Commentary

Predicting the range of linkage disequilibrium

Jurg Ott

Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10021-6399

A much pondered question in the Middle Ages was, ''How many angels can dance on the head of a pin?'' Today, we pose more sensible questions but answers are not necessarily getting easier.

A generally accepted strategy in the mapping of a disease gene is to initially apply linkage analysis for an approximate estimate of the location of the trait gene and to subsequently make use of linkage disequilibrium (association) for a more accurate localization. The thinking behind this approach has been that disequilibrium extends over much shorter distances from a disease gene than does linkage. On the other hand, biotechnology companies are gearing up to develop large numbers of single nucleotide polymorphism (SNP) markers (1) to localize disease genes by the disequilibrium mapping approach alone, for example, in case-control studies. It is then of interest to know how many such markers will be required on a genome-wide basis. Thus, the question is, how rapidly does disequilibrium decay as one moves away from a disease locus? The answer to this question is obviously of more than academic interest. Whether several 1,000 or several 100,000 SNP markers are needed will have a major impact on such association mapping studies. An early prediction was that in large outbred populations, disequilibrium should be detectable within 100 kb of a disease locus (2). So, it is timely that an extensive study addresses this important point in a recent issue of PNAS (3).

Collins and colleagues (3) collected information from the literature on autosomal haplotypes identified through family studies. They focused on haplotypes for pairs of loci, with the two members of a pair being a disease and a marker locus or two marker loci. Markers with multiple alleles were dichotomized into two-allele systems for a unique estimate of association designated by ρ. Distances (d in kb) between loci were determined on the scale of the physical map. For a disease gene with uncertain map position, its map location was estimated under the Malecot model with the aid of the ALLASS program (4). This approach previously has been shown to provide excellent estimates for the locations of trait genes (5). All in all, approximately 250 locus pairs were worked up.

Table 1. Sample data on linkage disequilibrium (ρ) and distance (d, in kb) between disease genes and nearby SNP markers

Disease gene	Population	ρ	d	χ_1^2
Huntington disease (14)	Canada	0.14	175	12.15
		0.11	175	8.05
Huntington disease (15)	England	0.29	175	11.93
		0.28	175	11.06
Cystic fibrosis (16)	Caucasians	0.19	9	6.15
		0.51	19	46.23
Limb girdle muscular dystrophy (17)	Reunion Islands	0.66	50	25.02
	Amish	0.80	50	33.63
Hemochromatosis (18)	Utah	0.80	0	164.68

Different classes of locus pairs were distinguished, in particular, those containing a disease locus and pairs of random SNP loci. For closely linked loci, the Malecot model allowed for the multiple-pairwise estimation of model parameters. One of these parameters, ε, depends on the number of generations since formation of the haplotypes and on the ratio between physical and genetic maps. The so-called swept radius, $1/\varepsilon$, estimates the distance in kb at which association falls to approximately 1/3 of its original level. Interestingly, this distance turns out to be very similar for disease and random haplotypes. For disease haplotypes, the swept radius is estimated between 300 and 500 kb and for random haplotypes it is somewhat smaller than 300 kb. The study (3) concludes with the suggestion that the number of SNPs required for a genome scan might be on the order of 30,000 or less.

On the basis of computer simulations, a recent progress report (6) predicted an extremely short range of useful disequilibrium, only about 3 kb. The report met with widespread skepticism even though it appeared in a so-called high-visibility journal. These predictions are clearly contradicted by observed data (3). The main reason for the discrepancy appears to be that the simulations were carried out under the assumption of a continuously expanding human population up to its present size of 5 billion, which seems unrealistic (3). A more likely scenario is that various bottlenecks and cycles of expansion and contraction have occurred in human history. Thus, it is reassuring that the study in a recent issue of PNAS (3) projects a required number of 30,000 SNPs rather than the figure of 500,000 resulting from the computer simulation.

Isolated populations often are considered advantageous for association mapping (7) but some examples have been found in which the extent of linkage disequilibrium is the same in small isolated and large outbred populations (1). A previous investigation of this situation concluded that isolated populations are to some, relatively small degree favorable for association analysis (8). On the other hand, there are well-known instances of strong disequilibrium in small populations (9). It seems to depend very much on the history of populations, be they large or small, whether disequilibrium will be extensive around disease loci.

In their figure 1, Collins and colleagues (3) show a graph of the association, ρ, versus distance, d, between loci on a haplotype. Visual inspection of the graph reveals two interesting features. First, at least at small distances, there seem to be two clusters of haplotypes, one corresponding to low and one with high association. Further, in each cluster, there does not seem to be a strong relationship between association and distance for approximate values of $d < 150$ kb. This finding confirms previous observations that disequilibrium and physical distance

See companion article on page 15173 in issue 26 of volume 96.

do not correlate significantly when $d < 60$ kb (10). My immediate reaction to these two clusters was that they correspond to isolated versus general populations. Thus, I obtained data for a few examples of each from the quoted web site (3). Table 1 shows the observed association values between disease and closest marker loci at very small distances. At least on the basis of this small sample, it appears quite convincing that increasing population isolation is more or less correlated with increasing association.

Special applications of disequilibrium mapping previously have furnished some spectacular results. A much-quoted example is that of the locus for diastrophic dysplasia, which was predicted to be 60 kb from the best marker (11) and was localized at a distance of 70 kb from it (12). On the other hand, the same method has been much less successful in other instances (13). A very dense map of SNPs can be expected to yield rather accurate results. Of course, if candidate genes are available for a trait, analysis of SNP markers in or very near these genes provides a cost-effective solution.

1. Jorde, L. B., Watkins, W. S., Kere, J., Nyman, D. & Eriksson, A. W. (2000) *Hum. Hered.* **50,** 57–65.
2. Bodmer, W. F. (1986) *Cold Spring Harb. Symp. Quant. Biol.* **51,** 1–13.
3. Collins, A., Lonjou, C. & Morton, N. E. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 15173–15177.
4. Collins, A. & Morton, N. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1741–1745.
5. Lonjou, C., Collins, A., Beckmann, J., Allamand, V. & Morton, N. E. (1998) *Hum. Hered.* **48,** 333–337.
6. Kruglyak, L. (1999) *Nat. Genet.* **22,** 139–144.
7. Peltonen, L. (2000) *Hum. Hered.* **50,** 66–75.
8. Lonjou, C., Collins, A. & Morton, N. E. (1999) *Proc. Natl. Acad. Sci USA* **96,** 1621–1626.
9. Laan, M. & Pääbo, S. (1997) *Nat. Genet.* **17,** 435–438.
10. Jorde, L. B. (1995) *Am. J. Hum. Genet.* **56,** 11–14.
11. Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. & Lander, E. (1992) *Nat. Genet.* **2,** 204–211.
12. Hästbacka, J., de la Chapelle, A., Mahtani, M. M., Clines, G., Reeve-Daly, M. P., Daly, M., Hamilton, B. A., Kusumi, K., Trivedi, B., Weaver, A., *et al* (1994) *Cell* **78,** 1073–1087.
13. Vesa, J., Hellsten, E., Verkruyse, L. A., Camp, L. A., Rapola, J., Santavuori, P., Hofmann, S. L. & Peltonen, L. (1995) *Nature (London)* **376,** 584–587.
14. Andrew, S., Theilmann, J., Hedrick, A., Mah, D., Weber, B. & Hayden, M. R. (1992) *Genomics* **13,** 301–311.
15. Snell, R. G., Lazarou, L. P., Youngman, S., Quarrell, O. W., Wasmuth, J. J., Shaw, D. J. & Harper, P. S. (1989) *J. Med. Genet.* **26,** 673–675.
16. Kerem, B., Rommens, J. M., Buchana, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M. & Tsui, L.-C. (1989) *Science* **245,** 1073–1080.
17. Allamand, V., Broux, O., Richard, I., Fougerousse, F., Chiannilkulchai, N., Bourg, N., Brenguier, L., Devaud, C., Pasturaud, P., Pereira de Souza, A., *et al* (1995) *Am. J. Hum. Genet.* **56,** 1417–1430.
18. Ajioka, R. S., Jorde, L. B., Gruen, J. R., Yu, P., Dimitrova, D., Barrow, J., Radisky, E., Edwards, C. Q., Griffen, L. M. & Kushner, J. P. (1997) *Am. J. Hum. Genet.* **60,** 1439–1447.