Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 2582–2584, March 1999

Commentary

More on the X files

*Rosalind M. Harding**

Medical Research Council Unit of Molecular Haematology, Institute of Molecular Medicine, The John Radcliffe Hospital, Headington, Oxford OX3 9DS, United Kingdom

The study of diversity in the human genome has come far enough in the last few decades to give a sense of what to expect. However, when the unexpected turns up, there is an opportunity to learn something new. Eugene Harris and Jody Hey, who describe the pattern of allelic sequence polymorphism in part of the gene for the pyruvate dehydrogenase (PDH) *E1* α subunit, or PDHA1, in this issue of the *Proceedings* (1), have found something unusual. They report that the alleles at the PDHA1 locus completely segregate between samples of 16 Africans and 19 non-Africans and that the levels of diversity are entirely discordant between these samples, with diversity being much reduced in the non-Africans. These findings would not, at first glance, seem odd were the data from mtDNA or the Y chromosome. But PDHA1 is on the X chromosome, which, until this study, had shown the same kinds of patterns of polymorphism as loci from the autosomal chromosomes.

Genetic diversity in typical autosomal loci is not structured highly among human populations. The greatest proportion of genetic diversity, averaging over many autosomal loci, is found among individuals within populations and only about 10–15% is apportioned between populations. This pattern was described first after analyses of the classical protein markers and blood groups (2) and has since been observed over and over again in studies of DNA polymorphisms (3), including microsatellites (4) as well as craniometric variation (5). Recently, the same features have been reported for the dystrophin gene, an X chromosome locus (6). Although many of these same data also show that diversity within populations is greater for sub-Saharan Africa than elsewhere in the world, these differences are not large (4). In contrast to the patterns of apportionment for autosomal loci, variation at the PDHA1 locus between samples from sub-Saharan Africa and Eurasia accounts for 61.7% of total variation, and the level of variation among the sub-Saharan African individuals is 10-fold greater than that among the non-Africans.

The additional point to consider here is that the levels of diversity in autosomal and X chromosome genes, averaged over many loci, suggest expected coalescence times, under assumptions of neutrality and no recombination, of roughly 800,000 and 600,000 years (7), respectively. In comparison, coalescence times of less than 200,000 years have been estimated for diversity in mtDNA and the nonrecombining part of the Y chromosome, the loci that suggest some substantial population structure. Part of the reason that these loci do not have globally widespread distributions of common alleles is that their coalescent-time depths are too shallow. Another way to think about why haploid loci should be expected to show more structure is that they are subject to greater genetic drift. What makes the pattern of diversity at the PDHA1 locus unexpected is that this extreme structure is observed in a polymorphism with an estimated total coalescent-time depth of 1.86 million years. Thus, against the expectations suggested by most other polymorphism data, it is clear that something odd has happened—and perhaps is happening—at the PDHA1 locus.

It seems that PDHA1 diversity is, or has been, subject to a process of locus-specific selection. Such selection is entirely feasible, considering the function of PDHA1 (8). PDHA1 codes for a subunit of an enzyme essential for generating ATP from glucose oxidation. Mutations in this gene are recognized clinically as causing a range of disease phenotypes from severe to mild. Nearly all of these mutations occur in exons 6–11, overlapping the part of the gene (exons 7–10) that Harris and Hey have sequenced. "Mild" mutations lead to residual but deficient levels of PDH enzyme activity, compromising ATP production. Most tissues can compensate by generating ATP from alternative energy sources, such as protein and fat; however, the brain requires glucose to fulfill its energy needs, and insufficiency causes neurological impairment. In females, the level of ATP insufficiency depends not only on the particular mutation but also on the pattern of X chromosome inactivation, and Dahl (8) has suggested that there may be a significant number of undiagnosed, mildly affected, PDH-deficient females in the population. It also seems possible that mutations affecting function in the range of normal phenotypic variation could have arisen during hominid evolution (9). Harris and Hey observe that there is little evidence for recombination among the sequences they report; thus, any target site for selection easily could be outside of the region that they sequenced.

On the basis of a comparison between PDHA1 and β-globin in a commonly used neutrality test (10), Harris and Hey suggest that selection may account for the reduced diversity observed in the Eurasian samples. It may seem curious that β-globin was chosen as the standard in a neutrality test, because this gene provides one of the classic examples of selection (i.e., the sickle-cell hemoglobin variant that protects against severe malaria is caused by a mutation in the β-globin gene). However, although malarial selection has elevated frequencies not only of the sickle-cell mutation in the β-globin gene in Africa but also of a large number of thalassemia mutations in other regions of the world, levels of linked nonfunctional DNA diversity within and around the β-globin gene show no hint of perturbation. For example, both the total diversity in the β-globin gene and its variance among populations are entirely concordant with those measures from most other autosomal loci. There is still a lot to discover about the conditions under which selection on functional polymorphisms reduces or elevates diversity at nearby sites (7, 9).

For PDHA1, it may yet emerge that the apparently low diversity for Eurasians is increased substantially by adding more samples. Diversity could have been underestimated if there is further subdivision among non-African populations, as has been indicated by the dystrophin locus (6). However, even if the diversity is restored, selection still will be implicated by the unusual degree of population structure. This same reasoning was used to infer the impact of selection from the outliers of a large number of DNA polymorphisms analyzed for diversity and population structure by Bowcock *et al.* (3).

The companion to this Commentary begins on page 3320.

*To whom reprint requests should be addressed. e-mail: rharding@pinnacle.jr2.ox.ac.uk.

Most of the outliers showed unusually high levels of structure, as observed for the PDHA1 locus, but there were also some outliers that showed too little. Significantly reduced variation, both within and among populations, suggests a selective sweep. It may be that a selective sweep has reduced mtDNA variation, because, after correcting for haploidy to permit comparison with autosomal loci, mtDNA in fact shows a surprisingly low level of structure (11). The possibility that selection has made an impact on patterns of diversity in the Y chromosome requires further analysis. In a data set of single nucleotide polymorphisms analyzed from the Y chromosome, 52.7% of the total variation was apportioned between continents (12), a much higher proportion than that found for mtDNA. This level for the Y chromosome is also higher than a rough expectation of 36% for a haploid equivalent to 12.5% for autosomal loci. It seems that patterns of variation in PDHA1 have more in common with those observed for the Y chromosome than they have with those of autosomal loci.

Selection has played a part in the evolution of modern humans, for without selection there would be no adaptation. However, the impact of selection remains very poorly understood at the DNA level (9). The ability to recognize the footprints of selection on the genome would be valuable for supporting studies in functional genomics and taking their results one step further: from gene to biochemical pathway to the relative fitness of a phenotype in a population. The relevance of functional differences for an individual lies not just at the biochemical level but in how well his or her phenotype does in a given environment. Epidemiological analyses of polymorphisms that modify disease risk and protection are likely to have a few surprises in store (13).

The impact of selection is not the only inference that Harris and Hey make from their PDHA1 data. They conclude that the split between the African and non-African alleles at the PDHA1 locus dates to about 200,000 years ago. As they point out, most fossils from 120,000 to 190,000 years ago are considered to be transitional, showing mixtures of archaic and modern characteristics. Fully modern fossils only turn up toward the end of this period in east Africa; the oldest dates to about 130,000 years ago (14). The full range of fossils indicates a highly structured hominid population spread across Africa during the Late Middle Pleistocene period. Thus, it may be that different PDHA1 lineages can be traced back to different subpopulations from a date before the emergence of a fully modern morphology. This conclusion runs counter to the usual interpretation of human-diversity data, which maintains that structure was generated by the divergence of groups migrating into Eurasia within the last 100,000 years from a single ancestral source population of morphologically modern humans in sub-Saharan Africa.

The time scale for mutations generating PDHA1 diversity was estimated by Harris and Hey, who assumed a 5-million-year divergence time between humans and chimps and Kingman's coalescent model (15) for the population genetic history of the locus. The divergence time yields a mutation rate from a comparison of human and chimp sequences, assuming a molecular clock. The coalescent model estimates a compound parameter—the mutation rate multiplied by the effective population size (N_e)—from the human polymorphism data, as would other population genetic models. Dividing this compound parameter by the mutation rate provides a number for N_e. In a coalescent model, N_e is represented by a gene genealogy, connecting the 35 sampled copies of the PDHA1 gene back through their ancestral states and various mutation events to their single common ancestor. What is special about representing N_e with a genealogy rather than a single number is that such a model naturally gives a time scale, not only for the total coalescence time but also for the individual mutations. However, a number for N_e is still needed to convert the time units from the coalescent into years.

Many assumptions have to be made to impose time scales on genetic data. Furthermore, there is a big inferential leap between estimates for the ages of regionally localized alleles and for the onset of reduced gene flow between subpopulations. The uncertainty around all the assumptions is great, and, within this uncertainty, there are factors that increase estimates of ages, as well as those that decrease them. Let us consider a few of them. Humans and chimps underwent species divergence an estimated 5 million years ago. This estimate does not represent the split in the gene lineages, which may have occurred much earlier, depending on the size of the ancestral population. Further, the infinite-sites model underestimates total coalescence time if recurrent mutations can occur or if there is recombination. Likewise, random mating leads to underestimates of total coalescence times if there actually is population structure. A constant population size and a Wright–Fisher model of demography again lead to underestimates of total coalescence time compared with models assuming greater variance in the number of offspring. A finer point is that the coalescent model assumes strictly neutral variation, and an estimate of the mutation rate for these sites may be more appropriate than the evolutionary rate measured across the gene (16). If so, it is important to take into account functional constraints on the first and second positions in codons within exons and to remove them from the sequence over which the mutation rate is to be estimated when assembling divergence data. Applying this kind of correction to the PDHA1 data probably would increase the estimate for the mutation rate, decrease the estimate of N_e, and reduce the estimated coalescent-time depth from that reported. Within any such adjustments of the total coalescence time, the ages of the mutations cannot be predicted easily, because they are sensitive to the shape of the genealogy.

Until a lot more data are available, time estimates easily can range from half to twice as old, and this potential inaccuracy holds true for any of the estimates made from genetic data to date. These estimates certainly allow the possibility that humans living today inherited alleles from ancestors in a structured population 200,000 years ago. However, more importantly, uncertainty about the time scale changes neither the fundamental patterns observed in the PDHA1 data nor the conclusion that simplistic models of human evolution do not suffice to explain them. When the unexpected in science challenges those models that have served reasonably well, it may be that some calculations should be revised to incorporate the new data. However, there is also a chance that a different perspective and new models are needed. The most problematic simplification of current models of human evolution is that they are based on strict neutrality, with selection being dismissed other than to explain away unusual findings in the data. To make further progress, it will be necessary to integrate selection and its role in adaptation into the models that we use.

1. Harris, E. E. & Hey, J. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 3320–3324.
2. Lewontin, R. C. (1972) *Evol. Biol.* **6,** 381–398.
3. Bowcock, A. M., Kidd, J. R., Mountain, J. L., Hebert, J. M., Carotenuto, L., Kidd, K. K. & Cavalli-Sforza, L. L. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 839–843.
4. Jorde, L. B., Bamshad, M. & Rogers, A. R. (1998) *BioEssays* **20,** 126–136.
5. Relethford, J. H. & Harpending, H. C. (1994) *Am. J. Phys. Anthropol.* **95,** 249–270.
6. Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K. K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M., *et al*. (1998) *J. Mol. Evol.* **47,** 146–155.
7. Nachman, M. W., Bauer, V. L., Crowell, S. L. & Aquadro, C. F. (1998) *Genetics* **150,** 1133–1141.
8. Dahl, H.-H. M. (1995) *Am. J. Hum. Genet.* **56,** 553–557.
9. Eyre-Walker, A. & Keightley, P. D. (1999) *Nature (London)* **397,** 344–347.

10. Hudson, R. R., Kreitman, M. & Aguadé, M. (1987) *Genetics* **116,** 153–159.
11. Harpending, H., Relethford, J. & Sherry, S. T. (1996) in *Molecular Biology and Human Diversity*, eds. Boyce, A. J. & Mascie-Taylor, C. G. N. (Cambridge Univ. Press, Cambridge, U.K.), pp. 283–299.
12. Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. (1998) *Nat. Genet.* **20,** 278–280.
13. Williams, T. N., Maitland, K., Bennett, S., Ganczakowski, M., Peto, T. E. A., Newbold, C. I., Bowden, D. K., Weatherall, D. J. & Clegg, J. B. (1996) *Nature (London)* **383,** 522–525.
14. Lahr, M. M. & Foley, R. A. (1998) *Yearb. Phys. Anthropol.* **41,** 137–176.
15. Kingman, J. F. C. (1982) *Stochastic Processes Appl.* **13,** 235–248.
16. Crow, J. F. (1999) *Nature (London)* **397,** 293–294.