Classifying humans

Francesc Calafell

Unitat de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain.
e-mail: francesc.calafell@cexs.upf.es

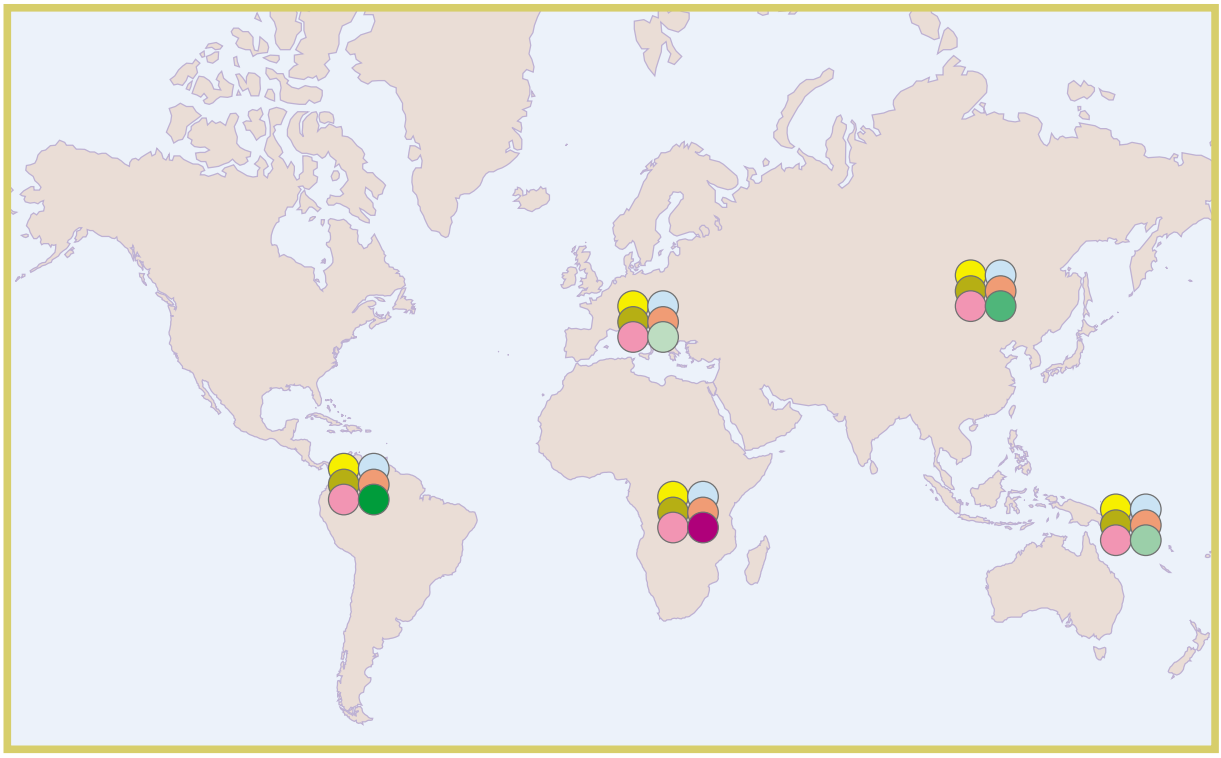
Recent papers have shown the feasibility of classifying humans into categorical populations from their genotypes. How can this be reconciled with the claim that human races are biologically meaningless, and what are the implications for medical genetics projects?

Recent papers in *Science*[1] and the *American Journal of Human Genetics*[2] have shown that genetic polymorphisms can be used to predict the population of origin of an individual. In both reports a large number of polymorphisms were genotyped in population samples from around the world, and a model-based clustering method[3] was used by the authors to ascertain how many distinct populations can be found in the global sample and estimate the probability that an individual belongs to one of these populations.

The algorithm used by the authors strips each sample of its original, self-reported population label and finds the groups of individuals that seem to cluster. Although the designers of the clustering method stated that these groups may not always have a clear biological interpretation[3], in both studies individuals in each group tended to fall into categories that corresponded to principal geographical (continental) divisions. At first glance, stating that the continental ancestry of each human can be identified seems to rehabilitate the discredited notion that humans can be classified typologically. Understanding what these studies really mean, however, requires a closer examination.

Classification the hard way

Assigning individuals to discrete groups requires a large number of polymorphic markers to be typed. Rosenberg *et al.*[1] estimated that the global classification would have required a minimum of 150 microsatellite polymorphisms per individual. Similarly, in the study by Bamshad *et al.*[2], 60 *Alu* insertion polymorphisms or 60 microsatellites were required to correctly group sub-Saharan Africans, Europeans and East Asians. It has been known for a long time that a mere 5–15% of the global genetic variation is accounted for by intercontinental differences (or F_{ST} for short; refs. 4–6). By genotyping such a large number of polymorphisms, the authors were able to delve into this component of genetic variation and accurately classify individuals into populations. The method used assigns each individual to a group with a probability, and often individuals have relatively high probabilities of belonging to more than one group. These probabilities could be interpreted as the fraction of each individual's genome coming from a group. Designation of broad continental ancestry, however, should be distinguished from socially defined ethnic groups residing in



BOB CRIMI

Global distribution of genetic diversity. Most genetic variation (shown here as color) is found within individuals of the same population, with a small fraction attributable to differences among populations.

the same territory, who are often classified as a race. Ascription to such groups is often thinly, or not at all, based on ancestry: for example, Hispanics in the United States are defined on the basis of language and may have ancestors from almost anywhere in the globe in various proportions.

Gene history

Interpopulation variation for each polymorphism may deviate from the average depending on a number of factors: type of marker and heterozygosity (for microsatellites, for instance, F_{ST} tends to decrease with heterozygosity), specific demographic factors, random variation and selection. This means that some polymorphisms may have large F_{ST} values and may be used more efficiently to ascribe individuals to continents[7].

Uniparentally transmitted markers (for example, mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome) present particular features relating to population structure and have been used extensively to infer population of origin, given that reasonable phylogeographies (the population distribution of haplotypes) are available for both genome segments. The Y chromosome is more geographically compartmentalized than is mtDNA, to the point that it can be used to track gene flow between closely related populations[8] or zero in on a particular population in a continent as a source of migrants[9]. Uniparental transmission, however, means that these loci capture a small fraction of overall ancestry: ten generations ago, in the absence of inbreeding, each individual had 1,024 ancestors but inherited mtDNA from only one of those. Thus, it is not surprising to find that 28% of Brazilians of light skin carry mtDNA sequences that originated in Africa[10]. Findings such as this reinforce the idea that features that may be used by some to define a 'racial' category have no relation to the overall genomic genotype.

Why it matters

Variation among continents in some loci can be explained by geographically restricted selection pressures, such as those related to climate or to some infectious diseases, such as malaria[11]. In fact, the right-hand tail of the F_{ST} distribution has been used to search for loci under selection[12]. Whether owing to drift or to selection, some genes with high F_{ST} values may be related to health issues, either as etiologic factors in common diseases or in drug metabolism[13]. Therefore, even if population genetic differences are small, they should not be ignored when trying to establish the genetic architecture of complex, common disease. Obviously, environmental factors may carry a greater weight in determining differences in prevalence of common diseases among socially defined ethnic groups.

One of the current main lines of attack for such a complex problem as the genetic basis of common diseases is a two-step approach: (i) define a minimum set of single-nucleotide polymorphisms (SNPs) that would capture the linkage disequilibrium landscape of the human genome and (ii) apply those SNPs to association studies. Human population genetic structure should be considered in both steps. The first step is being addressed by the HapMap project, in which millions of SNPs will be typed in samples of European, African and East Asian origin (see http://www.genome.gov/page.cfm?pageID=10005336). A careful consideration of sampling and population definition will be essential for this project. This division of humankind into three typologies would not capture a substantial fraction of genetic variation, both by ignoring subdivision within continents and by missing populations such as Native Americans and Oceanians that contribute as much as half of the global F_{ST} (ref. 2). This is particularly relevant as SNPs are being chosen with a frequency threshold set at 10% for the least frequent allele; a considerable fraction of common SNPs in Native Americans and Oceanians will probably be missed by the simplistic approaches to population designation[14,15].

General genetic differences between cases and controls may result in false positives in association studies. This can be avoided by carefully checking ancestry in cases and controls beyond the usual self-reported ethnic identities and by typing a routine battery of polymorphisms in both samples. Large differences in such polymorphisms can flag inadequate samples before the daunting task of whole-genome SNP typing is unleashed on them. Increasingly, medical genetics will rely on population-based approaches to studying disease. It is apparent that interpopulation genetic differences can be identified. Though they are small, these differences may be used to partly understand differences in disease risk among populations. The next natural level—differences among individuals—will pave the way for personalized medicine. □

1. Rosenberg, N.A. *et al. Science* **298**, 2381–2385 (2002).
2. Bamshad, M.J. *et al. Am. J. Hum. Genet.* **72**, 578–589 (2003).
3. Pritchard, J.K., Stephens, M. & Donnelly, P. *Genetics* **155**, 945–959 (2000).
4. Lewontin, R.C. *Evol. Biol.* **6**, 381–398 (1972).
5. Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L.L. *Proc. Natl. Acad. Sci. USA* **94**, 4516–4519 (1997).
6. Romualdi, C. *et al. Genome Res.* **12**, 602–612 (2002).
7. Shriver, M.D. *et al. Am. J. Hum. Genet.* **60**, 957–964 (1997).
8. Bosch, E. *et al.. Am. J. Hum. Genet.* **68**, 1019–1029 (2001).
9. Bosch, E. *et al. Hum. Genet.* (in the press).
10. Alves-Silva, J. *et al. Am. J. Hum. Genet.* **67**, 444–461 (2000).
11. Hamblin, M.T., Thompson, E.E. & Di Rienzo, A. *Am. J. Hum. Genet.* **70**, 369–383 (2002).
12. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. *Genome Res.* **12**, 1805–1814 (2002).
13. Wilson, J.F. *et al. Nat. Genet.* **29**, 265–269 (2001).
14. Carlson, C.S. *et al. Nat. Genet.* **33**, 518–521 (2003).
15. Reich, D.E., Gabriel, S.E. & Altshuler, D. *Nat. Genet.* **33**, 457–458 (2003).