^2 = 4r^2/(1 - r^2)$ (by inverting Rosenthal 1994, Eqs. 16–24). By analogy, we define another measurement of between-group genetic differentiation across a set of loci:

$$D_L^2 = \frac{4F_{ST}^k}{1-F_{ST}^k} = \frac{\overline{\delta^2}}{[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]/2}. \quad (15)$$

D_L^2 for a set of loci is not generally equal to the mean across loci of the value that would result by applying Eq. 15 to each locus separately. Rather, like F_{ST}^k, D_L^2 is a ratio of two means across loci—the mean of the δ_i^2 and the mean within-group variance in allelic type.

F_{ST}^k is a variance partition for allelic indicator variables representing one copy of a locus. At a diploid or polyploid biallelic locus, each copy of the locus provides information about population membership, so more information is available at the locus than is reflected in one copy. We thus define an analogue of F_{ST}^k for a set of ℓ-ploid loci by partitioning the variance of the sum of the number of 1 alleles at each locus into between-group and within-group components. For a single locus, the between-group variance of the sum is

$$\begin{aligned}
&\mathrm{Var}_M\left[E\left(\sum_{j=1}^{\ell} L_{ij}\,\middle|\,M\right)\right] \\
&= \sum_{m\in\{A,B\}} P(M=m)\left[E\left(\sum_{j=1}^{\ell} L_{ij}\,\middle|\,M=m\right) - E\left(\sum_{j=1}^{\ell} L_{ij}\right)\right]^2 \\
&= \frac{1}{4}\left[E\left(\sum_{j=1}^{\ell} L_{ij}\,\middle|\,M=A\right) - E\left(\sum_{j=1}^{\ell} L_{ij}\,\middle|\,M=B\right)\right]^2 \\
&= \frac{\ell^2}{4}\delta_i^2,
\end{aligned}$$

and by the independence of the allelic copies at a single locus, the within-group variance is

$$\begin{aligned}
&E_M\left[\mathrm{Var}\left(\sum_{j=1}^{\ell} L_{ij}\,\middle|\,M\right)\right] \\
&= \sum_{m\in\{A,B\}} P(M=m)\cdot \\
&\sum_{j=1}^{\ell} P(L_{ij}=1|\,M=m)[1-P(L_{ij}=1|\,M=m)] \\
&= \frac{\ell}{2}[p_i(1-p_i) + q_i(1-q_i)].
\end{aligned}$$

To define $F^k_{ST(\ell)}$, we sum these terms across loci to construct a ratio of the between-group variance to the total variance:

$$F^k_{ST(\ell)} = \frac{\sum_{i=1}^{k} \mathrm{Var}_M\left[E\left(\sum_{j=1}^{\ell} L_{ij} \,\middle|\, M\right)\right]}{\sum_{i=1}^{k} \mathrm{Var}\left(\sum_{j=1}^{\ell} L_{ij}\right)} = \frac{\ell\overline{\delta^2}}{2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \ell\overline{\delta^2}}. \quad (16)$$

We show in Appendix 1 that $F^k_{ST(\ell)} \in [F^k_{ST},\ \ell F^k_{ST})$ with $F^k_{ST(\ell)} = F^k_{ST}$ if and only if $F^k_{ST} = 0$ or $F^k_{ST} = 1$. Figure 3 shows the relationship between F^k_{ST} and $F^k_{ST(\ell)}$ for several values of ℓ, illustrating the relative increase in $F^k_{ST(\ell)}$ compared with F^k_{ST} as ℓ increases. Figure 3 also illustrates, as shown in Appendix 1, that $F^k_{ST(\ell)}$ is comparable to ℓF^k_{ST} for F^k_{ST} close to 0 and comparable to F^k_{ST} for F^k_{ST} close to 1.

Similarly, we can define an analogue of D^2_L for an ℓ-ploid locus:

$$D^2_{L(\ell)} = \frac{4F^k_{ST(\ell)}}{1-F^k_{ST(\ell)}} = \frac{\ell\overline{\delta^2}}{[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2]/2} = \ell D^2_L. \quad (17)$$

Whereas F^k_{ST} and D^2_L can be viewed as indices of the amount of information about population membership available in a single copy of a typical locus, $F^k_{ST(\ell)}$ and $D^2_{L(\ell)}$ assess the total amount of population membership information at a typical locus, considering all ℓ copies.

3.2 Simulation-Based Allele Frequency Differences

Because some of our results depend on specific characteristics of the p_i and q_i, we simulated allele frequencies under a model similar to that of Nicholson et al. (2002) to obtain suitable example distributions for the p_i and q_i (see Figure 4 for a schematic). Specifically, we generated allele frequencies for derived alleles in an ancestral population according to the neutral site frequency spectrum with $2N = 20{,}000$, choosing each allele frequency π_i according to $P(\pi_i = j/(2N)) \propto 1/j$ (Charlesworth and Charlesworth 2010, Eq. B6.6.1). To simulate drift after divergence, we produced postdivergence allele frequencies by adding to each "ancestral" allele frequency π_i an independently drawn Normal$(0, 0.3\pi_i(1-\pi_i))$ random number, where 0.3 is chosen so that F^k_{ST} approximates worldwide human F_{ST} estimates. Any postdivergence allele frequencies less than 0 or greater than 1 were set to 0 or 1, respectively. After simulating postdivergence frequencies of the derived allele independently in two populations, we assigned the frequencies of either the ancestral or the derived allele in each population to be p_i and q_i, requiring $q_i \geq p_i$. We generated 10^6 pairs of allele frequencies (p_i, q_i) after removing loci at which the same allele fixed in both populations. (Such loci do not contribute to F^k_{ST} or D^2_L.) For our simulated allele frequencies, $\bar{p} \approx 0.457$, $\bar{q} \approx 0.542$, $s_p^2 \approx s_q^2 \approx 0.191$, $\bar{\delta} \approx 0.086$, $\overline{\delta^2} \approx 0.025$, $\overline{\delta^4} \approx 0.006$, $F^k_{ST} \approx 0.099$, and $D^2_L \approx 0.440$. The F^k_{ST} value of 0.099 is similar to estimates of F_{ST} for human populations.

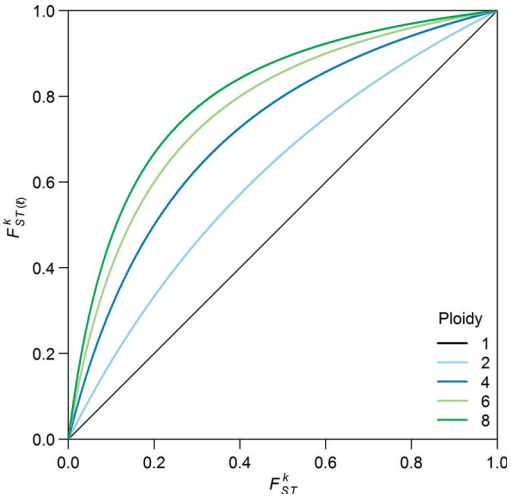

FIGURE 3. The relationship between F^k_{ST} (Eq. 14), which partitions the variance of allelic indicator variables representing a single copy of each locus, and $F^k_{ST(\ell)}$ (Eq. 16), which partitions the variance of sums of ℓ allelic indicator variables at each locus. Thus, for haploids ($\ell = 1$), $F^k_{ST(\ell)} = F^k_{ST}$. For higher ploidy, $F^k_{ST(\ell)} \in [F^k_{ST},\ \ell F^k_{ST})$, with $F^k_{ST(\ell)} = F^k_{ST}$ if and only if $F^k_{ST} = 0$ or $F^k_{ST} = 1$ (see Appendix 1). The plot is obtained from Eq. A1.4.

4. Properties of the Trait Value *T* Conditional on the Labeling X_i

We next consider the distribution and properties of the trait value T in each population. In this section, we condition on the labeling of the alleles at each locus $X_1, X_2, \ldots, X_k$. These labels determine, for each locus, whether the 1 or the 0 allele increases an individual's trait value. In Section 5, we remove this condition and consider the expected behavior of the trait value under random assignment of the labels.

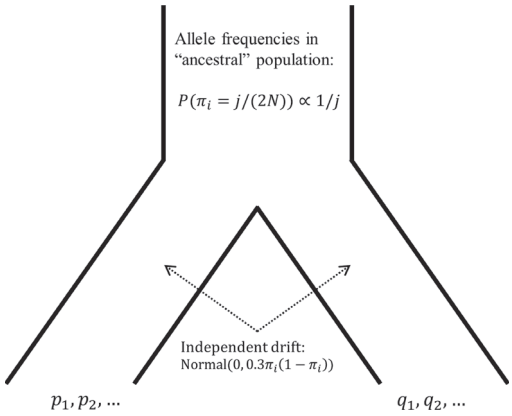

FIGURE 4. A schematic of the drift model used to simulate allele frequencies (see Section 3.2). Derived allele frequencies in an "ancestral" population are drawn according to the neutral site frequency spectrum. Following a split, the two subpopulations drift independently, with the drift represented by a truncated normal variate with expectation 0. After drift, for each locus *i*, the allele with greater frequency in population B than in population A is identified, its frequency in population A is labeled p_i, and its frequency in population B is labeled q_i.

It is convenient to define a transformation of the labels:

$$U_i = 2X_i - 1. \tag{18}$$

If $X_i = 1$ and the 1 allele is the + allele, then $U_i = 1$, and if $X_i = 0$ and the 1 allele is the – allele, then $U_i = -1$.

4.1 Distribution of *T* within Each Population Given the Labeling of the Alleles

In either population, conditional on the labeling of the alleles,

$$\begin{aligned}&(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\})\\&=\sum_{i:x_i=1}\sum_{j=1}^{\ell}L_{ij}+\sum_{i:x_i=0}\sum_{j=1}^{\ell}(1-L_{ij}).\end{aligned}$$

Because the L_{ij} are independent Bernoulli random variables with different success probabilities, T has a Poisson binomial distribution in each population. Specifically, within population A, each of the ℓ allelic copies at each locus at which $x_i = 1$ increases T by 1 with probability p_i, and each of the ℓ allelic copies at each locus at which $x_i = 0$ increases T by 1 with probability $1 - p_i$. Within population B, the same statement holds if p_i is replaced by q_i.

By the properties of the Poisson binomial distribution (Eq. 7), the expectations of T in populations A and B conditional on the labeling are then

$$\begin{aligned}&E(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=A)\\&\quad=\ell\left[\sum_{i:x_i=1}p_i+\sum_{i:x_i=0}(1-p_i)\right],\\&E(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=B)\\&\quad=\ell\left[\sum_{i:x_i=1}q_i+\sum_{i:x_i=0}(1-q_i)\right].\end{aligned} \tag{19}$$

The difference in the conditional expectations is then

$$\begin{aligned}&E(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=B)\\&-E(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=A)\\&=\ell\left[\sum_{i:x_i=1}q_i+\sum_{i:x_i=0}(1-q_i)\right]-\ell\left[\sum_{i:x_i=1}p_i+\sum_{i:x_i=0}(1-p_i)\right]\\&=\ell\left[\sum_{i:x_i=1}(q_i-p_i)-\sum_{i:x_i=0}(p_i-q_i)\right]=\ell\sum_{i=1}^{k}\delta_i u_i,\end{aligned} \tag{20}$$

where, analogously to Eq. 18, $u_i = 2x_i - 1$.

By Eq. 20 and the fact that the populations have equal size so that $P(M = A) = P(M = B) = ½$, the variance across populations of the conditional expectation of T is

$$\begin{aligned}&\mathrm{Var}_M[E(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M]\\&\quad=\frac{\ell^2}{4}\left(\sum_{i=1}^{k}\delta_i u_i\right)^2.\end{aligned} \tag{21}$$

By the properties of the Poisson binomial distribution (Eq. 7), the conditional variance of the trait in population A is

$$\begin{aligned}&\mathrm{Var}(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=A)\\&=\sum_{i:x_i=1}\sum_{j=1}^{\ell}p_i(1-p_i)+\sum_{i:x_i=0}\sum_{j=1}^{\ell}p_i(1-p_i)\\&=\ell\sum_{i=1}^{k}p_i(1-p_i)=\ell k[\bar{p}(1-\bar{p})-s_p^2];\end{aligned}$$

the last step follows from the simplification of the variance in Eq. 7. Because this quantity does not depend on $\{X_1, \ldots, X_k\}$, we can remove the condition on $\{x_1, \ldots, x_k\}$, giving

$$\mathrm{Var}(T|M = A) = \ell k[\bar{p}(1 - \bar{p}) - s_p^2]. \tag{22}$$

Similarly, the variance of T in population B is

$$\mathrm{Var}(T|M = B) = \ell k[\bar{q}(1 - \bar{q}) - s_q^2]. \tag{23}$$

We use the conditional expectations and variances of T in the two populations to define several measurements of the degree of difference between populations on the trait.

4.2 The Standardized Difference in Trait Means, D_T, Given the Labeling of the Alleles

We consider three indices of the degree of difference between populations on the trait—two here, and a third we defer to Section 6. The first is the

standardized difference in population means for the trait, D_T, which is the difference between population trait means divided by the square root of the mean across populations of within-population trait variances. D_T is an instance of the Cohen's d measure of effect size (Cohen 1988). In this case, conditional on the labeling, D_T is

$$\begin{aligned}&(D_T \mid \{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\})\\&=[E(T \mid \{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\}, M=B)\\&-E(T \mid \{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\}, M=A)]/\\&\sqrt{\frac{\mathrm{Var}(T \mid M=A)+\mathrm{Var}(T \mid M=B)}{2}}.\end{aligned}$$

The numerator is given in Eq. 20. By Eqs. 22 and 23 and the fact that the two populations are assumed to be the same size, the square of the denominator is

$$\begin{aligned}&E_M[\mathrm{Var}(T \mid M)]\\&=\frac{\mathrm{Var}(T \mid M=A)+\mathrm{Var}(T \mid M=B)}{2}\\&=\frac{\ell k}{2}[\overline{p}(1-\overline{p})-s_p^2+\overline{q}(1-\overline{q})-s_q^2].\end{aligned} \tag{24}$$

Therefore, combining Eqs. 20 and 24,

$$\begin{aligned}&(D_T \mid \{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\})\\&=\frac{\sqrt{\ell}\sum_{i=1}^{k}\delta_i u_i}{\sqrt{k[\overline{p}(1-\overline{p})-s_p^2+\overline{q}(1-\overline{q})-s_q^2]/2}},\end{aligned} \tag{25}$$

where again, as in Eq. 18, $u_i = 2x_i - 1$. In Section 5, we study the distribution of D_T across different labelings of the alleles.

4.3 Partitioning the Variance of the Trait Given the Labeling of the Alleles: ρ_T^2 and Q_{ST}

A second measure of between-population difference on the trait is the proportion of the trait's variance attributable to difference between populations. We label this proportion ρ_T^2, with

$$\begin{aligned}\rho_T^2&=\frac{\mathrm{Var}_M[E(T \mid M)]}{\mathrm{Var}(T)}\\&=\frac{\mathrm{Var}_M[E(T \mid M)]}{E_M[\mathrm{Var}(T \mid M)]+\mathrm{Var}_M[E(T \mid M)]}.\end{aligned} \tag{26}$$

The last step follows from the law of total variance. Conditional on the labeling $\{X_1, \ldots, X_k\}$, the numerator appears in Eq. 21, and the denominator is the sum of the expressions in Eqs. 21 and 24. Thus, by Eq. 26, conditional on the labeling of the alleles for a given trait,

$$\begin{aligned}&(\rho_T^2 \mid \{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\})\\&=\frac{\ell\left(\sum_{i=1}^{k}\delta_i u_i\right)^2}{2k[\overline{p}(1-\overline{p})-s_p^2+\overline{q}(1-\overline{q})-s_q^2]+\ell\left(\sum_{i=1}^{k}\delta_i u_i\right)^2}.\end{aligned} \tag{27}$$

The proportion ρ_T^2 is related to Q_{ST}, which is an analogue of F_{ST} developed for quantitative traits. For haploids, Q_{ST} is the proportion of the heritable variance in a quantitative trait attributable to genetic differences between populations (Whitlock 2008). Because we have assumed that the trait we examine is completely heritable, $\rho_T^2 = Q_{ST}$ for haploids. Q_{ST} is defined so that, like F_{ST}, it does not depend on ploidy, which means that $\rho_T^2 \neq Q_{ST}$ for ploidy $\ell > 1$ (unless $\rho_T^2 = 0$ or $\rho_T^2 = 1$). For diploids, again invoking the assumption of perfect heritability of the trait, Q_{ST} is

$$Q_{ST}=\frac{\mathrm{Var}_M[E(T \mid M)]}{2E_M[\mathrm{Var}(T \mid M)]+\mathrm{Var}_M[E(T \mid M)]}$$

(Whitlock 2008), and by analogy, for ℓ-ploid organisms,

$$\begin{aligned}Q_{ST}&=\frac{\mathrm{Var}_M[E(T \mid M)]}{\ell E_M[\mathrm{Var}(T \mid M)]+\mathrm{Var}_M[E(T \mid M)]}\\&=\frac{\left(\sum_{i=1}^{k}\delta_i u_i\right)^2}{2k[\overline{p}(1-\overline{p})-s_p^2+\overline{q}(1-\overline{q})-s_q^2]+\left(\sum_{i=1}^{k}\delta_i u_i\right)^2}.\end{aligned} \tag{28}$$

Thus, regardless of ploidy ℓ, Q_{ST} is obtained from the expression in Eq. 27 by setting ℓ to 1. The relationship between Q_{ST} and ρ_T^2 is exactly the same as the relationship between F_{ST}^k and $F_{ST(\ell)}^k$ (Figure 3 and Appendix 1); that is, if $Q_{ST} = F_{ST}^k$, then $\rho_T^2 = F_{ST(\ell)}^k$.

5. Properties of D_T and ρ_T^2 across Different Labelings of the Alleles

In this section, we consider properties of the trait value T across different traits, which may have different allelic labels (X_i), so that each trait has its own locus-specific effects for the alleles. Specifically, we consider two indices of the degree of difference between populations on the trait

defined in Section 4, the standardized group difference, D_T (Section 4.2), and the proportion of trait variance that is attributable to between-group differences, ρ_T^2 (Section 4.3).

5.1 Properties of $\Sigma_{i=1}^{k} \delta_i U_i$

One random variable that appears in expressions for both D_T (Eq. 25) and ρ_T^2 (Eq. 27) is $\Sigma_{i=1}^{k} \delta_i U_i$, where U_i is a function of the labels X_i that determine which allele at locus i is the + allele (Eq. 18), taking a value of either –1 or 1 with probability ½ each, and δ_i is the difference between populations in the frequency of the 1 allele (Eq. 8). We give the relevant moments of $\Sigma_{i=1}^{k} \delta_i U_i$ here for later reference.

We note first that, for all i and for integers $n \geq 0$, the odd and even moments of the U_i obey

$$E(U_i^{2n+1}) = 0, \tag{29}$$

$$E(U_i^{2n}) = 1. \tag{30}$$

Thus, by Eq. 29, for $n \in \{0, 1, 2, \dots\}$,

$$E[(\Sigma_{i=1}^{k} \delta_i U_i)^{2n+1}] = 0. \tag{31}$$

The second moment is

$$\begin{aligned} &E\left[\left(\sum_{i=1}^{k} \delta_i U_i\right)^2\right] \\ &= \sum_{i=1}^{k} \delta_i^2 E(U_i^2) + \sum_{i=1}^{k} \sum_{j \neq i} \delta_i \delta_j E(U_i U_j) \\ &= \sum_{i=1}^{k} \delta_i^2 = k\overline{\delta^2}, \end{aligned} \tag{32}$$

by Eq. 30 and because, by the independence of the U_i, $E(U_iU_j) = E(U_i)E(U_j) = 0$ for all $i \neq j$.

We show in Appendix 2 that the fourth moment is

$$E\left[\left(\sum_{i=1}^{k} \delta_i U_i\right)^4\right] = 3k^2\overline{\delta^2}^2 - 2k\left(\overline{\delta^2}^2 + s_{\delta^2}^2\right). \tag{33}$$

5.2 The Standardized Group Difference in Trait Means, D_T

Removing the condition on the labels in Eq. 25, u_i becomes the random variable U_i (Eq. 18), and D_T becomes the random variable

$$D_T = \frac{\sqrt{\ell} \sum_{i=1}^{k} \delta_i U_i}{\sqrt{k[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]/2}}. \tag{34}$$

By Eq. 29,

$$E(D_T) = \frac{\sqrt{\ell} \sum_{i=1}^{k} \delta_i E(U_i)}{\sqrt{k[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]/2}} = 0. \tag{35}$$

Equation 35 reflects the symmetry of the distribution of D_T around 0. By Eqs. 32, 34, and 35,

$$\begin{aligned} \mathrm{Var}(D_T) &= E(D_T^2) - E(D_T)^2 = E(D_T^2) \\ &= E\left[\left(\frac{\sqrt{\ell} \sum_{i=1}^{k} \delta_i U_i}{\sqrt{k[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]/2}}\right)^2\right] \\ &= \frac{\ell \overline{\delta^2}}{[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]/2}, \end{aligned} \tag{36}$$

where $\overline{\delta^2}$ is as defined in Eq. 10.

$E(D_T^2)$ (Eq. 36) is one measurement of the typical size of the between-group difference in trait means, irrespective of its direction. For fixed $\overline{p}$ and $\overline{q}$, $E(D_T^2)$ is usually larger if $k > 1$ than if $k = 1$ because of variation in the allele frequencies: $s_p^2 = s_q^2 = s_\delta^2 = 0$ if $k = 1$, but each may be positive if $k > 1$, and positive values of each of these terms increase $E(D_T^2)$. Nonetheless, $E(D_T^2)$ does not grow without bound as k increases, and it is equal to one of our indices of between-group genetic differentiation at a single locus, $D_{L(\ell)}^2$ (Eq. 17),

$$E(D_T^2) = D_{L(\ell)}^2 = \ell D_L^2. \tag{37}$$

Thus, though $E(D_T^2)$ increases with higher ploidy, it does not necessarily increase as the number of loci k influencing the trait increases (Figure 5A). Equation 36 reduces to the results we showed for D_T^2 in our previous work (Edge and Rosenberg 2015, Eqs. 37, 38) under the more restrictive assumptions we used there. The correspondences between the main results in this article and the main results in Edge and Rosenberg (2015) are summarized in Table 2.

In addition to the expectation of D_T^2, we may wish to know its variance—do traits influenced by many loci vary widely in their level of between-population difference? By Eq. 33,

$$\begin{aligned} E(D_T^4) &= \frac{\ell^2\left[3k^2\overline{\delta^2}^2 - 2k\left(\overline{\delta^2}^2 + s_{\delta^2}^2\right)\right]}{k^2[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]^2/4} \\ &= \frac{\ell^2}{[\overline{p}(1-\overline{p}) - s_p^2 + \overline{q}(1-\overline{q}) - s_q^2]^2/4} \\ &\quad \times \left(3\overline{\delta^2}^2 - 2\left(\overline{\delta^2}^2 + s_{\delta^2}^2\right)/k\right). \end{aligned} \tag{38}$$

The required variance, calculated as $E(D_T^4) - E(D_T^2)^2$, is, by Eqs. 36 and 38,

$$\mathrm{Var}(D_T^2) = \frac{2\ell^2}{[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2]^2/4} \times \left(\overline{\delta^2}^2 - \left(\overline{\delta^2}^2 + s_{\delta^2}^2\right)/k\right). \quad (39)$$

If $k = 1$, then $s_{\delta^2}^2 = 0$, and $\mathrm{Var}(D_T^2) = 0$. As k increases, $\mathrm{Var}(D_T^2)$ approaches

$$\lim_{k\to\infty} \mathrm{Var}(D_T^2) = \frac{2\ell^2\overline{\delta^2}^2}{[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2]^2/4} = 2\mathrm{Var}(D_T)^2. \quad (40)$$

Equations 39 and 40 indicate that as the number of loci k increases, the variance of D_T^2 does increase, but it asymptotes to a limit that does not depend on k (Figure 5B).

5.3 Properties of ρ_T^2 and Q_{ST}

Removing the condition on the labels in Eq. 27, u_i becomes the random variable U_i (Eq. 18), and ρ_T^2 becomes a random variable

$$\rho_T^2 = \ell\left(\sum_{i=1}^k \delta_i U_i\right)^2 / \{2k[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \ell\left(\sum_{i=1}^k \delta_i U_i\right)^2\}. \quad (41)$$

To describe the behavior of ρ_T^2 across different traits, we approximate $E(\rho_T^2)$ by replacing $(\Sigma_{i=1}^k \delta_i U_i)^2$ in Eq. 41 with its expectation, motivated by a Taylor approximation argument. Making the substitution $Y = (\ell/2)(\Sigma_{i=1}^k \delta_i U_i)^2$, we have

$$\rho_T^2 = \frac{Y}{k[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + Y} = g(Y). \quad (42)$$

Defining $\mu_Y = E(Y)$, a first-order Taylor series expansion for $g(Y)$ around $Y = \mu_Y$ gives

$$\rho_T^2 = g(Y) \approx g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y),$$

and taking the expectation gives

$$E(\rho_T^2) = E[g(Y)] \approx g(\mu_Y) + g'(\mu_Y)E(Y - \mu_Y) = g(\mu_Y).$$

By Eq. 32, $E[(\Sigma_{i=1}^k \delta_i U_i)^2] = k\overline{\delta^2}$. Substituting $E(Y) = (\ell/2)k\overline{\delta^2}$ for Y in Eq. 42 gives

$$E(\rho_T^2) \approx \frac{\ell\overline{\delta^2}}{2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \ell\overline{\delta^2}}. \quad (43)$$

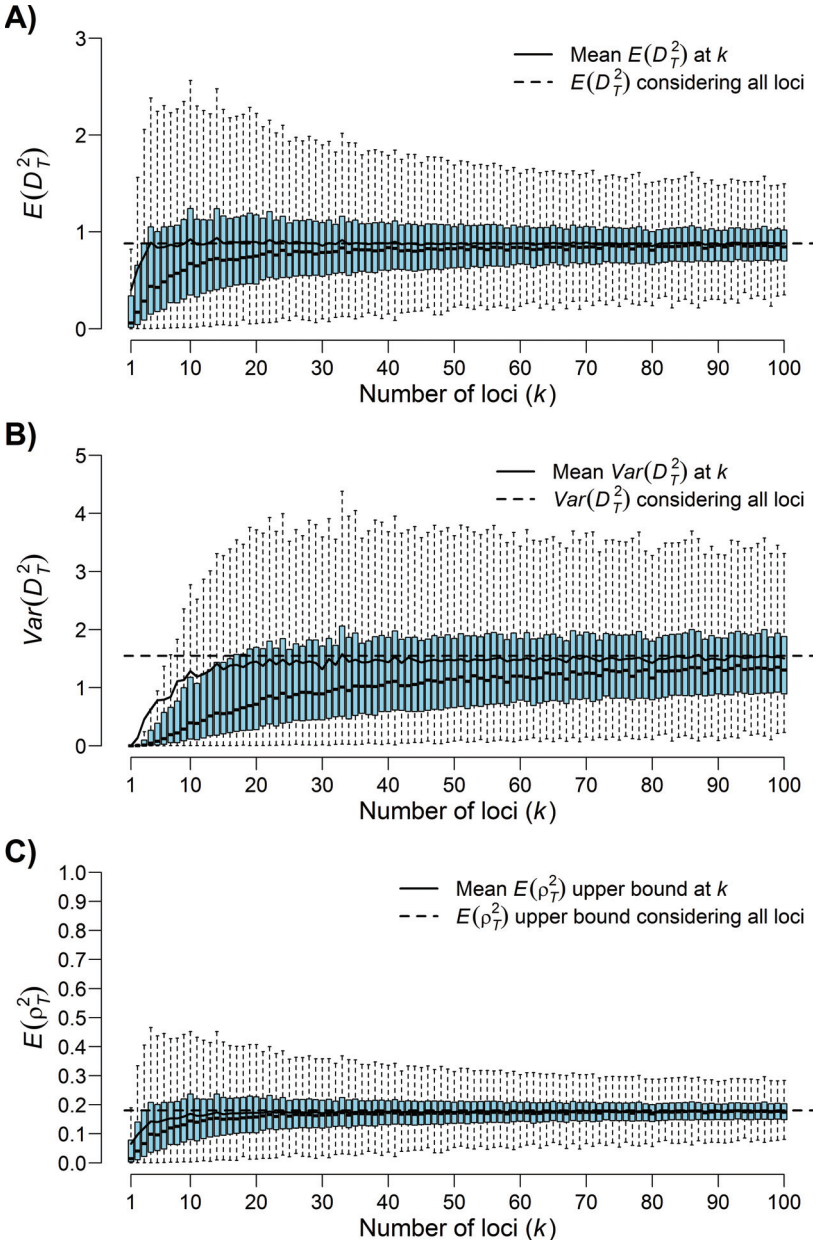

FIGURE 5. The behavior of expected measures of trait differentiation as the number of randomly selected loci influencing the trait increases. We simulated allele frequencies at 10^6 neutral loci for a pair of populations that have undergone independent drift since divergence from an ancestral population, with $F_{ST} \approx 0.1$ (see Section 3.2). For each $k \in \{1, \ldots, 100\}$, we selected 1,000 size-k random subsets of the 10^6 pairs of simulated allele frequencies and computed three quantities for each subset, assuming diploidy ($\ell = 2$). (A) The expected squared standardized trait difference between groups, $E(D_T^2)$ (Eq. 36). (B) The variance of the squared standardized trait difference between groups, $\mathrm{Var}(D_T^2)$ (Eq. 39). (C) The upper bound on (and approximate value of) the expected proportion of variance in a neutral trait attributable to allele-frequency differences between groups, $E(\rho_T^2)$ (Eqs. 43, 44). For each quantity, box plots of the 1,000 values for each k are shown. Boxes represent the middle 50% of data at each k, and whiskers extend 1.5 times the interquartile range beyond the edge of the box or to the most extreme observation, whichever is shorter. Outliers beyond 1.5 times the interquartile range from the edge of the box are not shown. For all three quantities, as k increases, the mean value at k loci (solid line) converges to the value obtained using all loci (dashed line) because larger random sets of loci more precisely reflect the overall degree of between-group differentiation than do smaller sets of loci.

Table 2. Correspondence between the Main Results in This Article and in Edge and Rosenberg (2015)

Result	Equation Number	
	This Article	Edge and Rosenberg (2015)
$E(D_T) = 0$ due to symmetry around 0 of the distribution of D_T.	35	36
$\mathrm{Var}(D_T) = E(D_T^2)$ does not increase without bound with the number of loci and is equal to D_L^2, where D_L is an analogue of D_T for the allelic count at a single locus.	36, 37	37, 38
$F_{ST} \approx Q_{ST}$.	45, 47, 48	42, 43
As the number of loci k increases without bound, the genetic misclassification rate approaches 0.	55	5
The expectation of the approximate trait-based misclassification rate is closely related to the genetic misclassification rate obtained using one locus.	57	47

The main results in this article reduce to the main results in Edge and Rosenberg (2015) under the following assumptions: (a) The allele frequencies are the same at each locus, meaning that $p_i = \bar{p} = p$ for all i, $q_i = \bar{q} = q$ for all i, $\delta_i = \bar{\delta} = q - p$ for all i, and $s_p^2 = s_q^2 = s_\delta^2 = 0$. (b) The allele frequencies are symmetric, meaning that $\bar{q} = 1 - \bar{p}$. In conjunction with assumption (a), (b) implies that $F_{ST} = \overline{\delta^2} = 1 - 4pq$. (c) The organisms are haploid, or in the present article's notation, $\ell = 1$. Assumption (c) implies that $Q_{ST} = \rho_T^2$ (Eqs. 26–28).

The expression on the right side of Eq. 43 is an approximation of $E(\rho_T^2)$, but it is also a strict upper bound on $E(\rho_T^2)$. To see that it is an upper bound, note that ρ_T^2 is concave in $Y = (\ell/2)(\Sigma_{i=1}^k \delta_i U_i)^2$ (Eq. 42). Thus, by Jensen's inequality, which states that if g is a concave function of a random variable X, then $E[g(X)] \le g[E(X)]$, we have

$$E(\rho_T^2) \le \frac{\ell\overline{\delta^2}}{2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \ell\overline{\delta^2}}. \quad (44)$$

As we observed for $E(D_T^2)$, if $\bar{p}$ and $\bar{q}$ are fixed, then $E(\rho_T^2)$ can take larger values if $k > 1$ than if $k = 1$ because increasing s_p^2 or s_q^2 increases the upper bound on $E(\rho_T^2)$. Nonetheless, $E(\rho_T^2)$ does not grow without bound as k increases (Figure 5C). Comparing Eqs. 43 and 44 with Eq. 16,

$$\begin{aligned} E(\rho_T^2) &\approx F_{ST(\ell)}^k, \\ E(\rho_T^2) &\le F_{ST(\ell)}^k. \end{aligned} \quad (45)$$

The expected value of ρ_T^2 is thus approximately equal to, and no greater than, the ratio of the mean across loci of the between-group variance of $\Sigma_{j=1}^{\ell} L_{ij}$ to the mean across loci of the total variance of $\Sigma_{j=1}^{\ell} L_{ij}$, where $\Sigma_{j=1}^{\ell} L_{ij}$ is a random variable representing the number of 1 alleles carried by an ℓ-ploid individual at locus i.

Because Q_{ST} is equal to the expression for ρ_T^2 in Eq. 41 with ℓ set to 1 (Eq. 28), Eqs. 43 and 44 imply that

$$\begin{aligned} Q_{ST} &\approx \frac{\overline{\delta^2}}{2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \overline{\delta^2}}, \\ Q_{ST} &\le \frac{\overline{\delta^2}}{2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2] + \overline{\delta^2}}. \end{aligned} \quad (46)$$

By Eq. 14, the expression on the right side of Eq. 46 is equal to F_{ST}^k, so

$$Q_{ST} \approx F_{ST}^k, \quad (47)$$

$$Q_{ST} \le F_{ST}^k. \quad (48)$$

Equation 47 is consistent with previous work on the relationship between F_{ST}^k and Q_{ST} under different models (e.g., Lande 1992; Whitlock 1999), and it provides one justification for the claim that the degree of between-group difference on a neutral trait is approximately equal to the degree of between-group genetic differentiation at a typical locus.

6. Adding Assumptions Inspired by Equal Drift since a Recent Divergence

Having addressed the relationship between neutral genetic and neutral phenotypic differentiation between populations in the context of standardized differences and variance partitioning (Sections 4 and 5), we now consider the accuracy with which individuals can be classified into populations using neutral genetic and phenotypic information. We require two assumptions that will allow us to consider approximate misclassification rates that would arise if we attempted to identify an individual's population of origin by examining k loci directly or by examining a trait determined additively by those k loci. The case in which these assumptions are met is a restriction of the general case we have been examining. The special case in this section can be viewed as the expectation under a model in which the allele frequencies in populations A and B have experienced equal amounts of drift since a recent divergence (see Appendix 3).

The first assumption is symmetry of average frequency of the 1 allele across loci in populations A and B,

$$\bar{q} = 1 - \bar{p}, \quad (49)$$

where both $\bar{p}$ and $\bar{q}$ are assumed to be nonzero. By Eq. 9, Eq. 49 implies that

$$\overline{\delta}^2 = 1 - 4\overline{p}\,\overline{q}. \tag{50}$$

The quantity in Eq. 50 is the F_{ST}^1 value of a locus that has the average frequency of the 1 allele in each population, lending $\overline{\delta}^2$ a new interpretation (Eq. 12). The second assumption is that the variance of the allele frequencies has the same value in each population:

$$s_p^2 = s_q^2. \tag{51}$$

6.1 Multilocus Classification

We now consider the problem of identifying the population of an individual of unknown origin. In this subsection, we examine the misclassification rates that arise from an examination of the number of 1 alleles carried by an individual across k loci, S.

Recall that the genotypic statistic S (Eq. 6) is the number of alleles carried by an individual that are more common in population B than in population A. If $X_i = 1$ for all i, then $T = S$. Within each population, the Ls are independent Bernoulli random variables with possibly different probabilities. Thus, within each population, S has a Poisson binomial distribution. By the properties of the Poisson binomial distribution (Eq. 7),

$$\begin{aligned}
E(S \mid M = A) &= \sum_{i=1}^{k}\sum_{j=1}^{\ell} p_i = \ell k \overline{p}, \\
\mathrm{Var}(S \mid M = A) &= \sum_{i=1}^{k}\sum_{j=1}^{\ell} p_i(1-p_i) \\
&= \ell k[\overline{p}(1-\overline{p}) - s_p^2]; \\
E(S \mid M = B) &= \sum_{i=1}^{k}\sum_{j=1}^{\ell} q_i = \ell k \overline{q}, \\
\mathrm{Var}(S \mid M = B) &= \sum_{i=1}^{k}\sum_{j=1}^{\ell} q_i(1-q_i) \\
&= \ell k[\overline{q}(1-\overline{q}) - s_q^2],
\end{aligned} \tag{52}$$

where $\overline{p}$, $\overline{q}$, s_p^2, and s_q^2 are as defined in Eqs. 2 and 3. The variances of S within each population are the same as the variances of T within each population (Eqs. 22, 23).

We consider the normal approximation of the misclassification rate obtained if the genotypic statistic S is used for classification. If the assumptions in Eqs. 49 and 51 hold, then the within-population variances of S in the two populations are equal:

$$\begin{aligned}
\ell k[\overline{p}(1-\overline{p}) - s_p^2] &= \ell k[\overline{q}(1-\overline{q}) - s_q^2] \\
&= \ell k[\overline{p}\,\overline{q} - s_p^2].
\end{aligned} \tag{53}$$

Further, when k is large, as a sum of independent Bernoulli variables the sum of whose variances increases without bound, the distribution of S is approximately normal:

$$\begin{aligned}
(S \mid M = A) &\sim \mathrm{Normal}(\ell k\overline{p}, \ell k[\overline{p}\,\overline{q} - s_p^2]), \\
(S \mid M = B) &\sim \mathrm{Normal}(\ell k\overline{q}, \ell k[\overline{p}\,\overline{q} - s_p^2]).
\end{aligned} \tag{54}$$

(Deheuvels et al. 1989, Theorem 1.1).

Denoting the normal density that approximates the distribution of S in population A by $f_A(s)$ and the corresponding normal density for population B by $f_B(s)$, then when we observe that $S = s$, we classify the individual into population A if $f_A(s) > f_B(s)$ and into population B if $f_A(s) < f_B(s)$. In this case, $f_A(s) > f_B(s)$ if $s < \ell k(\overline{p} + \overline{q})/2$ and $f_A(s) < f_B(s)$ if $s > \ell k(\overline{p} + \overline{q})/2$. We ignore the case of $s = \ell k(\overline{p} + \overline{q})/2$, which is negligible for large k. By the assumption in Eq. 49, $\overline{p} + \overline{q} = 1$, and we therefore classify an individual into population A if $S < \ell k/2$ and into population B if $S > \ell k/2$.

We represent the event that an individual is misclassified on the basis of S with the random indicator variable W_S, which equals 1 if and only if an individual is misclassified on the basis of S and equals 0 otherwise. In population A, the approximate probability of misclassification is

$$\begin{aligned}
P(W_S = 1 \mid M = A) &\approx P(S > \ell k/2 \mid M = A) \\
&\approx 1 - \Phi\left(\sqrt{\ell k}\,\frac{\overline{\delta}}{2\sqrt{\overline{p}\,\overline{q} - s_p^2}}\right),
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution. A similar calculation for population B gives the same misclassification rate. Thus, in both populations, the approximate misclassification probability obtained using S is

$$\begin{aligned}
P(W_S = 1) &\approx 1 - \Phi\left(\sqrt{\ell k}\,\frac{\overline{\delta}}{2\sqrt{\overline{p}\,\overline{q} - s_p^2}}\right) \\
&= 1 - \Phi\left(\sqrt{\ell k}\,\frac{\overline{q} - \overline{p}}{2\sqrt{\overline{p}\,\overline{q} - s_p^2}}\right).
\end{aligned} \tag{55}$$

As k increases, with $\overline{p}$, $\overline{q}$, and s_p^2 held constant, the argument to the cumulative distribution function in Eq. 55 approaches infinity, and the value of the cumulative distribution function approaches 1. Thus, as the number of loci increases, the

misclassification probability obtained when using the genotypic statistic S, $P(W_S = 1)$, approaches 0.

6.2 Trait-Based Classification

Next we consider the approximate misclassification rate obtained on the basis of an individual's trait value. We represent the event that an individual is misclassified on the basis of T with the random indicator variable W_T, which equals 1 if an individual is misclassified on the basis of its trait value and equals 0 otherwise. Using an argument similar to the one used to justify Eq. 55 (detailed in Appendix 4), the approximate trait-based misclassification rate, conditional on $X_1, \ldots, X_k$, is

$$P(W_T = 1 | \{X_1, \ldots, X_k\} = \{x_1, \ldots, x_k\}) \approx 1 - \Phi\left[\frac{\sqrt{\ell}\,|\sum_{i=1}^{k} \delta_i u_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right], \quad (56)$$

where Φ is the cumulative distribution function of the standard normal distribution.

To understand how the misclassification rate is expected to behave across different labelings of the loci, we consider the expectation of the normal approximation of the misclassification rate obtained using the trait value T, $P(W_T = 1)$. Removing the condition on the allelic labeling in Eq. 56 and rearranging gives

$$1 - P(W_T = 1) \approx \Phi\left[\frac{\sqrt{\ell}\,|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right],$$

where U_i is as defined in Eq. 18. Taking the expectation of both sides gives

$$1 - E[P(W_T = 1)] \approx E\left(\Phi\left[\frac{\sqrt{\ell}\,|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right]\right).$$

Noticing that Φ is concave for positive values of its argument gives, by Jensen's inequality,

$$E\left(\Phi\left[\frac{\sqrt{\ell}\,|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right]\right) \le \Phi\left[\frac{\sqrt{\ell} E|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right].$$

Because $|(\Sigma_{i=1}^{k} \delta_i U_i)| = [(\Sigma_{i=1}^{k} \delta_i U_i)^2]^{1/2}$, and because the square root is a concave function, Jensen's inequality gives

$$E\left(\left|\sum_{i=1}^{k} \delta_i U_i\right|\right) = E\left(\sqrt{\left(\sum_{i=1}^{k} \delta_i U_i\right)^2}\right) \le \sqrt{E\left[\left(\sum_{i=1}^{k} \delta_i U_i\right)^2\right]}.$$

Then

$$\Phi\left[\frac{\sqrt{\ell} E|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right] \le \Phi\left[\frac{\sqrt{\ell E\left[\left(\sum_{i=1}^{k} \delta_i U_i\right)^2\right]}}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right].$$

Because $E[(\Sigma_{i=1}^{k} \delta_i U_i)^2] = k\overline{\delta^2}$ (Eq. 32),

$$\Phi\left[\frac{\sqrt{\ell E\left[\left(\sum_{i=1}^{k} \delta_i U_i\right)^2\right]}}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right] = \Phi\left[\frac{\sqrt{\ell\overline{\delta^2}}}{2\sqrt{\bar{p}\bar{q} - s_p^2}}\right] = \Phi\left[\frac{\sqrt{\ell\left(\overline{\delta}^2 + s_\delta^2\right)}}{2\sqrt{\bar{p}\bar{q} - s_p^2}}\right].$$

We then have an approximate lower bound on the expected probability of misclassification on the basis of T:

$$E[P(W_T = 1)] \approx 1 - E\left(\Phi\left[\frac{\sqrt{\ell}\,|\sum_{i=1}^{k} \delta_i U_i|}{2\sqrt{k(\bar{p}\bar{q} - s_p^2)}}\right]\right) \ge 1 - \Phi\left[\frac{\sqrt{\ell\left(\overline{\delta}^2 + s_\delta^2\right)}}{2\sqrt{\bar{p}\bar{q} - s_p^2}}\right]. \quad (57)$$

If $s_\delta^2 = 0$, then Eq. 57 produces the same result as Eq. 55 with $k = 1$. The expectation of the approximate misclassification probability on the basis of the trait is therefore, if $s_\delta^2 = 0$, greater than or equal to the approximate misclassification probability obtained using a single locus.

For fixed $\bar{p}$, $\bar{q}$, and s_p^2, as k increases, the argument to the cumulative distribution function in Eq. 57 does not approach infinity. The value of the cumulative distribution function approaches an asymptotic value obtained as the increasingly many loci converge on large-k values of s_δ^2 and s_p^2. Thus, as the number of loci considered increases, the misclassification probability obtained when using the trait value T, $P(W_T = 1)$ does not approach 0.

7. Discussion

We have extended a model of multilocus allele frequency differences and polygenic trait differences between groups to accommodate more general allele-frequency distributions and arbitrary ploidy. Our results recapitulate our original conclusion (Edge and Rosenberg 2015): a single neutral trait provides approximately the same amount of information about population membership as does

a single neutral genetic locus. This general claim is reflected in three specific ways of asking about the relative magnitude of population-membership information in genotypes and in phenotypes: using the standardized trait difference between groups (D_T, Eq. 37), using the between-group variance in the trait (ρ_T^2 and D_{ST}, Eqs. 45, 47), and using the misclassification rate obtained when attempting to classify individuals into groups by their trait values ($P(W_T = 1)$, Eq. 57). We also provide two main updates to our previous work. First, under our model, within-population variation in allele frequency across loci tends to increase the degree of expected difference between groups on a polygenic trait (Eq. 36). Second, the degree of information about population membership is greater for a diploid (or polyploid) than for a haploid locus (Figure 3), and it is also correspondingly greater for a trait in a diploid (or polyploid) organism than in a haploid (Eq. 37).

What accounts for the difference between the ancestry information of multiple genetic loci and that of a trait governed by those same loci? When examining multiple genetic loci, information about population membership can be cumulated from multiple sites—for example, by counting in an individual the alleles that are more common in population A than in population B. In contrast, although an individual's trait value implicitly encodes information about its genotype at many loci, random genetic drift prevents the trait from accumulating information about population membership. In our model, for every locus at which the allele at higher frequency in population A is associated with larger trait values, there is likely to be another locus at which the "A-like" allele is associated with smaller trait values. The cumulative effect of this locus-by-locus shuffling of the choice of population associated with the higher trait value is that a single neutral trait is, in expectation, approximately as informative about population membership as a single neutral locus.

7.1 Models of Genotypic and Phenotypic Differentiation

Our results accord with those of previous efforts to address similar questions with different models (Felsenstein 1973, 1986; Rogers and Harpending 1983; Lande 1992; Spitze 1993; Lynch and Spitze 1994; Whitlock 1999; Berg and Coop 2014), which have repeatedly found that group or population differences in neutral, completely heritable traits mirror neutral genetic differentiation. Previous examinations of trait differentiation have often proceeded by relating assumptions about quantitative traits to models of evolutionary change in allele frequencies, such as a Wright–Fisher model (Felsenstein 1973), an island migration model (Lande 1992), or a coalescent framework (Whitlock 1999). The use of such evolutionary models can suggest connections with other areas of evolutionary genetics and can also provide insights with considerable generality; for example, Whitlock's (1999) results hold for coalescent models with arbitrary population structure.

In contrast with some previous models of phenotypic diversity, our models here and in our previous work (Edge and Rosenberg 2015) are more similar to the genetic classification models of Risch et al. (2002), Edwards (2003), and Tal (2012) in that we directly consider allele frequencies, using simple probabilistic arguments and minimal evolutionary assumptions. This approach complements earlier evolutionary work on the relationship of genetic and phenotypic differentiation in at least two ways. First, our model allows for computations with quantities that are of interest in epidemiological and biomedical studies but that do not necessarily arise naturally under evolutionary models, such as Cohen's d (equal to our D_T for a completely heritable trait) and the effect size r^2 (equal to our ρ_T^2 for a completely heritable trait). Second, the fact that our model makes only minimal evolutionary assumptions shows that similar results obtained under evolutionary models are robust in that they are also produced via a substantially different modeling approach.

7.2 Interpreting Group Differences in Phenotype

How can this work aid in the interpretation of phenotypic differences between human groups? Consider health outcomes, an important set of phenotypes for which genetic and phenotypic differentiation across populations have been of interest.

Among people in the United States, for example, well-established differences exist between socially defined racial groups in the incidence of many health conditions, including heart disease (e.g., Lloyd-Jones et al. 2010), various cancers (e.g.,

Ward et al. 2004; Siegel et al. 2013), and diabetes (e.g., LaVeist et al. 2009), with African Americans having worse health outcomes than European Americans across many domains (reviewed in Dressler et al. 2005; Adler and Rehkopf 2008). Such phenotypic differences between pairs of groups arise from a combination of interacting factors, which can be viewed as modifications of a baseline prediction made on the basis of neutral genetic differences. Evolutionary genetics can then contribute to understanding group differences in health outcomes by providing models that predict the degree to which phenotypic differences between human groups are likely to be based in neutral genetic differences. Such models do not necessarily explain the source of any particular phenotypic difference, but they do provide an idea of what patterns of difference are expected.

Using population-genetic models to build intuition about phenotypic differences between human groups does not require that group classifications are reducible to, caused by, or primarily based in genetic differences between populations. Rather, population-genetic models of neutral genetic variation are applicable to sets of groups that are correlated with some degree of genetic population structure and that thus distinguish groups that differ in allele frequency at sites across the genome. Although the relationship between population-genetic groupings and socially defined groupings is complex (e.g., Kittles and Weiss 2003; Bamshad et al. 2004; Kitcher 2007; Hunley et al. 2016), we can gain some intuition about the possible sources of a group difference in phenotype for socially defined groups by comparing its size to the associated degree of between-group differentiation at a typical genetic locus.

Two possible causes of group phenotypic differences that are larger or smaller than the degree of between-group differentiation at a typical selectively neutral locus are natural selection and environmental difference. If a trait has been under selection in two populations in a manner that leads to divergence—for example, if the trait is advantageous in one population and disadvantageous or neutral in the other—then the between-group difference on the trait will typically exceed the average between-group genetic difference. An example relevant to health differences between socially defined racial groups is skin pigmentation: lighter skin has been positively selected among populations at higher latitudes (Jablonski and Chaplin 2000; Relethford 2002; Berg and Coop 2014), but it is also a risk factor for skin cancer (Lin and Fisher 2007). In turn, European Americans, most of whose recent ancestors generally lived at high latitude, have substantially higher rates of skin cancer than do African Americans (Halder and Bridgeman-Shah 1995), a larger fraction of whose recent ancestors generally lived at lower latitudes. In contrast, many other health-related traits are likely associated with similar reproductive fitness wherever they occur. Such conditions would be expected to experience *convergent* selection, which would lead to *smaller* between-group trait differences than might be predicted from neutral genetic diversity. Thus, one explanation for larger-than-expected phenotypic differences among groups is divergent selection, and one explanation for smaller-than-expected phenotypic differences among groups is convergent selection. This reasoning is the basis for the use of comparisons between Q_{ST} and F_{ST} to test hypotheses about phenotypic evolution, a productive approach for model organisms that can be raised in a "common garden" situation (Whitlock 2008; Leinonen et al. 2013).

In humans, assessing whether selection has magnified health differences beyond the neutral prediction is difficult. In some specific cases, divergent selection is regarded as an important driver of group difference in disease burden—including, for example, sickle-cell disease, which occurs at higher rates in malarial regions of Africa and the Mediterranean (e.g., Piel et al. 2010). In many other cases of phenotypic difference, hypotheses of divergent selection—sometimes paired with gene–environment interaction—have also been proposed (Knowler et al. 1983; Meindl 1987; Zlotogora et al. 1988; Wilson and Grim 1991; Bindon and Baker 1997). Many such hypotheses have been criticized individually (Curtin 1992; Risch et al. 2003), and concern has been raised that selective explanations for differences in health outcomes often assume a degree of endurance and importance unwarranted by the evidence (Kaufman and Hall 2003). As new cases connecting selection pressures to molecular evidence of adaptation emerge (e.g., Fumagalli et al. 2015), the empirical basis for assessing the role of divergent selection in explaining differences in health outcomes will expand.

Much recent discussion has focused on an alternative approach to examining worldwide consequences of selection, considering demographic factors that influence the strength of selection against deleterious mutations in different populations (Lohmueller 2014; Henn et al. 2015). Recent studies have not found pronounced difference between groups of primarily European and African descent in the overall frequency of putatively deleterious alleles (Fu et al. 2014; Simons et al. 2014; Do et al. 2015), but populations differ in how these alleles are distributed among individuals, with Europeans carrying more genotypes homozygous for putatively deleterious alleles compared with Africans (Lohmueller et al. 2008; Fu et al. 2014; Do et al. 2015). Though these studies do not identify differences between populations in the influence of experienced selective pressures, their results suggest that a systematic difference across populations in the outcomes of selection on disease phenotypes, if it exists at all, would likely tilt toward a greater disease burden in non-Africans.

Finally, environmental differences between groups are important sources of between-group trait differences. Environmental differences and associated differences in the effect of gene–environment interactions can act in concert with or in opposition to any genetic differences that influence a trait, leading to between-group trait differences that are larger or smaller than would be expected on the basis of neutral genetic differences alone (Pujol et al. 2008). In the United States, the environments of people of different socially defined races differ in myriad factors that could contribute to differences in health outcomes (Williams and Jackson 2005), including socioeconomic status (Adler and Newman 2002), education (Non et al. 2012), residential segregation (Williams and Collins 2001), discrimination (Williams and Mohammed 2009), targeting by the criminal justice system (Iguchi et al. 2005), access to medical care (Mayberry et al. 2000), and doctor–patient communication (Ashton et al. 2003).

We can examine an environmental hypothesis about group differences in health outcomes in relation to the predicted pattern of differences across outcomes for our neutral model of two groups. Under the neutral model, each group has an equal chance of having a larger trait value for each trait. Thus, each population would have a larger mean value for roughly half the traits on which the groups differed, with each trait independent if there are no genetic correlations. An environmental explanation of differences in health outcomes might hold that social differences, such as differences in access to health care, are likely to cause a pattern

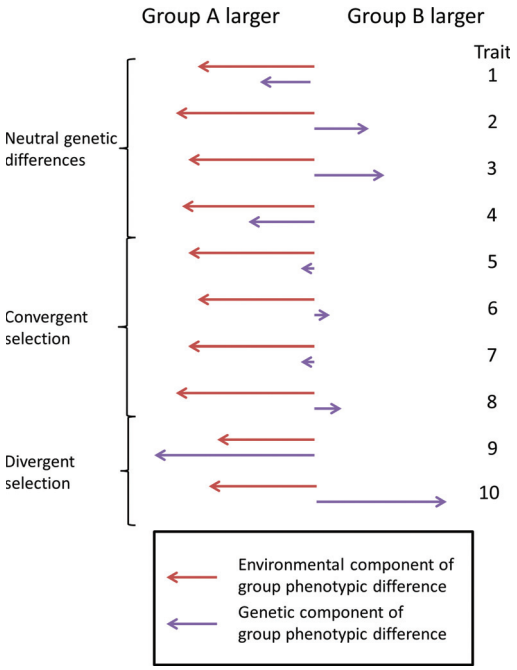

FIGURE 6. A schematic for thinking about health differences between socially defined racial groups in the United States. Groups such as African Americans face environmental differences likely to lead to worse health outcomes across a range of different diseases (red arrows). For health phenotypes that have been selectively neutral, genetic differences between groups will be random in direction and comparable in size to genetic differences at a single locus—modest for humans, but usually not zero, depending on the groups being considered (purple arrows in traits 1–4). For health phenotypes under convergent selection, such as those that lead to reduced reproductive success in most or all human environments, genetic differences will be random in direction and smaller than for neutral phenotypes (purple arrows in traits 5–8). Some health outcomes, such as skin cancer and sickle-cell disease, differ between groups in part because of divergent selection. (These differences do not necessarily coincide neatly with socially relevant racial divisions.) For such phenotypes, the genetic component of a group difference in phenotype can be large (purple arrows in traits 9 and 10). Not considered in this simple diagram are gene–environment interactions, which often are especially important in cases of divergent selection. For example, in the case of skin cancer, the degree to which genetic variants that lead to darker skin are protective depends on sun exposure.

in which differences between groups run in the same direction across many diseases and causes of mortality. Under this reasoning, the fact that African Americans suffer more than do European Americans from a wide variety of diseases is more consistent with environmental sources of phenotypic difference than with a neutral-genetic explanation. As an example, Wong et al. (2002) tabulated racial differences in causes of death in 36 categories over a 9-year period. After adjusting for age, sex, and years of education, they estimated that black Americans lost more life years than white Americans on average in 28 of those categories. Informally, assuming under our neutral model that no genetic correlation exists between phenotypic outcomes and that for each outcome the larger value has equal probability of occurring in either group, the binomial probability that one population would have a larger trait value than the other on at least 28 of 36 independent phenotypes is only 0.001. A single systematic environmental effect that simultaneously inflates many nongenetic risk factors in African Americans, on the other hand, can provide a simple explanation for such skewed outcomes.

7.3 Conclusions

Our model provides a general framework for describing the relevance of single-locus genetic diversity partitioning for predictions about the sources of phenotypic differences between groups. For neutral, heritable traits, group differences in phenotype will be random in direction and will reflect the degree of genetic difference at a single locus—modest in size for humans, but likely not zero—regardless of how many loci influence the trait. Such neutral differences are a baseline on top of which selection and environmental influences act (Figure 6). In the case of health-related differences between socially defined races in the United States, the occurrence of genetic differentiation as measured by F_{ST} suggests that neutral genetic differences are likely to exist for many heritable health outcomes that are not under selection. Such genetically based differences may run in the opposite direction of the apparent phenotypic difference between groups, and typical values of human F_{ST} suggest that they will likely be modest in size on average. Nonetheless, their existence supports the view that genetic research designs that capitalize on group differences (e.g., Winkler et al. 2010; Zaitlen et al. 2014) can be informative about genetic architecture or the genetic variants that influence phenotypes (Rosenberg et al. 2010; Teo et al. 2010). At the same time, any patterns of difference in which one group suffers more than others from the majority of many genetically independent diseases are unlikely to be explained by neutral genetic variation. For humans, our model supports the view that coordinated group differences across a preponderance of independent health-related traits suggest an important role for systematic differences in environmental risk factors.

ACKNOWLEDGMENTS

We thank the organizers for including us in this special issue, Jeff Long and Alan Rogers for stimulating discussion, and National Institutes of Health grant R01 HG005855, National Science Foundation grant DBI-1458059, and the Stanford Center for Computational, Evolutionary, and Human Genomics for support.

Received 26 February 2016; revision accepted for publication 13 April 2016.

LITERATURE CITED

Adler, N. E., and K. Newman. 2002. Socioeconomic disparities in health: Pathways and policies. *Health Aff.* (*Millwood*) 21:60–76.

Adler, N. E., and D. H. Rehkopf. 2008. US disparities in health: Descriptions, causes, and mechanisms. *Annu. Rev. Public Health* 29:235–252.

Ashton, C. M., P. Haidet, D. A. Paterniti et al. 2003. Racial and ethnic disparities in the use of health services: Bias, preferences, or poor communication? *J. Gen. Intern. Med.* 18:146–152.

Bamshad, M., S. Wooding, B. A. Salisbury et al. 2004. Deconstructing the relationship between genetics and race. *Nat. Rev. Genet*. 5:598–609.

Bamshad, M. J., S. Wooding, W. S. Watkins et al. 2003. Human population genetic structure and inference of group membership. *Am. J. Hum. Genet*. 72:578–589.

Barbujani, G., A. Magagni, E. Minch et al. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA* 94:4,516–4,519.

Berg, J. J., and G. Coop. 2014. A population genetic signal of polygenic adaptation. *PLoS Genet.* 10:e1004412.

Bindon, J. R., and P. T. Baker. 1997. Bergmann's rule and the thrifty genotype. *Am. J. Phys. Anthropol*. 104:201–210.

Bowcock, A. M., A. Ruiz-Linares, J. Tomfohrde et al. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455–457.

Brown, R. A., and G. J. Armelagos. 2001. Apportionment of racial diversity: A review. *Evol. Anthropol*. 10:34–40.

Chakraborty, R., and M. Nei. 1982. Genetic differentiation of quantitative characters between populations or species: I. Mutation and random genetic drift. *Genet. Res*. 39:303–314.

Charlesworth, B., and D. Charlesworth. 2010. *Elements of Evolutionary Genetics*. Greenwood Village, CO: Roberts.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum.

Curtin, P. D. 1992. The slavery hypothesis for hypertension among African Americans: The historical evidence. *Am. J. Public Health* 82:1,681–1,686.

Deheuvels, P., M. L. Puri, and S. S. Ralescu. 1989. Asymptotic expansions for sums of nonidentically distributed Bernoulli random variables. *J. Multivar. Anal*. 28:282–303.

Do, R., D. Balick, H. Li et al. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nat. Genet*. 47:126–131.

Dressler, W. W., K. S. Oths, and C. C. Gravlee. 2005. Race and ethnicity in public health research: Models to explain health disparities. *Annu. Rev. Anthropol*. 34:231–252.

Edge, M. D., and N. A. Rosenberg. 2015. Implications of the apportionment of human genetic diversity for the apportionment of human phenotypic diversity. *Stud. Hist. Philos. Biol. Biomed. Sci.* 52:32–45.

Edwards, A. W. F. 1960. The meaning of binomial distribution. *Nature* 186:1,074.

Edwards, A. W. F. 2003. Human genetic diversity: Lewontin's fallacy. *Bioessays* 25:798–801.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164:1,567–1,587.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet*. 25:471–492.

Felsenstein, J. 1986. Population differences in quantitative characters and gene frequencies: A comment on papers by Lewontin and Rogers. *Am. Nat*. 127:731–732.

Fox, J. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage.

Fu, W., R. M. Gittelman, M. J. Bamshad et al. 2014. Characteristics of neutral and deleterious protein-coding variation among individuals and populations. *Am. J. Hum. Genet*. 95:421–436.

Fumagalli, M., I. Moltke, N. Grarup et al. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349:1,343–1,347.

Halder, R. M., and S. Bridgeman-Shah. 1995. Skin cancer in African Americans. *Cancer* 75:667–673.

Henn, B. M., L. R. Botigué, C. D. Bustamante et al. 2015. Estimating the mutation load in human genomes. *Nat. Rev. Genet*. 16:333–343.

Holsinger, K. E., and B. S. Weir. 2009. Genetics in geographically structured populations: Defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet*. 10:639–650.

Hunley, K. L., G. S. Cabana, and J. C. Long. 2016. The apportionment of human diversity revisited. *Am. J. Phys. Anthropol.*, 160:561–569.

Iguchi, M. Y., J. Bell, R. N. Ramchand et al. 2005. How criminal system racial disparities may translate into health disparities. *J. Health Care Poor Underserved* 16:48–56.

Jablonski, N. G., and G. Chaplin. 2000. The evolution of human skin coloration. *J. Hum. Evol*. 39:57–106.

Kaufman, J. S., and S. A. Hall. 2003. The slavery hypertension hypothesis: Dissemination and appeal of a modern race theory. *Epidemiology* 14:111–118.

Kitcher, P. 2007. Does "race" have a future? *Philos. Public Aff.* 35:293–317.

Kittles, R. A., and K. M. Weiss. 2003. Race, ancestry, and genes: Implications for defining disease risk. *Annu. Rev. Genomics Hum. Genet*. 4:33–67.

Knowler, W. C., D. J. Pettitt, P. H. Bennett et al. 1983. Diabetes mellitus in the Pima Indians: Genetic and evolutionary considerations. *Am. J. Phys. Anthropol.* 62:107–114.

Lande, R. 1992. Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution* 46:381–389.

LaVeist, T. A., R. J. Thorpe Jr., J. E. Galarraga et al. 2009. Environmental and socio-economic factors as contributors to racial disparities in diabetes prevalence. *J. Gen. Intern. Med.* 24:1,144–1,148.

Leinonen, T., R. J. S. McCairns, R. B. O'Hara et al. 2013. Q_{ST}-F_{ST} comparisons: Evolutionary and ecological insights from genomic heterogeneity. *Nat. Rev. Genet*. 14:179–190.

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

Li, J. Z., D. M. Absher, H. Tang et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1,100–1,104.

Lin, J. Y., and D. E. Fisher. 2007. Melanocyte biology and skin

pigmentation. *Nature* 445:843–850.

Lloyd-Jones, D., R. J. Adams, T. M. Brown et al. 2010. Heart disease and stroke statistics—2010 update: A report from the American Heart Association. *Circulation* 121:e46–e215.

Lohmueller, K. E. 2014. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev*. 29:139–146.

Lohmueller, K. E., A. R. Indap, S. Schmidt et al. 2008. Proportionally more deleterious genetic variation in European than in African populations. *Nature* 451:994–997.

Lynch, M., and Spitze, K. 1994. Evolutionary genetics of *Daphnia*. In *Ecological Genetics*, L. Real, ed. Princeton, NJ: Princeton University Press, 109–128.

Mayberry, R. M., F. Mili, and E. Ofili. 2000. Racial and ethnic differences in access to medical care. *Med. Care Res. Rev*. 57(suppl. 1):108–145.

Meindl, R. S. 1987. Hypothesis: A selective advantage for cystic fibrosis heterozygotes. *Am. J. Phys. Anthropol*. 74:39–45.

Mountain, J. L., and L. L. Cavalli-Sforza. 1997. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet*. 61:705–718.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70:3,321–3,323.

Nicholson, G., A. V. Smith, F. Jónsson et al. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Stat. Soc. Series B Stat. Methodol.* 64:695–715.

Non, A. L., C. C. Gravlee, and C. J. Mulligan. 2012. Education, genetic ancestry, and blood pressure in African Americans and whites. *Am. J. Public Health* 102:1,559–1,565.

Piel, F. B., A. P. Patil, R. E. Howes et al. 2010. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun*. 1:104.

Pujol, B., A. J. Wilson, R. I. C. Ross et al. 2008. Are Q_{ST}-F_{ST} comparisons for natural populations meaningful? *Mol. Ecol.* 17:4,782–4,785.

Relethford, J. H. 2002. Apportionment of global human genetic diversity based on craniometrics and skin color. *Am. J. Phys. Anthropol*. 118:393–398.

Relethford, J. H. 2010. Population-specific deviations of global human craniometric variation from a neutral model. *Am. J. Phys. Anthropol*. 142:105–111.

Risch, N., E. Burchard, E. Ziv et al. 2002. Categorization of humans in biomedical research: Genes, race and disease. *Genome Biol.* 3:comment2007.

Risch, N., H. Tang, H. Katzenstein et al. 2003. Geographic distribution of disease mutations in the Ashkenazi Jewish population supports genetic drift over selection. *Am. J. Hum. Genet*. 72:812–822.

Rogers, A. R., and H. C. Harpending. 1983. Population structure and quantitative characters. *Genetics* 105: 985–1,002.

Roseman, C. C. 2004. Detecting interregionally diversifying natural selection on modern human cranial form by using matched molecular and morphometric data. *Proc. Natl. Acad. Sci. USA* 101:12,824–12,829.

Roseman, C. C., and T. D. Weaver. 2004. Multivariate apportionment of global human craniometric diversity. *Am. J. Phys. Anthropol*. 125:257–263.

Rosenberg, N. A., L. Huang, E. M. Jewett et al. 2010. Genome-wide association studies in diverse populations. *Nat. Rev. Genet*. 11:356–366.

Rosenberg, N. A., L. M. Li, R. Ward et al. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet*. 73:1,402–1,422.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber et al. 2002. Genetic structure of human populations. *Science* 298:2,381–2,385.

Rosenthal, R. 1994. Parametric measures of effect size. In *The Handbook of Research Synthesis*, H. Cooper and L. V. Hedges, eds. New York: Russell Sage Foundation, 231–244.

Siegel, R., D. Naishadham, and A. Jemal. 2013. Cancer statistics, 2013. *CA Cancer J. Clin*. 63:11–30.

Simons, Y. B., M. C. Turchin, J. K. Pritchard et al. 2014. The deleterious mutation load is insensitive to recent population history. *Nat. Genet*. 46:220–224.

Smouse, P. E., R. S. Spielman, and M. H. Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *Am. Nat.* 119:445–460.

Spitze, K. 1993. Population structure in *Daphnia obtusa*: Quantitative genetic and allozymic variation. *Genetics* 135:367–374.

Tal, O. 2012. The cumulative effect of genetic markers on classification performance: Insights from simple models. *J. Theor. Biol*. 293:206–218.

Teo, Y. Y., K. S. Small, and D. P. Kwiatkowski. 2010. Methodological challenges of genome-wide association analysis in Africa. *Nat. Rev. Genet*. 11:149–160.

Ward, E., A. Jemal, V. Cokkinides et al. 2004. Cancer disparities by race/ethnicity and socioeconomic status. *CA Cancer J. Clin*. 54:78–93.

Weaver, T. D., C. C. Roseman, and C. B. Stringer. 2007. Were Neandertal and modern human cranial differences produced by natural selection or genetic drift? *J. Hum. Evol*. 53:135–145.

Weir, B. S. 1996. *Genetic Data Analysis II*. Sunderland, MA:

Sinauer.

Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* 38:1,358–1,370.

Whitlock, M. C. 1999. Neutral additive genetic variance in a metapopulation. *Genet. Res*. 74:215–221.

Whitlock, M. C. 2008. Evolutionary inference from Q_{ST}. *Mol. Ecol*. 17:1,885–1,896.

Williams, D. R., and C. Collins. 2001. Racial residential segregation: A fundamental cause of racial disparities in health. *Public Health Rep*. 116:404–416.

Williams, D. R., and P. B. Jackson. 2005. Social sources of racial disparities in health. *Health Aff.* (*Millwood*) 24:325–334.

Williams, D. R., and S. A. Mohammed. 2009. Discrimination and racial disparities in health: Evidence and needed research. *J. Behav. Med*. 32:20–47.

Wilson, T. W., and C. E. Grim. 1991. Biohistory of slavery and blood pressure differences in blacks today: A hypothesis. *Hypertension* 17:I122–I128.

Winkler, C. A., G. W. Nelson, and M. W. Smith. 2010. Admixture mapping comes of age. *Annu. Rev. Genomics. Hum. Genet*. 11:65–89.

Wong, M. D., M. F. Shapiro, W. J. Boscardin et al. 2002. Contribution of major diseases to disparities in mortality. *New Engl. J. Med*. 347:1,585–1,592.

Zaitlen, N., B. Pasaniuc, S. Sankararaman et al. 2014. Leveraging population admixture to characterize the heritability of complex traits. *Nat. Genet*. 46:1,356–1,362.

Zlotogora, J., M. Zeigler, and G. Bach. 1988. Selection in favor of lysosomal storage disorders? *Am. J. Hum. Genet*. 42:271–273.

Appendix 1: F_{ST}^k and $F_{ST(\ell)}^k$

In this appendix, we derive some results about the relationship between F_{ST}^k (Eq. 14), which summarizes the information about population membership available in one copy of a typical locus chosen from within a set of k loci, and $F_{ST(\ell)}^k$ (Eq. 16), which summarizes the corresponding population membership information available in ℓ independent copies of a typical locus chosen from within the set of k loci.

For convenience, define ν as

$$\nu = 2[\bar{p}(1-\bar{p}) - s_p^2 + \bar{q}(1-\bar{q}) - s_q^2], \qquad \text{(A1.1)}$$

where $\bar{p}$, $\bar{q}$, s_p^2, and s_q^2 are as defined in Eqs. 2 and 3. Then by Eq. 14,

$$F_{ST}^k = \frac{\overline{\delta^2}}{\nu + \overline{\delta^2}}, \qquad \text{(A1.2)}$$

and by Eq. 16,

$$F_{ST(\ell)}^k = \frac{\ell\overline{\delta^2}}{\nu + \ell\overline{\delta^2}}, \qquad \text{(A1.3)}$$

where ℓ is the number of copies of the locus being considered and $\overline{\delta^2}$ is defined in Eq. 10. Because ν, $\overline{\delta^2}$, and ℓ are all nonnegative, $F_{ST}^k \in [0,1]$ and $F_{ST(\ell)}^k \in [0,1]$. Equations A1.2 and A1.3 also imply that $F_{ST}^k = 0$ if and only if $F_{ST(\ell)}^k = 0$.

By Eqs. 14 and 16, for $F_{ST(\ell)}^k > 0$,

$$\frac{F_{ST}^k}{F_{ST(\ell)}^k} = \frac{\dfrac{\overline{\delta^2}}{\nu + \overline{\delta^2}}}{\dfrac{\ell\overline{\delta^2}}{\nu + \ell\overline{\delta^2}}} = \frac{\ell\overline{\delta^2} + \nu}{\ell\left(\nu + \overline{\delta^2}\right)}.$$

Noting that

$$1 - F_{ST}^k = \frac{\nu}{\nu + \overline{\delta^2}},$$

we then have, for $F_{ST(\ell)}^k > 0$,

$$\frac{F_{ST}^k}{F_{ST(\ell)}^k} = \frac{\overline{\delta^2}}{\nu + \overline{\delta^2}} + \frac{\nu}{\ell(\nu + \overline{\delta^2})} = F_{ST}^k + \frac{1 - F_{ST}^k}{\ell} = \frac{1 + (\ell-1)F_{ST}^k}{\ell}.$$

Consequently, for $F_{ST(\ell)}^k > 0$,

$$\frac{F_{ST(\ell)}^k}{F_{ST}^k} = \frac{\ell}{1 + (\ell-1)F_{ST}^k}. \qquad \text{(A1.4)}$$

Because $F_{ST}^k \in [0,1]$ and $F_{ST}^k = 0$ if $F_{ST(\ell)}^k = 0$, Eq. A1.4 implies that

$$F_{ST(\ell)}^k \in [F_{ST}^k, \ell F_{ST}^k),$$

with $F_{ST}^k = F_{ST(\ell)}^k$ if $\ell = 1$ or if either $F_{ST}^k = 1$ or $F_{ST}^k = 0$, but with $\lim_{F_{ST}^k \to 0} F_{ST(\ell)}^k = \ell F_{ST}^k$. In other words, if $\ell > 1$, for very small but nonzero F_{ST}^k, $F_{ST(\ell)}^k \approx \ell F_{ST}^k$, but for F_{ST}^k near 1, $F_{ST(\ell)}^k \approx F_{ST}^k$. The relationship between F_{ST}^k and $F_{ST(\ell)}^k$ is plotted in Figure 3.

Appendix 2: The Fourth Moment of $\Sigma_{i=1}^{k}\delta_i U_i$

In this appendix, we show that the fourth moment of $\Sigma_{i=1}^{k}\delta_i U_i$ is equal to the expression in Eq. 33.

By the independence of the U_i and Eqs. 29 and 30,

$$\begin{aligned} &E\left[\left(\sum_{i=1}^{k}\delta_i U_i\right)^4\right] \\ &= E\left(\sum_{i=1}^{k}\delta_i^4 U_i^4 + 3\sum_{i=1}^{k}\sum_{j\neq i}\delta_i^2\delta_j^2 U_i^2 U_j^2\right) \\ &= \sum_{i=1}^{k}\delta_i^4 + 3\sum_{i=1}^{k}\sum_{j\neq i}\delta_i^2\delta_j^2. \end{aligned} \tag{A2.1}$$

To simplify the sum in Eq. A2.1, notice that $\Sigma_{j\neq i}\delta_j^2 = k\overline{\delta^2} - \delta_i^2$, so

$$\begin{aligned} &\sum_{i=1}^{k}\sum_{j\neq i}\delta_i^2\delta_j^2 \\ &= \delta_1^2(k\overline{\delta^2} - \delta_1^2) + \delta_2^2(k\overline{\delta^2} - \delta_2^2) + \ldots + \delta_k^2(k\overline{\delta^2} - \delta_k^2) \\ &= k\overline{\delta^2}\sum_{i=1}^{k}\delta_i^2 - \sum_{i=1}^{k}\delta_i^4 = k^2\overline{\delta^2}^2 - k\overline{\delta^4}, \end{aligned} \tag{A2.2}$$

where $\overline{\delta^4}$ is as defined in Eq. 11. Plugging the expression for $\Sigma_{i=1}^{k}\Sigma_{j\neq i}\delta_i^2\delta_j^2$ from Eq. A2.2 into Eq. A2.1 gives

$$\begin{aligned} E\left[\left(\sum_{i=1}^{k}\delta_i U_i\right)^4\right] &= 3k^2\overline{\delta^2}^2 - 2k\overline{\delta^4} \\ &= 3k^2\overline{\delta^2}^2 - 2k\left(\overline{\delta^2}^2 + s_{\delta^2}^2\right), \end{aligned} \tag{A2.3}$$

which proves the statement in Eq. 33.

Appendix 3: Allele Frequencies in a Drift Model

In this appendix, we show that the assumptions in Eqs. 49 and 51 are the expectations of allele frequencies under a population-genetic model in which the two populations experience equal degrees of drift since a recent divergence. The model we use for drift is similar to some models used in previous work (Nicholson et al. 2002; Falush et al. 2003). Note that, although the allele frequencies and p_i and q_i are treated as fixed quantities in the main text, they are treated as random variables in this appendix.

A.3.1 Drift Model

Let $\pi_1, \pi_2, \ldots, \pi_k$, represent the frequencies of one of two alleles at each of k loci in a predivergence population. The π_i may be outcomes of a random process with arbitrary distribution. After a divergence event, the predivergence population splits into two populations—populations A and B—that undergo drift. The amount of drift at each locus is represented by a set of continuous random variables, $\alpha_{A1}, \alpha_{A2}, \ldots, \alpha_{Ak}$ for population A and $\alpha_{B1}, \alpha_{B2}, \ldots, \alpha_{Bk}$ for population B. For all i, conditional on π_i, the α_{Ai} and α_{Bi} are independent, and $E(\alpha_{Ai}) = E(\alpha_{Bi}) = 0$. Further, if the two subpopulations have experienced equal amounts of drift, then conditional on π_i, the α_{Ai} and α_{Bi} are identically distributed. The postdrift allele frequencies in population A are $\pi_1 + \alpha_{A1}, \pi_2 + \alpha_{A2}, \ldots, \pi_k + \alpha_{Ak}$, and in population B they are $\pi_1 + \alpha_{B1}, \pi_2 + \alpha_{B2}, \ldots, \pi_k + \alpha_{Bk}$.

At each locus, define the 1 allele as the allele that is more frequent in population B than in population A. If the allele frequencies are the same in both populations at a locus, then the 1 allele at that locus is chosen randomly, with probability ½ for each allele. The frequency of the 1 allele at locus i is p_i in population A and q_i in population B.

A.3.2 Proposition 1

Proposition 1: If populations A and B have experienced equal amounts of drift since divergence, then $E(p_i) + E(q_i) = 1$.

Proof: Under the drift model outlined above,

(i) if $\alpha_{Ai} < \alpha_{Bi}$, then $p_i = \pi_i + \alpha_{Ai}$ and $q_i = \pi_i + \alpha_{Bi}$;

(ii) if $\alpha_{Ai} = \alpha_{Bi}$, then $p_i = q_i = \tau_i$, where τ_i is either $\pi_i + \alpha_{Ai}$ or $1 - \pi_i - \alpha_{Ai}$ with probability ½ of each possibility; and

(iii) if $\alpha_{Ai} > \alpha_{Bi}$, then $p_i = 1 - \pi_i - \alpha_{Ai}$ and $q_i = 1 - \pi_i - \alpha_{Bi}$.

If the two subpopulations have experienced equal amounts of drift since divergence, then conditional on π_i, the drift variables α_{Ai} and α_{Bi} are independent and identically distributed. Thus, $P(\alpha_{Ai} < \alpha_{Bi}) = P(\alpha_{Ai} > \alpha_{Bi})$, and therefore

$$P(p_i = \pi_i + \alpha_{Ai}) = P(p_i = 1 - \pi_i - \alpha_{Ai}) = ½.$$

Conditional on $\pi_i = w$, we have

$$E(p_i|\pi_i = w) = \tfrac{1}{2}(w + E[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) + \tfrac{1}{2}(1 - w - E[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]).$$

Similarly,

$$E(q_i|\pi_i = w) = \tfrac{1}{2}(w + E[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) + \tfrac{1}{2}(1 - w - E[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]).$$

And so, conditional on $\pi_i = w$,

$$\begin{aligned} E(p_i + q_i|\pi_i = w) &= \tfrac{1}{2}(w + E[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) \\ &\quad + \tfrac{1}{2}(1 - w - E[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) \\ &\quad + \tfrac{1}{2}(w + E[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) \\ &\quad + \tfrac{1}{2}(1 - w - E[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]) \\ &= 1. \end{aligned}$$

Because $E(p_i + q_i|\pi_i = w) = 1$ for all w, the unconditional expectation is

$$E(p_i + q_i) = E_{\pi_i}[E(p_i + q_i|\pi_i = w)] = 1.$$

This completes the proof of Proposition 1.

Thus, taking the mean across loci, $\mathrm{E}(\bar{q}) = 1 - \mathrm{E}(\bar{p})$, and the assumption in Eq. 49 characterizes the expectations of allele frequencies under the model of equal drift since a recent divergence.

A.3.3 Proposition 2

Proposition 2: If populations A and B have experienced equal amounts of drift since divergence, then $E(s_p^2) = E(s_q^2)$.

Proof: Define a random indicator variable G_i with the property

$$\begin{aligned} G_i = 0 &\Leftrightarrow p_i = \pi_i + \alpha_{Ai}, \\ G_i = 1 &\Leftrightarrow p_i = 1 - \pi_i - \alpha_{Ai}. \end{aligned}$$

This property implies that

$$\begin{aligned} G_i = 0 &\Rightarrow p_i = \pi_i + \min(\alpha_{Ai}, \alpha_{Bi}), \\ G_i = 1 &\Rightarrow p_i = 1 - \pi_i - \max(\alpha_{Ai}, \alpha_{Bi}), \end{aligned}$$

because $p_i \le q_i$. By the law of total variance,

$$\mathrm{Var}(p_i|\pi_i = w) = \mathrm{Var}_{G_i}[E(p_i|\pi_i = w, G_i = g)] + E_{G_i}[\mathrm{Var}(p_i|\pi_i = w, G_i = g)].$$

Under the assumption of equal drift since divergence, $P(G_i = 0) = P(G_i = 1) = \tfrac{1}{2}$. Thus, $\mathrm{Var}(p_i|\pi_i = w)$ is equal to the sum of

$$\begin{aligned} &\mathrm{Var}_{G_i}[E(p_i|\pi_i = w, G_i = g)] \\ &= \tfrac{1}{2}\Big[E(p_i|\pi_i = w, G_i = 0) - \tfrac{1}{2}[E(p_i|\pi_i = w, G_i = 0) + E(p_i|\pi_i = w, G_i = 1)]\Big]^2 \\ &\quad + \tfrac{1}{2}\Big[E(p_i|\pi_i = w, G_i = 1) - \tfrac{1}{2}[E(p_i|\pi_i = w, G_i = 0) + E(p_i|\pi_i = w, G_i = 1)]\Big]^2 \\ &= \tfrac{1}{4}\Big[E(p_i|\pi_i = w, G_i = 0) - E(p_i|\pi_i = w, G_i = 1)\Big]^2 \\ &= \tfrac{1}{4}\Big[2w - 1 + E[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w] + E[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]\Big]^2 \end{aligned}$$

and

$$\begin{aligned} &E_{G_i}[\mathrm{Var}(p_i|\pi_i = w, G_i = g)] \\ &= \tfrac{1}{2}\Big[\mathrm{Var}(p_i|\pi_i = w, G_i = 0) + \mathrm{Var}(p_i|\pi_i = w, G_i = 1)\Big] \\ &= \tfrac{1}{2}\Big[\mathrm{Var}[\min(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w] + \mathrm{Var}[\max(\alpha_{Ai}, \alpha_{Bi})|\pi_i = w]\Big]. \end{aligned}$$

A parallel calculation for q_i reveals that $\mathrm{Var}(q_i|\ \pi_i = w)$ is equal to the sum of two equivalent terms, meaning that for all i, $\mathrm{Var}(p_i|\ \pi_i = w) = \mathrm{Var}(q_i|\ \pi_i = w)$. Because this claim holds for all i and because π_i is by definition the same for populations A and B,

$$\mathrm{Var}(p_i) = \mathrm{Var}(q_i).$$

Finally, by Eq. 3, s_p^2 is the biased sample variance of the p_i and s_p^2 is the biased sample variance of the q_i. Applying Bessel's correction to the biased sample variance, the expectations of s_p^2 and s_q^2 under the drift model are, for k loci,

$$\begin{aligned} E(s_p^2) &= \frac{k-1}{k}\mathrm{Var}(p_i), \\ E(s_q^2) &= \frac{k-1}{k}\mathrm{Var}(q_i), \end{aligned}$$

and because $\mathrm{Var}(p_i) = \mathrm{Var}(q_i)$,

$$E(s_p^2) = E(s_q^2).$$

Thus, the assumption in Eq. 51 represents the expectation for properties of the variance of allele frequencies across loci under a model of equal drift since a recent divergence. This completes the proof of Proposition 2.

Appendix 4: The Approximate Trait-Based Misclassification Rate, Conditional on the Labeling of the Alleles

In this appendix, we justify Eq. 56 using an argument similar to the one used to justify Eq. 55. We assume the conditions that apply in Section 6, stated in Eqs. 49 and 51.

W_T is an indicator variable that equals 1 if an individual is misclassified on the basis of its value for T. Conditional on the allelic labels $X_1, \ldots, X_k$, T has a Poisson binomial distribution (see Section 4.1), which, for large k, is well approximated by a normal distribution (Deheuvels et al. 1989, Theorem 1.1). By the conditional expectations and variances in Eqs. 19, 22, and 23 and the assumptions in Eqs. 49 and 51, the large-k distributions of T in the two populations are approximately

$$\begin{aligned}
&(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=A)\\
&\sim \text{Normal}\left(\ell\left[\sum_{i:x_i=1} p_i+\sum_{i:x_i=0}(1-p_{i)})\right],\ell k[\overline{p}\,\overline{q}-s_p^2]\right),\\
&(T\,|\,\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=B)\\
&\sim \text{Normal}\left(\ell\left[\sum_{i:x_i=1} q_i+\sum_{i:x_i=0}(1-q_i)\right],\ell k[\overline{p}\,\overline{q}-s_p^2]\right),
\end{aligned}$$

conditional on the labeling of the alleles, $X_1, \ldots, X_k$.

Denoting the normal density that approximates the distribution of T in population A by $f_{A.T}(t)$ and the corresponding normal density for population B by $f_{B.T}(t)$, then after observing that $T = t$, we classify the individual into population A if $f_{A.T}(t) > f_{B.T}(t)$ and into population B if $f_{A.T}(t) < f_{B.T}(t)$. Because the variances of the two limiting normal distributions are equal, the relationship of the densities depends only on whether the observed trait value t is closer to its expectation in population A or population B. That is, defining

$$\begin{aligned}
\mu_{A.T} &= E(T|\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=A),\\
\mu_{B.T} &= E(T|\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},M=B),
\end{aligned}$$

we have

$$\begin{aligned}
f_{A.T}(t) > f_{B.T}(t) &\Leftrightarrow |t-\mu_{A.T}| < |t-\mu_{B.T}|,\\
f_{A.T}(t) < f_{B.T}(t) &\Leftrightarrow |t-\mu_{A.T}| > |t-\mu_{B.T}|.
\end{aligned}$$

Consider an individual drawn from population A, who will be misclassified if $f_{A.T}(t) < f_{B.T}(t)$. If $\mu_{A.T} < \mu_{B.T}$, then

$$f_{A.T}(t) < f_{B.T}(t) \Leftrightarrow t > (\mu_{A.T}+\mu_{B.T})/2,$$

whereas if $\mu_{A.T} > \mu_{B.T}$, then

$$f_{A.T}(t) < f_{B.T}(t) \Leftrightarrow t < (\mu_{A.T}+\mu_{B.T})/2.$$

(We defer for a moment the case $\mu_{A.T} = \mu_{B.T}$.) Thus, the approximate probability of misclassifying an individual from population A on the basis of its trait value is, if $\mu_{A.T} < \mu_{B.T}$,

$$\begin{aligned}
&P(W_T=1|\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},\mu_{A.T}<\mu_{B.T},M=A)\\
&\quad\approx P\left(T>\frac{\mu_{A.T}+\mu_{B.T}}{2}\right)\\
&\quad\approx 1-\Phi\left(\frac{\frac{\mu_{A.T}+\mu_{B.T}}{2}-\mu_{A.T}}{\sqrt{\ell k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right)\\
&\quad= 1-\Phi\left(\frac{|\mu_{B.T}-\mu_{A.T}|}{2\sqrt{\ell k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right)\\
&\quad= 1-\Phi\left(\frac{\sqrt{\ell}\left|\sum_{i=1}^{k}\delta_i u_i\right|}{2\sqrt{k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right),
\end{aligned}\tag{A4.1}$$

where the last step follows from Eq. 20. Similarly, if $\mu_{A.T} > \mu_{B.T}$, then

$$\begin{aligned}
&P(W_T=1|\{X_1,\ldots,X_k\}=\{x_1,\ldots,x_k\},\mu_{A.T}>\mu_{B.T},M=A)\\
&\quad\approx P\left(T<\frac{\mu_{A.T}+\mu_{B.T}}{2}\right)\\
&\quad\approx \Phi\left(\frac{\mu_{A.T}-\frac{\mu_{A.T}+\mu_{B.T}}{2}}{\sqrt{\ell k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right)\\
&\quad= 1-\Phi\left(\frac{|\mu_{B.T}-\mu_{A.T}|}{2\sqrt{\ell k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right)\\
&\quad= 1-\Phi\left(\frac{\sqrt{\ell}\left|\sum_{i=1}^{k}\delta_i u_i\right|}{2\sqrt{k\left[\overline{p}\,\overline{q}-s_p^2\right]}}\right).
\end{aligned}\tag{A4.2}$$

The expressions in Eqs. A4.1 and A4.2 are equal. Further, the expression they provide also applies to the case of $\mu_{A.T} = \mu_{B.T}$. If $\mu_{A.T} = \mu_{B.T}$, then $P(W_T = 1) = ½$ because the trait has the same distribution in each population. The expression in Eqs. A4.1 and A4.2 applies because if $\mu_{A.T} = \mu_{B.T}$, then by Eq. 20, $\Sigma_{i=1}^{k} \delta_i u_i = 0$, and $\Phi(0) = ½$ as required. A similar set of calculations for population B gives the same expression, so we remove the condition on population membership, arriving at the statement in Eq. 56:

$$P(W_T = 1 | \{X_1, \ldots, X_k\} = \{x_1, \ldots, x_k\}) \approx 1 - \Phi\left[\frac{\sqrt{\ell}\left|\sum_{i=1}^{k} \delta_i u_i\right|}{2\sqrt{k\left[\overline{p}\,\overline{q} - s_p^2\right]}}\right].$$