Genetic Similarities Within and Between Human Populations


* Dept. of Human Genetics, University of Utah Health Sciences Center, Salt Lake City Utah; † Dept. of Anthropology, University of Utah, Salt Lake City, Utah; ‡ The McDermott Center for Human Growth and Development, University of Texas Southwestern Medical Center, Dallas, Texas; § Dept. of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana.
ABSTRACT

The proportion of human genetic variation due to differences between populations is modest, and individuals from different populations can be genetically more similar than individuals from the same population. Yet sufficient genetic data can permit accurate classification of individuals into populations. Both findings can be obtained from the same data set, using the same number of polymorphic loci. This paper explains why.

Our analysis focuses on the frequency, $\omega$, with which a pair of random individuals from two different populations is genetically more similar than a pair of individuals randomly selected from any single population. We compare $\omega$ to the error rates of several classification methods, using data sets that vary in number of loci, average allele frequency, populations sampled, and polymorphism ascertainment strategy. We demonstrate that classification methods achieve higher discriminatory power than $\omega$ because of their use of aggregate properties of populations. The number of loci analyzed is the most critical variable: with 100 polymorphisms, accurate classification is possible, but $\omega$ remains sizable, even when using populations as distinct as sub-Saharan Africans and Europeans. Phenotypes controlled by a dozen or fewer loci can therefore be expected to show substantial overlap between human populations. This provides empirical justification for caution when using population labels in biomedical settings, with broad implications for personalized medicine, pharmacogenetics, and the meaning of race.
INTRODUCTION

Discussions of genetic differences between major human populations have long been dominated by two facts: (a) such differences account for only a small fraction of variance in allele frequencies, but nonetheless (b) multi-locus statistics assign most individuals to the correct population. This is widely understood to reflect the increased discriminatory power of multi-locus statistics. Yet Bamshad et al. (2004) showed, using multi-locus statistics and nearly 400 polymorphic loci, that (c) pairs of individuals from different populations are often more similar than pairs from the same population. If multi-locus statistics are so powerful, then how are we to understand this finding?

All three of the claims listed above appear in disputes over the significance of human population variation and “race.” In particular, the American Anthropological Association (1997) stated that “data also show that any two individuals within a particular population are as different genetically as any two people selected from any two populations in the world” (subsequently amended to “about as different.”) Similarly, educational material distributed by the Human Genome Project and Nature (2001) states that “two random individuals from any one group are almost as different [genetically] as any two random individuals from the entire world.” Previously, one might have judged these statements to be essentially correct for single-locus characters, but not for multi-locus ones. However, the finding of Bamshad et al. (2004) suggests that an empirical investigation of these claims is warranted.

In what follows, we use several collections of loci genotyped in various human populations to examine the relationship between claims a, b, and c above. These data sets vary in
the numbers of polymorphic loci genotyped, population sampling strategies, polymorphism
ascertainment methods, and average allele frequencies. To assess claim c, we define $\omega$ as the
frequency with which a pair of individuals from different populations is genetically more similar
than a pair from the same population. We show that claim c, the observation of high $\omega$, holds
with small collections of loci. It holds even with hundreds of loci, especially if the populations
sampled have not been isolated from each other for long. It breaks down, however, with data sets
comprising thousands of loci genotyped in geographically distinct populations: in such cases, $\omega$
becomes zero. Classification methods similarly yield high error rates with few loci, and almost
no errors with thousands of loci. Unlike $\omega$, however, classification statistics make use of
aggregate properties of populations, so they can approach 100% accuracy with as few as 100
loci.
MATERIALS AND METHODS

Data sets

Three data sets were used. Loci or individuals with more than 10% missing data were not included in any data set (loci were pruned first, then individuals). The first data set (“Insertions”) consists of 175 polymorphic transposable element insertion loci (100 Alu and 75 L1) previously genotyped in 253 individuals. The population sample consists of 101 individuals from sub-Saharan Africa, 53 East Asians, 61 individuals of northern European ancestry, and 38 individuals from Andhra Pradesh, India (Watkins et al. 2005; Witherspoon et al. 2006). The second data set (“Microarray”) consists of 9,922 biallelic single-nucleotide polymorphism (SNP) loci genotyped in 278 individuals (55 Africans, 42 African Americans, 40 Native Americans, 22 Indians, 20 East Asians, 62 Europeans, 18 Hispano-Latinos from Puerto Rico, and 19 individuals from New Guinea). This data set is derived from that of Shriver et al (2005). The third data set (“Resequenced”) is derived from the ten ENCODE regions of the HapMap project, release 16c.1 of phase I, June 2005 (International HapMap Consortium 2005). These regions were resequenced in 48 individuals in order to identify SNPs without ascertainment bias in favor of loci with common polymorphisms. These SNPs were then genotyped in 209 unrelated individuals: 60 Yoruba in Ibadan, Nigeria (YRI); 60 Utah residents with ancestry from northern and western Europe (CEU, from the CEPH diversity panel); and 89 Japanese in Tokyo, Japan, plus Han Chinese in Beijing, China (CHB+JPT). Our subset consists of 14,258 SNPs. All markers in all three data sets are biallelic. The proportions of missing genotypes are 2.5%, 2.1%, and 0.36%, respectively.
**Data subsampling**

To examine the effect of population sampling (i.e., the effects of comparing relatively isolated populations vs. more closely related or admixed ones), two subsets were constructed from each of the Insertions and Microarray main data sets: one consisting of the entire data set, with all its labeled populations, and another consisting of East Asian, European, and sub-Saharan African population groups only. The Resequenced data set consists only of the latter three population groups.

In order to investigate the effect of allele frequency, these five data subsets were subdivided according to three further treatments: loci with common polymorphisms (with minor allele frequency or MAF > 0.1); loci with rare polymorphisms (MAF < 0.1); and all polymorphic loci, regardless of frequency. Henceforth we refer to these classes of loci as rare polymorphisms, common polymorphisms, or all polymorphisms. For this classification, allele frequencies were computed across the entire sample in the parent data set. To investigate the effect of incrementally increasing the number of loci used, loci from each of these 15 data subsets were sampled (without replacement) to produce 200 independent data sets with numbers of loci varying in 21 steps on a logarithmic scale from 10 to the maximum.

**Pairwise Genetic Distance**

We use the “shared alleles” genetic distance (Bowcock *et al.* 1994; Chakraborty and Jin 1993; Mountain and Cavalli-Sforza 1997), which defines the distance between two individuals at a locus as one minus half the number of alleles they share. The genetic distance between individuals is the average of their per-locus distances. Pairs of individuals are classified as “within-population” or “between-population” according to whether the individuals were sampled from the same or different groups of populations as defined above.
Dissimilarity fraction $\hat{\omega}$

Let $\omega$ be the probability that a pair of individuals randomly chosen from different populations are genetically more similar than an independent pair chosen from any single population. We compute all possible pairwise genetic distances, classify them as within- or between-population distances (the sets $d_W$ or $d_B$, respectively), then calculate the frequency with which $d_W > d_B$ (that is, a within-population pair is more dissimilar than a between-population pair.) This fraction, $\hat{\omega}$, is an estimator of $\omega$. The expected value of $\hat{\omega}$ ranges from 0 to 0.5 (regardless of the number of populations). At $\hat{\omega} = 0$, individuals are always more similar to members of their own population than to members of other populations; at $\hat{\omega} = 0.5$, individuals are as likely to be more similar to members of other populations as to members of their own. The distributions of pairwise genetic distances implied here resemble the Common Ancestry Profiles proposed by Mountain and Ramakrishnan (2005), who use a different measure of genetic distance. The shared-alleles distance used here generally yields slightly lower values of $\hat{\omega}$.

Centroid misclassification rate $C_C$

The centroid classification method is also based on pairwise genetic distances, with one critical difference: every individual is compared to the centroid of each population, rather than to every other individual. The centroid is the genetic average of a population, an individual whose pseudo-genotypes at each locus are the frequencies of the genotypes in that population (not including the individual being compared to the centroid). This genetic distance is equivalent to the average of the genetic distances from an individual to all other individuals in the target population. Each individual is then assigned to the population with the closest centroid, as in Cornuet et al. (1999). These assignments are compared to the known populations of origin, and the proportion of individuals misclassified is reported as $C_C$. The expected classification error for
random assignment of individuals to populations is \(1 - 1/n\), where \(n\) is the number of populations.

**Population Trait Value misclassification rate \(C_T\)**

Our definition of \(C_T\) is implicit in the theoretical illustrations of Risch *et al.* (2002) and Edwards (2003). These authors used simplified models to show how modest differences between populations can nonetheless enable accurate classification. In both cases, population membership is treated as an additive quantitative genetic trait controlled by many loci of equal effect, and individuals are divided into populations based on their trait values.

This method is inherently limited to dividing individuals into just two clusters using only biallelic loci, so we limit our definitions to that situation. Consider individuals sampled from two populations, \(A\) and \(B\), and genotyped at many biallelic loci. At each locus, we identify the allele whose frequency is higher in population \(A\) and assign it a value of 0. The other allele (more frequent in \(B\) than in \(A\)) is assigned a value of 1. Let \(q_{ij}\) represent the genotype of individual \(i\) at locus \(j\), defined as the average of the assigned values of the two alleles carried by that individual at that locus. Now define \(q_i\) as the average of \(q_{ij}\) over all loci \(j\) (so \(q_i\) is a polygenic quantitative genetic trait.) Given these definitions, if populations \(A\) and \(B\) are typified by even slightly different allele frequencies at many loci, then \(q_i\) will usually be smaller for a member of population \(A\) than a member of population \(B\). Thus the value of the trait \(q_i\) indicates membership in one population or the other, so we call \(q_i\) the “population trait” value of individual \(i\).

Individuals are assigned to population \(A\) or \(B\) depending on whether their population trait value \(q_i\) falls below or above some dividing criterion \(q_C\), respectively. In the case of just two populations, these assignments are compared to the known origins of the individuals, and the proportion misclassified is reported as \(C_T\). The classification criterion \(q_C\) is chosen as follows.
Let $\overline{q}_A$ be the mean of $q_i$ taken over all individuals in population $A$, and define $\overline{q}_B$ similarly for population $B$. If the distributions of $q_i$ for individuals from the two populations are symmetric with equal variance, then letting $q_C = (\overline{q}_A + \overline{q}_B)/2$ minimizes misclassification (cf. Edwards 2003; Risch et al. 2002). To better account for unequal variances, we generalize slightly and solve for a criterion $q_C$ such that $r(q_C) = s(q_C)$ and $\overline{q}_A < q_C < \overline{q}_B$, where $r$ and $s$ are normal probability density functions with means and variances estimated from the distributions of $q_i$ for populations $A$ and $B$, respectively.

To extend this inherently pairwise approach to more than two populations, assignments for each individual are initially computed with reference to each possible pair of populations. The values (0 or 1) assigned to particular alleles, the criterion $q_C$, and all $q_i$ are calculated anew for each pair of populations. Individuals are finally assigned to a population only if they were assigned to it in all pairwise comparisons involving that population. The proportion of individuals misclassified (or not classified, since this method can fail to classify individuals) is reported as $C_T$. For comparison, a “single-locus” classification error rate is computed by using this method to classify individuals using each locus singly, then averaging the results over all loci.
RESULTS

**Distributions of Distances**

The statistics $\hat{\omega}$, $C_C$, and $C_T$ are closely related by design. To illustrate the relationships between them, the distributions of the genetic measures that underlie them are shown in Figure 1. For simplicity, only two populations (Europeans and sub-Saharan Africans) in the Insertions data set (175 loci) are used. The distributions of pairwise genetic distances for within- and between-population pairs of individuals (Figure 1A) overlap considerably even for these geographically isolated populations. The dissimilarity fraction, $\hat{\omega}$, is 23%, indicating that between-population pairs are more similar than within-population pairs nearly one-quarter of the time. In contrast, the distributions of individuals’ distances to the centroids of their own or different populations (Figure 1B) show much less overlap, resulting in $C_C = 4.9\%$. The population trait value distributions for Africans and Europeans overlap for just three individuals, yielding $C_T = 1.9\%$. Classifications using model-based methods such as *structure* (Pritchard et al. 2000) achieve 95% accuracy or better using the same data (Bamshad et al. 2003; Witherspoon et al. 2006).

The variances of the distributions are much greater for the individual-to-individual comparisons (Figure 1A) than for the centroid-to-individual comparisons (Figure 1B). The distribution means are nearly identical, however, so the distributions overlap more in panel A than in B, and thus $\hat{\omega} > C_C$. The difference in variances is due to the fact that each genetic distance to a centroid (each datum in panel B) is equivalent to the average of a sizable subset of pairwise genetic distances represented in panel A (see Methods). That averaging step eliminates considerable variation and produces the narrower distributions of panel B.
The simplifications introduced by Risch et al. (2002) and Edwards (2003) allow an alternative view, represented in Figure 1C. Here, each individual $i$ is assigned a unidimensional genetic location $q_i$ (the individual’s population trait value; see Methods.) The trait distance between any two individuals $x$ and $y$ is now just the horizontal distance between them, $|q_x - q_y|$. This simplification is possible only in the two-population case and requires a population-specific coding of allele states, so the trait distance is not equivalent to the genetic distances represented in panels A and B. Nonetheless, it is instructive to consider the analogy using panel C as a guide. For example, an African individual $x$ with $q_x = 0.52$ will be more similar to a European $y$ with $q_y = 0.55$ than to another African $z$ with $q_z = 0.4$. Yet that individual $x$ will still be closer to the population mean trait value for Africans ($q_A \cong 0.46$, the African centroid) than to the mean value of Europeans ($q_B \cong 0.61$). It follows that many individuals like this one will be correctly classified (yielding low $C_C$ and $C_T$) even though they are often more similar to individuals of the other population than to members of their own population (yielding high $\hat{\omega}$).

In order to empirically and quantitatively understand the relationships and contrasts between $\hat{\omega}$ and the misclassification rates $C_C$ and $C_T$, we examine three primary factors that influence them: the number of polymorphic loci used; the allele frequencies at those loci; and the degree of differentiation between the populations examined.

**Data Subset Statistics**

Three data sets, labeled “Insertions,” “Microarray,” and “Resequenced,” were used, and fifteen subsets were constructed from these to examine the effects of different data collection strategies (see Methods). Table 1 lists the fifteen data subsets and reports $\hat{\omega}$, $C_C$, and $C_T$ (each computed over all loci in each data subset) as well as the expected value of $C_T$ when only a single locus is used. Table 1 also gives values of five descriptive statistics for each data subset:
the proportion of genetic variance explained by interpopulation differences \((F_{ST})\); the observed proportions of heterozygotes (% Het); the absolute differences in allele frequencies between population pairs (averaged), \(\bar{\delta}\); the fraction of polymorphisms that are rare (MAF < 0.1) in at least one population and at the same time common (MAF > 0.1) in another population (% Rare & Common); and the fraction of loci that are monomorphic in at least one population and common polymorphisms in another (% Fixed & Common). The values observed are typical of human population genetic data sets (Dean et al. 1994; International HapMap Consortium 2005; Nei 1973; Shriver et al. 2005; Witherspoon et al. 2006).

\textbf{Dependency of \(\hat{\omega}\) on Number of Loci}

Figure 2 shows the dependency of \(\hat{\omega}, C_C\) and \(C_T\) on the number of loci for each of the 15 data subsets listed in Table 1. As the number of included loci increases, \(\hat{\omega}, C_C\) and \(C_T\) decrease. This is the expected behavior for \(C_C\) (Campbell et al. 2003; Manel et al. 2002; Smouse and Chakraborty 1986) and \(C_T\) (Edwards 2003; Risch et al. 2002). However, \(\hat{\omega}\) does not decrease nearly as rapidly. Figure 2A shows the results for a diverse sample of individuals genotyped at 175 insertion loci, a number that is typical of many studies of human genetic diversity published during the last decade. The downward trend in \(\hat{\omega}\) is apparent, but even with the full data set, it remains at 25% (with all four population groups; Table 1). Across all data sets and using fewer than 100 polymorphisms, \(\hat{\omega}\) generally exceeds 10% (Figure 2). With fewer than 100 loci, then, it will often be the case that two individuals from different populations are more similar to one another than are two individuals from the same population.

The power of large numbers of common polymorphisms is most apparent in the Microarray data set, comparing the European, East Asian and sub-Saharan African population groups (Figure 2C). \(\hat{\omega}\) approaches zero (median 0.12%) with 1,000 polymorphisms. This implies
that, when enough loci are considered, individuals from these population groups will always be
genetically most similar to members of their own group. In general, $C_C$ and $C_T$ decrease more
rapidly and to lower values than $\hat{\omega}$.

**Allele frequency effects**

The “Rare” polymorphism subsets defy this trend by converging toward high values of $\hat{\omega}$
as loci are added. This is largely because the frequencies of rare polymorphisms are necessarily
quite similar across populations, whereas higher-frequency polymorphisms have the potential to
differ more. For example, the frequency of an allele with an overall MAF of 5% can differ by at
most $\delta = 10\%$ between two populations (absent in one, at 10\% frequency in another). This
situation yields $\hat{\omega} > 0$ and very poor classification accuracy, since most between-population
pairs are identical but some within-population pairs differ. In contrast, an allele with an overall
frequency of 50\% across two populations could be fixed in one and absent in the other, resulting
in $\hat{\omega} = 0$ and allowing perfect classification. It is these frequency differences that allow
populations to be distinguished, so the data sets with lower $\bar{\delta}$ (and thus generally lower $F_{ST}$)
have lower classification power.

The sensitivity of these statistics to allele frequencies explains some differences between
the data sets. The Microarray data set exhibits strong ascertainment bias for common
polymorphisms, and it is with this data set that $\hat{\omega}$ drops most rapidly and to its lowest values
(Figure 2C and 2D). The Insertions data set exhibits a weaker ascertainment bias and includes
more rare polymorphisms, so $\hat{\omega}$ remains higher (Figure 2A and 2B). Similarly, $C_C$ and $C_T$ drop
more rapidly for the Microarray data set than for the Insertion data set. The Resequenced data set
polymorphisms were ascertained by resequencing a sizable panel of individuals from the
genotyped populations and thus include many rare polymorphisms, but this is partially offset by
the equally large number of common polymorphisms (Figure 2E). The classification methods are less affected by the inclusion of rare polymorphisms.

**Population Sampling Effects**

We contrast two choices: sets of populations that have been relatively isolated from each other by geographic distance and barriers since the earliest migrations of modern humans out of Africa; and sets that include populations that were founded more recently, are geographically closer to one another and therefore more likely to exchange migrants, or have recently experienced a large genetic influx from another population in the set. Sampling only from the more distinct populations yields lower $\hat{\omega}$ values, as expected. Figure 2, panels A, C, and E show the results of using only the three most distinct population groups (Europeans, East Asians, and sub-Saharan Africans.) Panels B and D expand the samples used in panels A and C to include recently founded and/or geographically intermediate populations (Indians in the Insertions data set; New Guinean, South Asian, and Native American in the Microarray data set) and “admixed” populations (i.e., those that have recently received many migrants from different populations, such as the African American and Hispano-Latino groups in the Microarray data set). With just 175 loci, choosing to sample distinct populations vs. more closely related ones only makes a modest difference (Insertions data set, compare Figure 2A to 2B; Table 1). The effect of population sampling becomes more pronounced when 1,000 or more loci are available. In the Microarray data set, $\hat{\omega}$ drops to zero at 1,000 loci if only distinct populations are sampled. With geographically intermediate and admixed populations added, however, $\hat{\omega}$ reaches an asymptotic value of 3.1%, $C_C$ remains well above zero, and even $C_T$ does not reach zero. (Microarray data, Figure 2C and 2D; Table 1.)
\( \hat{\omega} \) also appears to reach a nonzero asymptotic value in the Resequenced data set, instead of continuing to trend downward as would be expected given the distinct populations used. This may be due to the fact that many of the polymorphisms in that data set are physically linked and therefore nonindependent. Overall, the responses of the two classification methods to data set composition variables are qualitatively similar to the behavior of \( \hat{\omega} \) (Figure 2). The most apparent difference is that the misclassification rates \( (C_C \text{ and } C_T) \) decrease much more rapidly, and to lower values, than \( \hat{\omega} \) does as the number of loci considered increases.

**DISCUSSION**

It has long been appreciated that differences between human populations account for only a small fraction of the total variance in allele frequencies (typically presented as \( F_{ST} \) values of 10-15%; Barbujani et al. 1997; International HapMap Consortium 2005; Jorde et al. 2000; Latter 1980; Lewontin 1972; Nei and Roychoudhury 1972; Rosenberg et al. 2005; Watkins et al. 2003). Such observations triggered controversy from the outset. Some geneticists concluded the differences were negligible (Lewontin 1972); others disagreed (Mitton 1977). Despite the limited data, it soon became apparent that even a modest number of loci should allow accurate assignment of individuals to populations (Mitton 1978; Smouse et al. 1982).

More recently, the Human Genome Project (HGP 2001) highlighted the basic genetic similarity of all humans, yet subsequent analyses demonstrated that genetic data can be used to accurately classify humans into populations (Bamshad et al. 2003; Lao et al. 2006; Rosenberg et al. 2002, 2005; Tang et al. 2005; Turakulov and Easteal 2003). Risch et al. (2002) and Edwards (2003) used theoretical illustrations to show why accurate classification is possible despite the slight differences in allele frequencies between populations. These illustrations suggest that, if
enough loci are considered, two individuals from the same population may be genetically more similar (i.e. more closely related) to each other than to any individual from another population (as foreshadowed by Powell and Taylor 1978). Accordingly, Risch et al. (2002) state that “two Caucasians are more similar to each other genetically than a Caucasian and an Asian.” However, in a reanalysis of data from 377 microsatellite loci typed in 1,056 individuals, Europeans proved to be more similar to Asians than to other Europeans 38% of the time (Bamshad et al. 2004; population definitions and data from Rosenberg et al. 2002).

With the large and diverse data sets now available, we have been able to evaluate these contrasts quantitatively. Even the pairwise relatedness measure, $\hat{\omega}$, can show clear distinctions between populations if enough polymorphic loci are used. Observations of high $\hat{\omega}$ and low classification errors are the norm with intermediate numbers of loci (up to several hundred). These results bear out the observations of Bamshad et al. (2004). The high $\hat{\omega}$ observed there was due primarily to the slow rate of decrease of $\hat{\omega}$ with increasing numbers of loci. Although Rosenberg et al. (2002) achieved a very low misclassification rate with the same data, far more loci would be needed to reduce $\hat{\omega}$ to similarly small values (assuming such values could be reached at all for those populations.)

Thus the answer to the question “How often is a pair of individuals from one population genetically more dissimilar than two individuals chosen from two different populations?” depends on the number of polymorphisms used to define that dissimilarity and the populations being compared. The answer, $\hat{\omega}$, can be read from Figure 2. Given ten loci and three distinct populations, the answer is $\hat{\omega} \approx 0.3$, or nearly one-third of the time. With 100 loci, the answer is about 20% of the time; and even using 1,000 loci, $\hat{\omega} \approx 10\%$. However, if genetic similarity is
measured over many thousands of loci, the answer becomes “never” when individuals are sampled from geographically separated populations.

On the other hand, if the entire world population were analyzed, the inclusion of many closely related and admixed populations would increase $\hat{\omega}$. This is illustrated by the fact that $\hat{\omega}$ and the classification error rates, $C_C$ and $C_T$, all remain greater than zero when such populations are analyzed, despite the use of more than 10,000 polymorphisms (Microarray data set, Table 1; Figure 2D). In a similar vein, Romualdi et al. (2002) and Serre and Pääbo (2004) have suggested that highly accurate classification of individuals from continuously sampled (and therefore closely related) populations may be impossible. However, those studies lacked the statistical power required to answer that question (see Rosenberg et al. 2005).

How can the observations of accurate classifiability be reconciled with high between-population similarities among individuals? Classification methods typically make use of aggregate properties of populations, not just properties of individuals or even of pairs of individuals. For instance, the centroid classification method computes the distances between individuals and population centroids, then clusters individuals around the nearest centroid. The population trait method relies on information about the frequencies of each allele in each population to compute individual trait values and on the means and variances of the trait distributions to classify individuals. The structure classification algorithm (Pritchard et al. 2000) also relies on aggregate properties of populations, such as Hardy-Weinberg and linkage equilibrium. In contrast, the pairwise distances used to compute $\hat{\omega}$ make no use of population-level information and are strongly affected by the high level of within-groups variation typical of human populations. This accounts for the difference in behavior between $\hat{\omega}$ and the classification results.
Since an individual’s geographic ancestry can often be inferred from his or her genetic makeup, knowledge of one’s population of origin should allow some inferences about individual genotypes. To the extent that phenotypically important genetic variation resembles the variation studied here, we may extrapolate from genotypic to phenotypic patterns. Resequencing studies of gene-coding regions show patterns similar to those seen here (e.g. Stephens et al. 2001), and many common disease-associated alleles are not unusually differentiated across populations (Lohmueller et al. 2006). Thus it may be possible to infer something about an individual’s phenotype from knowledge of his or her ancestry.

However, consider a hypothetical phenotype of biomedical interest that is determined primarily by a dozen additive loci of equal effect whose worldwide distributions resemble those in the Insertion data set (e.g., with $\delta = 0.15$; Table 1). Given these assumptions, the genetic distance used in computing $\hat{\omega}$ and $C_C$ is equivalent to a phenotypic distance, so Figure 2 can be used to analyze this hypothetical trait. Figure 2A shows that a trait determined by 12 such loci will typically yield $\hat{\omega} = 0.36$ (0.29-0.43) and $C_C = 0.32$ (0.20-0.47; medians and 90% ranges). About a third of the time ($\hat{\omega} = 0.36$), an individual will be phenotypically more similar to someone from another population than to another member of the same population. Similarly, individuals will be more similar to the average or “typical” phenotype of another population than to the average phenotype in their own population with a probability of about one-third ($C_C = 0.32$). It follows that variation in such a trait will often be discordant with population labels.

The population groups in this example are quite distinct from one another: Europeans, sub-Saharan Africans, and East Asians. Many factors will further weaken the correlation between an individual’s phenotype and their geographic ancestry. These include: considering more closely related or admixed populations; studying phenotypes influenced by fewer loci;
unevenly distributed effects across loci; nonadditive effects; developmental and environmental effects; and uncertainties about individuals’ ancestry and actual populations of origin. The typical frequencies of alleles that influence a phenotype are also relevant, as our results show that rare polymorphisms yield high values of $\hat{\omega}, C_C$ and $C_T$, even when many such polymorphisms are studied. This implies that complex phenotypes influenced primarily by rare alleles may correspond poorly with population labels and other population-typical traits (in contrast to some Mendelian diseases). However, the typical frequencies of alleles responsible for common complex diseases remain unknown. A final complication arises when racial classifications are used as proxies for geographic ancestry. Although many concepts of “race” are correlated with geographic ancestry, the two are not interchangeable, and relying on racial classifications will reduce predictive power still further.

The fact that, given enough genetic data, individuals can be correctly assigned to their populations of origin is compatible with the observation that most human genetic variation is found within populations, not between them. It is also compatible with our finding that, even when the most distinct populations are considered and hundreds of loci are used, individuals are frequently more similar to members of other populations than to members of their own population. Thus, caution should be used when using geographic or genetic ancestry to make inferences about individual phenotypes.

ACKNOWLEDGMENTS

We thank Jinchuan Xing, Michael Bamshad, Dennis O’Rourke and Thomas Doak for thoughtful comments. This work was supported by National Science Foundation BCS-0218338 (MAB), BCS-0218370 (LBJ), EPS-0346411 (MAB), National Institutes of Health GM-59290
(LBJ and MAB) and Louisiana Board of Regents Millennium Trust Health Excellence Fund HEF (2000-05)-05 (M.A.B.), (2000-05)-01 (M.A.B.) and (2001-06)-02 (M.A.B.), and by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health.
LITERATURE CITED


<table>
<thead>
<tr>
<th>Data set</th>
<th>MAF class</th>
<th>Populations</th>
<th>Ind. Loci</th>
<th>$\hat{\omega}$%</th>
<th>$C_C$%</th>
<th>$C_T$%</th>
<th>$C_T$, Single-Locus%</th>
<th>$F_{ST} \times 10^0$</th>
<th>% Rare</th>
<th>% Fixed</th>
<th>$\delta_{100}$</th>
<th>% Het.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insertions</td>
<td>All</td>
<td>Distinct (3)</td>
<td>215</td>
<td>175</td>
<td>23</td>
<td>6.5</td>
<td>6.5</td>
<td>59</td>
<td>9.5</td>
<td>15</td>
<td>2.3</td>
<td>15</td>
</tr>
<tr>
<td>Insertions</td>
<td>Rare</td>
<td>Distinct (3)</td>
<td>215</td>
<td>21</td>
<td>47</td>
<td>57</td>
<td>44</td>
<td>66</td>
<td>10</td>
<td>48</td>
<td>9.5</td>
<td>4.9</td>
</tr>
<tr>
<td>Insertions</td>
<td>Common</td>
<td>Distinct (3)</td>
<td>215</td>
<td>154</td>
<td>21</td>
<td>5.1</td>
<td>5.6</td>
<td>59</td>
<td>9.5</td>
<td>11</td>
<td>1.3</td>
<td>16</td>
</tr>
<tr>
<td>Insertions</td>
<td>All</td>
<td>All (4)</td>
<td>253</td>
<td>175</td>
<td>25</td>
<td>14</td>
<td>11</td>
<td>68</td>
<td>8.2</td>
<td>17</td>
<td>2.9</td>
<td>13</td>
</tr>
<tr>
<td>Insertions</td>
<td>Rare</td>
<td>All (4)</td>
<td>253</td>
<td>20</td>
<td>48</td>
<td>69</td>
<td>63</td>
<td>72</td>
<td>2.7</td>
<td>50</td>
<td>15</td>
<td>3.1</td>
</tr>
<tr>
<td>Insertions</td>
<td>Common</td>
<td>All (4)</td>
<td>253</td>
<td>155</td>
<td>24</td>
<td>11</td>
<td>9.9</td>
<td>68</td>
<td>8.3</td>
<td>12</td>
<td>1.3</td>
<td>14</td>
</tr>
<tr>
<td>Microarray</td>
<td>All</td>
<td>Distinct (3)</td>
<td>137</td>
<td>9,922</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>56</td>
<td>15</td>
<td>26</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>Microarray</td>
<td>Rare</td>
<td>Distinct (3)</td>
<td>137</td>
<td>920</td>
<td>17</td>
<td>7.3</td>
<td>0</td>
<td>65</td>
<td>10</td>
<td>87</td>
<td>56</td>
<td>9.5</td>
</tr>
<tr>
<td>Microarray</td>
<td>Common</td>
<td>Distinct (3)</td>
<td>137</td>
<td>9,002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>15</td>
<td>19</td>
<td>7.8</td>
<td>21</td>
</tr>
<tr>
<td>Microarray</td>
<td>All</td>
<td>All (8)</td>
<td>278</td>
<td>9,922</td>
<td>3.1</td>
<td>10</td>
<td>1.1</td>
<td>82</td>
<td>13</td>
<td>44</td>
<td>27</td>
<td>18</td>
</tr>
<tr>
<td>Microarray</td>
<td>Rare</td>
<td>All (8)</td>
<td>278</td>
<td>851</td>
<td>36</td>
<td>72</td>
<td>1.4</td>
<td>84</td>
<td>8.9</td>
<td>98</td>
<td>88</td>
<td>7.9</td>
</tr>
<tr>
<td>Microarray</td>
<td>Common</td>
<td>All (8)</td>
<td>278</td>
<td>9,071</td>
<td>3.3</td>
<td>8.3</td>
<td>1.1</td>
<td>81</td>
<td>13</td>
<td>39</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Resequenced</td>
<td>All</td>
<td>All (3)</td>
<td>209</td>
<td>14,233</td>
<td>8.2</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>13</td>
<td>40</td>
<td>29</td>
<td>11</td>
</tr>
<tr>
<td>Resequenced</td>
<td>Rare</td>
<td>All (3)</td>
<td>209</td>
<td>89</td>
<td>29</td>
<td>45</td>
<td>0.96</td>
<td>59</td>
<td>9.6</td>
<td>51</td>
<td>43</td>
<td>4.6</td>
</tr>
<tr>
<td>Resequenced</td>
<td>Common</td>
<td>All (3)</td>
<td>209</td>
<td>5,844</td>
<td>5.9</td>
<td>0.96</td>
<td>0</td>
<td>53</td>
<td>14</td>
<td>23</td>
<td>8.4</td>
<td>19</td>
</tr>
</tbody>
</table>
a Sets of loci with MAF < 0.1 (Rare), MAF > 0.1 (Common), or any MAF (All).

b Sets of individuals representing all populations in a data set or only the more distinct populations.

c Dissimilarity fraction.

d Centroid misclassification rate.

e Population trait value misclassification rate.

f Trait value misclassification rate based on a single locus, averaged over all loci.

g $F_{ST}$, the proportion of variance in allelic frequencies attributable to population differences.

h Percentage of loci with MAF < 0.1 in one population and MAF > 0.1 in another.

i Percentage of loci that are monomorphic in one population and polymorphic with MAF > 0.1 in another.

j Average difference in allele frequency between populations, averaged across alleles, loci and population pairs.

k Percentage of individuals heterozygous at each locus, averaged across all loci.
Figure 1

Frequency distributions of the underlying genetic measures used to compute $\hat{\omega}$, $C_C$, and $C_T$, for 175 loci genotyped in 101 sub-Saharan African and 61 European individuals of the Insertions data set. The measures shown are (A) 13,041 pairwise genetic distances for within- and between-population pairs of individuals (in blue and red, respectively); (B) 324 genetic distances between each individual and the centroid of each population, for an individual’s known population of origin (blue) or other population (red); and (C) 162 population trait values $q_i$ for individuals computed relative to the African vs. European population pair. Alleles more common in Africa than in Europe are assigned a value of 0; those more common in Europe are assigned a value of 1. The classification criterion $q_C$ is marked. The $q_i$ distributions for Africans and Europeans are green and yellow, respectively. The areas of overlap between the distributions do not correspond directly to the dissimilarity fraction or misclassification rates. Distributions that do not overlap imply that $\hat{\omega} = 0$ and all individuals can be correctly classified. In (C), only three individuals are misclassified. Means and standard deviations are indicated above each distribution by vertical ticks and horizontal bars. The horizontal axes share the same scale. As the number of polymorphic loci used increases, the variances of these distributions decrease while their means remain roughly constant. As a result, the statistics $\hat{\omega}$, $C_C$, and $C_T$ decrease as more loci are used.
Figure 2

Behavior of the dissimilarity fraction (\( \hat{\omega} \)) and error rates of the “centroid” (\( C_C \)) and “population trait” (\( C_T \)) classification methods (red, blue and green lines, respectively) for each of 15 data subsets (see Table 1 and Methods.) The number of loci subsampled varies in 21 logarithmic steps from 10 to the maximum for each data subset. At each step, all three statistics were computed for 200 subsampled data sets. Lines indicate the medians of the resulting distributions. Within each panel, separate series represent three polymorphism frequency subsets: Rare (MAF < 10%; blue contours), Common (MAF > 10%; in green), and All (all polymorphisms; in black, see Legend.) Results computed from the data subsets derived from the Insertion, Microarray, and Resequenced data sets are shown in panels (A, B), (C, D), and (E), respectively. Panels A and C show results from analyses that use only the three most distinct population groups (Europeans, East Asians, and sub-Saharan Africans, abbreviated Eu, EA, and Af), while panels B and D show results based on all populations in the Insertion and Microarray data sets, respectively (Indian, Native American, New Guinean, African-American, and Hispano-Latino, abbreviated In, NA, NG, AfAm, and HL.) Panel E uses all three population groups in the Resequenced data set.
A. Pairwise Genetic Distances within and between populations
\( \delta = 0.23 \)

B. Means of Pairwise Distances within and between populations
\( C_C = 0.049 \)

C. Population Trait Values, \( q_i \)
for individuals of two populations
\( C_T = 0.019 \)