METHODS

Evolving Methods in Genetic Epidemiology. I. Analysis of Genetic and Environmental Factors in Family Studies

T. H. Beaty

INTRODUCTION

Using family data (either sibships, nuclear families, or extended pedigrees) to test models of inheritance is the essence of human genetics. The rigor and precision of this testing process can range from a strictly observational process (eliminating simple single-locus models one at a time on individual pedigrees) through formal tests of hypotheses about the expected proportion of affected sibs up to maximum likelihood estimation of parameters in a generalized model of inheritance. The goal of this process is to use family data to identify Mendelian mechanisms. If a phenotype (either a quantitative or a qualitative trait) is directly controlled by gene(s), the phenotypic distribution in families (and ultimately in populations) should follow the predictions of Mendelian genetics. Unfortunately, the information content of family data from observational studies of humans is limited, and the statistical tools available have important limitations.

As the observed phenotype becomes further and further removed from the direct action of genes, however, its distribution in families will be more influenced by environmental factors and is less likely to follow predictable Mendelian patterns. Such phenotypes are traditionally called "multifactorial" because both genetic and environmental factors control the trait (n.b., the term "complex" disease is often used as a synonym). Many important diseases fall under this multifactorial label, including most of the major chronic diseases (cardiovascular disease, cancer, diabetes, some birth defects, etc.). Epidemiologic studies of most, if not all, of these multifactorial diseases have identified one or more observable environmental factors also associated with risk, and it is the combination of genetic and environmental factors that is important. It remains a major challenge to the field of genetic epidemiology to develop and implement study designs and analytic approaches capable of combining information from observable environmental risk factors into a comprehensive analysis of etiologic mechanisms.

LIKELIHOOD STRATEGIES FOR ANALYSIS OF FAMILY DATA

While the classic Neyman-Pearson strategy of stating the null hypothesis and testing its predictions against observed data remains the statistical standard, much of genetic epidemiology focuses on fitting models to data rather than testing a single, pre-specified null hypothesis. This reliance on some form of "goodness-of-fit" statistic reflects a scientific prejudice in favor of genetic models (1). The overall strategy is to fit a series of models of inheritance and choose the best fitting, simplest model (i.e., the most parsimonious model) to explain the data at hand.

The likelihood of any one model given the data, \( L(Model|Data) \), is proportional to the probability of observing the data predicted under that model (2), and the algorithm for computing these likelihoods takes advantage of the repeating internal structure of families (i.e., nuclear family units are connected in various ways to create more extended pedigrees) (3-4). Conceptually, this likelihood can be visualized as a series of joint probabilities summed over all possible combinations of essential "types," either Mendelian genotypes or more general "ousiotypes." There are two major classes of parameters: 1) frequencies of "types" and 2) penetrances, which specify the probability of displaying a phenotype conditional on the unobserved "type." Thus, an intrinsic parameter is the number of "types." It is important to note that frequencies of genotypes among founder individuals (i.e., persons without parents in the dataset) are simple functions of allele frequencies (e.g., often predicted by Hardy-
Weinberg equilibrium in the reference population). Nonfounders (i.e., persons whose parents are in the dataset) have genotypic frequencies solely determined by the respective genotypes of their parents, as summarized by transmission parameters. The number of transmission parameters is determined by the number of “types” of parents. For example, under a single locus, two allele model there are three transmission parameters, one for each parental genotype (AA, AB, and BB), and these represent the probability of transmitting the A allele to an offspring.

More generally, if there are \( m = 1 \ldots M \) “types,” let \( f(m) \) represent the frequency of the \( m \)-th “type” and \( P(Y|m) \) be the penetrance function (where either \( P(Y|m) \) is a conditional probability that person \( I \) has a qualitative phenotype \( Y \), or an analogous conditional density function, if \( Y \) is quantitative). Then any founder individual \( I \) in a pedigree contributes

\[
\sum_{m=1}^{M} P(Y|m)f(m)
\]

to the likelihood function. In other words, the likelihood function is merely the product of the prior probability of having a given “type” times the conditional probability of having the observed phenotype given this “type,” summed over all possible “types.” If matings are independent, the joint probability of two founder parents \( I \) and \( J \) can be written as

\[
\prod_{i=1}^{M} \prod_{j=1}^{M} P(Y|i)f(i)P(Y|j)f(j) = \sum_{i} P(Y|i)P(Y|j)f(i)f(j)
\]

Furthermore, letting the transmission parameters \( \tau(k|i,j) \) describe the probability of the \( k \)-th “type” in offspring \( K \) of parents \( I \) and \( J \) allows the likelihood of a parent-offspring trio to be written as

\[
\prod_{i=1}^{M} \prod_{j=1}^{M} \prod_{k=1}^{M} P(Y|i)f(i)P(Y|j)f(j)P(Y|k)\tau(k|i,j) = \sum_{i} \sum_{j} \sum_{k} P(Y|i)f(i)P(Y|j)f(j)P(Y|k)\tau(k|i,j)
\]

These sequential summations over all possible “types” for parent-offspring trios accumulate a non-normalized conditional probability (or likelihood) of the observed data on parents \( I \) and \( J \) and their offspring \( K \), weighted by the appropriate prior probabilities (5). In general, if a pedigree contains \( N \) individuals, \( N_1 \) of whom are founders (and \( N_2 \) of whom are nonfounders), the likelihood function of the entire pedigree can be written as

\[
\sum_{i_1} \sum_{i_2} \sum_{i_3} \prod_{j=1}^{N} P(Y|j)\prod_{k=1}^{M} f(i_1)\prod_{m=1}^{M} \tau(i_m|i_1,i_2,i_3)
\]

Note the first product (that for the penetrance parameters) is over all \( N \) members of the pedigree, while the middle product (for frequency parameters) is over the \( N_1 \) founders, and the last product (for transmission parameters) is over the \( N_2(= N - N_1) \) nonfounders (6).

Thus, the likelihood function can be considered a complete enumeration of all possible combinations of genotypes in the pedigree considering both their prior probabilities (either in terms of allele frequencies for founders, or transmission probabilities for nonfounders) and the conditional probability of their observed phenotypes. This likelihood is a probability for a qualitative trait, while for a quantitative trait the probabilities of having any given genotype are multiplied by the height of the normal curve at the individual’s observed phenotypic value. The In-likelihood values are summed up over each family in a dataset to give a single value for the model on the entire dataset. By changing the parameter values underlying the model, these In-likelihood functions can be maximized to give estimators with desirable statistical properties.

Various models of inheritance are constructed by altering the basic parameters described in table 1. For example, the only difference between a codominant model and a dominant model lies in the penetrance function of the heterozygote (i.e., either it is identical to one homozygote or it is distinct). Similarly, the difference between a Mendelian model of inheritance and a nongenetic model lies in the transmission parameters (i.e., they are Mendelian or not). Single-locus Mendelian inheritance specifies that these parameters take on the values shown in table 1, but it is straightforward to specify two-locus Mendelian models. Models with no residual heritability or other correlation are called major gene models, because there is no familial resemblance beyond that due to the Mendelian gene(s). The “mixed model” of inheritance includes penetrance functions for both a major gene component and residual familial correlations based on a polygenic component, where any correlation not due to the major gene is a function of an additive polygenic component. This component reflects the action of many independent Mendelian loci, and the final residual correlation is the product of kinship coefficients between pairs of relatives (which specify the probability of sharing one allele identical by descent) and the ratio of the additive
TABLE 1. Parameters of a general model of inheritance

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Symbols</th>
<th>Alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Penetrance</td>
<td>Conditional probability of displaying phenotype given genotype</td>
<td>( P(\text{affected}</td>
<td>\mu, \sigma^2) )</td>
</tr>
<tr>
<td>Frequencies</td>
<td>For founders: population &quot;allele&quot; frequency</td>
<td>( p^2, 2pq, q^2 )</td>
<td>Hardy-Weinberg equilibrium</td>
</tr>
<tr>
<td></td>
<td>For nonfounders: transmission parameters</td>
<td>( p_1, p_2, (1 - p_1 - p_2) )</td>
<td>Arbitrary</td>
</tr>
<tr>
<td>Residual</td>
<td>Familial correlations conditional on major locus &quot;type&quot;</td>
<td>( p_{\text{spouse}}, p_{\text{per-off. sib}} )</td>
<td>Mendelian</td>
</tr>
<tr>
<td>Correlative</td>
<td>Regression coefficients for observable covariates</td>
<td>( p = \phi h^2 )</td>
<td>Non-genetic: ( t_1 = t_2 = t_3 = t )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Polygenic or mixed</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>May be &quot;type&quot; specific</td>
</tr>
</tbody>
</table>

polygenic component to the total phenotypic variance (i.e., the heritability). Regressive models of inheritance (7) allow an arbitrary correlation under an auto-regressive structure for nuclear families, where each individual is conditioned on his/her predecessors in the family (i.e., the first child is conditioned on the parents, the second child on the first child plus the parents, etc.). While these regressive models can be constrained to be identical to the mixed model for nuclear families (by forcing the residual sib correlation to be equal to the parent-offspring correlation), extending regressive models to second-degree relatives and beyond implies a lower residual correlation than predicted under a mixed model with a residual polygenic component.

Testing hypotheses and selecting the most parsimonious model

The likelihood method described above allows specific hypotheses to be tested by comparing competing models under the likelihood ratio test. The likelihood ratio test is computed as twice the difference between \( \ln \)-likelihoods of a complete model and a reduced model, and asymptotically, this likelihood ratio test statistic follows a \( \chi^2 \) distribution with degrees of freedom equal to the difference in the number of unconstrained parameters fit. Not all models of inheritance can be compared directly under the likelihood ratio test, however, because the hierarchical structure necessary for valid inferences may not be met (i.e., the model of interest may not be a subset of the other). However, the likelihood method can still be used to compare various models by employing Akaike's information criterion, which is simply \(-2\ln(L(\text{Model}|\text{Data}))\) plus twice the number of parameters estimated (8). This penalty for adding more parameters to the model is useful for identifying the most parsimonious model—it is simply the one with the smallest Akaike's information criterion value.

Inferences about genetic control are built up through a sequential process of showing 1) that there is evidence of distinct "types" of individuals (commingling analysis), 2) that the transmission of these "types" of individuals is compatible with Mendelian inheritance, and 3) that transmission of these "types" is not compatible with a simplistic nongenetic equal probability model (9, 10). The limits of statistical inference are strict: statistical analysis can never prove a biologic mechanism exists, it can only assess if observed data are compatible with predictions of a given model of inheritance. The probability of correctly identifying a Mendelian model depends on many uncontrollable factors, i.e., the size and structure of the families, the phenotypic separation between genotypes, and the true underlying model of inheritance including its allele frequencies. Among those factors that can be controlled is the total number of families and, to some extent, the structure of the families (e.g., nuclear families can become extended pedigrees through greater effort).

It is not a simple matter, however, to estimate statistical power (i.e., the probability of wrongly accepting a null hypothesis when the alternative is true) in a multistage testing strategy. While analysis of recurrence risks in sibships can be formulated as simple null and alternative hypotheses for qualitative phenotypes (11), this is not possible for quantitative phenotypes. Simulation studies suggest that dominant mixed models for quantitative traits can be identified with confidence with samples of moderate-sized families ascertained through a proband (12, 13). A sequential sampling strategy was suggested, whereby the proband would be identified through an extreme phenotype (e.g., beyond the 95th percentile) and all first-
degree relatives (parents, children, full sibs) of the proband are then recruited. If one of these relatives also had an extreme phenotype, the family would be extended by recruiting all his/her first-degree relatives (second-degree or more distant relatives of the proband) until the extreme phenotype was no longer found among additional pedigree members. This sequential approach maximizes the proportion of extreme phenotypes, and thus increases the power of statistical tests. It is not clear that this strategy would be optimal for recessive traits, however. Borecki et al. (14, 15) conducted a simulation study of randomly sampled nuclear families and showed that the probability of correctly inferring Mendelian control (when it does exist) ranges from 22 to 50 percent under dominant models (for sample sizes of 50 to 300 nuclear families). Thus, it is much easier to identify dominant mechanisms controlling quantitative traits than recessive mechanisms. Furthermore, there is a high probability of failing to detect Mendelian mechanisms in small data sets.

Limitations of this strategy

Since the first goal in this type of genetic analysis is to identify the best-fitting, most parsimonious model of inheritance, one obvious limit arises from the breadth of models being examined. If all models considered are essentially incorrect, the best fitting incorrect model will have limited utility. Thus, it is desirable to not only compare the maximum likelihoods within a series of models and choose the best, but to have some assessment of its fit to the observed data, especially for the final model selected. While a few goodness-of-fit statistics have been developed under general multifactorial models (16), these are designed to identify either outlier individuals or families rather than assess the overall fit of the model. One way to judge the utility of a model of inheritance for a quantitative trait is to graphically display the observed phenotypic distribution against the predicted genotypic distributions under the best model of inheritance (see, for example, figure 1 in Prenger et al. (17)). Critical analysis of this type of display can reveal much about the appropriateness of the best model, at least in terms of the marginal phenotypic distributions. Further analysis of observed familial correlations compared with those predicted under the best model of inheritance, or critical analysis of the observed phenotypes among individuals with various predicted genotypes, offer another way to assess how good the best model really is.

Maximizing the likelihood function will provide statistically consistent and unbiased estimators, but questions remain about the robustness of these estimators to mis-specification of the model (see the section on Alternative approaches, below). The real problem is that a number of competing models of inheritance can explain a given pattern of familial aggregation, and any single pedigree has limited power to distinguish among models. The dilemma was first encountered when statistical tests were developed to compare single major locus models with more general polygenic models (18). Since both are based on Mendelian inheritance of autosomal genes (one or many), it is not surprising that their predictions overlap.

Etiologic heterogeneity

In complex diseases there is always the possibility of etiologic heterogeneity, where all evidence in favor of or against a particular model comes from a subgroup of families, while other families may actually represent some different, possibly nongenetic, mechanism. It is possible to test for etiologic heterogeneity under this likelihood approach if some external grouping is available. For example, a heterogeneity statistic $\chi^2$ is computed as

$$\chi^2 = 2[\ln L(\text{Model|All data}) - \sum_{i=1}^{k} \ln L(\text{Model|subset i})]$$

(5)

which has $(k-1)$ degrees of freedom for a model with $k$ parameters. Typically, the $i = 1...k$ subsets are based on characteristics of the proband (e.g., mildly affected or severely affected), but these could be based on ethnic group or some other observable factor.

A somewhat more subjective strategy is to use the likelihood values themselves to sort families into groups that support one model of inheritance over another (2). Moll et al. (19) conducted an analysis of apolipoprotein A1 (apoA1) levels in randomly ascertained families and found that the entire data set could identify no single model of inheritance. However, when this “model choice” approach was used to group families, there was consistent evidence that a subgroup of families were segregating for a Mendelian gene leading to elevated apolipoprotein A1 levels. Coresh et al. (20, 21) used a similar approach to identify a subgroup of families ascertained through a proband suspected of having premature heart disease to test for Mendelian control of apolipoprotein B (apoB) levels. The subgroup of families segregating for a Mendelian locus were then used to test for linkage to markers in the apoB gene itself, a step obviously not appropriate for families where the phenotype is not under Mende-
lian control. This approach should be used cautiously, however, as the degree of support for one model over another is often very small for any one family, i.e., most families have limited ability to discriminate between competing models of inheritance.

INCORPORATING COVARIATES INTO A GENERAL MODEL OF INHERITANCE

Observed covariates (age, gender, exposures to environmental risk factors) may influence the phenotype, and these factors should be considered as part of the genetic analysis. Effects of observed covariates can be modeled as a fixed effect, where the influence of the covariate on the phenotype is constant among all individuals. If this situation exists, it is even possible to use a two-stage approach where multiple linear regression is used prior to the genetic analysis to adjust for covariate(s), and then residual or adjusted values are used to fit models of inheritance. While this two-stage approach is widely used, some effort should be made to appropriately rescale estimated penetrance parameters of the final model into meaningful values (i.e., in the original scale of measurement), and the relative importance of the covariates along with genetic components should be provided as a percent of the original variance.

It is also possible to directly incorporate a regression effect into the penetrance parameters as part of a general model. For example, a quantitative phenotype \( Y \) can be modeled as a linear function of a single covariate with a genotype specific intercept, i.e.,

\[
Y = \mu_i + \beta X + e, \tag{6}
\]

where \( \mu_i \) is the genotypic mean for the \( i \)-th genotype, \( X \) is the observed covariate, \( \beta \) is the corresponding regression coefficient, and \( e \) is a normally distributed residual error term with mean of zero and variance of \( \sigma^2_e \). Assuming these components are independent, the total variance of \( Y \) can be written as

\[
\sigma^2_Y = \sigma^2_{mg} + \beta^2 \sigma^2_z + \sigma^2_e, \tag{7}
\]

where the variance due to the major gene is computed as \( \sigma^2_{mg} = \sum f_i (\mu_i - \mu)^2 \) for \( i = 1\ldots I \) genotypes of frequencies \( f_i \), \( \mu \) is the grand mean, and \( \sigma^2_z \) is the observed variance of \( X \).

Models that incorporate both a major gene component and covariate effects along with residual familial correlations (either as arbitrary parent-offspring, sib, and spouse correlations, or as a polygenic component) are appealing as long as the list of covariates is short. The number of parameters fit often becomes a problem. For traits influenced by many covariates, which may require considerable modeling to check for inter-actions among the different covariates and possible nonlinear effects, the number of parameters becomes excessive. Investigators must justify inclusion of covariates and balance this against the practical constraints of numerical estimation.

GENE BY COVARIATE INTERACTION

The more interesting situation of gene-environment interaction arises, however, when the effect of the covariate depends on the individual's unobserved genotype. In quantitative genetics, gene-environment interaction is defined as genotypic instability across environments (22), implying that the expression of the gene depends on the observable environment. The quantitative phenotype \( Y \) with a single covariate \( X \) is now modeled as

\[
Y = \mu_i + \beta X + e, \tag{8}
\]

where both the means (\( \mu_i \)) and the regression coefficient (\( \beta \)) differ among genotypes. Under this model, the total variance of \( Y \) can be broken down as

\[
\sigma^2_Y = \sigma^2_{mg} + \sum f_i \beta^2_i \sigma^2_z + \sigma^2_e = \sigma^2_{mg} + \sigma^2_{ge} + \sigma^2_e. \tag{9}
\]

Thus, the total genetic component is composed of variation due to genotypic effects (\( \sigma^2_{mg} \)) plus that due to the interaction of genotype with the covariate (\( \sigma^2_{ge} \)). Tiret et al. (23) showed how ignoring this type of gene-environment interaction can lead to gross underestimation of the genetic effects on a quantitative trait, and can compromise tests of hypotheses about Mendelian control.

This form of gene-environment interaction is quite plausible for many biologic traits. Moll et al. (24) illustrated this in a large French-Canadian pedigree segregating for a Mendelian dyslipidemia, familial hypercholesterolemia (FH). Individuals identified as carriers for this FH gene not only showed a markedly higher baseline level of total serum cholesterol and low-density-lipoprotein cholesterol (LDL-c), but the regression slope of low-density-lipoprotein cholesterol with age was much greater in carriers compared with their noncarrier relatives. Ignoring this genotype specific effect of age, or adjusting for age using linear regression as the first of a two stage analysis would have lead to underestimating the impact of this FH allele on the quantitative phenotype of LDL-c.

When the major locus component is poorly defined, the consequences of not considering genotype-specific regression coefficients can be even greater. For example, Perusse et al. (25) used data from young families...
in Minnesota to test for possible Mendelian control of systolic blood pressure. In this study, age and gender were covariates, and a polynomial regression with age and age^2 terms was considered. When distinct genotypic means were considered with common regression coefficients, it was not possible to distinguish between genetic and nongenetic models. When regression coefficients for age were allowed to depend on genotype and gender, however, there was a significant improvement in the fit of the model and evidence for Mendelian control of systolic blood pressure was seen. The best fitting model of inheritance predicted that not only would the mean of systolic blood pressure increase with age, but its variance would also increase as genotypic means diverged with age. Thus, this putative major gene accounted for a very small proportion of the total variance among young children (1 percent among females and 6 percent among males at age 5 years) but a much larger amount among adults (61 percent among females and 55 percent among males at age 50 years). Furthermore, these results show that the common two-stage approach, where adjustment for covariates with multiple linear regression is conducted prior to the actual genetic analysis, can easily conceal evidence of Mendelian control of quantitative traits.

Mixed Mendelian models (those with major gene components and residual polygenic components) can be extended to multiple loci by expanding the number of genotypes, transmission parameters, and allele frequencies. Because the number of genotypes increases from three in a single-locus model to nine in a two-locus model, patterns of the penetrance parameters can become quite complex (26). Nonetheless, if these multiple penetrance functions change across different environmental conditions, it suggests gene-environment interaction. While this is difficult to document with human data, animal studies allow better control over the environment. A good example is in the work on apolipoprotein A1 levels in baboons under two dietary environments (basal diet and high cholesterol, saturated fat diet) done by Blangero et al. (27). Analysis of apolipoprotein A1 levels under these two diets showed 1) adding a second major locus improved the fit of the model compared with a single major locus model, 2) both major gene and polygenic components varied with diet, and 3) the pattern of genotypic means also varied with diet; in particular, while the genotype with highest apolipoprotein A1 levels stayed high under both diets, other predicted genotypic means decreased under the high-fat diet while some increased. Thus, segregation analysis revealed evidence of interaction between major-locus genotypes and the environment in these baboon families.

**COMPONENTS OF VARIANCE AND RESIDUAL CORRELATIONS**

It may be important to consider the effects of covariates on variability within a genotype, especially in cases where there may be gender-specific variance. Prenger et al. (17) conducted an analysis of apolipoprotein A1 in families ascertained through a proband at risk for premature heart disease, and found evidence of major gene control of apolipoprotein A1 levels but with significant heteroscedasticity between men and women. This study employed the two-stage analytic design, adjusting for a number of covariates (including gender) before the actual genetic analysis. This regression model did adjust for differences in the means of males and females, but did not affect the gender-specific variance. When gender-specific variances were examined as part of a series of models of inheritance, there was a significant difference between males and females. In the best fitting model of inheritance, the putative major gene component represented 49 percent of the variance among males and 37 percent among females.

Covariates such as gender may also be important in determining patterns of residual familial correlations. Typically, the mixed model considers a simple polygenic component underlying all residual correlation, but it is possible that the additive genetic component to this residual heritability could be influenced by gender. The study of genetic control of fat mass as measured by electrical bioimpedance, conducted by Comuzzie et al. (28), illustrates this situation. In this study of Mexican American families, there was evidence of gender differences in the genotypic means at a major locus. This putative locus accounted for 37 percent of the variance in males and 42 percent in females, and there was an additional gender difference in the residual polygenic component (18 percent in males and 35 percent in females). Thus, fat mass appeared to be under much greater total genetic control in females than in males (77 percent versus 55 percent). Analysis of another measure of obesity, body mass index, has also suggested similar major gene models with genotype specific gender effects (29).

**INTEGRATING INFORMATION FROM GENETIC MARKERS**

The premise of the genetic analysis described here is that an unobservable genotype controls phenotypic expression, possibly by interacting with an observable covariate. This covariate could be a genetic marker, and if so, the proportion of variation due to a "measured genotype" can be computed. If a large proportion of phenotypic variation can be attributed to allelic
differences at a marker locus, it could mean that the marker locus plays a causal role in controlling the phenotype or that it is tightly linked to (and in linkage disequilibrium with) a locus that does. Frequently, the marker is a polymorphism in or near to the gene coding for a protein whose serum levels are being measured. For example, the e2-e3-e4 polymorphism in the apolipoprotein E (apoE) locus accounts for most of the observed variation in apolipoprotein E levels (30), and polymorphic variation in the number of kringle repeats in the Lp(a) gene accounts for the majority of the variation in serum lipoprotein(a) levels (31).

Models combining information from both trait and the marker must specify 1) the recombination fraction between the unobserved trait locus and the observed marker locus and 2) some measure of gametic disequilibrium (i.e., the conditional probability of each allele at the trait locus given each marker allele). For example, Tiret et al. (32) conducted a segregation analysis of angiotensin I converting enzyme (ACE) levels in eight nuclear families and found evidence of a codominant major gene determining angiotensin I converting enzyme levels. When a polymorphic insertion/deletion marker in intron 16 of the ACE gene was used as the measured genotype, this marker alone accounted for 28 percent of the variance in angiotensin I converting enzyme levels. Fitting a combined model showed strong (but not complete) disequilibrium between the trait locus and this insertion/deletion marker, and a total of 44 percent of the variation in angiotensin I converting enzyme levels was attributed to a major locus near this marker. Analysis of another sample of 44 extended Jamaican families suggested that a second trait locus may exist, independent of the trait locus linked to the ACE gene marker (33). Together these two loci accounted for 79 percent of the observed variation in angiotensin I converting enzyme levels.

A VARIANCE COMPONENTS APPROACH FOR GENETIC MARKERS

Amos (34) developed a variance components approach that permitted a major gene component, covariate effects and a residual polygenic component to be combined with observed markers at candidate genes. The linear model is written as

\[ Y = \mu_i + g + \beta X + e, \]  

(10)

where the unobservable "polygenotype" \( g \) is normally distributed with mean zero and additive genetic variance \( \sigma^2_p \). The expected phenotype for any one individual is \( E(Y) = \mu_i + \beta X \), and the variance can be partitioned into

\[ \text{Var}(Y) = \sigma^2_{mg} + \sigma^2_p + \beta^2 \sigma^2_i + \sigma^2_r. \]  

(11)

To simplify decomposition of the major gene component (\( \sigma^2_{mg} \)), assume there is no covariate effect (i.e., \( \beta = 0 \)). While the polygenic component \( \sigma^2_p \) is assumed to be due to additive effects of many independent genes, the major gene component can be partitioned into its own additive and dominance components (i.e., \( \sigma^2_{mg} = \sigma^2_a + \sigma^2_d \)). Falconer (35) showed these two major gene components are functions of their allele frequency (\( p q = 1 - p \)) and the relative phenotypic values of the three genotypes. Specifically, if \( \mu_1 = +a \), \( \mu_2 = -a \), and \( \mu_3 = d \) on a phenotypic scale where the overall mean (\( \mu \)) is 0, then \( \sigma^2_a = 2pq(a - d(p - q))^2 \) and \( \sigma^2_d = 4p^2q^2d^2 \). This allows the covariance between any two relatives \( L \) and \( M \) to be written as

\[ \text{Cov}(Y_i, Y_m) = \begin{cases} \sigma^2_d + \sigma^2_a + \sigma^2_p + \sigma^2_i & \text{if } l = m \\ 2\Phi_{lm} \sigma^2_a + \Delta_{lm} \sigma^2_d + 2\phi_{lm} \sigma^2_p & \text{if } l \neq m \end{cases} \]  

(12)

where \( \Phi_{lm} \) is the kinship coefficient for the pair, \( \Delta_{lm} \) is the probability of both alleles at the major locus being identical by descent (36). This clearly shows the confounding between major gene and polygenic components under this general linear model.

To expand this model to incorporate observable marker loci, consider a marker linked to the major gene with recombination fraction \( \Theta \), where the probability of sharing marker alleles identical by descent for any pair of relatives \( L \) and \( M \) is denoted \( \pi_{lm} \). Given this observable identical by descent sharing, the covariance becomes

\[ \text{Cov}(Y_i, Y_m | \pi_{lm}) = \begin{cases} \sigma^2_d + \sigma^2_a + \sigma^2_p + \sigma^2_i & \text{if } l = m \\ f(\Theta, \pi_{lm}) \sigma^2_a + g(\Theta, \Delta_{lm}) \sigma^2_d + 2\Phi_{lm} \sigma^2_p & \text{if } l \neq m \end{cases} \]  

(13)

Thus, the observed covariance between relatives can be partitioned into components due to unobserved major loci linked to observed genetic markers by using identical by descent sharing information on pairs of relatives. Amos (34) gives \( f(\Theta, \pi_{lm}) \) in his table 1 for several common types of relatives, and presents the analogous functions \( g(\Theta, \Delta_{lm}) \) for all sibs. When only sib-pairs are considered, Amos (34) points out the parameters \( \sigma^2_{mg} \) and \( \Theta \) are confounded with \( \sigma^2_p \) and \( \sigma^2_i \), although it is still possible to test hypotheses about linkage if one is willing to assume \( \Theta = 0 \) (as would occur if the marker were in or very near a candidate
gene). Amos et al. (37) note that this approach, termed “interval mapping,” can be used to test for linkage by testing the hypothesis that the variance component due to linkage to a marker is zero. They also showed that maximum likelihood estimators of these components due to linkage to a major gene can be biased, but suggested a quasi-likelihood approach similar to the generalized estimating equations discussed below.

**ALTERNATIVE APPROACHES**

One key requirement of the likelihood method is that the model must be specified in practically every detail to evaluate the ln-likelihood function. The number of parameters and their assumed distributions are all dictated as part of the model. The ln-likelihood function (given some initial values) is computed on a given dataset, and this is numerically maximized by altering the parameter values to obtain maximum likelihood estimators, and then hypotheses can be tested. The justification in assuming a particular distribution for these parameters is often mathematical convenience rather than biologic relevance. Furthermore, it is difficult, if not impossible, to test the validity of all assumptions. Thus, there are serious questions about the robustness of the likelihood approach. One alternative to likelihood based estimation is the generalized estimating equation approach proposed by Liang and Zeger (38), where the likelihood function is replaced by more tractable estimating equations based on low-order marginal distributions of observed family data (39). Generalized estimating equations approaches use equations contrasting empirical observations to their theoretic expectations based on either first moment expectations (generalized estimating equation-1) or first and second moment expectations (generalized estimating equation-2), and estimators are obtained as solutions to these equations. By relying on only low-order moments (means, variances, and covariances) rather than the complete distribution used in the full likelihood approach, the final estimators should be more robust to mis-specification of the model.

There are problems in this extension, however, some of which are intrinsic to all estimation approaches and some of which are unique to the generalized estimating equations method. First, there is the difficult question of confounding among parameters, also called aliasing. Stram et al. (40) note there is complete or partial aliasing between penetrance parameters and allele frequencies when analyzing quantitative traits under certain genetic models. Essentially, it is not possible to separately estimate all parameters of a general model of inheritance from information on pairwise marginal distributions alone. Lee et al. (41) suggested a two-stage estimation process, much like a pseudo-likelihood approach, where one parameter (here, the allele frequency) is fixed during the first stage while penetrance parameters are estimated. In the second stage, the allele frequency is estimated using the estimators for penetrance parameters obtained from the first stage. Second, there is the problem of computational burden in the generalized estimating equation approach. Estimating equations have a structure determined by the size of each family: \( n \) means, \( n \) variances, and \( n(n - 1)/2 \) covariance terms. This places computational limits on the size of families which can be used. Lee and Stram (42) point out, however, opportunities exist for minimizing the dimensionality needed for larger families by grouping different types of relationships into unique categories. Zhao (43) suggests that, in most situations, the information provided by distant relatives can be ignored with little loss of statistical efficiency.

In spite of these problems, there is considerable promise in these alternative methods. Whittemore and Gong (39) propose score test statistics based on generalized estimating equations that can be used to test hypotheses and compare a series of models, including nongenetic models and single-locus models for qualitative phenotypes. Their simulation studies suggest a generalized estimating equations approach can be adapted for family studies based on epidemiologic designs, and they propose a hierarchical approach to identifying the best model of inheritance. Lee and Stram (42) suggest that some of the problems in uniquely identifying parameters of a general model of inheritance can be overcome by extending the generalized estimating equations approach to include third moments, and propose a multistage iterative procedure. While the full potential of the generalized estimating equations approach remains to be defined, given the shallow understanding of the biologic basis of many traits now being used to fit general models of inheritance with likelihood methods, it is only prudent to further develop more robust methods.

**CONSTRAINTS ON INCORPORATING ENVIRONMENT INFORMATION INTO GENETIC MODELS**

There are several important limits on incorporating environmental factors into genetic models of inheritance; some are intrinsic to family studies, some are analytic. The natural age structure of families creates distinctive patterns of censoring: younger generations can provide little or no information on late onset diseases, and older generations will be enriched for survivors who have escaped serious diseases. Thus, adult sibships will provide the most information about diseases of middle to late life; younger sibships will
provide the most information about childhood disorders. The natural age structure of pedigrees also has implications for phenotypes that show secular changes in risk. Dramatic changes in prevalence over time is the first clue that environmental factors are more important in disease etiology, since changes in the genetic constitution of a population occur extremely slowly. Nonetheless, diseases which do show a change in prevalence may have a major genetic component to their susceptibility since the combination of genes and environmental factors is critical. However, some accommodation should be made for interpreting phenotypes across generations. An affected individual in an earlier birth cohort or an older generation may not represent the same pathologic process as an affected individual in a later birth cohort or a younger generation. Adjustment of a quantitative trait for covariates may need to similarly consider the generation of individuals, but this is confounded with age and cohort effects. In the analysis of extended families, the definition of generation can be somewhat arbitrary, since the age range of distant relatives can be very wide (e.g., first cousins can span a large age range).

A more practical limitation involves the amount of information available on family members, and how missing data are handled in the analytic model. The natural progression of family studies is to ascertain the family through an affected proband (or one with an extreme phenotype), collect information on close relatives of the proband (first through the proband’s own report, then by direct contact), then extend the family (either selectively or exhaustively) until information on all available relatives is collected. Clearly, the proband and his/her immediate relatives will be more motivated to participate in a research study and will provide the most complete data. Distant relatives will have an increasing amount of missing data, and only second-hand information will be available on many relatives (especially those who are dead). Any model considering covariates, such as in equations 6, 8, or 10, must make some accommodation for missing values. If individuals with missing data are dropped when the regression coefficients are estimated (as is done in most programs), the only individuals included in the final model will be those with complete data, and these may be a selected fraction of the total. In particular, the ln-likelihood function of a model with the covariate will be computed on this smaller number of individuals with complete data, and could be much larger (n.b., look as if it fit the data better). Careful examination of output from existing programs can prevent major misinterpretation, but sometimes the limitation is more subtle. For example, in a family study of congenital malformations, a reasonable covariate would be history of maternal miscarriage collected from interview of the mother. While it is reasonable to expect such data on sibs of an affected proband, this covariate would not be available for either parent (unless grandmothers of probands were interviewed). Thus, including this covariate effectively limits the analysis to sibships only.

Thus, it is not surprising that most work on covariate effects is currently limited to readily available observations such as age and gender. At least these covariates can be reliably obtained on all family members. Any strategy to incorporate more detailed covariates must balance the new information gained with the constraints of availability. Nonetheless, the biologic complexity of multifactorial diseases compels us to push forward statistical models to combine information on both genetic and environmental factors into a more comprehensive analysis.

REFERENCES

16. Hopper JL, Mathews JD. Extensions to multivariate normal