

Combined Linkage and Segregation Analysis Using Regressive Models

George Ebow Bonney,* G. Mark Lathrop,† and Jean-Marc Lalouel†

*Division of Biostatistics, Howard University Cancer Center, Washington, DC; and †Howard Hughes Medical Institute, University of Utah Medical Center, Salt Lake City

Summary

Regressive models for segregation analysis have been extended to include multivariate data and linked marker loci. The new models have been applied to data from two pedigrees segregating a gene for cardiovascular disease.

Introduction

The growing density of polymorphic marker loci on the map of the human genome is greatly increasing the power of linkage studies for the localization and identification of genes responsible for many human diseases. In addition to its usefulness for mapping single-gene disorders, linkage analysis can identify major genetic contributions to multifactorial diseases and other complex phenotypes. However, linkage analysis has usually been conducted under the assumption that the underlying major loci completely account for all correlations between relatives.

Segregation analysis, also, is designed to detect major genes underlying complex phenotypes; the modern methods account simultaneously for residual familial correlations due to polygenic and common environmental effects (Elston and Stewart 1971; Morton and MacLean 1974; Lalouel et al. 1983). In segregation analysis one commonly estimates components of the phenotypic variance under the assumption that the total variance can be partitioned to reflect effects of a major gene, of additive polygenic factors, of familial environment, and random envi-

ronmental factors (the mixed model; for details see Morton and MacLean 1974; Lalouel et al. 1983). The incorporation of residual correlations in linkage analysis will be useful to evaluate the contribution of major genes when other sources of familial correlation could be present, as they are in various multifactorial diseases.

Recently Bonney (1984, 1986) introduced regressive models as an alternative to the mixed model in segregation analysis. Regressive models account for family dependence by specifying (1) a regression relationship between a person's phenotype and genotype, (2) the phenotypes of parents and older brothers or sisters, and (3) other explanatory variables. This approach is intuitively appealing because it provides for a major gene and for complex patterns of familial correlation, including spouse-spouse and parent-offspring phenotypic correlations, without postulating an explicit genetic and environmental model to explain the residual variance. Using regressive models, one can control for important epidemiological variables by directly incorporating regressors into the segregation analysis. Computations under the regressive model are also simpler than those under the mixed model.

In this paper we extend the regressive models to incorporate linked marker loci, and we increase the scope of the models to include multivariate phenotypes and interactions between genes at different loci, either linked or unlinked. We describe applications of the extended model to the combined linkage and

Received December 7, 1987; revision received February 2, 1988.

Address for correspondence and reprints: G. M. Lathrop, Ph.D., Howard Hughes Medical Institute, 603 Wintrobe Building, University of Utah Medical Center, Salt Lake City, UT 84132.

© 1988 by The American Society of Human Genetics. All rights reserved. 0002-9297/88/4301-0005\$02.00

segregation analysis of two coronary heart disease pedigrees.

Multivariate Traits and Linked Marker Loci

As stated above, regressive models (Bonney 1984, 1986) account for patterns of dependence in family data by specifying a regression relationship between a person's phenotype (Y), the phenotypes of ancestors and older relatives (Y_A), the genotype (g), and other explanatory variables (X). In essence the classical penetrance function is generalized to include ancestral phenotypes. The generalized penetrance function is the conditional probability of Y given g , Y_A , and X , i.e.,

$$\Pr(Y|g, Y_A, X) . \quad (1)$$

Including Y_A in the regression allows for unspecified (or residual) factors, such as spouse correlations, effects of unspecified genes, and cultural and other environmental effects. Incorporation of transmission probabilities to test for compatibility with Mendelian inheritance (Elston and Stewart 1971; Lalouel et al. 1983) is straightforward. Moreover, we have identified broad patterns of dependence that allow Y_A to include only a few relatives, which simplifies the calculations greatly. For example, in the class A pattern of dependence parents alone account for residual correlations among sibs, so that Y_A includes only the phenotypes of mother (Y_M), father (Y_F), and spouse (Y_S) if spouse correlations are not zero.

In the following discussion, the generalization of the regressive model to a multivariate trait Y with an arbitrary number of marker loci (q) is considered. *The underlying major genotype can be composed of several linked or unlinked loci.* Let the phenotype, Y , be a p -valued vector of measurements on a single individual, i.e., $Y = (Y_1, Y_2, \dots, Y_p)$. The most important cases are as follows: I, one trait and q marker phenotypes; II, one quantitative phenotype, $p - 1$ qualitative phenotypes, and q marker loci; and III, p continuous traits. Clearly, p qualitative traits can be considered as a single polychotomous trait and modeled as a multiple logistic function (Bonney et al., in press).

Case I: One Trait and q Marker Phenotypes

Let Y denote the value of the trait and $M = (M_1, M_2, \dots, M_q)$ denote the q marker phenotypes. The genotype, g , in formula (1) spans all the loci underly-

ing Y and M . Moreover, there is no loss of generality if one assumes that the marker phenotypes are completely determined by g . Note that if a person is not typed for a given marker, then the corresponding phenotype vector, M , is shortened for that person. With these considerations in mind, formula (1) can be decomposed and simplified as follows:

$$\begin{aligned} \Pr(Y, M|g, Y_A, X) &= \Pr(Y|g, Y_A, X)\Pr(M|Y, g, Y_A, X) \\ &= \Pr(Y_1|g, Y_A, X) \times 1 . \end{aligned} \quad (2)$$

The second factor is unity because M is completely determined by g . Thus, in linkage-to-marker studies we only need to specify the penetrance function (2) for the genotype underlying the univariate trait (Y_1). If Y_1 is continuous, penetrance function (2) is a normal density function (Bonney 1984); if it is qualitative, penetrance function (2) is specified as a logistic function (Bonney 1986; Bonney et al., in press).

Case II: One Quantitative Trait and $p - 1$ Qualitative Traits

Let the quantitative trait be Y_1 . Equation (2) can be decomposed in two ways; either as

$$\begin{aligned} \Pr(Y_1, Y_2, \dots, Y_p|g, Y_A, X) \\ = \Pr(Y_1|g, Y_A, X)\Pr(Y_2, \dots, Y_p|Y_1, g, Y_A, X) \end{aligned} \quad (3a)$$

or as

$$\begin{aligned} \Pr(Y_1, Y_2, \dots, Y_p|g, Y_A, X) \\ = \Pr(Y_2, \dots, Y_p|g, Y_A, X) \\ \times \Pr(Y_1|Y_2, \dots, Y_p, g, Y_A, X) . \end{aligned} \quad (3b)$$

The first factor in equation (3a) is the univariate penetrance function in equation (2); the second factor can be modeled by a multiple logistic function. Alternatively, in equation (3b) the first factor is assumed to be a logistic function and the second factor is modeled by the normal density function with the mean (and possibly the variance) dependent on $Y_2, \dots, Y_p$, g , Y_A , and X .

Case III: p Continuous Traits

Continuous variables can usually be modeled by the normal density function. By way of illustration we shall present the bivariate case ($p = 2$), assuming

a class A pattern of dependence (Bonney 1984) in which sib-sib residual correlations are due solely to common parentage. The results here and the extensions to $p > 2$ follow from standard regression formulas.

Three types of penetrance functions must be specified, involving the following phenotypes: (1) a person randomly selected from the population, (2) a person conditional on the spouse, or (3) a person conditional on the parents. There is no loss of generality in assuming that the three persons are members of a nuclear family, so for convenience the randomly selected person in (1) will be taken to be the mother; (2) will be the father conditional on the mother; and (3) will be the offspring conditional on the parents.

Before defining the penetrance function we must introduce some notation. Each of the three penetrances will be specified by a bivariate normal density function. We shall denote generically the ordinate at X_1 and X_2 of the bivariate normal with means μ_1 and μ_2 , variances σ_1 and σ_2 , and correlation ρ by

$$\zeta(X_1 - \mu_1, X_2 - \mu_2, \sigma_1^2, \sigma_2^2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X_1 - \mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{X_1 - \mu_1}{\sigma_1}\right)\left(\frac{X_2 - \mu_2}{\sigma_2}\right) + \left(\frac{X_2 - \mu_2}{\sigma_2}\right)^2\right]\right\}. \quad (4)$$

The bivariate phenotypes and parameters will be denoted as shown in table 1. In general, the μ 's depend

Table 1**Notation for the Parameters of the Bivariate Phenotype Model**

Person	Phenotype	Means	Variances
Mother	Y_{M1}, Y_{M2}	μ_{M1}, μ_{M2}	σ_1^2, σ_2^2
Father	Y_{F1}, Y_{F2}	μ_{F1}, μ_{F2}	σ_1^2, σ_2^2
Offspring	Y_{i1}, Y_{i2}	μ_{i1}, μ_{i2}	σ_1^2, σ_2^2

on genotype and other explanatory variables. For a nuclear family the correlation matrix in table 2 is reasonable.

With the above notation the penetrance function for the mother as a randomly selected person from the population is

$$\zeta(Y_{M1} - \mu_{M1}, Y_{M2} - \mu_{M2}, \sigma_1^2, \sigma_2^2, \rho_{12}) . \quad (5)$$

The second penetrance function (father's phenotype conditional on mother's) requires the regression of Y_{F1}, Y_{F2} on Y_{M1}, Y_{M2} . In terms of standard regression theory, the father's phenotypes adjusted for the mother's phenotypes can be written as

$$\begin{aligned} Z_{F1} &= Y_{F1} - \mu_{F1} - b_{FM11}^*(Y_{M1} - \mu_{M1}) \\ &\quad - (\sigma_1/\sigma_2) b_{FM12}^*(Y_{M2} - \mu_{M2}) ; \\ Z_{F2} &= Y_{F2} - \mu_{F2} - (\sigma_2/\sigma_1) b_{FM21}^*(Y_{M1} - \mu_{M1}) \\ &\quad - b_{FM22}^*(Y_{M2} - \mu_{M2}) , \end{aligned} \quad (6)$$

Table 2**Correlation Matrix for a Bivariate Trait in a Nuclear Family with c Offspring**

	Y_{M1}	Y_{M2}	Y_{F1}	Y_{F2}	Y_{11}	Y_{12}	Y_{21}	Y_{22}	...	Y_{C1}	Y_{C2}
Y_{M1}	1	ρ_{12}	ρ_{FM11}	ρ_{FM12}	ρ_{MO11}	ρ_{MO12}	ρ_{MO11}	ρ_{MO12}	...	ρ_{MO11}	ρ_{MO12}
Y_{M2}	ρ_{12}	1	ρ_{FM12}	ρ_{FM22}	ρ_{MO12}	ρ_{MO22}	ρ_{MO12}	ρ_{MO22}	...	ρ_{MO12}	ρ_{MO22}
Y_{F1}	ρ_{FM11}	ρ_{FM12}	1	ρ_{12}	ρ_{FO11}	ρ_{FO12}	ρ_{FO11}	ρ_{FO12}	...	ρ_{FO11}	ρ_{FO12}
Y_{F2}	ρ_{FM12}	ρ_{FM22}	ρ_{12}	1	ρ_{FO12}	ρ_{FO22}	ρ_{FO12}	ρ_{FO22}	...	ρ_{FO12}	ρ_{FO22}
Y_{11}	ρ_{MO11}	ρ_{MO12}	ρ_{FO11}	ρ_{FO12}	1	ρ_{12}	ρ_{SS11}	ρ_{SS12}	...	ρ_{SS11}	ρ_{SS12}
Y_{12}	ρ_{MO12}	ρ_{MO22}	ρ_{FO12}	ρ_{FO22}	ρ_{12}	1	ρ_{SS12}	ρ_{SS22}	...	ρ_{SS12}	ρ_{SS22}
Y_{21}	ρ_{MO11}	ρ_{MO12}	ρ_{FO11}	ρ_{FO12}	ρ_{SS11}	ρ_{SS12}	1	ρ_{12}	...	ρ_{SS11}	ρ_{SS12}
Y_{22}	ρ_{MO12}	ρ_{MO22}	ρ_{FO12}	ρ_{FO22}	ρ_{SS12}	ρ_{SS22}	ρ_{12}	1	...	ρ_{SS12}	ρ_{SS22}
Y_{C1}	ρ_{MO11}	ρ_{MO12}	ρ_{FO11}	ρ_{FO12}	ρ_{SS11}	ρ_{SS12}	ρ_{SS11}	ρ_{SS12}	...	1	ρ_{12}
Y_{C2}	ρ_{MO12}	ρ_{MO22}	ρ_{FO12}	ρ_{FO22}	ρ_{SS12}	ρ_{SS22}	ρ_{SS12}	ρ_{SS22}	...	ρ_{12}	1

where the b^* 's are given by the matrix

$$\begin{bmatrix} b_{FM11}^* & b_{FM12}^* \\ b_{FM21}^* & b_{FM22}^* \end{bmatrix} = \begin{bmatrix} \rho_{FM11} & \rho_{FM12} \\ \rho_{FM12} & \rho_{FM22} \end{bmatrix} \begin{bmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{bmatrix}^{-1}, \quad (7)$$

which requires inversion of a 2×2 matrix. The associated conditional variances and correlation of Z_{F1} and Z_{F2} are given by

$$\begin{aligned} \sigma_1^2 &= \sigma_1^2(1 - b_{FM11}^* \rho_{FM11} - b_{FM12}^* \rho_{FM12}) ; \\ \sigma_2^2 &= \sigma_2^2(1 - b_{FM21}^* \rho_{FM12} - b_{FM22}^* \rho_{FM22}) ; (8) \\ \sigma_1 \sigma_2 \rho_{12} &= \sigma_1 \sigma_2 (\rho_{12} - b_{FM11}^* \rho_{FM12} \\ &\quad - b_{FM12}^* \rho_{FM22}) . \end{aligned}$$

The required penetrance function for a father conditional on the mother is thus

$$\zeta(Z_{F1}, Z_{F2}, \sigma_1^2, \sigma_2^2, \rho_{12}) . \quad (9)$$

The third type of penetrance function (offspring conditional on parents) is derived analogously. The adjusted variates are

$$\begin{aligned} Z_{i1} &= Y_{i1} - \mu_{i1} - b_{FO11}^*(Y_{F1} - \mu_{F1}) \\ &\quad - (\sigma_1/\sigma_2) b_{FO12}^*(Y_{F2} - \mu_{F2}) \\ &\quad - b_{MO11}^*(Y_{M1} - \mu_{M1}) \\ &\quad - (\sigma_1/\sigma_2) b_{MO12}^*(Y_{M2} - \mu_{M2}) ; \\ Z_{i2} &= Y_{i2} - \mu_{i2} - (\sigma_2/\sigma_1) b_{FO21}^*(Y_{F1} - \mu_{F1}) \\ &\quad - b_{FO22}^*(Y_{F2} - \mu_{F2}) - (\sigma_2/\sigma_1) b_{MO21}^* \\ &\quad \times (Y_{M1} - \mu_{M1}) - b_{MO22}^*(Y_{M2} - \mu_{M2}), \end{aligned} \quad (10)$$

where the b^* 's are given by

$$\begin{pmatrix} b_{MO11}^* b_{MO12}^* b_{FO11}^* b_{FO12}^* \\ b_{MO21}^* b_{MO22}^* b_{FO21}^* b_{FO22}^* \end{pmatrix} =$$

$$\begin{pmatrix} \rho_{MO11} \rho_{MO12} \rho_{FO11} \rho_{FO12} \\ \rho_{MO21} \rho_{MO22} \rho_{FO21} \rho_{FO22} \end{pmatrix} \times \begin{pmatrix} 1 & \rho_{12} & \rho_{FM11} & \rho_{FM12} \\ \rho_{12} & 1 & \rho_{FM12} & \rho_{FM22} \\ \rho_{FM11} & \rho_{FM12} & 1 & \rho_{12} \\ \rho_{FM12} & \rho_{FM22} & \rho_{12} & 1 \end{pmatrix}^{-1}, \quad (11)$$

which requires inversion of a 4×4 matrix. The conditional variances and correlations of Z_{i1} and Z_{i2} are given by

$$\begin{aligned} \bar{\sigma}_1^2 &= \sigma_1^2(1 - b_{MO11}^* \rho_{MO11} - b_{MO12}^* \rho_{MO12} \\ &\quad - b_{FO12}^* \rho_{FO11} - b_{FO11}^* \rho_{FO12}) ; \\ \bar{\sigma}_2^2 &= \sigma_2^2(1 - b_{MO21}^* \rho_{MO21} - b_{MO22}^* \rho_{MO22} \\ &\quad - b_{FO21}^* \rho_{FO21} - b_{FO22}^* \rho_{FO22}) ; \\ \bar{\sigma}_1 \bar{\sigma}_2 \bar{\rho}_{12} &= \sigma_1 \sigma_2 (\rho_{12} - b_{MO11}^* \rho_{MO12} \\ &\quad - b_{MO12}^* \rho_{MO22} - b_{FO11}^* \rho_{FO12} \\ &\quad - b_{FO12}^* \rho_{FO22}) . \end{aligned} \quad (12)$$

The required penetrance function (of offspring given parents) is

$$\zeta(Z_{i1}, Z_{i2}, \bar{\sigma}_1^2, \bar{\sigma}_2^2, \bar{\rho}_{12}) . \quad (13)$$

These equations can be simplified considerably if spouse correlations are assumed to be zero; in this case equation (9) is the same as equation (5), with the subscript F in place of M. Furthermore, the 4×4 matrix in equation (11) is easily inverted, so that after the necessary multiplication we obtain

$$\begin{aligned} b_{MO11}^* &= (\rho_{MO11} - \rho_{12} \rho_{MO12}) / (1 - \rho_{12}^2) , \\ b_{FO11}^* &= (\rho_{FO11} - \rho_{12} \rho_{FO12}) / (1 - \rho_{12}^2) ; \\ b_{MO12}^* &= (\rho_{MO12} - \rho_{12} \rho_{MO11}) / (1 - \rho_{12}^2) , \\ b_{FO12}^* &= (\rho_{FO12} - \rho_{12} \rho_{FO11}) / (1 - \rho_{12}^2) ; \\ b_{MO21}^* &= (\rho_{MO12} - \rho_{12} \rho_{MO22}) / (1 - \rho_{12}^2) , \\ b_{FO21}^* &= (\rho_{FO12} - \rho_{12} \rho_{FO22}) / (1 - \rho_{12}^2) ; \\ b_{MO22}^* &= (\rho_{MO22} - \rho_{12} \rho_{MO12}) / (1 - \rho_{12}^2) , \\ b_{FO22}^* &= (\rho_{FO22} - \rho_{12} \rho_{FO12}) / (1 - \rho_{12}^2) . \end{aligned} \quad (14)$$

Extension to Pedigrees

The penetrance equations presented thus far have assumed a nuclear family structure. For pedigrees with arbitrary structure, we need only allow for the fact that some persons have parents and a spouse. If the spouse correlations are zero, no modifications are needed; otherwise, the required modifications are simple because of the assumption that the phenotypes of an individual's spouse and of his or her ancestors and sibs are uncorrelated. Denoting the bivariate phenotype of a preceding spouse by Y_{S1} and Y_{S2} , with corresponding means μ_{S1} and μ_{S2} , we need only modify equation (10) by adding the term

$$-b_{FM11}^*(Y_{S1} - \mu_{S1}) - \sigma_1/\sigma_2 b_{FM12}^*(Y_{S2} - \mu_{S2}) \quad (15)$$

to Z_{i1} , and by adding

$$-\sigma_2/\sigma_1 b_{FM21}^*(Y_{S1} - \mu_{S1}) - b_{FM22}^*(Y_{S2} - \mu_{S2}) \quad (16)$$

to Z_{i1} ; then modify equation (12) by adding (to the terms within parentheses)

$$-b_{FM11}^* \rho_{FM11} - b_{FM12}^* \rho_{FM12} \quad (17)$$

to $\tilde{\sigma}_1^2$,

$$-b_{FM12}^* \rho_{FM12} - b_{FM22}^* \rho_{FM22} \quad (18)$$

to $\tilde{\sigma}_2^2$, and

$$-b_{FM11}^* \rho_{FM12} - b_{FM12}^* \rho_{FM22} \quad (19)$$

to $\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\rho}_{12}$.

Computer Implementation

Implementation of the models we have discussed requires merely that the penetrance function be modified in the existing computer programs for genetic analysis. Thus, all existing algorithms designed for the fast calculation of likelihoods over pedigrees can be employed (see Elston and Stewart 1971; Cannings et al. 1978). The bivariate forms of the class A models above have been implemented in the computer package LINKAGE (Lathrop et al. 1984). These programs allow both multilocus analysis of markers

and consideration of the interactions of several different linked or unlinked loci with univariate or bivariate phenotypes.

In LINKAGE, missing observations are currently handled as suggested by Bonney (1984). Thus, the penetrance is set to unity for missing persons, according to the method of Elston and Rao (1978). For an observed person, Y_A in the penetrance function (1) includes only the observed ancestral phenotypes. Thus formula (9) for a father is reduced to the form of formula (5) if the wife is missing; formula (13) is reduced to the form of formula (9) if one parent is missing and to the form of formula (5) if both parents are missing. A comprehensive study of missing observations in the regressive models will be communicated separately.

Applications

Hypercholesterolemia and the LDL Receptor Gene

Leppert et al. (1986) published data on total cholesterol and LDL receptor genotypes in a large pedigree with dominantly inherited hypercholesterolemia. The pedigree was ascertained through men with myocardial infarction prior to age 55 years.

To establish linkage with the LDL receptor locus, these authors assumed (a) contributions of a single major locus and (b) polygenic heritability of age- and sex-adjusted measurements of LDL cholesterol. By segregation analysis under the mixed model and by using the PAP program (Hasstedt and Cartwright 1981), genetic parameters were estimated in a larger set of pedigrees in which both the major gene and polygenic inheritance were found to contribute significantly to the phenotypic variance. Recombination between the LDL receptor genotypes and the major locus analysis in the single pedigree was undertaken with the other genetic parameters fixed at these estimates. The maximum lod score was 7.52 at $\theta = 0$.

We have analyzed the data on 57 members of this pedigree by using the extended regressive models. The pedigree and data are shown in figure 1; the second husband and a child of one member of the pedigree were not included in our analysis. We undertook no correction for ascertainment but, to avoid bias in the estimate of gene frequency, assumed a fixed frequency of .001 for a dominant major gene leading to elevated levels of LDL cholesterol.

The results of our analysis are shown in table 3.

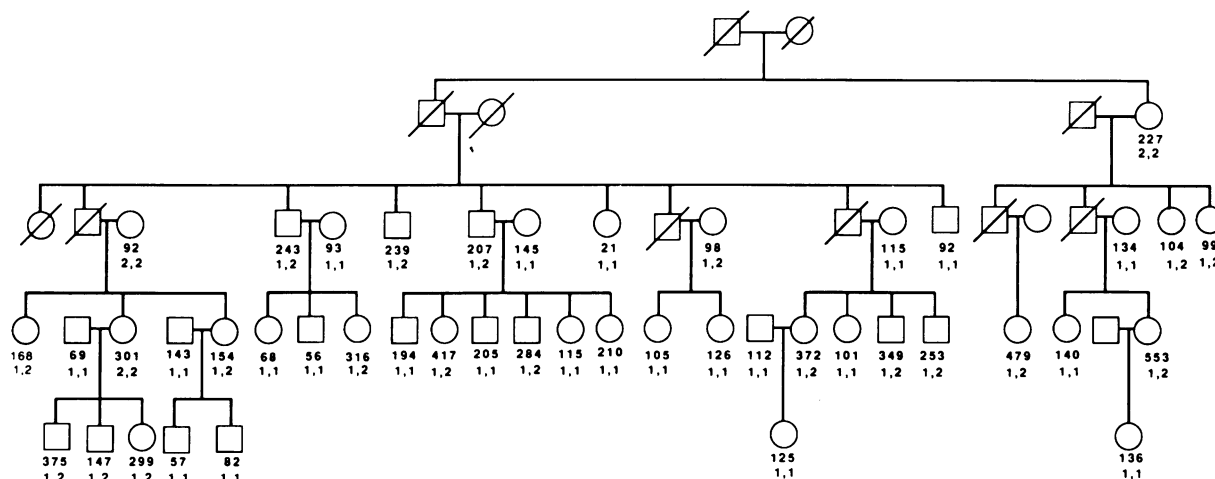


Figure 1 Linkage between hypercholesterolemia and the LDL receptor gene in a large pedigree. Total cholesterol levels adjusted for age and sex, and the genotypes for the LDL receptor, appear below the symbol for each sampled individual.

We found no evidence of a major gene in this pedigree when segregation analysis was performed assuming free recombination ($\theta = .5$) with the LDL receptor. Indeed, the estimate of the mean difference between heterozygotes and homozygotes for a hypothesized major gene is zero; therefore, the estimates of the residual correlations are the same with and without the inclusion of a major gene. Of these estimates, only that of the spouse correlation is significant ($p_{SS} = -.83$, $\chi^2_1 = 4.2$). A possible explanation for the negative value of this estimate is ascertainment; the pedigree contains many core members with hypercholesterolemia who have married persons drawn from the general population, in which the disease is infrequent.

In the combined segregation and linkage analysis, the estimate of recombination between a major gene for elevated cholesterol levels and the LDL receptor gene is 0. Table 3 shows that neither the spouse-spouse nor the parent-offspring correlation is significant when the recombination fraction is estimated simultaneously. As the spouse correlation is significant when $\theta = .5$, however, we advocate that the test for linkage include this correlation. The likelihood ratio test then gives $\chi^2_1 = 27.2$, equivalent to a lod score of 5.9. The test of linkage without residual correlations is also significant ($\chi^2_1 = 31.7$ —or a lod score of 6.9).

Hypobetalipoproteinemia and the Apolipoprotein B Gene

Leppert et al. (in press) have ascertained a 52-member pedigree through a single family in which

several members exhibit reduced plasma concentrations of LDL cholesterol, triglycerides, and apolipoprotein B, traits characteristic of heterozygous hypobetalipoproteinemia. Figure 2 shows the pedigree, with genotypic data for two polymorphisms of the apoB gene, along with age-adjusted triglyceride and total cholesterol values; ascertainment was through the family of the proband indicated in the figure. Analysis of the quantitative measurements by means of the PAP (Hasstedt and Cartwright 1981) and LINKAGE (Lathrop et al. 1984) programs showed significant evidence of linkage to the apoB gene. Here we use the regressive model for segregation and linkage

Table 3

Results of the Combined Segregation and Linkage Analysis of Total Cholesterol and the LDL Receptor Locus

Linkage Assumption and Segregation Model	-2 ln Likelihood
No linkage:	
No residual correlation	346.8
Spouse correlation	342.3
Parent-offspring correlation	346.1
All residual correlations	341.7
Recombination estimated:	
No residual correlation	315.1
Spouse correlation	315.0
Parent-offspring correlation	315.0
All residual correlations	315.0

NOTE.—Under the assumption of no linkage ($\theta = .5$), the estimated mean difference for the distribution of total cholesterol in normal homozygotes and disease heterozygotes is zero.

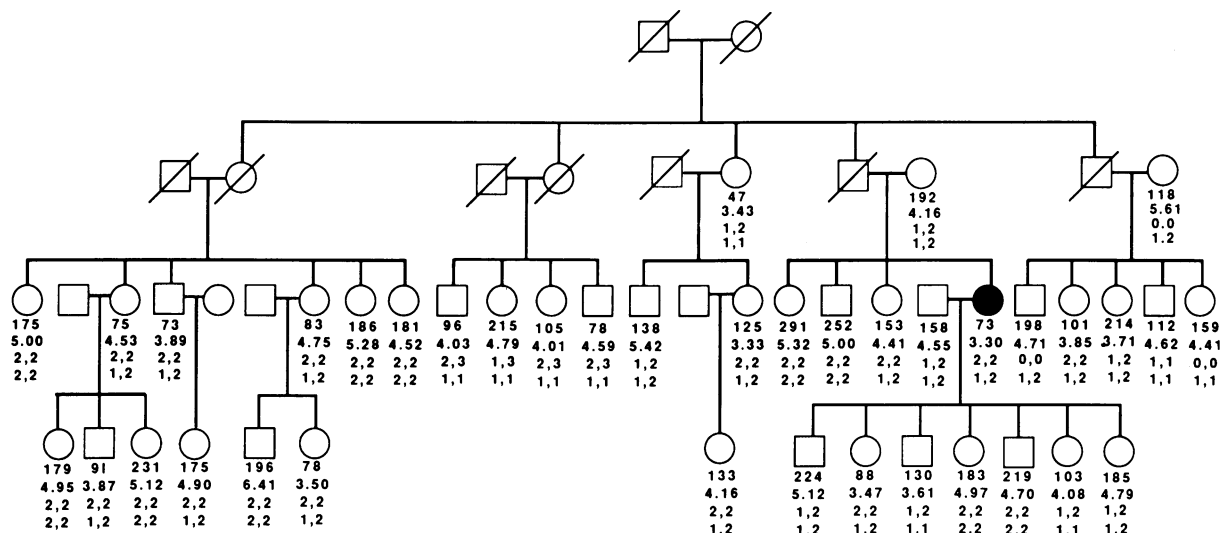


Figure 2 Linkage between hypobetalipoproteinemia and the apolipoprotein B gene in a large pedigree. The proband is designated by the blackened circle. Below the symbol for each sampled individual are, from top to bottom, (1) age- and sex-adjusted values of total cholesterol, (2) age- and sex-adjusted values of triglycerides, (3) genotypes for ApoB (*Msp*I), and (4) genotypes for ApoB (*Xba*I).

analysis of triglyceride and total cholesterol data with the apoB genotypes. Again, other than to assume a fixed gene frequency of .001 for the disease allele, no ascertainment correction has been applied to the data.

We applied a logarithmic transformation to adjust for skewness of the triglyceride values and standardized the data prior to analysis because of differences in the scale of the triglyceride and cholesterol measurements. As the haplotypes for apoB cannot be de-

duced for all typed members of the pedigree, the markers were considered as separate but completely linked loci in linkage equilibrium.

The results of the univariate analyses of triglycerides and total cholesterol are shown in table 4. The recombination estimates are .0 with both variables. For triglycerides the estimated spouse correlation approaches -1.0 when estimated simultaneously with the recombination fractions ($\chi^2_1 = 6.6$); otherwise, the residual correlations are not significant. As in the

Table 4

Likelihood Results for the Combined Segregation and Linkage Analysis of Total Cholesterol and Triglycerides with ApoB

LINKAGE ASSUMPTION AND SEGREGATION MODEL	-2 ln LIKELIHOOD		
	Total Cholesterol	Triglycerides	Bivariate
No linkage:			
No residual correlation	170.0	166.0	188.0
Spouse correlation	165.9	165.9	184.2
Parent-offspring correlation	165.6	165.2	179.0
All residual correlations	164.6	165.2	179.0
Recombination estimated:			
No residual correlation	142.3	154.8	155.7
Spouse correlation	140.7	148.2	147.0
Parent-offspring correlation	141.7	150.9	152.2
All residual correlations	138.3	147.9	145.4

NOTE.—Under the assumption of no linkage ($\theta = .5$), the estimated mean difference for the distributions of total cholesterol and triglyceride in normal homozygotes and disease heterozygotes is zero.

previous example, the negative spouse correlation could be due to the method of ascertainment. Linkage with apoB is significant for both total cholesterol ($\chi^2_1 = 25.2$ —or a lod score of 5.5) and triglycerides ($\chi^2_1 = 11.4$ —or a lod score of 3.8) when the spouse correlations are included in the latter test. Interestingly, if linkage is tested without accounting for this residual correlation, the lod score for triglycerides and apoB is 2.42 ($\chi^2_1 = 11.1$). By convention the demonstration of linkage requires a lod score ≥ 3.0 ; therefore, the linkage between triglycerides and apoB is not conclusive when modeled without residual spouse correlations.

Analysis of triglycerides and total cholesterol under the bivariate model gives results similar to those of the univariate analyses (table 4). The spouse-spouse correlations are marginally significant when estimated simultaneously with the recombination fraction ($\chi^2_3 = 7.9$ at $\theta = 0$). The test for linkage when these correlations are included gives $\chi^2_1 = 36.5$ —or a lod score of 7.9—compared with $\chi^2_1 = 32.3$ —or a lod score of 7.0—when residual correlations are not included. Thus, the lod score from the bivariate analysis is greater than the lod scores from either of the univariate analyses.

Discussion

The extension of regressive models to include multivariate data and an arbitrary number of linked marker loci allows an effective combination of segregation and linkage analysis for studies of multifactorial disease. Linkage studies of multifactorial disease can be conducted either with candidate genes—such as the LDL receptor and apoB polymorphisms in the examples presented here—or with arbitrary DNA segments that may show linkage with the phenotype concerned. The detailed genetic maps that are now available for many chromosomes will allow one to combine these two approaches by selecting for linkage studies several highly polymorphic markers in the vicinity of a candidate gene. Joint analysis of several markers in the region, as provided for in our implementation of the regressive models, will be important for the detection of linkage and for precise localization of genes involved in multifactorial diseases.

Regressive models have several advantages over the mixed model for the study of multifactorial disease; these advantages include allowing for complicated patterns of familial correlation, including spouse-spouse and parent-offspring phenotypic cor-

relations, the possibility of introducing regressor variables directly into the analysis, and computational efficiency (Bonney 1984, 1986). For example, it is possible to estimate and test the hypothesis of different father-offspring and mother-offspring residual correlations under regressive models but not under the mixed model. In segregation analysis, the regressive and mixed models give similar results when phenotypic correlations are due to a major gene and to additive polygenic factors (Bonney 1984).

The approach we have adopted—to estimate recombination, parameters of a major gene, and residual correlation simultaneously—is likely to be important for studies of diseases in which polygenic and environmental sources of familial correlation could be significant. Indeed, detection of linkage may in some cases require the incorporation of residual covariance in the analysis to obtain significance, as illustrated in our univariate analysis of triglycerides in hypobetalipoproteinemia.

Linkage studies of rare disease will usually be undertaken in families selected for investigation because they contain many affected members. The spouses (or “marry-ins”) of affected pedigree members are highly unlikely to exhibit symptoms of disease. This is one possible source of significant negative spouse correlations. Although it is possible to incorporate explicit formulations for ascertainment correction (see, e.g., Cannings and Thompson 1979; Elston and Sobel 1979), such corrections may not account for negative spouse correlations. Since ignoring spouse correlations can apparently lead to a failure to detect linkage in some cases (as shown in one of our examples), allowing for residual correlations may be a simple way to account for certain types of ascertainment—even for traits determined totally by a single genetic locus. Further investigations are required before the importance of accounting for such correlations in linkage analysis can be fully assessed.

The popular linear combination approach to segregation analysis for multivariate traits (Goldin et al. 1980) does not allow for pleiotropy or for control of different phenotypes by different loci in a readily interpretable manner. With the methods presented here, bivariate traits can be analyzed either under a model that assumes pleiotropic effects of a single locus on two phenotypes or under a model that assumes two phenotypes controlled by separate linked or unlinked loci. The generalization to three or more traits is limited only by the practical problem of estimating a large number of parameters.

Acknowledgments

G.E.B. was partially supported by U.S. Public Service Research Grants 2P30CA14718 and R29GM39573. J.M.L. is an investigator of the Howard Hughes Medical Institute.

References

- Bonney, G. E. 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, G. E., G. M. Dunston, and J. Wilson. The use of regressive logistic models for ordered and unordered polychotomous traits: application to affective disorders. *Genet. Epidemiol.* (in press).
- Cannings, C., and E. A. Thompson. 1979. Ascertainment in the sequential sampling of pedigrees. *Clin. Genet.* 12:208.
- Cannings, C., E. A. Thompson, and M. H. Skolnick. 1978. Probability functions on complex pedigrees. *Adv. Appl. Probability* 10:26–61.
- Elston, R. C., and D. C. Rao. 1978. Statistical modeling and analysis in human genetics. *Annu. Rev. Biophys. Bioeng.* 7:253–286.
- Elston, R. C., and E. Sobel. 1979. Sampling considerations in the gathering and analysis of pedigree data. *Am. J. Hum. Genet.* 31:62–69.
- Elston, R. C., and J. Stewart. 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.* 21:523–542.
- Goldin, L. R., R. C. Elston, J. B. Graham, and C. H. Miller. 1980. Genetic analysis of Von Willebrand's disease in two large pedigrees: a multivariate approach. *Am. J. Hum. Genet.* 6:279–293.
- Hasstedt, S. J., and P. E. Cartwright. 1981. PAP: pedigree analysis package. 2d rev. Tech. Rep. 13. Department of Medical Biophysics and Computing, University of Utah, Salt Lake City.
- Lalouel, J. M., D. C. Rao, N. E. Morton, and R. C. Elston. 1983. A unified model for complex segregation analysis of quantitative traits. *Am. J. Hum. Genet.* 35:816–826.
- Lathrop, G. M., J. M. Lalouel, C. Julier, and J. Ott. 1984. Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA* 81:3443–3446.
- Leppert, M. F., J. Breslow, L. Wu, S. Hasstedt, P. O'Connell, M. Lathrop, R. Williams, R. White, and J. M. Lalouel. Inference of a molecular defect of apolipoprotein B in a hypobetalipoproteinemia by linkage analysis in a large kindred. *J. Clin. Invest.* (in press).
- Leppert, M., S. J. Hasstedt, T. Holm, P. O'Connell, L. Wu, O. Ash, R. R. Williams, and R. White. 1986. A DNA probe for the LDL receptor gene is tightly linked to hypercholesterolemia in a pedigree with early coronary disease. *Am. J. Hum. Genet.* 39:300–306.
- Morton, N. E., and C. J. MacLean. 1974. Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am. J. Hum. Genet.* 26:489–503.