BLUP Genotype Imputation for Case-Control Association Testing with Related Individuals and Missing Data

MARY SARA McPEEK

ABSTRACT

We consider the problem of case-control association testing in samples that contain related individuals, where we assume the pedigree structure is known. Typically, for each marker tested, some individuals will have missing genotype data. The MQSL method has been proposed for association testing in this situation. We show that the MQSL method is equivalent to an approach in which missing genotypes are imputed using the best linear unbiased predictor (BLUP) based on relatives' genotype data. Viewed this way, the MQSL exactly corrects for the imputation error and for the extra correlation due to imputation. We also investigate the amount of additional power for detecting association that is provided by this BLUP imputation approach.

Key words: GWAS, pedigrees, quasi-likelihood, score test.

1. INTRODUCTION

STANDARD METHODS FOR CASE-CONTROL ASSOCIATION TESTING typically assume that sampled individuals are unrelated. However, many ongoing genetic studies involve families, so genetic association analysis of such studies requires some way of dealing with related individuals. The simplest way is to select a subsample of unrelated individuals, but that clearly results in a loss of information. Family-based tests (Rabinowitz and Laird, 2000) can be used in some contexts. Such approaches have the benefit that they protect against population structure, but they can also be underpowered relative to case-control association tests in some circumstances (Risch and Teng, 1998; Thornton and McPeck, 2010). A third approach is to perform a standard χ^2 test with a post-hoc correction factor applied. Slager and Schaid (2001) and Bourgain et al. (2003) have derived explicit correction factors for χ^2 tests of association performed in samples of related individuals. Another way to obtain such a correction factor in the context of a genome screen would be to use the genomic control method of Devlin and Roeder (1999), which works well if the rate of missing genotypes is similar at all markers. A fourth approach is to try to make more efficient use of the available information to obtain a more powerful test than the one that results from correcting the standard χ^2 test. This approach has been taken

Departments of Statistics and Human Genetics, University of Chicago, Chicago, Illinois.

This article is dedicated to Simon Tavaré in celebration of his 60th year and to Michael Waterman in celebration of his 70th year. I am grateful that they gave me the opportunity to be part of their computational biology group at USC during 1994–1995. It was a scientifically exciting and welcoming environment in which I met many interesting people.

by Bourgain et al. (2003) and by Thornton and McPeck (2007), who showed that there can be a nonnegligible power gain for their M_{QLS} method over the corrected χ^2 .

One purpose of this article is to better understand the nature of this power gain. One way in which the M_{QLS} improves on the corrected χ^2 is that it uses a more efficient estimator of allele frequency, which is a nuisance parameter in the analysis. Both the M_{QLS} and corrected χ^2 could also make use of unphenotyped controls who are not related to phenotyped individuals, but who are genotyped, to improve the nuisance parameter estimation. Additionally, the M_{QLS} can make use of genotyped individuals who are unphenotyped but who have phenotyped relatives in the sample as well as phenotyped individuals who have missing genotypes but who have genotyped relatives in the sample. It is this last aspect of power gain, the gain from using information on related individuals with missing genotype or phenotype, that we study in this article. We derive a connection between the M_{QLS} and imputation of missing genotypes based on genotypes of relatives, and we consider the question of how much power can be gained by squeezing this information from the data.

2. METHODS

2.1. Best linear unbiased prediction of genotypes

First, we ignore phenotype and focus on the problem of predicting missing genotypes based on the observed genotypes of relatives. Suppose our study consists of $n + m$ individuals, where we assume, without loss of generality, that the first n of the $n + m$ individuals have non-missing genotype data at the marker under consideration, while the last m have missing genotype data at the marker. The $n + m$ individuals can be arbitrarily related, with the pedigree(s) that specify the relationships assumed to be known. The method we describe is feasible even for large complex pedigrees with inbreeding loops. Unrelated individuals can also be included in the sample. If their genotypes are known, they will contribute to the prediction of missing genotypes in others by contributing to the allele frequency estimate. However, if individuals with missing genotypes are unrelated to others in the sample, then their genotypes will not be predicted by our method.

For the moment, we assume that the marker under consideration is an autosomal binary marker (e.g., a SNP) with alleles labeled “0” and “1.” We extend to multiple alleles in the next subsection. Let $\mathbf{G} = (G_1, \dots, G_{n+m})^T$ be the vector of true genotypes at the marker under consideration, where $G_i = 0, 1$, or 2, according to whether individual i has, respectively, 0, 1, or 2 copies of allele 1 at the marker. We assume that the first n entries of $\mathbf{G}$ are observed, while the last m entries are unobserved and are to be predicted. We use the notation $\mathbf{G} = (\mathbf{G}_N^T, \mathbf{G}_M^T)^T$ to denote the partition of the vector $\mathbf{G}$ into the n -vector $\mathbf{G}_N$ of observed genotypes and the m -vector $\mathbf{G}_M$ of unobserved genotypes.

Let p represent the frequency of allele 1 in the population from which the pedigree founders are assumed to be drawn, where $0 < p < 1$. Then, we have $E(\mathbf{G}) = p\mathbf{1}_{n+m}$, where $\mathbf{1}_{n+m}$ is a column vector of length $n + m$ with every entry equal to 1. If we further assume Hardy-Weinberg equilibrium (HWE) in the population from which the pedigree founders are drawn, then $\text{Var}(\mathbf{G}) = \sigma^2\Phi$, where $\sigma^2 = \frac{1}{2}p(1-p)$ and Φ is the kinship matrix given by

$$\Phi = \begin{pmatrix} 1+h_1 & 2\phi_{1,2} & \dots & 2\phi_{1,n+m} \\ 2\phi_{1,2} & 1+h_2 & \dots & 2\phi_{2,n+m} \\ \vdots & \dots & \dots & \vdots \\ 2\phi_{n+m,1} & 2\phi_{n+m,2} & \dots & 1+h_{n+m} \end{pmatrix}, \quad (1)$$

where $\phi_{i,j}$ is the kinship coefficient between individuals i and j , and h_i is the inbreeding coefficient of individual i . To make the approach more robust to deviations from HWE, we can remove the assumption that $\sigma^2 = \frac{1}{2}p(1-p)$ and simply assume $\text{Var}(\mathbf{G}) = \sigma^2\Phi$, where σ^2 is an unknown parameter that we estimate. Corresponding to the partition of $\mathbf{G}$ into $\mathbf{G}_N$ and $\mathbf{G}_M$, we have the following partition of the Φ matrix:

$$\Phi = \begin{pmatrix} \Phi_N & \Phi_{NM} \\ \Phi_{MN} & \Phi_M \end{pmatrix}, \quad (2)$$

where Φ_N is the $n \times n$ kinship matrix for the individuals with observed genotypes at the marker, $\Phi_{NM} = \Phi_{MN}^T$ is the $n \times m$ matrix of kinship coefficients for pairs of individuals in which one individual has observed genotype and the other has unobserved genotype, and Φ_M is the $m \times m$ kinship matrix for the individuals with unobserved

genotypes at the marker. Provided that the set of individuals with observed genotypes does not contain both members of any monozygotic (MZ) twin pair, the matrix Φ_N is invertible (Thornton and McPeck 2007).

Under these conditions, the best linear unbiased estimator (BLUE) of p is given by $\hat{p} = (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1} \mathbf{G}_N$, and the variance of this estimator is $\text{Var}(\hat{p}) = \sigma^2 (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1}$. McPeck et al. (2004) show that this is the BLUE based on the allele indicators for the set of genotyped individuals in the situation when parental origin of allele is not known.

Now we propose to predict $\mathbf{G}_M$ by finding its best linear unbiased predictor (BLUP) based on $\mathbf{G}_N$. That is, among all fixed $m \times n$ real matrices $\mathbf{R}$, we find the one that minimizes

$$E[\{\mathbf{c}^T (\mathbf{G}_M - \mathbf{R} \mathbf{G}_N)\}^2], \quad (3)$$

for every fixed m -vector $\mathbf{c}$, subject to

$$E(\mathbf{R} \mathbf{G}_N) = E(\mathbf{G}_M). \quad (4)$$

Equivalently, we find the $\mathbf{R}$ satisfying condition (4) such that $\text{Var}(\mathbf{G}_M - \tilde{\mathbf{R}} \mathbf{G}_N) - \text{Var}(\mathbf{G}_M - \mathbf{R} \mathbf{G}_N)$ is a positive semidefinite matrix for all $\tilde{\mathbf{R}}_{m \times n}$ satisfying condition (4). In our case, condition (4) reduces to $\mathbf{R} \mathbf{1} = \mathbf{1}_m$, where we economize on subscripts by letting $\mathbf{1}$ always denote a vector of length n with every entry equal to 1, while we let $\mathbf{1}_m$ denote a vector of length m with every element equal to 1. In the Appendix, we show that the resulting $\mathbf{R}$ is given by

$$\mathbf{R} = \mathbf{1}_m (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1} + \Phi_{MN} \Phi_N^{-1} - \Phi_{MN} \Phi_N^{-1} (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1}, \quad (5)$$

so that the BLUP of $\mathbf{G}_M$ is given by

$$\hat{\mathbf{G}}_M = \mathbf{R} \mathbf{G}_N = \hat{p} \mathbf{1}_m + \Phi_{MN} \Phi_N^{-1} (\mathbf{G}_N - \hat{p} \mathbf{1}). \quad (6)$$

The variance of the BLUP is

$$\text{Var}(\hat{\mathbf{G}}_M) = \sigma^2 \Phi_{MN} \Phi_N^{-1} \Phi_{MN}^T + \sigma^2 (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} (\mathbf{1}_m \mathbf{1}_m^T - \Phi_{MN} \Phi_N^{-1} \mathbf{1} \mathbf{1}^T \Phi_N^{-1} \Phi_{MN}^T). \quad (7)$$

If p were known, only the first term of equation (7) would be present; the second term of the variance results from estimation of p .

The above development holds for the case when the set of individuals with observed genotypes does not contain both members of any MZ twin pair. If the set of genotyped individuals contains both members of one or more MZ twin pairs, then Φ_N is not invertible, but the above results hold with Φ_N^{-1} replaced by the Moore-Penrose generalized inverse, which we write as Φ_N^- . Provided that the two members of any MZ twin pair have identical genotypes, use of Φ_N^- in place of Φ_N^{-1} in the above formulas is mathematically equivalent to setting the genotype of one of the two members of each MZ twin pair to be missing, in order to obtain invertible Φ_N .

Advantages of the BLUP, $\hat{\mathbf{G}}_M$, in equation (6), as a predictor of $\mathbf{G}_M$ are that (1) it is extremely fast to calculate, even in large complex pedigrees with inbreeding loops, making it feasible to use in studies with large numbers of people and markers; and (2) its variance-covariance matrix, given in equation (7), is also very easy to calculate, making it relatively easy to incorporate the prediction uncertainty and correlation between predicted values into a genetic analysis. Note that for different markers, the sets N and M of individuals with, respectively, non-missing and missing genotype, will change, and so the vectors $(\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1}$ and $\mathbf{R}$ will differ from marker to marker.

2.2. BLUP for multi-allelic markers

Now suppose the marker under consideration is observed to have a distinct alleles. Let $\mathbf{p} = (p_1, \dots, p_{a-1})$ be the vector of allele frequencies for alleles 1 through $a - 1$. (Note that $p_a = 1 - p_1 - \dots - p_{a-1}$, so p_a is redundant and can be dropped from $\mathbf{p}$.) For this subsection only, we redefine $\mathbf{G} = (\mathbf{G}^{(1)T}, \dots, \mathbf{G}^{(a-1)T})^T$, where $\mathbf{G}^{(j)T} = (G_1^{(j)}, \dots, G_{n+m}^{(j)})$, with $G_i^{(j)}$ equal to 0, 1/2 or 1, according to whether individual i has 0, 1, or 2 copies of the j th allele, $1 \leq j \leq a - 1$. Thus, the vector $\mathbf{G}$ now has length $(n + m)(a - 1)$. As before, the first n entries of each $\mathbf{G}^{(j)}$ are observed, while the last m entries are to be predicted. We write $\mathbf{G}^{(j)} = (\mathbf{G}_N^{(j)T}, \mathbf{G}_M^{(j)T})^T$, to denote the partition of the vector $\mathbf{G}^{(j)}$ into the n -vector $\mathbf{G}_N^{(j)}$ of observed genotypes and the m -vector $\mathbf{G}_M^{(j)}$ of unobserved genotypes.

In the multi-allelic case, we have $E(\mathbf{G}) = (\mathbf{I}_{a-1} \otimes \mathbf{1}_{n+m}) \mathbf{p}$, with $\otimes$ denoting Kronecker product and $\mathbf{I}_{a-1}$ denoting the identity matrix of dimension $a - 1$. Under HWE, we also have $\text{Var}(\mathbf{G}) = \mathbf{F} \otimes \Phi$, where

$\mathbf{F}_{(a-1) \times (a-1)}$ has (i,j) th entry equal to $\frac{1}{2}p_i(1-p_i)$ if $i=j$ and $-\frac{1}{2}p_i p_j$ if $i \neq j$. As shown by McPeck et al. (2004), the BLUE of $\mathbf{p}$ is given by $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_{a-1})$, where $\hat{p}_j = (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1} \mathbf{G}_N^{(j)}$, which is the BLUE for the frequency of allele j that would be calculated if the marker were treated as biallelic with one allele being j and all the other alleles being collapsed into a single “not- j ” allele. Furthermore, $\text{Var}(\hat{\mathbf{p}}) = (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{F}$.

The BLUP of $\mathbf{G}_M$ based on $\mathbf{G}_N$ is found to be $\hat{\mathbf{G}}_M^T = (\hat{\mathbf{G}}_M^{(1)T}, \dots, \hat{\mathbf{G}}_M^{(a-1)T})$, where $\hat{\mathbf{G}}_M^{(j)} = \hat{p}_j \mathbf{1}_m + \Phi_{MN} \Phi_N^{-1} (\mathbf{G}_N^{(j)} - \hat{p}_j \mathbf{1})$ is the BLUP of $\mathbf{G}_M^{(j)}$ that would be calculated if the marker were treated as biallelic with one allele being j and all the other alleles being collapsed into a single “not- j ” allele. We have

$$\text{Var}(\hat{\mathbf{G}}_M) = \mathbf{F} \otimes (\Phi_{MN} \Phi_N^{-1} \Phi_{MN}^T) + \mathbf{F} \otimes [(\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} (\mathbf{1}_m \mathbf{1}_m^T - \Phi_{MN} \Phi_N^{-1} \mathbf{1} \mathbf{1}^T \Phi_N^{-1} \Phi_{MN}^T)]. \quad (8)$$

Equation (8) is similar to equation (7) with σ^2 replaced by $\mathbf{F} \otimes$.

An important feature of the multiallelic BLUE and BLUP calculations is that they can be performed by, for example, calculating $(\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Phi_N^{-1}$ and $\mathbf{R}$ only once for a given multiallelic marker and then taking the inner products of these with $a-1$ different vectors $\mathbf{G}_N^{(j)}, j=1, \dots, a-1$.

2.3. Overview of the M_{QLS} method for case-control association testing with related individuals

Now we suppose that, in addition to having genotype data, we also have case-control phenotype data, where we allow some phenotype values to be unknown. Let $\mathbf{D}$ denote the phenotype data on the $n+m$ individuals, with each individual categorized as “affected,” “unaffected,” or “unknown phenotype.” Here, the designation “unknown phenotype” could be used to refer to, for example, an unphenotyped individual taken from a generic control panel. Alternatively, it could refer to an individual whose phenotype has not yet become apparent (e.g., for an age-related trait). For $1 \leq i \leq n+m$, we define the i th component of $\mathbf{D}$ to be $D_i = 1$ if i is affected, $D_i = 0$ if i is unaffected, and $D_i = k$ if i is of unknown phenotype, where k is an externally obtained estimate of the population prevalence of the trait. We write $\mathbf{D} = (\mathbf{D}_N^T, \mathbf{D}_M^T)^T$ to denote the partition of $\mathbf{D}$ into the vectors $\mathbf{D}_N$ and $\mathbf{D}_M$, corresponding to individuals with non-missing and missing genotype, respectively, at the marker being tested. For simplicity, we describe the M_{QLS} method for the case when the marker being tested is biallelic. The multiallelic case is given in Thornton and McPeck (2007).

We analyze the data retrospectively, i.e., we condition on $\mathbf{D}$ and treat $\mathbf{G}_N$ as random in the analysis. The retrospective approach is appropriate, for example, with either random or phenotype-based ascertainment. Our null hypothesis is that there is no association and no linkage between the marker being tested and the trait. Under the null hypothesis, we assume that $E_0(\mathbf{G}_N | \mathbf{D}) = p \mathbf{1}$ and $\text{Var}_0(\mathbf{G}_N | \mathbf{D}) = \sigma^2 \Phi_N$. The test statistic for the M_{QLS} method is given by

$$M_{QLS} = \frac{(\mathbf{V}^T \mathbf{G}_N)^2}{\hat{\sigma}^2 \mathbf{V}^T \Phi_N \mathbf{V}}, \quad (9)$$

with

$$\mathbf{V} = [\mathbf{I}_n - \Phi_N^{-1} \mathbf{1} (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} \mathbf{1}^T] [\mathbf{D}_N - k \mathbf{1} + \Phi_N^{-1} \Phi_{N,M} (\mathbf{D}_M - k \mathbf{1}_m)], \quad (10)$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. We typically use $\hat{\sigma}^2 = (n-1)^{-1} [\mathbf{G}_N^T \Phi_N^{-1} \mathbf{G}_N - (\mathbf{1}^T \Phi_N^{-1} \mathbf{1})^{-1} (\mathbf{1}^T \Phi_N^{-1} \mathbf{G}_N)^2]$, as proposed by Thornton and McPeck (2010). Under the null hypothesis of no linkage and no association between the marker and the trait, the M_{QLS} statistic given in equation (9) is asymptotically χ_1^2 distributed.

2.4. Previous interpretations of the M_{QLS} method

There are several possible ways of understanding the M_{QLS} statistic of equation (9). The original development of the M_{QLS} came from the fact that it is the quasi-likelihood score test of the null hypothesis $H_0 : \gamma = 0$ in the retrospective model $E(\mathbf{G}_N | \mathbf{D}) = p \mathbf{1} + \gamma [\Phi_N (\mathbf{D}_N - k \mathbf{1}) + \Phi_{N,M} (\mathbf{D}_M - k \mathbf{1}_m)]$, where γ represents the association parameter. For a genotyped individual i , this conditional expectation can be written

$$E(G_i | \mathbf{D}) = p + \gamma \sum_{j=1}^{n+m} 2\phi_{i,j} (D_j - k), \quad (11)$$

where we let $2\phi_{i,i} = 1 + h_i$. In expression (11), the association parameter, γ , is multiplied by a weighted sum of centered phenotype values, where the weight of individual j 's centered phenotype is twice the

kinship coefficient of individual j with individual i . This results in an enrichment effect, i.e., individuals with multiple affected relatives are assumed to have a higher chance of carrying a causal allele than individuals without affected relatives. For outbred individuals, it can be shown (Thornton and McPeck, 2007) that this retrospective model holds, up to terms of order $o(\gamma)$, assuming any prospective, two-allele, disease model, when the effect size γ is close to 0.

A later development (Wang and McPeck, 2009) shows that the M_{QLS} is closely connected to a retrospective likelihood score test based on the following prospective model: $P(\mathbf{D}|\mathbf{G}) = P_0(\mathbf{D})c(\mathbf{G}, r) \prod_i e^{rG_i D_i}$. Here $P_0(\mathbf{D})$ is the null model for the joint distribution of phenotype values in the absence of association with the marker being tested, where arbitrary dependence among phenotypes of related individuals is allowed. It is actually not necessary to specify the form of $P_0(\mathbf{D})$. $c(\mathbf{G}, r)$ is simply a normalizing constant. Logistic regression can be obtained as a special case of this prospective model. If we take into account incomplete data in deriving a retrospective likelihood score test for $r = 0$ under this class of models, then it is asymptotically equivalent to the M_{QLS} test, where the difference between them arises from the fact that the M_{QLS} test estimates the nuisance parameter p by the best linear unbiased estimator (BLUE), while the retrospective likelihood score test uses the maximum likelihood estimator of p under the null hypothesis.

2.5. A new interpretation of the M_{QLS} method

We describe a novel interpretation of the M_{QLS} statistic, which is somewhat different in flavor from previous interpretations and, therefore, can be illuminating. We note that the expression $\mathbf{V}^T \mathbf{G}_N$ in the numerator of the M_{QLS} test statistic of equation (9) can be rewritten as

$$\begin{aligned} \mathbf{V}^T \mathbf{G}_N &= (\mathbf{D}_N - k\mathbf{1})^T (\mathbf{G}_N - \hat{p}\mathbf{1}) + (\mathbf{D}_M - k\mathbf{1}_m)^T (\hat{\mathbf{G}}_M - \hat{p}\mathbf{1}_m) \\ &= \sum_{i \in L \cap N} (D_i - k)(G_i - \hat{p}) + \sum_{j \in L \cap M} (D_j - k)(\hat{G}_j - \hat{p}), \end{aligned} \quad (12)$$

where L is the set of phenotyped individuals, $\hat{p}$ is the BLUE of allele frequency, and $\hat{G}_j$ is the BLUP of G_j given by equation (6). The first sum is taken over all individuals who are both phenotyped and genotyped, and it represents the inner product of their genotypic and phenotypic residuals, where phenotype is centered around the externally derived prevalence estimate, k , and genotype is centered around the BLUE of allele frequency. The second sum is over all phenotyped individuals who have missing genotype at the marker being tested, in which case we replace the missing genotype G_j by its BLUP under the null hypothesis, $\hat{G}_j$. Thus, we can interpret the M_{QLS} as involving a form of imputation of missing genotypes by their BLUPs based on genotyped relatives. The main advantage of this form of imputation is that, under the null hypothesis, the uncertainty in imputation and dependence in imputation across individuals is exactly taken into account in the variance that appears in the denominator of equation (9).

2.6. How much power does BLUP imputation of missing genotypes add to the analysis?

To assess the effect, on the analysis, of BLUP imputation of missing genotypes, we perform analytical power calculations using a noncentral χ^2_1 approximation to the alternative distribution of the M_{QLS} statistic of equation (9). We illustrate this computation in two examples. In each case, we obtain a noncentrality parameter λ , and then calculate power at level 10^{-3} as $1 - R_{\lambda,1}(K)$, where $R_{\lambda,1}$ is the cumulative distribution function of a noncentral χ^2_1 with noncentrality parameter λ , and $K \approx 10.82757$ is the upper 10^{-3} quantile of a χ^2_1 distribution, i.e., $1 - R_{01}(K) = 10^{-3}$.

Example 1. Sib pairs. In this example, we assume that there are f sampled families, with each family consisting of a sib pair whose phenotypes are known. We assume that, in the resulting sample of size $2f$, the proportion of affected individuals is μ , and the correlation of affection status between sibs in the study is ρ . (Note that both μ and ρ reflect ascertainment. For example, if discordant sib pairs were preferentially ascertained, then ρ could be negative in the sample, even if there were positive sib-sib correlation for the trait in the general population.) For all 3 cases below, we use the mean model given in equation (11). We define a value s , which we call the *scaled genetic effect*, by $s = \sigma^{-2} f \mu (1 - \mu) \gamma^2$, and we assume $s > 0$. All the noncentrality parameters we calculate are proportional to s .

Case 1. Everyone genotyped. When all $2f$ individuals are both phenotyped and genotyped at the marker being tested, then we have $n = 2f$, $m = 0$, and

$$\Phi_N = \mathbf{I}_f \otimes \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}. \quad (13)$$

Plugging into equation (10), we obtain $\mathbf{v}_{complete} = \mathbf{D}_N - \mu \mathbf{1}_{2f}$. Then the noncentrality parameter is

$$\lambda_{complete} = \frac{(\mathbf{v}_{complete}^T E(\mathbf{G}_N | \mathbf{D}))^2}{\sigma^2 \mathbf{v}_{complete}^T \Phi_N \mathbf{v}_{complete}} = s(2 + \rho),$$

where s is defined in the previous paragraph.

Case 2. One sib in each pair genotyped, BLUP used. In this case, both sibs in each pair are phenotyped, and the sib who has missing genotype in each pair is chosen at random. Then $n = m = f$, $\Phi_N = \mathbf{I}_f$, $\Phi_{N,M} = \frac{1}{2} \mathbf{I}_f$, and, plugging into equation (10), we obtain $\mathbf{v}_{partial} = \mathbf{D}_N + \frac{1}{2} \mathbf{D}_M - \frac{3}{2} \mu \mathbf{1}_f$. Then the noncentrality parameter that results is

$$\lambda_{partial} = \frac{(\mathbf{v}_{partial}^T E(\mathbf{G}_N | \mathbf{D}))^2}{\sigma^2 \mathbf{v}_{partial}^T \Phi_N \mathbf{v}_{partial}} = s \left(\frac{5}{4} + \rho \right).$$

Case 3. One sib in each pair genotyped, missing sib discarded. In this case, although both sibs' phenotypes are observed, in each family, the sib with missing genotype is discarded from the analysis. This is in contrast to Case 2, in which that individual is still included. The calculation of $E(\mathbf{G}_N | \mathbf{D})$ in case 3 is exactly the same as for case 2. However, for calculation of $\mathbf{v}_{dropped}$, we use $n = f$, $m = 0$, and $\Phi_N = \mathbf{I}_f$. Then, plugging into equation (10), we obtain $\mathbf{v}_{dropped} = \mathbf{D}_N - \mu \mathbf{1}_f$. The resulting noncentrality parameter is $\lambda_{dropped} = s(1 + \frac{1}{2} \rho)^2$.

By comparison of cases 1, 2, and 3 for sib pairs, we see that $\lambda_{complete} > \lambda_{partial} \geq \lambda_{dropped}$, with equality between $\lambda_{partial}$ and $\lambda_{dropped}$ only when $\rho = 1$ or -1 . The greatest difference between $\lambda_{partial}$ and $\lambda_{dropped}$ occurs when $\rho = 0$. This makes sense because $\rho = 1$ or -1 corresponds to the case when the phenotype information on the ungenotyped individuals is completely redundant and provides no new information, while $\rho = 0$ corresponds to the phenotype information on the ungenotyped individuals being maximally informative (or at least, not well-predicted based on the phenotypes of their sibs). Power plots examining these three cases are given in Results.

Example 2. Sib quartets. The reason for considering the sib quartets example is that it is similar to the sib pairs example in many respects, except that, with two typed sibs per family, there is more information available on the missing genotypes, so the BLUP imputation might be expected to be more informative. We assume that there are f sampled families with each family consisting of a sib quartet whose phenotypes are known. In the resulting sample of $4f$ individuals, we assume that the proportion of affected individuals is μ , and the correlation of affection status between sibs in the study is ρ . For the sake of comparison, we use the same scaled genetic effect, $s = \sigma^{-2} f \mu (1 - \mu) \gamma^2$, that is used in the sib pair example.

Case 1. Everyone genotyped. When all $4f$ individuals are both phenotyped and genotyped at the marker being tested, then we have $n = 4f$, $m = 0$, and

$$\Phi_N = \mathbf{I}_f \otimes \begin{pmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{pmatrix}. \quad (14)$$

We obtain $\mathbf{v}_{complete} = \mathbf{D}_N - \mu \mathbf{1}_{4f}$, and $\lambda_{complete} = s(4 + 6\rho)$. Note that with four sibs, if every pair has correlation ρ , then we have the constraint $-1/3 \leq \rho \leq 1$, so $\lambda_{complete} \geq 2s > 0$.

Case 2. Two sibs in each quartet genotyped, BLUP used. In this case, all four sibs in each quartet are phenotyped, and the 2 sibs who have missing genotype in each quartet are chosen at random. Then $n = m = 2f$, Φ_N has the same form as in equation (13), and

$$\Phi_{N,M} = \mathbf{I}_f \otimes \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}. \quad (15)$$

We obtain $\mathbf{v}_{partial} = \mathbf{D}_N - \mu \mathbf{1}_{2f} + \frac{1}{3} (\mathbf{D}_M - \mu \mathbf{1}_{2f}) + \frac{1}{3} (\tilde{\mathbf{D}}_M - \mu \mathbf{1}_{2f})$, where

$$\tilde{\mathbf{D}}_M = \left(\mathbf{I}_f \otimes \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right) \mathbf{D}_M \quad (16)$$

is a permutation of the vector $\mathbf{D}_M$ such that the two ungenotyped sibs in each family have their phenotypes interchanged. Then $\lambda_{\text{partial}} = s \left(\frac{8}{3} + \frac{17}{3} \rho \right)$.

Case 3. Two sibs in each quartet genotyped, missing sibs discarded. In this case, although all four sibs' phenotypes are observed, in each family, the two sibs with missing genotypes are discarded from the analysis. We calculate $E(\mathbf{G}_N | \mathbf{D})$ in the same way as in case 2. However, for calculation of $\mathbf{v}_{\text{dropped}}$, we use $n = 2f$, $m = 0$, and the same Φ_N as in equation (13). Then $\mathbf{v}_{\text{dropped}} = \mathbf{D}_N - \mu \mathbf{1}_{2f}$, and $\lambda_{\text{dropped}} = s \frac{(2+3\rho)^2}{2+\rho}$.

By comparison of Cases 1, 2, and 3 for sib quartets, we see $\lambda_{\text{complete}} > \lambda_{\text{partial}} \geq \lambda_{\text{dropped}}$, with equality between λ_{partial} and λ_{dropped} only when $\rho = 1$ (recall the constraint $-1/3 \leq \rho \leq 1$ for sib quartets). Power plots for all cases are given in Results.

3. RESULTS

Figure 1 shows power results, at significance level 10^{-3} , for f sampled families, where each family is either a sib pair or a sib quartet. In each case, the solid line represents the situation when all members of the sibship are available to be analyzed (Case 1 above). The dotted line represents the situation when all individuals are phenotyped, but only half the members of the sibship are genotyped, and the individuals with missing genotype are simply dropped from the analysis (Case 3 above). The dashed line represents the situation when phenotype data are available on all members of the sibship, but genotype data are available on only half the members, and the M_{QLS} is used to analyze the data, which is equivalent to BLUP imputation of the missing genotypes for the other half of the sibship (Case 2 above). We expect the dashed line (BLUP imputation) to be intermediate between the dotted line (individuals with missing genotype discarded) and the solid line (complete information on the individuals), and it is of interest to get an idea of how much of the full information can be recovered by the BLUP, in the context of association testing.

FIG. 1. How much power for association is recovered by using BLUPs of missing genotypes? Approximate power at level 10^{-3} , as calculated using a non-central chi-square approximation, vs. scaled genetic effect, given by $\sigma^{-2}f\mu(1-\mu)\gamma^2$, for sib pairs or sib quartets and for different values of the correlation between sampled sib phenotypes, where this correlation would depend on ascertainment. In each plot, the solid line represents the situation in which the phenotypes and genotypes of all sibs are observed, representing the gold standard of perfect recovery of missing genotypes. The dashed line represents the situation in which the genotypes of half of the sibs (1 sib in a sib pair or 2 sibs in a sib quartet) are not observed, and they are incorporated into the M_{QLS} statistic, which is equivalent to BLUP imputation. The dotted line represents the situation in which the genotypes of half the sibs are not observed, and the ungenotyped sibs are simply removed from the analysis.

The sib quartet study design represents a larger sample, so it is to be expected that power should be higher than for the sib pair study design. In addition, we can see that the power is higher overall when the study design is such that the correlation between sampled sibs' phenotypes is higher. One explanation for this is the enrichment effect: for complex traits, an affected individual with an affected sibling is more likely to carry a particular variant associated with the phenotype than is an affected individual with an unaffected sibling. By the same token, an unaffected individual with an unaffected sibling is more likely to carry a particular protective variant than is an unaffected individual with an affected sibling. Therefore, we might expect to have higher power to detect a genetic effect when we sample individuals who have relatives with similar trait values. By comparing the results from complete data on the sibship (solid lines) to the results that ignore half of the sibs (dotted lines) for different values of the correlation, we see that the higher the phenotype correlation between the sibs, the less important it is to have the missing sib(s) in the analysis (solid and dotted lines get closer). This also seems reasonable, because if the missing sibs' phenotype(s) are well-predicted by the observed sibs' phenotype(s), then there should be less new information by including the additional sib(s).

Finally, we can see that in all 6 cases, the use of the BLUP imputation for missing genotypes (i.e., MQLS test) can provide a moderate power increase over ignoring individuals with missing genotypes. In particular, the MQLS method seems more effective with the sib quartet design, which is to be expected, because in that case there is more information available, on related individuals, to predict the missing genotypes.

4. DISCUSSION

We describe an interesting connection between the M_{QLS} method, for case-control association testing in samples of related individuals, and the imputation of missing genotypes by the best linear unbiased predictor based on relatives' genotypes. In examples, we show that the use of BLUP to predict missing genotypes can add a reasonable amount of power to detect association. The amount of power added is higher when there are more typed relatives available to improve the prediction of missing genotypes.

The BLUP imputation described here is single-point. In contrast, most current genotype imputation methods (Scheet and Stephens, 2006; Browning, and Browning 2009) use information across many markers. However, with related individuals, imputed genotypes are dependent among relatives, where the dependence among imputed genotypes differs from the ordinary dependence among genotypes and is affected by the type and amount of information available for each individual. For association mapping, this complex dependence among imputed genotypes would need to be taken into account in the analysis in order to construct a valid test for association. A key feature of the single-point BLUP imputation we describe here is that the dependence among imputed genotypes is exactly taken into account in the construction of the MQLS test, in a way that is fast and computationally feasible even for large, inbred pedigrees.

The BLUP that we use is constructed assuming that there is no population structure beyond that captured by Φ . Because the genotype prediction for an individual is based on genotypes of close relatives, one would expect it to be robust to mild population structure. The main difficulty for the BLUP would seem to be the possibility that the BLUE of allele frequency, $\hat{p}$ which is the centering value for the BLUP, could be inappropriate in cases of highly differentiated markers, when the allele frequencies are very different in different subpopulations represented in the sample. If information on population structure were available (e.g. in the form of structure-capturing vectors), then this information could be used to replace the vectors $\hat{p}\mathbf{1}$ and $\hat{p}\mathbf{1}_m$ of equation (6) by vectors in which the entry for the i th individual is an ancestry-specific estimated allele frequency. Alternatively, with mild population structure, in the context of case-control association testing, the ROADTRIPS method (Thornton and McPeck, 2010) could be used. ROADTRIPS is a more robust form of MQLS in which an estimated structure matrix $\tilde{\Psi}$ is used to correct the variance of the test statistic for misspecified relationships in Φ as well as for mild population structure.

5. APPENDIX

5.1. Proof that the BLUP is given by equations (5) and (6) with variance in equation (7)

As mentioned in subsection 2.1, we need to find $\tilde{\mathbf{R}}_{m \times n}$ minimizing $g(\mathbf{c}, \tilde{\mathbf{R}}) = E[\{\mathbf{c}^T(\mathbf{G}_M - \tilde{\mathbf{R}}\mathbf{G}_N)\}^2]$ for every m -vector $\mathbf{c}$, subject to condition (4), which is $\tilde{\mathbf{R}}\mathbf{1} = \mathbf{1}_m$. Note that condition (4) implies

$E(\mathbf{G}_M - \tilde{\mathbf{R}}\mathbf{G}_N) = \mathbf{0}$, so we have $g(\mathbf{c}, \tilde{\mathbf{R}}) = \mathbf{c}^T \text{Var}(\mathbf{G}_M - \tilde{\mathbf{R}}\mathbf{G}_N)\mathbf{c}$. Define $\mathbf{R}$ by equation (5). We can trivially write $\text{Var}(\mathbf{G}_M - \tilde{\mathbf{R}}\mathbf{G}_N) = \text{Var}[(\mathbf{G}_M - \mathbf{R}\mathbf{G}_N) + (\mathbf{R} - \tilde{\mathbf{R}})\mathbf{G}_N]$.

Claim: $\text{Cov}(\mathbf{G}_M - \mathbf{R}\mathbf{G}_N, (\mathbf{R} - \tilde{\mathbf{R}})\mathbf{G}_N) = \mathbf{0}$.

Proof: We have $\text{Cov}(\mathbf{G}_M - \mathbf{R}\mathbf{G}_N, (\mathbf{R} - \tilde{\mathbf{R}})\mathbf{G}_N) = \sigma^2(\Phi_{MN} - \mathbf{R}\Phi_N)(\mathbf{R} - \tilde{\mathbf{R}})^T$. Applying equation (5), we get $\Phi_{MN} - \mathbf{R}\Phi_N = (\Phi_{MN}\Phi_N^{-1}\mathbf{1} - \mathbf{1}_m)(\mathbf{1}^T\Phi_N^{-1}\mathbf{1})^{-1}\mathbf{1}^T$. Condition (4) and equation (5) imply $\mathbf{1}^T(\mathbf{R} - \tilde{\mathbf{R}})^T = \mathbf{0}$, so we have $(\Phi_{MN} - \mathbf{R}\Phi_N)(\mathbf{R} - \tilde{\mathbf{R}})^T = \mathbf{0}$, which proves the claim. ■

Thus, $g(\mathbf{c}, \tilde{\mathbf{R}}) = \mathbf{c}^T \text{Var}(\mathbf{G}_M - \tilde{\mathbf{R}}\mathbf{G}_N)\mathbf{c} = \mathbf{c}^T \text{Var}(\mathbf{G}_M - \mathbf{R}\mathbf{G}_N)\mathbf{c} + \mathbf{c}^T \text{Var}[(\mathbf{R} - \tilde{\mathbf{R}})\mathbf{G}_N]\mathbf{c}$. This is minimized for every m -vector $\mathbf{c}$ by $\tilde{\mathbf{R}} = \mathbf{R}$. When Φ_N is invertible, $\mathbf{R}$ is the unique minimizer, and the unique BLUP is given by $\mathbf{R}\mathbf{G}_N$. (When the set of individuals with observed genotypes contains both members of one or more MZ twin pairs, Φ_N is not invertible, and a minimizer $\mathbf{R}$ can be obtained by replacing Φ_N^{-1} with Φ_N^- in equation (5). In that case, $\mathbf{R}$ is not the unique minimizer, but any other minimizer $\mathbf{R}^*$ satisfies $\mathbf{R}^*\mathbf{G}_N = \mathbf{R}\mathbf{G}_N$, provided that the two members of any MZ twin pair have identical genotypes, so $\mathbf{R}\mathbf{G}_N$ is still the unique BLUP.) Expression (6) follows immediately from expression (5), using $\hat{\mathbf{p}} = (\mathbf{1}^T\Phi_N^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Phi_N^{-1}\mathbf{G}_N$, where both $(\mathbf{1}^T\Phi_N^{-1}\mathbf{1})^{-1}$ and $\mathbf{1}^T\Phi_N^{-1}\mathbf{G}_N$ are scalars. Expression (7) follows from the fact that $\text{Var}(\hat{\mathbf{G}}_M) = \mathbf{R}\Phi_N\mathbf{R}^T$.

5.2. Proof of equation (12)

We have $\mathbf{V}^T\mathbf{G}_N = [\mathbf{D}_N - k\mathbf{1} + \Phi_N^{-1}\Phi_{N,M}(\mathbf{D}_M - k\mathbf{1}_m)]^T[\mathbf{I}_n - \mathbf{1}(\mathbf{1}^T\Phi_N^{-1}\mathbf{1})^{-1}\mathbf{1}^T\Phi_N^{-1}]\mathbf{G}_N = [\mathbf{D}_N - k\mathbf{1} + \Phi_N^{-1}\Phi_{N,M}(\mathbf{D}_M - k\mathbf{1}_m)]^T(\mathbf{G}_N - \hat{\mathbf{p}}\mathbf{1}) = (\mathbf{D}_N - k\mathbf{1})^T(\mathbf{G}_N - \hat{\mathbf{p}}\mathbf{1}) + (\mathbf{D}_M - k\mathbf{1}_m)^T\Phi_{M,N}\Phi_N^{-1}(\mathbf{G}_N - \hat{\mathbf{p}}\mathbf{1}) = (\mathbf{D}_N - k\mathbf{1})^T(\mathbf{G}_N - \hat{\mathbf{p}}\mathbf{1}) + (\mathbf{D}_M - k\mathbf{1}_m)^T(\mathbf{G}_M - \hat{\mathbf{p}}\mathbf{1}_m)$, which is the first line of equation (12). The second line of equation (12) follows by noting that for $i \in N \setminus L$, $D_i - k = 0$ and similarly for $i \in M \setminus L$.

ACKNOWLEDGMENTS

This study was supported in part by the National Institutes of Health (grant R01 HG001645).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

- Bourgain, C., Hoffjan, S., Nicolae, R., et al. 2003. Novel case-control test in a founder population identifies P-selectin as an atopy susceptibility locus. *Am. J. Hum. Genet.* 73, 612–626.
- Browning, S.R., and Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097.
- Devlin, B., and Roeder, K. 1999. Genomic control for association studies. *Biometrics* 55, 997–1004.
- McPeck, M.S., Wu, X., and Ober, C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* 60, 359–367.
- Rabinowitz, D., and Laird, N. 2000. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.* 50, 211–223.
- Risch, N., and Teng, J. 1998. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Res.* 8, 1273–1288.
- Scheet, P., and Stephens, M. 2006. A fast and flexible statistical method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644.
- Slager, S., and Schaid, D.J. 2001. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am. J. Hum. Genet.* 68, 1457–1462.
- Thornton, T., and McPeck, M.S. 2007. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* 81, 321–337.
- Thornton, T., and McPeck, M.S. 2010. ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* 86, 172–184.

Wang, Z., and McPeck, M.S. 2009. An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. *JASA* 104, 1251–1260.

Address correspondence to:

*Dr. Mary Sara McPeck
Department of Statistics
University of Chicago
5734 South University Avenue
Chicago, IL 60637*

E-mail: mcpeek@galton.uchicago.edu