Heredity (2009) 103, 362–363
© 2009 Macmillan Publishers Limited All rights reserved 0018-067X/09 $32.00
www.nature.com/hdy

NEWS AND COMMENTARY

On population history and SNP variation

From the detection of population structure to the reconstruction of population history: the historical reading of the human genome

H Laayouni and J Bertranpetit

Heredity (2009) **103,** 362–363; doi:10.1038/hdy.2009.76; published online 22 July 2009

Genetics as a discipline has allowed a sudden flourish in reconstructing the history of the human species, including that of past continental groups and specific populations. Beyond doubt, our historical knowledge has received an unexpected boost thanks to the surge of population genetics methods in the fields of archaeology, prehistory, historical linguistics and other historical sciences. Genetics tools have improved with time thanks to the description of new genome variants (classical genetic markers, microsatellites, single nucleotide polymorphisms or SNPs), and the development of more powerful, quicker and cheaper technologies for large-scale genome typing. These allow the description of genetic differences according to some external variable of membership, for example, a geographic place, an ethnic group, a linguistic entity or a political unit. The overall procedure is simple: take some kind of genetic measure of difference (such as a genetic distance) and investigate whether the difference in membership corresponds to the amount of genetic difference expected under a given demographic scenario of drift, population expansion, migration and admixture. Only a comparative population history can be revealed, that is, the past of a given population in relation to others, according to the amount of genetic differentiation achieved.

From time to time, individuals working on human population genetics are able to shed light on previously unsolved problems thanks to advances in genome analysis techniques and in the numerical methods employed. From the conventional genetic-distance studies on classical genetic markers—mostly developed by Cavalli-Sforza *et al.* (1994)—the main steps forwards have been the inclusion of detailed data on the non-recombinant regions of the genome (mtDNA and non-recombinant part of the Y-chromosome) and the recent advent of lots of SNP data produced by commercial whole-genome analysis arrays that have been made easy to interpret thanks to the development of powerful statistical tools. These include the popular software, Structure (Pritchard *et al.*, 2000), and the application of principal component analysis (Reich *et al.*, 2008) and similar multidimensional scaling (MDS, a special representation of data that can facilitate interpretation and reveal relationships between variables). A series of papers have been published in this area (and this will continue throughout 2009) since the seminal papers on European population (Lao *et al.*, 2008; Novembre *et al.*, 2008). The recent paper by Salmela *et al.* (2008) is an analysis of the SNP allele frequency of a whole-genome scan in Finland and Sweden to uncover population structure and relate it to population history. Will this new perspective fill the gaps of previous genetic analysis? Will this be the ultimate genetic analysis to uncover population history in Scandinavia? Let us look at it in more detail.

The aim of this work was to characterize the genetic variation of Finland and Sweden, comparing it with Northern Germany and Great Britain, on a finer level than was previously possible. In addition to analysing the patterns of population differentiation, diversity and admixture in North Europe, the authors have a special interest in elucidating population structure within Finland. Indeed, population history reconstruction is a by-product of population genetics achieved by reading in historical terms the findings of structure and differentiation of the units of population defined *a priori*.

The results of this study revealed greater than expected population structure within Finland, and a small but significant differentiation between all the populations ($F_{\mathrm{ST}} = 0.0040$), with the Germans and British appearing especially genetically homogeneous. The fact that the Germans and British are genetically close to each other is consistent with earlier observations (Seldin *et al.*, 2006; Bauchet *et al.*, 2007). However, the German, British and CEU (Utah residents with ancestry from Northern and Western Europe) samples analysed here formed a single cluster, contrary to studies with a more comprehensive sampling from Central Europe (Lao *et al.*, 2008; Novembre *et al.*, 2008), possibly due to the lack of neighbouring reference populations. An interesting caution arising from the work is the limitation of the widely used MDS representations; the wider spread on the MDS plot observed in Swedes, compared with the other populations, was supported neither by diversity calculations nor by a more detailed analysis, and was at least partly an artefact of the MDS, where the representation in a few dimensions probably fails to capture all aspects of the complex data.

In contrast to the low divergence between the British and German populations, both of which have high diversity, the genetic distances between the Swedes and Eastern and Western Finns were larger, and the diversity of these populations was lower. Moreover, the genetic difference between Eastern and Western Finland ($F_{\mathrm{ST}} = 0.0032$) is substantial on a European scale and higher than most differences among more distant populations. Eastern Finland presented especially extreme features, such as high linkage disequilibrium, high similarity within the population, increased numbers of monomorphic markers and divergence from the other population. These results are in accordance with earlier studies (Service *et al.*, 2006) and are likely to be caused by population history: the young age of the population, founder and bottleneck effects, and substantial genetic drift attributable to small population size.

Comparisons with Asian HapMap samples revealed an interesting difference between the studied populations; the Nordic populations and Eastern Finns, in particular, seem to harbour a significantly stronger Asian affinity than that of Central Europeans. A similar eastern influence has been observed in Y-chromosomal, mitochondrial DNA and autosomal studies of the Finns (Lappalainen *et al.*, 2006), consistent with archaeological and linguistic data. Therefore, the possible eastern contribution observed among the Finns supports and extends the earlier studies carried out with a more limited number of markers. This study provides a good example of the power of genome-wide data sets to uncover hidden population structures and illustrates the pressing need for using high-coverage polymorphism data to identify and delimit isolated populations—such as the Finns—before their use as homogenous populations in gene mapping analyses. These are undoubtedly interesting findings, but, besides questions of detail, there still remain two important issues.

First, are these findings robust? Would other genetic analyses (of other individuals, other genome regions and other genome polymorphisms) reveal the same findings? The answer to that is simple—all the findings are 'supported' by the data, but none have null and alternative hypotheses to test using a specific statistical test. Thus these results are plausible stories that have been constructed using a variety of data (including non-genetic) that seem adequate: there is still a wide scope for alternative explanations of the same data, open to sharp minds to re-interpret.

Second, could the future exploitation of genome-wide diversity data dramatically change what has been found with SNP and other marker analyses to date? These analyses are similar to what Cavalli-Sforza did many years ago, just with more polymorphisms and with somewhat higher certainty that markers are neutral; and the methods used have changed rather little. Overall, the approach is to interpret the dominant patterns of genetic differences in a simplified multidimensional landscape. There have been significant advances in molecular evolutionary analysis in recent years (including coalescent analysis, sequence evolution, departure from neutral models due to demography and recombination analysis); the future and the challenge for the field is to integrate these methods with new sources of data. Together, the use of statistical methods to test alternative hypotheses and the huge range of genetic information available (soon, full genome sequences) will provide the power for genetic tools to dissect patterns of historical demography effectively, helping to fill the gap between finely resolved genetic analysis and *ad hoc* population history reconstruction.

Dr H Laayouni and Professor J Bertranpetit are at the Institut de Biologia Evolutiva IBE (UPF-CSIC), Centro de Investigación Biomédica en red Epidemiología y Salud Pública (CIBERESP) IBE, CEXS-UPF-PRBB, Doctor Aiguader 88, Barcelona, Catalonia 8003, Spain

e-mail: hafid.laayouni@upf.edu

Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K *et al*. (2007). Measuring European population stratification with microarray genotype data. *Am J Hum Genet* **80**: 948–956.

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

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A *et al*. (2008). Correlation between genetic and geographic structure in Europe. *Curr Biol* **18**: 1241–1248.

Lappalainen T, Koivumäki S, Salmela E, Huoponen K, Sistonen P, Savontaus ML *et al*. (2006). Regional differences among the Finns: a Y-chromosomal perspective. *Gene* **19**: 207–215.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A *et al*. (2008). Genes mirror geography within Europe. *Nature* **456**: 98–101.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Reich D, Price AL, Patterson N (2008). Principal component analysis of genetic data. *Nat Genet* **40**: 491–492.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A *et al*. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. *PLoS ONE* **3**: e3519; doi:10.1371/journal.pone.0003519.

Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G *et al*. (2006). European population substructure: clustering of northern and southern populations. *PLoS Genet* **15**: e143.

Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G *et al*. (2006). Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* **38**: 556–560.

Editor's suggested reading

Belle EMS, Benazzo A, Ghirotto S, Colonna V, Barbujani G (2008). Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations. *Heredity* **102**: 218–225.

Chikhi L (2008). Genetic markers: how accurate can genetic data be? *Heredity* **101**: 471–472.

Relethford JH (2008). Genetic evidence and the modern human origins debate. *Heredity* **100**: 555–563.

Romero IG, Manica A, Goudet J, Handley LL, Balloux F (2008). How accurate is the current picture of human genetic variation? *Heredity* **102**: 120–126.