Genetic Epidemiology 6:319–332 (1989)

Estimation of Genetic Model Parameters: Variables Correlated With a Quantitative Phenotype Exhibiting Major Locus Inheritance

Sandra J. Hasstedt and Patricia P. Moll

Department of Human Genetics, University of Utah, Salt Lake City (S.J.H.); Departments of Epidemiology and Human Genetics, University of Michigan, Ann Arbor (P.P.M.)

A major locus that is detected through its effect on one phenotype (a primary trait) may also affect other quantitative phenotypes or qualitative disease endpoints (secondary traits). The pattern of effects of the major locus on a set of primary and secondary traits suggests candidate defects for the mutant allele. The effects are directly estimable when "measured genotypes" or a tightly linked marker allow unambiguous assignment of major locus genotypes. When genotype assignments are ambiguous for a major locus detected through its effect on a quantitative primary trait, we propose estimators using genotypic probabilities. Making certain reasonable assumptions, we demonstrate asymptotic unbiasedness of these genotypic probability estimators of the genotypic means and variances for either the quantitative primary or secondary traits, of the covariances between quantitative primary and secondary traits, and of prevalences for the secondary qualitative traits. An important application of genotypic probability estimators is to define an effect of a major locus that cannot be detected upon analysis of the variable; for example, major locus effects may be defined for hypertension or blood pressure as secondary traits, but not detected as primary traits.

Key words: maximum likelihood estimators, likelihood analysis, genotypic probability estimators, pleiotropic gene effects, bivariate phenotypes

INTRODUCTION

A quantitative phenotype (primary trait), although many steps removed from a genetic locus, may exhibit effects of segregation of alleles at the locus. Other quantita-

Received for publication September 2, 1988; revision accepted December 12, 1988.

Address reprint requests to Sandra J. Hasstedt, Ph.D., Department of Human Genetics, University of Utah Medical Center, 50 N. Medical Drive, Salt Lake City, Utah 84132.

© 1989 Alan R. Liss, Inc.

tive traits and/or qualitative disease endpoints (secondary traits) may also exhibit effects of the same locus. This pleiotropic gene action reflects the highly integrated state of cellular and developmental metabolism. The magnitude of the effect on either a primary or secondary trait depends on the trait's distance from the primary gene action and the number of intervening steps influenced by other genes and environmental factors. While a correlation observed between two traits (primary or secondary) may result from a single pleiotropic locus, the correlation could also result from the same environmental exposure and/or other pleiotropic genetic loci. Attributing the source of the correlation may contribute more to understanding the underlying physiology of the locus than will a study restricted to the locus and a single primary trait.

Standard procedures allow estimation of a major locus effect when each individual can be unequivocally assigned a genotype, that is, the locus has "measured genotypes" [Sing and Davignon, 1985; Boerwinkle et al., 1986a, 1987; Boerwinkle and Sing, 1986; Hewett-Emmett et al., 1987; Boerwinkle and Utermann, 1988]. Genotypes can also be assigned using a genetic marker linked tightly to the locus of interest [Leppert et al., 1988]. A DNA sequence polymorphism detected by a cloned human gene sequence generally defines a marker with recombination below 10^{-4}, corresponding to a distance between the polymorphism and the site of mutation within the locus of a few thousand base pairs [Kan and Dozy, 1978]. However, since the polymorphism defining the marker differs from the hypothesized mutation, it is possible that unaffected family members may share the identifying marker allele seen in affected family members. When this occurs, a genotypic probability that is close to zero or one can be used to identify unaffected or affected family members, respectively. The genotypic probability equals the relative likelihood of the genetic model conditional on a particular genotype for the individual; the parameters of the genetic model are fixed at their maximum likelihood estimates including no recombination between the phenotype and marker loci. The probabilities therefore consider familial relationships and both phenotype and marker data. Assigning each individual as a heterozygote or homozygote based on his/her genotypic probability, Leppert et al. [1986] could estimate the effect of the low-density lipoprotein (LDL) receptor locus on the secondary traits, the lipid levels.

Upon inferring a major locus from likelihood analysis in the absence of a tightly linked marker, genotypes can he assigned using genotypic probabilities computed using familial relationships and the primary trait data, an approach called the "unmeasured genotype" or biometrical approach [Boerwinkle et al., 1986a; Sing et al., 1988]. Odenheimer [1985] tested the accuracy of genotype assignments using this approach for single locus genetic models and different pedigree configurations (nuclear families through large multigeneration pedigrees) using computer simulation. He assigned the carrier and noncarrier genotypes to individuals whose probability was above and below a given cutpoint, respectively. Sensitivity and specificity were defined as the proportion of individuals with a correctly assigned genotype among carriers and noncarriers, respectively. He found that when the genotypic means differed by at least three within genotype standard deviations, the inheritance was dominant, and when data were available on at least a five-member nuclear family (two parents and three offspring), both sensitivity and specificity exceeded .90 for a wide range of cutpoints. When the genotypic means differed by only two within genotype standard deviations, no cutpoint gave both sensitivity and specificity over .70 for any pedigree configuration consid-

ered, indicating that at least 30% of carriers would be classified as noncarriers or at least 30% of noncarriers would be classified as carriers. Therefore, the unmeasured genotype approach classifies individuals as carriers or noncarriers of a rare allele with reasonable accuracy only for dominant traits with genotypic means that differ by at least three within genotype standard deviations. However, for recessive traits and/or traits with means closer than three within genotype standard deviations, the unmeasured genotype approach misclassifies many individuals. A modified strategy of eliminating individuals with intermediate probabilities would reduce the sample size. However, either misclassification or elimination of individuals may produce biased estimates that could diminish large effects or exaggerate small effects of the locus.

Genotypes can be assigned unequivocally for major loci with measured genotypes, a tightly-linked marker, or nonoverlapping phenotypes within genotypes. However, for the majority of major loci of interest, we cannot accurately assign genotypes to individuals. With this in mind, we define estimators that use the genotypic probabilities from the unmeasured genotype approach as weighting factors. Intuitively, the genotypic probability estimators (GPEs) partially assign an individual to a given genotype. Thus, an individual with genotypic probabilities of .85, .14, and .01 for genotypes *dd*, *Dd*, and *DD* would contribute 85% of an observation to genotype *dd*, 14% to genotype *Dd*, and 1% to genotype *DD*.

In the present paper, we consider the effect on a quantitative or qualitative secondary trait of a major locus identified through a large effect on a quantitative primary trait. The designations as primary and secondary depend on the application and do not imply any specific causal relationships. We define the GPEs of the genotypic frequencies, means, and variances of primary or secondary traits, covariances between a primary and secondary trait, and prevalences of secondary traits. We then present applications of the GPEs to studies of apolipoprotein B [Hasstedt et al., 1987], intraerythrocytic sodium level [Hasstedt et al., 1988a], and sodium–lithium countertransport [Hasstedt et al., 1988b]. Finally, we state and discuss the assumptions made when deriving the expected values of the estimators.

THE GENOTYPIC PROBABILITY ESTIMATORS

Suppose that x_i, $i = 1, 2, \ldots, n$, represents a set of quantitative observations for the primary trait for which major locus inheritance was demonstrated using likelihood analysis and that observations y_i, $i = 1, 2, \ldots, n$ (quantitative) and z_i, $i = 1, 2, \ldots, n$ (dichotomous) are secondary traits available for the same sample. X, Y, and Z represent the corresponding random variables. Suppose further that genotypic probabilities of the J genotypes, $\hat{p}_{ij}$,

$$\sum_{j=1}^{J} \hat{p}_{ij} = 1,$$

have been computed as the relative likelihood conditional on individual i having genotype j with the parameters fixed at their maximum likelihood estimates for the primary trait X. MLE is used for both maximum likelihood estimators and estimates, and GPE is used for both genotypic probability estimators and estimates.

Genotypic Frequencies

If genotypes were unambiguously assigned, the genotypic frequency would be estimated as the number of sample members assigned the genotype divided by the sample size; for ambiguously assigned genotypes, we estimate the numerator by summing the portion of each individual assigned the genotype. Therefore, the GPE of the frequency of genotype j equals

$$\hat{f}_j = \sum_{i=1}^{n} \hat{p}_{ij}/n. \tag{1}$$

We can demonstrate that the GPE $\hat{f}_j$ equals the MLE of the genotypic frequency, f_j.

Genotypic Means

If genotypes were unambiguously assigned, the genotypic mean would be estimated as the sum of trait values for sample members assigned the genotype divided by the number assigned the genotype; for ambiguously assigned genotypes, we estimate the numerator as the sum of each trait value weighted by the portion of the individual assigned the genotype and the denominator as the sum of the portion of each individual assigned the genotype. Therefore, the GPE of the mean of the secondary trait variable Y for genotype j equals

$$\hat{\mu}_j = \sum_{j=1}^{n} \hat{p}_{ij}\, y_i/\hat{n}_j, \tag{2}$$

where $\hat{n}_j = n\,\hat{f}_j$. The weighted sum of $\hat{\mu}_j$ over the genotypes equals the sample mean, that is,

$$\sum_{j=1}^{J} \hat{f}_j\hat{\mu}_j = \sum_{j=1}^{J} \hat{f}_j \sum_{i=1}^{n} \hat{p}_{ij} y_i/\hat{n}_j = \sum_{i=1}^{n} y_i/n = \bar{y},$$

since

$$\sum_{j=1}^{J} \hat{p}_{ij} = 1.$$

We can demonstrate that GPE $\hat{\mu}_j$ is an asymptotically unbiased estimator of μ_j, that is, lim E $(\hat{\mu}_j) = \mu_j$. The variance of $\hat{\mu}_j$ is bounded between σ^2_{wg}/n_j and $\sigma^2_{wg}/n_j + (1 - f_j)\,\mu_j^2/n_j$, where σ^2_{wg} is the variance within genotypes. In addition, substituting x_i for y_i in Equation (2), the GPE equals the MLE of the genotypic mean for the primary trait variable X.

Variance Within Genotypes

If genotypes were unambiguously assigned, the variance within genotypes would be estimated as the sum of the squared deviations for sample members assigned the genotype divided by the number assigned the genotype; for ambiguously assigned geno-

types, we estimate the numerator as the sum of the squared deviations weighted by the portion of the individual assigned the genotype and the denominator as the sum of the portion of each individual assigned the genotype. Therefore, the GPE of the variance of the secondary trait variable Y within major locus genotype j equals

$$\hat{\sigma}^2_{wgj} = \sum_{i=1}^{n} \hat{p}_{ij} (y_i - \hat{\mu}_j)^2 / \hat{n}_j \tag{3}$$

and the GPE of the common variance within major locus genotypes equals

$$\hat{\sigma}^2_{wg} = \sum_{i=1}^{n} \sum_{j=1}^{J} \hat{p}_{ij} (y_i - \hat{\mu}_j)^2 / n. \tag{4}$$

We can demonstrate that both estimators underestimate σ^2_{wg}, but are asympototically unbiased. In addition, substituting x_i and μ_{xj} for y_i and μ_j, respectively, in Equations (4), the GPE equals the MLE of the within genotype variance of the primary trait variable X.

Variance Due to the Major Locus

The GPE of the variance for secondary trait Y due to the major locus is

$$\hat{\sigma}^2_{ml} = \sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_j - \bar{y})^2. \tag{5}$$

We can demonstrate that GPE $\hat{\sigma}^2_{ml}$ overestimates σ^2_{ml}, but is asymptotically unbiased.

As an alternative to the GPE, an ad hoc estimator of the proportion of variance of the secondary trait Y due to the major locus equals the product of the proportion of the variance of Y accounted for by the primary trait X (the squared correlation between X and Y) multiplied by the proportion of the variance of X due to the major locus [Sing et al., 1986]. We assume a constant expected slope of Y to X (that is, the relationship between X and Y is independent of the genotype for the primary trait X) when demonstrating asymptotic unbiasedness of the GPEs. We can demonstrate that this assumption also leads to equivalence of the ad hoc estimator and the parameter estimated.

Covariance Within Genotypes

If genotypes were unambiguously assigned, the covariance within genotypes would be estimated as the sum of the product of deviations for sample members assigned the genotype divided by the number assigned the genotype; for ambiguously assigned genotypes, we estimate the numerator as the sum of each deviation product weighted by the portion of the individual assigned the genotype and the denominator as the sum of the portion of each individual assigned the genotype. Therefore, the GPE of the covariance between primary trait X and secondary trait Y within major locus genotype j equals

$$\hat{\gamma}_{wgj} = \sum_{i=1}^{n} \hat{p}_{ij} (x_i - \hat{\mu}_{xj}) (y_i - \hat{\mu}_j) / \hat{n}_j, \tag{6}$$

and the GPE of the common covariance within genotypes equals

$$\hat{\gamma}_{wg} = \sum_{i=1}^{n} \sum_{j=1}^{J} \hat{p}_{ij} (x_i - \mu_{xj}) (y_i - \hat{\mu}_j)/n \tag{7}$$

where $\hat{\mu}_{xj}$ is the MLE of the mean of primary trait variable X for genotype j. We can demonstrate that both estimators are asymptotically unbiased estimators of γ_{wg}.

Covariance Due to the Major Locus

The GPE of the covariance between primary trait X and secondary trait Y due to the major locus is

$$\hat{\gamma}_{ml} = \sum_{j=1}^{J} \hat{f}_j (\hat{\mu}_{Xj} - \bar{x}) (\hat{\mu}_j - \bar{y}). \tag{8}$$

We can demonstrate that this estimator is an asymptotically unbiased estimator of γ_{ml}.

Prevalences

If genotypes were unambiguously assigned, the genotypic prevalence would be estimated as the number of affected sample members assigned the genotype divided by the number assigned the genotype; for ambiguously assigned genotypes, we estimate the numerator by summing the portion of all affected individuals assigned the genotype and the denominator by summing the portion of all individuals assigned the genotype. Therefore, the GPE of the disease prevalence for qualitative secondary trait Z for genotype j is

$$\hat{\phi}_j = \sum_{i=1}^{n} \hat{p}_{ij} z_i/\hat{n}_j, \tag{9}$$

where z_i equals 1 if individual i is affected and 0 otherwise. The weighted sum of $\hat{\phi}_j$ across genotypes equals the sample proportion, that is,

$$\sum_{j=1}^{J} \hat{f}_j \hat{\phi}_j = \sum_{j=1}^{J} \hat{f}_j \sum_{i=1}^{n} \hat{p}_{ij} z_i/\hat{n}_j = \sum_{i=1}^{n} z_i/n.$$

We can demonstrate that GPE $\hat{\phi}_j$ is an asymptotically unbiased estimator of ϕ_j with variance bounded by $(1 - \phi_j)/n_j$ and $\phi_j (1 - f_j \phi_j)/n_j$.

EXAMPLES

To estimate the effect of a major locus on a quantitative secondary trait, we can compute GPEs of its genotype-specific means, the GPE of the proportion of its variance due to the major locus, or the GPE of its correlation with the primary trait due to the locus. To estimate the effect of a major locus on a qualitative secondary trait, we can compute GPEs of the genotype-specific prevalences or GPEs of genotypic fre-

quencies within disease subgroups. To assess the fit of the genetic model, we can compute GPEs of the genotypic frequencies within age and sex groups. We demonstrate these applications of the GPEs for three unmeasured loci inferred by likelihood analysis. In each of these examples, the primary and secondary traits are thought to be related metabolically. We also present ad hoc estimates of the proportion of variance due to the major locus; differences between the GPE and ad hoc estimates are attributable to the violation of the constant expected slope assumption.

Apolipoprotein B

A codominant major locus with two alleles accounted for 43.2% of the variance in apolipoprotein B (apoB) level, a primary trait, as measured on 331 members of 36 pedigrees [Hasstedt et al., 1987]. MLEs of the genotypic means equaled 110.5, 141.9, and 208.1 mg/dl. As the primary protein constituent of low-density lipoprotein (LDL), apoB showed the expected high correlation with LDL cholesterol, a secondary trait ($r = .62$, $P < .01$). GPEs of the genotypic means of LDL cholesterol equaled 118.8, 141.0, and 191.5 mg/dl. The GPE of the variance due to the apoB locus was estimated to be 14% of the variance in LDL cholesterol, somewhat lower than the ad hoc estimate of 17% obtained by multiplying .432 (the proportion of the variance of apoB due to the major locus) by .384 (correlation .62 squared). Using either estimate, the unmeasured locus determining apoB levels explained more of the variability in LDL cholesterol level than did the LDL receptor locus (2.5%) [unpublished data] or the apoE locus (4.4%) [Boerwinkle and Sing, 1986].

Intraerythrocytic Sodium

A major locus accounted for 29.0% of the variance in intraerythrocytic sodium (RBC Na) level after adjustment and natural logarithm transformation in a sample of 1,800 normotensive members of 64 pedigrees [Hasstedt et al., 1988a]. Homozygotes for each of the four alleles in the genetic model expressed a different mean; heterozygotes expressed the lower of the two means associated with homozygotes for their two alleles. MLEs of the four genotypic means of the standardized natural logarithm-transformed RBC Na levels equaled -2.60, -0.21, 1.37, and 4.39 in standard deviation units. GPEs of the means in the original scale (mmole/liter RBC) equaled 4.32, 6.67, 9.06, and 12.19. Here we are using transformed RBC Na levels as a primary trait and untransformed RBC Na levels as a secondary trait.

RBC Na level is higher in hypertensive patients than in controls [Hilton, 1986]. Hypertensive members of our sample were removed before analysis because hypertensive medication affects RBC Na level. However, pedigrees ascertained through hypertensive probands might have a higher frequency of the alleles that elevate RBC Na level. GPEs of the summed frequency of the three genotypes for high RBC Na ranged from 0.135 to 0.170 in pedigrees ascertained through a hypertensive proband and from 0.061 to 0.103 in pedigrees ascertained through a normotensive proband [Hasstedt et al., 1988a]. Therefore, the GPEs provided evidence that when the proband is hypertensive, rather than normotensive, there is increased susceptibility for hypertension among normotensive relatives.

Various sodium transport systems regulate the sodium concentration in the cell, producing a correlation between transport variables (secondary traits) and RBC Na. Estimated correlations with RBC Na in normotensives were -0.43 ($P < .001$) for the

number of ouabain binding sites, -0.08 ($P < .001$) for sodium–potassium cotransport, and -0.04 ($P < .05$) for sodium–lithium countertransport. GPEs of the correlation with RBC Na due to the major locus equaled -0.12 for the number of ouabain-binding sites, -0.05 for sodium-potassium cotransport, and -0.02 for sodium–lithium countertransport [Hasstedt et al., 1988a]. Since the correlation with RBC Na due to the major locus is close to zero for all these secondary traits, the GPEs failed to implicate genes affecting the number of ouabain-binding sites, sodium–potassium cotransport, or sodium–lithium countertransport as responsible for the effect on RBC Na level.

Sodium–Lithium Countertransport

A codominant major locus with two alleles accounted for 34.4% of the variance in sodium–lithium countertransport (SLC) level in a sample of 1,880 members of 89 pedigrees [Hasstedt et al., 1988b]. In the analysis, equation $y = 6/P[(x/6 + 1)^P - 1]$ transformed the standardized levels of SLC (x in the equation) by a power (P) to obtain a variable y with an approximate normal distribution within genotypes [MacLean et al., 1976]. The MLEs of the genotypic means of transformed SLC equaled -0.27, -0.08, and 2.24 in units of standard deviates. If the relationship between the moments of x and y, assuming y is normally distributed, were known, we could transform the genotypic means to the original scale. In the absence of this relationship, GPEs of the genotypic means in the original scale (mmole/liter RBC/hr) were computed to equal 0.245, 0.286, and 0.530. Here we are using the transformed SLC as a primary trait and untransformed SLC as a secondary trait.

If SLC is a preclinical marker of essential hypertension [Canessa et al., 1980], the SLC major locus may be a hypertension susceptibility locus. Using SLC as the primary trait and hypertension as the secondary trait, the GPEs of hypertension prevalence equaled 0.04 and 0.09 in men aged 30 to 49 years, 0.23 and 0.36 in men aged 50 years and older, and 0.38 and 0.82 in women aged 50 years and older, in low and high homozygotes, respectively [Hasstedt et al., 1988b]. Therefore, the homozygote that expressed elevated SLC also had increased prevalence of hypertension.

Although the biological basis is unknown, SLC correlates with weight and triglyceride levels (.27 and .35, respectively [Hunt et al., 1986]). GPEs of the variances due to the SLC major locus were 1.93 and 4.02% of the variance of secondary traits weight and triglyceride levels, respectively [Hasstedt et al., 1988b]. Using the correlation computed on a different sample, the corresponding ad hoc estimates equal 2.5% for weight and 4.2% for triglyceride, both larger than the corresponding GPE.

High values of SLC occur less frequently in juveniles than in adults, suggesting delayed expression of elevated levels [Williams et al., 1983]. Similar GPEs of the genotypic frequencies within different age groups refute this suggestion and support the fit of the genetic model [Hasstedt et al., 1988b].

THE ASSUMPTIONS

The assumptions made to demonstrate asymptotic unbiasedness of the GPEs are 1) the observations are independent and identically distributed (iid), 2) the variance and correlation within genotypes are each constant across genotypes, 3) the primary phenotype for which major locus inheritance has been demonstrated and the secondary phenotype are described by a mixture of bivariate normal distributions, 4) the parame-

ters of the major locus model have been estimated by maximum likelihood in a large sample, and 5) the expected slope of the primary to the secondary trait is constant across the genotypes. These assumptions are not used in computing the GPEs, but only in demonstrating asymptotic unbiasedness. Each of these assumptions is described and discussed below.

Independent and Identically Distributed (iid) Observations

The likelihood of a major locus model on a primary quantitative trait contains for each pedigree member M a sum over all genotypes of a genotype-specific probability [Elston and Stewart, 1971]. M's genotype-specific probability is the product of two terms: 1) M's probability of genotype j and 2) the probability of M's phenotype conditional on genotype j. If the sample does not include M's parents, M's probability of genotype j equals the corresponding population genotypic frequency, f_j; if M is an offspring of pedigree members, M's probability of genotype j depends on M's parent's genotypes, inducing dependence in the observations. Upon assuming iid observations, M's probability of genotype j equals f_j for all pedigree members regardless of the inclusion of M's parents in the sample. Since M's genotype-specific probability depends only on M's genotype, each sum in the likelihood is independent of all others. Therefore, by assuming iid observations, we ignore the dependence between pedigree members due to the major locus; if the model includes polygenic inheritance, we also ignore that source of dependence between observations.

Because the likelihood function differs for every pedigree structure, obtaining MLEs requires maximizing the likelihood numerically. By assuming iid observations, we can obtain analytical MLEs by taking derivatives of the likelihood function.

A genotypic probability $\hat{p}_{ij}$ equals a ratio with denominator the complete likelihood and with numerator the likelihood excluding all genotypes for individual i other than genotype j. Assuming iid observations, $\hat{p}_{ij}$ equals f_j multiplied by the penetrance divided by the sum over all genotypes for that individual, since probabilities for other pedigree members cancel. Although the assumption of iid observations is obviously violated in this application, the equations are intractable without it. The only way to insure independence is to restrict the estimation to unrelated individuals with the resulting reduction in sample size. In a sample that includes related individuals, the GPEs need to be interpreted with caution since two familial phenotypes may both occur within a pedigree by chance alone, but have a large effect on the GPE if both are expressed in multiple family members.

Common Variance and Correlation

A common variance within major locus genotypes has been routinely assumed since the first implementation of the mixed model [Morton and MacLean, 1974]. When a positively skewed variable is transformed to approximate normality prior to analysis, the assumption of a common variance for the transformed variable corresponds to a larger variance for larger means in the original scale.

Likewise, a common correlation within major locus genotypes was assumed in the implementation of a bivariate mixed model [Morton et al., 1983]. However, the correlation between total serum cholesterol and triglyceride varied with the genotype at the apoE locus [Boerwinkle et al., 1987; Boerwinkle and Utermann, 1988]. When the genotype is unmeasured, using genotype-specific GPEs of the variances and corre-

lations will provide more tests of the validity of the assumption. However, since constant within genotype variances and correlations were assumed in demonstrating asymptotic unbiasedness, significant differences may result in biased estimates.

Bivariate Normal Mixture

We assume random variables X and Y are distributed as a bivariate normal density within each major locus genotype. Random variable Z assumes the values zero and one for values of Y below and above a threshold, respectively. Departures from normality generally affect estimation procedures little.

Maximum Likelihood Estimation in a Large Sample

We assume MLEs have been obtained for the parameters f_j, μ_{Xj}, and σ_{Xwg} in a large sample. Since MLEs are consistent and asymptotically unbiased [Edwards, 1972], in a large sample they approximately equal their parameter values. In deriving the expectations of the GPEs, the parameters f_j, μ_{Xj}, and σ_{Xwg} are substituted for their MLEs, with the justification that n, the sample size, is large.

Constant Expected Slope

We assume that the expected slope of Y to X does not depend on genotype. In other words, the two variables are directly related through their levels, rather than indirectly through the genotypic or environmental effects. Someone with a high level of the primary trait X will tend to have a high (or low for negative correlation) level of the secondary trait Y, regardless of whether the high level of X is due to the major locus, other genetic factors, or environmental factors. One consequence of this assumption is that the dominance relationships between the genotypes will be the same for all primary and secondary traits affected by the major locus.

DISCUSSION

Estimation of the effect of a genetic locus on a variable is straightforward when genotypes at the locus can be unambiguously assigned [Sing and Davignon, 1985; Boerwinkle and Utermann, 1988]. When genotypes cannot be unambiguously assigned and the major locus was revealed using likelihood analysis of a quantitative phenotype (primary trait), we propose GPEs to estimate the effect of the locus on secondary traits (other quantitative phenotypes or qualitative disease endpoints). We demonstrated asymptotic unbiasedness of the GPEs. However, the demonstration required many assumptions and the effect of violating those assumptions has yet to be explored. Estimates obtained using ambiguous genotype assignments probably reflect the ambiguity to some degree. Nevertheless, while recognizing the limitations of the GPEs, the estimates they produce may provide useful information about the nature of the genetic variability revealed through its effect on secondary traits. In particular, this approach can add insight about traits like hypertension and blood pressure for which single locus inheritance cannot be demonstrated [Sing et al., 1988].

The GPE approach consists of two stages; estimates for the primary trait are obtained in a univariate analysis, then estimates for the secondary trait are computed using genotypic probabilities. In a similar two-stage approach, Sing et al. [1986] and Boerwinkle et al. [1986b] assumed major locus and polygenic inheritance, respectively, first esti-

mated the genetic percentage of the variance in the primary trait, then multiplied the estimate by the percentage of variability in the secondary trait accounted for by variability in the primary trait to obtain an estimate of the genetic percentage of the variance in the secondary trait. In their applications, the same individuals were not measured for both the primary and secondary traits.

Pleiotropic effects have also been estimated from the simultaneous analysis of two (or more) traits. In such analyses, both traits contribute information about the inheritance and the estimates may differ from estimates produced by a univariate analysis of the primary variable and from GPEs for the secondary variable. Assuming polygenic inheritance, the effects of genes on multiple quantitative traits have been simultaneously estimated using factor analysis [Martin and Eaves, 1977], bivariate variance components analysis [Boehnke et al., 1986], and bivariate path analysis [Moll et al., 1978; Colletto et al 1981; Darlu et al., 1982; McGue, 1983; McGue et al., 1983; Vogler, 1985; Vogler and DeFries, 1985]. Another path analytic approach assumed a direct effect of one trait on another at the phenotypic level and requires longitudinal data [Hanis, 1981; Hanis et al., 1983]. While these approaches attribute correlation between variables to polygenic and shared environmental factors, none are free of assumptions or include a major locus.

However, a major locus (with or without polygenes) has been included in some bivariate analyses, thereby allowing estimation of the effect of the locus on both variables. Models have been applied to two quantitative traits [Elston et al., 1975; Morton et al., 1983; Williams et al., 1986] and to a quantitative trait and a disease endpoint [Morton and MacLean, 1974; Lalouel et al., 1985]. Simultaneous analyses of multiple traits often have the additional goal of increasing the amount of information to identify the major locus. That is the goal when analyzing a linear combination of variables constructed prior to analysis [Morton et al., 1978] or maximized in the analysis [Elston et al., 1975] or a linear combination of principal components of the variables in an approach termed pedigree discriminant analysis [Goldin et al., 1980].

Although these other approaches may have advantages over GPEs, they assume a priori relationships among the traits not required for GPEs. In addition, all of these approaches require many of the same assumptions as for the GPE approach and are not, in general, practical if estimates are desired for many pairs of traits.

Another application of GPEs in addition to estimating the effect of a major locus on secondary variables is to assess the fit of the genetic model to the primary variable. For example, the frequencies predicted by the genetic model for the total sample can be compared to GPEs of the frequencies computed within age and sex groups. Since the assumption of Hardy–Weinberg equilibrium implies equal frequencies for both sexes and all ages, age trends or sex differences in the GPEs of the frequencies suggest lack of fit to the genetic model. Similarly, age trends or sex differences in the GPEs of the variances or genotypic means may indicate a violation of the assumption of the genetic model that these are the same for all ages and sexes. Likewise, different GPEs of the variance within genotypes could indicate a violation of the assumption of a common variance.

Statistical tests would enhance the usefulness of the GPE approach. Tests of interest include the fit of the genetic model to the primary trait, the validity of the assumptions made for the GPE approach, and the significance of the effect of the major locus on the secondary trait. Testable assumptions of the GPE approach include a constant

expected slope and equality of the within genotype variances and correlations across genotypes. Traditional statistical tests cannot be used without demonstrating that the distribution theory of analysis of variance applies when individuals are only partially assigned to classes. In lieu of using existing distributional theory, one could develop randomization tests [Edgington, 1987]. Such tests have been applied to estimates of pedigree data to assign significance levels and to guide in interpretation [Karlin and Williams, 1984; Schwartz et al., 1988]. For example, one could permute the genotypic probabilities 1,000 times among the individuals in the study and recompute the GPE each time. The significance level equals the proportion of permutations for which the GPE exceeds (or is less than) the original GPE. We used this approach to test the significance of GPEs of the mean and prevalence of the genotype of interest [Hasstedt et al., 1989a; 1989b].

From a practical viewpoint, the GPEs have the advantages of simplicity of form and application. In form, the GPEs correspond to sample estimators, except that individuals are only partially assigned to a given class. Application of the GPE approach to many variables requires only a simple computer program and minimal computer time once the genotypic probabilities are available. Computer programs such as PAP [Hasstedt, 1988] are available to compute the genotypic probabilities as well as to perform the major locus analysis.

ACKNOWLEDGMENTS

The authors thank Dr. Michael Boehnke and Ms. Sharon Reilly for reading and commenting on the approach and manuscript. This research was supported by National Institutes of Health grants HL 24855 (SJH), HD 17463 (SJH), HL24489 (PPM), and HL39107 (PPM).

A detailed presentation of the derivation of the estimators is available upon request.

REFERENCES

Boehnke M, Moll PP, Lange K, Weidman WH, Kottke BA (1986): Univariate and bivariate analyses of cholesterol and triglyceride levels in pedigrees. Am J Med Genet 23:775–792.

Boerwinkle E, Sing CF (1986): Bias of the contribution of single-locus effects to the variance of a quantitative trait. Am J Hum Genet 39:137–144.

Boerwinkle E, Chakraborty R, Sing CF (1986a): The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50:181–194.

Boerwinkle E, Turner ST, Weinshilboum R, Johnson M, Richelson E, Sing CF (1986b): Analysis of the distribution of erythrocyte sodium lithium countertransport in a sample representative of the general population. Genet Epidemiol 3:365–378.

Boerwinkle E, Utermann G (1988): Simultaneous effects of the apolipoprotein E polymorphism on apolipoprotein E, apolipoprotein B, and cholesterol metabolism. Am J Hum Genet 42:104–112.

Boerwinkle E, Visvikis S, Welsh D, Steinmetz J, Hanash SM, Sing CF (1987): The use of measured genotype information in the analysis of quantitative phenotypes in man. II. The role of apolipoprotein E polymorphism in determining levels, variability, and covariability of cholesterol, betalipoprotein, and triglycerides in a sample of unrelated individuals. Am J Med Genet 27:567–582.

Canessa M, Adragna N, Solomon HS, Connolly TM, Tosteson DC (1980): Increased sodium-lithium countertransport in red cells of patients with essential hypertension. N Engl J Med 302:772–776.

Colletto GMDD, Krieger H, Magalhaes JR (1981): Estimates of the genetical and environmental determinants of serum lipid and lipoprotein concentrations in Brazilian twins. Hum Hered 31:232–237.

Darlu P, Rao DC, Henrotte JG, Lalouel JM (1982): Genetic regulation of plasma and red cell magnesium concentrations in man. I. Univariate and bivariate path analysis. Am J Hum Genet 34:874–887.

Edgington ES (1987): "Randomization Tests," 2nd Ed. New York: Marcel Dekker, Inc.

Edwards AWF (1972): "Likelihood." Cambridge: Cambridge University Press.

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.

Elston RC, Stewart J (1971): A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542.

Goldin LR, Elston RC, Graham JB, Miller CH (1980): Genetic analysis of von Willebrand's disease in two large pedigrees: A multivariate approach. Am J Med Genet 6:279–293.

Hanis CL (1981): Multivariate models for human genetic analyses: Development and application to systolic blood pressure and weight. PhD Dissertation, University of Michigan, Ann Arbor.

Hanis CL, Sing CF, Clarke WR, Schrott HG (1983): Multivariate models for human genetic analysis: Aggregation, coaggregation, and tracking of systolic blood pressure and weight. Am J Hum Genet 35:1196–1210.

Hasstedt SJ (1989): PAP: Pedigree Analysis Package, Rev. 3. Department of Human Genetics, University of Utah, Salt Lake City.

Hasstedt SJ, Wu L, Williams RR (1987): Major locus inheritance of apolipoprotein B in Utah pedigrees. Genet Epidemiol 4:67–76.

Hasstedt SJ, Hunt SC, Wu LL, Williams RR (1988a): The inheritance of intraerythrocytic sodium level. Am J Med Genet 29:193–203.

Hasstedt SJ, Wu LL, Ash KO, Kuida H, Williams RR (1988b): Hypertension and sodium-lithium countertransport in Utah pedigrees: Evidence for major locus inheritance. Am J Hum Genet 43:14–22.

Hasstedt SJ, Ramirez ME, Kuida H, Williams RR (1989a): Recessive inheritance of a relative fat pattern (submitted for publication).

Hasstedt SJ, Wu LL, Kuida H, Williams RR (1989b): Recessive inheritance of a high number of sodium pump sites associated with obesity, hypertension and diabetes. Am J Med Genet, (in press).

Hewett-Emmett D, Bertin TK, Hanis CL (1987): A micromethod for typing the human apolipoprotein E polymorphism: Effects of apo E phenotypes on levels of and correlations between apolipoproteins, lipoproteins and lipids in Mexican-American females (Abstract). Am J Hum Genet 41:A7.

Hilton PJ (1986): Cellular sodium transport in essential hypertension. N Engl J Med 314:222–229.

Hunt SC, Williams RR, Smith JB, Ash KO (1986): Associations of three erythrocyte cation transport systems with plasma lipids in Utah subjects. Hypertension 8:30–36.

Kan YW, Dozy AM (1978): Polymorphism of DNA sequence adjacent to human β-globin structural gene: Relationship to sickle mutation. Proc Natl Acad Sci USA 75:5631–5635.

Karlin S, Williams PT (1984): Permutation methods for the structured exploratory data analysis (SEDA) of familial trait values. Am J Hum Genet 36:873–898.

Lalouel JM, Le Mignon L, Simon M, Fauchet R, Bourel M, Rao DC, Morton NE (1985): Genetic analysis of idiopathic hemochromatosis using both qualitative (disease status) and quantitative (serum iron) information. Am J Hum Genet 37:700–718.

Leppert MF, Hasstedt SJ, Holm T, O'Connell P, Wu L, Ash O, Williams RR, White R (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.

Leppert M, Breslow JL, Wu L, Hasstedt S, O'Connell P, Lathrop M, Williams RR, White R, Lalouel J-M (1988): Inference of a molecular defect of apolipoprotein B in hypobetalipoproteinemia by linkage analysis in a large kindred. J Clin Invest 82:847–851.

MacLean CJ, Morton NE, Elston RC, Yee S (1976): Skewness in commingled distributions. Biometrics 32:695–699.

Martin NG, Eaves LJ (1977): The genetical analysis of covariance structure. Heredity 38:79–95.

McGue M (1983): Bivariate path analysis of plasma lipids. Hum Hered 33:145–152.

McGue M, Rao DC, Reich T, Laskarzewski P, Glueck CJ, Russell JM (1983): The Cincinnati Lipid Research Clinic Family Study. Bivariate path analyses of lipoprotein concentrations. Genet Res Camb 42:117–135.

Moll PP, Sing CF, Brewer GJ, Gilroy TE (1978): Multivariate analysis of the genetic effects on red cell

blood glycolysis. In Brewer GJ (ed): "The Red Cell." Progress in Clinical and Biological Research, Vol 21. New York: Alan R. Liss, Inc., pp 385–405.

Morton NE, MacLean CJ (1974): Analysis of family resemblance. III. Complex segregation of quantitative traits. Am J Hum Genet 26:489–503.

Morton NE, Gulbrandsen CL, Rhoads GG, Kagan A, Lew R (1978): Major loci for lipoprotein concentrations. Am J Hum Genet 30:583–589.

Morton NE, Rao DC, Lalouel JM (1983): "Methods in Genetic Epidemiology." New York: Karger, pp 96–97.

Odenheimer D (1985): An evaluation of complex segregation analysis in identifying an individual's genotype at a major locus. Ph.D. Dissertation, University of Michigan, Ann Arbor.

Schwartz AG, Boehnke M, Moll PP (1988): The family risk index as a measure of familial heterogeneity of cancer risk: A population based study in metropolitan Detroit. Am J Epidemiol 128:524–535.

Sing CF, Davignon J (1985): Role of the apolipoprotein E polymorphism in determining normal plasma lipid and lipoprotein variation. Am J Hum Genet 37:268–285.

Sing CF, Boerwinkle E, Turner ST (1986): Genetics of primary hypertension. Clin Exp Hypertens [A] 8:623–651.

Sing CF, Boerwinkle E, Moll PP, Templeton AR (1988): Characterization of genes affecting quantitative traits. In Weir BS, Eisen EJ, Goodman MM, Namkoong G (eds): "Proceedings of the Second International Conference on Quantitative Genetics." Sunderland, MA: Sinauer, pp 250–269.

Vogler GP (1985): Multivariate path analysis of familial resemblance. Genet Epidemiol 2:35–53.

Vogler GP, DeFries JC (1985): Bivariate path analysis of familial resemblance for reading ability and symbol processing speed. Behav Genet 15:111–121.

Williams RR, Hunt SC, Kuida H, Smith JB, Ash KO (1983): Sodium-lithium countertransport in erythrocytes of hypertension prone families in Utah. Am J Epidemiol 118:338–344.

Williams RR, Hasstedt SJ, Wilson DE, Ash KO, Yanowitz FF, Reiber GE, Kuida H (1986): Evidence that men with familial hypercholesterolemia can avoid early coronary death. An analysis of 77 gene carriers in four Utah pedigrees. JAMA 255:219–224.

Edited by D.C. Rao