A Multiple Locus Analysis of the COGA Dataset

(Running Title: A Multiple locus Analysis of the COGA Dataset)

Shili Lin*, Mark E. Irwin, and Fred A. Wright

Department of Statistics (S.L., M.E.I.) and Division of Human Cancer Genetics (F.A.W.), The Ohio State University, Columbus, OH

Parametric and nonparametric statistical methods have been applied to the alcohol dependence dataset collected in the Collaborative Study on the Genetics of Alcoholism (COGA). Our nonparametric linkage analyses (NPL) were based on the S_{all} statistic of GENEHUNTER [Kruglyak et al., 1996] and the improved NPL statistic of GENEHUNTER-Plus [Kong and Cox, 1997]. Based on likely regions for alcohol susceptibility genes identified from our nonparametric analyses, we reanalyzed the data using several two-locus models. We used the TMLINK program [Lathrop and Ott, 1990] in the LINKAGE package for these parametric analyses.

Key words: alcohol dependence, linkage analysis, GENEHUNTER-Plus, two-trait-locus models

INTRODUCTION

Recent linkage studies of genome-wide scan for genes affecting the risk for alcoholism provided evidence for linkage between susceptibility loci for alcohol dependence (AD) and regions on chromosomes 1, 2 and 7 [Reich et al., 1998]. These results were obtained based on nonparametric sib-pair methods, and the affected status phenotypes used were derived according to the COGA diagnostic criterion. Since breaking families into sib-pair data may lose information for linkage, we explored alternative approaches of linkage methods that make fuller usage of information provided by families as a whole. In addition to using phenotypes under the COGA criterion, we also use affected status data under the World Health Organization diagnosis (ICD10). Previous segregation analysis [Yuan et al., 1996] rejected the hypothesis of a single-locus major gene model with

* Address reprint requests to Dr. Shili Lin, shili@stat.ohio-state.edu, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210, USA.

Mendelian segregation for AD. This coupled with evidence of linkage on more than one chromosomal region led us to the exploration of two-locus models.

METHODS

Nonparametric Analysis

To determine a set of likely regions for alcohol susceptibility loci, nonparametric linkage analysis (NPL) based on the Kong and Cox lod score derived from the S_{all} statistic of GENEHUNTER-Plus (version 1.1) [Kruglyak et al., 1996; Kong and Cox, 1997] was used. This statistic compares the allele sharing by descent for all affected members of the pedigree. For all pedigrees with $2N - F \leq 16$, all members of the pedigree were used where N is the number of non-founders and F is the number of founders in the pedigree. In any pedigree with $2N - F > 16$, members of the pedigree were dropped based on the program defaults, which tries to select the least informative nonfounder members of the pedigree, until the cutoff of 16 was met. The analyses were done under both diagnostic criteria and all pedigrees were used to examine chromosomes 1 through 22. In addition, chromosome 4 was examined to see if there was excess sharing among the unaffected pedigree members to follow up on a previous finding [Reich et al., 1998].

Parametric Analysis Using Two-locus Models

Two-trait-locus, one-/two-marker-locus linkage analyses were carried out using the TMLINK program [Lathrop and Ott, 1990] of the LINKAGE package [Lathrop et al., 1984]. We assumed that the two postulated disease loci were diallelic and on different chromosomes. Locus 1 had two alleles, the disease allele a and the normal allele A, with frequencies p and $1 - p$, respectively. Locus 2 had two alleles, the disease allele b and the normal allele B, with frequencies q and $1 - q$, respectively. We considered two types of models: double recessive, and dominant/recessive. For the double recessive model, we assumed complete penetrance with $p = q = 0.1$. Therefore, those who were doubly homozygous for a and b must be affected and must be the only ones affected. For the dominant/recessive model, we assumed complete penetrance (at the recessive locus) for those who are homozygous bb. We then estimated allele frequencies p, q, and penetrance f for those who carried at least one copy of the disease gene a at the dominant locus. To estimate these parameters, we assumed that the overall population prevalence for AD was 0.1, and the relative risk of a child given that a parent was affected is 0.3. These prevalence and risk values implied that $p = 0.045$, $q = 0.228$, and $f = 0.574$. The dominant/recessive model was discussed in Schork et al. [1993]. See the Discussion Section for an explanation on the choices of parameters and additional models.

For two-trait-locus, one-marker-locus analysis, one of the trait loci was assumed to be linked to a marker locus at recombination fraction θ, while the location of the other trait locus was unspecified. We computed the lod score

$$\mathrm{LOD}(\theta) = \log_{10} \mathrm{L}(\theta,0.5)/\mathrm{L}(0.5,0.5); \quad 0 \leq \theta \leq 0.5.$$

Under the null hypothesis that $\theta = 0.5$, 2 log(10) times the maximum lod score is asymptotically distributed as a mixture of a point mass at zero and a χ_1^2 with mixing

parameters 1/2 and 1/2. For two-trait-locus, two-marker-locus analysis, each trait locus was assumed to be linked to a marker, each on a different chromosome. We computed the lod score

$$\mathrm{LOD}(\theta_1, \theta_2) = \log_{10} \mathrm{L}(\theta_1, \theta_2)/\mathrm{L}(0.5,0.5),$$

where θ_i is the recombination fraction between a pair of trait and marker loci. Under the null hypothesis that $\theta_1 = \theta_2 = 0.5$, 2 log(10) times the maximum lod score (maximizing over both θ_1 and θ_2) is asymptotically distributed as a mixture of a point mass at zero, χ_1^2, and χ_2^2 with mixing parameters 1/4, 1/2, and 1/4 respectively.

RESULTS

Nonparametric Analysis

No positions stand out strongly in the NPL analyses. Under the COGA criterion, only chromosome 1 and 6 had Kong and Cox lod scores of over 2. For the ICD10 criterion, no chromosomes achieved a lod score of over 2. The two chromosomes with the most significant results under this criterion were 8 and 10. Summaries of the results for these chromosomes are shown in table I. It is interesting that for the four loci mentioned, the lod scores patterns are quite different under the two diagnostic criteria. Our multipoint findings for chromosome 4 appear to agree with previous findings. For the unaffected members of the pedigrees, a lod score of 2.08 was found at marker D4S2393, which is in the same general region previously suggested for a potential protective locus [Reich et al., 1998]. The results for chromosomes 20 and 22 suggest that there is little evidence for trait loci on these chromosomes as the Kong and Cox lod scores are 0 across the intervals of markers observed. The remaining chromosomes give less clear results, with lod scores less than 2 in the regions studied. In particular, chromosome 7 did not have as significant results as seen in the analyses in Reich et al. [1998]. However the maximum lod score of 1.05 under the COGA criterion occurred in the same region as found by that group.

Table I. Results from GENEHUNTER-Plus for selected locations under both diagnostic criteria

			COGA		ICD10	
Chromosome	Position	Nearest Marker	Lod Score	p-value	Lod Score	p-value
1	176.3	D1S534	2.539	0.00031	0.820	0.026
6	23.0	D6S1006	2.390	0.00045	0.090	0.260
8	130.9	D8S594	0.051	0.314	1.717	0.0025
10	46.7	D10S1426	0.941	0.016	1.890	0.0016

Two-locus Analysis

We performed two-trait-locus, one-marker-locus analysis first to screen through sets of marker loci on chromosomes 1, 2, 7 and 8 that were identified as possible linked markers with AD susceptibility loci. Markers on chromosomes 1 and 8 were chosen based on preliminary multipoint NPL analyses and markers on chromosomes 2 and 7 were based on previous findings [Reich et al., 1998]. Both the COGA and ICD10 criteria were examined. For the dominant/recessive model, we explored the situation in which the dominant locus was linked to a marker and the situation in which the recessive locus was

linked to a marker. Our results revealed that, for all the three models considered, D1S534 and D8S1145 showed the strongest evidence for linkage among all the markers considered. Apart from the case where the dominant locus was assumed to be linked to marker D1S534, data under the ICD10 criterion always provided stronger evidence for linkage than data under the COGA criterion. Table II summarizes the results for markers D1S534 and D8S1145 under the ICD10 criterion. The maximum lod score, its associated recombination fraction, and the p-value are provided.

Table II. Results from two-trait-locus, one-marker-locus analyses. For the dominant/recessive model, we assumed that either the dominant locus was linked to the marker under consideration or the recessive locus was linked to the marker.

Model	Marker	θ	Lod Score	p-value
double recessive	D1S534	0.20	2.02	0.0011
	D8S1145	0.20	1.94	0.0056
dominant/recessive	D1S534	0.10	0.53	0.0590
(dominant locus linked)	D8S1145	0.05	1.06	0.0140
dominant/recessive	D1S534	0.10	1.96	0.0013
(recessive locus linked)	D8S1145	0.10	1.81	0.0019

The two-trait-locus, two-marker-locus linkage analysis was performed assuming that one of the trait locus was linked to D1S534 and the other linked to D8S1145 using data under the ICD10 criterion. For the double recessive model, the maximum lod score of 3.85 was achieved at $\theta_1 = \theta_2 = 0.2$. The lod score corresponds to a p-value of 4.8×10^{-5}. Figure 1 shows the entire lod score surface. For the dominant/recessive model with the dominant locus linked to D1S534 and the recessive locus linked to D8S1145, the maximum lod score of 1.95 was achieved at $\theta_1 = 0.2$ (between the dominant locus and D1S534) and $\theta_2 = 0.1$. This yields a p-value of 0.0042. For the dominant/recessive model with the dominant locus now linked to D8S1145 while the recessive locus linked to D1S534, the maximum lod score of 2.53 was achieved at $\theta_1 = \theta_2 = 0.1$. This yields a p-value of 0.0011, and the entire lod score surface is shown in Figure 2.

Figure 1. Lod score surface for two-trait-locus, two-marker-locus analysis for the double recessive model. Markers D1S534 and D8S1145 are assumed to be linked to the two trait loci.

Figure 2. Lod score surface for two-trait-locus, two-marker-locus analysis for the dominant/recessive model. Marker D1S534 is assumed to be linked to the recessive locus and marker D8S1145 is assumed to be linked to the dominant locus.

DISCUSSION

Results from our analyses provided evidence for AD susceptibility loci on chromosomes 1 and 6 under the COGA criterion. While under the ICD10 criterion, the results provided evidence for linkage on chromosomes 8 and 10. It is noteworthy that different chromosomes were identified as likely regions for AD loci under the two diagnostic criteria. Furthermore, chromosomes 6, 8 and 10 were not identified as likely regions for linkage in Reich et al. [1998]. On the other hand, while they found possible AD loci on chromosome 7, our results provided little evidence for that. This difference in findings may be due to Reich et al. [1998] performing sib-pair based analyses whereas our analyses examine the sharing among all affected individuals. It may also be due to false positives, or false negatives. Regardless, these results suggest that there might be

multiple genes responsible for AD, though some of these results do not meet statistical significance. This is consistent with results from segregation analyses performed by Yuan et al. [1996], leading to the hypothesis that the underlying genetic mechanism for AD may be oligogenic. We examined several two-locus models to investigate whether they were more adequate in describing the genetic mechanism of AD. Besides the two types of models whose results were reported in the current paper, we also examined three other types of models: dominant/dominant, threshold, and heterogeneity [Schork et al., 1993; Goldstein et al., 1996]. Other than the double recessive model, parameters of all the other models were estimated assuming population prevalence of 0.1 and relative risk of 0.3 for a child of an affected parent. The overall population prevalence of 0.1 was chosen because the data description provided the risk for AD to be about 17% in males and 4% in females [Reich]. In Yuan et al. [1996], they estimated the relative risk of affection for children of affected parents to be about twice as high as that for children of unaffected parents. On the other hand, our estimate of relative risk from the data was higher. Based on these, we chose the relative risk to be 3 times of the population prevalence. Regardless, these model parameters were not estimated simultaneously in our linkage analysis, guaranteeing that the asymptotic distribution of our test statistic to be correct. Of all the models examined, the only model that showed statistical significance for linkage at an α level of 0.0001 was the double recessive model, and we note that the significance level was not adjusted for multiple tests. However, there is no reason to believe that this model was adequate in describing the mechanism of AD. This model was chosen primarily for its simplicity rather than for its plausibility. For example, the population prevalence under this model was 0.0001, which was far from the true population characteristic. One possible explanation for the significance result is that we might have found a false positive result with an incorrectly specified model. Our results also suggest that the underlying true model for AD might be much more complex than any of the two-locus models considered.

ACKNOWLEDGMENTS

This work was supported in part by NSF grant DMS-9632117 (to S. Lin) and NIH grants GM58934 and P30CA16058 (to F. Wright).

REFERENCES

Goldstein AM, Goldin LR, Dracopoli NC, Clark WH, Jr (1996): Two-locus linkage analysis of cutaneous malignant melanoma/dysplastic nevi. Am J Hum Genet 58: 1050-1056.

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

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996): Parametric and Nonparametric Linkage Analysis: A Unified Approach. Am J Hum Genet 58: 1347-1363.

Lathrop GM, Lalouel JM, Julier C, Ott J (1984): Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81: 3443-3446.

Lathrop GM, Ott J (1990): Analysis of complex disease under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet 47: A188.

Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, Foroud T, Hesselbrock V, Schuckit MA, Bucholz K, Porjesz B, Li T-K, Conneally M, Nurmberger JI, Jr, Tischfield JA, Crowe RR, Cloninger CR, Wu W, Shears S, Carr K, Crose C, Willig C, Begleiter H (1998): Genome-Wide Search of Genes Affecting the Risk for Alcohol Dependence. Am J Med Genet 81: 207-215.

Schork NJ, Boehnke M, Terwilliger JD, and Ott J (1993): Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am J Hum Genet 53: 1127-1136.

Yuan H, Marazita ML, Hill SY (1996): Segregation analysis of alcoholism in high density families: a replication. Am J Med Genet 67:71-76.