RESEARCH ARTICLE

Dissecting the correlation structure of a bivariate phenotype: common genes or shared environment?

SAURABH GHOSH*

Human Genetics Unit, Indian Statistical Institute, 203 BT Road, Kolkata 700 108, India

Abstract

High correlations between two quantitative traits may be either due to common genetic factors or common environmental factors or a combination of both. In this study, we develop statistical methods to extract the genetic contribution to the total correlation between the components of a bivariate phenotype. Using data on bivariate phenotypes and marker genotypes for sib-pairs, we propose a test for linkage between a common QTL and a marker locus based on the conditional cross-sib trait correlations (trait 1 of sib 1 – trait 2 of sib 2 and conversely) given the identity-by-descent (i.b.d.) sharing at the marker locus. We use Monte-Carlo simulations to evaluate the performance of the proposed test under different trait parameters and quantitative trait distributions. An application of the method is illustrated using data on two alcohol-related phenotypes from a project on the collaborative study on the genetics of alcoholism.

[Ghosh S 2005 Dissecting the correlation structure of a bivariate phenotype: common genes or shared environment? *J. Genet.* **84**, 143–146]

Introduction

One of the current challenges of genetic epidemiology is to unravel the genetic architecture of complex traits. Heritable quantitative characters, possibly correlated, generally underlie complex traits. However, a high correlation between two quantitative traits need not necessarily imply a common quantitative trait locus (QTL) controlling both the traits, but may be due to common environmental factors. The aim of this study is to develop statistical methods to extract the genetic contribution to the total correlation between the components of a bivariate phenotype and hence, explore for common QTLs. Using data on bivariate phenotypes and marker genotypes for sib-pairs, we derive an expression, under certain assumptions, for the conditional cross-sib trait correlations (trait 1 of sib 1 – trait 2 of sib 2 and conversely) given the identity-by-descent sharing at a marker locus and develop a test for detecting linkage between the common QTL and the marker locus. Monte-Carlo simulations are included to assess the performance of the proposed procedure. We also present an application of our method to two correlated alcohol related endophenotypes using data from the Collaborative Study on the Genetics of Alcoholism (COGA) project.

Model

Suppose that two correlated quantitative traits Y and Z are controlled by a common biallelic QTL with alleles A and a. Suppose we have data on K sib-pairs. Let (y_{j1}, y_{j2}) and (z_{j1}, z_{j2}) denote the quantitative trait values of the j^{th} sib-pair corresponding to the first trait and the second trait, respectively. The model assumed is:

$$y_{ji} = G_{ji} + E_{ji};\ z_{ji} = H_{ji} + e_{ji};\ i = 1,2;\ j = 1,2,..., K$$

where, G_{ji} and H_{ji} are the genetic components and E_{ji} and e_{ji} are random errors with means 0, variances s_1^2 and s_2^2 and correlations between (E_{j1}, e_{j2}) and (e_{j1}, E_{j2}) being r. Suppose the expectations of Y and Z conditioned on the genotype at the QTL be (a_1, b_1, – a_1) and (a_2, b_2, – a_2), according as the genotype is (AA, Aa, aa). Then:

$$G_{ji} = d_{ji} + y_{ji};\ H_{ji} = \Delta_{ji} + f_{ji};\ i = 1,2;\ j = 1,2,..., K$$

where, d_{ji} and Δ_{ji} are the conditional expectations as mentioned above, y_{ji} and f_{ji} have means 0, and, variances t_1^2 and t_2^2. y_{ji} and f_{jl} are assumed to be uncorrelated for

*E-mail: saurabh@isical.ac.in.

$i \neq l$. Let $p_j = \{0, 0.5, 1\}$ denote the identity-by-descent (i.b.d.) sharing of the jth sib-pair at the QTL.

Statistical methods

Since correlations are scale-invariant, we can assume, without loss of generality, that $a_1 = 1$ and $a_2 = 1$. Then, using Table I of Haseman–Elston (1972), we can show that:

$$\text{Corr}\,(y_{j1}, z_{j2})\,|\,p_j$$
$$= \frac{\text{Cov}\,(d_{j1}, \Delta_{j2})\,|\,p_j + \text{Cov}\,(E_{j1,}\, e_{j2})}{\sqrt{\text{Var}(d_{j1} + y_{j1} + E_{j1})}\;\sqrt{\text{Var}(\Delta_{j2} + f_{j2} + e_{j2}}}$$
$$\sqrt{g_o + g_1 p_j + g_2\, I_{\{p_j = 0.5\}}}$$

where, $g_0 = r s_1 s_2 / 2pqV; g_1 = \{1 - (p-q)\,(b_1 + b_2) + (1-2pq)\; b_1 b_2\}/V;\;\; g_2 = 2p^2q^2 b_1 b_2/V;$

$$V = \sqrt{1 - 2(p-q)\; b_1 + (1-2pq)\; b_1^2 \; + \frac{t_1^2 + s_1^2}{2pq}} \times$$
$$\sqrt{1 - 2(p-q)\; b_2 + (1-2pq)\; b_2^2 \; + \frac{t_2^2 + s_2^2}{2pq}};$$

p and q are the frequencies of A and a.

By symmetry arguments, $\text{Corr}(z_{j1}, y_{j2})|p_j$ is identical to the above expression. We note that the genetic component of the correlation between y_{j1} and z_{j2} is the correlation between G_{j1} and H_{j2} and is always greater than g_1. Thus, it is easy to see that $[\text{Corr}(y_{j1}, z_{j2}) \mid p_j = 1] - [\text{Corr}(y_{j1}, z_{j2}) \mid p_j = 0]$ provides a lower bound for the genetic correlation between the two quantitative traits.

For deriving a method for detecting linkage between the common QTL and a marker locus, we assume that the dominance in the trait is negligible (i.e. $b_1 \approx 0$ and $b_2 \approx 0$). The effect of dominance is evaluated in the simulations section. Consider a marker locus in linkage equilibrium with the QTL with recombination fraction q. Suppose p_{mj} denotes the marker i.b.d. score for the jth sib-pair. Assuming complete parental genotypic information, the estimated marker i.b.d. score $\hat{p}_{mj}$ assumes 5 distinct values 0, 0.25, 0.5, 0.75, 1. If the trait genotypes of the jth sib-pair are T_{1j} and T_{2j}, respectively, we note that $P(T_{1j}, T_{2j}|\hat{p}_{mj}) = P(T_{1j}, T_{2j} \mid p_j)\; P(p_j \mid p_{mj})\; P(p_{mj}|\hat{p}_{mj})$. Thus, using the conditional probability distribution of p_j given p_{mj} [Table IV of Haseman and Elston (1972)] and that of p_{mj} given $\hat{p}_{mj}$ [multiallelic modification of Table V of Haseman and Elston (1972)], we can show that under no dominance at the trait:

$$r_{1\hat{p}} = \text{Corr}(y_{j1}, z_{j2})\,|\,\hat{p}_{mj} = a_0 + b_0 \hat{p}_{mj} \text{ and}$$
$$r_{2\hat{p}} = \text{Corr}(z_{j1}, y_{j2})\,|\,\hat{p}_{mj} = a_0 + b_0 \hat{p}_{mj}$$

Where, $a_0 = \{2q\,(1\text{–}q) + \dfrac{r s_1 s_2}{2pq}\}/V; b_0 = (1-2q)^2/V$

Thus, $q = 0.5 \Leftrightarrow b_0 = 0$ and $q < 0.5 \Leftrightarrow b_0 > 0$. We group the sib-pairs according to their estimated marker i.b.d. scores and compute the cross-sib correlations corresponding to each of the 5 groups.

We develop a least squares minimisation of

$$\sum_{i=1}^{2} \sum_{l=0}^{4} \{r_{i\hat{p}=\frac{l}{4}} - a_0 - b_0 \frac{l}{4}\}^2$$

with respect to a_0 and b_0. A test for linkage is equivalent to a test for $b_0 = 0$ versus $b_0 > 0$. The empirical p-value of the test is evaluated using permutation principles, that is, permuting the estimated i.b.d. scores among the sib-pairs. The estimated b_0 values for the different permutations are ranked in increasing order and the p-value is determined by the position of the observed b_0 value in that order.

Simulations

We perform simulations to evaluate the power of our proposed method. We generate data on the bivariate trait for 200 sib-pairs and marker genotypes at a marker, which is at recombination distance 0.01 with the common QTL and has 4 equifrequent alleles. In the *first* step, we generate the trait i.b.d. scores of the sib-pairs using a trinomial random number generator with cell probabilities (1/4, 1/2, 1/4). In the *second* step, we generate the QTL genotypes of the sib-pairs using a 9-variate random number generator with cell probabilities given by the conditional trait genotype distribution of sib-pairs given their trait i.b.d. score as provided in Table I of Haseman and Elston (1972). In the *third* step, we generate the marker i.b.d. scores of the sib-pairs using the conditional distribution of marker i.b.d. score given trait i.b.d. score as provided in Table IV of Haseman and Elston (1972). In the *fourth* step, we generate the estimated marker i.b.d. score of each sib-pair using the conditional distribution of the estimated marker i.b.d. score given the marker i.b.d. score using a multiallelic modification of Table V of Haseman and Elston (1972). In the *fifth* step, we generate the quantitative values of the two traits from (i) a bivariate normal distribution, (ii) a bivariate distribution with location-shifted chi-square marginals such that the mean vectors have components a_1, b_1 or – a_1 and, a_2, b_2 or – a_2, for the two traits respectively, according as the trait genotype is AA, Aa or aa. In all our simulations, we used fixed parameter values $a_1 = 5$, $a_2 = 3$, $s_1 = s_2 = 1$, $t_1 = t_2 = 0.5$ and $r = 0.5$.

The results of our simulations are presented in table 1 for normally distributed traits and in table 2 for chi-square distributed traits. The empirical powers of our proposed test procedure are based on 1000 replications. In each replication, the permutation test has size 0.05 (that is, we reject the null hypothesis of no linkage if the estimated b_0 value is not within the smallest 95% b_0 values generated via permutations). From both tables 1 and 2, we find that our method performs well especially for low dominance (b_1 and b_2) and high heterozysosity ($2pq$) at the trait locus. The linear relationship used in the least squares minimization is exact in the absence of dominance at the trait. Thus, the power of the test decreases with increase in dominance. Compared to normally distributed traits, we find that the powers of the tests are lower when the traits are distributed as chi-square for the same simulation parameter values. Thus, skewness in the trait distribution leads to a reduction in the power of the test.

An Application to COGA

The Collaborative Study on the Genetics of Alcoholism (COGA) is a multicenter research program established to detect and map susceptibility genes for alcohol dependence and related phenotypes. Genome-wide scans on two endophenotypes: maximum number of drinks in a 24 h period (Saccone *et al*. 2000) and the number of externalizing symptoms associated with the COGA (DSM-III-R + Feighner definite) alcoholism diagnosis (Ghosh, Beirut, Porjesz, Edenberg, Foroud, Goate *et al*. unpublished observations) have provided significant linkage findings near the ADH3 marker on Chromosome 4. Since this marker belongs to the alcohol dehydrogenase gene cluster (ADH-17), it is of potential interest as a candidate gene for alcohol-related endophenotypes. We used data on 171 independent sib-pairs to evaluate the performance of our proposed correlation-based method in detecting linkage with the ADH3 marker. The estimated correlation (considering one sib per sib-pair) between the two endophenotypes is 0.57. The proposed test for $b_0 = 0$ versus $b_0 > 0$ yielded a p-value < 0.0005, indicating the possibility of a common QTL for the two endophenotypes.

Table 1. Empirical power of the test procedure when the traits are distributed as normal with simulation parameter values $a_1 = 5$, $a_2 = 3$, $r = 0.5$, $t_1^2 = t_2^2 = 0.25$, $s_1^2 = s_2^2 = 1$, $q = 0.01$.

p	b_1	b_2	Power
0.5	0	0	0.889
	2.5	0	0.822
	0	1.5	0.828
	2.5	1.5	0.773
0.7	0	0	0.845
	2.5	0	0.786
	0	1.5	0.780
	2.5	1.5	0.731
0.9	0	0	0.774
	2.5	0	0.721
	0	1.5	0.718
	2.5	1.5	0.653

Table 2. Empirical power of the test procedure when the traits are distributed as located-shifted chi-square with simulation parameter values $a_1 = 5$, $a_2 = 3$, $r = 0.5$, $t_1^2 = t_2^2 = 0.25$, $s_1^2 = s_2^2 = 1$, $q = 0.01$.

p	b_1	b_2	Power
0.5	0	0	0.852
	2.5	0	0.787
	0	1.5	0.793
	2.5	1.5	0.738
0.7	0	0	0.819
	2.5	0	0.755
	0	1.5	0.753
	2.5	1.5	0.716
0.9	0	0	0.736
	2.5	0	0.698
	0	1.5	0.701
	2.5	1.5	0.637

Discussion

We have developed a distribution-free method of detecting linkage between a common QTL controlling two correlated traits and a marker locus. Since we are using cross-sib correlations between the bivariate traits for our analyses, the proposed method will be able to decipher whether the correlation between the traits is due to a common QTL or due to common environmental factors. If the two traits do not have a common QTL and the marker locus is linked to a QTL controlling only one of the traits, the estimated b_0 coefficient in the least squares minimization will not be significant. We wish to emphasize that the proposed methodology is aimed more at validating the existence of a common QTL as the source of the correlation between the quantiative traits rather than a bivariate linkage mapping procedure.

As in the classical Haseman–Elston regression method (1972) and its extensions, the performance of the method detoriates with increase in dominance at the trait. Dominance induces skewness in the trait and a deviation from the linear relationship between the cross-sib trait correlation and the estimated marker i.b.d. sharing.

When parental genotypes are not available, the estimation of marker i.b.d. scores involves knowledge of allele frequencies at the marker loci and these estimated scores will not be restricted to the set {0, 0.25, 0.5, 0.75, 1}. In such a case, the proposed method can not be applied directly. One way to modify is to group the estimated i.b.d. scores into 5 intervals: (0, 0.125), (0.125, 0.375), (0.375, 0.625), (0.625, 0.875), (0.875, 1); compute the cross-sib trait correlations for these 5 groups and use the midpoints of these intervals as the estimated i.b.d. score in the least squares minimization.

We are currently exploring possible extensions of the proposed methodology for multivariate phenotypes comprising more than two traits.

Acknowledgements

This work was supported by the grant R01-TW-6604-2 from the Fogarty International Center of NIH.

References

Haseman J. K. and Elston R. C. 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19.

Saccone N. L., Kwon J. M., Corbett J., Goate A., Rochberg N. and Edenberg H. J. *et al*. 2000 A genome screen of maximum number of drinks as an alcoholism phenotype. *Am. J. Med. Genet.* **96**, 632–637.

Received 12 January 2005