Heredity (2008) 101, 471–472

www.nature.com/hdy

NEWS AND COMMENTARY

Genetic markers

How accurate can genetic data be?

L Chikhi

Heredity (2008) **101,** 471–472; doi:10.1038/hdy.2008.106; published online 1 October 2008

Molecular markers come in different flavours—blood groups, allozymes, RFLPs, AFLPs, RAPDs, STRs, SNPs, you name them. Whether the focus is on specific populations or on worldwide patterns (Cavalli-Sforza *et al*., 1994), genetic data have become prominent in recent decades and have fundamentally changed our views on human evolution and prehistory. But what if some of these markers were biased? What if genetic markers, far from being more objective than other types of data, were producing a distorted view of human diversity, and, as a consequence, of human origins? And if that were the case, would it be possible to identify the best and least biased data sets around? These important questions are at the heart of an article by Romero *et al*. (2008) recently published in *Heredity*.

In technical terms, the issue addressed by Romero *et al*. is called ascertainment bias and it has been around for some time (Garrod, 1902). It refers to a statistical bias introduced during the collection (or ascertainment) of data, and started to catch the eye of human population geneticists some 15 years ago (Bowcock *et al*., 1994). In population genetic studies the main cause of ascertainment bias is an economic one. Genetic markers are usually selected on the basis that they should be polymorphic (that is, variable) in a reference sample. Understandably, their costly development is rarely carried out on large samples and, once identified, it would be hard to imagine colleagues who would be happy to spend their research budget genotyping whole populations at markers for which most individuals will be identical.

The first and most obvious consequence of this selection process is that, by eliminating the least variable markers, genetic diversity is overestimated. In itself this is not necessarily a major problem, if one keeps track of the markers that were eliminated. A second consequence is that genetic diversity is usually inflated in the reference population/s as shown by Bowcock *et al*. (1994) in humans. This effect was particularly strong in Europe compared to other regions with nuclear RFLPs (restriction fragment length polymorphism), allozymes and blood groups, but weak or absent in microsatellites. They wrote that 'a reasonable explanation […] is the bias introduced by their initial selection in Europeans.' They added that this 'bias is likely to be less serious for markers with large numbers of alleles such as microsatellites'. Interestingly, this second ascertainment problem is very general. Using cattle and sheep, Ellegren *et al*. (1997) elegantly showed that microsatellite markers developed in one species produced shorter repeats and lower diversity estimates in the other species. Importantly, this explained why humans appeared to have longer microsatellites than other apes without invoking directional selection in humans.

A third and more subtle consequence of ascertainment bias arises even when the reference sample comprises individuals from the whole species range, as is the case in the protocols used for single-nucleotide polymorphism (SNP) discovery in humans. The critical issue is that the number of individuals in the so-called 'discovery panel' is usually very small. Thus, rare alleles tend to be missed and selected SNPs typically have alleles with similarly high or medium frequencies (SNPs are typically biallelic). This is problematic because many demographic events leave specific signatures in the allele frequency distribution. For instance, population bottlenecks tend to eliminate rare alleles, whereas expanding populations exhibit more loci with rare alleles. Similarly, directional or balancing selection also either favour one allele or maintain the allele frequencies at some equilibrium value, respectively. In other words, this type of ascertainment bias can mimic balancing selection or demographic bottlenecks. It can thus either generate false signatures or mask existing ones.

What makes the study of Romero *et al*. important is that they not only try to identify biases in genomic data sets but they also suggest a way to identify 'unbiased' data sets. As an example, Romero *et al*. cite a study by Ray *et al*. (2005) who tried to infer the region of origin of modern humans using a large single-tandem repeats (STRs) data set and massive spatial simulations. Ray *et al*. (2005) found that the most likely region of origin was North Africa, a region for which there was no known support from archaeological or anthropological data. Their guess was that a bias similar to that identified by Bowcock *et al*. (1994) was somehow shifting the centre of origins towards Europe or regions genetically close to Europe. After correcting for this bias, East Africa became the most likely region. Although Ray *et al*. (2005)'s final result is very sensible, Romero *et al*. were not fully convinced that the STR markers used were biased in any particular way.

Romero *et al*.'s results can be divided into three main points. First by comparing three existing genomic data sets, namely 783 STRs, 2834 SNPs and 210 insertion deletion polymorphisms (indels), they showed that there are significant differences between them, and hence not all may properly reflect human neutral diversity. Then, by generating a new set of 16 STR markers in the least biased way possible, they used these new STRs as a benchmark against which the three genomic data sets could be compared. Finally, their comparisons showed that the genomic data set least biased was the STR data that Ray *et al*. had used.

Does that mean, as the authors claim, that the 783 STR markers 'suffer no discernable bias'? We need here to go back to the selection process followed to generate the 16 STRs. Romero *et al*. actually started by identifying 70 independent STRs. The difficulty to obtain reliably amplifying loci led to the elimination of 46 loci. Among the 24 remaining loci, eight (one-third) proved to be nearly monomorphic, and were discarded from the rest of the analyses. It is thus fair to ask whether discarding these loci would not affect parameter inference beyond the obvious overestimation of genetic diversity in human populations. In fact, there are good reasons to think that this would create a bias when populations have either gone through a bottleneck or a population expansion, because the very proportion of monomorphic loci is providing us with information on such events as I noted above. This had already been noticed by Beaumont (1999) in a bottlenecked population and has since been confirmed on other real data sets. As a quick test I also

performed some simulations (not shown), in which I had a set of 24 loci from which I then selected two sets of 16 loci: one by discarding the eight least variable loci, and the other by discarding eight loci randomly. I found that in an admixture model the admixture proportions did not seem to be biased, whereas in the population size change models the selection of the 16 most variable loci seemed to produce biases for some parameters, but not all. Altogether the previous studies and these (admittedly very limited) simulations thus suggest that even the STRs identified by Romero *et al*. are likely to produce some biases.

To conclude, Romero *et al*. have clearly demonstrated that significant problems exist with both indels and SNPs, and they have also shown that the STRs are probably the best loci available today (but see Nielsen *et al*. (2004) for possible corrections for SNPs). One should probably take with a pinch of salt their claim that their STRs were unbiased or that the biases identified by Ray *et al*. (2005) were not real. But clearly, Romero *et al*.'s study is a significant step towards proper population genetics inference.

Dr L Chikhi is a CNRS researcher (UMR CNRS/UPS 5174) at the Instituto Gulbenkian de Ciência, Rua da Quinta Grande, no 6, P-2780-156 Oeiras, Portugal.

e-mail: chikhi@igc.gulbenkian.pt

Beaumont MA (1999). Detecting population expansion and decline using microsatellites. *Genetics* **153**: 2013–2029.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL (1994). High resolution of human evolutionary history trees with polymorphic microsatellites. *Nature* **368**: 455–457.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). *The History and Geography of Human Genes*. Princeton University Press: Princeton, NJ.

Ellegren H, Moore S, Robinson N, Byrne K, Ward W, Sheldon BC (1997). Microsatellite evolution—a reciprocal study of repeat lengths at homologous loci in cattle and sheep. *Mol Biol Evol* **14**: 854–860.

Garrod E (1902). The incidence of alkaptonuria: a study in chemical individuality. *Lancet* **2**: 1616–1620.

Nielsen R, Hubisz MJ, Clark AG (2004). reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168**: 2373–2382.

Ray N, Currat M, Berthier P, Excoffier L (2005). Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* **15**: 1161–1167.

Romero IG, Manica A, Goudet J, Lawson Handley L, Balloux F (2008). How accurate is the current picture of human variation. *Heredity* (advance online publication, 3 September 2008; doi: 10.1038/hdy.2008.89).