Am. J. Hum. Genet. 61:810–812, 1997

INVITED EDITORIAL
What Is Significant in Whole-Genome Linkage Disequilibrium Studies?

Leonid Kruglyak

Whitehead Institute for Biomedical Research, Cambridge, MA

How will we find the susceptibility genes underlying complex human diseases? Genome scans for linkage have produced equivocal results. Currently, a great deal of hope is riding on the notion that systematic searches for linkage disequilibrium (LD) in isolated populations will provide a more powerful approach. The critical task of interpreting the significance of results from such studies poses thorny statistical problems, but—in contrast to linkage analysis—the issues have received little attention. An article by Durham and Feingold in this issue of the *Journal* takes a step toward remedying this situation.

LD mapping relies on the assumption that a single ancestral mutation is responsible for a large proportion of disease cases in a present-day population. The chromosome on which the mutation originally arose carried a particular set of marker alleles—the ancestral haplotype. With passing generations, this haplotype is whittled away by recombination (and possibly altered by mutation) but should be largely preserved in the region around the mutation. Detection of such a region of identity by descent (IBD) among affected individuals provides evidence that the region contains a disease gene. The strength of the evidence depends on the probability of IBD in a random (unrelated to disease) chromosomal region. The size of the conserved region decreases with the number of generations since the mutation was introduced into the population.

Note that the principle of LD mapping is the same as that of linkage analysis—affected individuals should, by virtue of their common ancestry, share alleles in a chromosomal region containing a susceptibility gene. In linkage analysis, the relationships among individuals are known and are directly exploited by tracing inheritance in families, while in LD mapping the relationships are more distant and typically are unknown, apart from estimates of the number of generations to a common ancestor and the degree of inbreeding. The line between the two types of studies is blurred in large kindreds with known but highly complex genealogies, for which traditional linkage calculations are not computationally feasible.

LD mapping has two potential uses: (1) to refine the location of a disease gene already mapped to a chromosomal region by a linkage study and (2) to map a disease gene de novo. The use of LD for fine mapping has received a lot of attention (Jorde 1995) and has seen a number of successful applications, beginning with the cloning of the cystic fibrosis gene (Kerem et al. 1989). Several quantitative methods for fine mapping have been developed (Hastbacka et al. 1992; Lehesjoki et al. 1993; Kaplan et al. 1995; Terwilliger 1995; Devlin et al. 1996) and used for improved localization and, in some cases, cloning of a number of disease genes. Rigorous analysis is complicated by the fact that a population represents a single example of a stochastic evolutionary process whose detailed properties are not known. It is therefore difficult to interpret the significance of the resulting statistics, or to place accurate confidence bounds on gene location, and we need a better handle on the statistical properties of the evolutionary dynamics. It would also be handy to have a truly multipoint method of analysis. Nonetheless, today's approaches work well in practice—although successful applications have been limited to simple Mendelian diseases.

By contrast, the use of LD mapping initially to localize disease genes has been limited to at most a few examples (notably by Houwen et al. 1994). Until now, the main reason for this has been technological—regions of LD are sufficiently small in all but the youngest populations that their systematic detection requires genotyping an impractically dense set of markers. However, new technologies are likely to make such dense scans feasible in the near future (Chee et al. 1996; Wang et al. 1996; Kruglyak 1997), and the key issues will shift to those of analysis and interpretation, just as they did for linkage studies with the emergence of genome scans a few years ago.

Durham and Feingold set out to address the issue of interpretation by considering the occurrence of false positives in a genome scan for LD in a young founder

Received August 7, 1997; accepted for publication August 13, 1997.

Address for correspondence and reprints: Dr. Leonid Kruglyak, Whitehead Institute for Biomedical Research, 1 Kendall Square, Building 300, Cambridge, MA 02139. E-mail: leonid@genome.wi.mit.edu

This article represents the opinion of the author and has not been peer reviewed.

0002-9297/97/6104-0007$02.00

population (or a large kindred). Specifically, they compute the probability that i of N distantly related affected individuals share by descent from a common ancestor at least one random chromosomal region (one that does not contain a disease gene) somewhere in the genome. They assume that one has a perfect method for recognizing regions of IBD and that each affected individual is m generations (meioses) removed from the common ancestor with independent lines of descent—that is, with no two individuals sharing a more recent common ancestor. They initially consider the special case of no inbreeding (single lines of descent) and then generalize the results to include inbreeding (d lines of descent from the common ancestor).

Under the assumption of independent lines of descent, computing the probability that i of N relatives share a given chromosomal region by descent is straightforward—it approximately follows a binomial distribution with parameter $d/2^m$. Extending this result to the probability of sharing a region somewhere in the genome requires techniques from the theory of stochastic processes, which are similar to those previously used to assess the issue of genomewide significance in linkage studies (Lander and Botstein 1989; Feingold 1993; Feingold et al. 1993; Lander and Kruglyak 1995). The resulting probability is given by equation (4) of Durham and Feingold for the general case that includes inbreeding. The calculation is approximate but agrees closely with empirical results from simulations.

What about real populations, where lines of descent are not independent? The authors test the applicability of their results through simulations and propose some general guidelines. In the absence of inbreeding, the best bet is to construct an approximate pedigree that matches the total number of meioses in the true kindred as closely as possible and then to use the calculation for that pedigree to estimate the false-positive rate. When inbreeding is present, the use of average values of m and d gives the best estimate. The authors do well to caution that the approximations are only as good as the best guess about population structure, that conservative parameter estimates—that is, less separation from a common ancestor and more inbreeding—should be used when this structure is unknown, and that common sense in interpreting results cannot be replaced with blind application of the formulas.

The calculations described so far assume a dense map of markers that provides perfect IBD information at every position in the genome. Durham and Feingold also address false-positive probabilities in two-stage studies in which the "hits" obtained in an initial sparse-map scan are followed up with a higher density of markers. They advocate the use of the genomewide dense-map probabilities for such studies. This is appropriate because the second stage almost invariably follows up the same regions that would have been picked up as false positives in a dense-map genomewide scan, and so the false-positive rates are virtually identical. The same recommendation has been made for whole-genome scans for linkage (Lander and Kruglyak 1995). It is important to note that when simulations are used to assess the significance of two-stage studies, the simulations must include the two-stage procedure of selectively increasing map density in initially positive regions—otherwise, the false-positive rate can be greatly underestimated.

We now come to the issue of establishing IBD. Affecteds can share markers across a region as a result of either identity by descent (IBD) or identity by state (IBS). A small probability of random IBD at the level observed is not strong evidence that a gene is present unless the probability of IBS is also small. Establishing IBD unambiguously for distant relatives can require a very dense map. For example, consider two fifth cousins ($m = 6$) who share alleles at 10 consecutive markers spaced at 1 cM across a 10-cM region, with each shared allele having a frequency of 25%. An exact calculation shows that, under the assumption of linkage equilibrium, the probability that this sharing reflects IBD is only 30%. The probability of IBD is higher if the region is shared by more than two relatives or if haplotype information is available, but caution should still be exercised. Ideally, we need an LD statistic that takes into account information from multiple partially informative markers in a probabilistic fashion instead of relying on unambiguous determination of IBD.

It is also important to consider the rate of background kinship in the population. In any real population, there is a nonzero probability that two apparently unrelated individuals in fact have a common ancestor and may therefore share regions of the genome by descent. The probability that chromosomes chosen at random from two members of a population are IBD at a given point in the genome is measured by the kinship coefficient. Kinship coefficients tend to be small ($<.001$) in large populations, but can be much higher ($>.01$) in genetic isolates characterized by a small number of founders (Bodmer and Cavalli-Sforza 1976), which are precisely the populations most likely to be useful for LD mapping. Suppose that a region of IBD is detected among affected individuals known to be related through a common ancestor, who is presumed to pass the disease. The sharing can arise either as a result of this relationship or as a result of background kinship—that is, because the individuals have other ancestors in common. If the relationship is very close (such as siblings), the former is much more likely, but as the relationship becomes more distant and the probability of sharing a region through the one common ancestor drops, sharing due to kinship begins to dominate. Put another way, the probability of sharing a random region does not decay to zero as the

number of generations away from the common ancestor increases, but rather approaches the level of background kinship.

As an illustration, consider 20 individuals, each of whom is connected through 10 lines of descent to a common ancestor 30 generations back. Using the results of Durham and Feingold, one can compute that some two of them will show IBD at a given point with probability $\sim 10^{-14}$ and somewhere in the genome with probability $\sim 10^{-10}$. These probabilities are likely to be smaller than the chance of IBS, unless the map covering the region is exceptionally dense and informative. They are also much smaller than the probability of IBD due to background kinship in any population. It is clear that when the probability of sharing a random region of IBD through a single common ancestor becomes negligible, even relatively small probabilities of IBS, as well as IBD due to background kinship, must be taken into account.

None of these caveats are meant as criticisms of the work of Durham and Feingold, which represents an important advance. It will be interesting to see the extent to which their ideas can be generalized to the cases of larger, older populations, more intense and complex inbreeding, incomplete IBD information, and background kinship. The ability to assess the significance of findings rigorously is essential if LD mapping is to achieve its full potential. Perhaps a deeper concern is precisely what this potential is. For simple Mendelian diseases, there is ample evidence that LD is detectable in a number of populations. There is little or no comparable evidence for common, genetically complex diseases. Locus and allelic heterogeneity can undermine the basic assumption of LD mapping—that a significant fraction of today's disease chromosomes derive from a common ancestor. Will the allelic complexity of common diseases turn out to be sufficiently low for LD mapping to work even in young, isolated populations? Only empirical studies are likely to answer this question, and great care in interpretation will be required if reality is to be distinguished from wishful thinking.

References

Bodmer WF, Cavalli-Sforza LL (1976). Genetics, evolution, and man. WH Freeman, San Francisco

Chee M, Yang Y, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–614

Devlin B, Risch N, Roeder K (1996) Disequilibrium mapping: composite likelihood for pairwise disequilibrium. Genomics 36:1–16

Feingold E (1993) Markov processes for modeling and analyzing a new genetic mapping method. J Appl Prob 30:766–779

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994) Genome scanning by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. Nat Genet 8:380–386

Jorde LB (1995) Linkage disequilibrium as a gene-mapping tool. Am J Hum Genet 56:11–14

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kerem B-S, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080

Kruglyak L (1997) The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–24

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lehesjoki AE, Koskiniemi M, Norio R, Tirrito S, Sistonen P, Lander E, de la Chapelle A (1993) Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet 2:1229–1234

Terwilliger JD (1995) A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56: 777–787

Wang D, Sapolsky R, Spencer J, Rioux J, Kruglyak L, Hubbell E, Ghandour G, et al (1996) Toward a third generation genetic map of the human genome based on biallelic polymorphisms. Am J Hum Genet Suppl 59:A3