Application of an Ordered Subset Analysis Approach to the Genetics of Alcoholism

Richard M. Watanabe, Soumitra Ghosh, Gunther Birznieks, William L. Duren, and Braxton D. Mitchell

Department of Biostatistics (R.M.W., W.L.D.), University of Michigan School of Public Health, Ann Arbor, Michigan; Genetics and Molecular Biology Branch (S.G., G.B.), National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland; Department of Genetics (B.D.M.), Southwest Foundation for Biomedical Research, San Antonio, Texas

For complex diseases, underlying etiologic heterogeneity may reduce power to detect linkage. Thus, methods to identify more homogeneous subgroups within a given sample in a linkage study may improve detection of putative susceptibility loci. In this study we describe an ordered subsetting approach that utilizes disease-related quantitative trait data to complement traditional linkage analysis. This approach uses family-based lod scores derived from the initial genome screen and a family-based descriptor of the trait of interest. The goal of the approach is to identify more homogeneous subgroups of the data by ranking families based on their quantitative trait data. Permutation testing is used to assess statistical significance. This approach can be adapted to a variety of linkage methods and may provide a means to dissect some of the underlying heterogeneity in complex disease genetics. © 1999 Wiley-Liss, Inc.

Key words: linkage analysis, permutation tests, quantitative traits, subsetting

INTRODUCTION

It is generally assumed that in complex diseases a number of genes of varying effect contribute to disease penetrance. Furthermore, the effect of any given gene may be influenced by environmental factors and/or interact with other loci along the genome. Thus, etiologic heterogeneity likely exists for complex diseases. This underlying heterogeneity may confound the identification of susceptibility loci using linkage

Address reprint requests to Dr. R. M. Watanabe, University of Michigan School of Public Health, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109-2029.

analysis approaches that rely upon disease status alone. In the presence of heterogeneity, maximum lod score curves can consist of broad peaks covering wide regions or multiple overlapping peaks in a narrow region along the chromosome. Thus, identification of unique peaks representing evidence for a potential susceptibility locus may be problematic. This scenario suggests that reliance upon affection status alone for genomic scans may be insufficient to localize susceptibility loci for complex disorders.

A systematic approach to identifying and characterizing subsets of these data that contribute significantly to linkage may prove useful in circumventing the problem of etiologic heterogeneity. To this end, we examine the efficacy of an ordered subsetting approach to assist in the identification of susceptibility loci for alcoholism using data collected for the Collaborative Study on the Genetics of Alcoholism (COGA) [Begleiter et al., 1995; Reich et al., 1998]. In this approach we attempt to increase our evidence for linkage by subsetting the data using quantitative trait data collected on affected subjects for the COGA study. Subsetting on disease-related trait values likely identifies a more phenotypically homogeneous subset of the sample and therefore likely approaches genetic homogeneity.

METHODS

Genome Scan

We performed a genome scan and subsequent ordered subset analysis on pedigrees from the COGA study [Begleiter et al., 1995; Reich et al., 1998]. This sample consisted of 450 affected and 533 unaffected subjects in 105 pedigrees. Alcohol dependency was defined according to the COGA criteria [Begleiter et al., 1995]. The initial genome scan (autosomes only) was performed using GENEHUNTER PLUS (GH+) [Kong and Cox, 1997].

Ordered Subset Analysis

Three sets of traits were examined for this study; age of onset for alcohol dependency, platelet monoamine oxidase (MAO) activity, and the P300 component of the event-related potentials (ERP). MAO activity has been reported to be associated with alcohol dependency; however, this association appears to be confounded by both gender and current smoking [Begleiter et al., 1995]. We therefore computed mean values after adjusting MAO activity data for gender and current smoking. For the P300 component, we examined all eight leads, with the understanding that some degree of correlation likely exists among these neurological tracings.

To perform the ordered subset analysis, we modified the GH+ software to output family-by-family the GENEHUNTER statistic ($\bar{Z}$) for each evaluation of the excess allele sharing parameter δ, as described by Kong and Cox [1997]. This was done for all putative disease locus positions along the genome.

Once the genome scan was completed, we performed our ordered subsetting analysis as follows. We computed a mean trait value for each family using data from the affected members only. The *m* families in our sample were then ranked by the trait mean in either ascending or descending order. The GH+ family-based statistics were then sequentially added in increasing rank order; denoted as Low→High. With each addition of a family, maximization was performed for the given subset and the maximum statistic value and estimated map position noted. Thus for each subset, we recalculated the GENEHUNTER Z_{lr} statistic as:

$$Z_{lr}(x) = \sqrt{2 \sum_{i=1}^{k} \ln[1 + \hat{\delta}\ \overline{Z}_i(x)]}$$

Here $\bar{Z}_i(x)$ is the family-specific GENEHUNTER statistic at disease position x for family i of the k families in the current subset, and $\hat{\delta}$ is the maximum likelihood estimate of the excess sharing parameter. The summation and maximization was repeated until all the families in the data set had been added and the parameter estimates and statistic value for the subset yielding the maximum evidence for linkage was then reported. We repeated the analysis adding statistics in decreasing rank order (High-Low) to ensure we did not miss subsets that might reside in either extreme of the trait distribution.

Statistical significance was assessed using a permutation approach [Good, 1994]. For our permutation test, the rankings for the family-based trait means are randomized and the permuted data were analyzed using the ordered subset approach. Five thousand replicates were generated to create the lod score distribution to allow determination of an empirical p-value.

RESULTS

Results from the initial genome scan are shown in Table I. We observed modest evidence for linkage on several chromosomes, with our strongest evidence found on chromosome 1. Several chromosomes appeared to show evidence for more than one locus. For example, the maximum lod score for chromosome 6 was observed at 23 cM; however, there was a secondary peak observed at 62.5 cM with a lod score of 1.17.

Our ordered subset results are shown in Table II. Subsetting on the trait data provided evidence for linkage on several chromosomes. It is of interest to note that of the initial six chromosomes for which we observed evidence for linkage from our genome scan, only chromosome 16 showed significant evidence for linkage in our subsetting results. For chromosome 16 the ordered subset approach identified 53 families with the highest mean age of onset showing the strongest evidence for linkage (lod = 3.17, 87 cM, p = 0.0144). It is of interest to note that the location of this peak is in a slightly different location on the chromosome compared to the result from the initial genome scan (87 vs. 79 cM).

The ordered subset approach did identify families showing strong evidence for linkage on those chromosomes showing no initial evidence for linkage in our genome scan (cf. Tables I and II). In all, eight chromosomes that previously showed no evidence for linkage showed significant evidence for linkage in some subset of the sample. In the ordered subset results, our strongest evidence for linkage was observed on chromosome 18 where we observed a lod of 3.89 in the 66 families with the lowest mean value for the t8 lead on the P300 ERP.

TABLE I. Maximum Lod Scores ≥ from GH+ Analysis on 105 Pedigrees

Chromosome	Lod scores	Position (cM)
1	2.80	176.5
6	1.95	23.0
7	1.00	100.5
8	1.54	20.5
11	1.49	52.0
16	1.02	79.0

TABLE II. Maximum Lod Scores from Ordered Subset Analysis

Chromosome	Stratification trait	Trait rank ordering	# of families	Lod	Position (cM)	Empirical p-value
2	O1 lead	High→Low	58	3.71	93.5	0.0004
	fp1 lead	High→Low	26	2.77	180.0	0.0190
	O2 lead	High→Low	50	2.70	92.0	0.0240
9	O1 lead	Low→High	76	2.23	145.5	0.0224
	Age-of-onset	Low→High	17	2.07	141.0	0.0368
10	fp2 lead	High→Low	68	3.08	53.0	0.0128
	fp1 lead	High→Low	73	2.78	54.0	0.0314
12	O1 lead	Low→High	42	2.10	177.0	0.0324
13	MAO	Low→High	40	2.52	89.0	0.0064
15	cz lead	Low→High	27	2.42	120.0	0.0308
16	Age-of-onset	High→Low	53	3.17	87.0	0.0144
18	t8 lead	Low→High	66	3.89	91.0	0.0006
	pz lead	Low→High	74	3.33	91.0	0.0078
22	t7 lead	Low→High	23	1.55	18.0	0.0038

DISCUSSION

In the search for susceptibility genes for complex diseases, etiologic heterogeneity may be an important component that actually reduces power to detect linkage. Given the hypothesized multigenic basis for many complex disorders, in a large-sample study it is likely that subsets of individuals, while contributing evidence for linkage in one part of the genome, may reduce evidence for linkage in another. A systematic approach to identifying homogeneous subgroups would provide a mechanism to assist in the identification of linkage peaks of importance.

We propose an ordered subset approach that utilizes quantitative trait data to identify subgroups from the overall sample. Because these subgroups cluster in the same region of the trait distribution, it is likely that they represent a more homogeneous group; although, clearly one cannot be assured that the increased homogeneity is genetic in nature. We must emphasize that although our ordered subset approach uses disease-related quantitative trait data, it is not an alternative to quantitative trait locus linkage analysis. In the ordered subset approach our basic linkage statistic is still based on disease affection status.

Our approach is family-based and uses the mean trait value derived from the affected individuals of the pedigree. We chose the mean value as our family-based descriptor variable for the purpose of simplicity, but clearly other descriptive measures, such as the trait median or variance, can be used. Similarly, for these analyses we chose to use the allele sharing approach of Kong and Cox [1997] for our linkage statistic, but the ordered subset approach can be easily adapted to other linkage methods. For example, we have applied this approach to affected sibling pairs from the Finland-United States Investigation of Non-insulin-dependent diabetes mellitus (FUSION) study [Valle et al., 1998] in which a linkage approach based on Risch's recurrence risk ratio was used [Hauser et al., 1996; Risch, 1990].

It is interesting to note that for some chromosomes (2, 9, 10, and 18) our subsetting approach showed evidence for linkage for more than one locus or more than one trait. The latter observation can be partially attributed to possible correlation among the traits. In fact, for chromosomes 9, 10, and 18 this may be the case as both subsets identified maximize at the same location on the chromosome and there is significant overlap in families in the subsets identified. However, for chromosome 2 it is noteworthy that while evidence for linkage maximizes at the same chromosomal location (~93 cM) when subsetting on the O1 and O2 leads, subsetting on the fp1 lead maximizes at a different location on the chromosome (180 cM). The fact that the O1 and O2 strata maximize at the same location is likely due to correlation between these two traits (49 common families between these subsets). However, when we checked for overlap among families in the fp1 subset compared to the O1 and O2 subsets, there were only 17 and 15 common families, respectively. Thus, some of the difference cannot be explained by correlation in the data.

One of the attractive properties of this subsetting approach is that use of disease-related trait information can lend itself to easy biologic interpretation. For example, low MAO activity is thought to be associated with early-onset forms of alcohol dependency [Begleiter et al., 1995]. With our subsetting approach we identified 40 families with low mean MAO activity, after adjustment for gender and smoking, who provide evidence for linkage on chromosome 13 (lod = 2.52, p = 0.0064) that did not appear in the original genome scan with the entire sample. Thus, these families could represent a subset with a slightly differing form of early-onset alcohol dependency.

A certain level of caution is required in the implementation of the approach and interpretation of the results. First, the *a priori* selection of traits should be carefully considered with respect to the relative informativity of the trait in relation to disease status. For example, in affected subjects trait values can be negatively impacted by treatment regimens for disease such that the trait, although related to the disease, may not be informative. In such cases, it may be preferable to stratify on unaffected offspring of affected subjects, as the offspring are "at risk" for disease and their trait data may be more informative. Alternatively, it may be more informative to rank on clusters of traits rather than individual traits. Second, the significance of the lod scores obtained from our subsetting approach requires careful interpretation. Our current permutation approach addresses the specific question regarding the probability of identifying a contiguous subset of families with the observed maximum lod score given a specific set of family-based lod scores and trait information. We have yet to incorporate the effect of multiple comparisons due to subsetting on multiple traits, or overall linkage information at a given locus from our original genome scan. In fact, when we make a simple Bonferroni correction for multiple comparisons, only three of the results reported in Table II remain statistically significant. Also, it is clear that alternative hypotheses can be addressed using this subsetting approach. For example, one may be interested in assessing statistical significance for the maximum lod score observed in the subset conditional on the maximum lod score observed in the initial genome scan. Also, the number of permutations to be performed should be selected with care. Third, although we have identified a contiguous subset of families with evidence for linkage, this obviously does not represent the absolute "best subset" of families. The current algorithm always starts with the family with the lowest (or highest) trait ranking regardless of whether that particular family provides evidence for linkage. Therefore, the current approach may report families with rankings 1 to 50 having the strongest evidence for linkage, when in reality only families 20 to 50 contribute the majority of evidence for linkage. It may be of use to include an optimization to identify the "best" contiguous subset of families contributing evidence for linkage from all possible contiguous subsets.

Finally, when a subset is identified, it would be of interest to repeat linkage analysis removing that subset from the data. Such an approach would not only provide an interesting assessment of the contribution of those families to any positive linkage result, but also may reveal other loci where the families in the subset were uniformly contributing negative evidence for linkage. An approach to assess statistical significance for the observed change in lod score under such a scenario is yet to be addressed.

In summary, we applied an ordered subset analysis approach to data from the COGA study after an initial genome scan using GH+. The ordered subset approach identified subsets of the original data that appeared to provide significant evidence for linkage on several chromosomes. Many of these results were obtained on chromosomes where the original genome scan did not yield any evidence for linkage. The use of quantitative trait information coupled with disease status in this approach may provide a means to diminish the impact of the underlying etiologic heterogeneity associated with complex diseases and allow for easier identification of potential disease susceptibility loci.

ACKNOWLEDGEMENTS

R.M.W. is supported by a Career Development Award from the American Diabetes Association. W.L.D. is supported by NIH grant HG-00376. B.D.M. is supported in part by NIH grants R01-AR43351 and P01-HL45522.

REFERENCES

Begleiter H, Reich T, Hesselbrock V, Porjesz B, Li T-K, Schuckit MA, Edenberg HJ, Rice JP (1995): The Collaborative Study on The Genetics of Alcoholism: the genetics of alcoholism. Alcohol Health Res World 19:228-236.

Good P (1994): "Permutation Tests." New York: Springer-Verlag, pp. 1-226.

Hauser E, Boehnke M, Guo S-W, Risch N (1996): Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. Genet Epidemiol 13:117-137.

Kong A, Cox N (1997): Allele-sharing models: lod scores and accurate linkage tests. Am J Hum Genet 61:1179-1188.

Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, Bucholz K, Porjesz B, Li TK, Conneally PM, Nurnberger JI Jr, Tischfield JA, Crowe RR, Cloninger CR, Wu W, Shears S, Carr K, Crose C, Willig C, Begleiter H (1998): Genome-wide search for genes affecting the risk for alcohol dependence. Am J Med Genet 81:207-215.

Risch N (1990): Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222-228.

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefield CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M (1998): Mapping genes for NIDDM. Design of the Finland-United States Investigation of NIDDM Genetics (FUSION) Study. Diabetes Care 21:949-958.