Editorial

The detection of regions of our genome under selection has increasingly relied on the use of genome scans, involving the genotyping of individuals from different populations with a large battery of markers. The markers linked to a region involved in local adaptation are expected to show a larger variance in allele frequencies across populations than neutral markers. Conversely, markers linked to a gene under balancing selection should show very similar frequencies between populations and thus lead to low levels of population differentiation (usually indicated by low F_{ST} values). This idea, originally put forward by Luca Cavalli-Sforza in the 1960s,[1] has only recently been applied on a large scale, allowing the detection of outlier loci from an empirical distribution of F_{ST} values computed over hundreds or thousands of loci.

Empirical tests of selection

A classical problem arising with any empirical distribution is that it necessarily has outliers — and therefore loci showing extreme F_{ST} values could be perfectly neutral. A potential way to remove, or at least lessen, this problem would be to build a large empirical distribution of F_{ST} based only on attested neutral markers, which would then serve as a reference for further tests of selection.

Unfortunately, this task is complex for several reasons: first, real neutral markers are difficult to identify, since any marker can be influenced by linked selected loci; secondly, the level of differentiation calculated by F_{ST} depends on the marker mutation model[2] and on the marker mutation rate;[3] thirdly, F_{ST} levels should only be compared for classes of markers having similar levels of diversity within a population;[4] and fourthly, F_{ST} levels usually depend on the composition of the population set, so that new markers to be tested should be genotyped on the same populations as those used to build the empirical distribution. It thus appears that empirical distributions are relatively poorly suited to assessing fine patterns of selection in the human genome or to discover the signature of recent adaptation (eg sickle cell anaemia in sub-Saharan Africa or lactose tolerance in northern European populations).

Model-based tests

In order to avoid the inherent problems of empirical distributions, it has been proposed to compare F_{ST} values with a null distribution generated under a given model of population evolution.[4,5] The advantages of this are that any locus can be tested independently for departure from neutrality — irrespective of previous analyses — and that the observed pattern of genetic diversity is only compared to simulated neutral diversity. A clear disadvantage of this approach, however, is that the shape and mode of the null distribution depends on the simulated evolutionary scenario. A model-based test of selective neutrality is, indeed, always a test of the assumed demographic scenario. For example, the pattern of genetic diversity within and among populations will be very different if populations are assumed to have remained constant in size or if they have gone through a recent bottleneck. It thus appears to be important for model-based approaches to be able to generate predictions under very realistic scenarios of human evolution, or simultaneously to estimate past demography and patterns of selection.[6] Some very simple models of human evolution have been used thus far — for example, a finite-island model with three representative populations of Africa, Asia and Europe assumed to be stationary. While this model is highly unrealistic, it has been shown to be remarkably robust — in the sense that more complex models did not really lead to very different null distributions of F_{ST}. It is nevertheless obvious that some specific demographic histories could mimic a given pattern of selection. For example, compared with a simple population split without change in population size, a bottleneck following population divergence would lead to a flatter F_{ST} distribution also shifted towards higher values. While all loci would be affected by such an event, it is clear that the power to detect adaptive selection at specific loci would be reduced. It is relatively certain that the past history of human populations has been much more complex than that, with a series of range expansions, bottlenecks, range contraction, re-expansions and demographic growth. The extent to which the distribution of genetic diversity would be affected by this complex historical demography remains to be shown.

Specific tests for different modes of selection

Genomic scans have now been used to try to detect genes involved in the recent adaptive selection which has occurred in non-African populations. The rationale is that, since modern humans are supposed to have left Africa some 60,000–100,000 years ago to colonise new territories, they should have simultaneously adapted to environments to which they had never been exposed. Under this view, no such new adaptive events are expected in African populations, since their environment remained relatively stable during that time. A number of studies have therefore been looking for loci with large associated F_{ST} values also showing reduced diversity within non-African populations. While recent adaptive events might well have occurred in Eurasia, one would nevertheless also expect many adaptive events to have occurred in Africa as a consequence of the global demographic increase having followed the transition from a hunter–gatherer way of life to farming and pastoralism. This is mainly for two reasons. The first is that selection is much more efficient in large than in small populations, implying that selective environmental pressures could have been neutralised by genetic drift. The second reason is that the sedentarisation and the densification of the populations offer more opportunities for transmissible pathogens to operate than in small, mobile and isolated populations. It is therefore believed that many new diseases have become much more prevalent in farming than in hunter–gatherer populations. Since there is a clear latitudinal gradient of pathogen density with a peak close to the equator,[7] large African populations living in inter-tropical regions should have been especially affected during this cultural transition, as exemplified by the many specific adaptations to malaria parasites that have recently appeared in Africa (eg haemoglobinopathies and the Duffy blood group). While searches for adaptive selection have been favoured in recent studies, evidence for balancing selection has been curiously neglected. This may perhaps be due to the fact that tests based on F_{ST} distribution have a low power to detect such a mode of selection,[8] when there is little differentiation among populations. In fact, the detection of significant reductions in levels of diversity would be much more efficient in a species showing higher levels of differentiation; for example, this is the case among common chimpanzee subspecies, which could also certainly shed light on similarly selected regions in humans.

Ascertainment bias

Since the efficiency of genome scans should be improved with a higher density of markers, scans based on large numbers of single nucleotide polymorphisms (SNPs) are desirable. Recent international and private efforts have produced an impressive list of millions of SNPs in the human genome, most of them being polymorphic in several populations. Due to the costs of SNP genotyping, however, it is likely that SNPs with very low frequencies of the minor allele will not be incorporated into future genome scans. It should be realised that such an ascertainment bias would enrich the distribution in markers showing low F_{ST} values and would thus lower the power for detecting genes involved in localised adaptive events, since local selective sweeps should lead to an increase in the frequency of some alleles around the selected genomic region. If ascertainment bias cannot be eliminated, it is important to document it precisely whenever possible, since it is often easy to take it into account when modelling the genetic diversity of SNP markers.

Some recommendations

The study of the genetic bases of adaptation and of the selective forces which have shaped our genome is of primary interest for understanding our past, and what makes us human. It is also important for medical genetics, where tools currently being developed for detecting adaptive selection could be used to detect genes involved in complex diseases by comparing cases and controls or individuals belonging to different phenotypic classes. Future genome scans potentially based on SNP chips will thus certainly become an important tool for discovering susceptibility genes for complex diseases. It therefore appears to be important that the potential problems associated with this method be clearly identified and minimised. The efficiency and adequacy of current genome scans could thus be checked and improved by a (non-exhaustive) series of measures, such as:

- Develop or use more realistic demographic scenarios of human evolution to eliminate some false-positive signals and perhaps increase the power of selective tests.
- Use many populations per geographical region to remove idiosyncrasies, lower the influence of particular populations and increase the potential to detect local adaptations.
- Develop or use statistics other than single-locus F_{ST}, since it has a large associated variance. For example, patterns of linkage disequilibrium among partially linked loci should also have some potential to detect genomic regions under selection.
- Do not only focus on adaptive selection, but improve the chances of detecting balancing regions as well — for example by performing studies in highly differentiated populations. The study of common chimpanzee populations could be useful in this respect, as we certainly not only share a large proportion of our genome with them, but also its selective constraints.
- Be sure that these methods can also detect regions known to be under selection (HLA, globin gene clusters etc).
- Take potential ascertainment bias in the choice of markers into account.

References

1. Cavalli-Sforza, L.L. (1966), ‘Population structure and human evolution’, *Proc R Soc Lond B Biol Sci*. Vol. 164, pp. 362–379.
2. Excoffier, L. and Hamilton, G. (2003), ‘Comment on Genetic Structure of Human Populations’, *Science* Vol. 300, pp. 1877b.

3. Flint, J., Bond, J., Rees, D.C. *et al.* (1999), 'Minisatellite mutational processes reduce F(st) estimates', *Human Genetics* Vol. 105, pp. 567–576.
4. Beaumont, M.A. and Nichols, R.A. (1996), 'Evaluating loci for use in the genetic analysis of population structure', *Proc. R. Soc. Lond. B* Vol. 263, pp. 1619–1626.
5. Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J. *et al.* (1994), 'High resolution of human evolutionary trees with polymorphic microsatellites', *Nature* Vol. 368, pp. 455–457.
6. Williamson, S.H., Hernandez, R., Fledel-Alon, A. *et al.* (2005), 'Simultaneous inference of selection and population growth from patterns of variation in the human genome', *Proc. Nat. Acad. Sci. USA*, Vol. 102, pp. 7882–7887.
7. Guernier, V., Hochberg, M.E. and Guegan, J.F. (2004), 'Ecology drives the worldwide distribution of human diseases', *PLoS Biol.* Vol. 2, pp. 740–746.
8. Beaumont, M.A. and Balding, D.J. (2004), 'Identifying adaptive genetic divergence among populations from genome scans', *Mol. Eco.* Vol. 13, pp. 969–980.

Laurent Excoffier
CMPG, Zoology Institute
University of Berne
Berne, Switzerland