

Inferring the population of origin of an individual from DNA data

Inferenza sulla popolazione d'origine di un individuo basata su dati DNA

Peter Donnelly¹

Department of Statistics

University of Oxford

donnelly@stats.ox.ac.uk

The genetic composition of human populations differs in different parts of the world. Many of these differences are the result of chance events in the histories of the populations. Some reflect adaptation to geographically differing environments. With advances in genetic technology, large amounts of data are being generated which document genetic variation within and between populations.

Three related statistical issues arise in this context. The first, given data from different populations, is to develop good statistical models which capture the differences between the populations, and through estimating parameters in these models, to quantify those differences. In this setting, the data includes information about which sampled individuals are from which populations. In some contexts, this information is not available, and the second statistical challenge is to jointly infer the population membership of each individual and the genetic structure of the different populations. This same problem arises in connection with so-called cryptic population structure: the possibility that a single geopolitical population (such as that of the UK, or Italy, or China) is more accurately modelled as the union of an unknown number of distinct subpopulations. The third statistical challenge (which we will not consider further here) for either type of data is to go beyond simply describing the differences between populations or subpopulations and to infer aspects of their demographic history. These three questions have attracted considerable recent attention, motivated by the growth of informative data, interest in human history, and the consequences of such population structure for association studies for human diseases - in such disease studies, cryptic population structure acts as a confounding factor and leads to false positive associations (e.g. Marchini *et al.* (2004)).

The simplest setting occurs when the genetic data is taken from a collection of markers widely spread across the human genome. Effectively due to the process of genetic recombination in the ancestors of sampled individuals, the types of sampled individuals at different markers will be statistically independent. Hierarchical statistical models are natural here, and have proved very successful (see for example Nicholson *et al.* (2002), and the associated discussion, and Pritchard *et al.* (2000), when subpopulation membership is not available). In statistical terms the problem is tantamount to fitting a mixture model. One interesting difference in the genetic context results from the admixture of populations. The effect of this is that "chunks" of an individual's genome come from different ancestral populations, adding a further level of structure to the model, and correlations between markers on the same chromosome. For markers which are very close in genomic terms (typically within tens of kilobases of each other) there is a further correlation due to the shared ancestry of the region of DNA on which they are found.

¹Address for correspondence: 1, South Parks Road, Oxford OX1 3TG, UK

It is now becoming possible to extract and type either nuclear or more typically mitochondrial DNA from samples of forensic interest, for example from unidentified individuals, or individual remains, which have resulted from crimes, or possibly from the culprits. In the forensic context it can often help investigators to know the population of origin of an unidentified victim, or of a culprit. We will describe the general statistical problem, and new work on inferring population of origin from DNA sequence, or closely linked marker data, in this forensic setting.

References

- Marchini J., Cardon L., Phillips M. and Donnelly P. (2004) The effects of human population structure on large genetic association studies, *Nature Genetics*.
- Nicholson G., Smith A.V., Jonsson F., Gustafsson O., Steffansson K. and Donnelly P. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data, *J. Roy. Stat. Soc. B. (with discussion)*, 64, 695–715.
- Pritchard J., Stephens M. and Donnelly P. (2000) Inference of population structure using multilocus genotype data, *Genetics*, 155, 945–959.