Identifying Influential Individuals in Linkage Analysis: Application to a Quantitative Trait Locus Detected in the COGA Data

Braxton D. Mitchell, Soumitra Ghosh, Richard M. Watanabe, Susan H. Slifer, Wen-Chi Hsueh, and Gunther Birznieks

Department of Genetics (B.D.M., S.H.S.), Southwest Foundation for Biomedical Research, San Antonio, Texas; National Human Genome Research Institute, (S.G., G.B.), Bethesda, Maryland; Department of Biostatistics (R.M.W.), University of Michigan School of Public Health, Ann Arbor, Michigan; Department of Epidemiology (W.-C.H.), Johns Hopkins University School of Hygiene and Public Health, Baltimore, Maryland

Once linkage is detected to a quantitative trait locus (QTL), the next step towards localizing the gene involved may be to identify those families, or individuals, in whom the putative mutations are segregating. In this paper, we describe a jackknife procedure for identifying individuals (and families) who contribute disproportionately to the linkage. Following initial detection of linkage to a QTL, the strategy involves sequentially removing each individual (or each family) from the analysis and recomputing the lod score associated with the linked region using data from all remaining subjects (or families). This procedure can be used to determine if particular observations have substantial impact on evidence for linkage. Identification of such observations may provide insights for further efforts to localize the QTL. ©1999 Wiley-Liss, Inc.

Key words: alcoholism, influence plot, outliers

INTRODUCTION

Detection of linkage between a genetic marker and a quantitative trait locus (QTL) represents the initial step in the process of mapping genes for complex traits. Once statistical evidence for linkage is observed, it may be desirable to identify families or subsets of individuals in whom the putative genes may be segregating. In this paper, we

Address reprint requests to Dr. B. Mitchell, Department of Genetics, Southwest Foundation for Biomedical Research, P.O. Box 760549, San Antonio, TX 78245-0549.

describe a strategy for identifying informative individuals who contribute disproportionately to the linkage. To illustrate the strategy, we used data from the Collaborative Study on the Genetics of Alcoholism (COGA) to identify individuals contributing to a linkage localized to the long arm of chromosome 6 for a QTL that influences amplitude of P300 event-related potentials (ERPs).

METHODS

We initially conducted a genome-wide scan of the cz-P300 trait in the COGA data set to identify regions of the genome containing alcoholism susceptibility genes. This initial analysis, based on 604 individuals from 105 families, revealed evidence for linkage of a QTL influencing variation in cz-P300 to a region on chromosome 6q, approximately 198 cM from pter. This linkage result has been previously described by Begleiter and colleagues [1998].

We then conducted subsequent multipoint linkage analyses that were restricted to this locus. Briefly, we used a jackknife procedure in which we sequentially removed each individual from the data set, and then recalculated the lod score for the entire sample less that one individual. Each linkage analysis was based on removal of only a single individual; thus, a total of 604 separate analyses were performed. Lod scores were saved from each analysis, and influential individuals were considered to be those for whom the lod scores were substantially altered when they were removed.

In addition to sequentially removing each individual from the analysis, we also performed an analogous procedure in which we sequentially removed each family from the analysis. Thus, we removed one at a time each of the 105 families and performed the linkage analysis on the remaining 104 families. This allowed us to determine which (if any) families contributed disproportionately to the linkage.

All linkage analyses were performed using a variance components framework, in which we used maximum likelihood methods to partition the phenotypic variation in cz-P300 into components attributable to covariates (age and sex), polygenetic effects, and locus-specific genetic effects [Almasy and Blangero, 1998]. The model does not require specific information about allele frequencies or penetrances at the putative trait locus, but we did assume for this analysis that the genetic effects were additive and that no interaction effects were present. Analyses were performed using the SOLAR software program [Almasy and Blangero, 1998]. The identity-by-descent (IBD) matrix was used to model the locus-specific genetic effects. The matrix was generated using the SOLAR software package, which uses information from all neighboring markers to impute the IBD probabilities. For the initial genome-wide scan, the multipoint IBD matrix was generated for each position (in 1-cM increments) along the genome. To reduce computational burden, we did not recompute the multipoint IBD probabilities after removing each individual, but rather used the IBD calculations computed from the full data set for all analyses. In each analysis, we used the pairwise IBD probabilities corresponding to the position at which peak evidence for linkage was observed (i.e., position 198 cM). The hypothesis of linkage was tested using the likelihood ratio test, which compares the likelihood of the data under the polygenic model (variation in cz-P300 attributable to age, sex, and family structure) with that of the data under the linkage model (variation in cz-P300 attributable to these same factors, as well as to allele-sharing at position 198).

RESULTS

The trait cz-P300 was measured in 604 individuals. Values of cz-P300 were approximately normally distributed, with mean 13.22 μV (median 13.04 μV) and standard deviation of 8.66.

In the full data set, multipoint linkage analysis of the trait cz-P300 revealed that peak evidence for linkage across the genome occurred at a region on chromosome 6q, approximately 198 cM from pter. The multipoint lod score at this position was 3.54. The linkage analysis was then repeated 604 times, following sequential removal of each individual from the analysis set. The corresponding lod scores associated with position 198 cM ranged from 3.95 to 2.73. Thus, depending on which individual was removed from the analysis, the lod score decreased by as much as 0.81 units or increased by as much as 0.41 units. The results of these analyses are summarized in Figure 1, which shows the value of each lod score according to the phenotypic value of the individual removed, expressed in terms of the number of standard deviations (SD) from the mean. In two of the analyses (based on removal of individuals 456 and 457), the lod scores decreased to less than 3.0. If both individuals were removed from the same analysis, the resulting lod score was 2.41.

The results of the linkage analysis for which entire families were sequentially removed are presented in Figure 2. When family 43 was removed, the lod score fell to 2.25. This family included 10 individuals, among whom were individuals 456 and 457. Removal of any other family from the analysis was associated with lod scores ranging from 2.92 to 4.00, depending on the family removed.

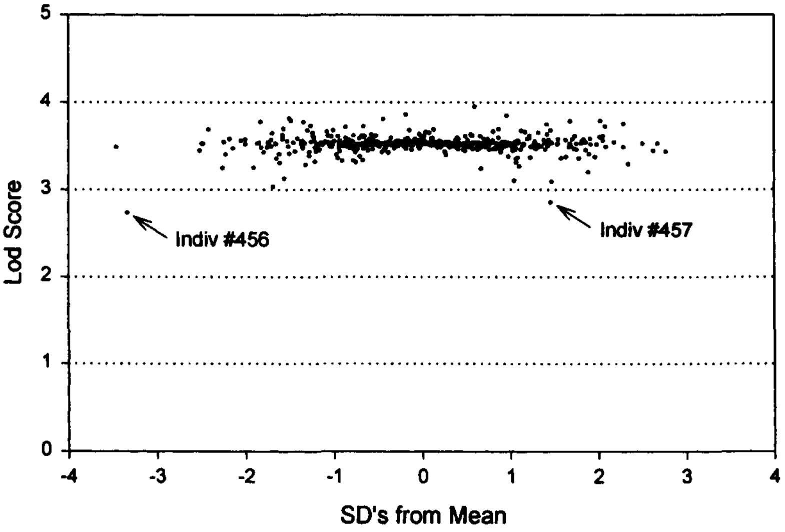

Fig. 1. Influence plot for linkage analysis of cz-P300 – effect of removing individuals from the analysis.

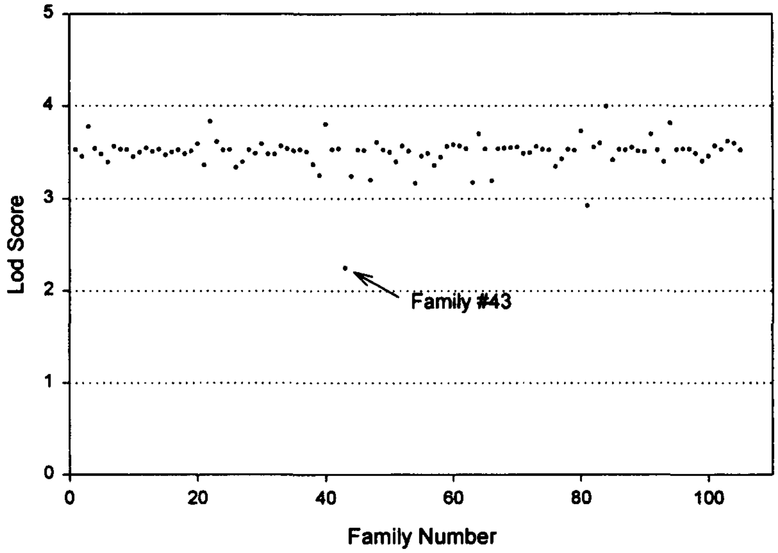

Fig. 2. Influence plot for linkage analysis of cz-P300 – effect of removing families from the analysis.

DISCUSSION

In the analysis of the full data set, we detected strong evidence for linkage between a QTL influencing variation in cz-P300 and markers on chromosome 6q. Through our analysis, we identified two siblings who had a substantial effect on the lod score as evidenced by the fact that removal of both individuals from the analysis caused a decrease in the lod score from 3.54 to 2.41. An immediate question that arises is whether inclusion of these influential individuals leads to a distortion in the evidence for linkage. On the one hand, it is possible that a QTL influencing variation in cz-P300 truly exists in this region, and that inclusion of the influential individuals enhanced our ability to detect the linkage. If true, then one might hypothesize that one or both of these individuals carries a mutation in this QTL. On the other hand, it is possible that no QTL influencing cz-P300 variability exists in this region, and that inclusion of these two influential individuals resulted in an overestimate of the evidence for linkage.

There are several reasons why one might want to identify influential observations in a linkage data set. The most obvious reason is to check for possible errors in the data. The phenotypic data for the outlying individuals should therefore be verified, perhaps by re-measuring the individuals' phenotypes. In the COGA data, one of the outlying individuals (456) had a relatively extreme value of cz-P300. In fact, this individual was one of only two whose phenotypic value deviated by more than three standard deviations from the mean (see Figure 1). If this phenotypic value is determined to be in error, then obviously the observation should be discarded. A second reason for determining whether the data include outlying observations is that their presence casts doubt on the nominal p-values associated with the computed lod score. In the presence of outliers, it might be advisable to compute p-values empirically (e.g., by simulation). The cost of doing such a simulation (say, with 10,000 replications), however, is not trivial in terms of computing time. Therefore, the identification of one or more influential observations in the data set may be one motivation for making the extra effort to determine p-values empirically.

Examination of outlier influence plots can provide other useful insights. First, if no outliers are observed, one may feel more confident about the observed linkages. If significant outliers are observed, and no obvious source of phnotypic error can be found, then additional follow-up is warranted. If removal of the outliers results in a decrease in the lod score, as is the case for individuals 456 and 457 in the COGA data, then one's view of the evidence for linkage may change accordingly. Some insights may be obtained by examining the IBD estimates between each of the outlying individuals and their relatives. For example, how informative is the marker information for the influential observations? If the precision of these IBD estimates is relatively low, then support for a true linkage might be diminished. One might examine also the genotypic and phenotypic correlations between the influential individuals. Is the individual influential because he or she is both phenotypically and genotypically discordant with his or her relatives, or is the individual influential because he or she is phenotypically and genotypically concordant with his or her relatives? The latter scenario may be more encouraging and potentially more informative if one is contemplating further efforts to localize the QTL and identify putative mutations because relatives who are concordant at the putative trait locus may carry the same mutation. In our example above, individuals 456 and 457 are phenotypically discordant (a 41.5μV difference), and their predicted IBD probability at position 198 is 0.02. Thus, if a mutation is segregating in this family, only one of these sibs is likely to carry it.

One could also check for individuals whose removal significantly increases evidence for linkage. As shown in Figure 2, removal of one family from the COGA analysis increases the lod score to 4.0. One potentially useful analysis might be to recalculate the multipoint linkage analysis across chromosome 6q with this family missing to determine whether the region of linkage can be more finely localized.

How lod score influence plots should be used to identify outlying individuals is an important issue. We regard our outlier analysis as a form of exploratory data analysis because the goal of the analysis is to gain insights into which parts of the data are providing evidence for linkage. The analysis is post-hoc, and outliers are identified with no prior biological hypotheses. A strength of this type of analysis, however, is that it can be easily automated and can be performed relatively rapidly. Despite the exploratory nature of this analysis, one could consider quantitative criteria for evaluating the lod score influence plots to assist in the identification of outlying individuals. For example, one simple-minded approach for defining outliers by statistical means would be to compute the mean of the lod scores and then to define an outlier as any value that deviates by more than some arbitrary amount from this mean. However, it is not evident what threshold to use because the lod scores are not independent and the expected variance of this distribution will be small. In our cz-P300 example, only three lod scores deviated by more than five SD from the mean of all lod scores (3.520 ± 0.469, μ ± 5 SD) – individual 922, who deviated by more than five SD from the mean, and individuals 457 and 456, who deviated by more than seven and eight SD from the mean, respectively. Had we used a cutpoint of four SD from the mean (± 0.375), we would have identified as outliers four additional individuals (three additional lod scores below the mean and one above the mean).

In summary, the procedure we have described may be useful for providing insights into linkages that are obtained from an initial analysis. If the data set includes individuals who contribute disproportionate evidence for (or against) linkage, their phenotypic values should be verified. P-values can be generated empirically if removal of such individuals significantly reduces evidence for linkage. In the analysis of cz-P300, two relatively influential individuals were identified, causing a decrease in the lod score from 3.54 to 2.41. The two influential individuals were siblings, and because they were phenotypically and genotypically discordant, it is unlikely that both carry the same mutation. One might hypothesize the presence of a rare mutation whose effect on phenotypic variation is large. One or both of these individuals might be candidates to include for more detailed analysis for sequencing or cloning studies.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants RO1-AR43351 and PO1-HL45522. R.M.W. was supported by an ADA Career Development Award. The authors thank Dr. John Blangero for his helpful suggestions.

REFERENCES

Almasy L, Blangero J (1998): Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198-1211.

Begleiter H, Porjesz B, Reich T, Edenberg HJ, Goate A, Blangero J, Almasy L, Foroud T, Van Eerdewegh P, Polich J, Rohrbaugh J, Kuperman S, Bauer LO, O'Connor SJ, Chorlian DB, Li TK, Conneally PM, Hesselbrock V, Rice JP, Schuckit MA, Cloninger R, Nurnberger JI Jr, Crowe R, Bloom FE (1998): Quantitative trait loci analysis of human event-related brain potentials: P3 voltage. Electroencephalogr Clin Neurophysiol 108:244-5.