Genome-Wide Association Mapping With Longitudinal Data

Nicholas A. Furlotte,[1] Eleazar Eskin,[1,2]* and Susana Eyheramendy[3]

[1] *Department of Computer Science, University of California, Los Angeles, California*
[2] *Department of Human Genetics, University of California, Los Angeles, California*
[3] *Department of Statistics Pontificia, Universidad Catolica de Chile, Chile*

Many genome-wide association studies have been performed on population cohorts that contain phenotype measurements at multiple time points. However, standard association methodologies only consider one time point. In this paper, we propose a mixed-model-based approach for performing association mapping which utilizes multiple phenotype measurements for each individual. We introduce an analytical approach to calculate statistical power and show that this model leads to increased power when compared to traditional approaches. Moreover, we show that by using this model we are able to differentiate the genetic, environmental, and residual error contributions to the phenotype. Using predictions of these components, we show how the proportion of the phenotype due to environment and genetics can be predicted and show that the ranking of individuals based on these predictions is very accurate. The software implementing this method may be found at http://genetics.cs.ucla.edu/longGWAS/. *Genet. Epidemiol.* 36:463–471, 2012.

Key words: longitudinal; genome-wide association; mixed-model; statistical genetics

*Correspondence to: Eleazar Eskin, Departments of Computer Science and Human Genetics, University of California, Los Angeles, Mail Code: 1596 3532-J Boelter Hall, Los Angeles, CA 90095-1596. E-mail: eeskin@cs.ucla.edu

Received 2 December 2011; Revised 23 February 2012; Accepted 3 April 2012
Published online 11 May 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21640

INTRODUCTION

The use of genome-wide association study (GWAS) methodologies has become common practice in the analysis of complex traits. However, it is well known that the genetic variants identified for most common traits only account for a small portion of the heritability [Eichler et al., 2010; Manolio et al., 2009], implying that there are a large number of genetic associations yet to be identified. Many GWAS have been performed in cohorts, in which multiple time points are available for each individual [Aulchenko et al., 2008; Kamatani et al., 2010; Kathiresan et al., 2007; Sabatti et al., 2008]. However, current association methods only utilize one time point for each individual. This is accomplished by either selecting a single measurement [Sabatti et al., 2008] or by computing the average over all time points [Ionita-Laza et al., 2007; Kamatani et al., 2010; Kathiresan et al., 2007]. It is reasonable to assume that a method jointly considering all time points when performing association may have increased power over single time point approaches.

In this paper, we present a method for performing association mapping with longitudinal phenotypes and show that this method has increased power over single time point mapping approaches. Recently, there has been an interest in the prediction of the genetic contribution to traits using cohort data [Yang et al., 2010]. These methods utilize a single phenotype measurement for each individual and predict the genetic contribution to the phenotype by taking into account the relationships between individuals. We show that when utilizing multiple measurements, this genetic contribution can be differentiated from the environmental and error contribution and we show how each of these contributing factors can be accurately predicted. Our method utilizes a mixed effects approach to model phenotype measurements. Mixed-effects models have been used extensively for modeling correlated data and are an important tool in the analysis of longitudinal data [Harville, 1977; Laird and Ware, 1982]. This class of models has been used in the analysis of longitudinal data for twin studies [Wang et al., 2011] as well as pedigree-based family studies [de Andrade et al., 2002]. We propose a model that partitions each individual's trait measurements into both a genetic and environmental component. We refer to the genetic component as the "genetic influence" to the trait. Similarly, we refer to the environmental contribution as the "lifestyle value."

To evaluate our method, we first compare power with a traditional mapping procedure utilizing only one time point. Power is evaluated through an analytical approach similar to that introduced by Williams and Blangero [1999]. Using a set of individuals obtained from the Wellcome Trust Case Control Consortium (WTCCC), we show that our method has increased power over traditional approaches. Second, we evaluate the accuracy in calculating genetic influence and lifestyle values for individuals while varying the number of available time points. We show that for phenotypes heavily influenced by the environment, the accuracy in prediction of the proportion of the phenotype due to genetics and environment has large variation. However, when ranking individuals based on their predicted lifestyle values, we find that this ranking is highly concordant with the ranking obtained using the true lifestyle values. This implies that individuals may be effectively categorized by lifestyle based on these predicted values.

MATERIALS AND METHODS

LONGITUDINAL PHENOTYPES

In experiments adopting longitudinal designs, phenotype measurements are collected for each of n individuals at m time points. We expect that measurements acquired from the same individual will tend to be more correlated than those obtained from different individuals. This correlation is due to both genetic and environmental effects shared between measurements. In order to conceptualize this, we present a generative model for phenotype measurements, which is a model specifying the mathematical process by which measurements may be systematically generated.

$$y_{ij} = \mu + G_i + E_{ij} + \epsilon_{ij} \tag{1}$$

The generative model in equation (1) states that a phenotype measurement j from individual i is a function of the global phenotype mean μ, an individual-specific genetic effect G_i, a measurement and individual-specific environmental effect E_{ij} and an error term ϵ_{ij}, accounting for other unknown factors such as measurement error. The value of G_i is a function of the genetic variation for individual i, and is expected to remain constant over time as an individual's genetic make up does not change. This assumption may not hold if there exist, for example, gene-by-environment interactions. On the other hand, the value of E_{ij} may vary across measurements due to changing environmental conditions. The correlation between each pair of E_{ij}s will depend on the magnitude of environmental change between time points as well as the degree of influence environment has over the phenotype in question. The residual terms ϵ_{ij} are expected to be independent between measurements.

TRADITIONAL APPROACH TO ASSOCIATION MAPPING

The traditional approach to association mapping considers one measurement for each of n individuals and interrogates each genetic locus individually. The traditional model is given as follows

$$y_i = \mu + \beta_r x_{ir} + \epsilon_i \tag{2}$$

x_{ir} represents the state of SNP r (single nucleotide polymorphism r) for individual i and β_r its coefficient [Balding, 2006]. By testing the hypothesis $\beta_r = 0$, it is determined whether SNP r influences the trait or not. We note that, with respect to the model in equation (1), the model in equation (2) has folded many terms into the residual term ϵ_i. This model is represented using standard vector notation as follows:

$$\mathbf{y} = \mathbf{X}_r\beta + \boldsymbol{\epsilon} \tag{3}$$

$\mathbf{y}$ is a vector of all phenotype measurements and $\mathbf{X}_r = [\mathbf{1}_n \quad \mathbf{x}_r]$, where $\mathbf{x}_r$ is a vector representing the n SNP values for SNP r, β is a vector of coefficients and $\mathbf{1}_n$ is a column vector of ones. Other fixed effects may be added to $\mathbf{X}$ in order to account for additional confounding.

To apply this model to longitudinal data, the total set of mn measurements must be preprocessed into a set of n independent measurements. There are two common approaches taken and we refer to these as the single approach and the average approach. In the single approach, a single measurement from the set of m measurements is chosen for each individual i. In the average approach, the m measurements for each individual are averaged and the average value is used as the single phenotype measurement for that individual. Under the assumption that individuals are unrelated, both of these procedures result in a set of n independent measurements and the standard model may be applied.

MIXED EFFECTS MODEL FOR ASSOCIATION MAPPING

The traditional method for association mapping interrogates each genetic locus individually while using single time points. However, it is known that traits are often influenced by many loci and ignoring this fact may have a negative impact on association mapping results. In particular, global genetic similarities between individuals may be correlated with trait similarities and this global correlation may cause many genetic loci to appear to be associated with the trait, a problem often referred to as population structure or cryptic relatedness [Devlin et al., 2001; Voight and Pritchard, 2005]. One way to account for this structure is through the use of a variance component model, in which the global genetic relatedness, referred to as polygenic background, of individuals is accounted for by the introduction of a random variable into the simple model from equation (2) [Kang et al., 2010; Lange, 2002; Yu et al., 2006]. This model is summarized as follows.

$$y_i = \mu + \beta_r x_{ir} + u_i + \epsilon_i \tag{4}$$

The model is equivalently described in matrix notation using

$$\mathbf{y} = \mathbf{X}_r\beta + \mathbf{Zu} + \boldsymbol{\epsilon} \tag{5}$$

The random variable u_i is assumed to be normally distributed with mean zero and variance σ_g^2 and the $\text{cov}(u_i, u_j) = \sigma_g^2 K_{ij}$, where K_{ij} is the kinship coefficient for individual i and j, which is a value representing their genetic relatedness. The incidence matrix $\mathbf{Z}$ maps measurements from each individual to the phenotype vector $\mathbf{y}$, and in the case when there is only one measurement for each individual $\mathbf{Z} = \mathbf{I}_n$. This form is standard in the mixed-model literature. With this the $\text{var}(\mathbf{Zu}) = \sigma_g^2\mathbf{ZKZ}'$ and $\text{var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2\mathbf{I}$. The total variance of $\mathbf{y}$ is then given by

$$\boldsymbol{\Sigma} = \sigma_g^2\mathbf{ZKZ}' + \sigma_\epsilon^2\mathbf{I} \tag{6}$$

To test the hypothesis $\beta_r = 0$ using the model in equation (5), the two variance components σ_g^2 and σ_ϵ^2 must be estimated. Since there is no analytical solution, this is accomplished using a numerical search algorithm implemented in the program EMMAX [Kang et al., 2010]. EMMAX combines grid search with the Newton-Raphson algorithm, in order to find the optimal variance components σ_g^2 and σ_ϵ^2 in time linear in the number of measurements, given the singular value decomposition of $\mathbf{K}$. Furthermore, by assuming that each SNP only has a small to moderate effect on the phenotype, it is reasonable to assume that variance component estimates will be the same for each SNP. With this assumption it is only necessary to perform the variance component search once and thus feasible to perform the hypothesis test for each SNP within the genome.

There are many methods to compute the kinship matrix **K**. For a review of many standard relatedness estimators see [Oliehoek et al., 2006]. More recently, [Yang et al., 2010] proposed a method for adjusting the relatedness matrix, to account for the fact that the true causal SNPs may not be strongly correlated with the genotyped SNPs. Such issues are beyond the scope of this paper and thus we will only use the identity by state (IBS) allele sharing matrix [Kang et al., 2008]. Furthermore, these issues and the choice of kinship matrix do not affect our simulation results. In general, the methodology introduced in this paper may be utilized as long as the kinship matrix is positive semi-definite.

ASSOCIATION MAPPING WITH LONGITUDINAL DATA

The models from equations (2) and (4) do not take advantage of the availability of multiple time points and do not directly account for both genetic and environmental factors. We suggest a model that directly accounts for each term using all time points by extending the model in equation (4) .

$$y_{ij} = \mu + \beta_r x_{ir} + u_i + v_{ij} + \epsilon_{ij} \tag{7}$$

The random variable v_{ij} is introduced to represent the contribution of the environment to the phenotype measurement (E_{ij} from equation (1)). The matrix version is as follows.

$$\mathbf{y} = \mathbf{X}_r\beta + \mathbf{Zu} + \mathbf{v} + \boldsymbol{\epsilon} \tag{8}$$

We assume, without loss of generality, that the mean of the random components **u** and **v** are equal to zero and that the variance structure is as follows, where **D** is a known matrix representing the covariance between environmental components.

$$\text{var}\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\sigma_g^2\mathbf{K} & 0 & 0\\ 0 & \sigma_v^2\mathbf{D} & 0\\ 0 & 0 & \sigma_\epsilon^2\mathbf{I}\end{bmatrix} \tag{9}$$

With this we define the variance of **y**.

$$\text{var}(\mathbf{y}) = \boldsymbol{\Sigma} = \sigma_g^2\mathbf{ZKZ}' + \sigma_v^2\mathbf{D} + \sigma_\epsilon^2\mathbf{I} \tag{10}$$

In general, the matrix **D** will depend on the level of correlation between individual time points and can be determined through estimation techniques or by fitting parametric models, such as models of the autoregressive class [Jennrich and Schluchter, 1986]. Most commonly **D** will take the form of a block diagonal matrix, so that environmental components between individuals will be independent. The variance of **v** is then given as $\text{var}(\mathbf{v}) = \mathbf{D} = \mathbf{E} \otimes \mathbf{I}$, where $\otimes$ represents the Kronecker product of two matrices, and **E** is an $m \times m$ matrix representing the covariance between the set of m time points for each individual.

MISSING DATA

One complication that often arises when dealing with longitudinal data is that of unbalanced or missing data [McCulloch and Searle, 2001]. When a study is unbalanced, meaning that all individuals do not have the same number of measurements, the model notation becomes slightly more complicated. Let us consider that individual i has m_i measurements and define $\mathbf{m} = [m_1 \quad m_2 \ldots m_n]$. The incidence matrix **Z** will still map genetic components to measurements and its structure will be dictated by the vector **m**. For example, the first m_1 rows of **Z** will have a 1 in the first column and the second m_2 rows of **Z** will have a 1 in the second column and so on.

To avoid complicated notation, we suggest a simple scheme for defining the model. Define the model using $m = max(\mathbf{m})$, so that the assumed number of measurements is equal to nm and the vector **y** has missing values. Select the measurements that are missing at each individual and remove the rows and columns of the full covariance matrices of each component (genetic, environmental, and residual error) that correspond to the indices of these entries in the vector **y** of size nm. The resulting covariance matrices will then correspond to a new vector $\tilde{\mathbf{y}}$, defined as the vector **y** with the missing values removed. Modeling fitting procedures are then readily adaptable to such matrices and hypothesis testing procedures can be easily applied.

ESTIMATING VARIANCE COMPONENTS

To fit the models in equations (5) and (8), we must estimate a set of variance components. For the model in equation (5), a linear time search algorithm based on maximum likelihood exists, which is able to identify the optimal variance components. However, no linear time method exists to find the three variance components required for the model in equation (8). Therefore, we utilize an approach suggested by [Listgarten et al., 2010], in which we use the EMMAX algorithm inside of a linear time search.

First, we rewrite the variance of **y** as shown in equation (11), letting $\tau^2 = \sigma_v^2 - \sigma_g^2$ and $w = \sigma_g^2/(\sigma_v^2 - \sigma_g^2)$. Then, given a value w between zero and one, we apply the EMMAX search algorithm to find optimal variance components τ^2 and σ_ϵ^2. If we search q different values of w, then our approach will be q times slower than EMMAX. More specifically, EMMAX has a one time computational cost of $O(N^3)$ followed by a cost of $O(rN)$ for r search iterations, where N is the total number of measurements or the size of the vector **y**. In comparison, our approach will have a one time cost of $O(qN^3)$ followed by a cost of $O(qrN)$. This is compared to the basic Newton-Raphson technique, which has a total computation cost of $O(qrN^3)$.

$$\begin{aligned}\text{var}(\mathbf{y}) &= \tau^2(w\mathbf{ZKZ}' + (1-w)\mathbf{D}) + \sigma_\epsilon^2\mathbf{I}\\ &= \tau^2\mathbf{K}^* + \sigma_\epsilon^2\mathbf{I}\end{aligned} \tag{11}$$

Additional cost will be incurred if **D** has to be estimated, such as in the case when an auto-regressive model is utilized. For example, in the case where **D** is determined through an auto-regressive model, the total computational time would be multiplied by the size of the search space for the additional auto-regressive parameter. If the number of iterations required to optimize this parameter were $O(p)$, then the total computational cost will be $O(pqN^3 + qrN)$.

PREDICTING LIFESTYLE VALUES AND GENETIC INFLUENCE

The realization of **u** is a vector of values representing the genetic contribution to the phenotype measurement. That is, **u** is a random variable and at the time the phenotype was measured a value for **u** was sampled from a multivariate distribution. This is the realized value. Just as fixed effects are estimated, the value of a random variable can be predicted using the best linear unbiased predictors (BLUPs) introduced by Henderson [1950]. The BLUP for **u**, denoted by $\tilde{\mathbf{u}}$, is given by

$$\tilde{\mathbf{u}} = \sigma_g^2 \mathbf{K}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}_r\hat{\beta}) \tag{12}$$

where $\boldsymbol{\Sigma}$ is given by equation (10). The realized value of **u** represents the overall genetic contribution to the phenotype for each individual. By determining the value of u_i for each individual, we are able to determine what proportion of each phenotypic measurement is due to genetics and what proportion is due to other factors, specifically fixed effects and error. This enables us to compare individuals based on the magnitude of their genetic contributions. Certain individuals may have a stronger genetic effect than others. When this large genetic effect causes phenotypes to become harmful, such as in high cholesterol, we may see this difference as indicator of increased risk. When large genetic effects lead to beneficial phenotypes, this may indicate a sort of genetic robustness, a phenomenon often referred to colloquially as having "good genes." Therefore, we refer to the realized value of **u** as the genetic influence.

Just as it is possible to predict **u**, it is also possible to predict the realized value of **v** from equation (8). The realized value of **v** for each individual represents the environmental contribution to the phenotype. By comparing realized values of **v** it is possible to uncover differences in individual environment, which may be an indicator of an individual's lifestyle. For this reason, we refer to the realized value of **v** as the vector of lifestyle values. When calculating the realized values for **u** and **v**, we are able to partition each phenotype measurement into genetics, environment, fixed effects and error, and give the proportion that each factor contributes.

$$\begin{bmatrix} \hat{\beta} \\ \tilde{\mathbf{u}} \\ \tilde{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}' \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\alpha_1 & \mathbf{Z}' \\ \mathbf{X} & \mathbf{Z} & \mathbf{I} + \mathbf{D}^{-1}\alpha_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{y} \end{bmatrix} \tag{13}$$

The BLUPs for **u** and **v** are obtained by solving the system of equations given in equation (13) [Henderson, 1973; Mrode and Thompson, 2005], where $\alpha_1 = \sigma_\epsilon^2/\sigma_g^2$ and $\alpha_2 = \sigma_\epsilon^2/\sigma_v^2$. These solutions are obtained by maximizing the joint likelihood of **y**, **u**, and **v** under the assumption of normality. After solving the so-called mixed-model equations (MME), we obtain a prediction of the random variable v_{ij} for each individual i and time point j, as well as predictions for u_i for each individual i.

Theoretical accuracy of the random effects may be analyzed by evaluating the variance in the difference between the true and predicted effects, which are calculated by

$$\mathrm{var}(\tilde{\mathbf{u}} - \mathbf{u}) = \sigma_g^2\mathbf{K} - \sigma_g^2\mathbf{K}\mathbf{Z}'\mathbf{P}\mathbf{Z}\sigma_g^2\mathbf{K} \tag{14}$$

$$\mathrm{var}(\tilde{\mathbf{v}} - \mathbf{v}) = \sigma_v^2\mathbf{D} - \sigma_v^2\mathbf{D}\mathbf{P}\sigma_v^2\mathbf{D} \tag{15}$$

where $\mathbf{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}$.

ANALYTICAL POWER FOR MIXED-EFFECTS MODELS

One common approach to evaluate methods for performing association is through the analysis of statistical power [de Bakker et al., 2005]. Although power is easily calculated analytically when assuming the model from equation (2), it is not well-known how to calculate power when using a mixed effects model. For this reason, time consuming simulations are often employed in order to estimate statistical power [Bennett et al., 2010]. Williams and Blangero [1999] introduced a likelihood-ratio based technique to compute power in variance component models used for linkage analysis. We introduce a similar derivation based on the *F*-test.

Let **y** be a vector of size n and assume that it has a normal distribution with mean $\mathbf{X}\beta$ and variance $\boldsymbol{\Sigma}$, where **X** is an $n \times q$ matrix of fixed effects, β is a $q \times 1$ vector of coefficients and $\boldsymbol{\Sigma}$ is an $n \times n$ covariance matrix. In order to test a hypothesis about β, we define a $q \times 1$ matrix **R**, which defines a linear combination of the elements of β. For example, if **X** only encodes global mean and SNP, then we define $\mathbf{R} = [0 \quad 1]$, so that $\mathbf{R}\beta$ results in the single SNP coefficient. Given **R** we define the hypothesis test $\mathbf{R}\beta = \mathbf{r}$. The generalized least squares (GLS) *F*-statistic is then given by

$$\phi_F = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{q} \tag{16}$$

We show that under the alternative hypothesis $\mathbf{R}\beta = \mathbf{r} + \delta$, ϕ_F follows an *F*-distribution with $n - q$ numerator and q denominator degrees of freedom and noncentrality parameter λ (see Appendix), given by

$$\lambda = \delta'[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta \tag{17}$$

It is important to note that we have assumed an optimal estimate for the true covariance matrix $\boldsymbol{\Sigma}$. Given the noncentrality parameter in equation (17), power is calculated as the area under the curve of the distribution defined by the noncentrality parameter that is beyond the null rejection region.

SIMULATIONS

The power gain was estimated by generating random covariance matrices for the environmental components and then by averaging analytical power over 1,000 randomly selected SNPs with minor allele frequencies in the range of 1– 5%. The expected power gain was calculated as the average power gain over 1,000 such randomly generated covariance matrices. The covariance matrices were generated by randomly selecting a vector of size $m - 1$ from a uniform (0,1) distribution, where m is the number of time points. The covariance between time points i and j $(i > j)$ is then given as the $(i - j - 1)$th entry in this vector. We then define an $m \times m$ matrix **E** using this scheme and define the full environmental covariance matrix $\mathbf{D} = \mathbf{E} \otimes \mathbf{I}$.

Phenotypes with multiple time points were generated by sampling both genetic and environmental components from their respective multivariate normal distributions, having a mean of zero and with variance as specified in equation (9). This results in both a genetic and environmental contribution value for each individual. These values are used as the individual's mean phenotype value to which random noise is added to generate each time point.

Random effects were predicted by fitting the model from equation (8), with only a mean effect, and then by obtaining the solution to equation (13). The environmental covariance matrix was estimated by calculating the correlation between time points using all individuals. This procedure works well, when either the genetic effect is small or the sample has little population structure. In the case where the population has a large amount of structure, we employ a simple iterative scheme. Starting with an estimate of the environmental covariance matrix calculated on the raw data, we predict the random genetic effect. This effect is regressed from the phenotype values and a new environmental covariance matrix is computed with these new phenotype values. This procedure is repeated until the environmental covariance matrix converges.

RESULTS

MULTIPLE MEASUREMENTS PROVIDE INCREASED POWER OVER TRADITIONAL APPROACHES

We evaluated the gain in power achieved when using the proposed method (the full method) over using an averaging approach or single approach. In the single approach, a single time point is selected for each individual, while in the average approach time point values are averaged for each individual. Power gain is evaluated by comparing the ratios of the power achieved with one method to that of the power achieved with the single approach. Figure 1 summarizes these results. Power gain was calculated for each effect size by averaging the power gain over 1,000 iterations, in which a different randomly selected environmental covariance matrix was used in the analytical calculation of power. This power was averaged over 1,000 SNPs with minor allele frequency in the range of 1–5%. All calculations assume that the environment accounts for 80% of the phenotypic variance while both the genetic background and residual error account for 10%.

Figure 1 compares both the power curves (Figures 1A and C) and power gain (Figures 1B and D) for 1,000 and 2,000 individuals randomly selected from the WTCCC. We see that on average the full method has increased power when compared to the average and single methods. This increased power is seen more clearly in the power gain plots, which show that the full method has as much as an eightfold gain in power over the single approach, compared to a roughly 4.5-fold power gain achieved by the average approach.

MULTIPLE MEASUREMENTS ALLOW FOR THE PREDICTION OF INDIVIDUAL LIFESTYLE

Another benefit of using the full method over approaches utilizing only one time point is the ability to predict the phenotypic contribution due to environment. We evaluate this ability through simulation. We simulated phenotypes for 1,000 individuals in which the environment accounts for 80% of the variance, while genetic and error each contribute 10%. These phenotypes represent those that are largely influenced by environment. For each phenotype, we predict the environmental contribution and the proportion of the phenotype that this value accounts for. More specifically, each phenotype can be seen as a linear combination of mean, genetic contribution, environmental contribution, and error. Using random effect predictions, as summarized in the methods, we obtain predicted values for the genetic and environmental contributions at each time point. The prediction of random effects for each time point scales with the cube of the number of individuals times the number of time points. In practice, 1,000 individuals and two time points require a running time under 5 min for one phenotype, whereas the running time for 1,000 individuals with five time points is just under 50 min. However, we note that recent advances in the computational theory behind linear mixed models can serve to decrease these running times [Lippert et al., 2011].

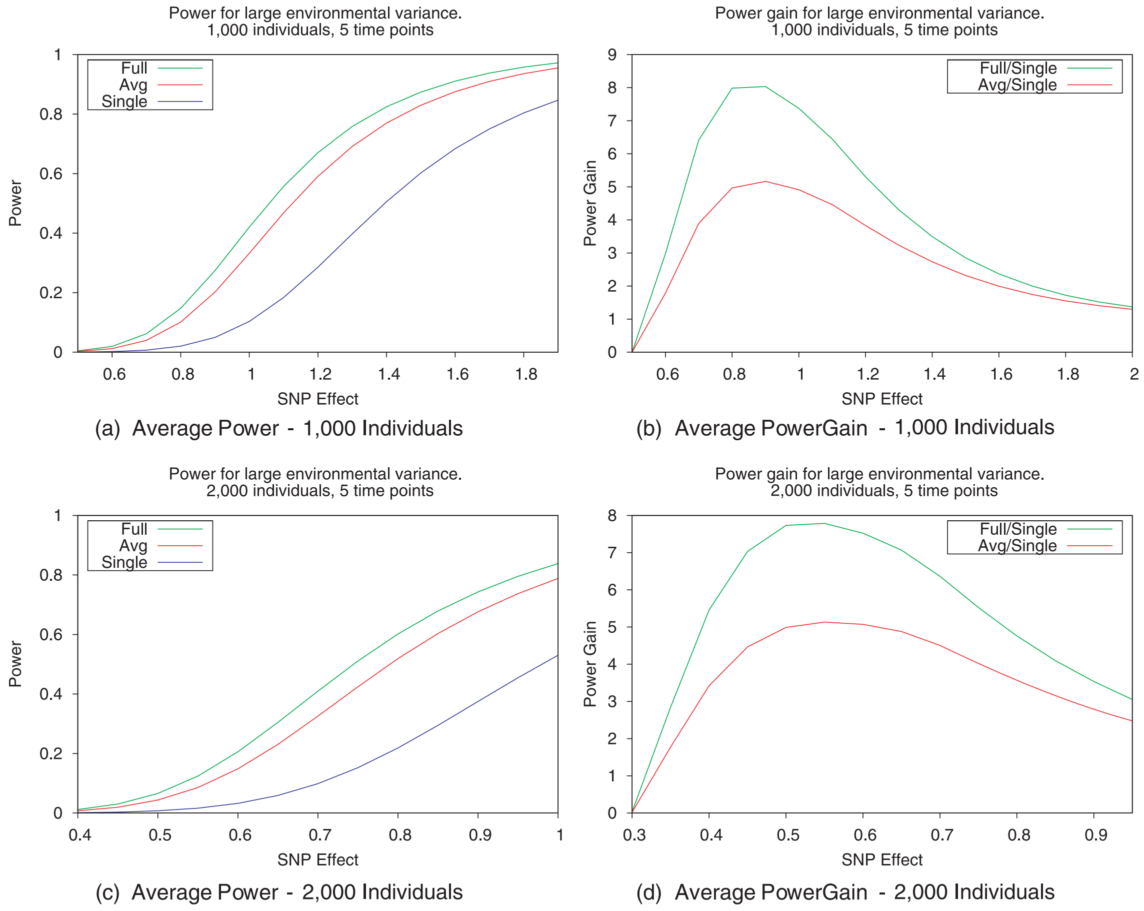
The results of this simulation are summarized in Figure 2. First, we compare the accuracy of the predictions by summarizing the difference in the true proportion of the phenotype contributed by the environmental component with that of the predicted proportion (Figure 2A) for 1,000 randomly generated environmental covariance matrices. From this plot, we see that the difference between the true and predicted proportion hovers around 25%, while increasing the number of time points available shifts this mean toward zero. Despite the high variation in accuracy, Figure 2B, showing the distribution of correlations between the true lifestyle values with the predicted, shows that on average the predictions have a rank correlation of 0.94. This indicates that although accuracy is not always high, the relative ranking of individuals based on their predicted lifestyle values is highly concordant with their true ranking. This implies that individuals may be effectively ranked based on their predicted lifestyle values.

Figure 3 shows the accuracy and correlations for predicted genetic values. We find that the predicted proportions behave very similarly to that of the lifestyle values, except that the availability of additional measurements does not increase accuracy in this case. However, the correlation between the true and predicted genetic effects is on average very small and has a strange pattern. Although, we find that as the number of time points increases, the average correlation increases.

DISCUSSION

In this paper, we introduce a mixed-model based approach to perform association mapping in GWAS, when multiple measurements are available. We show that by utilizing multiple measurements, our method achieves increased power over methods that either select a single measurement or average measurements for individuals. Furthermore, we show that when multiple measurements for each individual are available, it is possible to differentiate the genetic contribution from the environmental contribution. We call these quantities the genetic influence and lifestyle values, respectively.

The ability to partition a phenotype into its constituents may be useful for future phenotype prediction and treatment selection. For example, some individuals might



(a) Average Power - 1,000 Individuals

(b) Average PowerGain - 1,000 Individuals

(c) Average Power - 2,000 Individuals

(d) Average PowerGain - 2,000 Individuals

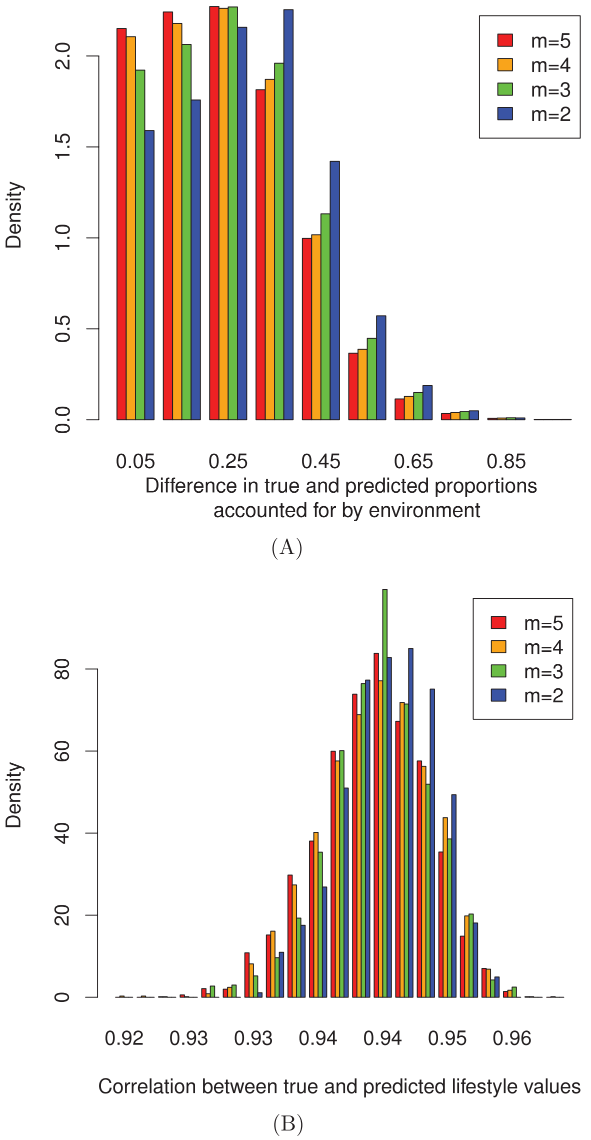
Fig. 1. Association mapping utilizing multiple measurements leads to an increase in power over traditional approaches. We compare the average power gain for the proposed full model with that of the average model (using averaged measurements for each individual). Power gain is defined as the ratio of the power of a given method to that achieved with the single approach (i.e., mapping with only one measurement for each individual) and was calculated by averaging power gain over 1,000 randomly selected SNPs with minor allele frequency (MAF) in the range of 1–5% and over 1,000 randomly selected covariance structures for the multiple measurements ($m = 5$). Simulations were performed with the environmental effect accounting for 80% of the variance while the genetic background and residual error accounted for the remaining 20%.

gain substantially from a decrease in dietary cholesterol when the largest part of their cholesterol is due to their intake. On the other hand, some individuals who are genetically predisposed to high cholesterol might stand to gain little from decreasing their dietary cholesterol, but instead might require medication in order to alter their overall cholesterol levels. With knowledge of the individual contributions to total cholesterol, the appropriate treatment options may be put in place in order to alter the future of the phenotype.

The previous cholesterol example extends very naturally to explain how prediction of genetic influence and lifestyle may be useful for risk prediction. For example, it may be discovered through these methods that an individual has a very large lifestyle component for cholesterol. In this case, their risk for developing cardiovascular disease may be predicted based on the magnitude and direction of this value, such that the resulting prediction may be different from that obtained by using the total level of cholesterol.

Another interesting aspect of the ability to predict lifestyle values is the subsequent ability to categorize individuals. For example, when evaluating a trait such as lung capacity, certain individuals will have decreased lung capacity due to long-term smoking, while others will have relatively normal capacity given their age and genetic makeup. This might be easily discerned by using a series of questions, but it is well-known that the truth is not always told when answering such questions. In this case, the lifestyle value may help to categorize individuals based only on their phenotype measurements and genotypes.

The model we propose is tested under a certain set of assumptions, however, the structure is very general and may work for a larger class of problems. For example, there are cases when it is not reasonable to assume that a phenotype follows a normal distribution, and a simple transformation such as log is not sufficient to obtain a normally distributed measure. For example, binary or categorical outcomes cannot be expected to follow a normal distribution. In this case, the phenotype may be modeled using a link function, such is done in logistic regression [McCulloch and Searle, 2001]. The models presented here may then be utilized in this space. Furthermore, there may be other factors to include in the model, such as gene-by-environment interactions, or even more complicated treatment-genetic-environmental

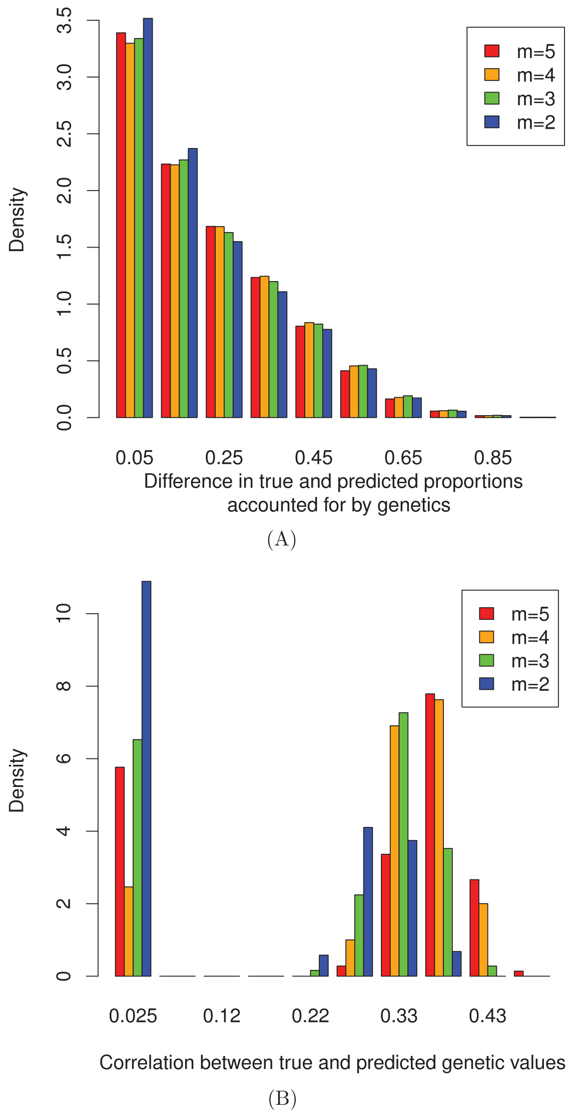


(A)

(B)

Fig. 2. The accuracy in prediction of lifestyle values varies, while the ranking remains consistent. For each of 1,000 iterations, lifestyle values were predicted and compared with their known true values, through simulation. Each simulation assumes 1,000 individuals and m time points. (A) Evaluates the difference between the proportion accounted for by the environment as determined by the true lifestyle effect with that of the predicted lifestyle effect. This result indicates that the accuracy of these predictions has a high variation, but that by increasing the number of time points it is possible to obtain more accurate predictions. (B) Shows the distributions of Spearman rank correlations between the true lifestyle and predicted lifestyle values. This result indicates that the ranking of individuals based on their predicted lifestyle is highly concordant with the true lifestyle ranking.

interactions. Perhaps some individuals have increased variance when subjected to certain environments and certain treatments but not with others. The model presented in this paper can be used as a base to explore such conditions, which may require more complex models with additional random effects.



(A)

(B)

Fig. 3. The accuracy in prediction of genetic values is similar to that of lifestyle. For each of 1,000 iterations, genetic values were predicted and compared with their known true values, through simulation. (A) Shows a very similar result to that found in lifestyle values, where the accuracy of these predictions has a high variation and has a relatively uniform distribution across different numbers of time points. However, the result in (B) is much different than that found in the lifestyle value prediction. There is not a clear pattern, although the average correlation does increase as the number of time points increases.

ACKNOWLEDGMENTS

N.F. and E.E. are supported by National Science Foundation grants 0513612, 0731455, 0729049, 0916676, and 1065276, and National Institutes of Health grants K25-HL080079, U01-DA024417, P01-HL30568, and PO1-HL28481. N.F. is supported by National Institute of Health training grant T32-HG002536. S.E. is supported by FONDECYT grant 11085012. This research was supported in part by the University of California, Los Angeles subcontract of contract N01-ES-45530 from the National Toxicology Program and National Institute of Environmental Health Sciences to Perlegen Sciences.

REFERENCES

Aulchenko Y, Ripatti S, Lindqvist I, Boomsma D, Heid I, Pramstaller P, Penninx B, Janssens A, Wilson J, Spector T, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin M-RR, Gyllensten U, Campbell H, Rudan I, Johansson Å, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, Falchi M, Willemsen G, Hottenga JJ, de Geus EJC, Montgomery GW, Whitfield J, Magnusson P, Saharinen J, Perola M, Silander K, Isaacs A, Sijbrands EJG, Uitterlinden AG, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin MR, Gyllensten U, Campbell H, Rudan I, Johansson Å, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, Falchi M, Willemsen G, Hottenga JJ, de Geus EJC, Montgomery GW, Whitfield J, Magnusson P, Saharinen J, Perola M, Silander K, Isaacs A, Sijbrands EJG, Uitterlinden AG, Witteman JCM, Oostra BA, Elliott P, Ruokonen A, Sabatti C, Gieger C, Meitinger T, Kronenberg F, Döring A, Wichmann HE, Smit JH, McCarthy MI, van Duijn CM, Peltonen L. 2008. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 41(1):47–55.

Balding D. 2006. A tutorial on statistical methods for population association studies. Nat Rev Genet 7(10):781–791.

Bennett B, Farber C, Orozco L, Min Kang H, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, Truong A, Yang WP, He A, Kayne P, Gargalovic P, Kirchgessner T, Pan C, Castellani LW, Kostem E, Furlotte N, Drake TA, Eskin E, Lusis AJ. 2010. A high-resolution association mapping panel for the dissection of complex traits in mice. Genome Res 20(2):281.

de Andrade M, Guéguen R, Visvikis S, Sass C, Siest G, Amos C. 2002. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 22(3):221–232.

de Bakker P, Yelensky R, Pe’er I, Gabriel S, Daly M, Altshuler Det al. 2005. Efficiency and power in genetic association studies. Nat Genet 37(11):1217–1223.

Devlin B, Roeder K, Bacanu S. 2001. Unbiased methods for population-based association studies. Genet Epidemiol 21(4):273–284. doi:10.1002/gepi.1034.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6):446–50.

Harville D. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72(358):320–338.

Henderson C. 1950. Estimation of genetic parameters. Ann Math Stat 21:309.

Henderson CR. 1973. Sire evaluation and genetic trends. J Anim Sci 1973(no. Symposium):10–41.

Ionita-Laza I, McQueen M, Laird N, Lange C. 2007. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. The American Journal of Human Genetics 81(3):607–614.

Jennrich RI, Schluchter MD. 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42(4):805–820.

Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, Nakamura Y, Kamatani N. 2010. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat Genet 42(3):210–215.

Kang H, Zaitlen N, Wade C, Kirby A, Heckerman D, Daly M, Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics 178(3):1709.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SYY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42(4):348–54.

Kathiresan S, Manning A, Demissie S, D’Agostino R, Surti A, Guiducci C, Gianniny L, Burtt N, Melander O, Orho-Melander M, Arnett DK, Peloso GM, Ordovas JM, Cupples LA. 2007. A genome-wide association study for blood lipid phenotypes in the framing ham heart study. BMC Med Genet 8(Suppl 1):S17.

Laird N, Ware J. 1982. Random-effects models for longitudinal data. Biometrics 38(4):963–974.

Lange K. 2002. Mathematical and Statistical Methods for Genetic Analysis (2nd ed.). New York: Springer Verlag.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D, Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. Fast linear mixed models for genome-wide association studies. Nat Methods 8(10):833.

Listgarten J, Kadie C, Schadt EE, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. Proc Natl Acad Sci USA 107(38):16465–16470.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747–753.

McCulloch C, Searle S. 2001. Generalized, Linear, and Mixed Models. New Jersey: Wiley-Interscience.

Mrode R, Thompson R. 2005. Linear Models for the Prediction of Animal Breeding Values (2nd ed.). Cambridge, MA: Cabi.

Oliehoek PA, Windig JJ, van Arendonk JAM, Bijma P. 2006. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173(1):483–496.

Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, Jones CG, Zaitlen NA, Varilo T, Kaakinen M, Sovio U, Ruokonen A, Laitinen J, Jakkula E, Coin L, Hoggart C, Collins A, Turunen H, Gabriel S, Elliot P, McCarthy MI, Daly MJ, Jarvelin MR, Freimer NB, Peltonen L. 2008. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41(1):35–46.

Voight B, Pritchard J. 2005. Confounding from cryptic relatedness in case-control association studies. PLoS Genet 1(3):32. doi:10.1371/journal.pgen.0010032.

Wang X, Guo X, He M, Zhang H. 2011. Statistical inference in mixed models and analysis of twin and family data. Biometrics 67(3):987–995.

Williams JT, Blangero J. 1999. Power of variance component linkage analysis to detect quantitative trait loci. Annals of Human Genetics 63(6):545–563.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565–569. doi:10.1038/ng.608.

Yu J, Pressoir G, Briggs W, Vroh Bi I, Yamasaki M, Doebley J, McMullen M, Gaut B, Nielsen D, Holland JB, Kresovich S, Buckler ES. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208.

APPENDIX: CALCULATING POWER UNDER A MIXED EFFECT MODEL

Consider that $\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1})$ from equation (16). Now consider that the hypothesis test $\mathbf{R}\beta = \mathbf{r}$, while in truth $\mathbf{R}\beta = \mathbf{r} + \boldsymbol{\delta}$ ($\mathbf{r} = \mathbf{R}\beta - \boldsymbol{\delta}$). To derive a χ^2 statistic, we first derive a Z-score statistic for the test $\mathbf{R}\beta = \mathbf{r}$.

$$\begin{aligned} Z &= [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\beta} - \mathbf{r}) \\ &= [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\beta} - \mathbf{R}\beta + \delta) \\ &= [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \end{aligned}$$

We know the distribution of $\hat{\beta}$ and therefore $\mathbf{R}\hat{\beta}$; thus,

$$\begin{gathered} [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta)) \sim N(0, I) \\ [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \\ \sim N([\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}\delta, I) \end{gathered}$$

Squaring Z, we obtain a χ^2 statistic. Let $\mathbf{W} = [\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']$.

$$\begin{aligned} \phi_c &= (\mathbf{R}(\hat{\beta} - \beta) + \delta)'\mathbf{W}^{-1}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \\ &= (\mathbf{R}\hat{\beta} - \mathbf{r})'\mathbf{W}^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(q; \delta'\mathbf{W}^{-1}\delta) \qquad \text{(A1)} \end{aligned}$$

Thus, ϕ_c is a χ^2 statistic with q degrees of freedom and a noncentrality parameter of $\delta'\mathbf{W}^{-1}\delta$. Now consider that $\mathbf{\Sigma}$ is actually unknown and that we will use an estimate $\hat{\mathbf{\Sigma}}$, such that $\mathbf{\Sigma} = \sigma_c^2\hat{\mathbf{\Sigma}}$, where σ_c^2 is an unknown scalar. Given this, we know that $(n - q)\hat{\sigma}_c^2/\sigma_c^2 \sim \chi^2(n - q)$ and may obtain the following statistic by dividing ϕ_c by this quantity.

$$\begin{aligned} \phi_F &= \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\hat{\mathbf{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{\hat{\sigma}_c^2 q} \\ \phi_F &\sim \mathcal{F}(q, n - q, \delta'\mathbf{W}^{-1}\delta), \qquad \text{(A2)} \end{aligned}$$

We use $\mathcal{F}(df_1, df_2, ncp)$ to represent the noncentral $\mathcal{F}$-distribution with numerator degrees of freedom df_1 and denominator degrees of freedom df_2 and noncentrality parameter ncp. With optimal variance component estimates, we expect that $\hat{\sigma}_c^2 = 1$.