Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 1170–1172, February 1999

Commentary

Genetic isolates: Separate but equal?

*Leonid Kruglyak**

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109

The tiny volcanic island of Tristan da Cunha is the most remote inhabited place on Earth. It lies in the South Atlantic, 1,700 miles and a week's journey by boat from Cape Town, South Africa. The island's population of 300 is descended from a handful of founders, mostly shipwreck survivors, who settled there in the 19th century. The small number of founders is reflected in the fact that only seven surnames and five mitochondrial lineages (1) are found among the inhabitants. At first glance, this population couldn't be more different from that of Europe, inhabited for tens of thousands of years and home to more than half a billion people and dozens of ethnic groups. So it may come as a surprise that a study in this issue of the *Proceedings* (2) reports that these two populations, along with a wide range of others, are remarkably similar in one important respect—the level of linkage disequilibrium between nearby genetic loci. Taken at face value, this study has significant implications for human geneticists who seek to use increased linkage disequilibrium in isolated populations as a tool for unraveling the genetics of common diseases.

Linkage disequilibrium refers to nonrandom association of alleles at two loci. To illustrate, consider locus A, with two equally frequent alleles A1 and A2, and locus B, with two equally frequent alleles B1 and B2. If the distributions of alleles at the two loci were independent of each other, the four haplotypes A1B1, A1B2, A2B1, and A2B2 would occur with equal frequency. A deviation of haplotype frequencies from these proportions defines linkage disequilibrium. For example, allele A1 may occur predominantly on chromosomes that carry allele B1. This might be the case if the mutation that gave rise to allele A1 at locus A originally occurred on a chromosome that carried allele B1 at locus B, and if the two loci are sufficiently closely linked that recombination hasn't had time to randomize the haplotypes.

The case of interest in human genetics is when locus A is a previously uncharacterized disease susceptibility locus, and locus B is a known polymorphic marker. If allele A1 confers increased risk of disease, then the frequency of A1 will be higher among patients—as will the frequency of B1 if there is linkage disequilibrium between the loci. The observed increase in frequency of B1 among patients then can be used to infer the presence of a nearby susceptibility locus.

How closely linked must a marker be to show linkage disequilibrium with a susceptibility locus? The answer depends on the number of generations since the susceptibility allele first arose, because recombination tends to reduce linkage disequilibrium with each generation. In the case of common diseases observed worldwide, it has been hypothesized that many of the underlying susceptibility alleles are common gene variants that experience little selective pressure (3–5). Such alleles likely date back to before the "out of Africa" expansion of modern humans $\approx$100,000 years ago. Over the thousands of generations since that time recombination will have erased linkage disequilibrium except for markers very near the susceptibility locus. Sufficient marker density for detecting such susceptibility alleles by linkage disequilibrium thus may be difficult to achieve in practice, at least in the immediate future (unpublished work).

One approach for getting around this problem is to take advantage of special populations that might show increased linkage disequilibrium. Two types of populations have been proposed for this purpose: those that are genetically isolated by geography or culture (6, 7), and those with a history of recent admixture between different ethnic groups (8, 9). In such populations, small numbers of founders, population bottlenecks caused by various catastrophic events, genetic drift, as well as interbreeding between groups with different allele distributions may alter haplotype frequencies and thereby create linkage disequilibrium anew, in effect resetting the clock. Does this, in fact, occur? It is certainly true that increased linkage disequilibrium is seen in genetic isolates around rare disease mutations (10), but the situation is far less clear for common variants. For example, even a narrow bottleneck may admit a large enough sample of chromosomes carrying a common variant that haplotype proportions are not significantly altered (unpublished work). Whether linkage disequilibrium around common variants is increased in special populations is currently an open question, as the demographic history of most populations is not known in sufficient detail for precise theoretical predictions, and few experimental studies have compared linkage disequilibrium across populations.

Lonjou *et al.* (2) set out to address the question empirically. They cull data from the literature (11–13) on haplotype frequencies at two genomic regions in a wide range of populations and compare levels of linkage disequilibrium. The regions are the MNS blood group system on chromosome 4 and the Rh blood group system on chromosome 1. [Lonjou *et al.* also include data on the CD4 gene on chromosome 12 (see ref. 14), but because no information about haplotype frequencies is available for special populations, these data cannot be used to examine the question of interest and will not be considered further.] Linkage disequilibrium is computed for four pairs of loci: MN-Ss from the MNS system and C-D, C-E, and D-E from the Rh system. The populations examined include a number of subpopulations from each of eight major geographic regions: Europe, Near East, India and Pakistan, Far East, sub-Saharan Africa, the Americas, Oceania, and North Africa. Some of the subpopulations have histories of recent admixture between Europeans and another ethnic group. Also examined are six populations Lonjou *et al.* consider to be genetic isolates: Jews, Basques, Lapps (Saami), Eskimos, the Ainu of Japan, and the islanders of Tristan da Cunha.

Lonjou *et al.* (2) examine the pattern of average linkage disequilibrium at the blood group loci and make the following observations. First, sub-Saharan Africa has consistently lower levels of linkage disequilibrium than all other populations. This observation is consistent with previous studies (14) and most likely reflects relatively recent common descent of non-African modern humans from an ancestral African population.

The companion to this Commentary begins on page 1621.

*To whom reprint requests should be addressed. e-mail: leonid@fhcrc.org.

PNAS is available online at www.pnas.org.

Second, there is little variation in linkage disequilibrium among the other seven geographic regions. Third, there is little variation in linkage disequilibrium among subpopulations from a single geographic region. Fourth, linkage disequilibrium is only slightly higher in the isolates than in the large populations, with the exception of significantly higher levels in the Ainu (see below). Fifth, linkage disequilibrium is no higher (and often lower) in admixed populations than in the "parent" populations. Based on these observations, Lonjou *et al*. (2) conclude that "the utility of isolated or F_1 hybrid [admixed] populations for a genome scan by allelic association [linkage disequilibrium] has been greatly exaggerated."

The generality of this strong conclusion is tempered by several factors. A key concern is that the set of locus pairs chosen for comparing linkage disequilibrium in different populations may not be representative of the entire genome. The sample is small, consisting of only two genomic regions and four pairs of loci, and the results therefore are subject to the vagaries of the evolutionary history of these loci. These include the effects of stochastic forces as well as selection (the latter may play a role because the loci encode red blood cell antigens). In addition, recombination between two of the four locus pairs (MN-Ss and Rh D-E) is frequent enough that little linkage disequilibrium is observed in any population. Only the Rh C-D and C-E locus pairs display sufficiently broad ranges of linkage disequilibrium values to allow meaningful comparisons between populations. As a result, Lonjou *et al*.'s conclusions hinge on linkage disequilibrium measurements for two adjacent pairs of loci in a single genomic region. The order of the three Rh loci and the distances between them are uncertain, and the distances estimated by Lonjou *et al.* ($\approx$0.16 centimorgans for D-E and $\approx$0.02 centimorgans for both C-D and C-E) are inconsistent with each other for any locus order.

Further, although Lonjou *et al*. emphasize the similarities between populations, some differences are apparent even in the limited set of comparisons. (These differences are largely hidden when Lonjou *et al.* present values of linkage disequilibrium that have been averaged over all locus pairs, all large populations, or all isolates.) For each locus pair, the highest value of linkage disequilibrium is observed in one of the isolates: the Ainu for MN-Ss and Rh C-E, the Basques for Rh C-D, and the islanders of Tristan da Cunha for Rh D-E. The Ainu show significantly higher linkage disequilibrium for three of the four locus pairs, perhaps as a consequence of severe population bottlenecks that occurred when these original inhabitants of Japan were overrun by subsequent migrants to the islands. For the two locus pairs that allow meaningful comparisons across all populations, linkage disequilibrium between Rh C-D is strong in one group of populations (Europe, India and Pakistan, Near East, and North Africa), and weak in another (Oceania, Far East, and the Americas), whereas the exact opposite pattern is observed for Rh C-E. It is tempting to speculate that this observation reflects shared evolutionary histories among populations within each of the two groups, although the geographic groupings are too imprecise and world population trees are too uncertain (15) to confirm this hypothesis.

It is also important to remember that results about the strength of linkage disequilibrium in some populations can never be extrapolated to other populations that have not been examined directly. Every one of the world's many populations is characterized by a unique demographic history that leaves its mark on the pattern of linkage disequilibrium in that population. Even if it were true that linkage disequilibrium between most locus pairs is not significantly greater in Jews, Basques, Lapps, Eskimos, and the islanders of Tristan da Cunha than in large outbred populations, it would not be valid to extend this conclusion to all other isolates. Some isolated or admixed populations may show a stronger increase in linkage disequilibrium than others, and thus prove more suitable for gene mapping. Indeed, in the present study the Ainu show consistently higher linkage disequilibrium and may turn out to be a favorable population to study. Similarly, a previous study noted higher linkage disequilibrium in the Saami (Lapps) than in Finns, Estonians, and Swedes (16), although it is debatable whether this population will be well suited for gene mapping (2).

Caveats aside, the study by Lonjou *et al.* (2) provides a welcome empirical test of the hypothesis that linkage disequilibrium is increased in special populations. The major strengths of the study include considering a broad range of populations, measuring linkage disequilibrium between common biallelic polymorphisms that approximate common disease genes, and quantifying the strength of linkage disequilibrium by a natural measure for comparisons (kinship). Subject to the limitations discussed above, the results set the baseline expectation of at best modest increases in linkage disequilibrium between most locus pairs in most special populations; this expectation now needs to be confirmed or refuted by stronger data.

What data are needed? A good starting point would be to measure linkage disequilibrium for a much larger number of locus pairs in a similar range of populations. Data must be collected on enough locus pairs to adequately sample the genome and assess the variability in levels of linkage disequilibrium from one genomic region to another. The spacing between pairs of loci should cover a range of distances sufficient for characterizing the relationship between linkage disequilibrium and inter-locus distance in each population. These initial aims could be accomplished by measuring linkage disequilibrium for a few dozen locus pairs scattered throughout the genome, with inter-locus distances in the range of 10–100 kb. Linkage disequilibrium should be measured in large outbred populations from the world's major geographic regions and in several isolated and admixed populations, including the isolates considered by Lonjou *et al.* as well as other special populations proposed for gene-mapping studies (a partial list includes Finns, Icelanders, Sardinians, different Anabaptist sects, African-Americans, and Mexican-Americans). Agreement on a common set of reference locus pairs would greatly facilitate this effort, because results generated by different researchers working on different populations could be compared directly.

In the longer term, this effort should be expanded in two directions. Linkage disequilibrium should be measured across a dense set of loci covering the entire genome. The number of populations examined should grow to include as many of the world's populations as possible. Such a project would be a natural marriage between the Human Genome Project, with its new emphasis on genetic variation (17), and the Human Genome Diversity Project (18, 19). The result can be viewed as a linkage disequilibrium map of the human genome. This map would chart populations (and genomic regions) with strong and weak linkage disequilibrium. The potential impact of such a map on population studies can be imagined by considering how detailed genetic and physical maps of the human genome have revolutionized family studies of disease genes over the past decade.

Early maritime explorers, such as the Portuguese navigator who first sighted Tristan da Cunha in 1506 and named it after himself, wandered the world's oceans more or less at random, at the mercy of winds and currents. This was not by choice; rather, they were hampered by a lack of accurate maps and navigational devices. Human geneticists who seek to use linkage disequilibrium for disease gene mapping currently find themselves in a similar situation. The distribution of linkage disequilibrium across the human genome in different populations is largely uncharted waters. Armed with a linkage disequilibrium map, geneticists will know whether special populations hold the key to mapping common disease genes and will be able to plot an appropriate course.

I am grateful to B. Coller, H. Coller, S. Kruglyak, C. Laird, and A. Motulsky for comments on the manuscript.

1. Soodyall, H., Jenkins, T., Mukherjee, A., du Toit, E., Roberts, D. F. & Stoneking, M. (1997) *Am. J. Phys. Anthropol.* **104,** 157–166.
2. Lonjou, C., Collins, A. & Morton, N. E. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 1621–1626.
3. Lander, E. S. (1996) *Science* **274,** 536–539.
4. Collins, F. S., Guyer, M. S. & Chakravarti, A. (1997) *Science* **278,** 1580–1581.
5. Risch, N. & Merikangas, K. (1996) *Science* **273,** 1516–1517.
6. Sheffield, V. C., Stone, E. M. & Carmi, R. (1998) *Trends Genet.* **14,** 391–396.
7. Terwilliger, J. D., Zollner, S., Laan, M. & Paabo, S. (1998) *Hum. Hered.* **48,** 138–154.
8. Stephens, J. C., Briscoe, D. & O'Brien, S. J. (1994) *Am. J. Hum. Genet.* **55,** 809–824.
9. Chakraborty, R. & Weiss, K. M. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9119–9123.
10. Jorde, L. B. (1995) *Am. J. Hum. Genet.* **56,** 11–14.
11. Roychoudhury, A. K. & Nei, M. (1988) *Human Polymorphic Genes: World Distribution* (Oxford Univ. Press, New York).
12. Tills, D., Kopec, A. C. & Tills, R. E. (1983) *The Distribution of Human Blood Groups and Other Polymorphisms* (Oxford Univ. Press, Oxford).
13. Mourant, A. E., Kopec, A. C. & Domaniewska-Sobczak, K. (1976) *The Distribution of Human Blood Groups and Other Polymorphisms* (Oxford Univ. Press, London).
14. Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M., *et al.* (1996) *Science* **271,** 1380–1387.
15. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, NJ).
16. Laan, M. & Paabo, S. (1997) *Nat. Genet.* **17,** 435–438.
17. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L. & the members of the DOE and NIH planning groups (1998) *Science* **282,** 682–689.
18. Harding, R. M. & Sajantila, A. (1998) *Nat. Genet.* **18,** 307–308.
19. Weiss, K. M. (1998) *Genome Res.* **8,** 691–697.