Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 11501–11503, September 1998

Commentary

The Chinese Human Genome Diversity Project

L. Luca Cavalli-Sforza

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

The Chinese population comprises one-fifth of the human species. The Chinese government officially recognizes 56 ethnic groups, one of which is the Han majority (1 billion and 100 million people), and the other 55 are ethnic minorities (totaling about 100 million). The latter are spread over most of China, but especially in the south. Close to half of the minorities are found in one of the 28 provinces of China, Yunnan. The distinction is primarily linguistic but corresponds closely to other cultural differences. The paper by Chu *et al*. published in this issue of the *Proceedings* (1) explores the genetic stratification of about half of the official ethnic subdivisions by means of microsatellites, a class of genetic markers recently discovered that has proved very useful for several purposes. The paper represents the collective effort of several institutes participating in the Chinese Human Genome Diversity Project (CHGDP). The broader Human Genome Diversity Project (HGDP) was generated in 1991 by the international Human Genome Organization (HUGO) and is regionally organized (see http://www.stanford.edu/group/morrinst/HGDP/html). The CHGDP has started collecting cell lines from the official ethnic groups and testing their DNAs. The 56 official ethnic groups do not exhaust current Chinese diversity, as there are more than 100 languages spoken in China, but they include the most important ones.

Microsatellites are repeats of short DNA segments, practically less than five nucleotides long. They have a high mutation rate and therefore a large number of alleles, which makes them perhaps three times more informative on average than the most common type of genetic polymorphisms, single nucleotide substitutions, which are mostly biallelic. They are used very widely in genetic linkage studies and have begun to be used in evolutionary analyses (e.g., refs. 2–4). Thirty microsatellites were tested by Chu *et al.* (1) for reconstructing a tree of 14 East Asian populations, which were studied along with 11 populations of a standard set representing the rest of the world. A subset of 15 of the same microsatellites were used to construct a second tree from 32 East Asian populations. These include the first 14 and are compared with the same 11 populations from the rest of the world.

Bootstrap (5, 6) values (measures of reproducibility of the tree branchings, varying from 0 to 100) are high in both trees for the fewer populations outside East Asia, which are rather remote both geographically and genetically from each other. These comparisons present the greatest genetic divergence, and their analysis by tree is therefore more reproducible. Results agree closely with a previous comparable analysis (2). The comparisons among East Asian populations involve much smaller genetic differences and, as expected, bootstrap values are much smaller. Because of their closer geographic proximity they are also likely to have had a much greater reciprocal gene flow than the more distant populations from the rest of the world. Studying populations much closer geographically and genetically puts analysis by tree to a more severe test. Even so, all East Asian populations cluster together in both trees. Their nearest genetic neighbors from the rest of the world are, not surprisingly, Native Americans. A little less close genetically is the small cluster formed by Australian aborigines and New Guineans, in agreement with the fact that Australia was settled before the Americas and had more time to differentiate (7, 8).

The first outlier within the East Asian cluster of the first tree is the Cambodian branch, and the second a small cluster made of two Altaic language-speaking populations (Buryat and Yakut). These populations live not too far from China, south and north of it, respectively. The other 11 East Asians form two fairly sharp clusters. One includes four Taiwan aborigines and two Chinese ethnic minorities from the western part of the Yunnan province. The other cluster includes Korean, Manchu, Japanese, and two groups of Han (one from Yunnan and the other from the United States). Usually, most Chinese immigrants to the U.S. (and to other countries, like Singapore, Malaysia, the Philippines, Taiwan, etc.) come from southern China, and this is certainly true of the cell lines from California residents from China born in the mainland, collected by Louise Chen and Alice Lin at Stanford and used in our surveys (2, 7, 8). Han living in the south of China mostly came originally from the north, but they did so at very different times, and thus had different times for gene flow from the earlier settlers, that is the minorities. In general, there is a correlation between the average genotype for protein polymorphisms of Hans from the different provinces and of local minorities, but there are exceptions (R. Du, H. Chungtze, E. Minch, and L.L.C.-S., unpublished work).

The second tree is based on more populations but fewer microsatellites, and the bootstraps are inevitably worse in the East Asian part of the tree. Conclusions therefore must be taken with greater caution. The southern group of populations falls into three clusters. S1 contains all four Taiwan aborigines and five Yunnan ethnic minorities. S2 contains Cambodians and six ethnic minorities from various southern provinces other than Yunnan, and also Han from the province of Henan, a north-central province on the north-south boundary. S3 is the tightest cluster and is made up of only two minorities, both from western Yunnan.

The northern group of populations falls into two clusters, N1 and N2. N1 is a classical northern cluster, with Japanese, Manchu, Korean, and Siberian. The Chinese are Han from the North—the northern Chinese by definition— and Han from the Yunnan, probably late immigrants who had no time to receive gene flow from the local people. There are also the Uyghur from the Xinjang province at the extreme west of China, who received a *ca.* 25% genetic contribution from ancestors of European origin, showing in their genes and, albeit qualitatively, in their phenotype and dresses (9). Their mummies, the oldest of which are from 3,800 years ago, show unquestionable evidence of European origins in their physical and cultural traits. They are probably descendants of people speaking Tocharian, an extinct Indo-European language. The residual 75% of their genotype must be from admixture with neighbors: 1% gene flow per generation (a very modest quantity) would be enough to cause the level of admixture observed (8).

The companion to this Commentary begins on page 11763.

N2 includes four minorities. Of these minorities, Evenki live in extreme northeast China but their origin is likely to be from Siberia. Tibetans are located in the southwest, but their origin from northern China is well established historically. The other two are minorities from a northern province and a south-central one. Strangely, N2 is part of the genetic cluster that includes all three southern groups, and in fact associates in the tree with S2. This finding is unexpected and requires an explanation. Chu *et al*. (1) acknowledge that statistical support of the N2-S2 relationship is weak and there may be a need of a greater number of microsatellites. Another possibility is the inappropriateness of a tree to represent a situation in which there is considerable admixture of the groups. Chu *et al.* have used the neighbor-joining method (NJ) of tree reconstruction (10), which has practical advantages, but it is hard to agree with their statement that NJ is "supposedly more robust in the presence of genetic admixture," except for the word "supposedly." In fact, I believe, on the basis of considerable simulation experience, published only in very small part (4), that admixture generates tree errors with NJ more easily than with other methods that we have tested. Chu *et al*. mention the possibility that the populations of cluster N2 were more exposed to southern admixture (excluding Ewenki).

Chu *et al.* draw a number of conclusions, the most general of which is: "It is now probably safe to conclude that modern humans originated in Africa constitute the majority of the current gene pool in East Asia." This should help refute the claim that there is a continuity of evolution from *Homo erectus* to modern humans in East Asia, as maintained by supporters of the multiregional hypothesis (11). The basis of this hypothesis came from paleoanthropological observations that have been criticized (12). Another stronghold of the multiregional hypothesis was the transformation of Neanderthal in modern humans in Europe, and also this has been falsified by an analysis of DNA of the Neanderthal *par excellence* (13).

Chu *et al*. strongly support the existence of a genetic difference between northern and southern Chinese, which, as mentioned in their paper, already was reached by a variety of other approaches, archeological, craniometric, and dental. The first genetic claim of this kind known to me is the demonstration of a strong difference in the frequencies of Gm markers (14). This is likely to be tied to a strong epidemiological difference. Other "classical" protein polymorphisms (blood groups, enzymes, and HLA) gave results very similar (8, 15) to those obtained with DNA markers in the present work.

Another source of information is surnames. They are transmitted like Y chromosomes and therefore may give results somewhat discrepant from those obtained by genes transmitted biparentally. Characteristics transmitted patrilinearly tend to be more highly clustered geographically than those transmitted matrilinearly like mtDNA and may be more useful on average than other DNA markers for reconstructing more ancient migrations (16).

In China surnames are particularly useful, being on average much older than in other parts of the world (15). In older times, however, some surnames were in part transmitted matrilinearly, as seems reasonable to infer from the presence of a female, or a male symbol in the characters of some older surnames, and from other more direct historical evidence. A China-U.S. team has analyzed surnames from a 1/2,000 random sample of the Chinese population, by standard techniques of population genetics, and the picture is largely superimposable on the genetic one. In fact, it is much more detailed given the magnitude of the sample and the number of "alleles" (surnames). The northern provinces are more homogeneous than the southern ones, among which three major subclusters seem fairly clear cut. The most distinct one is a group of four eastern provinces, including Shanghai. The far south is divided into two clusters. The three coastal provinces, Fujian, Guangdong, and Guangxi, form one, and the six others the rest. Tibet is not included in this analysis, for linguistic reasons. The greater geographic homogeneity of the north is shown especially by the difference between the linear regressions of the average distance between surnames on geographic distance. The slope of the northern provinces is at least four times smaller than that of the southern ones.

That the south of China is more heterogeneous than the north of China seems to be true without exception, from history to geography, ecology and culture, and now genetics. The greater heterogeneity of southern China is likely to reflect the greater geographic fragmentation of this area, resulting in greater isolation of local populations, probably mostly determined by the nature of the environment.

The surname border between north and south China is approximately intermediate between the two major rivers, the Yellow and the Yang-Tze. The discontinuity already is found in the paleolithic (17). Also neolithic developments were different and largely independent of each other in north and south China, probably for ecological reasons. Different plants and animals were domesticated. There is substantial agreement between archeological findings, genetic, and surname data.

At the end Chu *et al*. (1) discuss possible patterns of prehistoric expansions in East Asia, and in particular the question of whether people speaking Altaic languages originated from Middle Asia or East Asia. They give reasons why the latter seems preferable. As they acknowledge, their analysis suffers from lack of mid-Asian data. Nevertheless, their conjecture may be correct for another reason. Expansions from Africa to the rest of the world did not, or not necessarily, occur through the Middle East. When the earliest modern humans first settled the Middle East from Africa around 100,000 years ago, they had not yet developed the behavioral adaptations that helped them in their expansion out of Africa (18). They probably later abandoned the area, which was inhabited by Neanderthals around 60,000 years ago. But this is the most likely time when the major expansions of behaviorally modern human from Africa to Asia began. At least some of these may have started from nearer to the equator, perhaps from East Africa. If the European neolithic expansion can serve as a model of a much earlier one, it is useful to remember that it spread most easily along the coasts of the Mediterranean or along major rivers of central Europe. To settle Australia about 40,000 or 50,000 years ago (19), some navigation skills were necessary for crossing multiple tracts of sea (8). If such skills were already available to East Africans, the settlement of south Asia from East Africa might have begun along its southern coast, perhaps 10,000 years earlier or more (19). This would have given modern humans a chance to reach Southeast Asia fairly rapidly and from there, both Australia and East Asia, without major changes in food procurement techniques or climate adjustments. It also would favor the idea that Middle Asia was reached in the sequence Southeast Asia → East Asia → Middle Asia. From East Asia, Northeast Asia also could be reached and finally America.

It is very encouraging to see a cooperative effort of this magnitude beginning to take place in this most important part of the world, and Chu *et al*. are to be warmly congratulated for it. It is also important that their experience has made them aware that the number of markers must be greatly increased. This applies to practically every other paper recently published. For a long time, markers were simply not available, or difficult to study, but the situation is changing rapidly and very significantly. Bootstrap values demonstrate that large numbers of genetic markers are necessary for really solid conclusions. Variety of markers is also important (20). This shows that, in spite of the need of small amounts of DNA for PCRs, the strategy of collecting cell lines remains a necessary part of an HGDP program.

Whether one uses for research DNA extracted from blood, or other biological materials, including cell lines, there arise ethical problems that have been widely discussed. The North American Region of the HGDP has prepared a model ethical protocol (see http://www.stanford.edu/group/morrinst/HGDP/html and ref. 21), which examines these issues in great detail. The UNESCO International Bioethics Committee's Subcommittee on Bioethics and Population Genetics (see http://www.biol.tsukuba.ac.jp/~macer/PG.html and ref. 22), the Committee on Human Genome Diversity convened by the U.S. National Research Council (23), and the HUGO Committee on Ethical, Legal and Social Issues (see http://hugo.gdb.org: 80/conduct.htm) all have praised the model ethical protocol, while offering their own suggestions about appropriate ethical constraints on this kind of work. These issues obviously play a crucial role in such research everywhere in the world, although the exact ethical problems and solutions may differ among cultures.

1. Chu, J. Y., Huang, W., Kuang, S. Q., Wang, J. M., Xu, J. J., Chu, Z. T., Yang, Z. Q., Lin, K. Q., Li, P., Wu, M., *et al.* (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11763–11768.
2. Bowcock, A., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. (1994) *Nature (London)* **36,** 455–457.
3. Jorde, L. B., Bamshad, M. J., Watkins, W. S., Zenger, R., Fraley, A. E., Krakowiak, P. A., Carpenter, K. D., Soodyall, H., Jenkins, T., Rogers, A. R., *et al.* (1995) *Am. J. Hum. Genet.* **57,** 523–538.
4. Ruiz-Linares, A., Minch, E., Meyer, D. & Cavalli-Sforza, L. L. (1995) in *The Origin and Past of Humans as Viewed from DNA*, eds. Brenner, S. & Hanimara, K. (World Scientific, Teaneck, NJ).
5. Efron, B. (1982) *The Jackknife, Bootstrap, and Other Resampling Plans* (Society for Industrial and Applied Mathematics, Philadelphia).
6. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
7. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6002–6006.
8. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton).
9. Mair, V. (1998) *The Bronze Age and Early Iron Age Peoples of Eastern Central Asia* (Institute for the Study of Man, Washington, DC).
10. Du, R. F. & Xiao, C. J. (1997) *Science in China Series C* **40,** 613–621.
11. Wolpoff, M. H. (1989) in *The Human Revolution*, eds. Mellars, P. & Stringer, C. (Edinburgh Univ. Press, Edinburgh), pp. 62–108.
12. Lahr, M. M. (1996) *The Evolution of Modern Human Diversity: A Study on Cranial Variation* (Cambridge Univ. Press, New York).
13. Krings, M., Stone, A., Schmitz, R. W., Krainitzki, H., Stoneking, M. & Paabo, S. (1997) *Cell* **90,** 19–30.
14. Schanfield, M. S. & Gershowitz, J. (1973) *Am. J. Hum. Genet.* **25,** 567–574.
15. Du, R., Yuan, Y., Hwang, J., Mountain, J. & Cavalli-Sforza, L. L. (1992) *J. Chinese Linguistics, Monograph Series No. 5.*
16. Seielstad, M., Minch, E. & Cavalli-Sforza, L. L. (1998) *Nat. Genet.,* in press.
17. Chang, K. C. (1977) *The Archeology of Ancient China* (Yale Univ. Press, New Haven, CT).
18. Klein, R. (1999) *The Human Career: Human Biological and Cultural Origins* (Univ. of Chicago Press, Chicago), revised ed.
19. Connell, J. F. & Allen, J. (1998) *Evol. Anthropol*. **6,** 132–146.
20. Cavalli-Sforza, L. L. (1998) *Trends Genet.* **14,** 60–65.
21. North American Regional Committee, Human Genome Diversity Project (1997) *Houston Law Rev.* **33,** 1431–1473.
22. Macer, D., Fleming, J., Keyeux, G. & Knoppers, B. M. (1996) *Nature (London)* **379,** 11.
23. National Research Council, Committee on Human Genome Diversity (1997) *Evaluating Human Genetic Diversity* (Natl. Acad. Press, Washington, DC).