Ann. Hum. Genet. (1986), **50**, 181–194
Printed in Great Britain

The use of measured genotype information in the analysis of quantitative phenotypes in man

I. Models and analytical methods

BY E. BOERWINKLE,* R. CHAKRABORTY† AND C. F. SING*
* *Department of Human Genetics, University of Michigan, Ann Arbor, Michigan 48109-0015*
† *Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77225*

SUMMARY

Improved laboratory methods allow one to investigate the contribution of measured allelic variability at a locus physiologically involved in determining the expression of a quantitative trait. We present statistical methods that incorporate measured genotype information into the analysis of a quantitative phenotype that allows one simultaneously to detect and estimate the effects of a measured single locus and residual polygenic effects. Likelihoods are presented for the joint distribution of the quantitative phenotype and a measured genotype that are appropriate when the data are collected as a sample of unrelated individuals or as a sample of nuclear families. Application of this method to the analysis of serum cholesterol levels and the concentration of the group specific component (Gc) are presented. The analysis of the contribution of the common *Gc* polymorphism to the determination of quantitative variability in Gc using samples of related and unrelated individuals presents, for the first time, the simultaneous estimation of the frequencies and the effects of the genotypes at a measured locus, and the contribution of residual unmeasured polygenes to phenotypic variability.

INTRODUCTION

The study of the genetics of a quantitative phenotype in humans can be viewed as utilizing one of two basic strategies. The first strategy we refer to as the unmeasured genotype or biometrical approach. This approach utilizes information about the distribution of the phenotype among related individuals to estimate genetic parameters. The second strategy uses information about the biochemical basis of the phenotype to identify loci physiologically involved in the etiology of a quantitative phenotype, and then determines the contribution of genetic variability at these loci to variability of the trait. We refer to this second strategy as the measured genotype approach. Employing such a strategy can yield direct detailed information about the genetic architecture of a quantitative trait: the number of loci involved, the frequencies and effects of their alleles and the type of loci (i.e. structural genes or regulatory genes). The measured genotype approach has at least two requirements: first, an understanding of the biology of the phenotype that enables one to identify the candidate genes that may be contributing to phenotypic variability, and secondly, the ability to measure allelic variability at the identified loci. By employing methods presented here that relate this measured genetic variability to phenotypic variability, the genetic architecture of a quantitative phenotype can begin to be elucidated.

Because of the shortage of mutational variation at loci involved in the metabolism of

interesting quantitative phenotypes, the measured genotype approach has had little impact on human quantitative genetics in the past. One classic example is the work by Hopkinson and his colleagues (Hopkinson *et al.* 1964) relating variability in red-cell acid phosphatase activity to electrophoretic variability in the same protein. With improved ability to measure variants at loci with known biological roles, the measured genotype approach will begin to play a larger role in studying the genetics of human quantitative phenotypes.

We present here the maximum-likelihood methods for the measured genotype approach to investigating the genetics of a quantitative phenotype and compare this method to the unmeasured genotype approach. We consider data collected as a sample of unrelated individuals or as a sample of nuclear families. The likelihood that jointly parameterizes the measured genotype and the quantitative phenotype collected as a sample of related individuals allows one, for the first time, to estimate simultaneously the frequency and the effects of the genotypes at the measured single locus, and the contribution of residual unmeasured polygenic effects to the total phenotypic variability. An example application of this method using the group-specific component (Gc) is presented. Using a sample of related individuals, we estimate the contribution of the common *Gc* polymorphism and residual polygenic effects to the variability in plasma Gc concentration.

METHODS

Modelling the quantitative phenotype

We begin by defining the general model for the quantitative phenotype of the ith individual as an additive combination of effects,

$$y_{ij} = \mu_j + G_i + E_i, \tag{1}$$

where μ_j is the mean of the jth genotype at a single locus ($j = 1 \ldots J$). If there are two alleles at the locus, j ranges from one to three. If individuals have been typed at more than one locus involved in the determination of the quantitative trait, multilocus information can be included in the measured genotype likelihood function. f_j is the relative frequency of the jth genotype in the general population and $\Sigma_{j=1}^{J} f_j = 1$. We define G_i as the effect of the polygenotype of the ith individual. It is assumed to be the consequence of the action of a large number of genes, each with small effects, acting additively and independently. E_i represents the totality of all environmental effects specific to the ith individual. We assume that G and E are random effects, each normally distributed in the population with mean 0 and variance σ_G^2 and σ_E^2, respectively. The components of the model are assumed to be uncorrelated and combine additively. The parameters of the model to be estimated include the $J-1$ genotype frequencies, the J genotype-specific means, the polygenic variance and the environmental variance. Other genetic and environmental random effects, such as dominance or a common environment, that could affect the phenotypic variance among individuals with the same single locus genotype and the covariance between individuals, may be added to the model with straightforward extensions.

We next present the likelihood functions and the strategies for estimating the parameters of the model when either the unmeasured or measured genotype approaches are considered, and when the data are collected either as a sample of unrelated individuals or as a sample of nuclear families. Sampling is assumed to be random throughout our presentation, so that observations from different sampling units are uncorrelated.

Unrelated individuals – unmeasured genotype

Here we consider the likelihood function for observations taken on a random sample of individuals drawn from a mixture of normal distributions. This function has been utilized by geneticists to determine if there is evidence in a sample of individuals for an underlying mixture of distributions, possibly due to the effect of a single unmeasured locus (Lalouel *et al.* 1983*a*; Turner *et al.* 1985). Unfortunately, the literature on the admixture model applied in human genetics is disperse. An admixture model has been presented by Day (1969) that parameterizes a mixture of two multivariate normal distributions. Hasselblad (1966) presented a model that classifies the data into groups and allows for unequal variances within the component distributions. Neither parameterization adequately addresses the applications found in human genetics. For completeness, we present here the likelihood function and parameter estimation approach for the admixture model that has been applied to human quantitative phenotypes.

Consider a sample $y_1 \ldots y_n$ drawn from a mixture of J normal distributions. We let n be the total number of individuals in a given sample and n_j the number of individuals of genotype j. The distribution of the ith observation, conditional on the jth genotype at the single locus, is normal, with mean μ_j and variance σ^2 ($\sigma^2 = \sigma_G^2 + \sigma_E^2$) for all $j = 1 \ldots J$. This distribution has density ζ_{ij}, where

$$\zeta_{ij} = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{\frac{-1}{2\sigma^2}(y_{ij}-\mu_j)^2\right\}. \tag{2}$$

The conditional distribution of the phenotype, given a particular genotype, is referred to as the penetrance function. The unconditional likelihood of the ith observation considers all possible genotypes weighted by their relative frequencies and is written

$$\zeta_i = \sum_{j=1}^{J} f_j \zeta_{ij}. \tag{3}$$

When the observations are randomly sampled and uncorrelated, the likelihood of the sample is the product of the likelihoods for the individuals.

The maximum likelihood estimates of the parameters of the model can be found using the Newton–Raphson method. This scheme, a version of a more general EM algorithm previously used in admixture problems (Dempster *et al.* 1977), uses the iterative relationship

$$\boldsymbol{\Theta}^{s+1} = \boldsymbol{\Theta}^s - \mathbf{H}(\boldsymbol{\Theta}^s)^{-1}\mathbf{D}(\boldsymbol{\Theta}^s), \tag{4}$$

where $\boldsymbol{\Theta}^s$ is the vector of parameter estimates on the sth iteration, $\mathbf{D}(\boldsymbol{\Theta}^s)$ is the vector of first partial derivatives of the likelihood evaluated at $\boldsymbol{\Theta}^s$, and $\mathbf{H}(\boldsymbol{\Theta}^s)$ is the matrix of second partial derivatives evaluated at $\boldsymbol{\Theta}^s$. The values of $\boldsymbol{\Theta}$ that maximize the likelihood function are taken to be the maximum likelihood estimates. An estimate of the asymptotic variance–covariance matrix of the parameter estimates for a sample of data is $-\mathbf{H}^{-1}$ evaluated at the maximum likelihood estimates. The first derivatives of the natural logarithm of the sample likelihood function, denoted Z, are given in equations (5).

$$\frac{\partial Z}{\partial \mu_j} = \sum_{i=1}^{n} \frac{1}{\zeta_i} f_j \zeta_{ij} \frac{y_{ij}-\mu_j}{\sigma^2}, \qquad j = 1, \ldots, J \tag{5a}$$

$$\frac{\partial Z}{\partial f_j} = \sum_{i=1}^{n} \frac{1}{\zeta_i}(\zeta_{ij} - \zeta_{iJ}), \qquad j = 1, \ldots, J-1 \tag{5b}$$

$$\frac{\partial Z}{\partial(\sigma^2)} = \sum_{i=1}^{n} \frac{1}{\zeta_i} \sum_{j=1}^{J} f_j \zeta_{ij} \left\{ \frac{-1}{2\sigma^2} + \frac{(y_{ij}-\mu_j)^2}{2\sigma^4} \right\}. \tag{5c}$$

The second partial derivatives are given by

$$\frac{\partial^2 Z}{\partial \mu_j \partial \mu_{j'}} = \sum_{i=1}^{n} \frac{-1}{\zeta_i^2} f_{j'} f_j \zeta_{ij'} \zeta_{ij} \frac{(y_{ij}-\mu_j)(y_{ij}-\mu_{j'})}{\sigma^4} + \delta_{jj'} \sum_{i=1}^{n} f_j \zeta_{ij} \left(\frac{(y_{ij}-\mu_j)^2}{\sigma^4} - \frac{1}{\sigma^2} \right) \tag{6a}$$

$$\frac{\partial^2 Z}{\partial \mu_j \partial f_{j'}} = \sum_{i=1}^{n} \frac{-1}{\zeta_i^2} f_j \zeta_{ij} \frac{(y_{ij}-\mu_j)(\zeta_{ij'}-\zeta_{iJ})}{\sigma^2} + (\delta_{jj'} - \delta_{jJ}) \sum_{i=1}^{n} \frac{1}{\zeta_i} \zeta_{ij} \frac{y_{ij}-\mu_j}{\sigma^2} \tag{6b}$$

$$\frac{\partial^2 Z}{\partial \mu_j \partial(\sigma^2)} = \sum_{i=1}^{n} \frac{1}{\zeta_i} f_j \zeta_{ij} (y_{ij}-\mu_j) \left(\frac{-1}{\sigma^2} - \frac{1}{2\sigma^2} + \frac{(y_{ij}-\mu_j)^2}{2\sigma^4} \right) - \left\{ \frac{1}{\zeta_i^2} f_j \zeta_{ij} \frac{y_{ij}-\mu_j}{\sigma^2} \sum_{j=1}^{J} f_j \zeta_{ij} \left(\frac{-1}{2\sigma^2} + \frac{y_{ij}-\mu_j}{2\sigma^4} \right) \right\} \tag{6c}$$

$$\frac{\partial^2 Z}{\partial f_j \partial f_{j'}} = \sum_{i=1}^{n} \frac{-1}{\zeta_i^2} (\zeta_{ij} - \zeta_{iJ})(\zeta_{ij'} - \zeta_{iJ}) \tag{6d}$$

$$\frac{\partial^2 Z}{\partial f_j \partial(\sigma^2)} = \sum_{i=1}^{n} \frac{1}{\zeta_i} \left(\zeta_{ij} \left(\frac{-1}{2\sigma^2} + \frac{y_{ij}-\mu_j}{2\sigma^4} \right) - \zeta_{iJ} \left(\frac{-1}{2\sigma^2} + \frac{(y_{ij}-\mu_J)^2}{2\sigma^4} \right) \right) - \left\{ \frac{1}{\zeta_i^2} \left(\sum_{j=1}^{J} f_j \zeta_{ij} \left(\frac{-1}{2\sigma^2} + \frac{(y_{ij}-\mu_j)^2}{2\sigma^4} \right) \right) (\zeta_{ij} - \zeta_{iJ}) \right\} \tag{6e}$$

$$\frac{\partial^2 Z}{\partial(\sigma^2)^2} = \sum_{i=1}^{n} \frac{1}{\zeta_i} \left\{ \sum_{j=1}^{J} f_j \zeta_{ij} \left(\frac{1}{2\sigma^4} - \frac{(y_{ij}-\mu_j)^2}{\sigma^6} + \left(\frac{-1}{2\sigma^2} + \frac{(y_{ij}-\mu_j)^2}{2\sigma^4} \right)^2 \right) \right\} + \frac{-1}{\zeta_i^2} \left\{ \sum_{j=1}^{J} f_j \zeta_{ij} \left(\frac{-1}{2\sigma^2} + \frac{(y_{ij}-\mu_j)^2}{2\sigma^4} \right) \right\} \tag{6f}$$

where j and j' each range from 1 to J. The Kronecker delta function, $\delta_{jj'}$, is equal to one when j is equal to j', and zero otherwise.

Tests of hypotheses may be carried out using a likelihood ratio criterion. The test statistic, Λ, is the ratio of the value of the likelihood function maximized under the full model (H_1) to the maximum of the likelihood function under some reduced model (H_0) in which one or more of the parameters is restricted to a hypothesized value. Because the exact distribution of Λ is not known, a chi-square approximation to the distribution of Λ is utilized to establish whether the value(s) of the restricted parameter(s) deviates significantly from the hypothesized value(s). When H_0 is true, the value of $-2 \ln \Lambda$ is distributed approximately as a chi-square with degrees of freedom equal to the number of parameters constrained to a hypothesized value.

Unrelated individuals – measured genotype

Here we consider the situation where the value of the quantitative phenotype and the single-locus genotype are measured for each member of a sample of unrelated individuals. This approach has been used in plant and animal genetics, and to a lesser extent in human genetics, to investigate the effect of variability at a measured locus on a quantitative phenotype and to estimate the genotype frequencies. This application is equivalent to the single-classification

linear statistical model used in the analysis of variance (Neter & Wasserman, 1974; Scheffé, 1959) where the classifications are determined by the genotypes at the measured locus. Estimates of the genotype frequencies are obtained by counting the number of individuals with a specific genotype. Estimates of the means of the measured single-locus genotypes are given by the average values of individuals sharing the same genotype. The best estimate of the random variance component is obtained from the within-class error mean square of the analysis of variance. The F ratio from the analysis of variance can be used to test the hypothesis of equal genotype means.

One of the first applications of the measured genotype approach in humans studied the relationship between red cell acid phosphatase (RCAP) activity and the discrete electrophoretic types of the same enzyme. Hopkinson *et al.* (1964) found mean differences in RCAP activity among the five different genotypic classes determined by electrophoresis of the RCAP protein. We have reported similar examples involving the Gc protein (Daiger *et al.* 1984), *Apo E* and cholesterol (Sing & Davignon, 1985), and *α-glycerophosphate* and a vector of glycolytic intermediates in *Drosophila* (Clark *et al.* 1983).

The experimental design in each of these four examples (*RCAP*, *Gc*, *Apo E*, *α-GPD*) addresses the effects of the measured locus only and does not provide an estimate of the contribution of other unmeasured loci that may also affect the phenotype. We consider below likelihoods that parameterize a sample of observations collected on a set of related individuals, so that one may simultaneously estimate the relative frequencies and effects of genotypes at a single locus, and the contribution of unmeasured polygenic loci to variability in the phenotype.

Related individuals – unmeasured genotype

When the data are collected on a sample of related individuals it is possibe to construct a likelihood function that includes parameters that define the contribution of unmeasured shared genes to the phenotypic correlations between relatives. The distribution of the phenotype among a sample of related individuals is used to obtain information about the values of the parameters of the model. Evidence supporting an unmeasured single locus with a large effect comes from the distribution of the phenotype in the population and from the transmission of the phenotype from parents to their offspring. Analysis of the distribution of a quantitative phenotype among members of a pedigree for evidence of an underlying large single-gene effect has been termed complex segregation analysis (for a review see Elston, 1980).

The likelihood function developed by Elston & Stewart (1971) or the alternative parameterizations developed by Morton & MacLean (1974) and Lalouel *et al.* (1983*b*) may be used to relate the distribution of the quantitative phenotype in a sample of nuclear families to the parameters of the model. A chi-square approximation to the likelihood ratio test provides a criterion for assessing support for competing hypotheses. This model has been used extensively to investigate the genetic contribution to a quantitative phenotype (Boerwinkle *et al.* 1984; Moll *et al.* 1984; Lalouel *et al.* 1983*a*).

Related individuals – measured genotype

By sampling groups of related individuals and measuring values of the quantitative phenotype and the genotypes at a single locus known to be involved in the quantitative trait,

one can partition the overall genetic variability into a proportion attributable to the measured single locus and a proportion due to the segregation of unmeasured polygenes. The joint likelihood of observing the phenotypes and the measured single-locus genotypes on the members of a family can be partitioned as

$$L(\mathbf{y} \text{ and } \mathbf{j}) = L(\mathbf{j})\, L(\mathbf{y}|\mathbf{j}), \tag{7}$$

where $\mathbf{y}$ and $\mathbf{j}$ are the vectors of phenotypes and measured single-locus genotypes, respectively. $L(\mathbf{y}|\mathbf{j})$ is the likelihood of observing the phenotypes on each member of the family conditional on their single-locus measured genotypes and is a function of the means and variance components of the model. $L(\mathbf{j})$ is the likelihood of observing the genotypes of the family members and is a function of the measured single-locus genotype frequencies. From the factorization theorem (Bickel & Doksum, 1977), the information in the data about the measured single-locus genotype frequencies is all contained in $L(\mathbf{j})$, and the information in the data about the means and variance components of the model is all contained in $L(\mathbf{y}|\mathbf{j})$.

The likelihood of the measured genotypes in a nuclear family can be partitioned into the likelihood of observing the genotypes of the parents and the product of likelihoods for the genotypes of the children conditional on the genotypes of the parents. The value of these likelihoods depends on the single-locus genotypes of the parents and Mendelian transmission probabilities. Assuming random mating, the likelihood of observing a certain mating type is the product of the likelihoods for each parent. The likelihood of the constellation of measured genotypes in a nuclear family with C children is

$$L(\mathbf{j})' = L(j_m, j_f, j_{c-1} \cdots j_{c-C}) = f_{j_m} f_{j_f} \prod_{c=1}^{C} L(j_c | j_m j_f) \tag{8}$$

where j_m, j_f and j_c are the measured genotypes of the mother, father and children, respectively. To illustrate an application, we consider the four-member pedigree and data given in Fig. 1. Assuming Hardy–Weinberg equilibrium and Mendelian transmission, the likelihood of the constellation of measured genotypes is $p^3q/2$. Maximum-likelihood estimates of the relative genotype frequencies from a sample of nuclear families may be obtained by counting the number of parents with each genotypic class and dividing by the total number of parents.

The second part of the likelihood function given in equation (7) considers the distribution of the phenotypes among pedigree members, conditional on their measured genotypes. Lange *et al.* (1976*b*) and Smith (1980) parameterized the distribution of a quantitative phenotype measured on related individuals as a multivariate normal distribution but did not include measured genotype information. When measured genotype information is available a multivariate normal parameterization can be extended to estimate genotype means and other genetic and environmental parameters, as suggested by Moll *et al.* (1979) and Hopper & Mathews (1982). Their work parameterizes the likelihood of a quantitative phenotype measured on a sample of related individuals conditional on the genotypes at the single measured locus. They do not consider the joint distribution of the quantitative phenotype and the measured genotype among pedigree members. The amount of information contained in a sample about the parameters of the model depends on whether the phenotype is modelled in the likelihood as being conditional on the measured genotype or jointly with it.

The distribution of the quantitative phenotype among family members is assumed to follow a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The natural

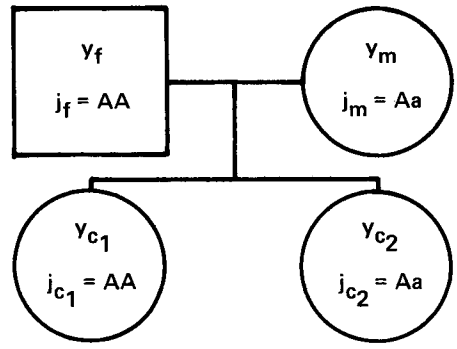

Fig. 1. Example of a four-member pedigree with quantitative phenotype data (y) and measured genotype data (j) at a hypothetical two-allele (A,a) locus. f, m, c_1 and c_2 represent the father, mother, first child and second child, respectively.

logarithm of the likelihood function of the quantitative phenotype for one pedigree conditional on their measured single-locus genotypes is given by

$$L(\mathbf{y}|\mathbf{j}) = \frac{-1}{2} \ln|\boldsymbol{\Sigma}| - \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}), \tag{9}$$

where the constant term has been omitted to simplify the notation. The dimensions of the mean vector and the covariance matrix are determined by the size of the family. The mean vector takes the values of the genotypic means determined by the measured genotypes of the individuals in the pedigree. The elements of $\boldsymbol{\Sigma}$ are functions of the variance components of the model and the transmission of the unmeasured polygenotype from parents to offspring. The conditional covariance between any two individuals i and i' is

$$\text{cov}(i, i') = 2\phi_{ii'}\sigma_G^2 + \delta_{ii'}\sigma_E^2, \tag{10}$$

where $\phi_{ii'}$ is the kinship coefficient between individuals i and i' (Malécot, 1948) and $\delta_{ii'}$ is equal to one when i is equal to i' and zero otherwise. The elements of the data vector ($\mathbf{y}$), and the parameters of the mean vector ($\boldsymbol{\mu}$) and covariance matrix ($\boldsymbol{\Sigma}$), for the family in Fig. 1 are given below.

$$\mathbf{y} = \begin{bmatrix} y_f \\ y_m \\ y_{c_1} \\ y_{c_2} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_{AA} \\ \mu_{Aa} \\ \mu_{AA} \\ \mu_{Aa} \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_G^2+\sigma_E^2 & 0 & \frac{1}{2}\sigma_G^2 & \frac{1}{2}\sigma_G^2 \\ 0 & \sigma_G^2+\sigma_E^2 & \frac{1}{2}\sigma_G^2 & \frac{1}{2}\sigma_G^2 \\ \frac{1}{2}\sigma_G^2 & \frac{1}{2}\sigma_G^2 & \sigma_G^2+\sigma_E^2 & \frac{1}{2}\sigma_G^2 \\ \frac{1}{2}\sigma_G^2 & \frac{1}{2}\sigma_G^2 & \frac{1}{2}\sigma_G^2 & \sigma_G^2+\sigma_E^2 \end{bmatrix}$$

When the data are collected from randomly selected and uncorrelated pedigrees, the likelihood of the sample is the sum of the ln likelihoods for the pedigrees.

Estimation of the variance components and the vector of genotype-specific means of the likelihood can be carried out by Fisher's scoring algorithm. If $\boldsymbol{\Theta}^s$ is the vector of parameter estimates at the sth iteration Fisher's scoring algorithm updates $\boldsymbol{\Theta}$ using

$$\boldsymbol{\Theta}^{s+1} = \boldsymbol{\Theta}^s + \mathbf{I}(\boldsymbol{\Theta}^s)^{-1}\mathbf{D}(\boldsymbol{\Theta}^s), \tag{11}$$

where $\mathbf{D}(\boldsymbol{\Theta}^s)$ is the vector of first partial derivatives of the likelihood evaluated at $\boldsymbol{\Theta}^s$ and $\mathbf{I}(\boldsymbol{\Theta}^s)$ is the information matrix evaluated at $\boldsymbol{\Theta}^s$. An estimate of the asymptotic variances and

covariances of the parameter estimates is given by the elements of the negative of the inverse of the information matrix evaluated at the maximum-likelihood estimates. Using the generalized form of the first derivatives presented by Lange *et al.* (1976*b*), the first derivatives of equation (9) with respect to each parameter are given by:

$$\frac{\partial L(\mathbf{y}|\mathbf{j})}{\partial(\sigma_G^2)} = -\tfrac{1}{2}\,\mathrm{tr}\,(\boldsymbol{\Sigma}^{-1}2\boldsymbol{\Phi}) + \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}2\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}), \tag{12a}$$

$$\frac{\partial L(\mathbf{y}|\mathbf{j})}{\partial(\sigma_E^2)} = -\tfrac{1}{2}\,\mathrm{tr}\,\boldsymbol{\Sigma}^{-1} + \tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}), \tag{12b}$$

$$\frac{\partial L(\mathbf{y}|\mathbf{j})}{\partial \mu_j} = \mathbf{c}_j'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}), \tag{12c}$$

where $\boldsymbol{\Phi}$ is the matrix of kinship coefficients among individuals in the vector $\mathbf{y}$ and $\mathbf{c}_j' = (c_{1j}\ldots c_{nj})$ is a design vector of indicator functions where $c_{ij} = 1$ if individual i has the measured single-locus genotype j and zero otherwise. These derivatives make up the elements of the score vector (**D**). The information matrix (**I**), which is the negative of the expectations of the matrix of second partial derivatives, is more difficult to obtain. The expectations must be taken over both the observed phenotype and the observed measured single-locus genotype distributions. According to the principle of conditional expectation, the expectation of a function (f) with respect to the random variables y and j can be obtained in successive expectations as $E_{yj}(f) = E_j(E_{y|j}(f))$. The expectations of the second derivatives with respect to the distribution of the observation vector, conditional on the value of the measured genotype, are:

$$E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{\partial\mu_j\,\partial\mu_{j'}}\right) = -\mathbf{c}_j'\boldsymbol{\Sigma}^{-1}\mathbf{c}_{j'}, \tag{13a}$$

$$E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{\partial\mu_j\,\partial(\sigma_G^2)}\right) = E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{\partial\mu_j\,\partial(\sigma_E^2)}\right) = 0, \tag{13b}$$

$$E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{[\partial(\sigma_G^2)]^2}\right) = -\tfrac{1}{2}\,\mathrm{tr}\,(\boldsymbol{\Sigma}^{-1}2\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1}2\boldsymbol{\Phi}), \tag{13c}$$

$$E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{\partial(\sigma_G^2)\,\partial(\sigma_E^2)}\right) = -\tfrac{1}{2}\,\mathrm{tr}\,(\boldsymbol{\Sigma}^{-1}2\boldsymbol{\Phi}\boldsymbol{\Sigma}^{-1}), \tag{13d}$$

$$E_{\mathbf{y}|\mathbf{j}}\left(\frac{\partial^2 L(\mathbf{y}|\mathbf{j})}{[\partial(\sigma_E^2)]^2}\right) = -\tfrac{1}{2}\,\mathrm{tr}\,(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}). \tag{13e}$$

The expectations of equations (13) with respect to the measured genotype (**j**) only affect the second partial derivatives with respect to the means (equation 13*a*). This expectation can be rewritten as $\mathrm{tr}\,[E_j(\mathbf{c}_j\,\mathbf{c}_j')\,\boldsymbol{\Sigma}^{-1}]$. The expectations of the joint indicator functions over the random variable j are equal to the joint probabilities that individual i will have genotype j and individual i' will have genotype j'. These probabilities are a function of the allele or genotype frequencies and the transmission probabilities of the measured single-locus genotypes. They can be obtained for general pairs of relatives using Jacquard's coefficients of identity (Jacquard, 1974) or the ITO method of Li & Sacks (1954). Some examples are: $E(c_{f_{AA}}\,c_{m_{Aa}}) = 2p^3q$, $E(c_{f_{AA}}\,c_{c_{Aa}}) = p^2q$,

where f_{AA} in the subscript refers to the father having genotype AA for example. m and c are mother and child, respectively.

The question of primary interest is to obtain estimates of the means of the measured single-locus genotypes and the variance components of the model. One may then wish to test the hypothesis that the means are all equal. This, and other hypotheses, can be tested using the chi-square approximation to the distribution of the likelihood ratio statistic.

APPLICATIONS

Sing & Davignon (1985) recently presented an application of the measured genotype approach using the *apolipoprotein E* (*Apo E*) isoelectric focusing polymorphism and a profile of lipid measures. They reported significant differences in LDL-cholesterol and total serum cholesterol among *Apo E* genotypes. We examined the same sample of data to determine if an unmeasured genotype analysis could detect the effects of variability in *Apo E*. In these data, the *Apo E* effects on these phenotypes were only detectable by considering the measured genotype information.

We next present an application of the related individuals–measured genotype likelihoods and compare the results with those obtained using an unmeasured genotype analysis on the same set of data. The sample of data consists of related individuals measured for the concentration of the group-specific protein (Gc) and the common polymorphism for *Gc* published by Daiger *et al.* (1984). The measured genotype analyses presented here on the Gc data for the first time simultaneously estimate the effects of a measured locus and unmeasured polygenes on a quantitative phenotype.

Gc is approximately 50000 molecular weight and reversibly binds vitamin D and its metabolites during transport. There is quantitative phenotypic variation among individuals for the concentration of Gc in the plasma. The *Gc* locus is polymorphic. There are two common alleles, Gc^1 and Gc^2, separated by gel electrophoresis. The common *Gc* electrophoretic polymorphism is a logical tool to begin to investigate the genetic architecture underlying variability in plasma Gc concentration.

Daiger *et al.* (1984) have investigated the effect of the *Gc* polymorphism on Gc concentration from a sample of 89 unrelated individuals. They also estimated the polygenic heritability from a sample of 44 twin pairs. The measured genotype likelihoods presented here allow one to use the combined samples of unrelated individuals and twins, both samples measured for the common *Gc* polymorphism and Gc concentration, to estimate simultaneously the single-locus genotype effects of the *Gc* electrophoretic polymorphism and the residual polygenic component on Gc concentration. We compare the results from an application of the measured genotype approach to an unmeasured genotype analysis of these same data.

Analyses were carried out on Gc concentration adjusted for the linear effects of sex and age. A complete model defining three means, one for each of the three *Gc* genotypes underlying the phenotypic distribution, and a reduced model constraining the three means equal to the same value, were fitted to these data. Both models include separate variance components for the polygenic effects and the random environmental effects. The variance components describe the variability within a mode and the covariance between pairs of individuals conditional on their *Gc* locus genotypes. From the sample of 89 unrelated individuals, the genotype frequencies for the Gc polymorphism were 0·58, 0·30 and 0·12 for the *1-1*, *1-2* and *2-2* genotypes, respectively.

Table 1. *Results of the measured genotype analysis on Gc concentration using the common Gc electrophoretic polymorphism measured on 89 unrelated individuals and 44 twin pairs*

	Models	
Parameters	1 Mode	3 Modes
$\mu_{1\text{-}1}$	30·4	31·7
$\mu_{1\text{-}2}$	—	29·5
$\mu_{2\text{-}2}$	—	26·1
σ^2_G	13·6	10·1
σ^2_E	2·5	2·5
ln likelihood	−230·3	−217·9
Difference		12·4

Table 2. *Parameter estimates and ln likelihoods from an admixture analysis for Gc concentration measured on 89 unrelated individuals*

	Models		
Parameters	1 Mode	2 Modes	3 Modes
μ_1	30·30	26·57	25·90
μ_2	—	31·63	29·71
μ_3	—	—	32·14
σ^2	16·89	11·90	11·49
ln likelihood	−170·27	−169·99	−169·99
Difference		0·28	0.0
f_1	1·0	0·264	0.182
f_2	—	0·736	0·289
f_3	—	—	0·529

Estimates of the means and variance components and the relative ln likelihoods for the measured genotype analysis are given in Table 1. The complete model, with $\mu_{11} \neq \mu_{12} \neq \mu_{22}$, fitted the data significantly better than the reduced model, with $\mu_{11} = \mu_{12} = \mu_{22}$ ($\chi^2_2 = 24{\cdot}8$, $P < 0{\cdot}001$). The maximum likelihood estimates of the parameters of the model were obtained using all of the data and the iterative relationship defined in equations (11) to (13). The estimated means of the *1-1*, *1-2* and *2-2* genotypes were 31·7, 29·5 and 26·1, respectively. The estimate of the polygenic variance component was 10·1 and the estimate of the environmental variance component was 2·5. The ratio of $\sigma^2_G/(\sigma^2_G+\sigma^2_E)$, or the residual polygenic heritability, was 0·80. The point estimates given in Table 1 are similar to those reported by Daiger *et al.* (1984). The variance of the estimates of the measured genotype means is, however, lower for the analyses presented here because we used information from all the data whereas Daiger *et al.* used only the unrelated subset of the sample. For example, the variance of the estimate of μ_{11} was 0·217 for the analysis reported here and 0·255 for the analyses reported by Daiger *et al.*

We carried out an admixture analysis on the subsample of 89 unrelated individuals to investigate whether multiple modes could be detected in the distribution of Gc levels which correspond to the multiple modes detected in the measured genotype analysis. Three models were fitted to the data corresponding to one mode, two modes and three modes underlying the distribution of Gc levels. Parameter estimates and ln likelihoods for the unmeasured genotype analysis are given in Table 2. The model with one mode could not be rejected in favour of a model with more modes. It is interesting to note that the parameter estimates for the three-mode model correspond to the *Gc* effects estimated from the measured genotype analysis, even though

no significant effect could be detected by this admixture analysis. In these data, a statistically significant effect of the electrophoretic polymorphism at the *Gc* locus on Gc concentration could only be detected by a measured genotype analysis.

DISCUSSION

One of our goals in the study of the genetics of a quantitative phenotype in humans is to understand the genetic architecture underlying phenotypic variability. By this we refer to information concerning the number of loci affecting the quantitative phenotype, the number of alleles at each locus, the frequency and the size of the effects of these alleles and the type of loci making the contribution. These issues we feel can best be addressed by measuring an individual's genotype at loci which are known *a priori* to be involved in the metabolism of the phenotype and then relating variability at these measured loci to variability in the phenotype of interest.

Fisher's seminal work of 1918 forms the basis of the biometrical approach that uses the correlations between relatives to estimate the extent to which genetic variability among individuals is contributing to the quantitative phenotypic variation in the population. Although the biometrical approach implicates the presence and quantifies the effects of unknown loci, no information about the identity and action of alleles at specific loci is obtained. The biometrical approach is limited to detecting only those single-locus effects that are relatively large (MacLean *et al.* 1975; Go *et al.* 1978; Burns *et al.* 1984). It is likely that there is a spectrum of single-gene effects on the quantitative phenotype that form a continuum from large to small. A necessary assumption about the polygenic component considered by a biometrical analysis is that there is a large number of genes contributing to the quantitative phenotype of interest. Although this assumption seems to have been accepted as biological fact, findings from experimental animal work suggest that this assumption may not be valid, at least for certain traits (Thompson, 1975; Thompson & Thoday, 1979). As for the biological nature of the polygenic variation that was investigated, in Thompson's own words: 'Polygene is a nonspecific term used to describe what is probably a very heterogeneous set of gene effects.' There does not seem to be a clear pattern of one type of gene (e.g. structural, modifier, etc.) making up the class of polygenes. Establishing the genetic component of a quantitative trait through the direct estimation of the effects of specific loci is required to clarify these issues.

One of the primary applications of information obtained by a measured genotype analysis is to partition the total phenotypic variance into the contribution of separate loci. Estimators of the proportion of the total phenotypic variance attributable to variability at a single locus that replace parameters by estimators in a ratio have been suggested (Blackwelder & Elston, 1974; Lange *et al.* 1976*a*). These estimators have been shown to be biased (Neimann-Sørensen & Robertson, 1961; Boerwinkle, 1985; Boerwinkle & Sing, 1986). They will, on the average, overestimate the proportion of the phenotypic variance attributable to the single locus. It is important when determining the action of a locus on a quantitative phenotype to distinguish between the relationships that exist between the effect of a single locus on the level of a phenotype for an individual and the contribution to the variance of the phenotype by that locus in the population at large. Loci with very large effects on the individual, such as familial hypercholesterolaemia and its effects on serum cholesterol levels (Goldstein & Brown, 1979) for

example, do not contribute greatly to the phenotypic variance in the population because of the rare frequency of the variant genotype(s). The contribution to cholesterol variability by common alleles for marker loci such as *Secretor* and others (Sing & Orr, 1976) is also small because of their very minuscule effects on individual differences. It is the class of loci that have alleles with polymorphic frequencies and moderate effects on the level of the phenotype, such as the *apolipoprotein E* and its effect on cholesterol levels (Sing & Davignon, 1985), for example, that we hypothesize is contributing to the majority of the genetic variance of many quantitative phenotypes.

The measured genotype approach to examining quantitative variability has several advantages over the unmeasured genotype or biometrical approaches. First, the measured genotype approach allows one to address directly questions about the genetic variability underlying the variability of a quantitative phenotype. Secondly, the variance of an estimate of a genetic parameter of interest is smaller and the power of hypothesis tests about the parameters is greater in the measured genotype case than for the unmeasured case because of the additional available information (Boerwinkle, 1985). Thirdly, many of the assumptions about the frequencies and effects of the unmeasured loci, such as Hardy–Weinberg equilibrium and additivity among effects, are practically necessary for an application of the unmeasured genotype approach. These assumptions are not practically necessary for a measured locus. And fourthly, the application of multiple-locus models to quantitative phenotypes is feasible only with the measured genotype approach.

Advances in our understanding of the biology of many quantitative phenotypes have identified loci which are involved in their etiology. More sensitive measurement techniques, such as two-dimensional gel electrophoresis combined with silver staining (Celis & Bravo, 1984) and the restriction fragment length polymorphisms (Wyman & White, 1980), will allow one to identify genetic variants at these candidate loci. Using these measured genotypes and the measured-genotype analysis strategies presented here, the loci contributing to quantitative phenotype variability can be identified and their effects estimated. As more loci in a system are measured, the unmeasured polygenic random component will be reduced, and fundamentally important questions about the genetic architecture of quantitative phenotypic variability can begin to be addressed.

This work was completed as part of the Ph.D. requirement of Eric Boerwinkle. We wish to thank Drs P. P. Moll, R. J. Muirhead and J. V. Neel for their helpful comments. The work was supported in part by grants P01 CA26803, R01 HL24489 and R01 HL30428 to C.F.S., and NIH GM20293 to R.C.

REFERENCES

BICKEL, P. J. & DOKSUM, K. A. (1977). *Mathematical Statistics*. San Francisco: Holden-Day.

BLACKWELDER, W. C. & ELSTON, R. C. (1974). Comments on Dr. Robinson's communication. *Behavioral Genetics* **4**, 97–99.

BOERWINKLE, E. (1985). The use of measured genotype information in the genetic analysis of quantitative phenotypes. Ph.D. Thesis, University of Michigan, Ann Arbor.

BOERWINKLE, E. & SING, C. F. (1986). Bias of the variance contribution of single loci. *Am. J. Hum. Genet.* (Submitted.)

BOERWINKLE, E., TURNER, S. T. & SING, C. F. (1984). The role of the genetics of sodium lithium countertransport in the determination of blood pressure variability in the population at large. In *The Red Cell; Sixth Ann Arbor Conference* (ed. G. J. Brewer). New York: Liss.

BURNS, T. L., MOLL, P. P. & SCHORK, M. A. (1984). Comparisons of different sampling designs for the determination of genetic transmission mechanisms in quantitative traits. *Am. J. Hum. Genet.* **36**, 1060–1074.

CELIS, J. E. & BRAVO, R. (1984). *Two-Dimensional Electrophoresis of Proteins.* New York: Academic Press.

CLARK, R. L., BOERWINKLE, E., BREWER, G. J. & SING, C. F. (1983). Studies of enzyme polymorphism in the kamuela population of *Drosophila mercatorum*. III. Effects of variation at the α GPD locus and subflight stress on the energy charge and glycolytic intermediate concentration. *Genetics* **104**, 661–675.

DAIGER, S. P., MILLER, M. & CHAKRABORTY, R. (1984). Heritability of quantitative variation at the group specific component locus. *Am. J. Hum. Genet.* **36**, 663–676.

DAY, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56**, 463–474.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. Roy. Stat. Soc.* B **39**, 1–38.

ELSTON, R. C. & STEWART, J. (1971). A genetic model for the genetic analysis of pedigree data. *Hum. Hered.* **21**, 523–542.

ELSTON, R. C. (1980). Segregation Analysis. In *Current Developments in Anthropological Genetics, Vol. 1, Theory and Methods* (eds. M. H. Crawford and J. H. Mielke), pp. 327–352. New York: Plenum Press.

FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433.

GO, R. C. P., ELSTON, R. C. & KAPLAN, E. B. (1978). Efficiency and robustness of pedigree segregation analysis. *Am. J. Hum. Genet.* **30**, 28–37.

GOLDSTEIN, J. L. & BROWN, S. (1979). The LDL receptor locus and the genetics of familial hypercholesterolemia. *Ann. Rev. Genet.* **13**, 259–290.

HASSELBLAD, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics* **8**, 431–444.

HOPPER, J. L. & MATHEWS, J. D. (1982). Extension to multivariate normal models for pedigree analysis. *Ann. Hum. Genet.* **46**, 373–383.

HOPKINSON, D. A., SPENCER, N. & HARRIS, H. (1964). Genetical studies on human red cell acid phosphatase. *Am. J. Hum. Genet.* **16**, 141–154.

JACQUARD, A. (1974). *The Genetic Structure of Populations.* New York: Springer-Verlag.

LALOUEL, J. M., DARLU, P., HENROTTE, J. G. & RAO, D. C. (1983*a*). Genetic regulation of plasma and red blood cell magnesium concentration in man. II. Segregation analysis. *Am. J. Hum. Genet.* **35**, 938–950.

LALOUEL, J. M., RAO, D. C., MORTON, N. E. & ELSTON, R. (1983*b*). A unified model for complex segregation analysis. *Am. J. Hum. Genet.* **35**, 816–826.

LANGE, K., SPENCE, M. A. & FRANK, M. B. (1976*a*). Application of the LOD method to the detection of a linkage between a quantitative trait and a qualitative marker; a simulation experiment. *Am. J. Hum. Genet.* **28**, 167–173.

LANGE, K., WESTLAKE, J. & SPENCE, M. A. (1976*b*). Extensions to pedigree analysis. III. Variance components by the scoring method. *Ann. Hum. Genet.* **39**, 485–491.

LI, C. C. & SACKS, L. (1954). The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics* **10**, 47–360.

MACLEAN, C. J., MORTON, N. E. & LEW, R. (1975). Analysis of family resemblance. IV. Operational characteristics of segregation analysis. *Am. J. Hum. Genet.* **27**, 365–384.

MALÉCOT, G. (1948). *Les mathématiques de l'hérédité.* Paris: Masson.

MOLL, P. P., POWSNER, R. & SING, C. F. (1979). Analysis of genetic and environmental sources of variation in serum cholesterol in Tecumseh, Michigan. V. Variance components estimated from pedigrees. *Ann. Hum. Genet.* **42**, 343–354.

MOLL, P. P., BERRY, T. D., WEIDMAN, W. H., ELLEFSON, R., GORDON, H. & KOTTKE, B. A. (1984). Detection of genetic heterogeneity among pedigrees through complex segregation analysis: an application to hypercholesterolemia. *Am. J. Hum. Genet.* **36**, 197–211.

MORTON, N. E. & MACLEAN, C. J. (1974). Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. *Am. J. Hum. Genet.* **26**, 489–503.

NEIMANN-SØRENSEN, A. & ROBERTSON, A. (1961). The association between blood groups and several production characteristics in three Danish cattle breeds. *Acta Agricultura Scandinavica* **XI**, 163–196.

NETER, J. & WASSERMAN, W. (1974). *Applied Linear Statistical Models.* Homewood, Illinois: Irwin.

SCHEFFÉ, H. (1959). *The Analysis of Variance.* New York: Wiley.

SING, C. F. & ORR, J. D. (1976). Analysis of genetic and environmental sources of variation in serum cholesterol in Tecumseh, Michigan. III. Identification of genetic effects using 12 polymorphic genetic blood marker systems. *Am. J. Hum. Genet.* **28**, 453–464.

SING, C. F. & DAVIGNON, J. (1985). The role of Apoliprotein E genetic polymorphism in determining normal plasma lipid and lipoprotein variation. *Am. J. Hum. Gen.* **37**, 268–285.

SMITH, C. A. B. (1980). Estimating genetic correlations. *Ann. Hum. Genet.* **43**, 265–284.

THOMPSON, J. N., JR. (1975). Quantitative variation and gene number. *Nature* **258**, 665–668.

THOMPSON, J. N. & THODAY, J. M. (1979). Synthesis; Polygenic variation in perspective. In *Quantitative Genetic Variation* (eds. J. N. Thompson and J. M. Thoday). London: Academic Press.

TURNER, S. T., JOHNSON, M., BOERWINKLE, E., RICHELSON, E. & SING, C. F. (1985). Distribution of sodium lithium countertransport and its relationship to blood pressure in a large sample of blood donors. *Hyperten.* (In the Press.)

WYMAN, A. & WHITE, R. (1980). A highly polymorphic locus in human DNA. *PNAS* **77**, 6754–6758.