Ann. Hum. Genet. (1996), **60**, 261–263
Printed in Great Britain

Population stratifications can cause false positive linkage results if founders are untyped

D. CURTIS AND P. C. SHAM*
*Departments of *Psychological Medicine and Biostatistics, Institute of Psychiatry, De Crespigny Park, London SE5 8AF*

(Received 5.9.95. Accepted 10.11.95)

We wish to draw attention to the fact that population stratifications can produce a bias towards yielding false positive linkage results when founder genotypes are missing. It is now fairly well known that misspecification of allele frequencies can produce such a bias – if a common allele is misspecified as rare then the fact that it is frequently observed among affected individuals will be incorrectly taken as evidence of increased sharing of alleles identical-by-descent (Ott, 1992). What may be less widely recognized is that similar effects can occur even if the correct population allele frequencies are used. Independently of us, Miller *et al.* (1995) and Gu *et al.* (1995) have recently shown that population stratifications can produce false positive linkage results for Lange's (1985) identity-by-state affected sib-set analysis and for the affected-pedigree-member method (APM, Weeks & Lange, 1988). In fact, the problem applies to any method of linkage analysis when applied to datasets with untyped founders.

To see how the rate of false positives can increase, it may be helpful to consider an extreme example. If a population is stratified into two groups which never interbreed with each other, it may be that members of one group are all homozygous for one allele of a marker while members of the other group are all homozygous for another allele. The overall allele frequencies in the population will depend on the relative sizes of the two groups. If an affected sib pair analysis is carried out in which parental genotypes are missing, all sib pairs from these groups will share two alleles identical-by-state, leading to the conclusion that there is strong evidence for linkage between the disease and marker loci. This will be so even if the correct overall population frequencies for the alleles are used. To consider a slightly less extreme situation, suppose that there are four subgroups within a population such that within each group one allele of a four-allele marker has frequency of 0·7 while the other alleles have frequency 0·1, the common allele being different for each group. If the groups are equal sizes then the overall allele frequencies in the population will be equal. Using this population we simulated an unlinked Mendelian recessive disease in 1000 affected sib pairs using SLINK (Ott, 1989; Weeks *et al.* 1990) and analysed the results using MLINK (Lathrop *et al.* 1984). The maximum expected lod per sibship was 0·044 at a recombination fraction of 5%, whereas for an unlinked locus the maximum expected lod should be 0 at a recombination fraction of 50%. In this situation, with a sample of 100 sib pairs one would obtain on average a lod score of 4·4 at theta = 0·05, and the expected maximum lod score would be even higher.

Similar effects, albeit perhaps of a lesser magnitude, will occur for all methods of linkage analysis when founder genotypes are missing, providing that one concentrates more on phenotypic similarities than on differences. If a pedigree is drawn from some sub-population in which a certain allele has increased frequency then the probabilities for members of the pedigree to share that allele will be increased. If, as is generally the case, the pedigree is selected on the basis of containing multiple affected cases then this increased allele-sharing will be taken as evidence of linkage, since one will observe what is

apparently an increase in allele-sharing between affected subjects. Other pedigrees may be drawn from sub-populations in which different alleles have increased frequency, so that even if one estimates overall allele frequencies from the pedigree data set one may arrive at average values which are inappropriate for individual pedigrees, each of which may be biased with respect to different alleles.

Although we have begun by formulating the problem in terms of allele frequencies, it can be seen that the key feature which introduces bias is the underestimation of the probability for an untyped ancestor to be homozygous. This implies that there will be an exaggeration of the extent to which identity-by-state observations are taken to imply identity-by-descent. So another way of viewing the problem might be to say that bias can occur even with correct allele frequencies if there is deviation from Hardy–Weinberg equilibrium towards an increase in homozygosity. The first example given above can be seen to represent an extreme case where the entire population is homozygous. Any tendency for stratification to occur within a population is likely to increase homozygosity, and it is hard to conceive of mechanisms which would plausibly have a major effect in the opposite direction. Since modern western societies contain members having diverse origins who exhibit marked assortative mating along lines of race, religion, geography and class it would seem highly probable that many polymorphisms would indeed exhibit homozygosity above that expected from the overall population allele frequencies.

The extent to which these effects might cause problems in practice is difficult to judge. Differences in allele frequencies between sub-populations certainly occur, for example between Caucasians and African-Americans (Maliarik *et al.* 1995), and even a small bias can produce highly significant results in a large enough sample. The bias will be stronger the more ancestral genotypes are missing or ignored, and the more weight is given to affected subjects. If equal weight is given to unaffected subjects not sharing the same alleles as affecteds then the problem will be diminished, although selection of pedigrees on the basis of multiple affection may still introduce some bias. If founder genotypes are completely known then the analyses incorporating them will be unbiased.

Where ancestral genotypes are missing one should begin by recognizing that bias may occur. When characterizing a polymorphism it would make sense to estimate genotype frequencies as well as allele frequencies so that any deviation from Hardy–Weinberg equilibrium can be noted. Where such deviation does occur then logically one should specify the probabilities of individual genotypes for founders, rather than simply specifying allele frequencies, although currently available programs for performing linkage analysis do not allow this. One might attempt to separate pedigrees into homogeneous ethnic groups and to estimate allele frequencies in these groups separately. However, one could not go to the extent of estimating the allele frequencies for each pedigree individually, because this would produce a bias against the detection of linkage – if linkage were present then the frequency of the linked allele in each pedigree would be overestimated, unfairly reducing the evidence for linkage. Using more than one linked marker in analyses might also be helpful, since if one marker deviates from Hardy–Weinberg equilibrium then the other may not, and alarm bells will be sounded if only one of the markers produces evidence for linkage.

Overall, the effect we draw attention to is certainly capable of causing false positive linkage results in theory. The extent to which it might cause problems in practice is worthy of further investigation. For now, it seems reasonable to recognize the existence of the potential for problems and to exercise due caution when interpreting the results of analyses which do not incorporate founder genotypes.

REFERENCES

GU, C., MILLER, M. D. & REICH, T. (1995). The affected-pedigree-member method revisited under population stratification. *Psychiat. Genet.* **5**, S105–S106.

LANGE, K. (1986). A test statistic for the affected-sib-set method. *Ann. Hum. Genet.* **50**, 283–290.

LATHROP, G. M., LALOUEL, J. M., JULIER, C. & OTT, J. (1984). Strategies for multilocus linkage analysis in humans. *Proc. Natl Acad. Sci. USA* **81**, 3443–3446.

MALIARIK, M. J., KOST, J. A., HARRINGTON, D. W., POPOVICH, J. JUNIOR, MAJOR, M. L., RYBICKI, B. A. & IANNUZZI, M. C. (1995). Chromosome 6p microsatellite polymorphisms in African-Americans. *Hum. Hered.* **45**, 90–97.

MILLER, M. B., GU, C. & REICH, T. (1995). Sensitivity of Lange's (1986) identity-by-state method of linkage analysis to population stratification. *Psychiat. Genet.* **5**, S104.

OTT, J. (1989). Computer-simulation methods in human linkage analysis. *Proc. Natl. Acad. Sci. USA* **86**, 4175–4178.

OTT, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *Am. J. Hum. Genet.* **51**, 283–290.

WEEKS, D. E. & LANGE, K. (1988). The affected-pedigree-member method of linkage analysis. *Am. J. Hum. Genet.* **42**, 315–326.

WEEKS, D. E., OTT, J. & LATHROP, G. M. (1990). SLINK: a general simulation program for linkage analysis. *Am. J. Hum. Genet.* **47**, A204.