Genetic Diversity and Societally Important Disparities

Noah A. Rosenberg[1] and Jonathan T. L. Kang
Department of Biology, Stanford University, Stanford, California 94305-5020

ABSTRACT The magnitude of genetic diversity within human populations varies in a way that reflects the sequence of migrations by which people spread throughout the world. Beyond its use in human evolutionary genetics, worldwide variation in genetic diversity sometimes can interact with social processes to produce differences among populations in their relationship to modern societal problems. We review the consequences of genetic diversity differences in the settings of familial identification in forensic genetic testing, match probabilities in bone marrow transplantation, and representation in genome-wide association studies of disease. In each of these three cases, the contribution of genetic diversity to social differences follows from population-genetic principles. For a fourth setting that is not similarly grounded, we reanalyze with expanded genetic data a report that genetic diversity differences influence global patterns of human economic development, finding no support for the claim. The four examples describe a limit to the importance of genetic diversity for explaining societal differences while illustrating a distinction that certain biologically based scenarios do require consideration of genetic diversity for solving problems to which populations have been differentially predisposed by the unique history of human migrations.

KEYWORDS forensic DNA; genetic diversity; genome-wide association; human migration; transplantation matching

THE publication of an article suggesting that geographic patterns in economic development across countries worldwide have been driven by genetic diversity (Ashraf and Galor 2013) has generated considerable controversy (Callaway 2012; Chin 2012; Gelman 2013; Feldman 2014). Connecting data on measures of genetic diversity in different human populations to proxy measures of economic success, Ashraf and Galor (2013) argued that an optimal level of genetic diversity exists for enhancing the economic development of nations and that the optimum lies in an intermediate range characteristic of populations of Europe and Asia. In response to prepublication reports of the upcoming paper, a large interdisciplinary group of scholars vehemently criticized the methods and conclusions, objecting to the line of inquiry on genetic determination of economic outcomes on the grounds of its potential for misuse (d'Alpoim Guedes *et al.* 2013).

This controversial attempt to apply a population-genetic variable in an analysis of a societal outcome provides an occasion to examine the ways in which population differences in genetic diversity might contribute to consequential societal differences across populations. Several such examples have been reported. After reviewing the origins of differences across human populations in levels of genetic diversity, we describe three documented cases in which the variation in genetic diversity across populations interacts with social processes to produce population differences in important outcomes. We then return to the economic development study, investigating a genetic data set that expands beyond the data examined by Ashraf and Galor (2013). Our analysis finds that even when the same methods used by Ashraf and Galor are applied to this larger data set, no support for their claims of a major role of genetic diversity in economic development is evident. We discuss the characteristics that distinguish between this case in which no role for genetic diversity is observed and the three examples in which genetic diversity is seen to be important.

doi: 10.1534/genetics.115.176750
Manuscript received March 25, 2015; accepted for publication July 21, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176750/-/DC1
[1]Corresponding author: Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305-5020. E-mail: noahr@stanford.edu

Genetic Diversity in Humans

Measuring genetic diversity

We first clarify that the concept of genetic diversity of interest in the examples we consider is the diversity of genetic types observed among members of a population—and not the diversity in a collection of populations that is contributed by

differences across the constituent groups. This concept of the genetic diversity of a population is computed from data on that population alone, and it is unaffected by the composition of the larger data set of populations used for its calculation.

Within-population genetic diversity is most commonly measured by expected heterozygosity, the probability that two draws from a population at a specific site in the genome will produce different genetic types. Formally, suppose that a genetic locus l has K_l distinct alleles, with nonnegative frequencies $p_{l1}, p_{l2}, \ldots, p_{lK_l}$ such that $\sum_{i=1}^{K_l} p_{li} = 1$. The expected heterozygosity of locus l is defined by

$$H_l = 1 - \sum_{i=1}^{K_l} p_{li}^2. \tag{1}$$

For a collection of L loci, the mean expected heterozygosity across loci is

$$H = \frac{1}{L} \sum_{l=1}^{L} \left(1 - \sum_{i=1}^{K_l} p_{li}^2 \right). \tag{2}$$

When the allele frequencies p_{li} are estimated from data rather than treated as known parameters, a correction is introduced in order to obtain statistically unbiased estimators. Suppose that a sample of n_l observations is collected at locus l and that n_{li} of them have type i, with $\sum_{i=1}^{K_l} n_{li} = n_l$. Then $\hat{p}_{li} = n_{li}/n_l$ estimates the frequency of allele i, and expected heterozygosity is estimated by (Nei 1987)

$$\hat{H}_l = \frac{n_l}{n_l - 1} \left(1 - \sum_{i=1}^{K_l} \hat{p}_{li}^2 \right). \tag{3}$$

$$\hat{H} = \frac{1}{L} \sum_{l=1}^{L} \frac{n_l}{n_l - 1} \left(1 - \sum_{i=1}^{K_l} \hat{p}_{li}^2 \right). \tag{4}$$

Expected heterozygosity is a sensible diversity measure, with larger values indicating greater diversity (Figure 1). It has a natural interpretation in diploid organisms, measuring the probability that the two copies of a locus in an individual, treated as independent and identically distributed draws from a population with a specified allele frequency distribution, are distinct.

Origins of human genetic diversity

Surveys of genetic diversity in indigenous human populations worldwide have documented considerable variation in the level of heterozygosity present within a population (Bowcock *et al.* 1994; Rosenberg *et al.* 2002; Prugnolle *et al.* 2005; Ramachandran *et al.* 2005). These differences in heterozygosity follow a geographic pattern, with a systematic linear decline occurring as a function of increasing distance from East Africa, measured over land-based routes. The highest heterozygosities appear in populations from Africa, followed by populations from the Middle East, Europe, and Central and South Asia. Populations of East Asia have still lower

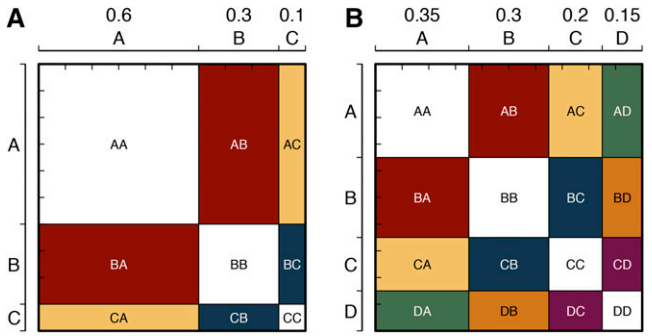

Figure 1 Expected heterozygosity as a measurement of diversity. Each axis in the unit square represents an allele frequency distribution, with each area representing the probability that an individual has a particular ordered pair of alleles. The shaded regions represent heterozygous combinations. The two loci shown represent different expected heterozygosity levels (equation 1). (A) A smaller heterozygosity (0.540). (B) A larger heterozygosity (0.725).

heterozygosities, and Pacific Islander and Native American populations, at the greatest geographic distance from Africa over migration paths traversed in human evolution, are the least heterozygous. The linear decrease in heterozygosity with increasing distance from Africa is a strong and replicable relationship, achieving correlation coefficients near -0.9 in a variety of studies of different genetic markers and sets of populations (Prugnolle *et al.* 2005; Ramachandran *et al.* 2005; Conrad *et al.* 2006; Jakobsson *et al.* 2008; Li *et al.* 2008; Pemberton *et al.* 2013).

Population-genetic models have explained the pattern of variation in human genetic diversity, with a decrease in heterozygosity at a greater distance from Africa, in relation to the relatively recent history of human migrations starting from an African origin. Under these models, during a geographic expansion, new regions are occupied not by expansion in the range of an existing population in its entirety but instead by a recursive procedure of new settlement formation by subgroups that separate from their parental colonies (Ramachandran *et al.* 2005; Liu *et al.* 2006; DeGiorgio *et al.* 2009, 2011; Deshpande *et al.* 2009). Each founding group carries only a subset of the total genetic diversity of its parental population, leading to a loss of genetic diversity in the new group (Figure 2). Newly established populations then generate their own subgroups that again separate to found new populations, and the founding process is repeated anew. In this model of serial founder effects, each founding event produces a loss of genetic diversity, so the populations at the greatest distance from the starting point possess the lowest heterozygosity.

Although other processes, such as admixture between populations and changes in population size, also affect genetic diversity patterns (DeGiorgio *et al.* 2009; Pickrell and Reich 2014), the general utility of serial founder models in human evolution as a first approximation for explaining the global pattern of genetic diversity is supported both by the strength of the correlation between heterozygosity and distance from

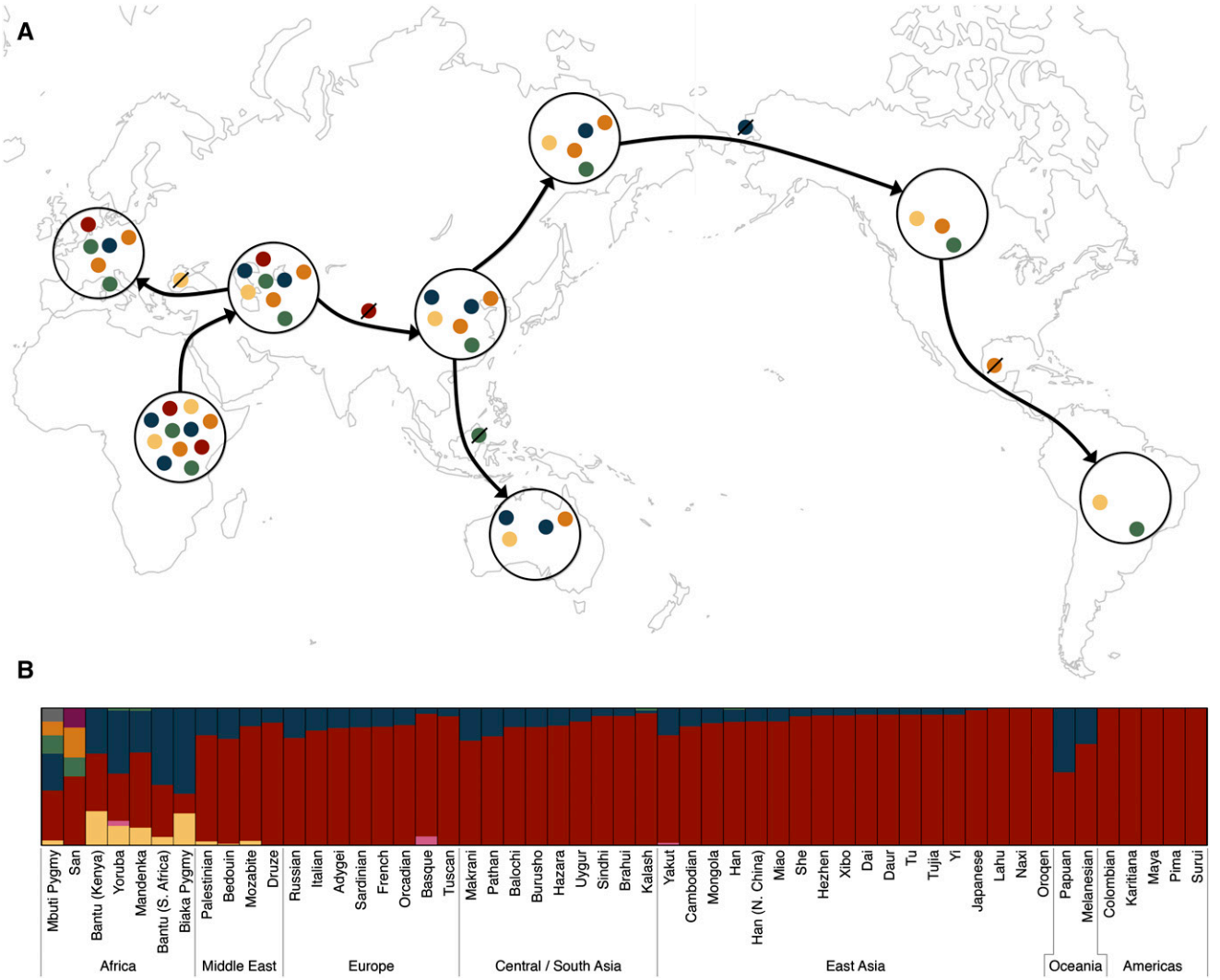

Figure 2 The serial founder model in human evolution. (A) A schematic of the model. Each color represents a distinct allele. Migration events outward from Africa tend to carry with them only a subset of the genetic diversity from the source population, and some alleles are lost during migration events. (B) An example of the model at a particular genetic locus, *TGA012*. Each set of vertical bars depicts the allele frequencies in a population, with different colors representing distinct alleles. Within continental regions, populations are plotted from left to right in decreasing order of expected heterozygosity at the locus [equation (3)]. This figure illustrates the loss of alleles across geographic regions; Native Americans all possess the same allele. The allele frequencies are taken from Rosenberg *et al.* (2005).

Africa and by an observation that within large geographic regions, source regions more accessible to colonizing populations along likely migration routes also have greater heterozygosity than more distant regions—for example, southern Europe compared to northern Europe (Lao *et al.* 2008), coastal Melanesia compared to inland Melanesia (Friedlaender *et al.* 2008), and northwest South America compared to the Amazon region (Wang *et al.* 2007). Further, serial founder models explain patterns in statistics of genetic differentiation and allelic correlation along the genome that other, substantially different models cannot explain (DeGiorgio *et al.* 2009, 2011).

We can therefore observe that population-genetic models of the spread of human populations explain the variation across human populations in levels of genetic diversity and that this variation is informative about the particular history of human migrations. We now turn to examining the effects of these genetic diversity differences on a variety of societally important scenarios.

Examples: Interactions of Genetic Diversity and Social Factors

Familial identification in forensic genetic testing

A comparatively new form of forensic DNA testing uses crime-scene samples to identify unknown suspects through genetic relatedness profiling (Bieber *et al.* 2006; Butler 2011; Gershaw *et al.* 2011). When no perfect DNA match of a crime-scene sample to an entrant in a database of potential suspects is found, investigators can test for a partial match to assess

whether the crime-scene sample might be from a genetic relative of an entrant in the database. A positive test leads investigators to consider as potential suspects genetic relatives of the person with the partial match.

The identification of relatives through partial matches raises new statistical and population-genetic issues largely absent in the standard setting of forensic profiling via exact matches. In the basic forensic scenario, a crime-scene sample is tested at a number of DNA markers that is small, but large enough that a false-positive match of a genetically unrelated noncontributor to the crime-scene sample at all the loci is extremely unlikely. Forensic marker systems are designed so that the false-positive probability is acceptably low for use as evidence in court irrespective of the actual alleles found in the genetic profile.

In familial identification, the false-positive rate is substantially higher because familial identification generally must rely on databases and marker sets designed for the simpler exact-match problem. In the absence of genotyping error, across a DNA marker set, the sibling of the actual contributor of a DNA sample, for instance, will match the crime-scene sample for a substantial fraction of the alleles. A sibling is expected to share both alleles identically by descent at a quarter of all loci—inheriting the same pair of alleles from shared parents—and one allele identically by descent at half the loci. Thus, the fraction of alleles shared with the sibling is one-half or more: on average, half the alleles are shared identically by descent, and an extra contribution arises from the chance that alleles not shared identically by descent have the same state nonetheless. The sibling, however, is not expected to match all alleles with the crime-scene sample; on average, at a quarter of loci, siblings share neither allele identically by descent, and these generally will not have the same allelic type. A partial match of the DNA profile is therefore expected for the sibling, with different loci having no, one, or two matching alleles. Thus, for a fixed marker set, because a true genetic relative of the contributor has only a partial match with the crime-scene sample, the chance of a false-positive match—the probability that a nonrelative also achieves the less stringent partial-match threshold—greatly exceeds the probability that the same nonrelative is a false exact match.

The underlying genetic diversity in a population affects the probability that a nonrelative of the DNA contributor produces a false-positive partial match close enough to appear to be a relative of the contributor of the crime-scene sample. Consider a hypothetical low-diversity population in which all members are homozygous at some locus for the same allele, implying a complete absence of genetic diversity. Suppose also that the contributor to the crime-scene sample has this same homozygous genotype. The locus contains no identifying information, and every individual in the population has an exact match at the locus—both genetic relatives and nonrelatives of the contributor.

Now consider a hypothetical high-diversity population with many rare alleles, so that individuals tend to have more distinctive genotypes at the locus. In this population, the contributor's homozygous genotype would be rare rather than common, and most of the individuals who possess a partial or exact match to the contributor at the locus could only do so if they had obtained the same alleles through shared familial lines of descent. The locus is highly informative for individual identification in this population, and it is primarily the genetic relatives who have a partial match.

These two extreme scenarios convey the idea that different levels of genetic diversity confer varying degrees of individual identifiability. The more genetically homogeneous a population is, the less identifying information an ostensible genetic match provides, and the conclusion that a partial match indicates a direct familial genetic relationship is more likely to represent a false-positive result (Figure 3).

The connection between genetic diversity and the distinguishability of genetic relatives and unrelated individuals was demonstrated by Rohlfs *et al.* (2012, 2013), who studied the effect of differences in genetic diversity on the extent to which relatives and nonrelatives can be distinguished in a familial identification context, employing the 13 Combined DNA Index System (CODIS) forensic identification loci widely used as the standard marker panel in forensic testing. Relying on allele-frequency distributions at the CODIS loci for each of five population samples that represent different levels of genetic diversity—in decreasing order, largely following the serial founder model (Figure 2), African American, Latino, European American, Vietnamese, and Navajo—the authors simulated unrelated pairs of individuals and sibling pairs, generating sibling pairs by transmitting alleles through pedigrees with shared unrelated parents. Next, they computed a likelihood ratio to quantify whether a partial genetic match between two individuals is more likely under the hypothesis of a familial relationship—a sibling relationship in this case—or under the hypothesis of a chance partial match of unrelated individuals. For each population, Rohlfs *et al.* (2012) measured a "distinguishability statistic" based on the likelihood-ratio distributions for the simulated unrelated pairs and sibling pairs, arguing that significant overlap between the two distributions indicates reduced potential to distinguish between siblings and unrelated pairs—and more chance partial matches of nonrelatives—as more pairs are assigned likelihood ratios compatible with either category.

Computing the distinguishability statistic in each of the five populations, Rohlfs *et al.* (2012) found a strong relationship across populations between population-level heterozygosity and the distinguishability measure (squared correlation of 0.95), confirming that a higher false-positive rate occurs in low-diversity populations. The authors then took their result one step further. In performing the likelihood-ratio computation, a distribution of allele frequencies must be specified for the population to which the crime-scene sample belongs. In practical conditions, this population might be unknown, so the allele frequencies used in the likelihood-ratio computation might be misspecified. The authors examined the distinguishability statistic in the context of misspecified allele frequencies, for each choice of population using each of the

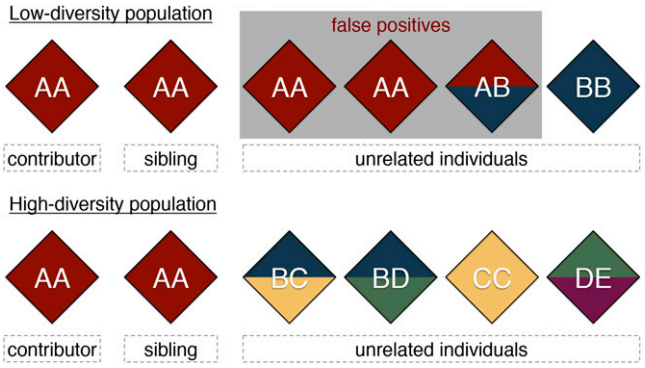

Figure 3 Familial identification in forensic testing. A contributor to a crime scene DNA sample has genotype *AA* at a locus. A sibling of the contributor is likely to share more alleles with the contributor than are unrelated individuals; the probability of an exact match at a locus, as shown, exceeds 25% for a sibling. This figure illustrates that in a low-diversity population, the chance of a false-positive match of an unrelated individual to a crime-scene contributor at a locus is greater than in a high-diversity population. In the low-diversity population, two nonrelatives have exact matches, and one has a partial match, whereas in the high-diversity population, the nonrelatives do not have exact or partial matches.

four other populations to misspecify the allele frequencies. In simulations with the misspecified allele frequencies, they found that distinguishability was lower than in the case in which allele frequencies were properly specified, particularly when individuals from a less genetically diverse population were erroneously assumed to belong to a more genetically diverse population.

The analysis of Rohlfs *et al.* (2012) and subsequent work extending beyond sibling relationships (Rohlfs *et al.* 2013) illustrate that in familial identification based on a fixed marker system shared across all groups, populations with lower genetic diversity are likely to have higher false-positive match rates: the genetic diversity of the population has a direct impact on the familial identification setting. The results further suggest that similar issues are relevant to other forensic problems involving partial matches, such as when the crime-scene sample represents a DNA mixture from many individuals rather than a single person, potentially with degraded DNA, missing genotypes, or genotyping errors (Balding 2013; Steele and Balding 2014). In this context, which relies on partial matches to determine whether a test individual might be included in the mixture, the false-positive probability that a noncontributor is erroneously regarded as a contributor is likely to depend on genetic diversity in a parallel manner to the familial identification setting. As in familial identification, a chance partial match of a mixed crime-scene sample with a random individual has greater probability in a low-diversity population, so the probability of a false-positive partial match is likely to exceed the corresponding probability in a more diverse population.

The demonstration of variability across populations in false-positive match rates is purely population-genetic, using population-genetic theory to evaluate the influence of genetic relatedness and population allele frequencies on the probability of DNA matches, but the result exists in a context in which substantial differences exist across groups in the probability that an individual appears in forensic databases. In the United States, representation of an individual in a database is related to the past experience of the individual with criminal investigation, a factor that varies across populations. The probability that an individual has a close genetic relative in a forensic database—and therefore has a DNA profile accessible to investigators through familial identification—then also varies by population. The potential inequalities that could arise from this variation have been much discussed (Greely *et al.* 2006; Garrison *et al.* 2013). The analysis of Rohlfs *et al.* (2012), however, indicates that outcomes of familial identification analyses depend not only on inequality across groups in representation in criminal investigations, which affects the chance that a genetic relative of the DNA contributor is in the database, but also on the difference across groups in genetic diversity, which affects the distinguishability of DNA profiles from genetic relatives and unrelated individuals. Genetic diversity and its interaction with variation across populations in representation in the justice system are therefore both essential to determining and improving the utility and fairness of a familial identification test.

Bone marrow transplantation matching

Genetic diversity has a quite different impact in another area that also relies on match probabilities: transplantation matching. In medical transplantation, an immunologic match between a recipient and donor reduces the risk that the recipient immune system will recognize the donor cells as foreign and therefore produce an undesirable immune response. The problem is particularly salient in bone marrow transplantation, which involves a transplant of donor cells from the immune system itself—cells that can recognize the recipient as foreign.

In bone marrow transplantation, the degree of matching is assessed using multilocus genotypes at a set of protein variants encoded by the genes of the human leukocyte antigen (HLA) system on chromosome 6. The HLA system contains six highly polymorphic loci whose alleles determine the core of an individual HLA multilocus genotype and that are generally matched for bone marrow transplantation. The number of alleles at highly polymorphic HLA loci can run into the thousands, and as of 2014, the database of HLA alleles (Robinson *et al.* 2013) records more than ~12,000 distinct alleles across the six major genes. An already large number of potential types at each locus increases to tens or hundreds of millions as multilocus types are considered to ensure a lower chance for rejection.

Owing to the fact that HLA alleles are codominantly expressed, the aim in transplantation matching is to match as many alleles as possible between donor and recipient. Close genetic relatives of a potential recipient have the greatest match probability because a recipient and a relative share a substantial fraction of their genomes identically by descent.

Given the high genetic diversity that exists in the HLA system, however, the probability is low that two unrelated individuals would match perfectly at important loci.

Unlike the setting of forensic familial identification, in which low genetic diversity generates problematic false-positive matches, the challenge of transplantation matching is exacerbated by *high* diversity in the population to which the recipient belongs: a population with high levels of HLA genetic diversity will possess a large number of HLA multilocus combinations. Each unique combination then appears at a lower frequency, and the probability that any given pair of individuals has an exact HLA match is reduced. Conversely, a population that is more homogeneous at the HLA loci has fewer unique combinations, and the chance of a match for a random pair of individuals is increased.

In the United States, the National Marrow Donor Program (NMDP) registry contains HLA profiles of millions of potential donors, each of whom can be queried if a matching recipient is in need of a donation. In parallel with the analysis of forensic profiles by Rohlfs *et al.* (2012), analyses of NMDP profiles identify an effect of population differences in genetic diversity on match probabilities for HLA (Cao *et al.* 2001). They also illustrate the interaction between genetic diversity and social phenomena in influencing the probability that a match exists for potential recipients from each of a series of populations.

An investigation by Bergstrom *et al.* (2009) highlights the key issues. Using NMDP three-locus genotype frequencies reported by Mori *et al.* (1997), Bergstrom *et al.* (2009) computed the theoretical match probabilities for pairs of HLA profiles drawn from each of five populations. High-diversity African Americans, with their high fraction of African ancestry at the source of the serial founder expansion (Figure 2), had the lowest theoretical match probability, a value substantially lower than in the other groups. In increasing order, the match probabilities were greater in the Hispanic and Asian-American populations and greatest in white and Native American groups. Although the groups studied by Bergstrom *et al.* (2009) and Rohlfs *et al.* (2012) do not align exactly (nor do they align with the indigenous groups in the global characterization of genetic diversity in Figure 2), the pattern of decreasing transplantation match probabilities largely reverses the sequence describing increasing numbers of false-positive matches in familial identification. Similar general patterns are observed when considering theoretical match probabilities between recipient HLA profiles chosen from one population and donor profiles chosen from another.

As in the analysis of Rohlfs *et al.* (2012), the theoretical computation of HLA match probabilities from transplantation database frequencies is a calculation under a model constructed from population-genetic principles showing that from population-genetic considerations alone, higher-diversity populations are expected to have lower transplantation match probabilities. Also similar to the analysis of Rohlfs *et al.* (2012), for the practical setting, population-genetic match probabilities appear in a context of population differences in the frequencies with which individuals are represented in transplantation databases. Bergstrom *et al.* (2009, 2012) comment on a number of factors that vary across populations: the overall size of the population, the rate at which members of a population choose to contribute profiles to the database, and the rate at which potential donors participate in a transplantation when queried. Incorporating these factors, including the nontrivial role played by the difficulty of characterizing HLA variation in minority populations with smaller sizes, the chance that no donor match is found is greatest for African Americans, followed by the Asian-American, Hispanic, Native American, and white groups. As in the forensic case, the population genetics of genetic diversity, together with societal factors that vary across populations, contributes to the quantity of ultimate interest. Both genetic diversity and its interaction with factors that affect participation in transplantation are important in increasing the probability that any given recipient can find a successful match.

Genome-wide association studies

A third area of influence for differences in genetic diversity is in representation in the development and application of research resources in human genomics. Genome-wide association studies (GWA studies) are genomic investigations of the statistical correlations between genetic variation among members of a population and an observed phenotype (Hirschhorn and Daly 2005; McCarthy *et al.* 2008; Stranger *et al.* 2011). In humans, these studies, which seek to uncover genetic factors that underlie a phenotype, typically compare the genotypes of two groups of individuals—cases, who have a disease, and control individuals, who do not. An allele found more frequently in cases than in controls is said to be associated with the disease. In recent years, GWA studies have proliferated rapidly, producing thousands of successes in the identification of disease-associated loci (Hardy and Singleton 2009; Hindorff *et al.* 2009).

GWA studies rely on linkage disequilibrium (LD), the association between allelic states at different loci along a chromosome: association with a disease occurs not only for a risk mutation but also for other alleles located proximate on the genome to the susceptibility allele (Figure 4). Thus, nearby alleles in the ancestor in whom a disease mutation originally occurred are transmitted to diseased descendants with greater probability than are distant alleles. At the same time, recombination breaks down the correlation—the LD—between the disease allele and distant alleles. As a result, among diseased descendants, alleles that remain associated with the disease allele—and, consequently, with the disease phenotype—are likely to lie close on the genome to the disease mutation. The premise that LD enables variants to "tag" their neighbors (Carlson *et al.* 2004; International HapMap Consortium 2005) and, hence, to facilitate the discovery of disease variants through the separate associations of a disease variant with both the disease phenotype and a tag-SNP proxy, made it possible to begin performing GWA studies without requiring full genome sequences in every sampled individual.

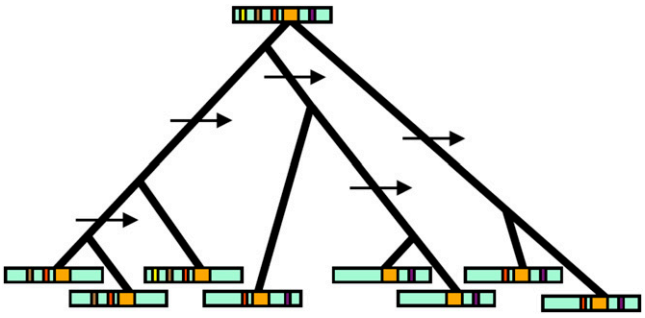

Figure 4 The principle of linkage disequilibrium that underlies genome-wide association studies. This figure depicts a series of individuals with a disease, tracing the genealogy of the section of the chromosome on which a disease-causing allele is located. A disease allele (orange) occurs on an ancestral chromosome containing several marker alleles (yellow, brown, red, and purple). Recombination events (arrows) break down correlations between the disease mutation and marker alleles, so the closer a marker allele is to the mutation, the more likely it is to be found in present-day disease cases.

LD varies across human populations, however, with populations that possess a greater diversity of haplotypes having a lower probability that a genotype at one site on the genome will be informative about other nearby sites. In parallel with the decrease in genetic diversity with increasing distance from Africa, a decrease in haplotype diversity and a concomitant increase in LD exist with increasing distance from Africa (Conrad *et al.* 2006; Jakobsson *et al.* 2008; Li *et al.* 2008). This pattern has had the consequence that sets of tag SNPs used in early GWA studies are less successful in tagging genetic variants in low-LD African populations at the source of the serial founder human expansion (Figure 2), generating reduced potential for uncovering disease associations in these groups (Conrad *et al.* 2006; Debakker *et al.* 2006). The problem has persisted as tag-SNP approaches have been replaced with genotype imputation studies, which rely on LD to impute unmeasured genotypes that can be tested for disease association in a similar manner to genotypes that have actually been measured: low-LD African populations generate lower imputation accuracy (Huang *et al.* 2009, 2011).

The heterogeneity in genetic diversity, as reflected in the low LD for African populations and its consequences in generating relatively low tag-SNP "portability" and genotype imputation accuracy, has rendered African genomes comparatively less well suited to GWA studies based on LD. Partly as a result of this phenomenon, GWA studies have been implemented unevenly across human populations, generating concerns that the benefits of human genetics research will not accrue equally in different groups (Need and Goldstein 2009; Rosenberg *et al.* 2010; Bustamante *et al.* 2011). Most GWA studies have focused on populations of European ancestry, and other populations have been underrepresented, quite dramatically in some cases.

Differences in genetic diversity that have influenced GWA studies have interacted with sociological factors in the scientific community that have also prioritized the use of European samples (Rosenberg *et al.* 2010; Teo *et al.* 2010). Because GWA studies are expensive, early studies focused on a small number of populations for which shared sets of genomic resources—standardized marker panels, shared controls, and shared databases of densely genotyped samples with deep characterization of genetic variation—could be generated. Well-developed networks of investigators in countries of Europe and North America with the resources to conduct GWA studies generally had easiest access to patient populations of European descent, further contributing to an emphasis on these populations in early studies. Though the incorporation of non-European populations has increased, initial inequalities across populations in GWA representation have persisted because subsequent investigations continue to build on patient populations, funded projects, and researcher networks from earlier studies (Burchard 2014).

This interaction of a form of genetic diversity and societal variables in the structure of the scientific research enterprise has led to a situation in which one estimate recorded 96% of GWA subjects as having European ancestry (Bustamante *et al.* 2011). Though this disparity has a basis partly in variation in access to populations generated by the structure of scientific collaboration networks and the distribution of research funding, it has been exacerbated by considerations of genetic diversity; indeed, a feedback loop exists between differences in societal variables and genetic diversity phenomena because initial differences among populations in practical and technical feasibility have contributed to overemphasis on European populations in developing technical capabilities, making further European overrepresentation enticing to researchers and funding panels. In parallel with the variation across populations observed in the familial identification and bone marrow transplantation scenarios, a consequential practical difference across populations in representation in genomic studies arises from the interaction of genetic diversity with social factors.

An Effect of Genetic Diversity on Economic Development?

We have described three examples that each involve an interaction of differences in genetic diversity with population differences in society to produce a difference in an important phenomenon—false-positive matches in forensic genetics, match probabilities in transplantation, and research efforts in GWA studies. Each of these settings involves a problem that is fundamentally biological—DNA-based identification, transplantation, and genetics of disease. In each setting, principles from population-genetic theory in which aspects of genetic diversity feature prominently underlie the contribution of genetic diversity: theories of forensic and transplantation matching explicitly produce an inverse relationship between match probabilities and genetic diversity, and GWA statistics rely on models of the decay of genetic diversity and production of LD during human migrations. When genetic diversity appears as a variable in a context in which no similar theory exists, in which theoretical constructs are

drawn from outside population genetics, is genetic diversity similarly important? How far do implications of genetic diversity extend for societal phenomena, in scenarios that *a priori* have no evident connection to biology? We have reexamined the study of Ashraf and Galor (2013) to evaluate their hypothesis that genetic diversity is a key determinant of economic development.

Economic development

Ashraf and Galor (2013) advanced the claim that genetic diversity levels have had a persistent long-term effect on comparative economic development. They argued that genetic diversity at the high and low extremes—characteristic of African and Native American populations, respectively—has been detrimental for development, whereas the intermediate genetic diversity of European and Asian populations has, however, facilitated development. In other words, economic development has a "hump-shaped" negative quadratic relationship with genetic diversity.

To argue for their hypothesis, Ashraf and Galor (2013) relied on short-tandem-repeat genetic markers genotyped in 53 worldwide populations from the Human Genome Diversity Panel (Ramachandran *et al.* 2005; Rosenberg *et al.* 2005), using expected heterozygosities previously reported for each population according to equation (4). They adopted the distances of Ramachandran *et al.* (2005) of each population from East Africa, taking into account geographic waypoints to approximate migratory paths of human populations outward from Africa.

In their economic analysis, the 53 populations were grouped into 21 present-day countries based on geographic coordinates. For each of these countries, an "observed diversity" was computed as the mean of the expected heterozygosities of populations sampled within the country. Several ordinary-least-squares regressions then were performed using as the dependent variable the natural logarithm of population density in 1500 CE, treated as a proxy for economic development, and as the independent variables the observed diversity, its square, and control variables relating to geography and to the local timing of the Neolithic transition. The regressions produced a generally significant quadratic relationship between the dependent variable and observed diversity, even after conditioning on various control variables (Table 1 and Supporting Information, Table S1). The authors used these results to claim that economic development has a statistically significant "hump-shaped" dependence on observed diversity.

Next, Ashraf and Galor (2013) extended their analysis to a worldwide sample of 145 countries. For most of the countries, however, information on expected heterozygosity was unavailable. In place of actual data on expected heterozygosity for most of these countries, the authors used as the observed diversity the *predicted* expected heterozygosity from the linear regression of expected heterozygosity in 53 populations with migratory distance from East Africa. They justified this choice on the grounds that expected heterozygosity has a strong relationship with distance from East Africa, enabling heterozygosity predictions for unsampled populations and, because the 21-country analysis produced a significant economic effect for diversity, suggesting the plausibility of using an estimated value of this quantity in place of actual genetic data. It was then possible in the absence of genetic data on additional countries to enable incorporation of economic variables on those countries.

To calculate predicted diversity, for each of the 145 countries, the migratory distance from East Africa of its capital city was substituted into the regression of expected heterozygosity on migratory distance. Using predicted diversity in regressions of the economic development variable similar to those performed with observed diversity, a broadly significant quadratic relationship between economic development and predicted diversity was observed, even after controlling for other variables. On the basis of this analysis, Ashraf and Galor (2013) claimed that a "hump-shaped" effect of genetic diversity on economic development from the 53-population data set was a general worldwide phenomenon.

The reanalysis

We sought to examine the argument of Ashraf and Galor (2013) on its own terms using their assumptions, methods, economic variables, and regression models—all contested elsewhere (d'Alpoim Guedes *et al.* 2013; Gelman 2013; Feldman 2014)—and changing the analysis only by expanding the genetic data. In particular, we revisited their analysis with a recently assembled data set that largely subsumes the earlier 53-population data set. These data consisted of 237 populations studied by Pemberton *et al.* (2013), excluding from a larger set of 267 the populations with unknown or ambiguous geographic assignments, populations with sample size ≤ 5, and populations from Micronesia and Samoa, for which Ashraf and Galor (2013) did not provide values of the economic variables. The 237 populations represent 39 countries. From Pemberton *et al.* (2013), we used the expected heterozygosities reported using 645 loci in the full 5795-individual data set; the calculation is analogous to the earlier expected heterozygosities computed with 783 loci and 1048 individuals (Ramachandran *et al.* 2005).

Repeating the regressions of the economic development variable on "observed diversity"—with the only difference from Ashraf and Galor (2013) being use of the data on 237 populations in 39 countries instead of 53 populations in 21 countries—we observe that the quadratic relationship is no longer close to statistically significant (Table 1). The magnitude of the effects for observed diversity and its square are much reduced, and none of the regressions involving either variable generates significance (Table S2). This result suggests that the "hump-shaped" effect of observed diversity was limited by the particular set of countries and populations covered by the earlier available data: with an expansion of the number of countries, the observed diversity variable fails to produce an effect.

We further investigated this claim in two ways. First, we subsampled only the 136 populations studied by Pemberton

Table 1 ***P*-values for multiple regressions of "log population density in 1500 CE" on "observed diversity" and "observed diversity squared"**

	Regression 1: genetic variables only	Regression 4: genetic variables and nongenetic covariates	Regression 5: genetic variables, nongenetic covariates, and continent fixed effects
53 populations in 21 countries: same countries and populations as Ashraf and Galor (2013)			
Observed diversity	0.000483***	0.00856***	0.0609*
Observed diversity squared	0.000634***	0.0124**	0.0973*
136 populations in 21 countries: same countries as Ashraf and Galor (2013), more populations			
Observed diversity	0.000233***	0.00916***	0.101
Observed diversity squared	0.000297***	0.0122**	0.147
237 populations in 39 countries: more countries, more populations			
Observed diversity	0.515	0.145	0.642
Observed diversity squared	0.639	0.266	0.719

The nongenetic covariates are "log Neolithic transition timing," "log percentage of arable land," "log absolute latitude," and "log land suitability for agriculture." Each variable was computed and employed as in Ashraf and Galor (2013) using their regression models and the values they reported for nongenetic variables. Regression models 1, 4, and 5 are the three models of Ashraf and Galor (2013) that use genetic data. The analysis of 53 populations in 21 countries recomputes the same analysis as in Table 1 of Ashraf and Galor (2013) using scripts they provided. Significance at the 10, 5, and 1% levels is represented by *, **, and ***, respectively. Full regression tables appear in Table S1, Table S2, and Table S3.

et al. (2013) from countries with geographic coordinates that placed them in the earlier set of 21 countries, repeating the same regressions, again computing observed diversity for a country by averaging values for its constituent populations. This analysis represents an expansion of the data used by Ashraf and Galor (2013), but examining only the same 21 countries they considered. For some models, as in the analysis of Ashraf and Galor (2013), the 21-country analysis continues to produce a significant effect for observed diversity and its square when more populations are considered (Table 1 and Table S3).

Next, to assess whether the significant result for observed diversity from the 21-country subset was anomalous among possible choices of countries, we considered alternative 21-country subsamples of the Pemberton *et al.* (2013) data set, repeating the regressions for each subsample. We randomly selected 1000 subsamples of 21 countries among the 39 countries available, maintaining the continental distribution in the original 21-country data set (eight in Africa, four in Europe, three in the Americas, and six in Asia and Oceania). Note that because the Pemberton *et al.* (2013) data set does not cover any countries in Europe beyond the earlier genetic data, all 1000 subsets use the same four European countries (France, Italy, Russia, and the United Kingdom). If a regression represents a true effect of observed diversity on the dependent variable irrespective of the subsample of countries, then we would expect a low *P*-value in most of the 1000 replicates. If, however, the 21-country subsample of Ashraf and Galor (2013) is an anomalous false-positive result, then we expect relatively few replicates to produce small *P*-values, with a uniform distribution of *P*-values across replicates occurring under the null hypothesis that observed diversity has no effect.

For each regression, *P*-values for the observed diversity and observed diversity squared variables in the 1000 subsamples appear in Figure 5. For both variables, across regressions, only at most ∼27% of subsamples produce a significant effect at the 5% level. For the most complete regression, regression 5, accounting for nongenetic covariates and continent fixed effects, the *P*-value distributions are nearly uniform, with 4–5% of replicates producing $P < 0.05$. Thus, had a different set of 21 countries been used by Ashraf and Galor (2013), effects for observed diversity and its square are unlikely to have been seen.

This analysis thus finds that the 21-country set of Ashraf and Galor (2013) is unlike both the enlarged 39-country data and most other 21-country data sets in producing significant effects for genetic diversity. Recalling that Ashraf and Galor (2013) used significance in the 21-country analysis as a basis for replacing actual values of genetic diversity with "predicted diversity," our reanalysis both contests the result of the first component of the genetic diversity analysis of Ashraf and Galor (2013) and undermines the rationale for the second component. We can conclude that even if the suitability of the methods and data of Ashraf and Galor (2013) to studying the effect of genetic diversity on economic development is left unquestioned, the "hump-shaped" effect for genetic diversity does not persist with an expanded genetic data set.

Perspective

The variability of genetic diversity across different human populations, a vestige of the history of human migrations, is consequential for population differences in a variety of settings of societal interest. These differences across populations, each of which might be viewed through a purely social-scientific lens, involve a population-genetic contribution of the properties of genetic diversity.

In the familial identification, transplantation matching, and GWA representation examples that we have examined, addressing inequalities across populations in the phenomena of ultimate interest requires a particular effort to overcome not only the sociological determinants of inequality across populations but also the intrinsic inequalities that arise from differences in genetic diversity. Thus, Bergstrom *et al.* (2009) study the relative value of efforts to enhance

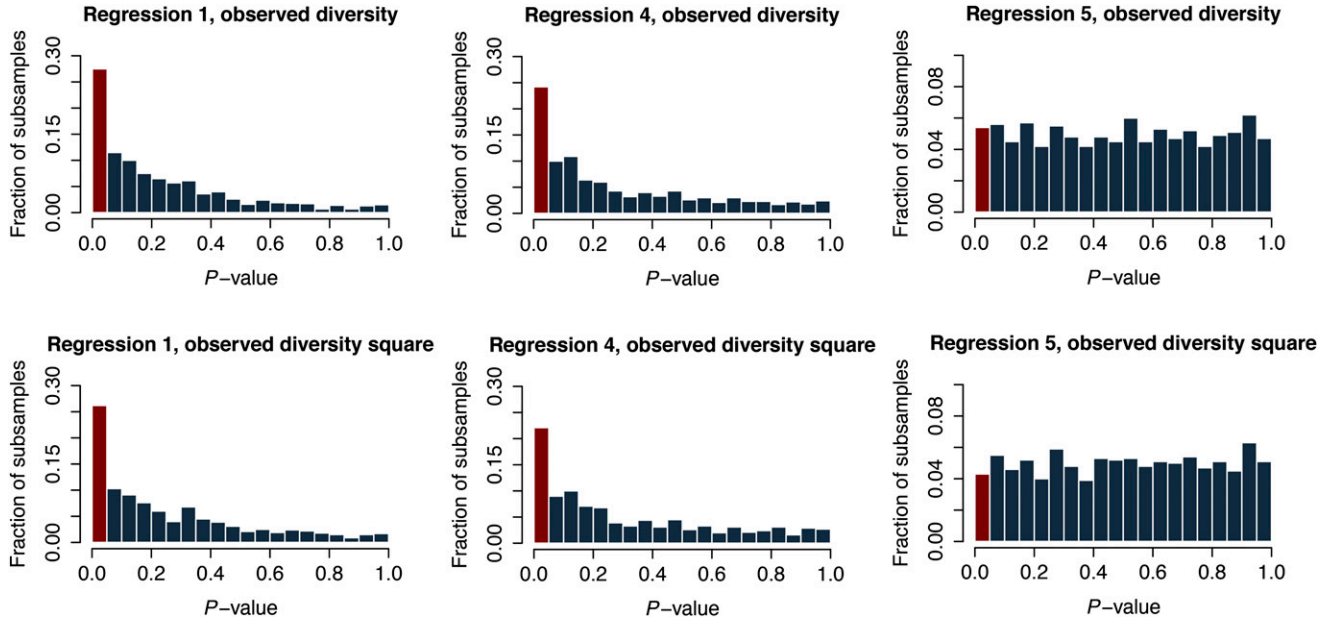

Figure 5 The distribution across 1000 replicate subsamples of regression *P*-values for the influence of observed genetic diversity and its square on a proxy for economic development. Each panel represents a regression model (regressions 1, 4, or 5, as in Table 1 and Table S1, Table S2, and Table S3) and a variable whose significance is tested (observed diversity or its square). Each replicate subsample considers 21 countries. The red bar indicates the fraction of subsamples for which the *P*-value is smaller than 0.05.

representation of high-diversity African Americans not only for the goal of achieving equality of representation but also because each African American added can have a greater chance than other individuals of providing the only database match for a potential transplantation recipient. For GWA studies, not only is reduction of inequality in participation a desirable goal achievable via mechanisms such as funding priorities emphasizing underrepresented groups, improvement in the ultimate outcome—genetic understanding of disease for all populations—can be achieved by generating new genomic resources for additional populations (International HapMap 3 Consortium 2010), developing statistics for enhancing GWA designs and analyses in underrepresented and high-diversity populations (Teo *et al.* 2010; Huang *et al.* 2011), and employing studies that capitalize on unique features of genetically admixed groups (Winkler *et al.* 2010; Seldin *et al.* 2011). In forensic familial identification, improvements toward the goal of equally minimal false-positive matches in forensic casework can be achieved by dissemination in the legal system of knowledge about false-positive matches and the role of genetic diversity, transparency in applications of familial identification methods, and development of new marker sets with lower error rates (Rohlfs *et al.* 2012, 2013; Garrison *et al.* 2013).

Each of these settings can be viewed from an economic perspective: cost differences across populations can arise from the differential pursuit by law enforcement agencies of false-positive forensic identifications, the variable rates of success or failure to find transplantation matches, and the potential inequalities in the success of treatments arising from genomic medicine. Indeed, Bergstrom *et al.* (2009, 2012) adopted an explicitly economic perspective in analyzing improvements in transplantation matching, estimating a cost and benefit for each additional registrant added to the NMDP database.

Nevertheless, despite this view that economic consequences can be traced to variation in genetic diversity, we have found no support for the claim of Ashraf and Galor (2013) that genetic diversity has been important in contributing to differences across human populations in levels of economic development. Our reanalysis has focused exclusively on the genetic data in their study, not repeating objections raised elsewhere about their demographic and economic data, statistics, and interpretations, or about the suitability of their data and genetic variables to addressing the question at hand (d'Alpoim Guedes *et al.* 2013; Gelman 2013; Feldman 2014). Whereas genetic diversity affects differences among human populations in other scenarios, reproducing the work of Ashraf and Galor (2013) on its own terms using expanded genetic data challenges the claim for a role of genetic diversity in economic development.

What distinguishes the forensic, transplantation, and GWA scenarios in which genetic diversity has a demonstrable impact from the economic development problem? The former scenarios are each tightly connected to biological phenomena. For these cases, computations from population genetics prominently feature genetic diversity; in fact, it can be argued that population genetics suggests that proper analysis of population differences in these scenarios is incomplete without consideration of genetic diversity. In the case of economic development, however, genetic diversity is merely another variable alongside nongenetic variables in a multiple

regression; although it is plausible that genetic diversity could affect the regression in the same way that nongenetic variables plausibly contribute to economic development, principles from population genetics produce no theory of the economic development of nations and thus do not contribute to this plausibility. The work of Ashraf and Galor (2013) is one of the first among recent studies seeking to identify an effect of a variable from population genetics on global economic outcomes. Given the novelty of population-genetic variables in attempts to address long-standing economic questions, such studies are likely to proliferate and deepen in methodologic sophistication. As genetic diversity and its interaction with social phenomena are considered in new contexts across different areas of inquiry, however, it will be important to take note of the distinction between fundamentally nonbiological uses of population-genetic variables and cases in which their utility is grounded in biology.

Acknowledgments

We thank E. Bendavid, W. Bodmer, M. Edge, K. Hunley, P. Norman, R. Rohlfs, M. Turelli, and an anonymous reviewer for comments on a draft of the manuscript; J. Cohen, M. Feldman, D. Laitin, and A. Saperstein for discussions; and M. Dey for research assistance.

Literature Cited

Ashraf, Q., and O. Galor, 2013 The "Out of Africa" hypothesis, human genetic diversity, and comparative economic development. Am. Econ. Rev. 103: 1–46.

Balding, D. J., 2013 Evaluation of mixed-source, low-template DNA profiles in forensic science. Proc. Natl. Acad. Sci. USA 110: 12241–12246.

Bergstrom, T. C., R. J. Garratt, and D. Sheehan-Connor, 2009 One chance in a million: altruism and the bone marrow registry. Am. Econ. Rev. 99: 1309–1334.

Bergstrom, T.C., R. J. Garratt, and D. Sheehan-Connor, 2012 Stem cell donor matching for patients of mixed race. B.E.J. Econ. Anal. Policy 12: 30.

Bieber, F., C. Brenner, and D. Lazer, 2006 Finding criminals through DNA of their relatives. Science 312: 1315–1316.

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

Burchard, E. G., 2014 Missing patients. Nature 513: 301–302.

Bustamante, C. D., E. G. Burchard, and F. M. De la Vega, 2011 Genomics for the world. Nature 475: 163–165.

Butler, J. M., 2011 *Advanced Topics in Forensic DNA Typing: Methodology*. Academic Press, Waltham, MA.

Callaway, E., 2012 Economics and genetics meet in uneasy union. Nature 490: 154–155.

Cao, K., J. Hollenbach, X. Shi, W. Shi, M. Chopek *et al.*, 2001 Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of genetic diversity in these loci and contrasting distribution patterns in these populations. Hum. Immunol. 62: 1009–1030.

Carlson, C. S., M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak *et al.*, 2004 Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74: 106–120.

Chin, G. J., 2012 The long shadow of genetic capital. Science 337: 1150.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat. Genet. 38: 1251–1260.

d'Alpoim Guedes, J., T. C. Bestor, D. Carrasco, R. Flad, E. Fosse *et al.*, 2013 Is poverty in our genes? A critique of Ashraf and Galor, "The 'Out of Africa' hypothesis, human genetic diversity, and comparative economic development," *American Economic Review* (Forthcoming). Curr. Anthropol. 54: 71–79.

DeBakker, P. I. W., N. P. Burtt, R. R. Graham, C. Guiducci, R. Yelensky *et al.*, 2006 Transferability of tag SNPs in genetic association studies. Nat. Genet. 38: 1298–1303.

DeGiorgio, M., J. H. Degnan, and N. A. Rosenberg, 2011 Coalescence time distributions in a serial founder model of human evolutionary history. Genetics 189: 579–593.

DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg, 2009 Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. Proc. Natl. Acad. Sci. USA 106: 16057–16062.

Deshpande, O., S. Batzoglou, M. W. Feldman, and L. L. Cavalli-Sforza, 2009 A serial founder effect model for human settlement out of Africa. Proc. Biol. Sci. 276: 291–300.

Feldman, M. W., 2014 Echoes of the past: hereditarianism and *A Troublesome Inheritance*. PLoS Genet. 10: e1004817.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, K. K. Kidd, J. R. Kidd *et al.*, 2008 The genetic structure of Pacific Islanders. PLoS Genet. 4: e19.

Garrison, N. A., R. V. Rohlfs, and S. M. Fullerton, 2013 Forensic familial searching: scientific and social implications. Nat. Rev. Genet. 14: 445.

Gelman, A., 2013 Ethics and statistics: they'd rather be rigorous than right. Chance 26: 45–49.

Gershaw, C., A. Schweighardt, L. Rourke, and M. Wallace, 2011 Forensic utilization of familial searches in DNA databases. Forensic Sci. Int. Genet. 5: 16–20.

Greely, H. T., D. P. Riordan, N. A. Garrison, and J. L. Mountain, 2006 Family ties: the use of DNA offender databases to catch offenders' kin. J. Law Med. Ethics 34: 248–262.

Hardy, J., and A. Singleton, 2009 Genomewide association studies and human disease. N. Engl. J. Med. 360: 1759–1768.

Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta *et al.*, 2009 Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106: 9362–9367.

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95–108.

Huang, L., M. Jakobsson, T. J. Pemberton, M. Ibrahim, T. Nyambo *et al.*, 2011 Haplotype variation and genotype imputation in African populations. Genet. Epidemiol. 35: 766–780.

Huang, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis *et al.*, 2009 Genotype-imputation accuracy across worldwide human populations. Am. J. Hum. Genet. 84: 235–250.

International HapMap Consortium, 2005 A haplotype map of the human genome. Nature 437: 1299–1320.

International HapMap 3 Consortium, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype, and copy-number in worldwide human populations. Nature 451: 998–1003.

Lao, O., T. T. Lu, M. Nothnagel, O. Junge, S. Freitag-Wolf *et al.*, 2008 Correlation between genetic and geographic structure in Europe. Curr. Biol. 18: 1241–1248.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.

Liu, H., F. Prugnolle, A. Manica, and F. Balloux, 2006 A geographically explicit genetic model of worldwide human-settlement history. Am. J. Hum. Genet. 79: 230–237.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9: 356–369.

Mori, M., P. G. Beaty, M. Graves, K. M. Boucher, and E. L. Milford, 1997 HLA gene and haplotype frequencies in the North American population: the National Marrow Donor Program donor registry. Transplantation 64: 1017–1027.

Need, A. C., and D. B. Goldstein, 2009 Next generation disparities in human genomics: concerns and remedies. Trends Genet. 25: 489–494.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Pemberton, T. J., M. DeGiorgio, and N. A. Rosenberg, 2013 Population structure in a comprehensive data set on human microsatellite variation. G3 3: 891–907.

Pickrell, J. K., and D. Reich, 2014 Toward a new history and geography of human genes informed by ancient DNA. Trends Genet. 30: 377–389.

Prugnolle, F., A. Manica, and F. Balloux, 2005 Geography predicts neutral genetic diversity of human populations. Curr. Biol. 15: R159–R160.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman *et al.*, 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA 102: 15942–15947.

Robinson, J., J. A. Halliwell, H. McWilliam, R. Lopez, P. Parham *et al.*, 2013 The IMGT/HLA database. Nucleic Acids Res. 41: D1222–D1227.

Rohlfs, R. V., S. M. Fullerton, and B. S. Weir, 2012 Famiilial identification: population structure and relationship distinguishability. PLoS Genet. 8: e1002469.

Rohlfs, R. V., E. Murphy, Y. S. Song, and M. Slatkin, 2013 The influence of relatives on the efficiency and error rate of familial searching. PLoS One 8: e70495.

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

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard *et al.*, 2005 Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 1: 660–671.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298: 2381–2385.

Seldin, M. F., B. Pasaniuc, and A. L. Price, 2011 New approaches to disease mapping in admixed populations. Nat. Rev. Genet. 12: 523–528.

Steele, C. D., and D. J. Balding 2014 Statistical evaluation of forensic DNA profile evidence. Ann. Rev. Stat. Appl. 1: 361–384.

Stranger, B. E., E. A. Stahl, and T. Raj, 2011 Progress and promise of genome-wide association for human complex trait genetics. Genetics 187: 367–383.

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.

Wang, S., C. M. Lewis, M. Jakobsson, S. Ramachandran, N. Ray *et al.*, 2007 Genetic variation and population structure in Native Americans. PLoS Genet. 3: 2049–2067.

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

Communicating editor: M. Turelli

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.176750/-/DC1

Genetic Diversity and Societally Important Disparities

Noah A. Rosenberg and Jonathan T. L. Kang

DOI: 10.1534/genetics.115.176750

	Log population density in 1500 CE				
	(1)	(2)	(3)	(4)	(5)
Continent fixed effects	No	No	No	No	Yes
Number of countries	21	21	21	21	21
Genetic variables					
Observed diversity	413.504*** (97.320) 0.000483			225.443*** (73.781) 0.00856	203.817* (97.637) 0.0609
Observed diversity square	−302.647*** (73.344) 0.000634			−161.160** (56.155) 0.0124	−145.720* (80.413) 0.0973
Non-genetic variables					
Log Neolithic transition timing		2.396*** (0.272) 3.92×10^{-8}		1.214*** (0.373) 0.00578	1.135 (0.658) 0.112
Log percentage of arable land			0.730** (0.281) 0.0188	0.516*** (0.165) 0.00749	0.545* (0.262) 0.0617
Log absolute latitude			0.145 (0.178) 0.427	−0.162 (0.130) 0.230	−0.129 (0.174) 0.475
Log land suitability for agriculture			0.734* (0.381) 0.0711	0.571* (0.294) 0.0729	0.587 (0.328) 0.101
Optimum diversity	0.683			0.699	0.699
R^2	0.417	0.540	0.568	0.894	0.903

Table S1. **Regressions of "log population density in 1500 CE" on a series of variables, as performed by Ashraf & Galor (2013).** Each variable was employed and computed as in Ashraf & Galor (2013), using values they reported for the non-genetic variables and 53 population-specific values of genetic diversity from Ramachandran et al. (2005) and Rosenberg et al. (2005). The 53 populations represent 21 countries. Each entry of the table contains an estimate of a regression coefficient, a heteroscedasticity-robust standard error in parentheses, and the *P*-value. Significance at the 10, 5, and 1 percent levels is represented by *, **, and ***, respectively. Each column represents a regression performed with different subsets of independent variables. "Optimum diversity" is the diversity value at which the log population density is at its maximum. This table has been recomputed as in Table 1 of Ashraf & Galor (2013) using scripts they provided.

	Log population density in 1500 CE				
	(1)	(2)	(3)	(4)	(5)
Continent fixed effects	No	No	No	No	Yes
Number of countries	39	39	39	39	39
Genetic variables					
Observed diversity	30.943 (47.026) 0.515			37.691 (25.230) 0.145	28.855 (61.403) 0.642
Observed diversity square	−17.143 (36.238) 0.639			−23.088 (20.408) 0.266	−19.796 (54.530) 0.719
Non-genetic variables					
Log Neolithic transition timing		2.076*** (0.362) 1.45×10^{-6}		1.693*** (0.380) 9.63×10^{-5}	1.324*** (0.354) 0.000796
Log percentage of arable land			0.991*** (0.262) 0.000574	0.456** (0.190) 0.0220	0.487** (0.205) 0.0240
Log absolute latitude			−0.167 (0.197) 0.404	−0.173 (0.181) 0.348	−0.334* (0.184) 0.0799
Log land suitability for agriculture			0.253 (0.379) 0.510	0.540* (0.269) 0.0535	0.497** (0.224) 0.0345
Optimum diversity	0.903			0.816	0.729
R^2	0.101	0.458	0.443	0.762	0.825

Table S2. Regressions of "log population density in 1500 CE" on a series of variables, computed as in Table S1, except that 237 populations from Pemberton et al. (2013), representing 39 countries, were used. Unlike in Table S1, the observed diversity and observed diversity square variables are not significant.

	Log population density in 1500 CE				
	(1)	(2)	(3)	(4)	(5)
Continent fixed effects	No	No	No	No	Yes
Number of countries	21	21	21	21	21
Genetic variables					
Observed diversity	598.189*** (130.670) 0.000233			335.137*** (110.942) 0.00916	265.482 (148.199) 0.101
Observed diversity square	−432.029*** (96.698) 0.000297			−237.527** (82.622) 0.0122	−183.002 (117.198) 0.147
Non-genetic variables					
Log Neolithic transition timing		2.396*** (0.272) 3.92×10^{-8}		1.257*** (0.371) 0.00442	1.183* (0.655) 0.0984
Log percentage of arable land			0.730** (0.281) 0.0188	0.500** (0.172) 0.0114	0.459* (0.252) 0.0957
Log absolute latitude			0.145 (0.178) 0.427	−0.212 (0.145) 0.167	−0.145 (0.208) 0.501
Log land suitability for agriculture			0.734* (0.381) 0.0711	0.588* (0.297) 0.0680	0.631* (0.324) 0.0773
Optimum diversity	0.692			0.705	0.725
R^2	0.411	0.540	0.568	0.891	0.900

Table S3. Regressions of "log population density in 1500 CE" on a series of variables, computed as in Table S1, except that 136 populations from Pemberton et al. (2013), representing the same 21 countries in Table S1, were used. In models 1 and 4 but not 5, the observed diversity and observed diversity square variables are significant.