BIOMETRICS 60, 359–367
June 2004

# Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees

**Mary Sara McPeek,[1,2,*] Xiaodong Wu,[1,2,†] and Carole Ober[2]**

[1]Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago, Illinois 60637, U.S.A.
[2]Department of Human Genetics, University of Chicago, 920 E. 58th Street, Chicago, Illinois 60637, U.S.A.
[†]*Current Address:* Department of Preventive Medicine and Epidemiology, Loyola University,
2160 S. First Avenue, Maywood, Illinois 60153, U.S.A.
[*]*email:* mcpeek@galton.uchicago.edu

SUMMARY. Many types of genetic analyses depend on estimates of allele frequencies. We consider the problem of allele-frequency estimation based on data from related individuals. The motivation for this work is data collected on the Hutterites, an isolated founder population, so we focus particularly on the case in which the relationships among the sampled individuals are specified by a large, complex pedigree for which maximum likelihood estimation is impractical. For this case, we propose to use the best linear unbiased estimator (BLUE) of allele frequency. We derive this estimator, which is equivalent to the quasi-likelihood estimator for this problem, and we describe an efficient algorithm for computing the estimate and its variance. We show that our estimator has certain desirable small-sample properties in common with the maximum likelihood estimator (MLE) for this problem. We treat both the case when parental origin of each allele is known and when it is unknown. The results are extended to prediction of allele frequency in some set of individuals $S$ based on genotype data collected on a set of individuals $R$. We compare the mean-squared error of the BLUE, the commonly used naive estimator (sample frequency) and the MLE when the latter is feasible to calculate. The results indicate that although the MLE performs the best of the three, the BLUE is close in performance to the MLE and is substantially easier to calculate, making it particularly useful for large complex pedigrees in which MLE calculation is impractical or infeasible. We apply our method to allele-frequency estimation in a Hutterite data set.

KEY WORDS: Allele-frequency estimation; BLUE; BLUP; Complex pedigree; Quasi-likelihood.

## 1. Introduction

Chromosomes sampled from different individuals in a population typically show DNA sequence variation (i.e., polymorphism) at a number of locations (loci) throughout the genome. Polymorphisms are scientifically important for many reasons, including their usefulness for genetic mapping studies and for studies of population history. We refer to the copy of a polymorphic locus inherited by an individual from a particular parent as an *allele*, and we consider it a random variable. The $m$ possible observed types for an allele of a locus will be called *allelic types*. *Genotype data* refers to the observation of the two alleles for each individual in a sample. For a population of interest, the frequency distribution of allelic types, $a = (a_1, \ldots, a_m)^T$, where $a_i > 0$ is the frequency of the $i$th allelic type, $\sum_i a_i = 1$, is known as the *allele-frequency distribution*, and the $a_i$'s are known as *allele frequencies*. Many methods used in genetic mapping studies and studies of population history require estimates of allele frequencies. For instance, methods for genetic linkage analysis or association mapping often require such estimates, and the analyses may be sensitive to allele-frequency misspecification (e.g., Ott, 1992; see also Lockwood, Roeder, and Devlin, 2001).

Consider the empirical allele-frequency distribution based on genotype data from a sample of individuals from a population. This estimator is given by $\tilde{a} = (\tilde{a}_1, \ldots, \tilde{a}_m)^T$, with $\tilde{a}_i = (2n)^{-1} \sum_{j=1}^{2n} X_j^{(i)}$, where $X_j^{(i)}$ is the indicator of the event that the $j$th observed allele is of allelic type $i$, $n$ is the number of sampled individuals, and there are $2n$ alleles because each sampled individual has two alleles at each locus, one inherited from his or her mother and one inherited from his or her father. We will call $\tilde{a}$ the *naive estimator* of the allele-frequency distribution $a$. In the case of a simple random sample from an infinitely large, outbred population, assuming that alleles of randomly sampled individuals are independent and that the population is in Hardy–Weinberg equilibrium (i.e., the two alleles of an individual are independent draws from the allele-frequency distribution), then the only dependence in the data arises from the fact that $\sum_{i=1}^{m} X_j^{(i)} = 1$ for each $j$ in $1, \ldots, 2n$. In that case, the naive estimator, $\tilde{a}$, corresponds to the MLE of $a$ and has $\text{Cov}(\tilde{a}_i, \tilde{a}_j)$ equal to $a_i(1 - a_i)/2n$ if $i = j$ and $-a_i a_j/2n$ otherwise.

Data collection for mapping studies often involves sampling of families, rather than of individuals, in which case the naive estimator can typically be improved upon by taking into account the relatedness of individuals. In the extreme case of an isolated founder population, all individuals in the population may be related, and the population itself can often be thought of as a single extended family. We consider the problem of allele-frequency estimation based on individuals sampled from an isolated founder population. Here, again, sampling is typically family based rather than simple and random, and we perform inference conditional on the pedigree information. (We assume that the choice of sample is independent of the $X$'s, conditional on the pedigree, as is the case in the Hutterite data set we analyze.) Our results apply quite generally to allele-frequency estimation from data on individuals of known relationship, not only in isolated founder populations.

In a population that is so large that its size can be treated as effectively infinite, the definition of allele frequency may be relatively straightforward, at least in the absence of major population substructure. In contrast, in a finite-size population in which randomly sampled individuals are detectably related, genetic drift may have a greater effect, leading to allele frequencies that fluctuate noticeably across generations. Possible notions of allele frequency that may be of interest in an isolated founder population include (i) the allele frequency in the founding population, i.e., the population from which the founders' alleles are assumed to be randomly drawn (where founders are defined to be individuals in the pedigree whose parents are not in the pedigree); and (ii) the allele frequency corresponding to choosing an individual at random from the extant population and then choosing one of the individual's two alleles at random. Notion (ii) can be generalized to the problem of prediction of the allele frequency in a given subset of the pedigree. Notion (ii) is obviously of interest for, e.g., cross-population comparisons of allele frequencies, while linkage analysis typically requires (i). For use in various types of linkage disequilibrium studies, (i) is often the most useful, e.g., in the method of Abney, Ober, and McPeek (2002) for isolated founder populations. The naive estimator is commonly used for both the estimation problem in (i) and the prediction problem in (ii). The naive estimator is, in fact, an unbiased estimator of (i) and an unbiased predictor of (ii), but it makes inefficient use of the data (see results in Section 5).

Previously proposed approaches to allele-frequency estimation include maximum likelihood estimation (Fisher, 1940; Ceppellini, Siniscalco, and Smith, 1955; Boehnke, 1991) and various types of linear estimation (Cotterman, 1947; Chakraborty, 1978; Olson, 1994). Finney (1948a,b) proposes an interactive scoring method. Broman (2001) compares five different estimators for allele frequency based on sibship data. In principle, maximum likelihood estimation could be applied to an isolated founder population, with the entire population treated as a single large family. This would yield the MLE of (i), the allele frequency in the founding population. However, in the Hutterite pedigree we consider, calculation of the MLE, by either "exact" or Markov-chain Monte Carlo methods, is impractical due to the computationally intensive nature of the calculations and the frequency with which they must be carried out on a routine basis.

Our estimator for problem (i) above and our predictor for problem (ii) above are the best linear unbiased estimator (BLUE) and best linear unbiased predictor (BLUP), respectively, under the assumption that the founder alleles are independent draws from a common distribution and that the loci follow the laws of Mendelian inheritance conditional on the pedigree. Thus, our estimators are at least as efficient and usually more efficient than the previously proposed linear estimators under those assumptions. In the case when the founding population does not satisfy the assumption of Hardy–Weinberg equilibrium, our estimators are still unbiased, as are the previously proposed linear estimators for problem (i). Note that we do not require Hardy–Weinberg equilibrium within the isolated founder population itself; indeed, even with random mating in a sufficiently small isolated founder population, Hardy–Weinberg equilibrium would not hold because of inbreeding (in the case when allele frequency is defined to be the frequency in the founding population). The effects of the pedigree structure, including inbreeding, are explicitly taken into account in our estimators. Our methods are computationally simple and applicable to any type of family data, including large, complex pedigrees. We apply the methods to a Hutterite data set involving ∼800 individuals drawn from a 13-generation, 1623-member pedigree, with virtually all individuals related through multiple lines of descent.

## 2. BLUE for Estimation of Frequency in Founding Population

We first consider problem (i), estimation of the allele-frequency distribution in the founding population. In Section 2.1, we derive the BLUE in the case when, for each allele of an individual, we have the information of whether it was inherited from the individual's mother or the individual's father. In Section 2.2, we treat the case when this additional information of parent-of-origin for each allele is not available. Section 3 gives some properties shared by the BLUE and MLE. Then, in Section 4, we consider problem (ii), prediction of the allele frequency in a given subset of the pedigree, for which we derive the BLUP.

### 2.1 *BLUE When Parental Origin of Allele Is Observed*

For simplicity of presentation, we initially suppose that the number of allelic types, $m$, at the locus is two, so that the parameter of interest is a scalar. To streamline the notation, we write $a$ in place of $a_1$, $X_i$ in place of $X_i^{(1)}$, and let $X = (X_1, \ldots, X_{2n})^T$. Then we have $E(X_j) = a$ and $\text{Var}(X_j) = a(1 - a)$ for each $j$, and there is correlation between the indicators that results from the relationships among the individuals in the pedigree.

A linear estimator of $a$ is an estimator $\delta$ of the form $\delta = w^T X$, where $w$ is a known $2n \times 1$ weight vector. Consider the class of linear unbiased estimators of $a$, $\{\delta\colon \delta = w^T X$ with $w^T 1 = 1\}$, where 1 is a column vector of 1's of length $2n$. Let $C$ be the covariance matrix with $(i, j)$th entry $C_{(i,j)} = \text{Cov}(X_i, X_j)$, and assume that $C$ is known up to a constant multiple and invertible. Then the BLUE, i.e., the estimator with smallest variance among all unbiased estimators that are linear in $X$, is $(1^T C^{-1} 1)^{-1} 1^T C^{-1} X = \sum_{j=1}^{2n} w_j X_j$, where $w_j = (1^T C^{-1} 1)^{-1} (1^T C^{-1})_j$ (e.g., see Lehmann and Casella, 1998, p. 130). That is, $w_j$ is proportional to the sum of the $j$th row (equivalently, column) of $C^{-1}$. (More generally, if $E(X) = da$, where $d$ is a known, nonzero $2n \times 1$ vector, then the BLUE of $a$ is $(d^T C^{-1} d)^{-1} d^T C^{-1} X$.) The naive estimator is obtained in the special case in which $(1^T C^{-1})_j$ is the same for all $j$, e.g., when the indicator random variables are uncorrelated or when they are correlated but exchangeable or when there is a group of permutations of the alleles that acts transitively and that preserves the covariance matrix.

Under the assumptions that the founder alleles are independent draws from a common distribution and that the loci follow the laws of Mendelian inheritance conditional on the pedigree, the covariance matrix $C$ can be written as $a(1-a)K$, where $K$ is the correlation matrix. $K$ has 1's on the diagonal and $K_{i,j}$, $i \neq j$, can be described as follows: Suppose that allele $i$ represents the allele individual $k$ inherited from parent $m$ and allele $j$ represents the allele individual $l$ inherited from parent $n$. Then $K_{i,j} = \phi_{m,n}$, the kinship coefficient between individuals $m$ and $n$, where the kinship coefficient between two individuals is the chance that a randomly chosen pair of alleles, one from each individual, is identical by descent, i.e., is an inherited copy of the same founder allele. This value is determined solely from the pedigree graph, without reference to genotype data. For example, the kinship coefficient between an outbred parent–offspring or sib pair is 1/4, while that between an outbred grandparent–grandchild, avuncular, or half-sib pair is 1/8. Kinship coefficients for all pairs of individuals in a pedigree can be efficiently computed by a recursive algorithm (Boyce, 1983). We note that if the sampled individuals are distinct, then $K$ is necessarily nonsingular except in the case when the sample contains monozygous twins (in which case we suggest that genotype data on only one twin be used, leading to a nonsingular $K$). Then the BLUE can be written as

$$\hat{a} = (1^T K^{-1} 1)^{-1} 1^T K^{-1} X, \tag{1}$$

where $K$ does not depend on $a$. The variance of $\hat{a}$ is easily seen to be

$$\mathrm{Var}(\hat{a}) = (1^T K^{-1} 1)^{-1} a(1-a). \tag{2}$$

In comparison, the variance of the naive estimator $\tilde{a}$ is $(2n)^{-2}(1^T K 1)a(1-a)$, so the relative efficiency of the BLUE compared to the naive estimator is $(1^T K^{-1} 1)(1^T K 1)/(2n)^2 \geq 1$, and the correlation between these two estimators is $2n\{(1^T K^{-1} 1)(1^T K 1)\}^{-1/2}$. We note that, in the special case when the sample consists of $\eta$ independent families having indicator vectors $Y^1, \ldots, Y^\eta$ and correlation matrices $K_1, \ldots, K_\eta$, then in the formulae above, $1^T K^{-1} X$ reduces to $\sum_{i=1}^{\eta} 1^T K_i^{-1} Y^i$, $1^T K^{-1} 1$ reduces to $\sum_{i=1}^{\eta} 1^T K_i^{-1} 1$, and $1^T K 1$ reduces to $\sum_{i=1}^{\eta} 1^T K_i 1$.

Now suppose that the number of allelic types $m > 2$. Let $a = (a_1, \ldots, a_{m-1})^T$ and $X = (X^{(1)T}, \ldots, X^{(m-1)T})^T$, where $X^{(i)T} = (X_1^{(i)}, \ldots, X_{2n}^{(i)})$. An obvious extension of the notion of a BLUE is that if $\hat{a} = AX$ with $A$ a known $(m-1) \times 2n(m-1)$ matrix, then $\hat{a}$ is a BLUE of $a$ if and only if $c^T\hat{a}$ is a BLUE of $c^T a$ for every choice of real vector $c$ of length $m-1$. In our case, we have $E(X) = \tilde{D}a$, where $\tilde{D} = I_{m-1} \otimes 1_{2n}$, $I_{m-1}$ is the identity matrix of dimension $m-1$, $1_{2n}$ is a column vector of 1's of length $2n$, and $\otimes$ is Kronecker product. Let $\mathrm{Var}(X) = \tilde{C}$. First suppose that $\tilde{C}$ is a known, invertible matrix. Then it can be shown that the BLUE $\hat{a}$ of $a$ is given by

$$\hat{a} = (\tilde{D}^T \tilde{C}^{-1} \tilde{D})^{-1} \tilde{D}^T \tilde{C}^{-1} X. \tag{3}$$

In fact, in our case $\tilde{C} = F \otimes K$, where $F$ is an $(m-1) \times (m-1)$ matrix with $F_{ij} = a_i(1-a_i)$ if $i = j$ and $-a_i a_j$ if $i \neq j$, so $\tilde{C}$ depends on $a$. However, by manipulation of equation (3) with the expressions for $\tilde{D}$ and $\tilde{C}$ plugged in, we find that

$$\begin{aligned}\hat{a} &= \left[I_{m-1} \otimes \left\{\left(1_{2n}^T K^{-1} 1_{2n}\right)^{-1} 1_{2n}^T K^{-1}\right\}\right] X \\ &= ((1^T K^{-1} 1)^{-1} 1^T K^{-1} X^{(1)}, \ldots, (1^T K^{-1} 1)^{-1} 1^T K^{-1} X^{(m-1)}),\end{aligned} \tag{4}$$

which depends only on $X$ and the pedigree, not on $a$, and which can be equivalently expressed as follows: For each $i$, let $\hat{a}_i$ be the estimator obtained from equation (1) with $X^{(i)}$ in place of $X$ (i.e., collapsing the observed alleles into two classes, $i$ and not-$i$). Then the BLUE of $a$ is $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_{m-1})^T$, and $\sum_{i=1}^{m} \hat{a}_i = 1$. From either equation (3) or (4), the variance of the BLUE is easily found to be

$$\mathrm{Var}(\hat{a}) = (1^T K^{-1} 1)^{-1} F. \tag{5}$$

*Remark* 1. Equation (3) gives a convenient mathematical form for $\hat{a}$, but it is expressed in terms of the inverse of a $2n(m-1)$-dimensional matrix. Equation (4) shows that this calculation reduces to $(m-1)$ calculations, each of which involves the inverse of the same $2n$-dimensional matrix.

*Remark* 2. Our definition of a multivariate BLUE $\hat{a}$ of $a$ (which was that $c^T\hat{a}$ be a BLUE of $c^T a$ for every choice of real vector $c$ of length $m-1$) can be shown to be equivalent to the superficially weaker requirement that each component of $\hat{a}$ be a one-dimensional BLUE of the corresponding component of $a$. This equivalence follows from the fact that a (one-dimensional) linear unbiased estimator is BLUE if and only if it is uncorrelated with every linear unbiased estimator of 0 (see Lehmann and Casella, 1998, p. 130). However, this fact, combined with equation (1), does not directly yield equation (4). To obtain equation (4) by this reasoning, one would need the further information that the BLUE for $a_i$ based on $X$ is the same as the BLUE for $a_i$ based on $X^{(i)}$.

*Remark* 3. It can be seen from equations (3) and (4) that the BLUE is equivalent to the quasi-likelihood estimator (Wedderburn, 1974; see McCullagh and Nelder, 1989 for details) for this problem. The quasi-likelihood score function can generally be written as $U(a) = D^T V^{-1}(X - \mu)$, where $\mu = E(X)$ is assumed to be a known function of $a$, $D$ is the matrix whose $(i, j)$th element is $\partial\mu_i/\partial a_j$, and $V$ is the covariance matrix of $X$, where $V$ is assumed to be a known function of $a$ (possibly up to an unknown scale factor) and invertible. In our case, $\mu = a \otimes 1_{2n}$, $D = \tilde{D}$, and $V = \tilde{C}$. The quasi-likelihood estimator for $a$ is the solution of $U(a) = 0$, which, in our case, is given by equations (3) and (4).

*Remark* 4. The variance of the BLUE $c^T\hat{a}$ is $(1^T K^{-1} 1)^{-1} c^T F c$, while the variance of the naive estimator $c^T\tilde{a}$ is $(2n)^{-2}(1^T K 1) c^T F c$. Thus, the relative efficiency of the BLUE $c^T\hat{a}$ compared to the naive estimator $c^T\tilde{a}$ is again $(1^T K^{-1} 1)(1^T K 1)/(2n)^2 \geq 1$, and the correlation between these two estimators is again $2n\{(1^T K^{-1} 1)(1^T K 1)\}^{-1/2}$, just as for the case $m = 2$, with neither expression depending on $c$.

### 2.2 *BLUE When Parental Origin of Allele Is Not Observed*

A difficulty is that we do not actually observe which of an individual's two alleles is maternally inherited and which is paternally inherited, although this can sometimes be inferred from the data. To construct a linear estimator in that case, we assume that this information is always unavailable and that an individual's two alleles are given an arbitrary labeling that is independent of which is maternally and which is paternally inherited. In that case, the off-diagonal elements of the correlation matrix $K$ take a different form. If $i$ and $j$, $i \neq j$, index the two alleles within a single individual $k$, then $K_{i,j} = h_k$, the inbreeding coefficient of individual $k$, which is equal to the kinship coefficient between the parents of $k$, or, equivalently, the probability that the two alleles of $k$ are identical by descent. If $i$ and $j$ index alleles taken from two different individuals $k$ and $l$, then $K_{i,j} = \phi_{k,l}$, the kinship coefficient between individuals $k$ and $l$. Except for this change in $K$, all the results of the previous subsection still hold.

We now briefly describe the computational issues involved in obtaining the BLUE and show that when parental origin of allele is not observed, the calculations can be simplified, resulting in faster and less memory-intensive computations. To obtain the BLUE, we need to calculate $(1^T K^{-1} 1)^{-1} 1^T K^{-1} X^{(i)}$ for $i = 1, \ldots, m-1$, where $X^{(i)} = (X_1^{(i)}, \ldots, X_{2n}^{(i)})^T$. We efficiently compute the BLUE by taking the Cholesky decomposition of $K$, i.e., finding upper triangular $B$ such that $B^T B = K$, using an algorithm that simultaneously computes $b_0 = B^{-T}1$ and $b_i = B^{-T}X^{(i)}$, $i = 1, \ldots, m-1$ at little extra cost (Graybill, 1976). Then we can take $\hat{a}_i = (b_0^T b_i)/(b_0^T b_0)$, and the BLUE is $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_{m-1})^T$. Note that with a sample of $n$ individuals, the matrix $K$ is $2n \times 2n$. For instance, in one of the Hutterite samples we consider, $n = 806$, so $K$ is $1612 \times 1612$. Thus, the Cholesky decomposition step can be slow. Typically, the calculation would be performed for each of a large number of loci throughout the genome. At different loci, different individuals may have missing genotype data, and the estimation at each locus is based only on the nonmissing data for that locus. Thus, $K$ typically differs from locus to locus, and the Cholesky decomposition must be recomputed at each locus. We are able to obtain an improvement in efficiency and reduction in needed dynamic memory by using the following result:

*Notation.* Recall that the BLUE is given by $\hat{a} = (\hat{a}_1, \ldots, \hat{a}_{m-1})^T$, with $\hat{a}_i = (1_{2n}^T K^{-1} 1_{2n})^{-1} 1_{2n}^T K^{-1} X^{(i)}$, where $1_{2n}$ is $2n \times 1$, $K$ is $2n \times 2n$, and $X^{(i)}$ is $2n \times 1$. Let $L$ be the $n \times n$ matrix with $(i, j)$th element equal to $1 + h_i$ if $i = j$ and $2\phi_{ij}$ if $i \neq j$, let $Z^{(i)} = (Z_1^{(i)}, \ldots, Z_n^{(i)})^T$, where $Z_j^{(i)} = \frac{1}{2}$ (the number of copies [0, 1, or 2] of allele $i$ held by individual $j$), and let $1_n$ be $n \times 1$.

PROPOSITION 1: *We have* $\hat{a}_i = (1_n^T L^{-1} 1_n)^{-1} 1_n^T L^{-1} Z^{(i)}$.

*Proof.* See the Appendix.

*Remark* 1. This proposition allows us to perform our Cholesky decomposition on an $n \times n$ matrix, $L$, instead of a $2n \times 2n$ matrix, $K$, in the case when parental origin of allele is not observed.

*Remark* 2. The variance of $\hat{a}$ can also be more efficiently computed using this simplification. As a consequence of the proposition, we have $\mathrm{Var}(\hat{a}) = \frac{1}{2}(1_n^T L^{-1} 1_n)^{-1} F$.

## 3. Some Properties Shared by the BLUE and MLE

Suppose one or more pedigrees are sampled, and it is desired to estimate the allele frequency in the founding population. In the case when the set of genotyped individuals includes all founders of the pedigree(s), it is natural to take as the estimator the observed frequency in the founders, because the data from succeeding generations do not add additional information on allele frequency in the founding population. It turns out that both the MLE and the BLUE result in the natural estimator in this case. A related result, which also holds for both the MLE and BLUE, is that if an entire subpedigree within a pedigree is genotyped, then only the founders of the subpedigree need be considered. In the case of the BLUE, these results follow from the following more general result:

*Notation.* Let $X_{k\times 1}$ be random with $E(X) = Da$ and $\mathrm{Var}(X) = \Sigma$, where $a_{r\times 1}$ is unknown, $D_{k\times r}$ is known, and $\Sigma_{k\times k}$ is known up to a constant multiple. Write $X = (X_1^T, X_2^T)^T$, where $X_1$ is $k_1 \times 1$ and $X_2$ is $(k - k_1) \times 1$, write $D = (D_1^T, D_2^T)^T$, where $D_1$ is $k_1 \times r$ and $D_2$ is $(k - k_1) \times r$, and write

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix},$$

where $\Sigma_{11}$ is $k_1 \times k_1$, $\Sigma_{12}$ is $k_1 \times (k - k_1)$, and $\Sigma_{22}$ is $(k - k_1) \times (k - k_1)$. Here, $E(X_i) = D_i a$, $\mathrm{Var}(X_i) = \Sigma_{ii}$ for $i = 1, 2$, and $\mathrm{Cov}(X_1, X_2) = \Sigma_{12}$.

PROPOSITION 2: *If* $\Sigma_{11}$ *is invertible, rank* $(D_1) = r$, *and*

$$\Sigma_{12}^T \Sigma_{11}^{-1} D_1 = D_2, \tag{6}$$

*then the BLUE of* $a$ *depends only on* $X_1$ *and is given by* $(D_1^T \Sigma_{11}^{-1} D_1)^{-1} D_1^T \Sigma_{11}^{-1} X_1$.

*Proof.* See the Appendix.

*Remark.* For our problem, if we let $X_1$ represent the genotype data of certain individuals, then equation (6) is equivalent to $K_{12}^T K_{11}^{-1} 1 = 1$, where $K_{11} = \mathrm{Corr}(X_1^{(i)})$ and $K_{12} = \mathrm{Corr}(X_1^{(i)}, X_2^{(i)})$.

Let $\hat{a}_A$ be the BLUE of allele frequency based on all the genotyped individuals. Now consider the following recursive algorithm for pruning the data set: For each individual $i$ who does not have any genotyped descendants in the data set, if $i$'s parents are both genotyped, then remove $i$ from the data set. This algorithm can be started at the bottom of a pedigree and can be applied recursively up the pedigree. Let $\hat{a}_B$ be the BLUE of allele frequency based on the data remaining after pruning. Then we have the following corollary.

COROLLARY TO PROPOSITION 2: (i) $\hat{a}_A = \hat{a}_B$, *i.e., the BLUE for the pruned data set is the same as the BLUE for the original data set.* (ii) *When the set of genotyped individuals* *includes all the founders of the sampled pedigrees, the BLUE is the observed frequency in the founders.*

*Proof.* See the Appendix.

*Remark* 1. These results apply both in the case when the information of which allele is maternal and which is paternal is available and in the case when this information is not available.

*Remark* 2. The MLE also has these properties (see the Appendix). It is obvious that the naive estimator does not.

*Remark* 3. Result (i) still holds when the pruning algorithm is generalized to remove from the data set any allele $i$ such that every directed path in the pedigree from any founder allele to $i$ passes through at least one genotyped allele besides $i$.

## 4. BLUP for Estimation of Frequency in a Given Subset of Individuals

Let $Q$ be the set of individuals on whom pedigree information is available, and let $R \in Q$ be the subset of individuals for whom genotype data are available. Suppose we wish to predict the allele frequency in a target set of individuals $S \in Q$, where $S$ might intersect with $R$.

*Notation.* Let $|R| = r$, $|S| = s$, $|R \cap S| = u$. For $i = 1, \ldots, m$, let $X^{(i)}_{R\backslash S}$ be the $2(r - u)$-vector of indicators for allele $i$ among the (ordered) individuals in $R\backslash S$. Similarly, let $X^{(i)}_{R\cap S}$ and $X^{(i)}_{S\backslash R}$ be the $2u$-vector and $2(s - u)$-vector indicators for allele $i$ among the individuals in $R \cap S$ and $S\backslash R$, respectively. Define $X = (X^{(1)T}, \ldots, X^{(m-1)T})^T$, where $X^{(i)} = (X^{(i)T}_{R\backslash S}, X^{(i)T}_{R\cap S}, X^{(i)T}_{S\backslash R})^T$. Define $X_R = (X^{(1)T}_R, \ldots, X^{(m-1)T}_R)^T$ and $X_S = (X^{(1)T}_S, \ldots, X^{(m-1)T}_S)^T$, where $X^{(i)}_R = (X^{(i)T}_{R\backslash S}, X^{(i)T}_{R\cap S})^T$ and $X^{(i)}_S = (X^{(i)T}_{R\cap S}, X^{(i)T}_{S\backslash R})^T$. For any integer $k$, define $\tilde{D}_k = I_{m-1} \otimes 1_k$.

Let $Y = (2s)^{-1}\tilde{D}^T_{2s} X_S$ be the quantity we wish to predict, i.e., $Y$ is just the vector of length $m - 1$ with $i$th component equal to the sample average, over all individuals in $S$, of the indicators for allele $i$. We observe the quantity $X_R$, and we propose to predict $Y$ by finding the BLUP, i.e., the linear combination of $X_R$, $AX_R$, that minimizes $E[\{c^T(Y - AX_R)\}^2]$, for all $(m - 1)$-vectors $c$, subject to $E(AX_R) = E(Y)$. We apply the following result:

*Notation.* Let $X$, $D$, $\Sigma$, and $a$ be defined as in Proposition 2. Suppose $W = GX$ is observed where $G_{p\times k}$ is known, and it is desired to predict $BX$, where $B_{s\times k}$ is known. Consider linear predictors $AW$ of $BX$, where $A$ is $s \times p$. Let $\Gamma = \{A\colon E(AW - BX) = 0\}$. Find $A^* \in \Gamma$ that minimizes $E[\{c^T(AW - BX)\}^2]$ for all real $r$-vectors $c$.

Proposition 3: *If $G\Sigma G^T$ and $D^T G^T (G\Sigma G^T)^{-1} GD$ are invertible, then there is a unique minimizer $A^*$, and the resulting predictor is $A^*W = B\{D\hat{a} + \Sigma_{X,W}\Sigma^{-1}_W(W - GD\hat{a})\}$, where $\hat{a} = (D^T G^T (G\Sigma G^T)^{-1} GD)^{-1} D^T G^T (G\Sigma G^T)^{-1} W, \Sigma_W = G\Sigma G^T$, and $\Sigma_{X,W} = \Sigma G^T$.*

*Proof.* See the Appendix.

In our case, $D = \tilde{D}_{2(r+s-u)}, G = I_{m-1} \otimes (I_{2r}, 0_{2r\times 2(s-u)})$ where $(I_{2r},\ 0_{2r\times 2(s-u)})$ is a $2r \times 2(r + s - u)$ matrix whose first $2r$ columns are the identity matrix $I_{2r}$ and whose last $2(s - u)$ columns are 0, $B = (2s)^{-1}[I_{m-1} \otimes (0^T_{2(r-u)},\ 1^T_{2s})]$, where $(0^T_{2(r-u)},\ 1^T_{2s})$ is a row vector with first $2(r - u)$ elements equal to 0 and last $2s$ elements equal to 1, and $\Sigma = F \otimes K$, where $F$ is as before and $K$ is the correlation matrix of $X^{(i)}$ which can be written as

$$K = \begin{bmatrix} K_{R,R} & K_{R,S\backslash R} \\ K_{S\backslash R,R} & K_{S\backslash R,S\backslash R} \end{bmatrix},$$

where, e.g., $K_{R,R} = \operatorname{Corr}(X^{(i)}_R)$ and $K_{R,S\backslash R} = \operatorname{Corr}(X^{(i)}_R, X^{(i)}_{S\backslash R})$.

Corollary to Proposition 3: *For our problem, we obtain the following BLUP for $Y$:*

$$\begin{aligned} AX_R = (2s)^{-1}\Big[&\tilde{D}^T_{2u} X_{R\cap S} + 2(s-u)\hat{a} \\ &+ \Big\{I_{m-1} \otimes \Big(1^T_{2(s-u)} K_{S\backslash R,R} K^{-1}_{R,R}\Big)\Big\}(X_R - \tilde{D}_{2r}\hat{a})\Big], \end{aligned}$$

*where $\hat{a} = (1^T_{2r} K^{-1}_{R,R} 1_{2r})^{-1}\{I_{m-1} \otimes (1^T_{2r} K^{-1}_{R,R})\} X_R$ is the BLUE for a in terms of $X_R$ (see equation [4]).*

*Remark* 1. The $i$th component of $AX_R$ predicts the sample average of the indicator of allele $i$ in population $S$ and is equal to $(2s)^{-1}\{1^T_{2u} X^{(i)}_{R\cap S} + 2(s - u)\hat{a}_i + 1^T_{2(s-u)} K_{S\backslash R,R} K^{-1}_{R,R} (X^{(i)}_R - 1_{2r}\hat{a}_i)\}$, where $1^T_{2u} X^{(i)}_{R\cap S}$ is just the sum of indicators for the individuals in $R \cap S$, and the remaining part of the expression comes from the generalized regression of $X^{(i)}_{S\backslash R}$ on $X^{(i)}_R$.

*Remark* 2. The standard error of prediction of the BLUP is given by

$$\begin{aligned} \text{s.d.}\big(c^T(AX_R - Y)\big) = (2s)^{-1}(c^T F c)^{1/2} & \\ \times\Big[1^T K_{R\backslash S,R\backslash S} 1 + 2\big(1^T K_{R\backslash S,S\backslash R} 1\big)& \\ + 1^T K_{S\backslash R,S\backslash R} 1 + 4(s-u)^2\alpha - \gamma& \\ + \alpha\beta^2 - 4(s-u)\alpha\beta\Big]^{1/2}&, \end{aligned}$$

where $\alpha = (1^T K^{-1}_{R,R} 1)^{-1}$, $\beta = 1^T K^{-1}_{R,R} K_{R,S\backslash R} 1$, and $\gamma = 1^T K_{S\backslash R,R} K^{-1}_{R,R} K_{R,S\backslash R} 1$.

*Remark* 3. For comparison, we can define the *naive predictor* of $Y$ by $(2s)^{-1}(1^T_{2u} X_{R\cap S} + 2(s - u)\bar{X}_R)$, which is also an unbiased predictor, with standard error of prediction equal to $(2s)^{-1}(c^T F c)^{1/2}[-1^T K_{R\cap S,R\cap S} 1 + ((s - u)^2/r^2) 1^T K_{R,R} 1 + 1^T K_{S,S} 1 - 2(1^T K_{R\cap S,S\backslash R} 1) - (2(s - u)/r) 1^T K_{R,S\backslash R} 1]^{1/2}$.

## 5. Application to Allele-Frequency Estimation in the Hutterites

### 5.1 *Data Analysis*

The data are from a Hutterite population, with the approximately 800 sampled individuals virtually all related through multiple lines of descent. Their relationships are characterized by a known, 13-generation, 1623-person pedigree (Abney,

**Table 1**

*Allele-frequency estimates for five SNPs in the Hutterites. Naive estimate and BLUE are estimates of founder allele frequency, while naive prediction and BLUP are predictions of frequency in full sample of size* 858 *based on genotyped subsample.*

| SNP | Subsample size | Naive estimate (s.e.$_1$) | BLUE (s.e.$_1$) | Naive prediction (s.e.$_2$) | BLUP (s.e.$_2$) |
|---|---|---|---|---|---|
| TLR4+896 | 755 | 0.04 (0.04) | 0.05 (0.04) | 0.036 (0.002) | 0.034 (0.002) |
| IL13-110 | 734 | 0.17 (0.08) | 0.12 (0.06) | 0.170 (0.005) | 0.167 (0.003) |
| IFNG-1616 | 729 | 0.19 (0.08) | 0.19 (0.07) | 0.195 (0.006) | 0.199 (0.003) |
| IL4R-3223 | 383 | 0.33 (0.10) | 0.30 (0.09) | 0.328 (0.014) | 0.322 (0.009) |
| IL10-1117 | 736 | 0.43 (0.10) | 0.40 (0.09) | 0.425 (0.007) | 0.431 (0.004) |

Note: s.e.$_1$ is standard error of estimate of founder allele frequency, s.e.$_2$ is standard error of prediction (equal to root mean-squared error of prediction) of frequency in full sample.

**Table 2**

*Allele-frequency estimates and standard errors for four STRPs in the Hutterites. BLUE of founder allele frequency is compared to naive estimate.*

| D17S928 | | D1S468 | | D14S1426 | | D10S1225 | |
|---|---|---|---|---|---|---|---|
| Naive | BLUE | Naive | BLUE | Naive | BLUE | Naive | BLUE |
| 0.25 (0.09) | 0.21 (0.08) | 0.33 (0.10) | 0.34 (0.09) | 0.42 (0.10) | 0.35 (0.09) | 0.38 (0.10) | 0.42 (0.09) |
| 0.21 (0.08) | 0.22 (0.08) | 0.20 (0.08) | 0.20 (0.08) | 0.29 (0.09) | 0.35 (0.09) | 0.28 (0.09) | 0.24 (0.08) |
| 0.20 (0.08) | 0.27 (0.08) | 0.15 (0.07) | 0.16 (0.07) | 0.12 (0.07) | 0.12 (0.06) | 0.21 (0.08) | 0.21 (0.08) |
| 0.11 (0.07) | 0.10 (0.06) | 0.14 (0.07) | 0.13 (0.06) | 0.09 (0.06) | 0.11 (0.06) | 0.11 (0.07) | 0.10 (0.06) |
| 0.11 (0.06) | 0.12 (0.06) | 0.13 (0.07) | 0.11 (0.06) | 0.08 (0.06) | 0.06 (0.04) | 0.02 (0.03) | 0.02 (0.03) |
| 0.07 (0.05) | 0.04 (0.04) | 0.04 (0.04) | 0.06 (0.04) | 0.003 (0.01) | 0.01 (0.02) | – | – |
| 0.02 (0.03) | 0.01 (0.02) | 0.002 (0.01) | 0.002 (0.01) | – | – | – | – |
| 0.02 (0.03) | 0.01 (0.02) | 0.002 (0.01) | 0.004 (0.01) | – | – | – | – |
| 0.01 (0.02) | 0.01 (0.02) | – | – | – | – | – | – |

McPeek, and Ober, 2000). All the individuals in this highly complex, inbred pedigree are descended from 64 founders. Genotype data, as well as data on a variety of qualitative and quantitative traits, were collected for the purpose of mapping loci predisposing to these traits (Ober, Abney, and McPeek, 2001). For many mapping methods, estimates of allele frequencies for a large number of loci must first be obtained, and this is the problem that we address here.

Table 1 gives results of allele-frequency estimation for five single-nucleotide polymorphisms (SNPs) located in candidate genes for asthma and other inflammatory diseases. Table 2 gives similar results for four short tandem repeat polymorphisms (STRPs) that are used for mapping but that do not correspond to any particular candidate genes. Table 2 compares naive estimates and BLUEs, which estimate the frequency in the founding population. Table 1 also compares naive predictions and BLUPs, which predict the frequency in a set of 858 individuals, based on the genotyped subset. The differences in the estimates are nontrivial in many cases, and the reduction in standard error from the naive estimator to the BLUE is consistent with the theoretical calculations that we describe next.

### 5.2 *Analytical and Simulation-Based Comparison of Estimators*

Note that for any given pedigree or set of pedigrees, we can directly calculate the sampling variances of the naive estimator and BLUE as a function of allele frequency, to compare their efficiencies. This is an analytical result that is not based on simulation or data. We would like to compare the MLE as well. Unlike the naive estimator and the BLUE, the MLE is biased, so we compare its mean-squared error (MSE) to the variances of the other estimators. For the MLE, calculation of MSE is based on simulation. In the Hutterites, it is not currently feasible to calculate the MLE at all, so we compare only the naive estimator and the BLUE in 806 individuals from that population. To compare all three estimators, we consider a family-based sample of individuals drawn from an outbred population. This sample is composed of 85 two-, three-, and four-generation outbred pedigrees consisting of 996 individuals. We assume that data are missing for 190 individuals from the top two generations. Analytical variance calculations for the naive estimator and BLUE are made based on the pedigrees. To approximate the MSE for the MLE in the outbred pedigrees, we performed 5000 simulations at each of five allele frequencies (given in Table 3), in which founder

**Table 3**

*Ratios of MSE in the outbred pedigrees* (*obtained analytically for naive estimator and BLUE and by simulation for MLE*)

| Allele frequency | Naive estimator vs. MLE | BLUE vs. MLE |
|---|---|---|
| 0.50 | 1.41 | 1.01 |
| 0.30 | 1.43 | 1.02 |
| 0.20 | 1.44 | 1.03 |
| 0.15 | 1.45 | 1.04 |
| 0.10 | 1.45 | 1.04 |

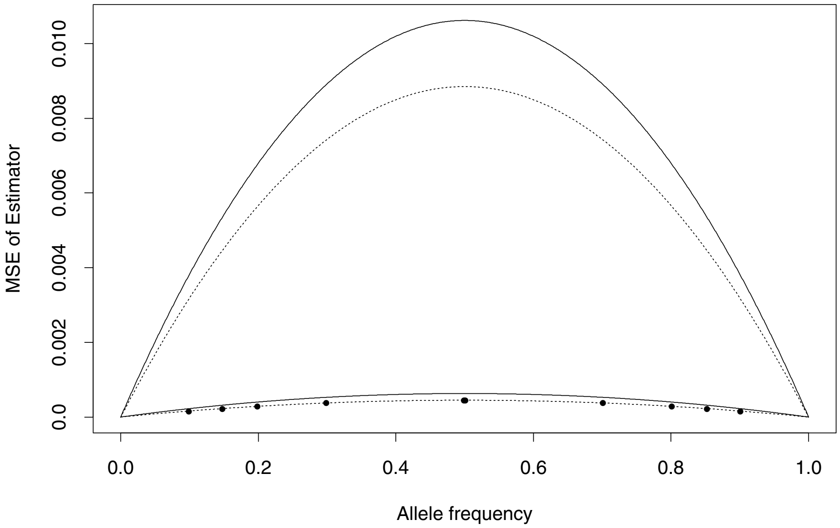

**Figure 1.** MSE versus allele frequency in Hutterite and outbred samples. Upper solid line represents variance of naive estimator in Hutterite sample, upper dotted line represents variance of BLUE in Hutterite sample, lower solid line represents variance of naive estimator in outbred sample, lower dotted line represents variance of BLUE in outbred sample, where all variances are obtained analytically, and points represent MSE of MLE obtained by simulation in outbred sample.

alleles were chosen at random with replacement, and alleles were dropped down the pedigree. The USERM13 module of MENDEL (Lange, Weeks, and Boehnke, 1988; Boehnke, 1991), was used to calculate the MLE in each simulated realization. In the simulations, the bias in the MLE is too small to be detected above the sampling variability based on 5000 realizations.

Figure 1 shows the MSE for each estimator in each sample, while Figure 2 magnifies the results for the outbred sample. A striking feature of Figure 1 is how much larger the variances of the naive estimator and BLUE are in the Hutterite sample than in the outbred sample, although the apparent sample sizes are the same (both 806). Recall that the Hutterite sample descends from only 64 founders, so even with all 1623 members typed, the effective sample size would be only 128 alleles. Furthermore, in the 13-generation Hutterite pedigree, the genotyped individuals lie in the last few generations, and much of the variance in the estimator is caused by genetic drift, i.e., random variation in the numbers of alleles of each type transmitted to succeeding generations, which, in effect, leads to noisier data. The outbred pedigrees are much shallower (i.e., smaller numbers of generations), so the effects of genetic drift are much less, leading to the ability to estimate more precisely the allele frequency in the founding population.

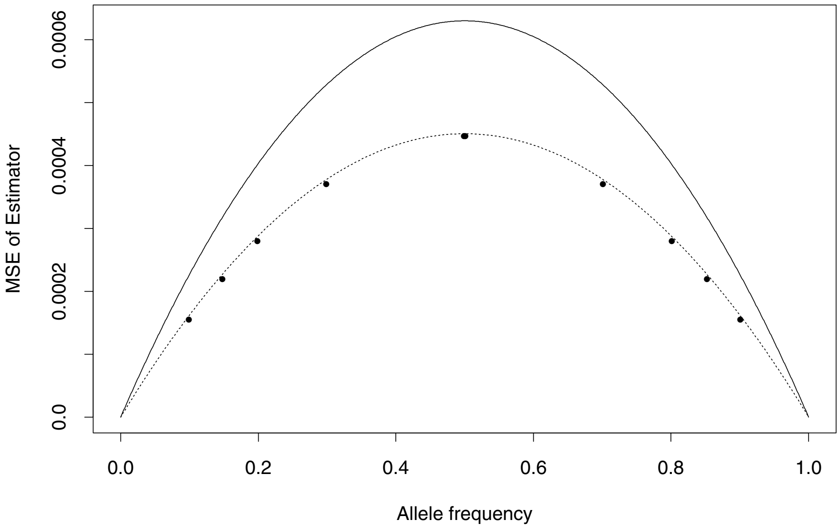

**Figure 2.** MSE versus allele frequency in an outbred sample, a rescaling of the bottom portion of Figure 1. Solid line represents variance of naive estimator, dotted line represents variance of BLUE, where both variances are obtained analytically, and points represent MSE of MLE obtained by simulation.

In the Hutterite pedigree, the efficiency of the BLUE relative to the naive estimator is 1.20, while in the outbred pedigree it is 1.40. The smaller ratio for the Hutterites is likely due to the fact that, as mentioned above, the data are inherently noisier, so improving the estimator does not reduce the variance by as high a percentage as in the outbred case, although in absolute value the reduction in variance is much greater for the Hutterites than for the outbred population. Table 3 shows that in the outbred population, based on the simulations, the ratio of the MSE of the naive estimator to the MSE of the MLE is between 1.41 and 1.45, while the ratio of the MSE of the BLUE to the MSE of the MLE is between 1.01 and 1.04. Thus, the additional improvement of the MLE over the BLUE is small compared to the improvement of the BLUE over the naive estimator. This simulation suggests that much of the advantage of the MLE can be obtained by the use of the computationally much simpler BLUE.

## 6. Discussion

We consider the problem of allele-frequency estimation from data on related individuals, with particular attention to the case when the relationships among the individuals are specified by a large, complex pedigree. For the problem of estimating the allele frequency in the founding population, we derive the BLUE and show how this may be efficiently computed in a large, complex pedigree. Both the case in which parental origin of allele is known and the case in which it is unknown are treated. We show that when all founders of the pedigree(s) are genotyped, the BLUE coincides with the MLE, which is just the frequency in the founders, and we derive a related result that applies, for example, when all individuals in a subpedigree of the original pedigree are genotyped. We extend our results to address the problem of prediction of allele frequency in some subset $S$ of individuals in the pedigree, based on genotype data from some subset $R$ of individuals in the pedigree, for which we derive the BLUP.

We compare the performances of the BLUE and the naive estimator (sample frequency) in a Hutterite pedigree, using both data analysis and analytical results, and we use both simulations and analytical results to compare the MLE, BLUE, and naive estimator in an outbred sample. Our results suggest that the BLUE provides substantial improvement over the naive estimator and performs very similarly to the MLE. In large complex pedigrees such as the Hutterites, the MLE may be infeasible to compute, especially as one may need to repeat this computation for hundreds or thousands of loci, e.g., for a genome screen. In contrast, the BLUE may be very efficiently computed even in large complex pedigrees.

The methods presented are for codominant, autosomal loci, but can be extended to X-linked loci by suitable modification of the indicator vector $X$ and recalculation of the covariance matrix, which would then be singular in general. This can be remedied by the use of an appropriate generalized inverse or by removing redundant entries of $X$. The extension to Y-linked or mitochondrial loci is trivial because the pedigree specifies the IBD relationships.

We have approached the problem of allele-frequency estimation by taking into account only the first and second moments of the allele-frequency indicator vector. This was done because the full-likelihood calculation quickly becomes infeasible in large, complex pedigrees. More broadly, one can think of taking a similar approach to solve other related inference problems in large, complex pedigrees. For instance, the problem of detecting association between a binary trait and a locus and the problem of detecting deviation from Hardy–Weinberg can both be formulated as hypothesis tests involving allele frequencies. As we noted above, the BLUE for the problem of allele-frequency estimation is the same as the quasi-likelihood estimator. As a result, the framework of quasi-likelihood could be used to extend our results on allele-frequency estimation to create quasi-likelihood score tests for the problems of case-control association testing (Bourgain et al., 2003) and Hardy–Weinberg testing (work in preparation).

Related software is incorporated into the CC-QLS package available at `http://galton.uchicago.edu/~mcpeek/software`.

## Acknowledgements

Support from NIH grants HG001645, DK55889, HL56399, and HL66533 and from an NSF GIG grant to the University of Chicago Department of Statistics is gratefully acknowledged.

## Résumé

De nombreux types d'études génétiques reposent sur l'estimation des fréquences alléliques. Nous considérons le problème de l'estimation des fréquences alléliques à partir de données sur des individus apparentés. La motivation de ce travail est l'analyse de données dans la population Huttérites, une population fondatrice isolée, et en conséquence nous nous intéresserons plus particulièrement au cas où les relations entre les différents individus de l'échantillon sont spécifiées par un arbre généalogique étendu et complexe pour lequel l'estimation par maximisation de la vraisemblance n'est pas réalisable. Dans ce cas, nous proposons d'utiliser le meilleur estimateur linéaire non biaisé (BLUE) de la fréquence allélique. Nous dérivons cet estimateur, qui dans ce problème est équivalent à l'estimateur de quasi-vraisemblance, et nous décrivons un algorithme efficace pour calculer cet estimateur et sa variance. Nous montrons que cet estimateur à certaines propriétés désirables en commun avec l'estimateur du maximum de vraisemblance (MLE) pour ce type ce problème. Nous traitons les deux cas où l'origine parentale de chaque allèle est connue ou non. Les résultats sont étendus à la prédiction de fréquence allélique dans certains ensembles d'individus S à partir de données collectées sur un ensemble d'individus R. Nous comparons l'erreur quadratique moyenne du BLUE, de l'estimateur naïf communément utilisé (fréquence de l'échantillon) et du MLE lorsqu'il est possible de le calculer. Les résultats indiquent que bien que le MLE présente les meilleures performances, le BLUE en est très proche tout en en étant substantiellement plus facile à calculer, ce qui le rend particulièrement utile pour l'analyse des grandes généalogies complexes où le MLE est incalculable. Finalement, nous appliquons notre méthode pour estimer les fréquences alléliques chez les Huttérites.

## References

Abney, M., McPeek, M. S., and Ober, C. (2000). Estimation of variance components of quantitative traits in inbred populations. *American Journal of Human Genetics* **66,** 629–650.

Abney, M., Ober, C., and McPeek, M. S. (2002). Quantitative trait homozygosity and association mapping and empirical genome-wide significance in large complex pedigrees:

Fasting serum insulin level in the Hutterites. *American Journal of Human Genetics* **70,** 920–934.

Boehnke, M. (1991). Allele frequency estimation from data on relatives. *American Journal of Human Genetics* **48,** 22–25.

Bourgain, C., Hoffjan, S., Nicolae, R., Newman, D., Steiner, L., Walker, K., Reynolds, R., Ober, C., and McPeek, M. S. (2003). Novel case-control test in a founder population identifies *P-selectin* as an atopy susceptibility locus. *American Journal of Human Genetics* **73,** 612–626.

Boyce, A. J. (1983). Computation of inbreeding and kinship coefficients on extended pedigrees. *Journal of Heredity* **74,** 400–404.

Broman, K. W. (2001). Estimation of allele frequencies with data on sibships. *Genetic Epidemiology* **20,** 307–315.

Ceppellini, R., Siniscalco, M., and Smith, C. A. B. (1955). The estimation of gene frequencies in a random mating population. *Annals of Human Genetics, London* **20,** 97–115.

Chakraborty, R. (1978). Number of independent genes examined in family surveys and its effect on gene frequency estimation. *American Journal of Human Genetics* **30,** 550–552.

Cotterman, C. W. (1947). A weighting system for the estimation of gene frequencies from family records. *Contributions from the Laboratory of Vertebrate Biology* **33,** 1–21.

Finney, D. J. (1948a). The estimation of gene frequencies from family records. I. Factors without dominance. *Heredity* **2,** 199–218.

Finney, D. J. (1948b). The estimation of gene frequencies from family records. II. Factors with dominance. *Heredity* **2,** 369–390.

Fisher, R. A. (1940). The estimation of the proportion of recessives from tests carried out on a sample not wholly unrelated. *Annals of Eugenics* **10,** 160–170.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* North Scituate, Massachusetts: Duxbury Press.

Lange, K., Weeks, D., and Boehnke, M. (1988). Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genetic Epidemiology* **5,** 471–472.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd edition. New York: Springer.

Lockwood, J. R., Roeder, K., and Devlin, B. (2001). A Bayesian hierarchical model for allele frequencies. *Genetic Epidemiology* **20,** 17–33.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. New York: Chapman and Hall.

Ober, C., Abney, M., and McPeek, M. S. (2001). The genetic dissection of complex traits in a founder population. *American Journal of Human Genetics* **69,** 1068–1079.

Olson, J. M. (1994). Robust estimation of gene frequency and association parameters. *Biometrics* **50,** 665–674.

Ott, J. (1992). Strategies for characterizing highly polymorphic markers in human gene mapping. *American Journal of Human Genetics* **51,** 283–290.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61,** 439–447.

*Received July* 2003. *Revised December* 2003.
*Accepted January* 2004.

## Appendix

*Proof of Proposition* 1. We show that the BLUE gives the same weight to both alleles of an individual when information on parental origin of allele is unavailable. The proposition then follows from the standard result on calculation of the BLUE. Fix an individual in the sample. Let $P$ be the $2n \times 2n$ permutation matrix that interchanges the rows corresponding to the two alleles of the given individual. Then when information on parental origin of allele is unavailable, we have $PKP^T = K$. For any permutation matrix $P$, we also have $(1^T(PKP^T)^{-1}1)^{-1}1^T(PKP^T)^{-1}P = (1^TK^{-1}1)^{-1}1^TK^{-1}$. Combining these gives $(1^TK^{-1}1)^{-1}1^TK^{-1}P = (1^TK^{-1}1)^{-1}1^TK^{-1}$, which shows that the BLUE gives the same weight to both alleles of the individual. ■

*Proof of Proposition* 2. Let $T_1 = (D_1^T\Sigma_{11}^{-1}D_1)^{-1}D_1^T\Sigma_{11}^{-1}X_1$. Any LUE $AX$ of $a$ can be written as $AX = T_1 + T_2$, where $T_2 = WX$ and $WD = 0$. $\mathrm{Cov}(T_1,\, T_2) = (D_1^T\Sigma_{11}^{-1}D_1)^{-1}(D_1^T,\, D_1^T\Sigma_{11}^{-1}\Sigma_{12})W^T$, which $= 0$ if equation (6) is satisfied. In that case, $\mathrm{Var}(AX) \geq \mathrm{Var}(T_1)$. ■

*Proof of Corollary to Proposition* 2. (i) First suppose $m = 2$. Suppose individual $i$ has no genotyped descendants and has both parents genotyped. Consider the case in which maternal and paternal alleles can be distinguished and are listed in that order for each individual. Reorder the vector $X$, so that $X_2$ represents individual $i$'s data, and $X_1$ represents all other data. In $X_1$, let $\alpha$ and $\alpha + 1$ index the two alleles of $i$'s mother and $\beta$ and $\beta + 1$ index the two alleles of $i$'s father. Then, from standard recursive formulae for kinship coefficients, we have that the first column of $\Sigma_{12}$ is the average of the $\alpha$th and $\alpha + 1$st columns of $\Sigma_{11}$ and the second column of $\Sigma_{12}$ is the average of the $\beta$th and $\beta + 1$st columns of $\Sigma_{11}$. Thus, $\Sigma_{12}^T\Sigma_{11}^{-1}$ has entries $(1,\, \alpha)$, $(1,\, \alpha + 1)$, $(2,\, \beta)$, and $(2,\, \beta + 1)$ equal to $1/2$ and all others 0. Equation (6) follows. This argument is applied recursively to prove the Corollary. The case $m > 2$ follows from equation (4). The proof for the case when parental origin of allele cannot be distinguished is similar. For (ii), note $\Sigma_{11} = I$ and use the fact that every allele is IBD with exactly one founder allele. That these properties hold for the MLE follows in each case by decomposing the likelihood into two factors $L_1$ and $L_2$, where $L_1$ is the likelihood for $X_1$, $L_2$ is the conditional likelihood for $X_2$ given $X_1$, and $L_2$ does not depend on the allele-frequency parameter. For part (i), apply this argument recursively. ■

*Proof of Proposition* 3. Any $A \in \Gamma$ can be written as $A = A^* + A_0$ with $A_0GD = 0$. First consider the case $\Sigma = I$. Let $H = G^T(GG^T)^{-1}G$. Note that $\mathrm{Cov}(A^*W - BX, A_0W) = 0$ can be obtained by using the decomposition $A^*W - BX = B(H - I)X + (A^*W - BHX)$ to break the covariance into two terms, plugging in the expression for $A^*$ and using $A_0GD = 0$. Thus, $E[\{c^T(AW - BX)\}^2] = E\{(c^TA^*W - c^TBX)^2\} + E\{(c^TA_0W)^2\}$, which is minimized when $E\{(c^TA_0W)^2\} = 0$, which holds for all $c$ iff $A_0G = 0$. This proves the result for the case $\Sigma = I$. More generally, suppose $\Sigma = J^TJ$ is invertible. Let $\tilde{X} = J^{-T}X, \tilde{D} = J^{-T}D, \tilde{G} = GJ^T$, and $\tilde{B} = BJ^T$. Then $\tilde{W} = \tilde{G}\tilde{X} = W, \tilde{B}\tilde{X} = BX, E(\tilde{X}) = \tilde{D}a$, and $\mathrm{Var}(\tilde{X}) = I$. The proposition follows by application of the previous case. ■