Genetic Epidemiology 19(Suppl 1):S22-S28 (2000)

Variance-Components QTL Linkage Analysis of Selected and Non-Normal Samples: Conditioning on Trait Values

Pak Chung Sham,[1]* Jing Hua Zhao,[1] Stacey S. Cherny,[2] and John K. Hewitt[2]

[1]Social, Genetic and Developmental Psychiatry Research Center and Department of Psychiatry, Institute of Psychiatry, London, United Kingdom
[2]Institute for Behavioral Genetics, University of Colorado, Boulder, Colorado

Standard variance-components quantitative trait loci (QTL) linkage analysis can produce an elevated rate of type 1 errors when applied to selected samples and non-normal data. Here we describe an adjustment of the log-likelihood function based on conditioning on trait values. This leads to a likelihood ratio test that is valid in selected samples and non-normal data, and equal in power to alternative methods for analyzing selected samples that require knowledge of the ascertainment procedure or the trait values of non-selected individuals. Genet. Epidemiol. 19(Suppl 1):S22-S28, 2000. © 2000 Wiley-Liss, Inc.

Key words: quantitative trait loci (QTL), linkage, selection, normality, conditioning

INTRODUCTION

Standard variance-components quantitative trait loci (QTL) linkage analysis [Schork, 1993; Amos, 1994; Kruglyak and Lander, 1995; Eaves et al., 1996; Almasy and Blangero, 1998; Fulker et al., 1999] is not robust to non-random ascertainment [Dolan et al., 1999] or non-normality [Allison et al., 1999]. We propose that modifying the likelihood function by conditioning on trait values will overcome these problems. This method is demonstrated here for sib-pairs, but can be generalized to pedigrees.

*Correspondence to: Dr. P.C. Sham, SGDP Research Center, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF, United Kingdom. E-mail: p.sham@iop.kcl.ac.uk

METHODS

Model Specification

The trait values of a sib-pair, denoted $\mathbf{x} = (x_1, x_2)^T$, conditional on the proportion of alleles identical-by-descent (IBD) at a test locus, π, are assumed to be bivariate normal. The log-likelihood function of a sib-pair is therefore

$$\ln L(\mathbf{x} \mid \pi) = -\frac{1}{2}\left(\ln|\Sigma_\pi| + (\mathbf{x}-\mu)^T \Sigma_\pi^{-1}(\mathbf{x}-\mu)\right),$$

where μ and Σ_π are the predicted mean vector and covariance matrix, respectively [Fulker et al., 1999; Sham et al., 2000]. The covariance matrix is given by

$$\Sigma_\pi = \begin{bmatrix} \sigma_A^2 + \sigma_S^2 + \sigma_N^2 & \pi\sigma_A^2 + \sigma_S^2 \\ \pi\sigma_A^2 + \sigma_S^2 & \sigma_A^2 + \sigma_S^2 + \sigma_N^2 \end{bmatrix},$$

where σ^2_A, σ^2_S and σ^2_N are variance components due to QTL, residual shared effects, and residual non-shared effects, respectively. The residual shared and non-shared variances (σ^2_S and σ^2_N) are not of primary interest but must be estimated with the QTL variance. We suggest a different way of writing the covariance matrix as

$$\Sigma_\pi = \begin{bmatrix} v & rv + (\pi - .5)\sigma_A^2 \\ rv + (\pi - .5)\sigma_A^2 & v \end{bmatrix},$$

where v is the variance and r the sib correlation of the trait in the population. The admissible range for σ^2_A is $[0, 2vr]$ if $r \leq ½$, $[0, 2v(1-r)]$, if $r>½$. The parameters μ, v and r are fixed at values obtained from previous studies of the same trait, or from preliminary analysis of the sib-pair data. Estimates of μ, v and r obtained from modeling selected sib-pair data will be unbiased only if an appropriate adjustment for ascertainment is made. Otherwise, an unbiased estimate of r can be obtained if correct values of μ and v are specified, in samples ascertained via probands [see Sham, 1997, Page 243-244]. Misspecification of μ, v and r will reduce the power to detect linkage.

In practice, π is estimated to various degrees of certainty from the genotypes (G) at marker loci in the vicinity of the test locus. If we assume that the likelihood is dependent on G only through π, then

$$L(\mathbf{x} \mid G) = \sum_\pi L(\mathbf{x} \mid \pi) P(\pi \mid G) \approx L(\mathbf{x} \mid \hat{\pi}),$$

where the summation is over $\pi = 0, ½, 1$, and $\hat{\pi}$ is the expected proportion of IBD sharing given G. The approximation based on $\hat{\pi}$ is adequate when marker data are highly informative.

Conditioning on Trait Values

For the analysis of selected samples [Eaves and Meyer, 1994; Risch and Zhang, 1995; Gu et al., 1996; Dolan and Boomsma, 1998; Purcell et al, 2000] and mildly non-normal data, we propose defining the likelihood of the genotype data conditional on trait values as

$$L(G \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid G)P(G)}{L(\mathbf{x})} \propto \frac{\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi \mid G)}{\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi)} \approx \frac{L(\mathbf{x} \mid \hat{\pi})}{\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi)},$$

where $P(G)$ is omitted because it does not involve any parameter. Note that the $\hat{\pi}$ approximation cannot be used for the denominator, because here $\hat{\pi}$ =½, and the approximate likelihood is independent of the QTL variance σ^2_A.

Alternative Methods for Analyzing Selected Samples

We consider two alternative methods for analyzing selected samples. The first is that suggested by Eaves et al. [1996], of imputing the prior IBD probabilities of ¼, ½, and ¼ for non-selected sib-pairs with known trait values, but unknown marker genotypes. This method was shown to produce an unbiased test for linkage, provided that one uses the exact "weighted likelihood" rather than the $\hat{\pi}$ approximation [Dolan et al., 1999].

The second is the classical method of ascertainment correction, which involves conditioning the family data on the event that the family is selected, under an assumed model of ascertainment [Fisher, 1934; Morton, 1959; Morton and MacLean, 1974]. Here we assume that all sib-pairs whose trait values fall within certain predefined ranges (denoted R) have equal probability of being selected; the probability being proportional to the integral of the density function of sib pair trait values over R. The ascertainment-adjusted likelihood function is then defined as

$$L_A(\mathbf{x} \mid G) = \frac{\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi \mid G)}{\int_R \left(\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi) \right) d\mathbf{x}} \approx \frac{L(\mathbf{x} \mid \hat{\pi})}{\int_R \left(\sum_{\pi} L(\mathbf{x} \mid \pi)P(\pi) \right) d\mathbf{x}}$$

The integration is straightforward when the selection region is simple. This is the case for the method proposed by Risch and Zhang [1995]. This method defines the lower and upper limits of selection to be the fixed values t_1 and t_2. The selection region R consists of the 4 quadrants: ($x_1<t_1$, $x_2<t_1$), ($x_1>t_2$, $x_2>t_2$), ($x_1<t_1$, $x_2>t_2$), ($x_1>t_2$, $x_2<t_1$).

An "approximation" to the integral can be used when the cut-offs for selection are unknown or non-linear. This approximation is based entirely on the trait data of the selected sib-pairs. By definition, these sib-pairs must fall within the region of selection, so that the distribution of trait values among them should provide information on the boundaries of the selection region. Furthermore, if the selected sib-pairs were evenly

scattered in the region of selection, a reasonable approximation to the integral then would be the average likelihood function of the trait values of the sib-pairs. Since the expected number of selected sib-pairs in a region is proportional to the likelihood, a reasonable weight for the contribution of a sib-pair with trait value x is the inverse of $L(x \mid \pi = ½)$. This gives the "approximation"

$$\sum_{i=1}^{n}\left(\frac{\sum_{\pi} L(\mathbf{x}_i \mid \pi)P(\pi)}{L\left(\mathbf{x}_i \mid \pi = \frac{1}{2}\right)}\right),$$

where the outer summation is over all n selected sib-pairs.

Simulation Studies

We simulated trait values and marker genotypes under both the null and a range of alternative hypotheses. Genotype data were simulated for 4 marker loci with 4 equally frequent alleles, equally spaced at 5 cM intervals. Simulations under an alternative hypothesis involved a diallelic QTL accounting for 10% of the phenotypic variance ($\sigma^2_A = 0.1$) at the midpoint between the second and third markers, and an overall sibling correlation $r = 0.2$. For each replicate, trait and marker data for 20,000 sib-pairs were simulated, and the trait was standardized to have mean 0 and variance 1. Each replicate represented a random sample from which sib-pairs with both members more than one standard deviation away from the mean (approximately 10% of sib-pairs) were selected to form a subsample.

We simulated 500 replicate data sets under the null hypothesis ($\sigma^2_A = 0$) to assess the empirical type 1 error rates of the different tests. We also simulated 100 replicate data sets under each of a range of alternative hypotheses: an additive QTL with equal or unequal (0.1, 0.9) allele frequencies, and a dominant QTL with equal and unequal (dominant 0.1, recessive 0.9) allele frequencies.

We used Mapmaker/SIBS [Kruglyak & Lander, 1995] to compute the IBD probabilities of each sib-pair, given the marker genotype data, at the true position of the QTL. The phenotype data and these IBD probabilities were then used as input to a SAS program for the calculation of the different test statistics.

To assess the impact of non-normality, we repeated all the simulations but inserted a transformation between data generation and analysis. The transformation simply consisted of cubing the standardized trait values, followed by re-standardization. The transformed variable therefore had unchanged mean and variance (namely 0 and 1), but was markedly leptokurtic and had a reduced sib correlation of about 0.12. We set thresholds of selection at $\pm$ 0.3 to keep the proportion of selected sib-pairs at about 10%.

RESULTS

The results of the simulation studies are presented in Table I. For normally distributed data, all methods designed for analyzing selected samples led to correct type 1 error rates. Conditioning on trait values did not reduce power when applied to complete samples, and was as powerful as the imputation method (which requires the trait values

TABLE I. Properties of QTL Linkage Tests in Simulated Samples of 20,000 Sib Pairs*

	Significance	Power H_1	H_2	H_3	H_4
Normal data					
Complete samples					
Standard variance components	0.04	0.91	0.93	0.90	0.89
Conditioning on trait values	0.04	0.91	0.93	0.90	0.89
Selected samples					
Standard variance components	0.11	*0.87*	*0.87*	*0.81*	*0.86*
Conditioning on trait values	0.05	0.66	0.53	0.56	0.58
Imputing prior IBD distribution	0.05	0.68	0.54	0.55	0.58
Ascertainment correction 1	0.05	0.66	0.53	0.57	0.57
Ascertainment correction 2	0.05	0.66	0.53	0.56	0.58
Non-normal data					
Complete samples					
Standard variance components	0.13	*0.44*	*0.64*	*0.46*	*0.56*
Conditioning on trait values	0.04	0.09	0.19	0.13	0.22
Selected samples					
Standard variance components	0.15	*0.33*	*0.47*	*0.33*	*0.34*
Conditioning on trait values	0.06	0.07	0.15	0.14	0.12
Imputing prior IBD distribution	0.25	*0.83*	*0.91*	*0.75*	*0.83*
Ascertainment correction 1	0.15	*0.32*	*0.47*	*0.32*	*0.33*
Ascertainment correction 2	0.05	0.04	0.11	0.12	0.11

*Significance and power are the proportions of replicates with likelihood ratio test statistic exceeding 2.70 (i.e., $p = 0.05$) and 9.55 (i.e., $p = 0.001$), respectively. Power estimates in italic are not valid due to inflated type 1 error rate. Alternative hypotheses: H_1 additive, equal allele frequencies; H_2 additive, allele frequencies 0.1 and 0.9; H_3 dominant, equal allele frequencies; H_4 dominant, frequency of dominant allele 0.1. Ascertainment correction 1 involves integration. Ascertainment method 2 uses an approximation of the integral.

of the non-selected sib-pairs) and the integration method (which requires knowledge of the selection criteria) when applied to selected samples.

For non-normal data, conditioning on trait values and approximating the integral led to correct type 1 error rates, whereas the imputation and integration methods were liberal. However, both valid tests (conditioning on trait values and approximating the integral) lost considerable power as compared to similar analyses of the original data before transformation.

DISCUSSION

Conditioning on trait values is an application of the "model-free" method of ascertainment correction [Ewens and Shute, 1986]. It is also implicit in classical parametric linkage analysis [Hodge and Elston, 1994]. Conditioning on trait values is therefore expected to enjoy the same benefits as classical parametric linkage analysis, i.e., robustness to the ascertainment procedure [Williamson and Amos, 1995] and model-misspecification [Clerget-Darpoux et al., 1986], as well as minor violations of distributional assumptions. Gross non-normality will substantially reduce power; thus, either alternative methods of analysis should be used [see Allison et al., 1999, for a discussion of alternative approaches], or the data transformed to approximate normality before applying the proposed method.

It is interesting to note that the log-likelihood function after conditioning on trait values is 0 when $\sigma^2_A = 0$. This means that the adjusted log-likelihood function maximized under the alternative hypothesis ($\sigma^2_A > 0$) is itself the likelihood ratio test statistic. It is as if the adjustment term represents the log-likelihood function under a null hypothesis that is not $\sigma^2_A = 0$, but one that assumes the presence of a QTL with an IBD distribution of ¼, ½ and ¼ rather than the values estimated from the marker genotype data. Thus, the existence of a QTL is assumed, and one is merely testing whether the QTL is linked to the test locus.

ACKNOWLEDGMENTS

This work was supported by the Medical Research Council (G9700821) and the Wellcome Trust (055379) in the United Kingdom and the National Institutes of Health (EY12562, AA07330, AA10556, DA11015, MH43899 and MH53668) in the United States. Software is available at http://alpha.iop.kcl.ac.uk/qtl.

REFERENCES

Allison DB, Neale MC, Zannolli R, Schork NJ, Amos CI, Blangero J. 1999. Testing the robustness of the likelihood ratio test in a variance component quantitative trait loci mapping procedure. Am J Hum Genet 65:531-544.

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

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535-543.

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J. 1986. Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393-399.

Dolan CV, Boomsma DI. 1998. Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. Behav Genet 28:197-206.

Dolan CV, Boomsma DI, Neale MC. 1999. A simulation study of the effects of assignment of prior identity-by-descent probabilities to unselected sib pairs, in covariance-structure mapping of a quantitative trait locus. Am J Hum Genet 64:268-280.

Eaves L, Meyer J. 1994. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotpying. Behav Genet 24:443-455.

Eaves LJ, Neale MC, Maes H. 1996. Multivariate multipoint linkage analysis of quantitative trait loci. Behav Genet 26:519-525.

Ewens WJ, Shute NCE. 1986. A resolution of the ascertainment sampling problem. Theor Popul Biol 30:388-412.

Fisher RA. 1934. The effect of methods of ascertainment upon the estimation of frequencies. Ann Eugenics 6:13-25.

Fulker DW, Cherny SS, Sham PC, Hewitt JK. 1999. Combined linkage and association analysis for quantitative traits. Am J Hum Genet 64:259-267.

Gu C, Todorov A, Rao DC. 1996. Combining extremely concordant sib pairs with extremely concordant sib pairs provides a cost effective way to linkage analysis of quantitative traits. Genet Epidemiol 13:513-533.

Hodge SE, Elston RC. 1994. Lods, wrods, and mods: the interpretation of lod scores calculated under different models. Genet Epidemiol 11:329-342.

Kruglyak L, Lander E. 1995. Complete multipoint sib pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454.

Morton NE. 1959. Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1-16.

Morton NE, MacLean CJ. 1974. Analysis of family resemblance III. complex segregation analysis of quantitative traits. Am J Hum Genet 26: 489-503.

Purcell S, Cherny SS, Hewitt JK, Sham PC. 2000. Optimal sibship selection for genotyping in QTL linkage analysis. Human Hered (in press).

Risch N, Zhang H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589.

Schork NJ. 1993. Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am J Hum Genet 55:1306-1319.

Sham PC. 1997. Statistics in human genetics. London: Edward Arnold.

Sham PC, Cherny SS, Purcell S, Hewitt JK. 2000. Power of linkage versus association analysis of quantitative traits using variance components models for sibship data. Am J Hum Genet 66:1616-1630.

Williamson JA, Amos CI. 1995. Guess LOD approach: sufficient conditions for robustness. Genet Epidemiol 12:163-176.