Current Biology, Vol. 13, R134–R136, February 18, 2003, ©2003 Elsevier Science Ltd. All rights reserved. PII S0960-9822(03)00074-5

Human Diversity: Our Genes Tell Where we Live

Dispatch

Laurent Excoffier

A detailed genetic analysis of more than a thousand human subjects clusters them into five groups corresponding to major geographical regions. This new study shows that self-reported ancestry is a good predictor of one's genetic make-up.

Since the 18th century, there has been much controversy on how one should classify human individuals, and on what basis: physical appearance, skin color or, recently, genetic diversity. The first genetic data on blood group and protein diversity showed that racial categories based on quantitative traits were arbitrary, as the human populations could not be simply divided into a few categories, but rather formed a continuum resulting from the settlement history of our species [1]. More recent molecular studies have allowed us to detail this complex migration and expansion process and put it into a time frame [2–4]. Genetic studies have also shown that most of the genetic variability in our species is due to differences between individuals within populations, rather than to differences between populations [5,6]. This might be because human populations have not been independently evolving entities, but rather have maintained connections through the exchange of migrants; it also implies that the definition of a human population is somewhat unclear. Despite this ambiguity, most genetic studies have involved the comparison of gene frequencies among different samples assumed to be drawn from different population subdivisions.

The novelty of the recent work of Rosenberg *et al.* [7] is precisely that they have checked the validity of the population-sampling approach and tried to define the genetic structure of the human population without using *a priori* information on the geographic origin of the individuals. For that purpose, they used the *structure* program [8], which attempts to find, for each individual, the proportion of its genome that comes from a given 'population', whose unknown genetic constitution is estimated in the same process. This procedure is performed successively with the assumption of an increasing number of 'populations' or clusters (K): $K = 2, 3, 4$ and so on [8].

Rosenberg *et al.* [7] applied this procedure to 1056 individuals analyzed for 377 autosomal short tandem repeat (STR) loci. This data set is the first outcome of the analysis of a cell-line panel of 52 worldwide populations [9] managed by the French Center for the Study of Human Polymorphism (CEPH) in the framework of the Human Genome Diversity Project initiated by Luca Cavalli-Sforza. This is by far the largest multi-locus data set presently available for humans.

The results obtained by Rosenberg *et al.* [7] are quite remarkable. For $K = 2$ case, where it is assumed that there are two clusters, a contrast is found between individuals from sub-Saharan Africa and native Amerindians. Individuals from other regions seem to harbor various proportions of 'African' genes, with a tendency to a dilution of these genes with distance from Africa (Figure 1 in [7]). A gradient of decreasing levels of STR genetic diversity with larger distance from Africa has been mentioned before [4], and was interpreted as the result of a series of historical bottlenecks during the colonization process of the world, starting in Africa. The opposition between Africa and the Americas, also observed when comparing blood groups, allozymes and immunological marker frequencies [10], could thus result from the same process.

Also interesting is the observation that with two clusters, individuals found in populations from Africa, Europe, North Africa, the Middle East and the Indian sub-Continent (mainly Pakistani populations) present a large majority of their genes as coming from the same population, whereas genes from the other hypothetical population are at a majority in individuals from East-Asia, Oceania and the Americas. This first division of the world (Figure 1, barrier 1) is at odds with previous results where a first split has often been observed between sub-Saharan Africans and non-Africans [11–13].

Assuming that three populations are present ($K = 3$) leads to a split of individuals found in sub-Saharan Africa from those found in Europe, North-Africa, the Middle East and Pakistan (Figure 1, barrier 2). With $K = 4$, a cluster of Asiatic and Oceanian individuals separates from Amerindians (Figure 1, barrier 3). With $K = 5$, an Oceanian cluster appears (Figure 1, barrier 4), and we are left with the pleasant picture of a world divided into genetic clusters that closely correspond to five geographic regions: sub-Saharan Africa, East Asia, Oceania, the Americas and the rest, comprising Europe, North Africa and West Asia. With $K = 6$, a new genetic cluster made up essentially of individuals from a single Pakistani population emerges, showing that with the invocation of further clusters, single populations with peculiar allele frequencies stand out, probably because of isolation and founder effects.

It thus seems that these five groups do correspond to major subdivisions of the human population. Rosenberg *et al.* [7] then attempted to examine further the internal genetic structure of these subdivisions. Sub-Saharan Africa presents clear additional levels of subdivisions, in keeping with previous results [14]. Amerindian populations also present substantial subdivisions corresponding to the five sampled populations. The other regions present less clear subdivisions, in the sense that the recovered populations do not correspond to collections of individuals found

Computational and Molecular Population Genetics Lab,
Zoological Institute, University of Bern, Baltzerstrasse 6, 3012
Bern, Switzerland. E-mail: laurent.excoffier@zoo.unibe.ch

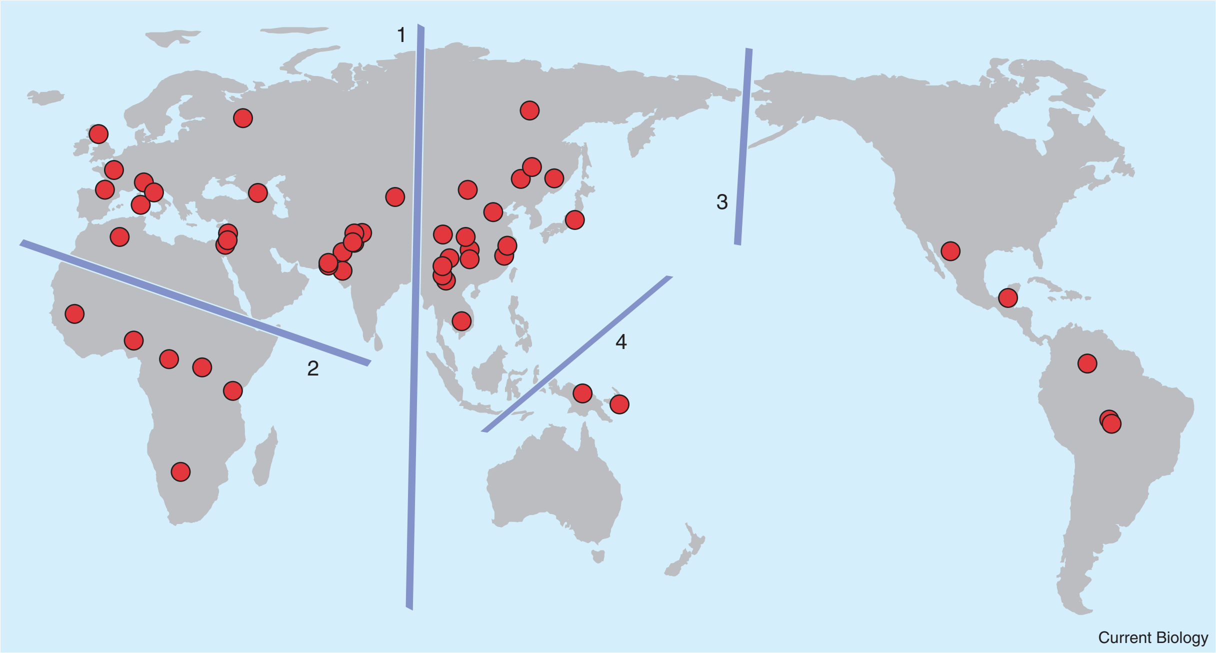

Figure 1. Geographical location of the 52 population samples studied by Rosenberg *et al.* [7].
The barriers numbered 1 to 4 correspond to the sequential partition of the sampled populations into genetic clusters.

at the same location, or that individuals have genes originating from several clusters.

The main conclusion of this work is that there is a very good agreement between the geographic and genetic assignment of individuals. The five major genetic clusters do correspond to five geographic regions. In other words, sampled sub-Saharan Africans are 'all' genetically sub-Saharan Africans and native Amerindians are 'all' genetically Amerindians. It would be highly misleading to conclude that Rosenberg *et al.* [7] have just rediscovered five basic races. The concept of race indeed assumes that members of a race are much more similar to each other than they are from members of other races: this is *not* what is found here. On the contrary, this study estimates that, if you consider two genes from two individuals in the same geographic region, they are *on average* only 4% more similar than two genes drawn from individuals belonging to different regions.

This figure is actually surprisingly smaller than what was found previously (9–13% [6,15]), and shows that identical STR alleles can be found in very distant individuals. It is only because a very large number of loci were studied that individuals could be correctly assigned to the five clusters. Previous attempts at assigning individuals to similar continental groups were much less successful [15,16]. For instance, about 30% of misclassifications were observed when only 21 biallelic markers were used [15], while Rosenberg *et al.* [7] showed that about 150 loci were needed to have five stable clusters at the world level, and thus correct assignments.

With these new methodologies allowing the apportionment of an individual genome to hidden populations or subdivisions, it would be tempting to carry out anthropological or epidemiological studies without care for the ethnic or geographic origin of an individual, with the hope that this assignment will be done later solely on genetic basis. This would probably be a mistake for the following reasons. First, while genetic assignment to global geographic regions *only* requires the genotyping of 150 markers, more loci are needed to resolve finer subdivisions [7], because differences between nearby populations are usually very small. Self-reported ancestry is thus much less costly and Rosenberg *et al.* [7] find it is often as accurate as large-scale genotyping. Second, grouping individuals according to their genotypes would be equivalent to creating 'pure' breeds in agronomy. It does not correspond at all to the real nature of human groups, which incorporate new immigrants each generation, and which are all made of individuals of mixed ancestry [17].

It is thus likely that statistically reconstructed populations do not correspond to real entities. The definition of these virtual entities actually depends on the sampling scheme, the number of genotyped loci and the variability of the markers used [7,8,16,18]. Finally, the definition of groups (case-control or others) in epidemiological studies on a pure genetic basis may be problematic, because disease susceptibility genes or genes controlling drug responses might interact with social or cultural factors that can be readily identified from simple queries, leading to potentially false genetic associations if missing [19].

The value of the *structure* approach arises precisely when confronting geographic and genetic information, such as recognizing populations and individuals of mixed ancestry. This approach can lead to more powerful case-control studies taking into account sample internal stratification [20], and can successfully identify individuals and populations with mixed ancestry [7,16,17], which can provide important insights concerning past migration and colonization events, and thus help to reconstruct the settlement history of our species.

References

1. Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1994). The History and Geography of Human Genes. (Princeton University Press).
2. Templeton, A. (2002). Out of Africa again and again. Nature *416*, 45–51.
3. Excoffier, L. (2002). Human demographic history: refining the recent African origin model. Curr. Opin. Genet. Dev. *12*, 675–682.
4. Harpending, H. and Rogers, A. (2000). Genetic perspectives on human origins and differentiation. Annu. Rev. Genom. Hum. Genet. *1*, 361–385.
5. Lewontin, R.C. (1972). The apportionment of human diversity. Evol. Biol. *6*, 381–398.
6. Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L. (1997). An apportionment of human DNA diversity. Proc. Natl. Acad. Sci. U.S.A. *94*, 4516–4519.
7. Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.
8. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics *155*, 945–959.
9. Cann, H.M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A. *et al.* (2002). A human genome diversity cell line panel. Science *296*, 261–262.
10. Cavalli-Sforza, L.L., Menozzi, P. and Piazza, A. (1993). Demic expansions and human evolution. Science *259*, 639–646.
11. Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J. (1988). Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. Proc. Natl. Acad. Sci. U.S.A. *85*, 6002–6006.
12. Nei, M. and Roychoudhury, A.K. (1993). Evolutionary relationships of human populations on a global scale. Mol. Biol. Evol. *10*, 927–943.
13. Watkins, W.S., Ricker, C.E., Bamshad, M.J., Carroll, M.L., Nguyen, S.V., Batzer, M.A., Harpending, H.C., Rogers, A.R. and Jorde, L.B. (2001). Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am. J. Hum. Genet. *68*, 738–752.
14. Yu, N., Chen, F.C., Ota, S., Jorde, L.B., Pamilo, P., Patthy, L., Ramsay, M., Jenkins, T., Shyue, S.K. and Li, W.H. (2002). Larger genetic differences within Africans than between Africans and Eurasians. Genetics *161*, 269–274.
15. Romualdi, C., Balding, D., Nasidze, I.S., Risch, G., Robichaux, M., Sherry, S.T., Stoneking, M., Batzer, M.A. and Barbujani, G. (2002). Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res. *12*, 602–612.
16. Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N. and Goldstein, D.B. (2001). Population genetic structure of variable drug response. Nat. Genet. *29*, 265–269.
17. Chikhi, L., Nichols, R.A., Barbujani, G. and Beaumont, M.A. (2002). Y genetic data support the Neolithic demic diffusion model. Proc. Natl. Acad. Sci. U.S.A. *99*, 11008–11013.
18. Cornuet, J.M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics *153*, 1989–2000.
19. Risch, N., Burchard, E., Ziv, E. and Tang, H. (2002). Categorization of humans in biomedical research: genes, race and disease. Genome Biol. *3*, 2007.
20. Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000). Association mapping in structured populations. Am. J. Hum. Genet. *67*, 170–181.