DECONSTRUCTING THE RELATIONSHIP BETWEEN GENETICS AND RACE

Michael Bamshad[*‡], *Stephen Wooding*[*], *Benjamin A. Salisbury*[§] *and J. Claiborne Stephens*[§]

The success of many strategies for finding genetic variants that underlie complex traits depends on how genetic variation is distributed among human populations. This realization has intensified the investigation of genetic differences among groups, which are often defined by commonly used racial labels. Some scientists argue that race is an adequate proxy of ancestry, whereas others claim that race belies how genetic variation is apportioned. Resolving this controversy depends on understanding the complicated relationship between race, ancestry and the demographic history of humans. Recent discoveries are helping us to deconstruct this relationship, and provide better guidance to scientists and policy makers.

The classification of humans by race has a long and tumultuous history in biomedical research, partly because the way in which race has been defined and applied in research (BOX 1) has had important implications for science and society[1–4]. The biological information captured by concepts of race is confounded by personal experiences and varied perceptions of race as social, economic and political identities[5]. Although highlighting genetic differences among people might unfortunately reinforce the stereotypic features of these identities, exploring the genetic influence on common health-related traits and disparities could also be beneficial to human health. For example, it might lead to better prevention strategies and treatment options for common maladies such as infection, heart disease and cancer. This tension highlights the urgent need for and the challenges of determining the relationships between race, patterns of human genetic variation and inferences of individual ancestry.

Investigation of the biological differences between populations that are defined by ethnic and racial labels has intensified as an increasing number of polymorphisms putatively associated with morphological characteristics, disease susceptibility and environmental response are reported[6]. Much of the contention about using such labels as proxies for genetic relatedness has been fostered by the conflation of several issues: whether individuals can be reliably allocated into valid genetic clusters in which all members share more recent common ancestry than members of other clusters; whether descriptors such as race or ethnicity capture any of the genetic differences between such clusters; and whether these differences are meaningful for health-related variation among groups. Although none of these issues can be neatly resolved given our current knowledge of human genetic variation, it has become increasingly clear that genetic data can be used to distinguish groups and allocate individuals into groups. This raises several important questions that we address in this review: What are the most effective ways of inferring genetic groups and individual ancestry? What is the correspondence between groups that are inferred using explicit genetic data versus groups that are distinguished using ethnic and racial labels? Do these groups frequently share common polymorphisms? For which of these groups does membership more accurately predict health-related traits such as disease susceptibility or drug response?

Our focus here is to describe, and in some instances interpret, how patterns of human genetic variation relate to these questions and to clarify some of the misconceptions about how human genetic variability is

*Department of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA. ‡Department of Pediatrics, University of Utah, Salt Lake City, Utah 84112, USA. §Genaissance Pharmaceuticals, New Haven, Connecticut, 06511, USA. Correspondence to M.B. e-mail: mike@genetics.utah.edu
doi:10.1038/nrg1401

Box 1 | **Labelling human populations**

The term 'race' probably derives from the old French word *rasse* and the Italian word *razza*, which roughly translate into breed or strain. It is difficult to say exactly when the term was first used to refer to differences in appearance between human populations[78]. Some scholars believe that the term became widespread in Europe in the sixteenth and seventeenth centuries[79]. One of the best-known classifications was that of Johann Blumenbach, who in his 1775 book *On the Natural Varieties of Mankind* defined five races: the Caucasian, or white, race, which included most European nations and those of Western Asia; the Mongolian, or yellow, race, which included China and Japan; the Ethiopian, or black, race occupying most of Africa (except North Africa), Australia, Papua and other Pacific Islands; the American, or red, race comprising the Indians of North and South America; and the Malayan, or brown, race, which included the islands of the Indian Archipelago. The use of these 'traditional' terms is often context-dependent, and therefore controversial: for example, a person who is described as black in one society might be considered non-black in another.

Many naturalists, including Blumenbach, struggled with the biological meanings of racial categories in some of the same ways that we do today. He acknowledged that morphological variation varied widely within each race and often overlapped with variation observed in other races, that boundaries between races were not discrete and that races could not be defined solely by geographical boundaries between continents or otherwise. But scientists have continued to often treat groups identified by commonly used racial and ethnic labels as biological categories. Proponents of such classification argue that race and genetic differences are strongly associated, justifying the use of race as a proxy for POPULATION STRUCTURE in the design of experiments and medical application[80,81]. Specifically, they contend that individuals who are assigned to the same racial category share more of their recent ancestry and therefore are more similar genetically to each other than individuals from different racial categories, and that the accuracy of race as a proxy for ancestry is good enough to be useful as a research variable. But others argue that race is neither a meaningful concept nor a useful heuristic device[19,82], and even that genetic differences between groups are biologically meaningless[83] or that genetic differences among human populations do not exist[84]. Many people prefer to use 'ethnic group,' frequently defined as a group that has shared religious, social, linguistic and cultural heritage/identity, instead of race, but the two terms suffer from the same conceptual and heuristic problems and questions.

measured and distributed. We also point out some of the limitations of what we know about the distribution of human genetic variation and indicate areas in which further investigation is needed. We emphasize that any framework that is designed to study the relationship between patterns of genetic variation and notions of race must consider the impact of these relationships on concepts of identity, lay perceptions of race, the application of justice and the development of public policy. These topics are beyond the scope of this review so the reader is referred elsewhere[7,8].

Population history and genetic variation

We commonly sort individuals into groups based on characteristics such as physical appearance or language. Most Danes, for example, resemble each other more than they do Italians. But Danes and Italians resemble each other more than either group resembles sub-Saharan Africans. Human phenotypic variation is therefore organized in a sort of geographical pattern. Human genetic variation shows similar geographical relationships. Members of the same local group are typically more closely related to each other than to members of groups who live in different geographical areas, and people who live in the same principal geographical region are more similar than those separated by key geographical barriers such as mountains. So how is genetic variation in humans geographically distributed?

Since the 1980s, there have been indications that the genetic diversity of humans is low compared with that of many other species[9]. This has been interpreted to mean that humans are a relatively young species, so populations have had relatively little time to differentiate from one another[10]. For example, 2 randomly chosen humans differ at ~1 in 1,000 nucleotide pairs, whereas two chimpanzees differ at ~1 in 500 nucleotide pairs[11]. Nevertheless, on average, 2 humans differ at ~3 million nucleotides. Most of these polymorphisms are neutral or almost neutral, but a fraction of them are functional and influence phenotypic differences between people[12].

The distribution of neutral polymorphisms among humans reflects the demographic history of our species[10]. Genetic and archaeological evidence indicates that, over the past 100,000 years, the population size of humans increased markedly and humans dispersed from Africa to colonize other parts of the world. This process affected the geographical distribution of genetic variation in two important ways. First, FOUNDER POPULATIONS typically carried with them only a subset of the genetic variation found in their ancestral population. Second, as these founders became more widely separated from one another, the probability of two individuals mating with one another became increasingly variable. Individuals were more likely to mate with one another if they lived closer to each other, and members of the same founder group typically lived closer to each other than members of different groups. This ASSORTATIVE MATING restricted the shuffling of polymorphisms among individuals who lived in different geographical regions, and over time led to genetic differentiation between groups.

The dispersal of humans throughout the world distributed genetic variation in two other ways. First, the degree to which the frequencies of neutral polymorphisms fluctuate is inversely related to population size — there is more GENETIC DRIFT in smaller populations. This increased the VARIANCE of neutral polymorphism frequencies among groups. Second, polymorphisms that arose in one group had a greater chance of remaining restricted to that group because their spread was limited by the degree of gene flow between groups. The founding and dispersion of new populations eventually led to the sorting of different polymorphisms and polymorphisms with different frequencies among populations. This raises several questions. Can boundaries between groups be inferred using genetic data? If so, what is the most precise way to do it? What is the most reliable way to infer individual ancestry? The following sections describe the genetic tools that are being used to address these questions and illustrate some of their limitations.

Inference of population structure

In the 1970s and 1980s, it became increasingly common to use multilocus genotypes to distinguish different human groups[13], and in the 1990s, to allocate individuals to groups[14]. Repeatedly, it was shown that restriction site polymorphisms[15], short tandem repeat polymorphisms

POPULATION STRUCTURE
Organization of a population into sub-populations as a consequence of factors such as finite population size and geographical subdivision.

FOUNDER POPULATION
A group of individuals that establishes a new population.

ASSORTATIVE MATING
Nonrandom choice of mates based on phenotypic characteristics such as geographical proximity, skin colour, height or religion.

GENETIC DRIFT
Fluctuations of allele frequencies over time due to chance alone.

VARIANCE
A statistic that quantifies the dispersion of data about the mean.

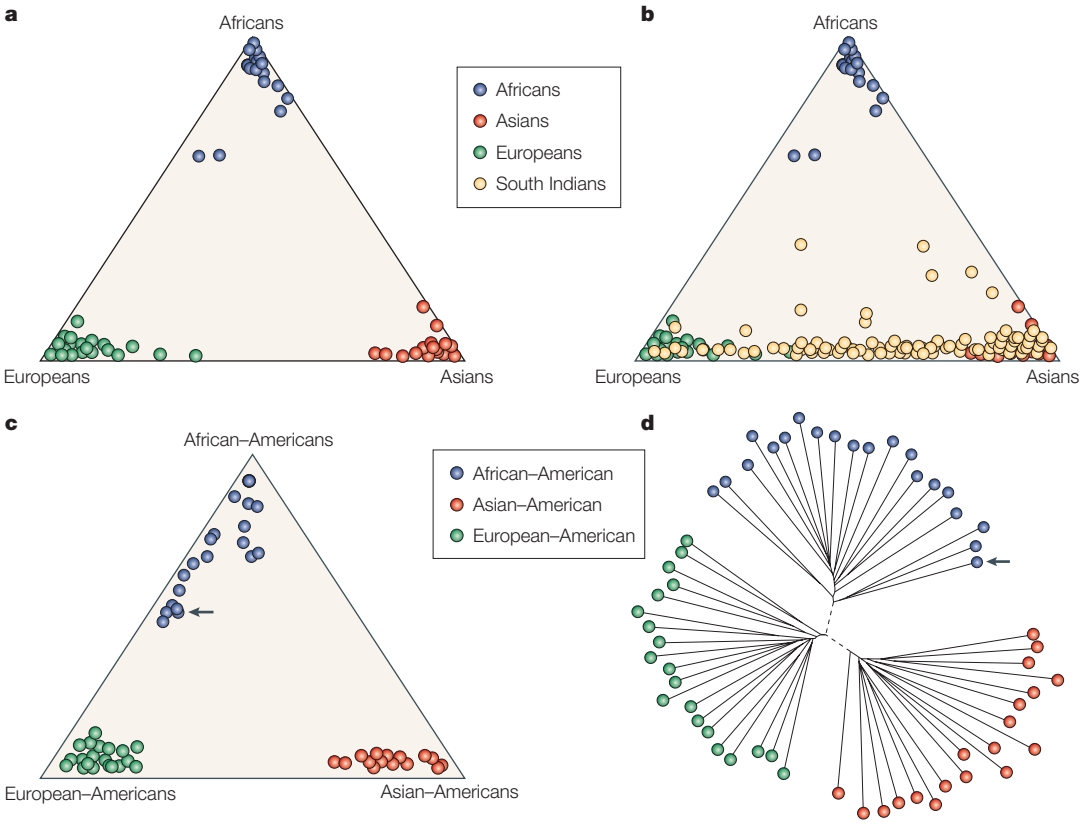


Figure 1 | **Inference of individual ancestry proportions from genetic data. a** | Inferred ancestry proportions for 107 sub-Saharan African, 67 East Asian and 81 Western European individuals (filled circles) genotyped for 100 *Alu* insertion polymorphisms[24] (see online supplementary information S1 (data)). The distance between a circle and each side is proportional to the fraction of an individual's ancestry shared with Africans, Asians and Europeans. For many individuals, the proportion of ancestry from a single population is almost 100%. **b** | Inferred ancestry proportions for the same individuals in **a** plus 263 individuals from South India. The proportion of ancestry shared with Europeans and Asians varies widely among individuals from South India. **c** | Inferred ancestry proportions for individuals who identified themselves as African (n = 20), Asian (n = 19) and European–American (n = 20) in the Genaissance panel (see online supplementary information S2 (data)), genotyped for 500 coding SNPs with a minor allele frequency of 10%. The circles are less tightly clustered because the proportion of ancestry among individuals is more varied. For example, one African–American (arrow) is estimated to share ~60% of his ancestry with other African–Americans and ~40% of his ancestry with European–Americans. **d** | A network that depicts the genetic relatedness among individuals from **c** using 250 SNPs (see online supplementary information S3 (data)). The length of each branch (black lines) is proportional to genetic distance. The distance between any two circles of the same colour (solid lines) is large and contributes to high within-group variance, whereas the distance between clusters (dotted lines) is small and contributes to low between-group variance. Branches that connect individuals with a high proportion of ancestry from more than one population (arrow) are connected directly to the branches between clusters.

(STRs)[16], SNPs[17] or *Alu* insertion polymorphisms[18] could be used to delimit groups and to assign individuals to specific groups. They also showed that groups that live on the same continent were typically more similar to each other than groups from different continents. However, in all of these studies, the identities of groups and individuals were assigned *a priori*. In other words, ancestry information such as race, ethnicity or geographical origin was used in conjunction with genetic data to infer group boundaries and allocate individuals to groups. If individuals were stripped of all prior information about ancestry (such as geographical location of origin, race, ethnic group) and assigned to groups *a posteriori* using only genetic data, it was less clear that geographical origin or racial categories provided reliable information about population structure[19]. However, the accurate inference of an individual's ancestry using genetic data depends on several factors, including the number of genotypes used, the degree of differentiation between groups and how each group is sampled — the magnitude of the effect of each of these variables in human studies is just beginning to be explored[20,21].

Judicious choice of markers. The statistical power with which ancestry can be inferred reliably depends on the informativeness of each marker tested. Informativeness can be measured in many ways, although the performance of different statistics varies depending on the circumstances[22]. For a biallelic marker such as a SNP, informativeness is maximized if one allele is limited to a single population — if an individual has this allele, there is no uncertainty about his origin. On average, dinucleotide STRs are approximately 5–8 times more informative for inferring ancestry than random SNPs[22], and

markers that are informative for inferring regional ancestry (for example, within continents) are, in large part, also informative for inference among populations in these regions. This result is to be expected because populations from neighbouring regions typically share more recent common ancestors. Therefore, their allele frequencies are more highly correlated, a pattern that is commonly manifest as a CLINE of allele frequencies. The occurrence of such clines is often offered as evidence that individuals cannot be allocated into genetic clusters[23]. Clines for some loci are steep, whereas others are gradual, with the overall proportion of each reflected by the level of genetic differentiation among groups. Similar to distinguishing between similar temperatures, low levels of differentiation between groups make distinguishing them more difficult, but not impossible.

How many markers? If people from different continents — chosen to maximize the level of genetic differentiation among groups — are stripped of ancestry information, how many markers are required to distinguish groups and reliably allocate individuals into these groups? For a sample of ~200 individuals from sub-Saharan Africa, Europe and East Asia, correct allocation to the continent of origin with a mean accuracy of 90% requires ~60 randomly selected *Alu* insertion polymorphisms or STRs[24]. The mean accuracy of allocation improves to 99–100% with the use of a modest ~100–160 markers (FIG. 1). What if individuals are sampled from broader geographical regions? Rosenberg *et al.* studied 52 ethnic groups distributed worldwide and allocated each of ~1,000 individuals into 1 of 5 different genetic clusters using 377 randomly selected STRs[25]. Each cluster represented people whose ancestors were typically isolated by large geographical barriers: sub-Saharan Africans; Europeans and Asians west of the Himalayas; East Asians; inhabitants of New Guinea and Melanesia; and Native Americans. These studies confirmed that there is a relationship between patterns of genetic variation and geographical ancestry; with a high degree of accuracy and reliability using a relatively modest number of multilocus genotypes, individuals can indeed be allocated to groups that represent broad geographical regions.

Effects of population sampling. Do these results provide us with any guidance about the relationship between genetic variation and race? Insofar as geographical ancestry corresponds to some notions of race, patterns of genetic variation will also co-vary with these notions. However, the genetic clusters inferred were composed of individuals who lived in widely separated geographical regions, and individuals who were allocated to clusters most accurately were from non-ADMIXED populations. Both studies failed to accurately allocate individuals from Central or South Asia into a genetic cluster that corresponded to common concepts of race (FIG. 1c), presumably because these populations are historically admixed with populations from both Europe and Asia[25,26]. Similarly, if individuals were sampled continuously from region to region, it might be more difficult to infer genetic clusters that were inclusive of all or even most individuals in large geographical regions. Many hundreds of polymorphisms might have to be examined to distinguish between groups that separated within the past several thousand years or admixed with one another thousands of years ago.

The limited guidance offered by these studies highlights the deficits in our basic knowledge about the geographical distribution of human genetic variation. The sampling of individuals from many parts of the world (such as sub-Saharan Africa, India, North and South America) has been extremely limited even though genetic diversity in some of these regions (such as Africa or India) seems to be higher than in many parts of the world and there is substantial genetic structure among African populations[24,25,27]. In India alone, for example, thousands of different castes/sub-castes and ~450 tribal groups comprising approximately one-fifth of the world population have been documented, but molecular genetic data are available for only a handful[28]. Indeed, much of the genetic data available on many populations is limited to genotypes of the mitochondrial genome and the non-recombining portion of the Y chromosome. But assessments of patterns of variation based on single-locus analyses fundamentally lack power. What is needed instead is an unbiased sampling of variation (for example, through resequencing) across the genome from individuals in well-characterized communities sampled from contiguous geographical regions throughout the world.

Relying on polymorphisms subject to selection. Inferences of human population structure based on genetic data often differ from inferences based on phenotypic characteristics. So, even if humans can be allocated into groups using genetic data, these groups might correspond only crudely, if at all, to common notions of race based on physical traits. For instance, although facial features and skin pigmentation are routinely used to group people by race, populations that share similar physical characteristics as a result of natural selection can be very different genetically. For example, the degree of skin pigmentation in some sub-Saharan African, South Indian and Melanesian populations is similar because of adaptive evolution. But, genetically these populations are quite dissimilar. By contrast, the Ainu of northern Japan are morphologically different (for example, they have less skin pigmentation and more body hair) from other East Asian populations, but are genetically very similar to them[29]. Overall, the degree of differentiation in QUANTITATIVE TRAITS often exceeds that observed for neutral markers, indicating that these traits have been subject to natural selection[30]. Therefore, their distribution does not necessarily reflect the distribution of neutral polymorphisms nor are they good predictors of group membership. Indeed, these characteristics might imply that there is a closer degree of relatedness than exists or vice versa.

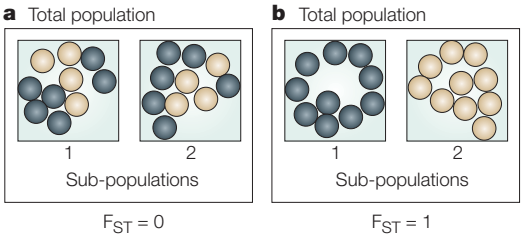
Inferences of population structure based on polymorphisms that have been subject to natural selection can also be unreliable. Some polymorphisms that have been exposed to LOCAL POSITIVE SELECTION have increased

CLINE
A gradient in the frequency of an allele.

ADMIXTURE
The mixing of two or more genetically differentiated populations.

QUANTITATIVE TRAIT
A measurable trait that depends on the cumulative action of many genes and that can vary among individuals over a given range to produce a continuous distribution of phenotypes. Common examples include height, weight and blood pressure.

Box 2 | The apportionment of human diversity: implications for inferences of ancestry



F_{ST} is a statistic that compares the level of genetic variation within two or more sub-populations relative to all sub-populations combined (that is, the total population). F_{ST} is often estimated as: $F_{ST} = (T–S)/T$, where T is the average difference between pairs of alleles (or allele frequencies) drawn at random from the total population, and S is the average difference between pairs of alleles (or allele frequencies), both drawn at random from the same sub-population. The figure illustrates two examples of how pairwise differences within and among populations can be distributed. If mating in the total population is random, then T and S are expected to have the same value and F_{ST} will be 0 (panel **a**). However, if mating between sub-populations is nonrandom (that is, they are partially reproductively isolated), genetic variation in each sub-population will be less than variation in the total population, and F_{ST} will increase to a maximum of 1.0 or there will be complete differentiation between sub-populations (panel **b**). Therefore, F_{ST} is a measure of the departure from random mating caused by population structure.

Estimates of F_{ST} can be calculated in several ways. One common method is to count the number of nucleotide differences between DNA sequences in each population. If sub-Saharan Africa, Europe and East Asia are used to delimit sub-populations in humans, F_{ST} values estimated from DNA sequence data are often around 5–15% (REF. 16). That is, on average, the difference between randomly chosen human DNA sequences is roughly 15% greater than the difference between sequences from the same region. The average difference between humans living in these different regions is slightly greater than would be expected if humans were mating at random.

Although F_{ST} is informative about average differences within and among populations, it tells us little about the consistency, or variance, of these differences. If variances within sub-populations are relatively high (that is, if a similar number of coloured balls are present in each sub-population in panel **a**), it might be difficult to determine the population to which an individual belongs. Conversely, if variances within sub-populations are relatively low (that is, if balls of different colours are restricted to different sub-populations, as in panel **b**), the origin of an individual might be easily determined. However, it is possible for two different pairs of sub-populations to have the same F_{ST}, but have different within- and between-population variances. Therefore, F_{ST} does not necessarily reflect the probability that two DNA sequences drawn at random from the same sub-population are more similar than two sequences drawn randomly from the total population.

differentiation among populations (for example, the *FY*O* allele of the *DUFFY* locus that confers protection against *Plasmodium vivax* malaria) compared with estimates based on neutral polymorphisms, and are highly informative for inferring ancestry[31]. But some such polymorphisms are sometimes found in genetically dissimilar groups because of admixture or CONVERGENT EVOLUTION. For example, the *FY*O* allele seems to have arisen independently in both sub-Saharan Africans and New Guinea highlanders[32]. By contrast, the level of differentiation among groups can be substantially lower if estimated from polymorphisms that have been subject to BALANCING SELECTION. Polymorphisms in the 5′-*cis*-regulatory region CC chemokine receptor 5 (*CCR5*) and the coding region of the phenylthiocarbamide sensitivity gene (*TAS2R38*) are good examples[33,34]. Nevertheless, if enough loci are analysed, polymorphisms subject to selection can be used to make accurate inferences of ancestry. For example, 500 SNPs with a frequency of ≥10% randomly chosen from the regulatory and coding regions of 3,931 genes sequenced in 20 African–, 20 European– and 19 Asian–Americans (hereafter referred to as the Genaissance panel; see online supplementary information S2 (data)) allocated individuals into clusters that were entirely concordant with self-assessed ancestry (FIG. 1c). Therefore, despite assertions to the contrary[35], ancestry inferences are robust using a modest number of polymorphisms in either coding or non-coding regions.

LOCAL POSITIVE SELECTION
A type of natural selection in which favoured variants increase in frequency in a localized geographical region.

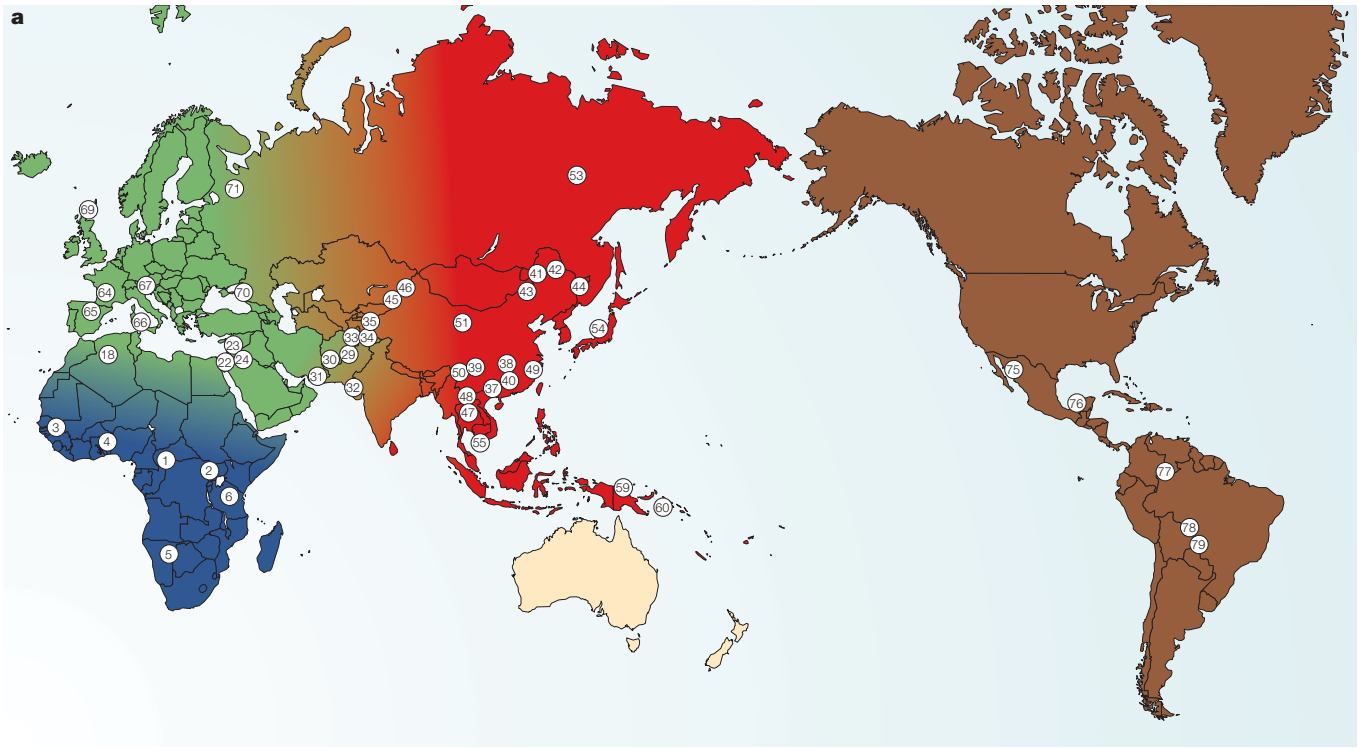
DUFFY BLOOD GROUP
This group is defined by variants in a chemokine receptor that is present on the surface of several types of cell, including red blood cells. This receptor must be present for *Plasmodium vivax* to invade cells and cause malaria.

CONVERGENT EVOLUTION
A process in which traits evolve to a similar state in two or more genetically distinct populations, typically as an adaptive response.

BALANCING SELECTION
A selection regime that results in the maintenance of two or more alleles at a single locus in a population.

The effects of population structure

Given that genetic data can be used to infer population structure and frequently assign individuals to populations that correspond with their self-identified geographical ancestry, what does this tell us about the way that polymorphisms are distributed among groups? Many studies over the past 30 years have shown that most genetic variation is distributed among individuals within populations rather than among populations[36–38]. The fraction of total genetic variance that is distributed among populations is frequently measured by a statistic known as Wright's Fixation Index or F_{ST} (BOX 2)[39,40]. Among sub-Saharan Africans, East Asians and Northern Europeans, F_{ST} estimates based on neutral, autosomal markers typically vary between ~5 and 15% depending on the population sampling strategy and marker characteristics[24,25,41]. These are relatively low values (compared with other species), which critics of race have, for several reasons, suggested are too low to justify the existence of human races[42]. Frequently, it is erroneously contended that the high (85–95%) within-group variance of human populations is inconsistent with the existence of races because differences between individuals are greater than differences between groups[43]. Such low F_{ST} values are sometimes misinterpreted to mean that genetic differences between individuals within sub-Saharan Africa, Asia or Europe are typically greater than differences between individuals on different continents[44,45]. A positive F_{ST} indicates,



b

	Africans	Europeans	Asians
Europeans	0.635		
Asians	0.645	0.617	
Native Americans	0.739	0.666	0.650

c

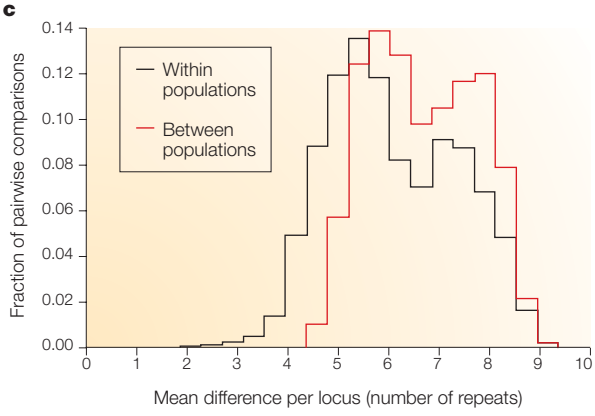


Figure 2 | **Comparison of genetic differences among individuals in different 'racial' populations. a** | Map of the world illustrating the geographical boundaries that historically have been used to classify populations into the different racial groups: Africans (blue), Europeans (green), Asians (red) and Native Americans (brown)[78]. Africa and Asia are the continents of origin of more than one racial group, and the boundaries between racial groups are not discrete. The geographical origin of each local population from the CEPH HUMAN DIVERSITY PANEL is numbered, and the populations are classified according to the definition of the US Office of Management and Budget (see online supplementary information S4 (data)). **b** | Matrix of probabilities of randomly choosing two individuals from different 'racial' populations (see text for details) who are more dissimilar, based on the number of shared short tandem repeat (STR) alleles, than two randomly chosen individuals from the same population. For example, an African will be more dissimilar to an Asian individual compared with another African individual ~65% of the time. **c** | Distribution of the mean number of pairwise differences of STRs per locus of individuals within (black line) and between (red line) populations summed across Africans, Asians, Europeans and Native Americans. There is substantial overlap between the distributions, but the mean number of repeats per locus between individuals from different populations is frequently greater than between individual from the same population.

CEPH HUMAN DIVERSITY PANEL
A resource of 1,064 cultured lymphoblastoid cell lines from individuals in 51 different world populations that are banked at the Foundation Jean Dausset (CEPH) in Paris, France.

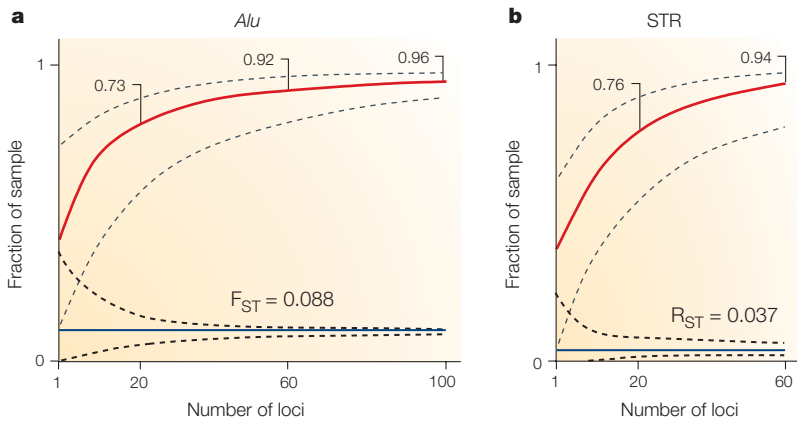


Figure 3 | **Marker performance for inference of individual ancestry versus estimation of F_{ST} or R_{ST}.** Proportion of sub-Saharan Africans, Europeans and East Asians assigned correctly (red lines) to their population of origin (data from REF. 24), and predicted F_{ST} estimated from 1–100 *Alu* insertion polymorphisms (blue line in **a**; see online supplementary information S1 (data)) or predicted R_{ST} estimated from 1–60 short tandem repeats (STR; blue line in **b**; see online supplementary information S5 (data)). The traces are bounded by 95% confidence intervals (dotted lines). The estimate of F_{ST} or R_{ST} remains almost the same as the number of markers increases, whereas the probability of correctly classifying an individual increases.

however, that individuals from different populations are, on average, slightly more different from one another than are individuals from the same population (FIG. 2).

To quantify how often an individual is more similar to another individual from the same group versus a different group, we defined racial groups *a priori* using a particularly crude and contentious classification scheme proposed by the Office of Management and Budget (OMB) and adopted by many federal agencies including the National Institutes of Health and the Food and Drug Administration[35,46]. The OMB defined five categories for race based on physical features and/or country of origin: African–American, 'White', American–Indian or Alaska Native, Asian, and Native Hawaiian or Pacific Islander. We categorized the populations of the CEPH Human Diversity Panel (see online links box) according to these OMB categories, and estimated how often an individual was more different from a person in a different 'racial group' than a person from the same 'racial group' based on 377 STR genotypes[25]. Regardless of the racial group to which an individual belonged, two people from different racial groups were more different than two individuals from the same racial group approximately two-thirds of the time (FIG. 2b,c). Only approximately one-third of the time were two people from the same racial group more different that two individuals from different racial groups. This estimate is probably conservative as the proportion of comparisons in which two individuals from different racial groups were more similar would have been higher if admixed populations such as African–Americans had also been sampled.

Differentiation versus sorting. If the proportion of genetic variation distributed among populations is so low (that is, 10–15%), how can genetic differences be used to sort individuals into groups reliably? The explanation lies in the way that each genotype is used to estimate F_{ST} versus sorting individuals into genetically inferred groups. Each genotype contributes equally to the estimate of between-population and within-population variation, so the proportion remains the same no matter how many loci are sampled. Therefore, F_{ST} remains about the same regardless of the number of genotypes analysed, although the confidence interval around the estimated F_{ST} gets smaller (FIG. 3). By contrast, the contribution of each genotype to the allocation of individuals into genetically inferred groups is cumulative over loci because allele frequencies among loci within a population are correlated. Accordingly, as the number of genotypes used to assign individuals to genetic clusters increases, so does the proportion of individuals allocated correctly (FIG. 3). So, although there might be little variation among groups, it is highly structured and therefore useful for distinguishing groups and allocating individuals into groups[47].

Distribution of common polymorphisms. Understanding the structure of neutral human genetic variation provides insights about the allelic structure of health-related genetic variation. Population genetics theory predicts that most polymorphisms are rare — singletons being the most common — and consequently confined to a single population. Therefore, most genetic variation is, arguably, population-specific. However, such rare polymorphisms are not practical for inferring ancestry (that is, such polymorphisms might be found in only a single individual or family) and do not contribute appreciably to disease risk in a population. Several investigators have suggested that, with the exception of alleles that have risen to a high frequency as a result of local positive selection, few alleles common enough to be of medical significance for an entire group are likely to be confined to one population[48]. This prediction is based, in part, on a model of complex disease in which the underlying causal variants are common — the common-disease/common-variant hypothesis (CD/CV) — and are therefore old and found in multiple groups, rather than being rare and population-specific[49]. Are most common variants found in multiple groups? Several analyses of allele sharing among groups have indicated that the answer is yes, but in most of these studies, alleles were identified by genotyping common SNPs that were ascertained in a single or at most several groups[50]. This ASCERTAINMENT strategy can cause an upwards bias in the estimate of allele sharing among groups[51]. What proportion of SNPs and HAPLOTYPES are shared among groups if alleles are ascertained in an unbiased fashion?

We analysed 63,724 SNPs that were found by resequencing the regulatory and coding regions of 3,931 genes in the Genaissance panel (see online supplementary information S6 (data)). Comparing just the SNPs that are polymorphic in African– and European–Americans (n = 50,736) and defining a common SNP as one with a minor allele frequency of 10% in one or both populations, we found 20,409 common SNPs (7,776 common only in African–Americans and 2,802 common only in European–Americans). Of these SNPs, 4,704 (23.1%) were private (that is, population-specific)

R_{ST}
A statistic similar to F_{ST} that is used to estimate differentiation among groups by using microsatellite markers.

ASCERTAINMENT
The selection of samples (such as markers, individuals, populations) through a process that often deviates from random sampling and can therefore introduce bias.

HAPLOTYPE
The combination of alleles or genetic markers that is found on a single chromosome of a given individual.

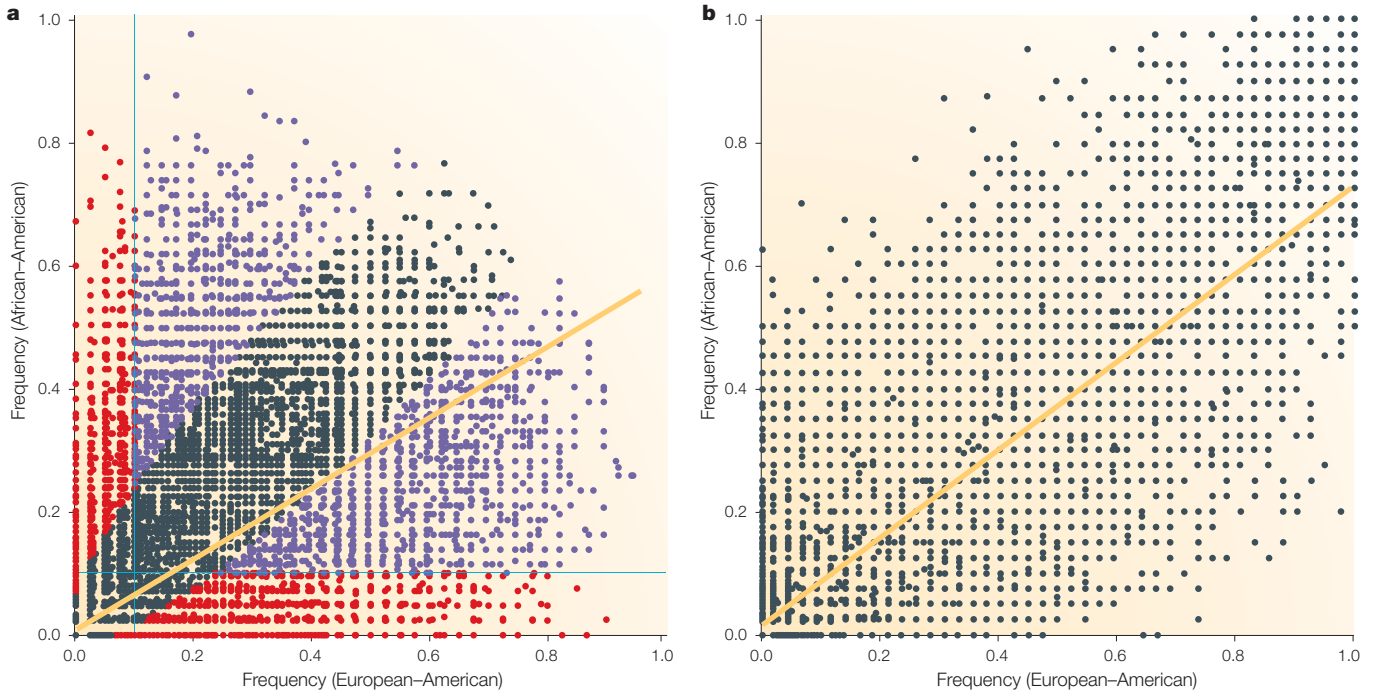


Figure 4 | **Comparison of polymorphism frequencies between African– and European–Americans. a** | Comparison of the frequencies of SNPs shared among African–American and European–American populations from the Genaissance panel in which 3,931 genes were resequenced (see online supplementary information S6 (data)). The less frequent allele in the combined population was designated as the minor allele, and the frequency of the minor allele was calculated in each population for the 50,735 SNPs analysed. A linear correlation of the minor allele frequency in European–Americans with the minor allele frequency in African–Americans had an R^2 (coefficient of determination) of 0.37. Sites where the allele frequency in each population was not significantly different ($|Z| < 1.65$, $p > 0.05$) are shown as black circles. Sites with significant differences in allele frequency ($|Z| > 1.65$, $p < 0.05$) between populations that are common in both populations are shown as blue circles. Sites with significant differences in allele frequency ($|Z| > 1.65$, $p < 0.05$) between populations that are common in only one population are shown as red circles. **b** | Comparison of the frequencies of 37,749 haplotypes found in African–American and European–American populations. A linear correlation of the haplotype frequency in African–Americans with the haplotype frequency in European–Americans had an R^2 of 0.79.

in African–Americans, and 585 (2.9%) were found only in European–Americans. Only 9,831 (48%) SNPs were common in both populations, and of these SNPs, 4,015 (41%) had significantly different ($|Z| > 1.65$, $p < 0.05$) allele frequencies (FIG. 4a). If 'common' is redefined as a minor allele frequency of 20%, 12,641 common SNPs were found (4,322 common only in African–Americans and 2,902 common only in European–Americans). Of these SNPs, 1,220 (9.7%) were private in African–Americans and 117 (1%) were found only in European–Americans (see online supplementary information S6 (data)). Therefore, most of the common SNPs in this data set are either private or common in only a single population. The pattern was similar for haplotypes inferred from these 3,931 genes, with only 51% of 8,876 haplotypes with a frequency of ≥10% being shared by African– and European–Americans (FIG. 4b). Of these 8,876 common haplotypes, 2,687 (30%) were common only in African–Americans and 1,703 (19%) were common only in European–Americans. Whether these findings can be generalized to the entire genome and all main human populations is unclear, but they do indicate that more data are needed before we can conclude that common variants are typically shared by all main human populations. If they are not, it might be necessary to develop initiatives to identify additional alleles that are common specifically in each population to be studied[52,53]. Moreover, these results indicate that even if common variants are shared among groups, their frequencies often differ substantially. This underscores the need to account for population structure in study designs that can be confounded by population stratification.

Race as a predictor of individual ancestry proportions. If some individuals can be sorted broadly into genetic groups that are concordant with their self-assessed racial identity, is race a good predictor of individual ancestry? Most people who identify themselves as African–Americans have relatively recent ancestors from West Africa, and West Africans, like most sub-Saharan Africans, have a pattern of polymorphism frequencies that can be used to distinguish them from Europeans, Asians and Native Americans. However, the fraction of variation that an African–American individual shares with West Africans varies considerably because, over the past few centuries, African–Americans have admixed to variable degrees with groups that originate from other geographical regions. Several studies have shown that the West African contribution to individual

African–American ancestry is on average ~80%, although it ranges from ~20–100% (REF. 54). The genetic composition of self-identified European–Americans also varies, with ~30% of European–Americans estimated to have <90% European ancestry[54]. Accordingly, membership in a genetically inferred cluster does not mean that all members of the cluster necessarily have a similar genetic composition. This observation is important because knowing the proportion of an individual's ancestry that originated in different populations can be useful for identifying genetic and environmental factors that underlie common diseases for which risk varies among populations[55–57]. To this end, several hundred loci that are particularly informative for estimating ancestry proportions in African, European, Asian, Hispanic and Native Americans have been identified[58,59]. These ancestry informative markers (AIMs) can be used to estimate individual ancestry proportions for forensic, clinical and scientific applications (see article by Shriver and Kittles[60] in this issue). It should be noted, however, that the development and validity of AIMs requires knowledge of the genetic composition of the original populations that contributed to an individual's ancestry. If these populations are unknown or defined incorrectly, estimates of an individual's ancestry proportions can be inaccurate; for example, inferring ancestry proportions in Native Americans using AIMs for Europeans and Africans would be inaccurate as in fact Native Americans are more closely related to East Asians. AIMs are currently available for distinguishing among individuals from only a few populations (African, Asian, European and Hispanic Americans). Whether it would be useful or practical to develop AIMs to distinguish closely related groups is unclear.

Implications of group membership

Knowing the proportion of recent genetic ancestry that an individual shares with members of one or more groups facilitates making accurate predictions about disease susceptibility and outcome, and can increase the power of studies designed to find environmental and genetic factors that underlie health-related traits. Most polymorphisms that differ in frequency between groups are neutral, functionally insignificant and probably of little relevance to phenotypic differences between individuals. Functionally significant polymorphisms also commonly differ in frequency between groups, and these differences might be larger compared with neutral alleles as functional alleles are almost always exposed to natural selection[61]. However, the relationship between specific functional variants and differences in morphology and/or health-related traits varies considerably. The usefulness of ancestry information depends on the nature of this relationship.

Benefits of ancestry information to medical research. Inference about an individual's ancestry can make it easier to predict how likely an individual is to have a disease-causing variant. The high frequency of the *HbS* allele in sub-Saharan Africans and Southern Europeans or the *C282Y-HFE* and *Δ508-CFTR* alleles, which cause HAEMOCHROMATOSIS and CYSTIC FIBROSIS, respectively, in Northern Europeans are well known examples, but many others have been discovered[62]. In these examples — as well as for most variants with known phenotypic consequences — each polymorphism is highly PENETRANT and has a relatively large effect (that is, it underlies a monogenic trait). Mapping and cloning such variants using traditional linkage methods is relatively straightforward. Once such a variant is identified, screening an individual for it directly makes inference of ancestry unnecessary. In some instances, however, the phenotype that is associated with an allele that underlies a monogenic trait varies between groups[63]. This variation is the result of genetic and/or environmental modifiers that differ between groups, and further study of individual ancestry could help to identify these modifier loci. Most common health-related phenotypes, such as susceptibility to diabetes, obesity, infection and cancer, are COMPLEX TRAITS. The presentation, natural history and outcome of these conditions often vary among groups[64–66]. For example, the incidence of both prostate and breast cancers and the rate of death from these cancers are significantly higher in African–Americans than European– Americans[64]. Similarly, increased susceptibility to both obesity[67] and abnormal levels of insulin secretion[68] has been associated with higher proportions of individual African ancestry. Ancestry information might help to identify the basis of these differences (for example, through ADMIXTURE MAPPING).

If the CD/CV hypothesis is correct, even in part, many of the risk alleles that influence these traits will be shared among populations. Will each of these risk alleles have the same effect in individuals with different proportions of recent common ancestors? So far, there are few examples of alleles that underlie common diseases, but some early results foreshadow what we might find. An allele of apolipoprotein E (*APOε4*) that is frequent in Africans, Asians and Europeans is associated in a dose-dependent manner with susceptibility to Alzheimer disease. However, the increased risk that is associated with homozygosity for *APOε4* is ~5-fold higher in individuals with Asian rather than African ancestry[69]. In individuals infected with human immunodeficiency virus type 1 (HIV-1), several polymorphisms in the 5′ *cis*-regulatory region of *CCR5* influence the rate of progression to acquired immunodeficiency syndrome and death[70]. Some *CCR5* haplotypes are associated with delayed disease progression in different populations, but for others, the effect is population-specific[71]. One *CCR5* haplotype (*HHE*) has been associated with delayed disease progression in European– Americans, but accelerated disease progression in African–Americans[71]. Therefore, even if the same risk allele for a complex trait is present in different groups, it might be associated with different outcomes. For some common diseases, differences in individual susceptibility among groups seem to be determined, in part, by risk alleles in different genes. For example, 3 key variants in *NOD2* (now known as *CARD15*) — R702W, G908R and 1007fs — have been

HAEMOCHROMATOSIS
An autosomal recessive condition that is common in Western Europeans and their descendants. It is characterized by excessive iron absorption by the gut, with subsequent accumulation in the liver, heart, joints and pancreas.

CYSTIC FIBROSIS
An autosomal recessive condition that is common in Western Europeans and their descendants. It is characterized by pancreatic insufficiency and obstruction of the lungs by thick, heavy mucus.

PENETRANCE
The proportion of individuals with a specific genotype who manifest this genotype at the phenotype level.

COMPLEX TRAIT
A trait that is influenced by the environment plus a combination of polymorphisms in at least several genes, each of which has a small effect.

ADMIXTURE MAPPING
A strategy for mapping loci for complex traits that differ in prevalence between two populations that have recently admixed with each other.

associated with Crohn disease (CD), an inflammatory bowel disorder, in European–Americans[72,73]. None of these variants or other variants in *CARD15* have been associated with CD in African–Americans or Asians[74]. Conditioning phenotypic variation on individual ancestry proportions could help to identify risk alleles that are shared or that differ among populations and/or the basis of phenotypic differences among groups.

Are descriptors such as race or ethnicity alone sufficient to infer ancestry to identify susceptibility alleles or predict risk of disease or drug response? In some cases, the accuracy of these inferences might be adequate[70], but in many cases, the inexact measure of ancestry that is afforded by these proxies and/or the overlap of phenotypes across groups will lower the chances of finding susceptibility loci and lessen the predictive value of clinical inferences. Such descriptors will become even more inaccurate as human populations become more intermixed, and/or colloquial usages change over time. Should susceptibility be estimated, instead, by testing for disease-causing variants alone? Most disease-related alleles have yet to be identified, so inference of an individual's ancestry will continue to provide researchers and clinicians with information about risk if the direct cause(s) of a health-related trait is unknown. What then is the most reliable way to make inferences about individual ancestry? Self-reported ancestry can be obtained less intrusively than explicit genetic data, but in many cases — particularly in the United States and Europe — it is less reliable than using explicit genetic data. However, genetic testing remains relatively expensive, and genetic screening has already raised difficult issues of equity, privacy and consent[75].

Future work

To further understand the relationships among genetic variation, populations and health, several additional important questions need to be considered. What further basic research needs to be done to adequately characterize human genetic variation? How do we best design studies to identify whether and how genetic differences between groups contribute to health disparities? Will already exaggerated beliefs in innate differences between groups be reinforced by the study of specific ethnic and racial groups or the development of genetic tests and treatments for specific groups? How should geneticists educate other researchers, clinicians, policy makers and the public at large about the nuances of ethnicity and race and their complex relationship with the distribution of genetic variation? How will a better understanding of individual ancestry affect public policies? Answering each of these questions will be complicated, costly and protracted — so what should researchers do in the interim?

Investigators should first determine whether the hypotheses to be tested necessitate distinguishing groups; group membership should not be used as a proxy for a factor that can be measured directly. If distinguishing groups is required, it might sometimes be helpful to use the social, political and economic identities of race in studies designed to find environmental factors that influence health-related disparities.

Because biological information that is captured by common notions of race varies depending on how race is defined, studies designed to identify genetic factors associated with health-related traits might need to carefully explain how race was defined and used[76,77]. Different notions of race and ancestry might be useful depending on the circumstance. However, the information about geographical ancestry captured by concepts of race is, in general, less than that obtained by making ancestry inferences from explicit genetic data (such as genome-wide SNPs, AIMs), and, for much of humanity (for example, Hispanics, Asian Indians), race is not a meaningful descriptor of biological ancestry. Therefore, using varied definitions of race might only further conflate racial identities and geographical ancestry. How then do we make progress towards understanding the relationship between notions of race and geographical ancestry?

For most biomedical research applications, there are little empirical data that compare the reliability of self-reported ancestry or biogeographical ancestry with ancestry inferences from explicit genetic data. Studies are needed to directly test how different ways of making ancestry inferences perform in various types of study design and to explore the conditions (such as improvement in the accuracy of ancestry inference or its cost) that affect performance. Nevertheless, the type and amount of data required to make ancestry inferences will vary depending on the level at which group membership needs to be resolved (for example, African versus European or East African versus West African). In addition, although using explicit genetic data might now be costly and inconvenient, it is becoming increasingly less so each year.

In the absence of studies to provide further guidance, how should ancestry be inferred? One strategy might be to gather as much information about ancestry as possible. Subsequently, study subjects could be stratified by the various communities in which an individual is a self-assessed member beginning first with the descriptor that reflects the most information about ancestry, then the next most informative, and so on. In most cases, the most informative descriptor of ancestry will probably be the geographical origin of an individual's ancestors, followed by their ethnic identity, and finally, the community in which a person resides. This strategy might not differ much from current practices — particularly for individuals who know little about their origins — but it underscores the need to take account of biogeographical ancestry and it de-emphasizes the use of racial categories.

Conclusions

For hundreds of years, we have based inferences of individual ancestry on proxies, such as differences in physical appearance, language, or derivative concepts of ethnicity and race. None of these characteristics is

determined entirely by genetic or environmental factors, but separating out the relative contribution of each will often require sorting individuals into ancestral groups — a research effort that has created great controversy.

However, descriptors such as race or ethnicity capture only some of the ancestral information about the biological and environmental factors that influence phenotypic characteristics. In addition, the amount of information captured by each varies depending on how race or ethnicity is defined, the specific groups being studied, and how a study is designed and executed. It is time to devote some political will, experimental innovation and financial resources to address crucial questions about patterns of genetic variation, geographical ancestry and disease. At the same time, geneticists need to facilitate the incorporation of this new information into complicated frameworks of knowledge that we already have and attitudes that manifest as public perceptions of race.

1. Provine, W. B. Genetics and the biology of race crossing. *Science* **182**, 790–796 (1973).
2. Gould, S. J. *The Mismeasure of Man* (W. W. Norton Press, New York, 1981).
3. Lewontin, R. C. *Human Diversity* (Scientific American Books, Inc., New York, 1982).
4. Bamshad, M. J. & Olson, S. E. Does Race Exist? *Sci. Am.* **289**, 78–85 (2003).
5. Smedley, A. *Race in North America: Origin and Evolution of a Worldview*, 2nd edn (Westview Press, Boulder, 1999).
6. Lohmueller, K. E., Pearce C. L., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
7. Elliott, C. & Brodwin, P. Identity and genetic ancestry gracing. *B. Med. J.* **325**, 1469–1471 (2002).
8. Foster, M. W. & Sharp, R. R. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.* **12**, 844–850 (2002).
An introduction to some of the problems and challenges of trying to understand the relationship between the social definitions of populations and biologically defined groups.
9. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
10. Harpending, H. & Rogers, A. R. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genom. Hum. Genet.* **1**, 361–385 (2000).
A good review of genetic evidence on the evolution of modern humans as it pertains to some of the fundamental questions about human demographic history and the impact of natural selection.
11. Fischer, A., Wiebe, V., Paabo, S. & Przeworski M. Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* **5**, 799–808 (2004).
12. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
13. Nei, M. & Roychoudhury A. K. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am. J. Hum. Genet.* **26**, 421–443 (1974).
14. Mountain, J. L. & Cavalli-Sforza, L. L. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* **61**, 705–718 (1997).
15. Bowcock, A. M. *et al.* Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Natl Acad. Sci. USA* **88**, 839–843 (1991).
16. Jorde, L. B. *et al.* The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66**, 979–988 (2000).
17. Shriver, M. *et al.* The genomic distribution of population substructure in four popoulations using 8,255 autosomal SNPs. *Hum. Genomics* (in the press).
18. Watkins, W. S. *et al.* Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13**, 1607–1618 (2003).
19. Wilson, J. F. *et al.* Population genetic structure of variable drug response. *Nature Genet.* **29**, 265–269 (2001).
20. Turakulov, R. & Easteal, S. Number of SNPS loci needed to detect population structure. *Hum. Hered*. **55**, 37–45 (2003).
21. Ramachandran, S., Rosenberg, N. A., Zhivotovsky, L. A. & Feldman, M. W. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Hum. Genom.* **1**, 87–97 (2004).
22. Rosenberg, N., Li, L. M., Ward, R. & Pritchard, J. K. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* **73**, 1402–1422 (2003).
A comprehensive analysis of worldwide human microsatellite data that examines the amount of information that multi-allelic markers provide about individual ancestry.
23. Kittles, R. A. & Weiss, K. M. Race, ancestry, and genes: implications for defining disease risk. *Annu. Rev. Genom.* **4**, 33–67 (2003).
24. Bamshad, M. J. *et al.* Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* **72**, 578–589 (2003).
25. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
A comprehensive analysis of global patterns of human population structure. Its shows that although there is substantial geographical structuring among populations, the proportion of ancestry of many individuals from one or more of these populations is highly variable.
26. Bamshad, M. J. *et al.* Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11**, 994–1004 (2001).
27. Tishkoff, S. A. & Verrelli, B. C. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genom. Hum. Genet.* **4**, 293–340 (2003).
28. Malhotra, K. C. & Vasulu, T. S. in *Human Population Genetics* (ed. Majumder, P. P.) 207–232 (Plenum Press, New York, 1993).
29. Tajima, A. *et al.* Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. *J. Hum. Genet.* **49**, 187–193 (2004).
30. Merilä, J. & Crnokrak, P. Comparison of genetic differentiation at marker loci and quantitative traits. *J. Evol. Biol.* **14**, 892–903 (2001).
31. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nature Rev. Genet.* **3**, 611–621 (2002).
32. Zimmerman P. A. *et al.* Emergence of *FY*A*^{null} in a *Plasmodium vivax*-endemic region of Papua New Guinea. *Proc. Natl Acad. Sci. USA.* **96**, 13973–13977 (1999).
33. Bamshad, M. J. *et al.* A strong signature of balancing selection in the 5′ *cis*-regulatory region of *CCR5*. *Proc. Natl Acad. Sci. USA* **99**, 10539–10544 (2002).
34. Wooding, S. *et al*. Natural selection and molecular evolution in *PTC*, a bitter-taste receptor gene. *Am. J. Hum. Genet.* **74**, 637–646 (2004).
35. Haga, S. B. & Venter, J. C. FDA races in wrong direction. *Science* **301**, 466 (2003).
36. Lewontin, R. C. The apportionment of human diversity. *Evol. Biol.* **6**, 381–398 (1972).
37. Cavalli, L. L. & Piazza, A. Analysis of evolution: evolutionary rates, independence, and treeness. *Theor. Pop. Biol.* **8**, 127–165 (1975).
38. Jorde, L. B., Watkins, W. S. & Bamshad, M. J. Human population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* **10**, 2199–2207 (2001).
39. Wright, S. The genetical structure of populations. *Annu. Eugenics* **15**, 323–354 (1951).
40. Weir, B. S. & Hill, W. G. Estimating F-statistics. *Annu. Rev. Genom. Hum. Genet.* **36**, 721–750 (2002).
41. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
42. Long, J. C. & Kittles, R. A. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* **75**, 449–471 (2003).
43. Templeton, A. R. Human races: a genetic and evolutionary perspective. *Am. Anthropol.* **100**, 632–650 (1999).
44. Steele, F. R. Genetic 'differences.' *Genomics* **79**, 145 (2002).
45. AAA. American Anthropological Association Statement on Race. *Am. Anthropol.* **100**, 712–713 (1999).
46. Office of Management and Budget. *Revisions to the standards for the classification of federal data on race and ethnicity* [online], <http://www.whitehouse.gov/omb/fedreg/ombdir15.html> (1997).
47. Edwards, A. W. Human genetic diversity: Lewontin's fallacy. *BioEssays* **25**, 798–801 (2003).
48. King, M. & Motulsky, A. G. Mapping human history. *Science* **298**, 2342–2343 (2002).
49. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 199–204 (2001).
50. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 22225–2229 (2002).
51. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genom*. **1**, 218–224 (2004).
52. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
53. Crawford, D. C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74**, 610–622 (2004).
54. Shriver, M. D. *et al.* Skin pigmentation, biogeographical ancestry, and admixture mapping. *Hum. Genet.* **112**, 387–399 (2003).
A clear example of how estimates of individual ancestry proportions can be used to identify genotypes that influence phenotypes that differ between populations.
55. Hoggart, C. J. *et al.* Control of confounding in genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
56. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population structure using multilocus genotype data. *Genetics* **155**, 945–959 (1999).
57. Hinds, D. A. *et al.* Matching strategies for genetic association studies in structured populations. *Am. J. Hum. Genet.* **74**, 317–325 (2004).
58. Parra, E. J. *et al.* Estimating African–American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
59. Frudakis, T. *et al.* A classifier for the SNP-based inference of ancestry. *J. Forens. Sci.* **48**, 1–12 (2003).
60. Shriver, M. D. & Kittles, R. A. Genetic ancestry and the search for personalized genetic histories. *Nature Rev. Genet.* **5**, 611–618 (2004).
61. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111 (2003).
62. Martinson, J. J., Chapman, N. H., Rees, D. C., Liu, Y. T. & Clegg, J. B. Global distribution of the *CCR5* gene 32-basepair deletion. *Nature Genet.* **16**, 100–103 (1997).
63. Hardy, J., Singleton, A. & Gwinn-Hardy, K. Ethnic differences and disease phenotypes. *Science* **300**, 739–740 (2003).
64. Wiencke, J. K. Impact of race/ethnicity on molecular pathways in human cancer. *Nature Rev. Cancer* **4**, 79–84 (2003).
65. Holden, C. Race and medicine. *Science* **302**, 594–596 (2003).
66. Yancy, C. D. Does race matter in heart failure. *Am. Heart J.* **146**, 203–206 (2003).
67. Fernandez, J. R. *et al.* Association of African genetic admixture with resting metabolic rate and obesity among women. *Obes. Res.* **11**, 904–911 (2003).
68. Gower, B. A. *et al.* Using genetic admixture to explain racial differences in insulin-related phenotypes. *Diabetes* **52**, 1047–1051 (2003).
69. Farrer, L. A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* **278**, 1349–1356 (1997).
70. Martin, M. P. *et al.* Genetic acceleration of AIDS progression by a promoter variant of *CCR5*. *Science* **282**, 1907–1911 (1998).
71. Gonzalez, E. *et al.* Race-specific HIV-1 disease-modifying effects associated with *CCR5* haplotypes. *Proc. Natl Acad. Sci. USA.* **96**, 12004–12009 (1999).

Author biographies

Michael Bamshad is an associate professor in the Departments of Paediatrics and Human Genetics at the University of Utah School of Medicine, USA. His research interests focus on population genetics, and specifically on the extent of human population structure and the impact of natural selection on disease-related genes. He has also worked extensively to identify the genes that underlie malformation syndromes, and particularly those that affect the limb.

Stephen Wooding is a postdoctoral fellow in the Department of Human Genetics at the University of Utah School of Medicine, USA. His research interests focus on molecular evolution and population genetics, and especially coalescence theory. He has a particular interest in the molecular evolution of environmentally responsive genes.

Benjamin Salisbury leads the Computational Genomics Group at Genaissance Pharmaceuticals, Connecticut, USA. His doctoral and postdoctoral work at the University of Michigan, USA, and at Yale University, USA, respectively, concerned the development of new frameworks for phylogenetic evolutionary analysis.

J. Claiborne Stephens is Vice President of Genetics at Genaissance Pharmaceuticals, Connecticut, USA, and comes from a background in population genetics and molecular evolutionary genetics. He has had previous posts at the National Cancer Institute, Maryland, USA, the Howard Hughes Medical Institute at Yale University, USA, and the Center for Demographic and Population Genetics at the University of Texas Health Science Center (UTHSC), Houston, USA.

Online summary

- Highlighting genetic differences among people could unfortunately reinforce stereotypical features of populations, but exploring the genetic influence on common health-related traits and disparities could also be beneficial to human health.
- Accurate inference of an individual's ancestry using genetic data depends on several factors, including the number of genotypes used, the degree of differentiation among groups and how each group is sampled.
- Inferences of human population structure based on genetic data often differ from inferences based on phenotypic characteristics.
- Although there might be little variation among groups, it is highly structured and therefore useful for distinguishing groups and allocating individuals into groups.
- Insofar as geographical ancestry corresponds to some notions of race, patterns of genetic variation will also co-vary with these notions.
- The inaccurate measure of ancestry afforded by proxies of genetic relationships such as race or ethnicity can sometimes be useful, but in other circumstances, might lower the chances of findings disease-susceptibility loci and lessen the predictive value of clinical inferences.

Online links/Supplementary data

Entrez

CCR5
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=1234

NOD2 (*CARD15*)
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=64127

TAS2R38
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene&cmd=Retrieve&dopt=Graphics&list_uids=5726

OMIM

Alzheimer disease
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=104300

Crohn disease
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=266600

cystic fibrosis
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=219700

haemochromatosis
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=235200

HIV-1
http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=143055

Further information

CEPH Human Diversity Panel
http://www.cephb.fr/HGDP-CEPH-Panel

Supplementary information

S1 (data)
S2 (data)
S3 (data)
S4 (data)
S5 (data)
S6 (data)

CFI statement (Web only)

Benjamin A. Salisbury and J. Claiborne Stephens are both employees and shareholders of Genaissance Pharmaceuticals, Inc.