Original Paper

Human Heredity

Hum Hered 2006;61:15–21
DOI: 10.1159/000091832

Received: December 14, 2005
Accepted after revision: January 26, 2006
Published online: March 7, 2006

Genomic Boundaries between Human Populations

Guido Barbujani Elise M.S. Belle

Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy

Key Words
Human genome diversity · Population structure · Microsatellites · Genetic boundaries · Randomization tests

Abstract
Different authors disagree on whether human genome variation should be described as continuous or discontinuous; in the latter case, by attributing an individual's genotype to one genetic cluster, one would also obtain information on the individual's genome in general. An analysis of 377 microsatellites of the CEPH human diversity panel was interpreted as evidence that most genotypes cluster into one of five distinct groups, approximately corresponding to continents, which were proposed by some authors as the major biological subdivisions of humankind. Here we analyse the same dataset by a specific numerical method, designed to detect genomic boundaries, i.e. zones of increased change in maps of genomic variation. We show that statistically significant boundaries can be described between groups of populations, but different clusters are identified, depending on the assumptions of the model. In addition, these clusters do not correspond to the clusters inferred from previous analyses of the same or of other polymorphisms. We conclude that it is indeed possible to cluster genotypes according to geography, but no study so far identified unambiguously anything that can be regarded as a major genetic subdivision of humankind, and hence discontinuous models of human diversity are unsupported by data.

Introduction

Human genetic diversity and population structure have long been studied for evolutionary purposes. Recently, however, a new interest in these subjects stemmed from the fact that knowledge of population structure was shown to be crucial in diverse areas such as epidemiology, pharmacogenetics, forensic science, and in the planning of association studies. Schematically, there are two main alternative views. Some authors think that, for practical purposes such as estimation of disease risks or gene hunting, it is convenient to regard humans as subdivided in genetically-distinct groups. These groups would roughly correspond to continents, races, or to what has been termed 'self-assessed ethnic identities' [1–4]. Indeed, it has been suggested that mutations leading to pathologies or predisposition to pathologies would be 'nearly always' race-specific [2]. Conversely, other authors regard variation among populations as essentially continuous and

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0001–5652/06/0611–0015$23.50/0

Accessible online at:
www.karger.com/hhe

Guido Barbujani
Dipartimento di Biologia, Università di Ferrara
Via L. Borsari 46
I–44100 Ferrara (Italy)
Tel. +39 0532 291312, Fax +39 0532 294761, E-Mail g.barbujani@unife.it

clinal, thus casting doubts on the existence of biological boundaries among human populations or groups thereof [2, 6–9] and ultimately on the usefulness of racial categorization in clinical genetic research.

Supporters of both views agree that the main fraction of human diversity, some 85% of the total, is represented by differences between members of the same population. As a consequence, individuals from different populations and different continents are expected to differ genetically, respectively, 5 and 10% more than two random members of the same population [5, 10–14]. The question is whether the differences between populations and continents, albeit representing a small fraction of the total, are large enough, and consistent enough across loci, to allow identification of clusters of biologically-differentiated individuals. If so, by analysing different sets of genetic data, or the same dataset with different methods, one should consistently find the same clusters, separated by boundaries of increased genomic change. If, on the contrary, no consistent clustering emerges, one should regard human genetic variation as essentially continuous in space. If variation is discontinuous, by attributing an individual to one genetic cluster, one would also obtain information on the individual's genome in general, whereas, if variation is continuous, the labels placed on such groups would be biologically arbitrary. Rather, these labels could reflect cultural or social differences, but would have little to do with clear-cut genetic differences, including differences at the genes involved in complex pathological traits.

In the largest study so far on human genomic diversity, Rosenberg et al. [13] analysed variation at 377 microsatellite loci by an algorithm implemented in the software Structure, which treats populations as hybrids among k parental populations defined by distinctive allele frequencies, and assigns individual genotypes to k clusters. Six clusters were identified, one comprising only the Kalash of Pakistan, and five approximately corresponding to continents, namely sub-Saharan Africa, East Asia, Oceania, the Americas, and Western Eurasia together with North Africa. The authors of the original paper drew prudent conclusions from their observations, but their results were interpreted by other authors as evidence that most humans can safely be attributed to one of five major genetic groups, roughly corresponding to continents [3, 4, 15] and ultimately to 'racial/ethnic groups' [16].

In fact, interpretation of these results is not straightforward. Structure estimates the likely genetic contributions of k parental populations to the current populations, but does not take geography into consideration, provides no information about the existence of boundaries of increased genetic change between populations, nor does it test for their statistical significance. In this study, we reanalysed the same dataset using a specific method that infers boundaries from genome diversity data [17]. We located the zones of highest genomic change on the world map and tested by two randomization schemes whether these zones do represent areas of significantly increased genetic change with respect to random locations on the world map, and with respect to the loci studied.

Materials and Methods

The dataset comprises 377 autosomal microsatellite loci distributed on all autosomes, typed in 1,056 individuals from 51 world populations (with all Bantu populations pooled) as part of a study of the cell-line diversity panel of the CEPH (Centre pour l'Etude des Polymorphismes Humains at the Institut Jean Dausset in Paris) [18].

Populations were placed on the map according to their coordinates, and a Delaunay triangulation (fig. 1) was drawn by the program Barrier ver.2.2 [17]. The algorithm used identifies neighboring populations and connects them so that the area studied is ultimately subdivided in triangles. The line separating adjacent populations, orthogonal to the edge connecting them, is the line across which the rate of genetic change is evaluated for every pair of populations. The network thus obtained was modified in peripheral areas in order to adapt it to specific features of the landscape such as the presence of seas or the known migration routes of the main human expansions [see for instance 19].

Matrices of F_{ST} [20] and of R_{ST} [21] genetic distances were computed between all populations by means of the Arlequin ver 2.0 software [22]. The two indices differ as for the underlying model of genetic differentiation. F_{ST} quantifies genetic diversity between populations assuming it reflects the interaction between genetic drift and gene flow. R_{ST} also considers mutation, and is specific for microsatellites in that it is based on the stepwise mutation model. Because it is calculated as the sum of the squared repeat differences between all pairs of haplotypes averaged across loci, R_{ST} provides a measure of genetic differentiation taking into account both allele-frequency differences between populations and molecular differences between alleles.

A graphic method, Monmonier's [23] maximum-difference algorithm, implemented in the software Barrier ver.2.2 [17], was used to identify putative boundaries. The genetic distances (F_{ST} or R_{ST}) between adjacent populations were associated with the respective edges of the network and ranked; putative boundaries were identified by initially tracing a perpendicular line across the edge of the Delaunay network showing the highest-ranking genetic distance, and were then extended across the adjacent edges, each time choosing the one associated with the highest genetic distances. This procedure was continued until the putative boundary hit the limits of the network, reached another pre-existing boundary, or closed on itself around one or more populations. There is no objective criterion to define the number of putative genetic boundaries to be recognized in this way, but there is a unique solution for each chosen

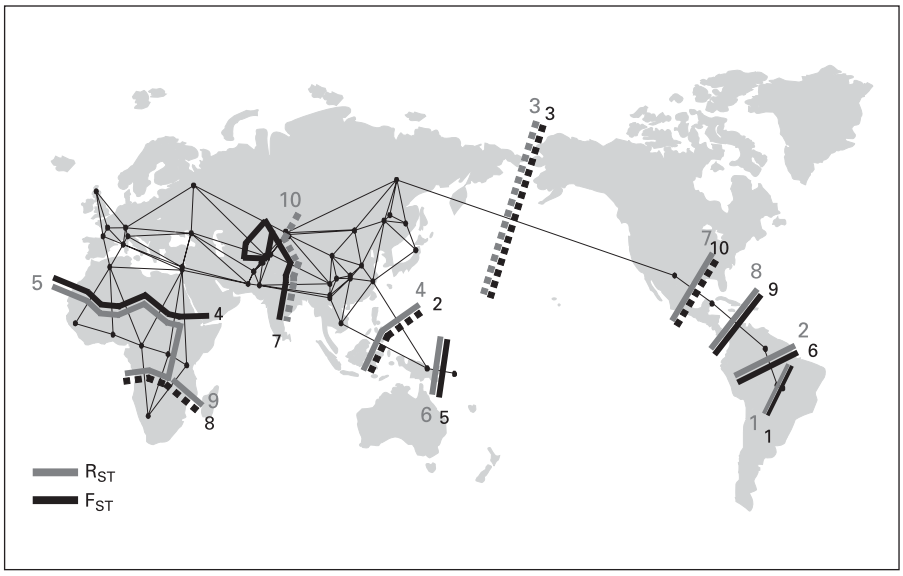

Fig. 1. Putative genomic boundaries between world populations represented by thick lines and numbered in decreasing order of importance. Sampling localities (black points) are connected by a Delaunay network (thin lines). Thick black lines indicate boundaries inferred from R_{ST} distances, thick gray lines those inferred from F_{ST} distances. Solid lines represent boundaries that are significant after both randomization tests, dotted lines the putative boundaries that do not reach statistical significance.

number of boundaries; we decided to stop at ten and, as we shall see, this was a conservative choice.

To assess the significance of the putative boundaries thus identified, we ran two independent randomization tests, considering respectively the individuals and the loci. By the first test we asked whether populations separated by putative boundaries are more genetically differentiated than random populations. In order to answer, we reassigned at random 1000 times individual genotypes to populations and populations to groups defined by the putative boundaries, thus carrying out an Analysis of MOlecular Variance (AMOVA) [24] implemented in the Arlequin ver 2.0 software [22]. At the end of each randomisation trial, for each locus, measures of genetic variance were estimated between individuals and populations, within and across each pair of regions separated by a putative boundary. The distribution of the 1000 sets of variances thus obtained was finally compared with the observed variances. Genetic change across a boundary was considered significant if the genetic variance observed across it fell into the top 2.5% of the distribution of the random variances, thus performing a two-tailed test.

For the second test, we asked whether putative boundaries represent zones of increased change for most polymorphisms considered, or only for a set of highly differentiated loci. In order to answer, the 377 loci were randomly resampled 100 times, using a customised version of the MsatBootstrap software [25] modified by the author so as to handle the large dataset analysed in the present study. In this way, 100 matrices of genetic distances derived from the bootstrapped loci were constructed, and boundaries were inferred from each matrix, repeating the whole procedure described above for the analysis of the original dataset, both for F_{ST} and for R_{ST} measures. Finally, we counted the number of times each segment of each putative boundary was also part of a boundary in the analysis of the bootstrapped data. We arbitrarily considered the boundary to be significant if a bootstrap value >70% was observed, that is to say if the boundary was found within the ten highest ranking boundaries more than 70 times out of 100.

Table 1. Test of significance of the putative boundaries by analysis of molecular variance (AMOVA)

Barrier number	Percentages of variance between groups and p values			
	R_{ST}		F_{ST}	
	percentage	p value	percentage	p value
1	15.4	0.0000	18.1	0.0000
2	12.6	0.0545	11.1	0.0000
3	6.7	0.0454	4.4	0.0273
4	11.9	0.0000	4.8	0.0000
5	12.6	0.0000	4.2	0.0000
6	10.5	0.0000	6.2	0.0000
7	6.6	0.0000	5.7	0.0000
8	2.4	0.1909	10.0	0.0000
9	10.5	0.0000	4.9	0.0000
10	5.2	0.0000	2.8	0.0000

Percentages of variance between groups at different sides of each identified boundary, either considering R_{ST} or F_{ST} distances, and p values. The significance threshold after Bonferroni correction is p = 0.0025.

Results

The ten highest-ranking putative genomic boundaries inferred from the analysis of 377 autosomal loci are presented in figure 1. The numbers associated with each boundary represent the ranking of that boundary, with 1 being the sharpest. Four of the putative boundaries occur within the New World, and two or three occur in Africa,

Table 2. Bootstrap values, considering either R_{ST} or F_{ST} genetic distances

Number of times as	1st		2nd		3rd		4th		5th		6th		7th		8th		9th		10th		Total	
	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}	F_{ST}	R_{ST}
Barrier 1	56	100	33	0	9	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	100	100
2	39	0	52	90	7	10	2	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100
3	3	0	6	10	28	90	25	0	27	0	8	0	2	0	1	0	0	0	0	0	100	100
4	0	0	4	0	22	0	39	71	24	10	9	13	1	5	0	1	0	0	0	0	100	100
5	2	0	4	0	32	0	18	9	22	37	11	28	9	17	3	2	0	1	0	0	100	100
6	0	0	1	0	1	0	13	5	23	30	5	29	11	20	0	15	0	1	0	0	100	100
7	0	0	0	0	0	0	1	1	1	3	6	12	28	36	26	41	5	7	4	0	71	100
8	0	0	0	0	1	0	1	0	1	0	6	1	21	5	30	25	10	57	8	12	78	100
9	0	0	0	0	0	0	0	16	0	22	0	11	0	12	3	10	37	16	39	10	79	97
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	5	1	12	10	49	11	67

The figures in each column represent the number of times each putative barrier ranked first, second, etc., over 100 randomization cycles.

considering respectively the R_{ST} or the F_{ST} distances. Only the seventh-ranking boundary considering the R_{ST} distances (or the tenth-ranking considering F_{ST}), roughly corresponds to the main partition of human populations identified by Rosenberg and colleagues [13] assuming the existence of $k = 2$ clusters.

In the first test of significance (AMOVA), based on random reallocation of individuals, 10 boundaries and two statistics were independently considered, and so a Bonferroni correction [26] brought the significance threshold to $p = 0.05/20 = 0.0025$. This analysis showed (table 1) that, considering R_{ST} distances, in seven cases out of ten the differences across barriers exceed those expected by chance, namely for the comparisons between Asia and the Americas (barrier 3); between subsaharan Africa and Eurasia (barrier 4); between Papuans and Melanesians (barrier 5); among East Asia, West Asia and Kalash (barrier 7); between Piapoco and Pima-Maya (barrier 9) and between Pima and Maya (barrier 10). The populations or groups separated by those significant barriers differ by a fraction representing between 2.7 and 12.6% of the global variance estimated at the 377 loci. Considering the F_{ST} distances, all boundaries remained significant except barrier 3 between East Asia and the America.

In the second test of significance (table 2), based on resampling of the loci, bootstrap values greater than 70% were observed for all barriers, except for the tenth-ranking barrier, which with R_{ST} distances was located between Pima and Maya (11%), and with F_{ST} distances separated East Asia from West Asia (67%).

In addition, we repeated the analysis using only the 100 most informative loci in the dataset, as defined in Ref. [13], and found that the results do not change qualitatively. The barriers were the same as previously inferred from the analysis of all loci using R_{ST} values, with just a few, minor differences in their relative ranking.

In summary, we initially identified about twelve zones of the world where genetic change is locally increased with respect to random locations on the map, either considering only allele frequencies or also taking into account the molecular distances between alleles. However, differences reached significance, with respect to both the loci and the individuals, only between eight pairs of populations (or groups thereof) considering R_{ST}, or nine considering F_{ST}. For the other putative boundaries, differences between the groups of populations did not exceed random expectations. These zones of significantly increased genetic change broadly define five genetic isolates (Kalash, Piapocco, Pima, Maya and Papuans), mostly in the Americas, and a few larger clusters of populations, which appear to differ substantially between the two analyses.

Discussion

This analysis of autosomal diversity shows that different clusters of populations can be identified in the same dataset, depending on the assumptions of the model. Under a model of divergence driven by genetic drift (and opposed by gene flow), which is implicit in the usage of Wright's F_{ST} statistics, we found nine groups, four of them in the Americas. Sharp differences among New World samples, even at close spatial distances, have been described by several authors [see e.g. 27], and are interpreted as a consequence of founder effects and strong genetic drift in populations occupying a highly-fragment-

ed habitat. A similar result was obtained using a completely different method by Corander et al. [28]. In addition, Africa seems subdivided in three clusters (which was not apparent in Rosenberg et al. [13]), in agreement with several studies [reviewed in 29] showing higher diversity in Africa than in other continents. F_{ST} shows no barrier between Western and Eastern Eurasia, so that a single, large cluster seems to encompass all populations from East Africa Bantus to the Pima of Mexico.

Sample sizes are small in several regions of the world, and notably in Africa and East Asia. As a consequence, it may be that we did not have enough statistical power to identify additional boundaries between those populations, and hence that the genetic landscape of Africa and East Asia is more fragmented than it appears in this study.

Under a model in which divergence is also due to the accumulation of STR mutations, implemented by estimating $R_{ST,}$ eight clusters are apparent. However, the main difference with respect to the previous model is not the number of clusters, but their scope. All populations from Subsaharan Africa form a single group, when analysed by R_{ST}. Western Eurasia is separated from East-Central Asia and Papua New Guinea by boundary 7 that closes on itself around the Kalash from Pakistan, thus defining a one-population cluster. This result is closer to the one obtained by Rosenberg et al. [13], although a greater level of subdivision is still evident in the Americas. F_{ST} and R_{ST} have different properties, and depending on the evolutionary scale of the study either one can be considered better. F_{ST} is known to reflect mostly events in the recent evolutionary history of populations, whereas the values of R_{ST} also depend on phenomena affecting the deepest branches of the evolutionary tree.

The CEPH cell lines are currently the largest resource available to study human genome diversity at the global level, but its coverage of the world populations is incomplete. India, Australia and Polynesia are not represented at all, as well as broad areas of the other continents. The inclusion of more samples from these areas may change the observed patterns and boundaries. As the number of populations typed increases, we would expect to find a greater number of differentiated isolates. Indeed, it has been shown that boundaries may be missed in non-sampled areas, whereas the opposite error is unlikely, i.e. recognizing a boundary in an area where genetic change is actually clinal [30].

But which is the real structure of human populations then? Even if we consider only the analyses of the CEPH diversity dataset, there is no single answer. Rosenberg et al. [13] concluded there are five major clusters, plus the Kalash as a genetic isolate, and confirmed their finding in a similar analysis of 993 loci in the same populations [31]. Corander et al. [28] analysed by a Bayesian Monte-Carlo Markov Chain approach the CEPH dataset. Besides showing that Structure may converge to different solutions when different values of k are predetermined, they found that more than six groups are needed to represent global human genomic diversity, with evidence for genetic isolates in South America, in addition to, once again, the Kalash. Serre and Pääbo [9] argued that these results could be largely accounted for by the discontinuous sampling design; they resampled individuals so as to approximate a random distribution of genotypes in the geographical space, and observed an increase of population differences with geographical distances, a pattern compatible with isolation by distance over much of the planet, without apparent biological barriers. Finally, Ramachandran et al. [32] also found a steady increase in genetic differentiation with geographic distances, suggesting genetic continuity between human groups. The present study, the only one looking explicitly for boundaries and testing for their statistical significance, showed two clusterings that do not correspond to any of the previously inferred ones.

Studies of different markers yield an even more complicated picture, where the only common element we can recognise is that each one is inconsistent with all the others [5, 16, 33]. The only way we see to interpret this contradictory set of results is to admit that its incongruences are not due to errors in the choice of the markers or of the methods, but rather represent a basic feature of human diversity. In other words, different genetic polymorphisms are differently distributed over the planet, and their distributions are not generally correlated. Clusterings are always possible, but the fact that two populations fall in the same cluster (or in different clusters) when described at loci A, B, C does not imply that they will fall in the same cluster (or in different clusters) based on loci X, Y, Z. In addition, differences between populations are often so subtle that the location of boundaries may change substantially even when the same data are analysed under different assumptions on the mutational model.

There is doubtless a geographic structure in human genome diversity; given a sufficient number of markers, allelic differences can reach significance between virtually any pairs of populations or groups thereof [15], including populations separated by very few kilometers [17, 34, 35]. The broad regions of genetic similarity observed in all studies presumably result from shared ancestry of popula-

tions that kept migrational contacts through most of their history. The additional boundaries identified in this study, and implicit in other studies, are the expected outcome of genetic drift in communities evolving in relative reproductive isolation, a consequence of geographical and/or cultural barriers. In the US, and in other areas where groups of very different geographical origin came in contact relatively recently and did not mix much, population-specific alleles occur even in groups of different ancestries sampled in the same locality [35, 36]. However, things are far more complicated at the world level, where processes of gene flow and admixture have been going on for millennia. The available genomic data show that zones of relatively rapid genetic change are scattered in an evolutionary landscape dominated by the continuous, clinal change resulting from isolation by distance [32]. An unambiguous clustering of humans in groups has so far proved impossible. Therefore, overemphasizing results that apparently suggest a deep subdivision of humans leads to an oversimplified view of human diversity, which can hardly be useful in biomedical research [37, 38].

Acknowledgements

Supported by a European Science Foundation OMLL grant through the Italian CNR and by funds from the University of Ferrara. We thank Pierre-Alexandre Landry for providing us with an upgraded version of MsatBootstrap.

References

1 Risch N, Burchard E, Ziv E, Tang H: Categorization of humans in biomedical research: Genes, race and disease. Genome Biol 2002;3: 1–12.
2 González Burchard E, Ziv E, Coyle N Gomez SL, Tang H, Karter AJ, Mountain JL, Perez-Stable EJ, Sheppard D, Risch N: The importance of race and ethnic background in biomedical research and clinical practice. N Engl J Med 2003;348:1170–1175.
3 Bamshad M, Wooding S, Salisbury BA, Stephens JC: Deconstructing the relationship between genetics and race. Nature Rev Genet 2004;5:598–609.
4 Mountain JL, Risch N: Assessing genetic contributions to phenotypic differences among 'racial' and 'ethnic' groups. Nat Genet 2004;36: S48–S53.
5 Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G: Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res 2002;12:602–612.
6 Cooper RS, Kaufman JS, Ward R: Race and Genomics. N Engl J Med 2003;348:1166–1170.
7 Kittles RA, Weiss KM: Race, ancestry, and genes: implications for defining disease risk. Annu Rev Genom Hum Genet 2003;4:33–67.
8 Calafell F: Classifying humans. Nature Genet 2003;33:435–436.
9 Serre D, Pääbo S: Evidence for gradients of human genetic diversity within and among continents. Genome Res 2004;14:1679–1685.
10 Lewontin RC: The apportionment of human diversity Evol Biol 1972:6:381–398.
11 Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL: An apportionment of human DNA diversity. Proc Natl Acad Sci USA 1997;94: 4516–4519.
12 Jorde LB, Watkins WS, Bamshad MJ Dixon ME, Ricker CE, Seielstad MT, Batzer MA: The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet 2000;66: 979–988.
13 Rosenberg NA, Pritchard JK, Weber JL Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: Genetic structure of human populations. Science 2002;298:2381–2385.
14 Excoffier L, Hamilton G: Comment on 'Genetic structure of human populations'. Science 2003;300:1877.
15 Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB: Human population genetic structure and inference of group membership. Am J Hum Genet 2003;72:578–589.
16 Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, Schork NJ, Risch NJ: Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet 2005;76: 268–275.
17 Manni F, Guerard E, Heyer E: Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by using Monmonier's algorithm. Hum Biol 2004;76: 173–190.
18 Cann HM: Human genome diversity. C R Acad Sci III 1998;321:443–446.
19 Simoni L, Gueresi P, Pettener D, Barbujani G: Patterns of gene flow inferred from genetic distances in the Mediterranean region. Hum Biol 1999;71:399–415.
20 Wright S: Isolation by distance. Genetics 1943; 28:114–138.
21 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. Genetics 1995;139:457–462.
22 Schneider S, Roessli D, Excoffier L: Arlequin ver. 2.0.: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.
23 Monmonier M: Maximum-difference barriers: An alternative numerical regionalization method. Geogr Anal 1973;3:245–261.
24 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data.. Genetics 1992;131:479–491.
25 Landry PA, Koskinen MT, Primmer CR: Deriving evolutionary relationships among populations using microsatellites and $\Delta\mu^2$: All loci are equal, but some are more equal than others. Genetics 2002;161:1339–1347.
26 Sokal RR, Rohlf, FJ: Biometry, ed 3. New York, Freeman and Company, 1995.
27 Mulligan CJ, Hunley K, Cole S, Long JC: Population genetics, history, and health patterns in native Americans. Annu Rev Genomics Hum Genet 2005;5:295–315.
28 Corander J, Waldmann P, Marttinen P, Sillanpaa MJ: BAPS 2: Enhanced possibilities for the analysis of genetic population structure. Bioinformatics 2004;20:2363–2369.
29 Tishkoff SA, Verrelli BC: Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 2003;4:293–340.
30 Barbujani G, Oden NL, Sokal RR: Detecting regions of abrupt change in maps of biological variables. Syst Zool 1989;38:376–389.
31 Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW: Clines, clusters, and the effect of study design on the inference of human population structure. PloS Genetics 2005;1:e70.

32 Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL: Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA 2005;102: 15942–15947.
33 Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB: Population genetic structure of variable drug response. Nat Genet 2001;29:265–269.
34 Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, Beckman G, Beckman L, Bertranpetit J, Bosch E, Bradley DG, Brede G, Cooper G, Corte-Real HB, de Knijff P, Decorte R, Dubrova YE, Evgrafov O, Gilissen A, Glisic S, Golge M, Hill EW, Jeziorowska A, Kalaydjieva L, Kayser M, Kivisild T, Kravchenko SA, Krumina A, Kucinskas V, Lavinha J, Livshits LA, Malaspina P, Maria S, McElreavey K, Meitinger TA, Mikelsaar AV, Mitchell RJ, Nafa K, Nicholson J, Norby S, Pandya A, Parik J, Patsalis PC, Pereira L, Peterlin B, Pielberg G, Prata MJ, Previdere C, Roewer L, Rootsi S, Rubinsztein DC, Saillard J, Santos FR, Stefanescu G, Sykes BC, Tolun A, Villems R, Tyler-Smith C, Jobling MA: Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet 2000;67:1526–1543.
35 Wooding S, Ostler C, Prasad BV, Watkins WS, Sung S, Bamshad M, Jorde LB: Directional migration in the Hindu castes: Inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet 2004;115:221–229.
36 Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE: Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 1997;60: 957–964.
37 Colhoun HM, McKeigue PM, Davey Smith G: Problems of reporting genetic associations with complex outcomes. Lancet 2003;361: 865–872.
38 Barnholtz-Sloan JS, Chakraborty R, Sellers TA, Schwartz AG: Examining population stratification via individual ancestry estimates versus self-reported race. Cancer Epidemiol Biomarkers Prev 2005;14:1545–1551.