Letters to the Editor

Am. J. Hum. Genet. 63:1243, 1998

How Sib Pairs Reveal Linkage

To the Editor:

The Haseman-Elston (1972) method, widely used for studying linkage, has been criticized for incomplete utilization of sib-pair information. As an alternative, Amos (1994) created and advocates the "variance-components" approach; Wright (1997), using a "likelihood argument," found that the phenotypic difference discards sib-pair linkage information; and Fulker and Cherny (1996) came to a similar conclusion after an analysis of sib-pair covariances (Fulker et al. 1995). Here, I propose an extension of the Haseman-Elston (1972) method that puts the sib-trait sum into linkage testing.

Suppose a trait X has a normal distribution with a mean genetically determined and environmental (residual) variance σ_e^2; each sib pair has i alleles identical by descent (IBD) at the trait locus, $i = 0$, 1, or 2; and the sib pair–trait vector $X^T \equiv (X_1, X_2)^T$ has joint normal (binormal) distribution:

$$\mathcal{F}(X) = \frac{1}{2\pi\sqrt{|\Sigma_x|}} \exp\left[-\tfrac{1}{2}(X-\mu)^T\Sigma_x^{-1}(X-\mu)\right] ,$$

where μ is the overall mean and the symbol T stands for "transpose." The matrix Σ_x^{-1} is the inverse of the variance-covariance matrix, which has the form

$$\Sigma_x = \begin{pmatrix} v & c \\ c & v \end{pmatrix} ,$$

where

$$\begin{aligned} v = \mathrm{var}(X) &= V_p + V_c + \tfrac{1}{2}V_e + V_a + V_d \\ &= V_p + V_c + \tfrac{1}{2}V_e + V_g , \end{aligned}$$

$$\begin{aligned} c = \mathrm{cov}(X_1, X_2) &= \tfrac{1}{2}V_p + V_c + \tfrac{1}{2}iV_a + \tfrac{1}{2}i(i-1)V_d \\ &= \tfrac{1}{2}V_p + V_c + \tfrac{1}{2}iV_g + \tfrac{1}{2}i(i-2)V_d , \quad (1) \end{aligned}$$

and the variances are as follows: polygenic, V_p; common environment, V_c; additive genetic, V_a; dominance genetic, V_d; residual, V_e $(= 2\sigma_e^2)$; and total genetic, $V_g = V_a + V_d$ (Malécot 1966, p. 320; Amos 1994; Fulker and Cherny 1996).

Let us introduce two new variables: $D = X_1 - X_2$, and $S = X_1 + X_2$. By use of matrix algebra methods, it is easy to show that the variance-covariance matrix of D and S is diagonal:

$$\Sigma = \begin{bmatrix} 2(v-c) & 0 \\ 0 & 2(v+c) \end{bmatrix} ;$$

thus, these new "coordinates" are uncorrelated, each of them having the normal distribution, and their joint distribution is

$$\mathcal{F}(D, S) = \frac{1}{2\pi\sigma_D\sigma_S} \exp\left[-\frac{D^2}{2\sigma_D^2} - \frac{(S-2\mu)^2}{2\sigma_S^2}\right] .$$

The variances are $\sigma_D^2 = 2(v-c)$ and $\sigma_S^2 = 2(v+c)$. Instead of variances, let us consider the squared pair-trait difference $Y \equiv D^2$ and the squared pair sum $Z \equiv S^2$:

$$\begin{aligned} \mathcal{E}(Y|i) &= \sigma_D^2 \\ &= (V_e + V_p + 2V_g) - iV_g + i(2-i)V_d , \quad (2) \end{aligned}$$

and

$$\begin{aligned} \mathcal{E}(Z|i) &= \sigma_S^2 + 4\mu^2 \\ &= (V_e + 3V_p + 4V_c + 2V_g + 4\mu^2) + iV_g \\ &\quad - i(2-i)V_d , \quad (3) \end{aligned}$$

where the symbol $\mathcal{E}$ stands for "expectation." The squared pair-trait difference, Y, has been studied (Haseman and Elston 1972; Blackwelder and Elston 1982).

Each of the variables (2) and (3) is a function of the number of alleles IBD, i, at the trait locus, and their expected values, conditional on the marker information, are of interest:

$$\mathcal{E}(\ldots|M) = \sum_{i=0,1,2} \mathcal{E}(\ldots|i)P(i|M) \ , \tag{4}$$

where $f_i \equiv P(i|M)$ is the probability of i alleles IBD (i = 0, 1, or 2) at the trait locus. The expectations are

$$\begin{aligned}\mathcal{E}(Y|M) &= \sum_{i=0,1,2} [(V_\mathrm{e} + V_\mathrm{p} + 2V_\mathrm{g}) - iV_\mathrm{g} \\ &\quad + i(2-i)V_\mathrm{d}]f_\mathrm{i} \\ &= (V_\mathrm{e} + V_\mathrm{p} + 2V_\mathrm{g}) - 2V_\mathrm{g}(\pi + \tfrac{1}{2}) \\ &\quad + V_\mathrm{d}(\varphi + \tfrac{1}{2}) \\ &= (V_\mathrm{e} + V_\mathrm{p} + V_\mathrm{g} + \tfrac{1}{2}V_\mathrm{d}) - 2V_\mathrm{g}\pi + V_\mathrm{d}\varphi\end{aligned}$$

and

$$\begin{aligned}\mathcal{E}(Z|M) &= \sum_{i=0,1,2} [(V_\mathrm{e} + 3V_\mathrm{p} + 4V_\mathrm{c} + 2V_\mathrm{g} + 4\mu^2) \\ &\quad + iV_\mathrm{g} - i(2-i)V_\mathrm{d}]f_\mathrm{i} \\ &= (V_\mathrm{e} + 3V_\mathrm{p} + 4V_\mathrm{c} + 2V_\mathrm{g} + 4\mu^2) \\ &\quad + 2V_\mathrm{g}(\pi + \tfrac{1}{2}) - V_\mathrm{d}(\varphi + \tfrac{1}{2}) \\ &= (V_\mathrm{e} + 3V_\mathrm{p} + 4V_\mathrm{c} + 3V_\mathrm{g} - \tfrac{1}{2}V_\mathrm{d} + 4\mu^2) \\ &\quad + 2V_\mathrm{g}\pi - V_\mathrm{d}\varphi \ ,\end{aligned}$$

where

$$\pi \equiv \tfrac{1}{2}f_1 + f_2 - \tfrac{1}{2} \quad \text{and} \quad \varphi \equiv f_1 - \tfrac{1}{2} \tag{5}$$

at the trait locus. These definitions of π and φ differ from those introduced by Haseman and Elston (1972) and used by Blackwelder and Elston (1982) by the term $\frac{1}{2}$. So defined, π and φ are proportional to the same functions (5) of $\{f_i\}$, calculated at the marker locus (Drigalenko, in press):

$$\pi = \eta\pi_\mathrm{m} \ , \qquad \varphi = \eta^2\varphi_\mathrm{m} \ , \tag{6}$$

where π_m and φ_m are calculated on the basis of relatives' marker phenotypes, $\eta = (1-2r)^2$, and r is the recombination coefficient between the trait locus and the marker locus that depends on the (unknown) distance between them. Finally, the regression equations become

$$\begin{aligned}\mathcal{E}(Y|M) &= (V_\mathrm{e} + V_\mathrm{p} + V_\mathrm{g} + \tfrac{1}{2}V_\mathrm{d}) - 2V_\mathrm{g}\eta\pi_\mathrm{m} \\ &\quad + V_\mathrm{d}\eta^2\varphi_\mathrm{m} \\ &= \alpha_D - \beta\pi_\mathrm{m} + \gamma\varphi_\mathrm{m}\end{aligned} \tag{7}$$

and

$$\begin{aligned}-\mathcal{E}(Z|M) &= -(V_\mathrm{e} + 3V_\mathrm{p} + 4V_\mathrm{c} + 3V_\mathrm{g} \\ &\quad - \tfrac{1}{2}V_\mathrm{d} + 4\mu^2) - 2V_\mathrm{g}\eta\pi_\mathrm{m} + V_\mathrm{d}\eta^2\varphi_\mathrm{m} \\ &= \alpha_S - \beta\pi_\mathrm{m} + \gamma\varphi_\mathrm{m} \ ,\end{aligned} \tag{8}$$

where $\alpha_D \equiv V_\mathrm{e} + V_\mathrm{p} + V_\mathrm{g} + \frac{1}{2}V_\mathrm{d}$, $\alpha_S \equiv -(V_\mathrm{e} + 3V_\mathrm{p} + 4V_\mathrm{c} + 3V_\mathrm{g} - \frac{1}{2}V_\mathrm{d} + 4\mu^2)$, $\beta \equiv V_\mathrm{g}\eta$, and $\gamma \equiv V_\mathrm{d}\eta^2$. So, consideration of the squared pair sum of the trait values (taken with the opposite sign) results in a regression line that is parallel to that for the squared pair difference. Since seven parameters are unknown (V_e, V_p, V_c, V_g, V_d, μ, and η) and four regression coefficients are independent (α_D, α_S, β, and γ), all the parameters cannot be estimated. Note that only the slopes, β and γ, are important for testing linkage (Haseman and Elston 1972; Blackwelder and Elston 1982) and that these are the same for the sum and the difference of the sib pair–trait values.

The method described here uses all the information from the sib pair. To demonstrate the gain obtained when the sum and the difference are used together, let us ignore dominance, suppose that the residuals have the same variance in (6) and (7), and use Student's t-statistic to test the hypothesis H_0: $\beta = 0$. Then, joint use of the sum and the difference (rather than the difference alone) doubles the number of points on the regression line and, therefore, doubles the estimated values of both β and its variance, so that the t-statistic is enlarged by a factor of $\sim\sqrt{2}$, increasing the power of the test. Fulker and Cherny (1996, fig. 1) obtained similar results using simulated data and maximum-likelihood estimation.

More explicitly, for N sib pairs, indexed by j ($j = 1, \ldots, N$), the regression equations (7) and (8) include residuals ε_D and ε_S, assumed to be normally distributed and common for each sib pair (the dominance is ignored):

$$Y_j = \alpha_D - \beta\pi_j + \varepsilon_D \ , \quad -Z_j = \alpha_D - \beta\pi_j + \varepsilon_S \ . \tag{9}$$

These regression lines give the least-squares estimates of the slope:

$$\begin{aligned}\hat{\beta}_D &= \frac{N\Sigma Y_j\pi_j - \Sigma Y_j\Sigma\pi_j}{N\Sigma\pi_j^2 - (\Sigma\pi_j)^2} \ , \\ \hat{\beta}_S &= \frac{N\Sigma Z_j\pi_j - \Sigma Z_j\Sigma\pi_j}{N\Sigma\pi_j^2 - (\Sigma\pi_j)^2} \ .\end{aligned} \tag{10}$$

Under the assumption that the residuals have the same variance in (9), $\mathrm{var}(\varepsilon_D) = \mathrm{var}(\varepsilon_S)$, it is easy to prove that the least-squares estimate of the slope based on combined data for D and S (denoted by $D\oplus S$) is

$$\hat{\beta}_{D\oplus S} = \frac{N\Sigma[(Y_j - Z_j)/2]\pi_j - \Sigma[(Y_j - Z_j)/2]\Sigma\pi_j}{N\Sigma\pi_j^2 - (\Sigma\pi_j)^2}$$

$$= \tfrac{1}{2}(\hat{\beta}_D + \hat{\beta}_S)\ , \qquad (11)$$

that is, the "combined" regression line is exactly between the two individual lines. Owing to the properties of variances,

$$\mathrm{var}(\hat{\beta}_{D\oplus S}) = \mathrm{var}\left[\tfrac{1}{2}(\hat{\beta}_D + \hat{\beta}_S)\right]$$

$$= \tfrac{1}{4}[\mathrm{var}(\hat{\beta}_D) + \mathrm{var}(\hat{\beta}_S)]\ ,$$

because $\mathrm{cov}(\hat{\beta}_D, \hat{\beta}_S) = 0$, which is easy to see from (10) under the condition of $\mathrm{cov}(Y_j, Z_j) = 0$, discussed above. Hence, the estimate based on combined data for D and S has the smallest variance, that is, it is the most effective.

Note that, for every pair, (11) is based on the half-difference of Y and Z, which is $\frac{1}{2}(Y - Z) = \frac{1}{2}[(X_1 - X_2)^2 - (X_1 + X_2)^2] = -2X_1X_2$. The half-sum of (7) and (8) gives the equation

$$\mathcal{E}(-2X_1X_2 \mid M) = \tfrac{1}{2}(\alpha_D + \alpha_S) - \beta\pi_{\mathrm{m}} + \gamma\varphi_{\mathrm{m}}\ ,$$

which may be easily derived from (1) and (4). Thus, the most clear estimate, $\hat{\beta}_{D\oplus S}$, is based on the pair-trait multiplication, because the linkage test depends on the number of alleles IBD (which is a characteristic of a pair rather than an individual); the covariance (1) gives the same information as any combination of the squared pair-trait difference and the squared pair sum. This explains the effectiveness of the variance-components method (Amos 1994).

Acknowledgments

I thank Dr. Robert Elston for pointing out this issue and for his help in the interpretation of the results. This work was supported by research grant F05 TW05285 from the Fogarty International Center and resource grant P41 RR03655 from the National Center for Research Resources, National Institutes of Health.

EUGENE DRIGALENKO
Department of Epidemiology and Biostatistics
Rammelkamp Center for Education and Research
MetroHealth Campus
Case Western Reserve University
Cleveland

References

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

Blackwelder WC, Elston RC (1982) Power and robustness of sib-pair linkage tests and extension to larger sibships. Commun Stat Theor Methods 11:449–484

Drigalenko E. Matrix representation of the Haseman-Elston method. Theor Popul Biol (in press)

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

Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci, using sib pairs. Am J Hum Genet 56:1224–1233

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

Malécot G (1966) Probabilités et hérédité. Presses universitaires de France, Paris

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

Address for correspondence and reprints: Dr. Eugene Drigalenko, Department of Epidemiology and Biostatistics, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Room R258, Cleveland, OH 44109-1998. E-mail: dei@darwin.mhmc.cwru.edu

0002-9297/98/6304-0046$02.00