

A Multivariate Method for Detecting Genetic Linkage, with Application to a Pedigree with an Adverse Lipoprotein Phenotype

C. I. Amos,^{*,†,‡} R. C. Elston,[§] G. E. Bonney,[‡] B. J. B. Keats,[§] and G. S. Berenson^{||}

^{*}Family Studies Section, National Cancer Institute, Bethesda, MD; [†]International Agency for Research on Cancer, Lyon; [‡]Division of Biostatistics and Epidemiology, Howard University Cancer Center, Washington, DC; and Department of [§]Biometry and Genetics and ^{||}Department of Medicine, Louisiana State University Medical Center, New Orleans

Summary

The robust or model-free method for detecting linkage developed by Haseman and Elston for data from sib pairs is extended to incorporate observations of multiple traits on each individual. A method is proposed that estimates the linear function that results in the strongest correlation between the squared pair differences in the trait measurements and identity by descent at a marker locus. The method is illustrated by the study of apolipoprotein and cholesterol levels in individuals from a large family that had many members diagnosed with coronary heart disease.

Introduction

Pleiotropism has been observed for many genetic diseases in man. Often, only the consequences of altered gene expression on a variety of traits are observable, rather than the gene product itself. In this situation, the presence of a major gene might be more easily detected if observations about more than one trait are analyzed simultaneously. Similarly, methods for detecting linkage between a marker locus and a major gene locus may be more powerful if they use the information from each of several traits that are affected by the same major gene.

Although many major genes have pleiotropic expression, use of multiple observations for each individual in linkage analysis has been difficult. Multivariate admixture and segregation analyses (Elston et al. 1975; Namboodiri et al. 1975; Lange and Boehnke 1983) can be performed by assuming that the data follow some multivariate distribution and by directly finding the maximum-likelihood estimates of the unknown pa-

rameters. Alternatively, pedigree discriminant analysis (Goldin et al. 1978; Zlotnik et al. 1983) can be used. This technique seeks the linear function of several traits that maximizes the likelihood of the data, when it is assumed that segregation at a major gene locus pleiotropically affects these traits. Values of the linear function of the trait measures are assumed to have a known univariate distributional form (e.g., normal, conditional on major genotype). Both of these procedures will provide a function that can then be used to combine, or score, the multivariate observations into univariate measures that can be analyzed subsequently for linkage to a marker. Pedigree discriminant analysis requires fewer assumptions about the distribution of the multivariate data and is computationally faster because fewer parameters need to be estimated. Nevertheless, pedigree discriminant analysis can be cumbersome to use because the appropriate likelihood may have many local maxima; finding the function that gives the largest likelihood can, therefore, be difficult. Techniques for implementing both of these multivariate analysis methods have been based on the further assumption that only a single major gene is segregating in the data.

Regressive models for segregation analysis (Bonney 1984, 1986) have recently been extended to allow joint segregation and linkage analysis (Bonney et al. 1988). The methods currently allow analysis of bivariate data,

Received August 2, 1989; revision received April 16, 1990.

Address for correspondence and reprints: C. I. Amos, Family Studies Section, National Cancer Institute, Suite 439, 6130 Executive Boulevard, North Bethesda, MD 20892.

© 1990 by The American Society of Human Genetics. All rights reserved.
0002-9297/90/4702-0009\$02.00

and possible linkage to a marker. Although these methods show great promise for resolving the multivariate effects of a single locus, data analysis methods in which individuals have been measured for more than two characteristics are currently unavailable. In addition, these methods require either explicit or implicit assumptions about the underlying genetic model.

As an alternative approach, we show here how the robust method of linkage analysis developed by Haseman and Elston (1972) can be extended to find a linear function having the strongest correlation to marker identity by descent among sib pairs. The linear function so derived can then be further analyzed as a univariate trait. Advantages of the methodology we develop include computational rapidity and the possibility of detecting linkage for traits with oligogenic inheritance, through the simple extension from linear to multiple linear regression (Bailey-Wilson et al. 1989). We first describe the Haseman-Elston method and then show how the approach can be generalized for multivariate data. Finally, we illustrate this methodology by applying it to a study of lipid metabolism in a single large pedigree.

The Haseman-Elston Method

Let X_{ij} , the value of a trait measured on the i th sib ($i = 1, 2$) for the j th noninbred pair of sibs, be composed of an overall mean, μ , a major genetic effect, g_{ij} , and an independent random effect, e_{ij} , that is normally distributed with zero mean. Assume that the major gene effect is due to a locus with two alleles which we shall label A and a. We shall consider the model $X_{ij} = \mu + g_{ij} + e_{ij}$, where

$$g_{ij} = \begin{cases} a & \text{if the } ij\text{th individual has genotype AA} \\ d & \text{if the } ij\text{th individual has genotype Aa} \\ -a & \text{if the } ij\text{th individual has genotype aa} \end{cases}$$

Let π_{mj} denote the proportion of genes (0, 1/2, or 1) that the j th sib pair shares identical by descent (ibd) at a marker locus, let I_{mj} denote the information available about ibd at the marker locus, and let f_{m1j} denote the probability that the sib pair shares one gene ibd. Let $Y_j = (X_{1j} - X_{2j})^2$, where X_{1j} and X_{2j} are the trait values for the two sibs composing the j th pair, and let the variance of the difference in residuals for the pair be σ^2 . Haseman and Elston (1972) showed that as long as there is no major gene by residual (environment) interaction, we have

$$E(Y_j | I_{mj}) = \alpha + \beta \hat{\pi}_{mj} + \gamma \hat{f}_{m1j}, \quad (1)$$

where $\hat{\pi}_{mj}$ and $\hat{f}_{m1j}$ are, respectively, estimates of π_{mj} and f_{m1j} . The coefficient γ was shown, by Blackwelder (1977) using simulation and by Amos et al. (1989) theoretically, to be generally negligible. Letting θ denote the recombination fraction between the trait and marker loci, the coefficient β is negative if $\theta < .5$ and the additive genetic variance is greater than zero ($a \neq 0$). Blackwelder and Elston (1982) showed that for sib pairs the ratio of the estimate of β to its standard error is asymptotically distributed as a standard normal variate, so that a one-sided test of linkage can be obtained by comparing this statistic with the appropriate t distribution. Haseman (1970) showed that for oligogenically influenced traits a valid test for linkage can be obtained by testing for a negative regression of the squared pair differences separately on to the proportion of genes ibd at each of several marker loci. Blackwelder and Elston (1982) showed that neither departures of the residual variation from the specified normal form nor nonindependence of the sib pairs greatly alters the properties of the test statistic. The interpretation of equation (1) is that the variability between a pair of sibs can be linearly modeled as a function of the genetic component of variance and the recombination fraction between the trait and marker loci.

The Multivariate Model

We now consider an extension of the Haseman-Elston approach for multivariate traits. Assume that p traits are studied on each of n sib pairs. Let $X_{ijk} = \mu_k + g_{ijk} + e_{ijk}$, where X_{ijk} is an observation on the i th ($i = 1, 2$) individual of the j th pair ($j = 1, 2, \dots, n$), for the k th trait ($k = 1, 2, \dots, p$). Assume that individuals may have one of three genotypes—AA, Aa, and aa—with effects:

$$g_{ijk} = \begin{cases} a_k & \text{if the } ij\text{th individual has genotype AA} \\ d_k & \text{if the } ij\text{th individual has genotype Aa} \\ -a_k & \text{if the } ij\text{th individual has genotype aa} \end{cases}$$

The covariances among the traits need not be zero. We shall assume that the pairs are independent (or at least that $\text{Cov}(e_{ijk}, e_{tuw}) = 0$ for all i, k, t, w , when $j \neq u$).

We wish to determine the linear function that maximizes the correlation between the squared sib-pair difference and $\hat{\pi}_{mj}$. That is, assuming a linear model of the form

$$[c_1(X_{1j1} - X_{2j1}) + c_2(X_{2j1} - X_{2j2}) + \dots + c_p(X_{1jp} - X_{2jp})]^2 = \alpha + \beta \hat{\pi}_{mj} + \epsilon_j, \quad (2)$$

we must find the values of the coefficients c_k that maximize the relationship as measured by the significance of the estimate β . Each of the trait pair differences has an expected value of zero, and by the law of large numbers, in large samples, the mean of the trait pair differences is zero. Thus, in large samples, the function obtained is the linear function that maximizes the linear regression of the variability between pairs of sibs on identity by descent at a marker locus. Assuming without loss of generality that $E(\epsilon_j) = 0$, expanding the left hand side of equation (2), and taking expectations of both sides, we have

$$E\left[\sum_k^p c_k^2 (X_{1jk} - X_{2jk})^2\right] + E\left[\sum_h^p \sum_{k < h}^{p-1} 2c_h c_k (X_{1jk} - X_{2jk})(X_{1jh} - X_{2jh})\right] = \alpha + \beta \hat{\pi}_{mj}. \quad (3)$$

We may now apply standard multivariate regression analysis methods (Anderson 1984; Seber 1988). To calculate the test statistic, let Y denote the $n \times (p^2 + p)/2$ matrix of all squares and cross products among traits of the pair differences. Let $\hat{\pi}$ denote the $n \times 1$ vector of estimates of the proportion of genes ibd at the marker locus for the n sib pairs. Let S_{11} be an estimate of the covariance matrix of rows of Y , let S_{12} with transpose S_{21} be an estimate of the vector of covariances between the rows of Y and $\hat{\pi}$; and let S_{22} be an estimate of the variance of $\hat{\pi}$. In other words, S_{11} , S_{12} , S_{21} , and S_{22} are the elements of the covariance matrix of the rows of $[Y; \pi]$:

$$\text{Cov}(Y, \hat{\pi}) = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

Letting $\mathbf{1}$ be an $n \times 1$ vector of 1's, letting $\bar{\hat{\pi}}$ be the mean of $\hat{\pi}$, and letting $\bar{Y}$ be a $1 \times (p^2 + p)/2$ mean vector (over j) of $(X_{1jk} - X_{2jk})(X_{1jh} - X_{2jh})$, one finds that the formulas for S_{11} , S_{12} , S_{21} , and S_{22} are

$$\begin{aligned} S_{11} &= \frac{1}{n-1}[(Y - \bar{Y})(Y - \bar{Y})'] \\ S_{12} &= \frac{1}{n-1}[(Y - \bar{Y})(\hat{\pi} - \bar{\hat{\pi}})'], \quad S_{21} = S_{12}' \\ S_{22} &= \frac{1}{n-1}[(\hat{\pi} - \bar{\hat{\pi}})(\hat{\pi} - \bar{\hat{\pi}})'] = \text{Var}(\hat{\pi}). \end{aligned}$$

Finally, let $Q_b = S_{12}S_{22}^{-1}S_{21}$ be the hypothesis sum-

of-squares matrix and let $Q_e = S_{11}^{-1} - S_{12}S_{22}^{-1}S_{21}$ be the error sum-of-squares matrix for testing the null hypothesis that $\beta = 0$. By employing the arrangement of the data given by equation (3), the model specified by equation (2) is the usual multivariate regression problem, except that we have imposed some additional constraints. We may obtain the linear function that maximizes the regression of the multivariate data on the estimates of the proportion of genes ibd by numerically finding the linear function d that maximizes the ratio:

$$r = \max_d \frac{d'Q_b d}{d'Q_e d}, \quad (4)$$

given the arbitrary restriction that $d'd = 1$, and with the additional constraints that (1), for the first p elements of d , $d_i = c_i^2$ (i.e., $d_i \geq 0$, for $1 \leq i \leq p$) and (2), for the last $p(p-1)/2$ elements of d , $d_i = 2c_k c_h$ [$p < i < p(p+1)/2$], where k and h are indices for the cross products among traits of the pair differences of equation (3). In the Appendix we show that if the first p constraints are relaxed, then, asymptotically, $(n-p-1)/p$ times the value of this ratio of quadratic forms is distributed as an F distribution having p and $n-p-1$ df. If the first p constraints are kept in the model, then the critical value obtained by comparison with an F distribution with the appropriate df must be conservative. Thus, in large enough samples, a conservative test of no significant regression of a linear function on the proportion of genes ibd can be obtained from the F statistic obtained from equation (4). Because we are only concerned with the one-sided test of no negative regression as expressed by equation (3), the F statistic obtained from equation (4) should be compared with the critical value from an F distribution with the appropriate df by using half the usual nominal value.

The value of the canonical correlation between the dependent observations and the proportion of genes that the sib pairs share ibd can be derived from the statistic r given by equation (4). According to the method of Seber (1988, p. 526 ff), r is the solution of the eigenvalue problem $(Q_e^{-1}Q_b)d = rd$. As shown by Anderson (1984), the canonical correlation, R , can be obtained from the eigenvalue r by the transformation $R = \sqrt{r/(1+r)}$.

Application to a Large Pedigree

Rosenbaum et al. (1986) described a large pedigree that included members with an adverse lipoprotein

phenotype which predisposed individuals to coronary artery disease. The pedigree, consisting of 200 sampled individuals, was ascertained through a child, seen as a participant in the Bogalusa Heart Study, whose mother was concerned by a high prevalence of heart disease on both sides of the child's family. Of the examined individuals, six were survivors of one or more myocardial infarctions (MI) and an additional 15 had heart disease. Previous analyses (Amos et al. 1987), using log-transformed variates, showed evidence for segregation of major gene(s) affecting levels of high-density lipoprotein cholesterol (HDL-C), apolipoprotein B (apo B), the residuals from the regression of HDL-C on apolipoprotein A1 (apo A1), and residuals from the regression of low-density lipoprotein cholesterol (LDL-C) on apo B. Pedigree discriminant analysis (Amos et al. 1986) of HDL-C and LDL-C levels and other environmental covariates suggested that a major gene affecting the ratio of HDL-C to LDL-C segregates in this pedigree. Analysis using maximum-likelihood linkage methods (Amos et al. 1987) showed evidence for linkage, with a lod score of 1.72, between haptoglobin and a linear function that was obtained from the coefficients of the regression of log HDL-C on log A1 and that was essentially the ratio of HDL-C to apo A1 (on the untransformed scale).

In the current study we apply the methodology described above to the marker-locus haptoglobin and to the traits HDL-C, LDL-C, apo A1, and apo B levels. The values of these four variables were first log_e transformed, adjusted statistically for effects attributable to age, sex, and cigarette, alcohol, and oral contraceptive use (as indicated by Amos et al. (1987)), and standardized by dividing by their SD prior to multivariate analysis. The linear function maximizing the correlation with identity-by-descent sharing for haptoglobin was obtained for each of the possible combinations of two, three, and four variables. The values of the *F* statistic formed from the ratio of the hypothesis sum of squares to the error sums of squares was used to select the best linear function of two, three, and four variables.

Results from Application to an Extended Pedigree

After log_e transformation and covariate adjustment, the values of the SDs were .159 for apo A1, .250 for apo B, .357 for HDL-C, and .255 for LDL-C. (These SDs are slightly different from those given by Amos et al. [1987] because the values for the six individuals who had suffered an MI were included in the calculation of

the SDs in the present analysis but were not included in the former analysis.) Results from all multivariate analyses refer to the standardized variates. For the purpose of using the robust linkage methods, the pedigree provided 108 sib pairs on whom data were available for both the marker and traits.

The linear functions, their canonical correlations, and the values of the *F* statistics from the univariate and multivariable regressions on ibd for haptoglobin are presented in table 1. Of the two-variable functions, the linear function of HDL-C and apo A1 had the strongest relationship with ibd at the haptoglobin locus. Since apo A1 and HDL-C values were log transformed prior to analysis, values of this linear function indicate that approximately the log of the ratio of apo A1 to HDL-C maximizes the relationship with ibd for haptoglobin. Of the functions incorporating three variables, a linear function of apo A1, HDL-C, and LDL-C was found to have the strongest relationship with ibd for haptoglobin. Exponentiation of this linear function indicates that approximately the square root of LDL-C times the ratio of apo A1 to HDL-C (i.e., $[\text{HDL-C}/\text{apo A1}] \times \sqrt{\text{LDL-C}}$) maximizes the relationship to ibd for haptoglobin. Inclusion of apo B values into the function had little impact on the strength of the relationship, as seen by the small coefficient for apo B in the last line of table 1 and by the lack of an increase of the canonical correlation when apo B levels were included in the analysis. For each of the derived variables, individuals with high values are at increased risk for cardiovascular disease, as they would have lower HDL-C levels or less cholesterol per HDL particle. For the linear function incorporating LDL-C levels, individuals with high measures also tend to have higher LDL-C levels, putting them at further increased risk for cardiovascular disease. Segregation analysis of the three-variable function (results not presented) showed evidence for a recessive major gene having a gene frequency of .166, which results in high levels of this linear function.

Discussion

We have developed a method for combining and using multiple observations of related traits for each individual in a study. This methodology has been developed with particular emphasis on quantitative traits. The method could also be applied to the study of qualitative traits or to mixtures of qualitative and quantitative traits, provided that the resulting linear function is approximately continuously distributed. Because the method

Table 1

Results for the Regression on ibd for Haptoglobin of Standardized Log_e HDL-C, LDL-C, Apo A1, and Apo B Levels and of Linear Functions of These Variables Derived from Multivariate Analysis

No. of Variables and Linear Function	R ^a	F Statistic	df
1:			
Apo A1	.126	1.70	1,106
Apo B	.000	.00	1,106
HDL-C	.224	5.62	1,106
LDL-C	.003	.11	1,106
2:			
(.798)Apo A1 + (.464)Apo B	.168	1.53	2,105
(.659)Apo A1 - (.618)HDL-C	.296	5.04	2,105
(.877)Apo A1 + (.357)LDL-C	.140	1.05	2,105
(.161)Apo B - (.974)HDL-C	.228	2.89	2,105
(.301)Apo B - (.911)LDL-C	.003	.05	2,105
(.897)HDL-C - (.326)LDL-C	.232	2.99	2,105
3:			
(.678)Apo A1 + (.252)Apo B - (.530)HDL-C	.312	3.77	3,104
(.822)Apo A1 + (.433)Apo B + (.019)LDL-C	.168	1.01	3,104
(.621)Apo A1 - (.528)HDL-C + (.351)LDL-C	.327	4.16	3,104
(.161)Apo B - (.974)HDL-C + (.016)LDL-C	.228	1.91	3,104
4:			
(.579)Apo A1 - (.024)Apo B - (.504)HDL-C + (.343)LDL-C	.327	3.09	4,103

^a Estimate of canonical correlation.

is computationally rapid, selection procedures can be easily implemented for choosing the best variables for incorporation into the linear function. In the example, we studied all possible combinations because we were working with relatively few variables. For studies involving more trait observations per individual, a forward selection procedure could be used in which variables are selected on the basis of the greatest change in the *R* statistic. Alternatively, a step-down procedure in which variables are sequentially removed from a full model incorporating all of the traits can be implemented. The advantage of this type of procedure is that it allows for the multiple testing problem inherent in testing all possible combinations of variables, by first assessing evidence that *any* linear function of the dependent variates is correlated with identity-by-descent sharing at the marker locus. In the current example, the overall test for a significant correlation between identity-by-descent sharing for haptoglobin and the set of all possible linear functions of the four traits has an *F* value (3.09) that, compared with the *F* distribution with four and 103 df, is significant at the .01 level (we use half the nominal level, since we are considering a one-sided test of the regression coefficient). In our

example, incorporation of apo B levels had no effect on the canonical correlation. Similarly, in other studies, these methods could be used to exclude variables having little or no impact on the resolution of linkage.

In the current study, on account of rather small sample size we should be cautious in assigning significance levels to the values of the *F* statistic that were obtained from the analyses. One can note, however, that values of the test statistic for the two- and three-variable functions are nominally significant at the .005 significance level, while the most significant univariate *F* test is nominally significant at the .01 level, suggesting that these multivariate functions are more strongly correlated to identity-by-descent sharing than is any single variable. In the univariate case, the test is identical to a *t* test, and Blackwelder (1977) found that critical values from the Student's *t* distribution were approximately valid for testing the hypothesis of no significant regression in this situation in samples as small as 50 sib pairs. We can generate the multivariate test from the univariate test by application of the union-intersection principle; thus, the convergence of the test statistic to an *F* distribution should also be rapid. Future simulation studies will be helpful in assessing what sample sizes

are needed in order to be assured that critical values from an F distribution with the appropriate df are not anticonservative.

Although other linkage analysis techniques such as the maximum-likelihood (e.g., LOD) method can be applied to the scores obtained from the linear function derived by application of this extension of the Haseman-Elston approach, the significance levels from such an approach would not be the same as those obtained from using a single univariate trait. Use of LOD scores in this setting would be appropriate for describing the data or to suggest evidence for linkage between the traits and the marker, but the critical values of the significance test would need to be adjusted. In any case, some type of adjustment should be made to the critical values when testing multiple traits with a single marker. Use of a cross-validation approach (Goldin et al. 1978) would assure that univariate significance levels are valid. To apply cross validation, the data are separated into two parts, a linear function is derived in one part, and then the data of the other part are scored with this function for subsequent maximum-likelihood segregation and linkage analysis. Cross validation assures that the critical values from performing a single univariate test are valid in the multivariate setting.

Robust methods for linkage analysis using multivariate data could also be formulated from an analysis-of-variance approach (Lowry and Schultz 1959; Jayakar 1970; Hill 1975; Smith 1975), but these methods use identity-by-state sharing and would have low power unless highly polymorphic markers are studied (Risch 1990). Chevalet et al. (1984) proposed a method for linkage analysis based on the covariance among relatives, conditional on their identity-by-descent sharing, a method which could be extended for studying multivariate traits. In the univariate case, his method is computationally more demanding than the Haseman-Elston method, as it requires obtaining maximum-likelihood estimates.

Results from this study are consistent with the hypothesis that, in this family, a locus near haptoglobin affects lipoprotein levels. Individuals with high values for this trait would have poor reverse transport of cholesterol from tissues to the liver, since they would have cholesterol-poor HDL particles and high levels of LDL-C. This finding has interest in light of the linkage between haptoglobin and a major gene resulting in lecithin:cholesterol acyltransferase (LCAT) deficiency (Teisberg et al. 1975; Albers and Uttermann 1981). Because persons with LCAT deficiency cannot transfer cholesterol from tissues to the HDL particle, they have protein-

rich, cholesterol-poor HDL particles (Glomset et al. 1983). Evidence has been presented that the cholesterol ester transfer protein (CETP) is located very close to haptoglobin (Lusis et al. 1987) on chromosome 16. Variability in expression of CETP would lead to altered levels of HDL-C (Faust et al. 1989). There is a need for further analyses of other large pedigrees and, for this pedigree, of the LCAT gene and other genes located near haptoglobin.

There are some limitations in the data that were analyzed. Although haptoglobin was informative in this pedigree, for relatively few of the sib pairs could complete information on identity-by-descent sharing be derived. Inclusion of information on alleles of haptoglobin that are determined by RFLP analysis (Kidd et al. 1986) would have improved our ability to discern identity-by-descent sharing at this locus. Data on many of the parental genotypes were not available, further restricting our ability to resolve identity-by-descent sharing among the sib pairs.

The method we have developed in the present study estimates a linear function of multivariate-trait observations having the strongest relationship to identity by descent at a marker locus. This method is computationally rapid and provides an objective method for finding a linear function of multivariate data for subsequent segregation analysis. It does not use all of the available information, because only data from sib pairs are analyzed. For pedigree data, as in the example that was studied, maximum-likelihood segregation and linkage methods could be directly applied. However, this requires simultaneously estimating a linear function of the trait values, a recombination fraction, and the segregation parameters. Although such an approach may be possible for bivariate data, it is not likely to be practical when several traits are considered, because of numerical difficulties. An approach that should avoid some difficulties in optimization but would use all available pedigree data can be developed using a three-step process. First, using all of the sib pairs, one would apply our extension of the Haseman-Elston method for multivariate data. The linear function so obtained could then be used to estimate trait-related parameters in segregation analysis using all of the pedigree data. Finally, these segregation analysis parameters could be used in maximum-likelihood linkage analysis to obtain a linear function maximizing linkage between the traits and the marker. Although this approach would allow use of all of the data, it would lose the robust properties of the Haseman-Elston approach, and so the estimates would again become influenced by departures from the

underlying genetic assumptions. For instance, it is not clear what linear function would result from such an analysis if the traits are affected by multiple, unlinked loci. Also, as has been noted above, a LOD score of 3 would not define the same critical region in the multivariate context as it does in the univariate case. The methodology we have developed provides a means of using multivariate data to obtain (1) a test of no relationship between a marker and set of multivariate traits and (2) a linear function maximizing the relationship between variability among sibs and identity by descent at the marker locus.

Acknowledgments

The authors thank Alexander F. Wilson for his helpful comments. This work was supported in part by U.S. Public Health research grants CA14718 from the National Cancer Institute and GM39573 and GM28356 from the National Institute of General Medical Sciences and by Division of Research Resources resource grant RR03655.

Appendix

In this appendix we develop a test for significance of the multivariate regression of the squared linear function of trait differences on the proportion of genes ibd given in equation (2). Assume the linear model expressed by equation (2), which we express in matrix notation as

$$[c'X_j]^2 = \alpha + \pi_j\beta + \varepsilon_j, \quad (A1)$$

where c' is a $1 \times p$ vector of constants and where X_j represents the p sib-pair differences for the j th sib pair, i.e., $X_j = [(X_{1j1} - X_{2j1}), (X_{1j2} - X_{2j2}), \dots, (X_{1jp} - X_{2jp})]$. We may assume without loss of generality that $\Sigma X_j = 0$. If this is not the case, we may linearly transform the model given by equation (A1) to satisfy this condition by subtraction of a vector of means from X_j ; this affects only the intercept term of the model.

We now transform the left-hand side of equation (A1) so that we work only with uncorrelated variables. As shown by Anderson (1986) and others, there exists an orthonormal linear transformation A (of dimensions $p \times p$ in this case) such that the columns of the product $A'X_j$ are orthogonal. Since A is orthonormal we have $AA' = I$. We now transform the left-hand side of the model given by equation (A1), using the transformation A to obtain $[c'AA'X_j]^2 = \alpha + \beta \pi_j + \varepsilon_j$. Letting $w' = c'A$ and $U_j = A'X_j$, we may rewrite this expression as

$$[w' U_j]^2 = \alpha + \pi_j\beta + \varepsilon_j. \quad (A2)$$

In the terms of equation (2), we have

$$[w_1(U_{1j1} - U_{2j1}) + w_2(U_{1j2} - U_{2j2}) + \dots + w_p(U_{1jp} - U_{2jp})]^2 = \alpha + \pi_j\beta + \varepsilon_j. \quad (A3)$$

Expanding equation (A3) and taking expectations as in equation (3), we have the model

$$E\left[\sum_{k=1}^p w_k^2 (U_{1jk} - U_{2jk})^2\right] = \alpha + \pi_j\beta.$$

The sum of the cross products are all zero because the columns of $A'X_j$ are orthogonal. For the j th pair, we shall denote the vector of squared pair differences of the U 's as U_j^* .

Now, as long as we allow estimates of the coefficients w_k^2 to be negative and make the further assumption that U_j^* has a multivariate normal distribution, a statistic for testing that β is different from zero can be obtained by standard application of the union-intersection approach (e.g., see Seber 1988). If we let ω be any $p \times 1$ vector, a test for no correlation between dependent and independent variates of the model $\omega'U_j^* = \alpha + \beta\pi_j$ can be obtained. If we define Q_b and Q_e to be as Q_b and Q_e in the text but substitute U^* in place of Y , then, for a given ω , $(n-2)\omega'Q_b^*\omega/\omega'Q_e^*\omega$ is distributed as an F with 1 and $n-2$ df. If we apply an arbitrary constraint that $\omega'Q_e^*\omega = 1$ and then maximize over ω , a union-intersection test statistic can be obtained. The test statistic $[p/(n-p-1)] \max_{\omega} [\omega'Q_b^*\omega/\omega'Q_e^*\omega]$ is distributed as an $F_{p, n-p-1}$ variate (Seber 1988). By design of the model, the U_j do not, however, follow a multivariate normal distribution. However, in large enough samples, the robustness that the F statistic has against nonnormality, when one tests the hypothesis of no significant regression, applies.

Use of constraints to require that each element of ω be positive can only result in the same or a smaller value of the ratio of the hypothesis sum of squares to the error sums of squares. Thus, for testing the null hypothesis of no relationship between the dependent traits and identity-by-descent sharing for a marker, the use of an F distribution to form the critical region of the statistic is conservative. In summary, in large enough samples, the test statistic given by equation (4) can be compared to an F distribution with the appropriate df. The resulting test of significance is, at worst, conservative whenever the constraints implied by the model given by equation (2) are employed.

Finally, it should be noted that in practice it is a simple matter to transform X_j to U_j and then obtain the maximizing coefficient vector ω as the characteristic vector corresponding to the largest characteristic root

of $Q_e^{-1}Q_h^*$; if all the elements of this vector are positive, then we immediately have real solutions to $w_k^2 = \omega_k$, and the resulting $\pm w_k$ can be used to obtain the desired c_k . It is, however, necessary to determine, for each coefficient, whether $+w_k$ or $-w_k$ leads to a maximum.

References

- Albers JJ, Utterman G (1981) Genetic control of lecithin:cholesterol acyltransferase: measurement of LCAT mass in a large kindred with LCAT deficiency. *Am J Hum Genet* 33:702–718
- Amos CI, Elston RC, Srinivasan SR, Wilson AF, Cresanta JL, Ward LJ, Berenson GS (1987) Linkage and segregation analyses of apolipoproteins A1 and B, and lipoprotein cholesterol levels in a large pedigree with excess coronary heart disease: the Bogalusa Heart Study. *Genet Epidemiol* 4:115–128
- Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust test for linkage. *Genet Epidemiol* 6:435–450
- Amos CI, Wilson AF, Rosenbaum PA, Srinivasan SR, Weber LS, Elston RC, Berenson GS (1986) An approach to the multivariate analysis of high-density-lipoprotein cholesterol in a large kindred: the Bogalusa Heart Study. *Genet Epidemiol* 3:255–268
- Anderson TW (1984) An introduction to multivariate statistical analysis, 2d ed. John Wiley & Sons, New York, pp 480–498
- Bailey-Wilson JE, Elston RC, Wilson AF, Amos CI (1989) A comparison of some sib-pair linkage methods and multiple locus extensions. *Prog Clin Biol Res* 157:129–134
- Blackwelder WC (1977) Statistical methods for detecting genetic linkage from sibship data. Ph.D. thesis, University of North Carolina, Chapel Hill. Institute of Statistics Mimeo series 1114. Chapel Hill, NC
- Blackwelder WC, Elston RC (1982) Power and robustness of sib-pair linkage tests and extension to larger sibships. *Commun Stat Theory Methods* 11:449–484
- Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models. *Am J Med Genet* 18:731–749
- (1986) Regressive logistic models for familial disease and other binary traits. *Biometrics* 42:611–625
- Bonney GE, Lathrop GM, Lalouel J-M (1988) Combined linkage and segregation analysis using regressive models. *Am J Hum Genet* 43:29–37
- Chevalet C, Gillois M, Khang VT (1984) Conditional probabilities of identity of genes at a locus linked to a marker. *Genet Sel Evol* 16:431–444
- Elston RC, Namboodiri KK, Glueck CJ, Fallat R, Tsang R, Leuba V (1975) Study of the genetic transmission of hypercholesterolemia and hypertriglyceridemia in a 195 member kindred. *Ann Hum Genet* 39:67–87
- Faust RA, Cheung MC, Albers JJ (1989) Secretion of cholesterol ester transfer protein-lipoprotein complexes by human HepG2 hepatocytes. *Atherosclerosis* 77:77–82
- Glomset JA, Norum KR, Gjone E (1983) Familial lecithin:cholesterol acyltransferase deficiency. In: Stanbury JB, Wynngaarden JB, Goldstein JL, Brown MS (eds) *The metabolic basis of inherited disease*, 5th ed. McGraw Hill, New York, pp 643–654
- Goldin LR, Elston RC, Graham JB, Miller CH (1978) Genetic analysis of vonWillebrand's disease: a multivariate approach. *Am J Med Genet* 6:279–293
- Haseman JK (1970) The genetic analysis of quantitative traits using twin and sib data. Ph.D. thesis, University of North Carolina, Chapel Hill
- Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 2:3–19
- Hill AP (1975) Quantitative linkage: a statistical procedure for its detection and estimation. *Ann Hum Genet* 38: 439–449
- Jayakar SD (1970) On the detection and estimation of linkage between a locus influencing a quantitative character and a marker locus. *Biometrics* 26:451–464
- Kidd KK, Kidd JR, Castiglione CM, Pakstis AJ, Sparkes RS (1986) Progress toward resolving the possible linkage of multiple endocrine neoplasia type 2A to haptoglobin and group specific component: use of restriction fragment polymorphisms extends the exclusion region. *Genet Epidemiol* 3:195–200
- Lange K, Boehnke M (1983) Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. *Am J Med Genet* 14:513–524
- Lowry DC, Schultz FT (1959) Testing association of metric traits and marker genes. *Ann Hum Genet* 23:83–90
- Lusis AJ, Zollman RS, Sparkes RS, Mohandas T, Klisak I, Drayna D, Lawn RM (1987) Assignment of the human gene for cholesteryl ester transfer protein to chromosome 16q12-21. *Genomics* 1:232–235
- Namboodiri KK, Elston RC, Glueck CJ, Fallat R, Buncher CP, Tsang R (1975) Bivariate analysis of cholesterol and triglyceride levels in families in which probands have type IIb lipoprotein phenotype. *Am J Hum Genet* 27:454–471
- Risch N (1990) Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* 46:242–253
- Rosenbaum PA, Amos CI, Shear CL, Elston RC, Sellers TA, Srinivasan SR, Berenson BS (1986) Description of a large pedigree with an adverse lipoprotein cholesterol phenotype: the Bogalusa Heart Study. *Genet Epidemiol* 3:241–254
- Seber GAF (1988) *Multivariate observations*. Wiley & Sons, New York
- Smith CAB (1975) A non-parametric test for linkage with a quantitative character. *Ann Hum Genet* 38:450–460
- Teisberg P, Gjone E, Olaisen B (1975) Genetics of LCAT (lecithin:cholesterol acyltransferase) deficiency. *Ann Hum Genet* 38:327–332
- Zlotnik LH, Elston RC, Namboodiri KK (1983) Pedigree discriminant analysis: a method to identify monogenic segregation. *Am J Med Genet* 15:307–313