Hum Genet (2005) 118: 366–371
DOI 10.1007/s00439-005-0039-3

ORIGINAL INVESTIGATION

Andrea Manica · Franck Prugnolle · François Balloux

Geography is a better determinant of human genetic differentiation than ethnicity

Received: 15 June 2005 / Accepted: 15 July 2005 / Published online: 28 September 2005

Abstract Individuals differ genetically in their susceptibility to particular diseases and their response to drugs. However, personalized treatments are difficult to develop, because disease susceptibility and drug response generally have poorly characterized genetic architecture. It is thus tempting to use the ethnicity of patients to capture some of the variation in allele frequencies at the genes underlying a clinical trait. The success of such a strategy depends on whether human populations can be accurately classified into discrete genetic ethnic groups. Despite the heated discussions and controversies surrounding this issue, there has been essentially no attempt so far to quantify the relative power of ethnic groups and geography at predicting the proportion of shared alleles between human populations. Here, we present the first such quantification using a dataset of 51 populations typed at 377 autosomal microsatellite markers, and show that pair-wise geographic distances across landmasses constitute a far better predictor than ethnicity. Allele-sharing between human populations worldwide decays smoothly with increasing physical distance. We discuss the relevance of these patterns for the expected distribution of variants of medical interest. The distribution patterns of gene coding for simple traits are expected to be highly heterogeneous, as most such genes experienced strong natural selection. However, variants involved in complex traits are expected to behave essentially neutrally, and we expect them to fit closely our predictions based on microsatellites. We conclude that the use of ethnicity alone will often be inadequate as a basis for medical treatment.

A. Manica
Evolutionary Ecology Group, Department of Zoology,
University of Cambridge, Downing Street,
CB2 3EJ Cambridge, UK

F. Prugnolle · F. Balloux
Theoretical and Molecular Population Genetics group,
Department of Genetics, University of Cambridge,
Downing Street, CB2 3EH Cambridge, England

F. Balloux (✉)
Department of Genetics, University of Cambridge,
Downing Street, CB2 3EH Cambridge, UK
E-mail: fb255@mole.bio.cam.ac.uk
Tel.: +44 1223 333964
Fax: +44 1223 333992

Introduction

In recent years, there has been renewed enthusiasm for the study of human settlement history. This has been fuelled by recent technical developments in molecular biology and statistical genetics and the increasing recognition that human genetic history has major medical relevance (Bamshad and Wooding 2003; Goldstein and Chikhi 2002; Tishkoff and Verreli 2003). While it is widely recognized that there is variation within and between human populations with respect to their susceptibility to many diseases (Burchard et al. 2003), focus has shifted towards ethnic-specific variation in drug processing, efficacy and frequency of adverse reactions (Rotimi 2004; Tate and Goldstein 2004; Xie et al. 2001). Adverse drug response is estimated to rank as high as the fourth most important cause of death in the USA (Lazarou et al. 1998) and is therefore a major public health concern. For traits with a well-understood genetic basis, patients can be typed at relevant loci, and the resultant genotype profile used for optimized treatment. While individually tailored treatments are viewed by some as the medicine of the future (Goldstein et al. 2003), their widespread realization will not occur for some time. Some authors have thus advocated taking a shortcut using the ethnic background of patients to capture a fraction of the genetic (and environmental) variation underlying traits of medical relevance (Burchard et al. 2003; Risch et al. 2002). How well such a strategy may perform depends on a series of factors, including the number of genes involved in the trait as well as patterns of past and present natural selection that affected those genes (Keita et al. 2004).

More crucially, the success of any strategy based on ethnicity depends on the ability to classify human populations into discrete ethnic groups. There is near consensus in the recent literature that while human genetic diversity is largely clinal, humans nevertheless cluster into five or six broad ethnic groups, roughly corresponding to continents (Jorde and Wooding 2004; Risch et al. 2002; Rosenberg et al. 2002; Tishkoff and Kidd 2004; Zhivotovsky et al. 2003). There is disagreement as to how meaningful those clusters are in a medical context, but there have been very few challenges to their existence. Serre and Pääbo (2004) did suggest that the clusters were artefacts generated by heterogeneous sampling and that they would vanish if more populations were analysed. Our recent work (Prugnolle et al. 2005a,b) provides some support to their claim. We show that geographic distance from East Africa along likely colonization routes is an excellent predictor of neutral genetic diversity in a large number of human populations ($n=51$; $R=93\%$). The smoothness of this relationship suggests no obvious macro-geographic pattern, such as a step-wise decrease in genetic diversity corresponding to a severe bottleneck following the colonization of a continent. This result does question the existence of previously defined ethnic groups (Prugnolle et al. 2005a); however, the power of the approach in detecting genetic discontinuities is arguably weak. Only genetic diversity within each population was considered; thus, information about allele sharing between populations was discarded. Here we build upon our previously developed, geographically explicit approach (Prugnolle et al. 2005a,b) and specifically quantify the amount of relative variation in population differentiation captured by ethnic clusters and geography. Statistical power is improved by correlating the proportion of shared alleles with geographic distance and previously defined ethnic groups, considering all possible pairs of populations.

Material and methods

We used the original dataset of 52 human populations distributed worldwide and typed at 377 autosomal microsatellite loci (Rosenberg et al. 2002), but the two Han populations were pooled (Han sampled in China and the USA). While this represents by far the largest available dataset of autosomal markers, the number of typed individuals per population is small (range: 7–51; mean ± SD: 22.9 ± 11.1). Geographic distances, which are meaningful in the context of migration between populations, are obtained by computing all shortest pair-wise distances through landmasses, under the assumption that most gene flow was the consequence of migrants moving over land until recent historical times. In Fig. 1, we illustrate the approach by giving as an example all routes connecting one focal population, represented by the larger blue dot, against all other populations. We chose as the focal population the Adygei, an ethnic group of the Russian Caucasus. Those distances were obtained using an algorithm based on graph theory that we previously developed (Prugnolle et al. 2005a). The advantage of this approach over conventional spatial statistics (as used in GIS software) is that we do not assume the data to be in a Cartesian coordinates system resulting from projecting a spherical surface onto a flat surface. While this approximation is quite accurate for relatively limited areas, it is problematic for problems that encompass the whole globe.

In a second step, we estimate genetic similarity between all pairs of populations as the proportion of shared alleles (Bowcock et al. 1994). We chose the allele sharing statistics (AS) rather than more classical genetic distances such as F_{ST} because allele sharing directly relates to the probability that a given variant will be found at similar frequencies in different populations. Allele sharing is a straightforward statistic, simply defined as the proportion of alleles over all loci shared between two populations divided by the number of loci typed. The relationship between geographic distance and allele sharing was investigated through a Mantel correlation. We then tested whether populations from the same ethnic group had a degree of genetic similarity greater than the one predicted by geographic distance alone. To do this, we first created an "ethnic" distance matrix—-where population pairs were scored as 0 if they came from the same ethnic group and 1 if they came from a different ethnic group—and then computed the first order partial Mantel correlation between the genetic and ethnic distance matrices, correcting for geographic distance. Here we report the results for the a priori and a posteriori ethnic clusters given by Rosenberg et al. (2002). The a priori clusters roughly correspond to continental origin, whereas the five a posteriori clusters were generated by the algorithm implemented in STRUCTURE (Pritchard et al. 2000).

In order to compare the pattern of isolation by distance within the different ethnic groups, we performed an analysis of similarity (ANOSIM), a technique analogous to an ANOVA but able to handle pair-wise distance matrices (Legendre and Legendre 1998). ANOSIM explicitly compares the level of similarity found in all pairs within each level of a factor against the similarity found in between-level pairs. Significance testing is obtained through randomizations (Legendre and Legendre 1998). To account for geographic distance in our ANOSIM, we used the residuals obtained from regressing genetic distance on geographic distance. All computations were performed in R 2.0.1 (R Core Development Team 2004), using packages ade4 (URL http://pbil.univ-lyon1.fr/ADE-4) and vegan (version 1.6–5; URL http://cc.oulu.fi/~jarioksa). We did not include Oceania for which only two populations are available (adding Oceania does not affect the results qualitatively, but the power is too low to draw any conclusions on this group).

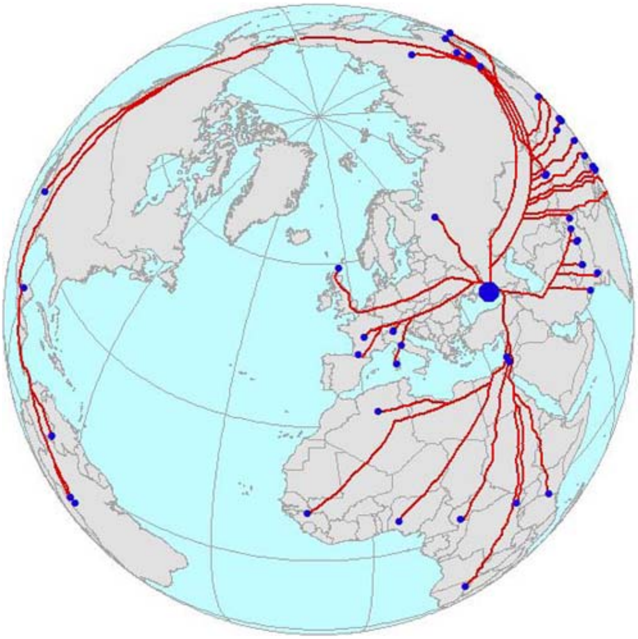

Fig. 1 Illustration of the shortest routes through landmasses and specified land bridges. For clarity, we limit ourselves to the routes connecting a single population against all others. We chose the Adygei, an ethnic group of the Russian Caucasus as focal population (*larger dot* at the centre of the web of routes)

Results

The allele-sharing distance [-log(AS)] and pair-wise geographic distance matrices are highly correlated ($R = 0.77$; $P < 0.001$; Fig. 2). After we have accounted for geography, we can ask whether previously defined ethnic groups explain additional variance in the pattern of genetic differentiation between populations. To test this, we recovered the residuals of the genetic differentiation matrix after geographic distance had been accounted for and tested for a correlation between those residuals and a matrix of ethnic groups. Both the a priori ($R = 0.13$) and a posteriori ($R = 0.42$) clusters defined by Rosenberg et al. (2002) were significantly correlated to the residuals of the allele-sharing matrix. We further considered a variety of other possible ethnic groupings. None of those alternative classifications explained more variance in allele sharing than the a posteriori clustering strategy.

Genetic differentiation among human populations at the scale of the globe is primarily dependent on geographic isolation. Discrete ethnic clusters can nevertheless be defined even after geographic distance is accounted for. As the a priori grouping explains far less variance, we will not consider this definition further, and for simplicity will refer hereafter to the a posteriori clusters as ethnic groups. The different ethnic groups are characterized by different patterns of genetic isolation by distance. While Eurasia displays a clear clinal decrease in allele sharing with geographic distance ($R = 0.36$; $P < 0.002$), isolation by distance within East Asia, Africa and America is not significant. It should however be noted that the statistical power is weak, as sample sizes for within-ethnic group comparisons are very small. The ANOSIM analysis allows us to get further insight into the within-ethnic group pattern (Fig. 3). The heterogeneity of the different ethnic groups is highly variable (ANOSIM $R = 0.51$; $P = 0.001$) (Fig. 3). Once geography has been accounted for, there is far less variance left to account for in Eurasia and to a lesser extent in Eastern Asia than in Africa and America.

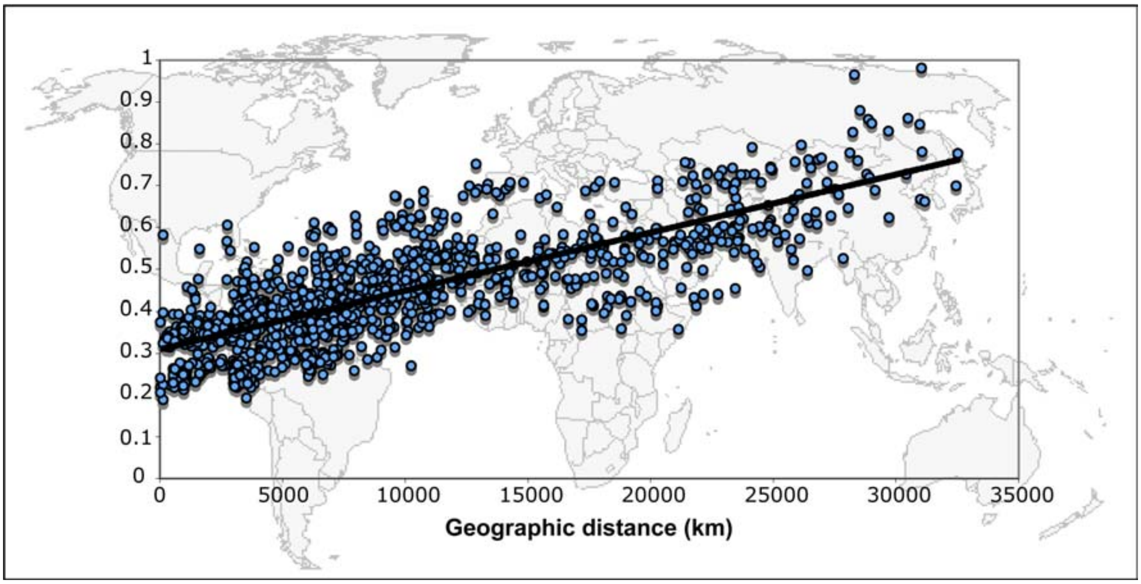

Fig. 2 Relationship between pair-wise allele sharing (-log(AS)) and geographic distance for all possible pairs of populations in the dataset. -log(AS) increases significantly with distance ($R = 0.77$, $P < 0.001$)

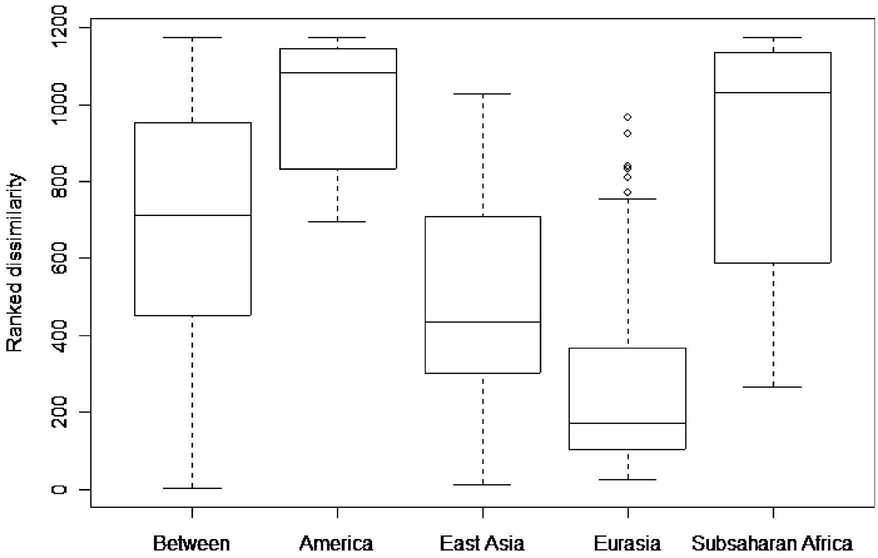

Fig. 3 Boxplot of ranked genetic dissimilarity corrected by geographic distance (i.e., residuals). Pair-wise comparisons were grouped as "between" ethnic groups or within ethnic group (individual groups shown in the plot). There are significant differences among the four ethnic groups with Eurasia and Eastern Asia being characterized by higher homogeneity (ANOSIM $R=0.51$, $P=0.001$). Boxes represent the median and interquartile range (IQR); whiskers extend to the most extreme data points up to 1.5 times the IQR; *open dots* represent outliers outside this range

So far we have been considering all alleles pooled together. It is likely that many variants of medical interest are at relatively low frequency. We have thus explored the relative power of geography and ethnicity at predicting allele sharing for classes of alleles at different frequencies (Fig. 4). Independent of the frequency of the alleles considered, geography is always a far better predictor than ethnicity for the proportion of shared variants between populations. The proportion of additional variance explained by ethnicity on top of geography is small for rare variants but increases with allele frequency. The overall proportion of variance explained increases with the inclusion of classes of more frequent alleles and peaks when all alleles below 20–30% are considered, before dropping when the most frequent alleles are included. Interestingly, the most powerful approach for predicting allele sharing considers ethnicity first and then geography, independent of the frequency class of the alleles considered (Fig. 4).

Discussion

In this paper, we have quantified the relative power of geography and ethnicity at predicting the proportion of shared alleles between populations by using neutral

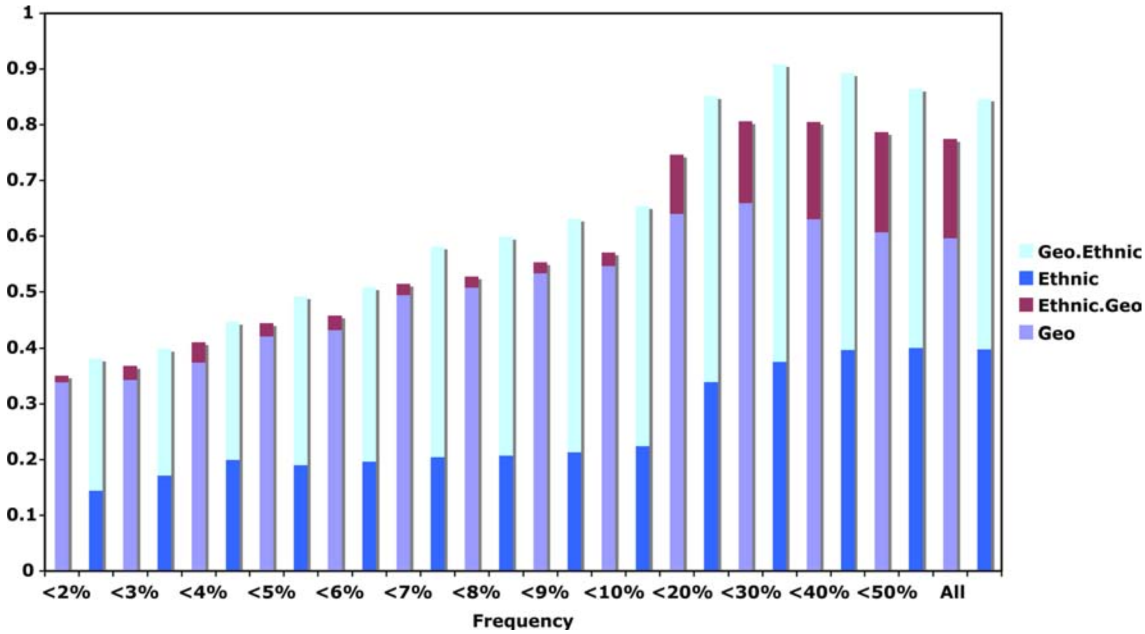

Fig. 4 Proportion of variance explained (R^2) in allele sharing between all pairs of populations by geographic distance along landmasses (*Geo*), ethnicity (*Ethnic*), ethnicity after geography has been accounted for (*Ethnic.Geo*) and geography after ethnicity was taken into account (*Geo.Ethnic*). The x-axis represents the cumulative pooling of alleles below a certain overall frequency in the 51 populations analysed

microsatellite loci. Our results demonstrate that geography is a far better proxy than ethnicity for predicting the probability of allele sharing between populations. While ethnic groups can be statistically defined even after geography has been accounted for, they explain little additional variance, especially for alleles at relatively low frequency.

Patterns of allele sharing within ethnic groups are more complex. In Eurasia, we observed a clear decrease in allele sharing with increasing physical distance ($R = 0.36$; $P < 0.002$). We could detect no such trend in Africa or America. However, the sample sizes for Africa and America are so small that this should not be interpreted as strong evidence against isolation by distance in those continents (six and five populations, respectively). The situation for East Asia is more intriguing. When we considered all 18 populations from East Asia, we did not detect significant isolation by distance despite the fact that the geographic scale is wider than for Eurasia. However if the Lahu are removed, the relation becomes significant ($R = 0.16$; $P < 0.03$). Interestingly, this population originates from the Tibetan plateau and has migrated to South East Asia over the last 200 years. Rosenberg et al. (2000) also reported that Lahu were the least heterozygous population in the region, and the researchers often separated the Lahu from other groups in southern China in their analysis. Even without the Lahu, the correlation between geography and genetic similarity remains weak. Our shortest-distance-along-landmasses approach may not be adequate to capture gene flow in East Asia due to the difficulty of crossing the Himalayan mountain range. This suggests that a possible improvement in future work may be achieved by giving high friction costs to mountain ranges. The within-continent patterns are corroborated by the ANOSIM analysis. Once geography has been accounted for, there is very little variance left to explain in Eurasia. On the other hand, Africa and America are characterized by strong heterogeneity between populations. Finally, East Asia displays an intermediate pattern. These results are also generally in line with previous evidence for differential spatial structuring of genetic diversity in each continent, as reviewed by Cavalli-Sforza and Feldman (2003).

Microsatellites are characterized by extremely high mutation rates in humans, typically lying around 10^{-3} (Ellegren 2000; Xu et al. 2000). Their mutation pattern is step-wise with the majority of mutations leading to an addition or deletion of one repeat motif. The high mutation rate together with the step-wise process leads to high homoplasy (Ellegren 2000; Xu et al. 2000). Alleles identical in state will not necessarily share the same ancestry, as the same allelic state can be obtained through different chains of mutational events. Homoplasy is expected to be highest for the most common alleles. This most likely explains the decrease in variance by both geography and ethnicity when we include alleles at highest frequency (Fig. 4). Higher homoplasy of frequent alleles may also drive the trend for ethnicity to explain a higher fraction of variance in allele sharing, as more and more common alleles are pooled (Fig. 4). It is thus likely that the geographic distribution of alleles with low homoplasy (e.g., SNPs) will be explained essentially by geography, similar to the microsatellite alleles at low frequency.

Microsatellites have no known function and are considered to be neutral, even if some might be in linkage disequilibrium (LD) with selected genes. Populations with smaller effective size will be characterized by higher genome-wide linkage disequilibrium, and thus by larger segregating LD blocks. The proportion of microsatellites segregating with variants under natural selection will thus be higher in populations with small effective size. For instance, African populations have higher effective population sizes than non-African populations and also lower levels of LD (Reich et al. 2001). Similarly, it has been shown that LD is higher in hunter-gatherers than in food producing societies (Kaessman et al. 2002). While the effect of population size on linkage disequilibrium should not be dismissed, we feel confident that the vast majority of the microsatellites in this dataset behave essentially neutrally. Indeed, using the same dataset, we previously reported a correlation of $R = 93\%$ between genetic diversity and geographic distance from East Africa along landmasses (Prugnolle et al. 2005b). Such a strong correlation would be unlikely if population size had a strong effect on the markers analysed. Furthermore, attempts to control for population size did not result in an increased correlation (FB unpublished results).

Whether medically relevant variants are characterized by the same patterns as we describe here for microsatellites mainly depends on the nature of the selective processes that have acted upon them. Many disease variants involved in simple traits have probably experienced strong selection in the past. If this selection has been geographically localized, the resulting distribution can be very patchy. An extreme example of this is the CCR5-Delta32 HIV resistance allele, which is essentially restricted to Northern Europe, where it is found at high frequency (Stephens et al. 1998). It has been suggested that drug response may generally have a simpler genetic architecture than disease susceptibility/resistance (Goldstein et al. 2003). While the fine geographic distribution of pharmacogenetic variants has not been studied, it is known they are also found at variable frequencies in different human populations (Goldstein and Hirschhorn 2004; Xie et al. 2001).

Current understanding of complex disease genes is still highly imperfect (Hirschhorn et al. 2002; Risch 2000), and the best-characterized examples may simply be "low hanging fruits" that do not display the full complexity of typical disease genes (Pritchard and Cox 2002). While variants involved in complex diseases seem to be at variable frequencies in different human populations (Ioannidis et al. 2004), there is essentially no information available about their fine geographic distribution. For complex traits involving many loci and/or alleles with low penetrance, the variants at those genes can be expected to be under weak selection and may thus conform to the predictions we define for neutral alleles.

In this paper we have shown that the proportion of neutral alleles shared between populations can to a large extent be predicted by geography. This prediction should remain true for any polymorphism under weak selection, and is thus expected to apply to variants underlying complex traits. The distribution of individual variants of medical interest involved in simpler traits is expected to be very variable and essentially impossible to predict. However, even for variants whose distribution cannot be accurately predicted by geography, there is no particular reason to believe the pattern could be better captured by some general ethnic classification. This strongly suggests that ethnic groups will generally be inadequate proxies for the distribution of traits of medical relevance. A more powerful approach when considering the host's genetic background in medicine might be to use individual geographical locations as a continuous variable, or even better, use both geography and ethnicity together.

Acknowledgments We thank Lori Lawson-Handley, Elizabeth Tyler and Graham Forster and four anonymous reviewers for very helpful comments and suggestions. FP is supported by a Lavoisier Fellowship from the Ministère Français des Affaires Etrangères. FB acknowledges support from the BBRSC.

References

Bamshad M, Wooding S (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4:99–111

Bowcock A, Ruiz-Linares A, Tonfohrde J, Minch E, Kidd J, Cavalli-Sforza L (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 386:455–457

Burchard E, Ziv E, Coyle N, Gomez S, Tang H, Karter A, Mountain J, Perez-Stable E, Sheppard D, Risch N (2003) The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 348:1170–1175

Cavalli-Sforza LL, Feldman W (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet 33(Suppl):266–275

Ellegren H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 16:551–558

Goldstein D, Chikhi L (2002) Human migrations and population structure: what we know and why it matters. Annu Rev Genomics Hum Genet 3:129–152

Goldstein D, Hirschhorn J (2004) In genetic control of disease, does "race" matter? Nat Genet 36:1243–1244

Goldstein D, Tate S, Sisodiya S (2003) Pharmacogenetics goes genomic. Nat Rev Genet 4:937–947

Hirschhorn J, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4:45–61

Ioannidis J, Ntzani E, Trikalinos T (2004) "Racial" differences in genetic effects for complex diseases. Nat Genet 36: 1312–1318

Jorde L, Wooding S (2004) Genetic variation, classification and "race". Nat Genet 36:S28–S33

Kaessmann H, Zollner S, Gustafsson A, Wiebe V, Laan M, Lundeberg J, Uhlen M, Paabo S (2002) Extensive linkage disequilibrium in small human populations in Eurasia. Am J Hum Genet 70:673–685

Keita S, Kittles R, Royal C, Bonney G, Furbert-Harris P, Dunston G, Rotimi C (2004) Conceptualizing human variation. Nat Genet 36:S17–S20

Lazarou J, Pomeranz B, Corey P (1998) Incidence of adverse drug reactions in hospitalized patients—a meta-analysis of prospective studies. JAMA 279:1200–1205

Legendre P, Legendre L (1998) Numerical ecology, 2nd English edn. Elsevier, Amsterdam

Pritchard J, Cox N (2002) The allelic architecture of human disease genes: common disease—common variant ... or not? Hum Mol Genet 11:2417–2423

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

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

Prugnolle F, Manica A, Charpentier M, Guégan J, Guernier V, Balloux F (2005b) Worldwide HLA diversity: human colonisation history and pathogen-driven selection. Curr Biol 15:1022–1027

Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D, Lavery T, Kouyoumjian R, Farhadian S, Ward R, Lander E (2001) Linkage disequilibrium in the human genome. Nature 411:199–204

Risch N (2000) Searching for genetic determinants in the new millenium. Nature 405:847–856

Risch N, Burchard E, Ziv E, Tang H (2002) Categorization of humans in biomedical research: genes, race and disease. Genome Biol 3(7):comment 2007

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Rotimi C (2004) Are medical and nonmedical uses of large-scale genomic markers conflating genetics and "race"? Nat Genet 36:S43–S47

Serre D, Pääbo S (2004) Evidence for gradients of human genetic diversity within and among continents. Genome Res 14:1679–1685

Stephens J, Reich D, Goldstein D, Shin H, Smith M, Carrington M, Winkler C, Huttley G, Allikmets R, Schriml L, Gerrard B, Malasky M, Ramos M, Morlot S, Tzetis M, Oddoux C, di Giovine F, Nasioulas G, Chandler D, Aseev M, Hanson M, Kalaydjieva L, Glavac D, Gasparini P, Kanavakis E, Claustres M, Kambouris M, Ostrer H, Duff G, Baranov V, Sibul H, Metspalu A, Goldman D, Martin N, Duffy D, Schmidtke J, Estivill X, O'Brien S, Dean M (1998) Dating the origin of the CCR5-Delta 32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62:1507–1515

Tate S, Goldstein D (2004) Will tomorrow's medicines work for everyone? Nat Genet 36:S34–S42

R Core Development Team (2004) R: a language and environment for statistical computing. R Foundation for statistical computing, Vienna, ISBN 3-900051-07-0, URL http://www.R-project.org

Tishkoff S, Kidd K (2004) Implications of biogeography of human populations for "race" and medicine. Nat Genet 36:S21–S27

Tishkoff S, Verreli B (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4:293–340

Xie H, Kim R, Wood A, Stein C (2001) Molecular basis of ethnic differences in drug disposition and response. Annu Rev Pharmacol Toxicol 41:815–850

Xu X, Peng M, Fang Z, Xu X (2000) The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24:396–399

Zhivotovsky L, Rosenberg N, Feldman M (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am J Hum Genet 72:1171–1186