Report

A Unified Haseman-Elston Method for Testing Linkage with Quantitative Traits

Xin Xu,[1] Scott Weiss,[1,3] Xiping Xu,[1,3] and L. J. Wei[2]

[1]Program for Population Genetics and [2]Department of Biostatistics, Harvard School of Public Health, and [3]Channing Laboratory, Brigham and Women's Hospital, Boston

The Haseman and Elston (H-E) method uses a simple linear regression to model the squared trait difference of sib pairs with the shared allele identical by descent (IBD) at marker locus for linkage testing. Under this setting, the squared mean-corrected trait sum is also linearly related to the IBD sharing. However, the resulting slope estimate for either model is not efficient. In this report, we propose a simple linkage test that optimally uses information from the estimates of both models. We also demonstrate that the new test is more powerful than both the traditional one and the recently revisited H-E methods.

The Haseman-Elston (H-E) method is widely used in genetic linkage studies that use sib pairs (Haseman and Elston 1972) to study quantitative traits. Specifically, let X_{ik} and π_{ik} be the squared trait difference and the proportion of alleles shared identical by descent (IBD) at the marker of interest for the kth sib pair of the ith family, $k = 1,\ldots, K_i$; $i = 1,\ldots, n$. Given π_{ik}, the H-E method specifies that

$$EX_{ik} = \alpha_1 - \beta\pi_{ik} \ , \qquad (1)$$

where EX is the expected value of X and where α_1 and β are unknown parameters to be estimated. A large observed value of the standardized estimate for β suggests a linkage between the trait and marker loci. Note that, for the large sample case, the test proposed by H-E is nonparametric, in the sense that the null distribution of the test statistic does not depend on the distribution of the trait values. However, if the traits are normally distributed, the H-E method can be substantially less efficient than the maximum-likelihood–estimation procedure (Amos et al. 1996; Fulker and Cherny 1996; Wright 1997). To increase the power of the H-E test for linkage, one may also consider the trait sums of sib pairs in the analysis. Let Y_{ik} be the squared mean-corrected trait sum for the kth sib pair of the ith family. Then, it follows from Drigalenko (1998) that

$$-EY_{ik} = \alpha_2 - \beta\pi_{ik} \ , \qquad (2)$$

where α_2 is an unknown intercept. Now, let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the least-squares estimators for β of models (1) and (2), respectively. Then, under the condition that the variances of the error terms of models (1) and (2) are equal, the estimator $\tilde{\beta} = (\hat{\beta}_1 + \hat{\beta}_2)/2$ is more efficient than $\hat{\beta}_1$ and $\hat{\beta}_2$ (Drigalenko 1998). This estimator is equivalent to the least-squares estimator of b for the model

$$2EZ_{ik} = \alpha_3 + \beta\pi_{ik} \ , \qquad (3)$$

where α_3 is an unknown intercept and Z_{ik} is the mean-corrected trait product for the kth sib pair of the ith family. On the basis of this observation and numerical studies, Elston et al. (2000) proposed to use model (3) for testing linkage. However, when the trait values among siblings are moderately or highly correlated, the test based on the estimate of β in model (3) may not be efficient. This motivates us to consider estimation procedures for β, which optimally use the information from $\hat{\beta}_1$ and $\hat{\beta}_2$.

Let us consider a class of linear estimators for β, using both model (1) and model (2) in the form of $w\hat{\beta}_1 + (1 - w)\hat{\beta}_2$, where w is a given weight, which may be data dependent. Note that $\hat{\beta}_1$, $\hat{\beta}_2$, and $\tilde{\beta}$ are special members in this class. Let $\hat{\sigma}_{12}$ be the estimated covariance of $\hat{\beta}_1$

Received June 16, 2000; accepted for publication August 11, 2000; electronically published August 28, 2000.

Address for correspondence and reprints: Dr. Xin Xu, Program for Population Genetics, FXB-103 Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115. E-mail: xin_xu@harvard.edu

0002-9297/2000/6704-0030$02.00

Table 1

Empirical Significance Level for the New Test

ρ AND n	SIBSHIP SIZE	NOMINAL SIZE OF TEST			
		.05000	.01000	.00100	.00010
0:					
50	2	.06854	.01734	.00250	.00042
	5	.04430	.00737	.00053	.00004
100	2	.05888	.01323	.00152	.00018
	5	.04254	.00658	.00041	.00003
200	2	.05435	.01143	.00115	.00014
	5	.04324	.00698	.00043	.00003
500	2	.05190	.01067	.00107	.00010
	5	.04492	.00763	.00058	.00004
.3:					
50	2	.06850	.01717	.00253	.00038
	5	.04286	.00718	.00055	.00004
100	2	.05815	.01243	.00136	.00015
	5	.04202	.00691	.00046	.00002
200	2	.05427	.01107	.00109	.00010
	5	.04311	.00715	.00050	.00004
500	2	.05179	.01040	.00099	.00009
	5	.04505	.00792	.00060	.00005

and $\hat{\beta}_2$ and let $\hat{\sigma}_{11}$ and $\hat{\sigma}_{22}$ be the estimated variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Then, for large n, the estimator $\hat{\beta}$ with weight $w = (\hat{\sigma}_{22} - \hat{\sigma}_{12})/(\hat{\sigma}_{11} + \hat{\sigma}_{22} - 2\hat{\sigma}_{12})$ has the smallest variance among all the linear combinations of $\hat{\beta}_1$ and $\hat{\beta}_2$ (Rao 1965; Wei and Johnson 1985). Furthermore, this estimator is approximately normally distributed, with mean β and variance $(\hat{\sigma}_{11}\hat{\sigma}_{22} - \hat{\sigma}_{12}^2)/(\hat{\sigma}_{11} + \hat{\sigma}_{22} - 2\hat{\sigma}_{12})$. Note that the weight w is data dependent. Essentially, we let the data guide us to choose the weight, regardless of the true correlation among sib-pair trait values. The distribution theory mentioned above is valid not only for the least-squares estimators but also for any consistent estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ for models (1) and (2) (e.g., see Hettmansperger 1984, for rank-estimation procedures for linear-regression models).

It is important to note that study families often include multiple siblings—for this case, $\hat{\sigma}$ cannot be obtained by use of the statistical software for the standard linear-regression model. If we use the least-squares principle to estimate parameters, consistent estimators $\hat{\alpha}_1$ and $\hat{\beta}_1$ for model (1) can be obtained by solving the following equation by the generalized estimating equations (GEE) techniques with an independent working model (Liang and Zeger 1986):

$$\sum_{i=1}^{n}\sum_{k}(1\pi_{ik})(X_{ik} - \alpha_1 + \beta\pi_{ik}) = 0 \ . \qquad (4)$$

Similarly, consistent estimators $\hat{\alpha}_2$ and $\hat{\beta}_2$ for model (2) can be obtained by solving the equation with X and α_1 in equation (4) being replaced by $-Y$ and α_2, respectively. With this technique, let $e_{ik1} = X_{ik} - \alpha_1 + \beta_1\pi_{ik}$ and let $e_{ik2} = -Y_{ik} - \alpha_2 + \beta_2\pi_{ik}$. Then,

$$\hat{\sigma}_{ml} = \frac{\sum_i\left[\sum_k e_{ikm}(\pi_{ik} - \bar{\pi})\sum_k e_{ikl}(\pi_{ik} - \bar{\pi})\right]}{\left[\sum_i\sum_k(\pi_{ik} - \bar{\pi})^2\right]^2} \ ,$$

where $m,l = 1,2$ and $\bar{\pi}$ is the average of π at the marker of interest among all the sib pairs in the study. Note that $\hat{\beta}_1$ and $\hat{\beta}_2$ are the ordinary least-squares estimators for β but that their estimated variances and covariance may be quite different from those obtained from the standard linear-regression analysis.

To investigate the performance of the proposed estimator $\hat{\beta}$, we conducted an extensive numerical study to examine whether the test based on $\hat{\beta}$ preserves the nominal type I–error probabilities and its power profile. Specifically, in our simulation, we assumed that, for the jth member in the ith family, its trait value T_{ij} follows a random-effects model:

$$T_{ij} = G_{ij} + C_i + \varepsilon_{ij} \ ,$$

where G_{ij} is the random variable from the locus-specific contribution to the trait defined by the genotypes and the genetic model, C_i is a random-effect variable that is the common component shared by the offspring in the same family, including both genetic and environmental factors, and e is the error term. These three random components are assumed to be independent. For G, we assume that there are two alleles (A and a, with frequencies p and $1 - p$, respectively) for the trait locus, and the trait contributions from the genotypes AA, Aa, and aa are 1, d, and -1, respectively. The value of d under the additive-, dominant-, and recessive-inheritance modes is 0, 1, and -1, respectively. The variance of G can be expressed as $V_a + V_d$, where $V_a = 2p(1 - p)[1 - d(2p - 1)]^2$ and $V_d = 4p^2(1 - p)^2d^2$. Furthermore, we assume that C and e are normally distributed, with mean 0 and variances V_c and V_ε. Now, let the locus-specific heritability $h = (V_a + V_d)/(V_a + V_d + V_c + V_\varepsilon)$ and let the sibling residual trait correlation $\rho = V_c/(V_a + V_d + V_c + V_\varepsilon)$. Then, V_c and V_ε are uniquely determined by V_a, V_d, h, and ρ. In our simulation, we assumed a fully informative marker at the disease locus, with zero recombination and with parental genotypes assigned randomly according to the allele frequencies. The offspring's haplotypes were derived by random transmission of parental alleles.

Under the null hypothesis, the variance of G is 0. To test the validity of our test constructed from $\hat{\beta}$, we examined whether the empirical levels of the test are close to their nominal counterparts. To this end, we considered 16 cases with $\rho = 0$, .3, two sibship sizes (2 and 5), and the number of families $n = 50$, 100, 200, and 500. For each case, we simulated 1 million realizations. The resulting empirical levels of the test for various nominal

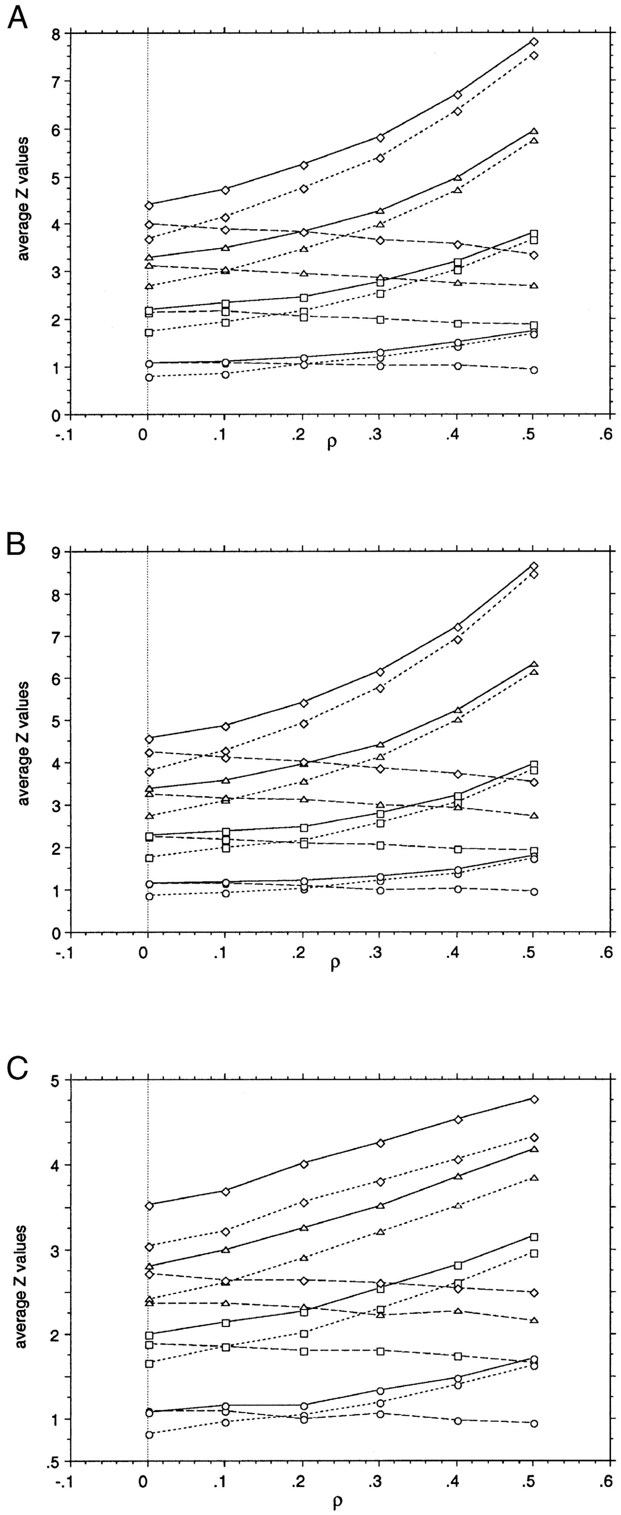

Figure 1 Power comparisons between the new procedure and the two H-E tests for linkage. *A*, Additive-trait locus with $P = .1$. *B*, Dominant-trait locus with $P = .1$. *C*, Recessive trait with $P = .2$. The average Z scores using traditional (model [1]), revisited (model [3]), and unified H-E methods are represented by dotted, dashed, and solid lines, respectively. Locus-specific heritability is $h = .1$ (*circle*), .2 (*square*), .3 (*triangle*), and .4 (*diamond*).

significance levels are reported in table 1. With respect to the type I–error rate, our procedure behaves reasonably well even when the nominal level is as low as .0001. For $n < 200$ and when each family has one sib pair, our procedure is slightly liberal, in the sense that the empirical levels tend to be higher than the nominal levels. On the other hand, when study families have multiple sib pairs, our method tends to be slightly conservative.

Although, in theory, the estimator $\hat{\beta}$ is more efficient than $\hat{\beta}_1$, $\hat{\beta}_2$, and $\tilde{\beta}$, it is important to know whether this optimality property holds for practical situation. To this end, we considered 360 cases, with three genetic models (additive, dominant, and recessive); $h = .1, .2, .3$, and .4; $P = .1, .2, .3, .4$, and .5; and $\rho = 0, .1, .2, .3, .4$, and .5. For each case, we simulated 1,000 realizations of 1,000 independent sib-pair observations. In figure 1, we summarize the power profiles for the H-E methods with models (1) and (3) and for our procedure, using the average Z score, based on 1,000 realizations. Figure 1*A* is for an additive-trait locus with $P = .1$, figure 1*B* is for a dominant-trait locus with $P = .1$, and figure 1*C* is for a recessive-trait locus with $P = .2$. The average Z scores for the additive and dominant models do not vary much with p, whereas those for the recessive model increase with p. On the basis of our numerical study, the power for the traditional H-E method based on model (1) increases with ρ, but the power of the revisited H-E method based on model (3) decreases with ρ. The revisited H-E method is more powerful than the traditional method when the sibling trait correlation is low, and the traditional method is more powerful when the sibling trait correlation is high. On the other hand, our procedure using models (1) and (2) simultaneously is more powerful than the two H-E methods mentioned above, especially when the sibling trait values are moderately correlated.

In this report, we have demonstrated that the distribution-free test for linkage based on a particular linear combination of the estimators for the slope β in models (1) and (2) is more powerful than the existing nonparametric tests. The new test procedure is valid for the case with multiple sibs in the study families. Although, to gain efficiency for estimation of β, one may use an elaborate working model for the GEE approach, it seems rather difficult to specify, for the estimating function, a weight function that reflects the true correlation among the squared differences and sums. The new method was implemented in an executable program named "XWXW," which runs on various computer platforms and is available from the FBAT Web site.

Acknowledgment

We thank Dr. Nan Laird for helpful comments on the manuscript.

Electronic-Database Information

The URL for data in this article is as follows:

FBAT, http://www.biostat.harvard.edu/~fbat/default.html (for the XWXW program)

References

Amos CI, Zhu DK, Boerwinkle E (1996) Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60:143–160

Drigalenko E (1998) How sib pairs reveal linkage. Am J Hum Genet 63:1242–1245

Elston RC, Buxbaum S, Jacobs KB, Olson JM (2000) Haseman and Elston revisited. Genet Epidemiol 19:1–17

Fulker DW, Cherny SS (1996) An improved multipoint sibpair analysis of quantitative traits. Behav Genet 26:527–532

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Hettmansperger TC (1984) Statistical inference based on ranks. John Wiley, New York

Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Rao CR (1965) Linear statistical inference and its applications. John Wiley, New York

Wei LJ, Johnson WE (1985) Combining dependent tests with incomplete repeated measurements. Biometrika 72:359–364

Wright FA (1997) The phenotypic difference discards sibpair QTL linkage information. Am J Hum Genet 60:740–742