46. Qureshi, N., Hapgood, R. & Armstrong, S. Continuous medical education approaches for clinical genetics: a postal survey of general practitioners. *J. Med. Genet.* **39**, e69 (2002).
47. Bankhead, C. *et al.* New developments in genetics — Knowledge, attitudes and information needs of practice nurses. *Fam. Pract.* **18**, 475–486 (2001).
48. Metcalfe, S., Hurworth, R., Newstead, J. & Robins, R. Needs assessment study of genetics education for general practitioners in Australia. *Genet. Med.* **4**, 71–77 (2002).
49. Mitchell, J. J., Capua, A., Clow, C. & Scriver, C. R. Twenty-year outcome analysis of genetic screening programs for Tay–Sachs and β-thalassemia disease carriers in high schools. *Am. J. Hum. Genet.* **59**, 793–798 (1996).
50. Vermeulen, E. G. *et al.* Effect of homocysteine-lowering treatment with folic acid plus vitamin B6 on progression of subclinical atherosclerosis: a randomised, placebo-controlled trial. *Lancet* **355**, 517–522 (2000).
51. Harrison, T. A. *et al.* Family history of diabetes as a potential public health tool. *Am. J. Prev. Med.* **24**, 152–159 (2003).
52. Gottesman, M. M. & Collins, F. S. The role of the human genome project in disease prevention. *Prev. Med.* **23**, 591–594 (1994).
53. Penrose, L. S. *Outline of Human Genetics* (Heinemann, London, 1959).
54. Christianson, A. & Modell, B. Medical genetics in developing countries. *Annu. Rev. Genomics Hum. Genet.* **5**, 219–265 (2004).
55. Department of Health, UK Department of Health Genetics White Paper — Our inheritance; realising the potential of genetics in the NHS <http://www.dh.gov.uk/assetRoot/04/01/92/39/04019239.pdf> (Department of Health, London, 2003).

Acknowledgements

We thank Maren Khan for allowing us to use her data in figure 3. N.Q. received a proportion of his funding from the NHS Research and Development levy and from the Commonwealth Fund, New York; the views expressed in this publication are those of the authors and not necessarily those of the NHS executive or the Commonwealth Fund, its directors, officers or staff.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

International database on the legal, social and ethical aspects of human genetics, maintained by the University of Montreal: http://www.humgen.umontreal.ca/en/

The following terms in this article are linked online to:

OMIM:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
Phenylketonuria | cystic fibrosis | Down syndrome | Tay–Sachs | thalassaemias

FURTHER INFORMATION

Accessible Publishing of Genetic Information:
www.chime.ucl.ac.uk/ApoGI/

Cambridge Genetics Knowledge Park, Public Health Genetics Unit: Education in genetics for health professionals:
http://www.phgu.org.uk/about_phgu/education.html

Contact a Family: http://www.cafamily.org.uk/

Ethical, legal and social issues (ELSI) surrounding the availability of genetic information:
http://www.ornl.gov/sci/techresources/Human_Genome/elsi/elsi.shtml

Genetics Resources on the Web:
http://geneticsresources.org/

National Coalition for Health Professional Education in Genetics: http://www.nchpeg.org/

NHS Sickle Cell & Thalassaemia Screening Programme:
http://www.kcl-phs.org.uk/haemscreening/Policy.htm

UK newborn screening programme centre:
www.ich.ucl.ac.uk/newborn

US National Newborn Screening and Genetic Resource Center: http://genes-r-us.uthscsa.edu/index.htm

US Office of Genomics and Disease Prevention:
http://www.cdc.gov/genomics/

Victorian (Australian) Government Genetic Health Education Resources:
http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/hc_geneticissues_general?OpenDocument

Access to this links box is available online.

OPINION

Beyond race: towards a whole-genome perspective on human populations and genetic variation

Morris W. Foster and Richard R. Sharp

Abstract | The renewed emphasis on population-specific genetic variation, exemplified most prominently by the International HapMap Project, is complicated by a longstanding, uncritical reliance on existing population categories in genetic research. Race and other pre-existing population definitions (ethnicity, religion, language, nationality, culture and so on) tend to be contentious concepts that have polarized discussions about the ethics and science of research into population-specific human genetic variation. By contrast, a broader consideration of the multiple historical sources of genetic variation provides a whole-genome perspective on the ways i n which existing population definitions do, and do not, account for how genetic variation is distributed among individuals. Although genetics will continue to rely on analytical tools that make use of particular population histories, it is important to interpret findings in a broader genomic context.

A prominent public message from the sequencing of the human genome is the 99.9% genetic similarity that is present among all individuals[1,2]. By contrast, post-sequence efforts to discover genes that contribute to disease susceptibility and drug response have emphasized human genetic variation, often characterized as varying by population[3,4]. The emphasis on how we differ from one another genetically (which has, in fact, been a focus of research throughout the history of the study of human genetics), has revived a long-standing debate about the biological importance of particular historical categories, such as race and ethnicity, that are frequently used to define membership of specific populations[5].

In the United States, much of this debate has centred on the biological meaning of race[6–8], an historically contentious concept that has polarized what might otherwise be a more nuanced consideration of the distribution and structure of genetic differences among humans. This polarization is not surprising in light of the importance that the public attaches to race. As a prominent way of defining population membership over the past 500 years, race has been used to advantage some groups over others. For that reason, race should not, and cannot, be avoided in considerations of issues such as access to care, exposure to environmental hazards and preferences regarding clinical interventions. However, when used to define populations for genetic research, race has the potential to confuse by mistakenly implying biological explanations for socially and historically constructed health disparities[9].

At the same time, some disparities in individuals' disease risks are the consequence of genetic differences among those individuals. New genomic tools are allowing scientists to explore inter-individual biological differences[10]. Each of these differences has its own history, which arises from an initial mutation event and is amplified through subsequent demographic and evolutionary processes that result in the inheritance of these differences at varying frequencies in different groups of individuals, at different points in time[11]. Often, however, the multiple histories that have contributed to each individual genome are subsumed in a more limited number of ancestries that are taken as defining human populations, usually as racial and ethnic groups. Here lies the problem — substituting a small number of racial and ethnic definitions of populations for multiple histories of genetic variants might lead to confusion about the biological significance of the former. Categorical population identities (such as African or European) imply single origins and mutually exclusive distributions for genetic variants. This complication has become especially problematic as population history is emerging as a powerful tool in using the completed human genome sequence to discover disease susceptibility and drug-response genes.

Although race tends to be less of an issue outside the United States, various pre-existing population definitions (including ethnicity, cultural affiliation, religion and nationality) are used by researchers throughout the world to identify participants of genetic studies. Both the general public and scientific researchers tend to categorize themselves and others as members of populations defined in historically selective ways; in other words, using categories and labels that have emerged from ongoing social processes rather than scientific investigation. Consequently, the population definitions that we commonly use today, and to which scientists often attribute genetic findings, are the same ones that have been the primary basis for establishing and maintaining socio-economic and other disparities for centuries. The ways in which we define populations can have significant implications for how we interpret the scientific meaning of genetic findings. Investments in genomic infrastructure such as the International HapMap Project and PROSPECTIVE COHORTS will soon generate large amounts of data that will allow complex quantitative analyses of human variation across the genome[12]. However, the sophistication of those analyses might be blunted if scientists continue to rely, without criticism, on pre-existing, selective and historically encumbered population definitions. This is primarily a conceptual, rather than analytical, problem that is rooted in how we are accustomed to think about individuals as members of groups. Emphasizing the multiplicity of historical contributions to inter-individual genetic variation will allow us to reconsider how we define populations in studies that attempt to discover genes and genetic variants that contribute to complex diseases and drug responses. It might also help us think differently about the ways in which we attribute biological significance to populations in which individuals are considered members (either by virtue of self-identification or by assignment of membership by others) and the ways in which genetic research on a population defined in one manner might, or might not, benefit populations defined in other manners.

Variant frequencies and gene discovery

Many studies looking for genes that contribute to disease risk and drug response use MICROSATELLITE or SNP markers that are randomly spaced throughout the genome (with the important exception of candidate gene studies, which use data from animal models and other studies to narrow genetic analysis to a particular gene, region or network of functionally related genes). In these whole-genome association studies, DNA samples are genotyped with the aim of identifying locations in which samples from people with a given disease or drug response show a statistically significant association with a particular marker (sometimes in comparison with DNA samples from individuals who do not have the disease or response). Typically, the same set of markers is used for all individual samples regardless of a donor's population identity or ancestry, or the disease being studied.

As we learn more about different sources of human genetic variation (or, perhaps more accurately, of differences in the frequencies of human genetic variants), it becomes possible to construct genomic resources that reflect historical patterns in sequence variation, both within and across populations. Unlike randomly spaced markers, these resources make use of our growing knowledge of how the landscape or architecture of the human genome varies between individuals and of how frequently those variations occur in a group of people that are defined in a particular way.

> "…when used to define populations for genetic research, race has the potential to confuse by mistakenly implying biological explanations for socially and historically constructed health disparities."

The most promising of these resources is the HapMap[12]. A haplotype is a linear sequence on the same chromosome (usually 100 to 100,000 bases) that is identifiable as one of a number of alternative structures for a given segment of that chromosome (by virtue of there being different SNPs for some bases in that segment). As such, nearby alleles in the same haplotype are inherited together more often than is expected by chance, a phenomenon that is called linkage disequilibrium (LD). Because of this co-inheritance, haplotypes offer a way of summarizing the genetic variation that is found within them. Using resources like HapMap, researchers need identify only a few SNPs (called haplotype-tagging SNPs or htSNPs) to identify the longer haplotype, which greatly reduces genotyping costs for gene discovery[13].

Once completed, the HapMap can be used to look for patterns or similarities in inter-individual genetic variation among those affected by a disease (that is, as an association study). The rationale for this strategy is that those similarities (most of which will not contribute to the disease or drug response in question) will allow researchers to narrow the possible chromosomal regions in which a susceptibility gene might lie[14,15]. htSNPs that are known to reliably predict parts of common haplotypes can be tested for their association with increased susceptibility to a disease, with the idea that the actual susceptibility gene(s) will be located near to the associated marker(s). Other techniques, such as positional cloning, can then be used to identify the gene, or genes, in question[16].

The associations that result from such studies depend on the existence of LD to generate a positive signal in a given set of DNA samples that include both affected individuals and matched healthy controls. However, not all alleles that co-occur more often than is expected by chance lie in regions that are on the same haplotypes[17]. Furthermore, because common haplotypes are conserved across populations, it is more efficient to use a haplotype map to harness LD for gene discovery than it is to attempt to catalogue or map all regions of LD (or, for that matter, to catalogue all SNPs). Information about haplotype structure and frequencies will allow researchers to improve the statistical power of association studies in a more cost-effective way; in other words, to get greater power for a given sample size, while genotyping fewer bases[18].

To work efficiently, a genomic resource such as a haplotype map should comprise common genetic variants that will be found in the DNA of most participants in a given genetic study. That degree of commonality, however, will depend on how the population for which it is calculated is defined. Interestingly, scientists tend to use pre-existing population definitions (such as African-American and European) to recruit and label study participants (who often identify themselves as members), whereas clinicians use these same definitions as indications that a patient might, or might not, have a particular disease or drug response. These pre-existing categories are not defined genomically (that is, by the degree to which individuals share genetic variants such as SNPs or haplotypes) nor are they defined strictly by ancestry (which would entail a range of gradations in combined ancestries, albeit combinations that would tend to cluster in particular patterns across geographical space). Rather, they are defined by complex social and historical

processes that have resulted in a small number of mutually exclusive (although often situational) ways of distinguishing one group from other groups. Although each of these pre-existing population categories has certain frequencies of genetic variants, it is important to keep in mind that the ways in which individuals are grouped together determine the genetic frequencies that are attributed to such populations, not that genetic frequencies determine how to group individuals into populations.

The alchemy of race

Despite the widespread use of categorical definitions of populations in human genetic studies, today's global genetic variation is the result of a combination of different historical sources of inter-individual differences (BOX 1). Consequently, variation among individuals is not coincident with any single population or ancestry, or with any single way of defining a population or ancestry. However, researchers interpret multiple historical sources of genetic variation in two contrasting ways: either as the basis for patterns of variation that are common across populations or as the basis for biological differences among populations. The former view has been developed as an argument for identifying globally common patterns to be used as resources (such as the HapMap) for gene discovery in all populations, whereas the latter view has been taken as support for the importance of using population-specific patterns for gene discovery. Support for each of these views depends on how populations are defined, a question that has, not surprisingly, become focused on the problem of race in the United States and on other pre-existing ways of defining populations elsewhere.

Advocates for an effort to map common haplotypes across populations argue that doing so will enable researchers to identify 80–85% of all human genetic variation[19,20]. They also argue that differences in haplotype structure between, for example, European and African populations can be used to link

Box 1 | **Historical sources of genetic variation**

Previous attempts to conceptualize human genetic variation, such as the Human Genome Diversity Project (see Online links box), have tended to rely on contemporaneous tribal, cultural, linguistic, racial and ethnic boundaries to summarize inter-individual genetic differences. However, the use of pre-existing population definitions has led scientists to think of each social identity as representing a distinctive and often unique biological history[46]. By contrast, we can think of the genetic variants in each individual's genome as having multiple historical sources that have crossed categorical population boundaries throughout human history.

Original variation

Humans who lived in Africa before any migrations to other continents developed a range of variant haplotypes, SNPs and alleles that constitute the original range of genetic variation among humans. Some, but not all, of these original variants were carried by migrants out of Africa and are found today in populations throughout the world. For instance, a study of DNA samples from 52 populations worldwide found that intra-population differences account for 93–95% of genetic variation, whereas differences between populations account for only 3–5% of variation[47] — evidence for the continuing influence of that original, pre-migration human variation. Some SNPs and, in particular, some disease-susceptibility and drug-response alleles were also part of the pre-migration range of variation, which is the basis for the COMMON DISEASE/COMMON VARIANT HYPOTHESIS[3,48,49]. Therefore, a resource, such as HapMap, that identifies patterns of variation common to most, or all, populations should be useful for discovering genetic variants that are also common in those populations. An allele of the *APOE* gene that increases risks for Alzheimer's disease and heart disease, is an example of a common variant present in populations throughout the world[48].

Continental diaspora

The initial migration out of Africa to other continents (beginning approximately 100,000 years ago) occurred as multiple, branching events and involved many FOUNDER EFFECTS in which rare haplotypes, SNPs and alleles appear to have increased in frequency in emigrant populations owing to genetic drift and different selection pressures. The founder effects, and particularly the decrease in genetic diversity resulting from continental migrations, also increased haplotype length in the emigrant populations[20]. Therefore, individuals who live in Africa today tend to have shorter haplotypes, because they usually have ancestors who have experienced more recombination events without population bottlenecks or founder effects. Individuals who live in Europe, however, tend to have longer haplotypes because they usually have ancestors who have experienced fewer recombination events, as humans migrated to Europe from Africa, bringing with them less variation in SNPs, alleles and haplotypes than there were among individuals who remained in Africa. Because of their historical relationship (and also perhaps because recombination events might preferentially occur at 'hotspots'), the shorter haplotypes are often subdivisions of the larger haplotypes, so that common African haplotypes can be correlated with common non-African haplotypes[50]. The shared historical boundaries of common haplotypes also make it possible to correlate those from different non-African populations with one another. Furthermore, the continental diaspora might also be the basis for the existence of multiple genetic variants that contribute to susceptibility for the same common disease but that vary considerably in frequency among people with differing continental ancestries[51–53], thereby requiring the development of differing tag SNPs for each population or sample studied[24,25].

Subsequent demographic events

Separation from a 'parental' population also meant that new recombination and mutation events in an emigrant population on one continent might not necessarily have been passed on to 'sibling' populations on other continents or, if so, might not necessarily have become common genetic features in the other populations. Historical patterns of intra-group mating create demographic circumstances that can amplify the frequencies of some variant features, especially as these groups experience changing circumstances that lead to such phenomena as population bottlenecks, genetic drift, admixture with other such groups and different selection pressures[54]. As with initial continental migrations, these subsequent demographic events have also created conditions in which some rare or new haplotypes, SNPs and alleles have become more common in some groups and their descendants, and not others[55]. Colon cancer has been suggested as an example of a common disease for which there might be multiple rare susceptibility variants that vary by ancestry[53]. Recent demographic events often have stronger, more easily detectable effects on the inter-individual distribution of haplotypes, SNPs and alleles because the genetic features involved will probably continue to be transmitted together from one generation to another (in other words, to provide evidence for linkage disequilibrium). Moreover, recently mutated SNPs and alleles and recently recombined haplotypes will probably be restricted to members of the same socially bounded groups, because there has been less time for them to be passed on to other such groups, or for the social dynamics that hold the initial groups together to dissipate. To the extent that these relatively young, geographically restricted rare variants contribute to disease susceptibility or drug response in one or a small number of populations, investments in genomic infrastructure that focus on common variation across populations (like HapMap) will be of limited benefit.

results of genetic studies of the same disease in different populations[21,22]. By contrast, others argue that migrational and subsequent demographic events make global catalogues and maps too imprecise for such studies so that haplotype and SNP analyses from one population cannot readily be extrapolated to others[6,8,23,24]. Emphasis on common variants might also not allow researchers to identify rarer variants that contribute to disease susceptibility in some populations but not in others[25,26].

Certainly, the migration of humans out of Africa created demographic conditions in which new and rare haplotypes, SNPs, LD regions and alleles became more frequent in some populations and not in others. So, for example, some researchers have identified particular non-coding SNPs as 'ancestral' from the original populations of Africa, Europe or Native America owing to their higher frequencies in contemporary populations that reside in or have recently migrated from those continents[27]. This approach, called ADMIXTURE mapping, argues that ancestral SNPs are more sensitive as markers for gene discovery than globally common variants because they take into account changes in frequencies of variation that occurred as a result of the continental migration[28]. Some scientists also point out that these continental variants correspond with racial categories used in epidemiological analysis[6,8], indicating that race is a valid (and, indeed, unavoidable) reflection of biological differences that result from the migration history of human populations out of Africa.

However, race and other history-related ways of defining identity might not be such comprehensive proxies for human population history as some would believe. Today, each individual, regardless of social identity or continent of residence, carries variant haplotypes, SNPs and alleles that have been generated in the context of the historical sources of variation experienced by all of his or her direct ancestors. Therefore, each of us has, in his or her personal genome, evidence of the initial genetic diversity that was present when humans lived only in Africa, evidence of at least one, and probably more than one, migration out of Africa to another continent (and perhaps of migrations from that continent to others), as well as evidence of many subsequent episodes of population expansions, contractions, and partial or temporary periods of reproductive isolation[29]. In addition, each of us is a member of one or more socially defined populations that pre-exist scientific investigation, in which these and other variant genomic features are shared among members at particular frequencies. These contemporary frequencies are also consequences of the various historical processes that all the ancestors of current members have experienced. In this sense, all populations and individual members are admixed.

"…race and other history-related ways of defining identity might not be such comprehensive proxies for human population history as some would believe."

Interestingly, however, scientists tend to characterize some populations as more historically heterogeneous and others as more homogeneous. For example, African-Americans are usually considered to be members of an 'admixed' population, and their DNA is analysed in a way that takes advantage of individuals having multiple ancestries[30]. In part, this is because much of the genetic literature on admixture analysis has been developed primarily through studies of African-Americans. At the same time, however, the social history of the concept of race has played a part in how scientists have defined and analysed different populations. For more than three centuries, anyone of mixed African and European ancestry in the American colonies, and (later) the United States, was legally categorized as African-American. Although there are no longer those legal requirements, a cultural practice of defining African-American populations as 'admixed' and European-American populations as 'outbred' clearly continues among genetic researchers[31]. Similarly, Native American tribes have often been characterized as 'isolated' or 'inbred' populations in genetic studies of their contemporary members[32,33], despite a history of extensive inter-marriage[34]. This characterization is a consequence of the way in which European-Americans have defined tribal membership by 'BLOOD QUANTUM' and romanticized Native identities and communities as self-contained historical entities[35].

African-Americans are 'admixed', European-Americans are 'outbred', and Native Americans are 'isolated' or 'inbred' populations for purposes of genetic analysis, primarily because that is how we have chosen to define membership in these categories, not because of intrinsic genomic structures or variant frequencies that made them such. These characterizations reveal the conceptual limitations that are inherent in using existing population definitions for genetic research without criticism and without carefully considering the social histories of those categories. Existing definitions bring with them several levels of meanings that are 'alloyed' into the same categories that even the purest science cannot transmute into gold standards for population genetics. Instead, without analytical investigation of their non-biological histories, they are only fool's gold.

Are populations unavoidable?

If race and other existing ways of defining populations are problematic as categories for genetic analysis, why bother trying to define populations for the purpose of recruiting participants for genomic resources and genetic studies?

Despite the problems, population identities can be heuristic proxies for identifying and approximating diversity in assembling genomic resources[5]. For example, SNP analysis has found that an existing public database contains nearly 80% of all common SNPs when tested using a sample of people of European ancestry, but only 50% when tested using an African-American sample[36]. This finding says less about the actual degree of overlap in common SNP frequencies between people of European ancestry and African-Americans than it does about the potential participant recruitment bias in the multiple studies that have contributed to the public SNP database. This disparity, with respect to inclusiveness, could have consequences for the discovery of disease-susceptibility and drug-response genes that are more frequent among African-Americans. The existing concepts of population identity become more problematic when genetic findings are used to characterize a specific population or to compare two or more populations. For instance, although informative about the inclusiveness of genomic resources, treating people of European ancestry and African ancestry as separate categories for genetic characterization tends to contribute to a public perception that the primary difference between these existing ways of defining populations is biological.

Nonetheless, most genetic studies of specific diseases and drug response do recruit participants using existing population identities. The frequencies of genetic variants found in populations defined in a particular way is one reason why such studies are often designed for participants who share a common contemporary identity or ancestry.

Researchers use those social identities as proxies to take advantage of demographic and evolutionary processes that can make a pattern or variant more evident, such as POPULATION BOTTLENECKS, which often create strong LD effects in subsequent generations. Such historical effects, and the population or ancestral identities that are proxies for them, can be useful in detecting genetic contributors to complex diseases that have proven difficult to investigate. In addition, population-specific studies can be used to confirm that a variant discovered in a population defined in one way also contributes to a disease or drug response in other populations that are defined in different ways. For downstream clinical applications, replicating and confirming a finding for different populations might be as important as obtaining the finding in the first place.

Researchers also use population identity to avoid problems of POPULATION STRATIFICATION in association studies[37,38]. Population stratification can result either in false-positive or false-negative findings, and is often given as the reason for the inability to replicate a positive association with participants recruited from differently defined populations[39]. However, because pre-existing population definitions are imperfect proxies for a shared biological history, self-reported or researcher-ascribed identities will not necessarily minimize genetic heterogeneity in recruiting participants[40]. For that reason, trio designs (in which DNA samples from parents and a child are genotyped, and the chromosomes that each parent did not pass on to the child are used as controls) and genome controls (in which randomly spaced markers are genotyped across the genome to determine the degree of common ancestry) are increasingly being used to confirm, or even replace, social identities in ascertaining population substructure.

However, neither trios nor genome controls can ensure that association-study participants with similar biological ancestries also have similar environmental exposures. If both genetic and environmental contributors are necessary for a phenotype to be expressed, cases and controls that are matched for allele frequencies, but are not matched for environmental exposures, can also result in spurious findings, potentially confusing the effects of environmental contributors with those of genetic contributors[41]. Moreover, differences in environmental contributors between two or more studies of the same disease can also hinder the detection of genetic effects, as well as limit the ability of researchers to confirm findings from an association study in one population in other populations. Although population identities are often used as proxies for environmental exposures, particularly in comparisons between populations, they can also conceal a considerable range of intra-group environmental variation, especially in the case of racial and ethnic categories that have many members.

Oddly, however, a saving grace for many genetic studies is that their participant recruitment is on a smaller scale than the social labels they use to identify these participants. Largely for reasons of logistics and expense, most genetic studies recruit participants from a small number of geographically proximate local sites or even just a single site (such as a hospital). Many studies also ask affected participants to help identify an appropriate unaffected control, which has the effect of limiting the latter to the social networks of the former. Consequently, most such studies already benefit from some *de facto* control over variation in environmental contributors between cases and controls. The localization of participant recruitment might also help control population stratification, to the extent that participants recruited from particular sites probably share more of the same multiple ancestries, which would mitigate some of the considerable genetic variation that lies in broad categories such as African-American or European-American.

Such *de facto* measures, however, cannot always be relied on even when recruitment is localized. In addition, they do nothing to mitigate broad associations of large-scale racial and ethnic identities with particular genetic findings, such as the recent claim that African-Americans do not respond to existing drugs for heart disease[7]. Moreover, as increasingly larger national and international cohorts are formed for genetic research, any implicit advantages from localized recruiting will disappear. Closed-end surveys that are often used to standardize ancestral and environmental information in large cohort studies are cost-efficient, but will probably provide less information about the multiple histories that shaped each individual participant's genome, or the local lifestyles and ecologies that contributed to each participant's contacts with toxins. Consequently, although larger cohorts with members who have diverse ancestries might, in fact, minimize effects of population stratification in association studies[40], discovery of contributing genetic variants or environmental factors that vary by ancestry or locality will be less likely.

A whole-genome perspective

Population-specific investigations of frequencies in common patterns of genetic variation, and of common variants discovered in some populations, have considerable potential for benefiting other populations, even in certain cases of rare diseases and drug responses, in which the contributing genes are located near common haplotypes[42]. However, progress in confirming gene discoveries, explaining functional mechanisms, and developing diagnostic tests and therapies will depend, in some considerable part, on how we can relate the findings from one study to those from others; that is, on how we are able to conceptualize different ways of defining human populations. More than one definition of ancestry or population will be necessary to account for all the variants that an individual's entire genome holds. Rather than treat populations as categorical entities into which individuals

Glossary

ADMIXTURE
Gene flow between differentiated populations.

BLOOD QUANTUM
A legal measure of degree of Native American ancestry. The designation of 'full blood' or some fraction such as 'quarter' or 'half' blood quantum depends on how one's nineteenth century Indian ancestors were designated (often arbitrarily) and the blood quanta of subsequent ancestors who were enrolled in federally recognized tribes.

COMMON DISEASE/COMMON VARIANT HYPOTHESIS
The view that one or a few genetic contributors account for significant numbers of cases of many common, complex diseases in most or all populations.

FOUNDER EFFECT
A situation in which a new population is founded by a small number of individuals. Similar to a bottleneck, the founder effect severely reduces genetic diversity, increasing the effect of random drift.

MICROSATELLITE
A class of repetitive DNA sequences that are made up of tandemly organized repeats that are 2–8 nucleotides in length. They can be highly polymorphic and are frequently used as molecular markers in population genetics studies.

POPULATION BOTTLENECK
A marked reduction in population size followed by the survival and expansion of a small random sample of the original population.

POPULATION STRATIFICATION Subdivision of a population into different subgroups with potentially different marker allele frequencies and different disease prevalences. This might result in participants with a disease having different allele frequencies than those without the disease that are recruited as controls.

PROSPECTIVE COHORT
Longitudinal study of individuals initially assessed for exposure to certain risk factors and then followed over time to evaluate the progression towards specific outcomes (often disease).

are placed, it might be better to think of individuals as having multiple ancestries that indicate several ways in which they might be grouped with others. The different ways of defining populations can be used by researchers to take advantage of different histories that affect contemporary frequencies of phenotypes, genetic patterns and genetic variants, as well as to devise various control methods for population stratification and environmental factors that might affect the accuracy and replicability of associations.

Some geneticists have begun to work towards such a perspective. A research design that uses individuals from different populations that tend to have longer and shorter common haplotypes for the same chromosomal regions has been proposed as a method to initially indicate and then increasingly refine and confirm regions of association[21]. A study of variable drug response has shown that individuals with the same racial, ethnic or other identity can be assigned to different genetic clusters that predict more accurately their ability to metabolize a drug than their shared, pre-existing population label[19]. At the same time, however, individuals who differ with respect to drug metabolism might be assigned to the same genetic cluster at a different level of resolution for other purposes, such as a study of genetic susceptibility for a particular disease.

Most scientists, however, have not yet developed robust conceptual tools to appreciate a fully genomic perspective on genetic variation. Instead, most researchers continue to rely on the same racial, ethnic and other ways of defining populations and ancestries[43]. Both scientists and the general public are accustomed to associating genetic variation with specific pre-existing social identities and ancestries. For example, arguments have been made for associating haplotypes with racial identities (based on differences in haplotype length and composition resulting from continental migrations)[6] and non-coding SNPs with a degree of racial admixture (also based largely on mutations that might be unique to continental migrations)[27]. Mitochondrial DNA and Y-chromosome DNA have also been analysed to associate the individuals who carry them with specific racial or ethnic ancestries[44]. The conceptual difficulty is that each of these associations involves selected genetic features that are located in relatively small segments of the genome, rather than the perspective of the whole genome of an individual[45].

Questions about the biological meaning of race (which is often conflated with the continental population migration), or of ethnic or linguistic groups (which are often conflated with subsequent demographic events), are narrowed to a limited number of genetic features that, when carefully chosen, might have a greater probability of co-variance with specific social identities. However, these selective types of questions necessarily ignore the full range of sources of genetic variation in every individual and population (regardless of how the latter is defined). In the end, the argument about the value of racial and ethnic classifications, and their role in the construction of genomic resources, is too narrow to fully appreciate the complexities of human genetic variation.

By choosing a limited number of ways of defining populations as the basis for compiling information on frequencies of genetic patterns and variants, researchers might be seen to be endorsing the selective categorizations of people, their identities and their ancestries as being more scientifically important or valid than other ways of doing so. Such scientific endorsements could indicate that these population definitions are more fundamental than other ways of defining populations; that genetic definitions of these 'fundamental' populations are more essential than other ways of defining membership in them; and that social identities such as race and ethnicity are ahistorical and mutually exclusive. Because so many genetic studies rely on existing selective, socially constructed population definitions and participant self-report of these population definitions, there is potential for mostly unintended implications of genetic findings for racial and ethnic identities to feed back into future genetic studies.

In contrast to such limited ways of thinking about human populations, discovering the genetic contributors to complex diseases and drug responses will require multiple cohorts, information about multiple historical sources of variation and sophisticated means for linking their analyses. To make these linkages, scientists will have to conceptualize the many ways in which genetic variants are distributed across different, overlapping definitions of contemporary and historical populations. This is a more complicated task compared with the simplified (and long-standing) practice of linking selected genetic features to a categorical population identity or ancestry. Human genetic variation and human social identity and ancestry are complex, multifaceted phenomena that promise great rewards when investigated in their entirety, but that have distinctive risks when one or both are reduced to less sophisticated typologies.

Morris W. Foster is at the Department of Anthropology, 455 W. Lindsey, Room 505C, University of Oklahoma, Norman, Oklahoma 73019, United States.

Richard R. Sharp is at the Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas 77030, United States.

Correspondence to M.W.F.
e-mail: morris.w.foster-1@ou.edu

doi:10.1038/nrg1452

1. Collins, F. S. & Mansoura, M. K. The human genome project: revealing the shared inheritance of all humankind. *Cancer* **91** (Suppl.), 221–225 (2001).
2. Subramanian, G. *et al.* Implications of the human genome for understanding human biology and medicine. *JAMA* **286**, 2296–2307 (2001).
3. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
4. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
5. Foster, M. W. & Sharp, R. R. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.* **12**, 844–50 (2002).
6. Risch, N. *et al.* Categorizations of humans in biomedical research: genes, race and disease. *Genome Biol.* **3**, 12 (2002).
7. Cooper, R. S., Kaufman, J. S. & Ward, R. Race and genomics. *N. Engl. J. Med.* **348**, 1166–1170 (2003).
8. Burchard, E. G. *et al.* The importance of race and ethnic background in biomedical research and clinical practice. *N. Engl. J. Med.* **348**, 1170–1175 (2003).
9. Freeman, H. P. The meaning of race in science — considerations for cancer research — concerns of special populations in the National Cancer Program. *Cancer* **82**, 219–225 (1998).
10. Collins, F. S. *et al.* A vision for the future of genomics research. *Nature* **422**, 835–847 (2003).
11. Goldstein, D. B. & Chikhi, L. Human migrations and population structure: what we know and why it matters. *Annu. Rev. Genomics Hum. Genet.* **3**, 129–152 (2002).
12. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
13. Johnson, G. C. L. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
14. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
15. Zhao, H. Y., Pfeiffer, R. & Gail, M. H. Haplotype analysis in population genetics and association studies. *Pharmacogenomics* **4**, 171–178 (2003).
16. Morton, N. E. Genetic epidemiology, genetic maps and positional cloning. *Phil. Trans. R. Soc. Lond. B* **358**, 1701–1708 (2003).
17. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
18. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
19. Wilson, J. F. *et al.* Population genetic structure of variable drug response. *Nature Genet.* **29**, 265–269 (2001).
20. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
21. Goldstein, D. B. Islands of linkage disequilibrium. *Nature Genet.* **29**, 109–111 (2001).
22. Daly, M. J. *et al.* High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
23. Calafell, F. Classifying humans. *Nature Genet.* **33**, 435–436 (2003).
24. Van den Oord, E. J. C. G. & Neale, B. M. Will haplotype maps be useful for finding genes? *Mol. Psychiatry* **9**, 227–236 (2004).
25. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).

26. Lai, E. *et al.* Medical applications of haplotype-based SNP maps: learning to walk before we run. *Nature Genet.* **32**, 353–54 (2002).
27. Shriver, M. D. *et al.* Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* **112**, 387–399 (2003).
28. Bamshad, M. J. *et al.* Human population genetic structure and inference of group membership. *Am. J. Hum. Genet.* **72**, 578–589 (2003).
29. Webster, M. T., Clegg, J. B. & Harding, R. M. Common 5′ β-globin RFLP haplotypes harbour a surprising level of ancestral sequence mosaic. *Hum. Genet.* **113**, 123–139 (2003).
30. Gower, B. A. Using genetic admixture to explain racial differences in insulin-related phenotypes. *Diabetes* **52**, 1047–1051 (2003).
31. Shifman, S. *et al.* Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet.* **12**, 771–776 (2003).
32. Lin, J. P. *et al.* Genealogy construction in a historically isolated population: application to genetic studies of rheumatoid arthritis in the Pima Indian. *Genet. Med.* **1**, 187–193 (1999).
33. Zhou, X. *et al.* Genome-wide association study of systemic sclerosis susceptibility in a Choctaw Indian population with high disease prevalence. *Arthritis Rheum.* **48**, 2585–2592 (2003).
34. Thornton, R. *American Indian Holocaust and Survival: A Population History since 1492* (Univ. Oklahoma Press, USA, 1987).
35. Sturm, C. *Blood Politics: Race, Culture, and Identity in the Cherokee Nation of Oklahoma* (Univ. California Press, USA, 2002).
36. 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).
37. Wacholder, S., Rothman, N. & Caporaso, N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl Cancer Inst.* **92**, 1151–1158 (2000).
38. Ziv, E. & Burchard, E. G. Human population structure and genetic association studies. *Pharmacogenomics* **4**, 431–441 (2003).
39. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
40. Cardon, L. R. & Palmer, L. J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
41. Foster, M. W. & Aston, C. E. A practice approach for identifying previously unsuspected environmental contributors to SLE and other complex diseases. *Environ. Health Persp.* **111**, 593–597 (2003).
42. Zondervan, K. T. & Cardon, L. R. The complex interplay among factors that influence allelic association. *Nature Rev. Genet.* **5**, 89–100 (2004).
43. Sankar, P. & Cho, M. K. Toward a new vocabulary of human genetic variation. *Science* **298**, 1337–1338 (2002).
44. Jorde, L. B., Watkins, W. S. & Bamshad, M. J. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* **10**, 2199–2207 (2001).
45. Broder, C. & Venter, J. C. Whole genomes: the foundation of new biology and medicine. *Curr. Opin. Biotechnol.* **11**, 581–585 (2000).
46. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, USA, 1994).
47. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
48. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease–common variant… or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
49. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
50. Sebastiani, P. *et al.* Minimal haplotype tagging. *Proc. Natl Acad. Sci. USA* **100**, 9900–9905 (2003).
51. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
52. Clark, A. G. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr. Opin. Genet. Devel.* **13**, 296–302 (2003).
53. Smith D. J. & Lusis A. J. The allelic structure of common disease. *Hum. Mol. Genet.* **11**, 2455–2461 (2002).
54. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142 (2002).
55. Tishkoff, S. A. & Williams, S. M. Genetic analysis of African populations: human evolution and complex disease. *Nature Rev. Genet.* **3**, 611–621 (2002).

Acknowledgements

This publication was made possible by grants from the National Institute on Environmental Health Sciences and from the National Human Genome Research Institute. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NHGRI or the National Institutes of Health. We also are grateful for comments made by L. D. Brooks and two anonymous reviewers on an earlier version.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

Entrez: http://www.ncbi.nih.gov/Entrez/

APOE

OMIM: http://www.ncbi.nlm.nih.gov/Omim/

Alzheimer disease

FURTHER INFORMATION

International HapMap Project:

http://www.nhgri.nih.gov/10001688

Human Genome Diversity Project:

http://www.stanford.edu/group/morrinst/hgdp.html

Access to this interactive links box is free online.

OPINION

Genetic equity

John Harris and John Sulston

Abstract | This paper proposes, elaborates and defends a principle of genetic equity. In doing so it articulates, explains and justifies what might be meant by the concept of 'human dignity' in a way that is clear, defensible and consistent with, but by no means the same as, the plethora of appeals to human dignity found in contemporary bioethics, and more particularly in international instruments on bioethics. We propose the following principle of genetic equity: humans are born equal; they are entitled to freedom from discrimination and equality of opportunity to flourish; genetic information may not be used to limit that equality.

The prestige of genetics and the fears provoked by genetic science are both possibly at their zenith. The sequencing of the human genome, the promises of PHARMACOGENETICS and GENE THERAPY and the success of genetic fingerprinting in the identification of criminals have placed genetic techniques at the cutting edge of science and popular interest in science. Equally, fears about REPRODUCTIVE CLONING, 'designer babies', genetic enhancements of humans and other species, and fears that the abuse of genetic information will lead to unfair discrimination or forms of eugenic genetic cleansing, have occasioned an almost unprecedented mistrust of science. Not all fears of science are either rational or justified, of course, and there is little or nothing that can be done about some of these irrational fears other than waiting for the passage of time. Jonathan Swift is often credited with having observed that "It is impossible to reason a man out of something that he has not been reasoned into." However, fears concerning the abuse of genetics resulting in discrimination or in violations of human rights are both rational and reasonable and it is, we believe, particularly incumbent on scientists to take a principled stand against such abuses. We therefore propose a firm commitment to genetic equity, and this paper elaborates the basis and the justification for such a principle. Some commentators consider that such an emphasis on genetics — so called GENETIC EXCEPTIONALISM — is misguided, but although we do not believe that genetics raises unprecedented or qualitatively different issues, we do think that there are good reasons to focus more general principles specifically on genetics.

In listening to, and participating in, many discussions about the social consequences of the rapidly increasing power of human genetic knowledge, we have found ourselves despairing at a tendency towards what might be termed genetic unexceptionalism: a feeling that 'the devil is in the detail' and that the best that can be done is to start from scratch in each case. It is true that grand generalizations are not a sufficient guide to good social conduct and law-making. However, detailed investigation is an insufficient guide to dealing with the future; we not only need to lift our gaze from the ground to the horizon, we urgently need a compass to indicate which points on the horizon are in the right direction. In developing a principle of genetic equity, we are proposing a compass point to guide the direction of genetics. We also propose a new and, we believe, more fundamental and coherent concept of human dignity.