Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00636.x

A Flexible Model for Association Analysis in Sibships with Missing Genotype Data

Frank Dudbridge[1,2,3]*, Peter A. Holmans[4] and Scott G. Wilson[5,6,7]

[1] *Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK*
[2] *Bloomsbury Centre for Genetic Epidemiology and Statistics, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK*
[3] *MRC Biostatistics Unit, Cambridge CB2 0SR, UK*
[4] *Biostatistics and Bioinformatics Unit, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University School of Medicine, Cardiff CF14 4XN, UK*
[5] *Department of Endocrinology & Diabetes, Sir Charles Gairdner Hospital, Nedlands, 6009, Australia*
[6] *School of Medicine & Pharmacology, University of Western Australia, Crawley, 6009, Australia*
[7] *Department of Twin Research & Genetic Epidemiology Research Department, King's College London, London SE1 7EH, UK*

Summary

A common design in family-based association studies consists of siblings without parents. Several methods have been proposed for analysis of sibship data, but they mostly do not allow for missing data, such as haplotype phase or untyped markers. On the other hand, general methods for nuclear families with missing data are computationally intensive when applied to sibships, since every family has missing parents that could have many possible genotypes. We propose a computationally efficient model for sibships by conditioning on the sets of alleles transmitted into the sibship by each parent. This means that the likelihood can be written only in terms of transmitted alleles and we do not have to sum over all possible untransmitted alleles when they cannot be deduced from the siblings. The model naturally accommodates missing data and admits standard theory of estimation, testing, and inclusion of covariates. Our model is quite robust to population stratification and can test for association in the presence of linkage. We show that our model has similar power to FBAT for single marker analysis and improved power for haplotype analysis. Compared to summing over all possible untransmitted alleles, we achieve similar power with considerable reductions in computation time.

Keywords: Family-based association, TDT, sib-TDT, FBAT, haplotype

Introduction

Family-based designs play an important role in genetic association analysis, primarily by allowing the comparison of subjects that are matched for shared risk factors, such as population membership, that might confound analyses (Whittaker & Morris, 2001; Tiwari et al., 2008). There are many family structures that are informative for association, but the most popular are based on the nuclear family consisting of two parents and their full offspring. In this case, the analysis method may depend on whether one, both, or no parents are genotyped, and the concordance in trait values among the siblings is a critical determinant of power (Abecasis et al., 2001).

Here we are concerned with samples of siblings without parents, which continue to be an important design for research into complex diseases. For example, in studies of late-onset disease, the parents of incident cases may be dispersed, non-consenting or deceased. Twin studies are an important source of sibling data, as large numbers are available in the general population and twin registers are maintained in many countries, making them an ideal resource for genetics research. By their registration in the twin study, these subjects are usually prequalified for participation in research, which can facilitate recruitment compared to a population-based study. Also, because twins are matched for age and for their developmental environment, many common important environmental

*Corresponding author: Dr Frank Dudbridge, Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK. Tel: +44 020 7927 2025; Fax: +44 20 7436 4230; E-mail: frank.dudbridge@lshtm.ac.uk

influences are controlled (MacGregor et al., 2000). In contrast, this information is not always available to a population study, or must be adjusted for statistically, which can increase the environmental variance in population studies compared to studies of siblings.

Several approaches have been proposed for association analysis of sibships. For binary traits an important early method was the sib-TDT (Spielman & Ewens, 1998), which compares allele frequencies between affected and unaffected sibs, stratifying on family. Conceptually similar approaches are the SDT (Horvath & Laird, 1998) and conditional logistic regression (Siegmund et al., 2000) treating each sibship as a matched case/control set. Another popular approach is to partition the total association into between- and within-family components, allowing joint modelling of linkage and association and combinations of sibships with complete nuclear families and unrelated subjects (Abecasis et al., 2000; Li et al., 2006). A further approach is to include all siblings in a mixed regression model including family-specific random effects (Allison et al., 1999; Jonasdottir et al., 2007). Recently, approximate but rapid methods have been developed for quantitative traits (Aulchenko et al., 2007; Chen & Abecasis, 2007). Finally a general approach for testing association in pedigrees can be applied to sibships (Lake et al., 2000; Rabinowitz & Laird, 2000) and is implemented in software called FBAT (Horvath et al., 2001).

These previous approaches generally do not allow for missing genotype data. If subjects with missing data are omitted from analysis, a loss of power can result particularly if the whole sibship is then rendered uninformative. A well-known form of missing data arises in the analysis of multilocus haplotypes, which cannot always be resolved from genotype data even in family samples with the parents present (Dudbridge et al., 2000). Another application of much recent interest is analysis of imputed markers that are not directly genotyped (Lin et al., 2008). A problem in family-based analysis is to account for missing data in such a way that the analysis remains robust to confounding while extracting the maximal information from the data. Established approaches include conditioning on sufficient statistics for the missing data (Horvath et al., 2004), maximum likelihood methods (Dudbridge, 2008), and conditional analyses of multiply imputed data (Croiseau et al., 2007). In the context of sibship analysis, two studies have used expected haplotype counts, estimated under the null hypothesis, in a form of weighted analysis (Jonasdottir et al., 2008; Stone et al., 2010). This type of approach has a history of application in genetic epidemiology (Cordell, 2006), although in principle its results will be biased towards the null hypothesis, and neither study considered the effect of population stratification on their models.

A general approach to missing data in nuclear families has been proposed (Dudbridge, 2008), which considers all possible completions of missing data with reasonable robustness to confounding by population stratification. Some advantages of this approach are that it provides proper estimates of effect sizes, incorporates covariates in a natural way, and can include data from additional unrelated subjects. A sibship can be treated as a nuclear family with two missing parents, and all parental genotypes are considered that are compatible with the sibling genotypes. This approach is implemented in software called UNPHASED (Dudbridge, 2006), and we will refer to it here as UNPHASED/default. Although this approach is quite feasible, it can be computationally intensive since there are missing data in each sibship, and the number of possible parental genotypes can be very large. Here we propose some modifications to this model that can significantly reduce the computation in sibships. The key aspect is to express the likelihood only in terms of haplotypes transmitted into the sibship: if a parental haplotype is not transmitted to any sibling we do not consider all of its possible values, as would be done in the previous approach. This desirable property is shared by other sibship methods, but our framework allows for missing data in the sibs themselves to be accounted for, as well as retaining an approach to testing association in the presence of linkage. Our modifications are implemented as a new option in UNPHASED, and will be referred to here as UNPHASED/sibship.

We compare UNPHASED/sibship to UNPHASED/default and also to the FBAT program (Lake et al., 2000) that conditions on the minimal sufficient statistic for the missing parents. We illustrate the methods on affection status data for a late-onset disease and quantitative trait data in a sample of sibships. We show that UNPHASED/sibship has power consistently as good as or better than FBAT, whereas UNPHASED/default has less consistent performance. We give an example of data in which FBAT had more significant results than either version of UNPHASED, but show using simulations that this result does not reflect a difference in power, but is likely to be a chance finding.

Methods

Statistical Model

First, we review the model of Dudbridge (2008). For simplicity we consider a binary trait with no covariates, but modifications for a quantitative trait will be indicated later. Consider a nuclear family having k children, with paternal genotype F, maternal genotype M, phased child genotypes $C = (C_1, \dots, C_k)$, and child trait vector $Y \in \{0, 1\}^k$. Genotypes may encompass multiple loci, and by "phased" we mean that the parent of origin is known for each allele at each genotyped locus. We use a retrospective likelihood given by the probability of genotypes, conditional on the child traits, as

$$\Pr(F = f, M = m, C = c \,|\, Y = y) = \Pr(c \,|\, f, m, y) \Pr(f, m \,|\, y) = \frac{\Pr(c \,|\, f, m) \Pr(y|f, m, c)}{\Pr(y|f, m)} \frac{\Pr(f, m) \Pr(y|f, m)}{\Pr(y)}. \tag{1}$$

Assume that child traits are conditionally independent of parental genotypes, given the child genotypes, so that $\Pr(y|f, m, c) = \Pr(y|c)$. Further assume that there is no transmission distortion so $\Pr(c \,|\, f, m)$ is constant for each mating type. Then we have the *conditional likelihood component*

$$\Pr(c \,|\, f, m, y) = \frac{\Pr(y|c)}{\sum_{c^* \in S(f,m)} \Pr(y|c^*)}, \tag{2}$$

and *marginal likelihood component*

$$\Pr(f, m \,|\, y) = \frac{\Pr(f, m) \sum_{c^* \in S(f,m)} \Pr(y|c^*) \Pr(c^*|f, m)}{\sum_{f^*,m^*} \Pr(f^*, m^*) \sum_{c^* \in S(f^*,m^*)} \Pr(y|c^*) \Pr(c^*|f^*, m^*)}, \tag{3}$$

where $S(f, m) = \{c : \Pr(c \,|\, f, m) > 0\}$ is the set of phased child genotypes consistent with parental genotypes f, m. We are primarily interested in the conditional likelihood (2), which is the probability of child genotypes given the parental genotypes. Inference based on the conditional likelihood is robust to mis-specification of the mating-type distribution $\Pr(f, m)$, which may occur through population stratification. However we will need the mating-type distribution in the marginal likelihood (3) to give a model for missing parental genotypes.

For $\Pr(y|c)$ in the conditional likelihood (2) we assume a log-linear model of the form

$$\Pr(Y_i = 1|c_i) = \exp(\mu + X'_{c_i} \beta), \tag{4}$$

where μ, β are fixed effect parameters and X_{c_i} is a design vector that codes for the genotype of child i. We make the simplifying assumption that the disease is rare, so that $\mu \approx -\infty$ and

$$\Pr(Y_i = 0|c_i) = 1 - \exp(\mu + X'_{c_i} \beta) \approx 1. \tag{5}$$

The above is also applicable if families are only ascertained according to affected sibs. With the additional assumption that sibling risks are independent given their genotypes, we can then use as the conditional likelihood contribution

$$\Pr(c \,|\, f, m, y) = \frac{\exp(y' X'_c \beta)}{\sum_{c^* \in S(f,m)} \exp(y' X'_{c^*} \beta)}, \tag{6}$$

in which β are fixed effects for the genotype log-relative risks, μ cancels and is no longer identifiable, and X_c is a design matrix that codes for genotypes with one column for each child. Various parameterisations and coding schemes are possible for the design matrices (Cordell & Clayton, 2002), the two most common being a saturated model having a parameter for each possible genotype, and an additive model having a parameter for each allele or haplotype.

In the marginal likelihood (3) we use a multinomial logistic model for $\Pr(f, m)$ since the mating type is a categorical variable. Importantly, we now use distinct parameters in $\Pr(y|c^*)$ from those used in the conditional likelihood. This means that when there are no missing data, inference on the parameters in (2) is independent of the model in (3), and thus of any mis-specification of the mating-type distribution. In terms of fixed effect parameters, the marginal likelihood contribution becomes

$$\Pr(f, m \,|\, y) = \frac{2^{h_{f,m}} \sum_{c^* \in S(f,m)} \exp(X'_{f,m}\alpha + y' X'_{c^*} \tilde{\beta}) \Pr(c^*|f, m)}{\sum_{f^*,m^*} 2^{h_{f^*,m^*}} \sum_{c^* \in S(f,m)} \exp(X'_{f^*,m^*}\alpha + y' X'_{c^*} \tilde{\beta}) \Pr(c^*|f^*, m^*)}, \tag{7}$$

where α is a vector of parameters for the mating-type frequency, $\tilde{\beta}$ are parameters for the genotype log-relative risks, and $X_{f,m}$ is a design vector that codes for the mating type. The term $h_{f,m}$ is the number of heterozygous parents: its inclusion allows a common interpretation of α for all mating types, and also convenient parameterisation of α in terms of genotype or haplotype frequencies (with appropriate assumptions of random mating and Hardy-Weinberg equilibrium). We have $\Pr(c^*|f, m) = 2^{-kh_{f,m}}$. The full likelihood contribution is then the product of (6) and (7), and when there are missing data the likelihood contribution is the sum over the set of consistent completions

$$\Pr(g_f, g_m, g_c \,|\, y) = \sum_{f,m,c} \Pr(g_f, g_m, g_c \,|\, y, f, m, c) \Pr(f, m, c \,|\, y) = \sum_{f,m,c \in \Gamma} \Pr(f, m, c \,|\, y), \tag{8}$$

where g_f, g_m, g_c denote the observed genotype data for father, mother, and children respectively, possibly including missing data, and $\Gamma = \{f, m, c : \Pr(F = f, M = m, C = c \,|\, g_f, g_m, g_c) > 0\}$.

For a sample of N families indexed by i, the total log-likelihood is then

$$\ell(\alpha, \beta, \tilde{\beta}) = \sum_{i=1}^{N} \log \sum_{f,m,c \in \Gamma_i} \Pr(f, m, c \,|\, y_i), \tag{9}$$

with the probability term given by the product of (6) and (7). This model is closely related to the method of Clayton (Clayton, 1999) and has been shown to be acceptably robust to mis-specification of $\Pr(f, m \,|\, y)$ when there are missing data (Dudbridge, 2008).

Modifications for Sibships

In the analysis of sibships, the parental genotypes are missing but can be partially or completely reconstructed from the sibs. The above model sums over all completions of a missing parent consistent with the sibs, but this can be computationally demanding. For example, consider a discordant sib pair, both of

Table 1 Possible completions of the missing parents, given two heterozygous sibs. Sib genotypes are given with the paternally transmitted allele first, but parental genotypes are unphased. The rightmost column is $2^{h_{f,m}}\,\Pr(c\,|\,f, m)$ in Equation (7).

Father	Mother	Sib 1	Sib 2	$2^{h_{f,m}(1-k)}$
1 2	1 2	1 2	1 2	1/4
1 2	1 2	1 2	2 1	1/4
1 2	1 2	2 1	1 2	1/4
1 2	1 2	2 1	2 1	1/4
1 2	1 1	2 1	2 1	1/2
1 1	1 2	1 2	1 2	1/2
1 2	2 2	1 2	1 2	1/2
2 2	1 2	2 1	2 1	1/2
1 1	2 2	1 2	1 2	1
2 2	1 1	2 1	2 1	1

which are heterozygous for a diallelic marker. Enumerating the possible completions and phased transmissions from the missing parents, there are 10 reconstructions consistent with the observed genotypes (Table 1). Thus, the computation of the likelihood contribution will take ten times as long as for a family with the transmissions known.

To reduce the computation in this situation, we now propose some modifications to the model that allow for incomplete parental reconstruction. Our aim is to express the likelihood only in terms of transmitted haplotypes, so that it is constant over all possible values of untransmitted haplotypes. The modifications we propose are first, to condition on the sets of haplotypes transmitted by each parent; second, to condition on the equivalence class of the inheritance vector; and finally, to use expected sib genotype frequencies in place of parental mating-type frequencies. We now examine each of these modifications in turn.

We first change the conditioning event in the conditional likelihood contribution (6). For given f, m, c we now consider the conditional probability of the child genotypes, given the parental genotypes and the sets of haplotypes transmitted into the sibship by each parent (for brevity we use the term "haplotype" interchangeably with "allele" when one marker is analysed). Of course if there is only one sib, this probability is one, but when there is more than one sib there are up to four haplotypes transmitted (assuming no recombination) in which case several values of c are possible given f, m. Define ϕ_c as the set of paternal haplotypes observed among phased child genotypes c, and ψ_c as the set of maternal haplotypes observed in c. Therefore, each child genotype c_i has its paternal haplotype from ϕ_c and its maternal haplotype from ψ_c. Then we now write the full likelihood contribution as

$$\begin{aligned}
&\Pr(f, m, c\,|\,y)\\
&\quad = \Pr(F = f, M = m, C = c, \phi_C = \phi_c, \psi_C = \psi_c\,|\,y)\\
&\quad = \Pr(c\,|\,f, m, \phi_c, \psi_c, y) \cdot \Pr(f, m, \phi_c, \psi_c\,|\,y). \qquad (10)
\end{aligned}$$

This changes the range of the sum in the conditional likelihood so that

$$\Pr(c\,|\,f, m, \phi_c, \psi_c, y) = \frac{\exp(y\prime X_c'\beta)}{\sum\limits_{\substack{c^*\in S(f,m);\\ \phi_{c^*}=\phi_c;\psi_{c^*}=\psi_c}} \exp(y' X_{c^*}'\beta)}, \qquad (11)$$

and the numerator of the marginal likelihood also changes so that

$$\begin{aligned}
&\Pr(f, m, \phi_c, \psi_c\,|\,y)\\
&\quad = \frac{\sum\limits_{\substack{c^*\in S(f,m);\\ \phi_{c^*}=\phi_c;\psi_{c^*}=\psi_c}} \exp(X_{f,m}'\alpha + y' X_{c^*}'\tilde{\beta})2^{h_{f,m}(1-k)}}{\sum\limits_{f^*,m^*}\sum\limits_{c^*\in S(f^*,m^*)} \exp(X_{f^*,m^*}'\alpha + y' X_{c^*}'\tilde{\beta})2^{h_{f^*,m^*}(1-k)}}. \qquad (12)
\end{aligned}$$

The second modification is necessary because conditioning on the sets of transmitted haplotypes creates a dependency among the transmissions to sibs. Given the haplotypes transmitted to one sib, the haplotypes transmitted to others must be such that the same final sets of transmitted haplotypes are obtained. This dependency is reminiscent of, but not the same as, the dependency that arises when testing association in the presence of linkage (Lake et al., 2000), and it results in violation of the assumption of constant transmission probability $\Pr(c\,|\,f, m)$ made in (2). Dudbridge (2008) proposed a conditioning step in the presence of linkage that can also be applied here. The phase of transmissions to the sibs is represented by an inheritance vector specifying the grand-parental origin of each haplotype transmitted to each sib. Because a nuclear family does not include grandparents, there are four inheritance vectors that are consistent with the observed transmissions, obtained by arbitrarily designating, in each parent, one haplotype as grand-paternal and the other as grand-maternal. These four inheritance vectors form an equivalence class (Kruglyak et al., 1996) that we denote $V_{f,m,c}$ for given parental and phased child genotypes. We now work with

$$\begin{aligned}
&\Pr(f, m, c\,|\,y)\\
&\quad = \Pr(f, m, c, \phi_c, \psi_c, V_{F,M,C} = V_{f,m,c}\,|\,y)\\
&\quad = \Pr(c\,|\,f, m, \phi_c, \psi_c, V_{f,m,c}, y) \cdot \Pr(f, m, \phi_c, \psi_c, V_{f,m,c}\,|\,y). \qquad (13)
\end{aligned}$$

To condition on the inheritance vector class $V_{f,m,c}$ we replace $S(f, m)$ above by the set of four genotype vectors formed by permuting the transmitted and untransmitted haplotypes in all siblings simultaneously (Dudbridge, 2008). If both haplotypes of a parent are transmitted into the sibship, then simultaneously exchanging the transmitted and untransmitted haplotypes to all sibs will ensure that both haplotypes remain transmitted into the sibship. But if only one haplotype of a parent is transmitted into the sibship, then conditioning on the transmitted haplotypes permits only that one contribution from the parent. It follows that this step maintains a constant $\Pr(c\,|\,f, m, \phi_c, \psi_c, V_{f,m,c})$ for each mating type and therefore a correct model. As an aside, note that conditioning on the inheritance vector class is not the same as permuting the sib genotypes, as the latter does not condition on the inheritance vector and could give rise to more than four genotype vectors.

These modifications ensure that the conditional likelihood involves only the haplotypes transmitted to the sibs. A parental haplotype is not deduced if it is never transmitted to a sib, but it does not then appear in the conditional likelihood, which is constant for all values of the unknown haplotype. Our final modification ensures that the marginal likelihood also involves only transmitted haplotypes. If necessary we reparameterise α in terms of genotypes or haplotypes, and we then replace $X_{f,m}$ in (12) with $\frac{2}{k} X_c \cdot \mathbf{1}$, which estimates the parental mating-type frequency by the square of the geometric mean sib genotype frequency. A similar approach has been previously proposed for this situation (Abecasis et al., 2000) and it gives unbiased estimation of α when there is random mating and no transmission distortion.

To define our final likelihood contribution for a sibship with two missing parents, we consider the set of possible phased child genotypes

$$\Gamma' = \{c : \Pr(C = c \,|\, g_c) > 0\}, \tag{14}$$

and use as the log-likelihood

$$\ell(\alpha, \beta, \tilde{\beta}) = \sum_{i=1}^{N} \log \sum_{c \in \Gamma'_i} \sum_{f,m \in Q(c)} \Pr(f, m, c \,|\, y_i), \tag{15}$$

where $Q(c) = \{f, m : \Pr(f, m|c) > 0\}$ is the set of possible parents of children c. Then

$$\begin{aligned}
\ell(\alpha, \beta, \tilde{\beta}) &= \sum_{i=1}^{N} \log \sum_{c \in \Gamma'_i} \sum_{f,m \in Q(c)} \Pr(f, m, c, \phi_c, \psi_c, V_{f,m,c} \,|\, y_i) \\
&= \sum_{i=1}^{N} \log \sum_{c \in \Gamma'_i} \sum_{f,m \in Q(c)} \Pr(c \,|\, f, m, \phi_c, \psi_c, V_{f,m,c}, y_i) \\
&\quad \times \Pr(f, m, \phi_c, \psi_c, V_{f,m,c} \,|\, y_i) \\
&= \sum_{i=1}^{N} \log \sum_{c \in \Gamma'_i} \frac{\exp(y' X'_c \beta)}{\sum_{c^* \in V_c} \exp(y' X'_{c^*} \beta)} \\
&\quad \times \frac{\left[\sum_{c^* \in V_c} \exp\left(\frac{2}{k} \mathbf{1}' X'_c \alpha + y' X'_{c^*} \tilde{\beta}\right)\right] \left[\sum_{f,m \in Q(c)} 2^{h_{f,m}(1-k)}\right]}{\sum_{f^*,m^*} \sum_{c^* \in S(f^*,m^*)} \exp\left(\frac{2}{k} \mathbf{1}' X'_{c^*} \alpha + y' X'_{c^*} \tilde{\beta}\right) 2^{h_{f^*,m^*}(1-k)})},
\end{aligned} \tag{16}$$

where

$$V_c = \{c^* \in S(f, m) : f, m \in Q(c); \phi_{c^*} = \phi_c; \psi_{c^*} = \psi_c; V_{f,m,c^*} = V_{f,m,c}\}. \tag{17}$$

With this form, the likelihood contribution can be calculated without enumerating all possible reconstructions of the missing parents. The missing parental genotypes only occur in the term $\sum_{f,m \in Q(c)} 2^{h_{f,m}(1-k)}$, but this can be calculated directly from the numbers of homozygous and heterozygous genotypes in $Q(c)$, which are trivial to obtain (see examples 2 and 4). In the case of two heterozygous siblings it turns out that only four configurations need be considered to cover all possible reconstructions of the parents (Table 2). Thus the amount of computation for this family is more than halved. Greater efficiencies are obtained for more polymorphic markers or haplotypes.

Table 2 Inferable reconstructions of the missing parents, given two heterozygous sibs. Sib genotypes are given with the paternally transmitted allele first, but parental genotypes are unphased. Asterisk (*) denotes either allele. The rightmost column gives the total weight due to the missing parents (see Table 1 and Equation (16)). For example the top row is the sum of rows 1, 6, 7, and 9 from Table 1.

Father	Mother	Sib 1	Sib 2	$\sum_{f,m \in Q(c)} 2^{h_{f,m}(1-k)}$
1 *	2 *	1 2	1 2	9/4
1 2	1 2	1 2	2 1	1/4
1 2	1 2	2 1	1 2	1/4
2 *	1 *	2 1	2 1	9/4

Examples

We illustrate our procedure on some examples. Consider two sibs genotyped at a marker with four alleles *A, B, C, D*. We assume the parental origin of all sib alleles is known; in practice it not known, but the different resolutions form the elements of the set Γ' that are summed over in (15).

In each example the likelihood contributions have the form of (16), differing in the sib genotypes c, consistent mating types $Q(c)$, and the elements of the conditioning set V_c.

Example 1. The sib genotypes are (*AC, BD*), where the paternally transmitted allele is given first. The set of paternally transmitted alleles is $\{A, B\}$ and that of the maternally transmitted alleles $\{C, D\}$. After conditioning on the inheritance vector class we obtain $V_c = \{(AC, BD), (AD, BC), (BC, AD), (BD, AC)\}$ and $2^{h_{f,m}(1-k)} = \frac{1}{4}$.

Example 2. The sib genotypes are (*AB, AC*). Then the set of paternally transmitted alleles is just $\{A\}$ whereas the maternally transmitted alleles are $\{B, C\}$. Further conditioning on the inheritance vector class gives $V_c = \{(AB, AC), (AC, AB)\}$. There are three heterozygous and one homozygous completions of the father, so $\sum_{f,m \in Q(c)} 2^{h_{f,m}(1-k)} = 3 \cdot 2^{-2} + 2^{-1} = \frac{5}{4}$.

Example 3. The sib genotypes are (*AB, BA*). For both parents the set of transmitted alleles is $\{A, B\}$ and similarly to example 1, $V_c = \{(AB, BA), (AA, BB), (BB, AA), (BA, AB)\}$ and $2^{h_{f,m}(1-k)} = \frac{1}{4}$.

Example 4. The sib genotypes are (*AB, AB*). The set of paternally transmitted alleles is just $\{A\}$ and the maternally transmitted alleles $\{B\}$. Therefore V_c is just $\{(AB, AB)\}$ and the conditional likelihood contribution is one whatever the values of the missing parental alleles. As in Example 2, there are three heterozygous and one homozygous completions for each parent, so $\sum_{f,m \in Q(c)} 2^{h_{f,m}(1-k)} = 9 \cdot 2^{-2} + 2 \cdot 3 \cdot 2^{-1} + 2^0 = \frac{25}{4}$.

Remark 1. Given unphased sib genotypes, our approach renders more sibships informative for association than the conditioning set out in Table 7 of (Rabinowitz & Laird, 2000). The reason is that although both methods condition on the sets of transmitted haplotypes, our approach does so after first allowing for uncertain phase in the sibs. Examples 3 and 4 show that, given two heterozygous sibs, the first scenario can be informative for association, whereas this family would be uninformative under the conditioning of (Rabinowitz & Laird, 2000), implemented in FBAT (Lake et al., 2000).

Remark 2. There are symmetries in the genotypes that are summed over, particularly if additive coding is used, in which genotype effects are represented by the sum of haplotype effects. In this case a sibship is only informative if there is variation in trait values y . This is similar to other sibship tests of association (Spielman & Ewens, 1998; Siegmund et al., 2000), though not all (Rabinowitz, 2002). Thus, in the case of a binary trait we require both unaffected and affected sibs, and in the case of sib pairs this reduces to a discordant sib-pair analysis. In our implementation, we only apply the sibship model if more than one trait is observed in the sibs; otherwise we apply the default model summing over all possible missing parents.

Remark 3. If there are missing data in the sibs, such as haplotype phase uncertainty, then we may simply enumerate all possible completions consisting of at most four distinct haplotypes, assuming no recombination if necessary. For each completion we then proceed as above by conditioning on the sets of transmitted haplotypes and the inheritance vector class. Thus if we had two heterozygous sibs, the likelihood contribution would be the sum of the contributions described in examples 3 and 4.

Quantitative Traits

For analysis of a quantitative trait, a simple extension is to allow y to take any real value. This effectively makes models (6) and (7) into multinomial logistic models including an interaction between genotype and y, an approach that has been explored by several authors (Waldman et al., 1999; Kistner & Weinberg, 2004; Wheeler & Cordell, 2007). The main difficulty with this approach is interpretation of the estimated effect, and a more accurate approach is to use a normal or other distribution for $\Pr(y|c)$ throughout. Our implementation assumes that sib traits are multivariate normal with the mean vectors given by linear predictors $X_c'\tilde{\beta}$ and $X_c'\beta$, constant variance σ^2, and constant pairwise residual covariance ρ. In order to identify all parameters under all hypotheses, it is convenient to fix σ^2 and ρ *a priori* (Dudbridge, 2008). When genotype effects β are small, second-order terms are negligible and the model is well approximated by the multinomial model. The two approaches therefore have similar power for small effects, and either model is valid for testing the null hypothesis of no effect. Exactly the same modifications are used for sibships as described above for binary traits.

Simulations and Data

We have implemented the proposed sibship model as an option in UNPHASED. The previous approach, which treats a sibship as a family with two missing parents and sums over all possible parents, is retained as the default option. We refer to the previous model as UNPHASED/default and our new model as UNPHASED/sibship. We compared these two models to each other and also to the FBAT software under a number of scenarios. We chose FBAT because it also has a unified approach to binary and quantitative traits, can perform haplotype analysis, deal with arbitrary pedigree structures, and can test for association in the presence of linkage. Of other available software that could be applied to sibships, TRANSMIT (Clayton, 1999) gives very similar results to UNPHASED/default, PLINK/DFAM (Purcell et al., 2007), gives very similar results to FBAT in sibships, and QTDT (Abecasis et al., 2000) requires prior computation of IBD probabilities and cannot perform haplotype analysis (for parent-child trios, Gauderman has shown similar power of QTDT to the retrospective model implemented in UNPHASED [Gauderman, 2003]). We therefore felt that FBAT provided the most relevant general-purpose comparison.

We simulated a binary disease trait and selected 500 sib-pairs that were discordant for affection status. The disease locus was diallelic with risk allele frequency 0.3 and allelic relative risk varied between 1.0 and 1.5, under a multiplicative model. We also simulated a quantitative trait in 500 unselected sib-pairs, having the standard normal distribution with each minor allele of the trait locus having an additive effect that was varied between 0.0 and 0.5. The trait locus had increaser allele frequency 0.3. For each value of the relative risk and additive QTL effect we estimated power by directly testing the trait locus in 1000 simulated data sets.

To assess the effect of population stratification, we simulated a second sample of 500 sib pairs in which the allele frequencies were equal, the mean trait value was 2.0, and there was no additive effect of genotype. This was then combined with the first sample and quantitative trait analysis was performed on the combined sample of 1000 sib pairs. This is an extreme example of stratification, as there are two equally sized subpopulations with a strong difference in allele frequency and a clearly bimodal marginal trait distribution; realistic situations of latent stratification will be far more subtle. Type-1 error was estimated at $\alpha = 5\%$.

We then repeated the above simulations for a haplotype analysis, simulating a second diallelic marker with no disease effect and in linkage equilibrium with the disease or trait locus. Under this model, there are two associated haplotypes with the same effect size. We estimated power for the omnibus test of haplotype effects.

We then compared the methods on two data sets with substantial numbers of sibships. The first consists of a set of diverse nuclear family structures typical of those seen in multiplex families ascertained for the study of a late-onset disease. There are 828 nuclear family structures consisting of a proband and at least one of: a parent, an unaffected sib, or two affected sibs (up to eight sibs). The distribution of structures was based on a real data set

(D. F. Levinson, personal communication), but no genotypic data were provided, only a list of representative structures. A pertinent question for this type of data set is to identify the best method of analysis for a genome-wide association scan.

In such a mixture of nuclear family types, FBAT adjusts its conditioning scheme according to how many parents are genotyped (Rabinowitz & Laird, 2000). UNPHASED/default sums over all completions of any missing parent, whereas UNPHASED/sibship applies the model described above when both parents are missing and more than one trait is seen in the sibs. Following Remark 2 above, we always use the UNPHASED/default model for families in which all sibs have the same trait, or if only one parent is missing, so as to render all families informative for association. Of the 828 nuclear family types considered here, our sibship model is applied to 81.

In these families, we simulated a disease locus with risk allele frequency 0.3 and relative risk 1.3 under a multiplicative model. We estimated power by directly testing this locus in 1000 simulated data sets, and we checked type-1 error by simulating a marker locus that was completely linked to, but not associated with, the disease locus, and which had the same allele frequency. To allow for the linkage we conditioned on the inheritance vector class in both options of UNPHASED and used the empirical variance option in FBAT.

The second data set consists of 335 female sibships, generally ascertained on the basis of an osteoporotic proband. The sample contains a total of 769 women collected in Australia and the United Kingdom with up to seven sibs per family. The sibships have been genotyped at a number of candidate genes and phenotyped for quantitative measurements of bone mineral density for studies of bone and calcium metabolism (Mullin et al., 2008).

We are particularly interested in some SNP associations that, in our initial analyses, were more significant using FBAT than with UNPHASED/default. We compared UNPHASED/sibship to FBAT and UNPHASED/default both for the original data and for simulated trait values based on the same genotypes. This allowed us to assess whether the difference in results seen in the original data was consistent with an overall difference in power, possibly due to violations of model assumptions such as random mating and trait normality. We simulated trait values from the standard normal distribution with an additive effect of 0.3 for each copy of the minor allele. We estimated power from 1000 simulated data sets, both using the simulated normal traits and also using their exponentiated values in order to consider a skewed (log-normal) trait distribution.

Results

Table 3 shows the estimated power for a simulated disease in 500 discordant sib pairs, and Table 4 the power for a simulated quantitative trait in 500 unselected sib pairs. There are no significant differences in power between methods, which is not surprising as the sib genotypes are all directly observed. Although UNPHASED can use information on untransmitted alleles that FBAT does not, this has no discernible effect in the situations considered here.

Table 3 Power (%) to detect association (at $\alpha = 0.05$) of a diallelic disease locus with minor allele frequency 0.3 in a sample of 500 discordant sib pairs. Power is estimated from 1000 simulated data sets. Standard error is 1.35% at power 5% and 95%, 3.1% at power 50%.

Relative risk	UNPHASED/ default	UNPHASED/ sibship	FBAT
1.0	4.5	4.8	5.0
1.1	10.7	10.9	11.3
1.2	29.3	31.5	31.7
1.3	52.6	55.6	56.2
1.4	78.7	80.1	81.1
1.5	92.2	92.2	92.2

Table 4 Power (%) to detect association (at $\alpha = 0.05$) of a diallelic quantitative trait locus with increaser allele 0.3 in a sample of 500 unselected sib pairs. Power is estimated from 1000 simulated data sets. Standard error is 1.35% at power 5% and 95%, 3.1% at power 50%.

Additive value	UNPHASED/ default	UNPHASED/ sibship	FBAT
0.0	4.5	3.9	3.8
0.1	16.9	16.7	17.4
0.2	54.9	52.4	53.2
0.3	88.7	85.3	86.1
0.4	98.8	98.3	98.2
0.5	99.6	99.7	99.8

Under population stratification UNPHASED/default had type-1 error 5.2%, UNPHASED/sibship 5.3%, and FBAT 4.6%, all of which were within the 95% confidence interval (3.65%, 6.35%) from 1000 simulated data sets.

Tables 5 and 6 show the power of the haplotype analyses. There are again no significant differences between UNPHASED/sibship and FBAT, but UNPHASED/default had a small but significantly increased power for the quantitative trait. This small improvement results from a gain in information by using the sib trait values to infer parental haplotypes that are not transmitted to any sib.

Under population stratification in the haplotype analysis, UNPHASED/default had type-1 error 5.7%, UNPHASED/sibship 5.0%, and FBAT 4.6%, all of which are again within the expected range.

For the first data set of mixed nuclear family types, we show results for type-1 error and power in Table 7. Here UNPHASED/sibship has similar power to UNPHASED/default. The power of FBAT is significantly lower, owing to many families with one or both parents missing, for which

Table 5 Power (%) to detect association (at $\alpha = 0.05$) of a two-locus haplotype in a sample of 500 discordant sib pairs. Each locus has minor allele frequency 0.3. The first locus has the relative risk shown, the second has no effect, and is in linkage equilibrium with the first. Power is estimated from 1000 simulated data sets. Standard error is 1.35% at power 5% and 95%, 3.1% at power 50%.

Relative risk	UNPHASED/ default	UNPHASED/ sibship	FBAT
1.0	4.6	4.2	4.7
1.1	8.5	7.9	7.8
1.2	19.1	20.1	20.7
1.3	36.1	39.3	39.1
1.4	62.0	63.5	63.0
1.5	81.5	82.6	82.4

Table 6 Power (%) to detect association (at $\alpha = 0.05$) of a two-locus haplotype to a quantitative trait in a sample of 500 unselected sib pairs. Each locus has minor allele frequency 0.3. The minor allele of the first locus has the additive effect shown. The second locus has no effect and is in linkage equilibrium with the first. Power is estimated from 1000 simulated data sets. Standard error is 1.35% at power 5% and 95%, 3.1% at power 50%.

Additive effect	UNPHASED/ default	UNPHASED/ sibship	FBAT
0.0	5.3	5.2	5.0
0.1	13.5	13.0	12.6
0.2	39.9	36.1	35.4
0.3	77.5	72.0	70.5
0.4	98.3	95.1	94.7
0.5	99.8	99.6	99.7

Table 7 Type-1 error and power (% at $\alpha = 0.05$) for a simulated diallelic disease locus with risk allele frequency 0.3 in a typical sample of 828 nuclear family structures. Standard error is 1.35% for type-1 error and 2.5% at power 80%.

Relative risk	UNPHASED/ default	UNPHASED/ sibship	FBAT
1.0	4.9	5.1	4.6
1.3	77.3	75.5	66.3

UNPHASED gains information by considering the possible parents but which are less informative for FBAT. In view of Tables 3 and 4, these are unlikely to be the families treated as sibships, but are more likely to be those with one parent and at least one unaffected sib, for which UNPHASED has been shown to have greater power than FBAT (see Table 1 of [Dudbridge, 2008]).

For the second data set of sibships with quantitative traits, Table 8 shows the *P*-values for tests of association of two SNPs with normalised measurements of bone mineral density at three anatomical locations (Mullin et al., 2009). We noticed that five of the *P*-values from FBAT were more significant than those from UNPHASED/default, which affects the nominal significance of *rs17080528* with the hip measurement, and may affect whether the femoral neck result remains significant after correcting for multiple testing. UNPHASED/sibship also produced less significant results than FBAT. Since our simulations have shown that UNPHASED should have power at least as high as FBAT, we wondered whether the reversed pattern seen here was due to random variation or some violation of assumptions, such as nonnormality in the trait or nonrandom mating in the parents, that we had not considered in our simulations.

To address this question we fixed the genotypes in all sibships and simulated both a normal trait and a log-normal trait as described in Methods. The power estimates shown in Table 9 suggest that UNPHASED/default has the same or greater power than the other two analyses for a normal trait, but is less stable for a nonnormal trait. UNPHASED/sibship had the same or higher power than FBAT in each case. However, the observed distributions of all three traits were close to normal, and the *P*-values for UNPHASED/default were more similar to those of FBAT for *rs17595772* than for *rs17080528*. The discrepancy in *P*-values seen in Table 8 is likely due to chance, unless there are other features in the data that we have overlooked. Taken together, our results suggest that *rs17595772* is more likely to be a true association than *rs17080528*.

We compared the speed of the three analyses for some representative data sets. The comparison is not a strict benchmark since the programs have different overheads including input/output, book-keeping, and for UNPHASED, the number of evaluations needed to maximise the likelihood, which varies from one data set to the next and also varies between UNPHASED/default and UNPHASED/sibship for a given data set. Very roughly, for single locus analysis in 500 discordant sib-pairs we found the running times for UNPHASED/default, UNPHASED/sibship, and FBAT to be in the ratio 21:8:1, and for a quantitative trait the ratio was 46:12:1. For the nuclear family structures the running times were in the ratio 42:40:1, and for the quantitative trait sibships the ratio was 32:22:1. The improvement in speed is greater for haplotype analysis: for a two-marker haplotype in 500 discordant sib-pairs the running times were in the ratio 32:7:1 and for a quantitative trait the ratio was 35:5:1.

Discussion

We have described a model for association analysis in sibship data that is more computationally efficient than the full

Table 8 *P*-values for association between two SNPs and quantitative measurements of bone mineral density at three locations, in a sample of 335 female sibships.

	rs17080528			rs17595772		
	Spine	Hip	Femoral neck	Spine	Hip	Femoral neck
UNPHASED/ default	0.103	0.223	0.0280	0.00732	0.0181	0.00203
UNPHASED/ sibship	0.0545	0.0426	0.00456	0.043	0.0226	0.00881
FBAT	0.0867	0.0258	0.0031	0.0358	0.0143	0.00127

Table 9 Power (at $\alpha = 0.05$) to detect association of a quantitative trait locus with additive effect 0.3 (exp[0.3]) on a standard normal (log-normal) trait simulated from the fixed genotypes of 335 female sibships. Standard error is 2.5% at power 80%, and 3.1% at power 50%.

	rs17080528		rs17595772	
	Normal	Log-normal	Normal	Log-normal
UNPHASED/ default	0.863	0.701	0.806	0.430
UNPHASED/ sibship	0.782	0.577	0.807	0.594
FBAT	0.768	0.497	0.808	0.617

missing-parents approach previously proposed (Dudbridge, 2008). The efficiency comes from conditioning on the sets of haplotypes transmitted into the sibship by each parent, which means that we do not have to sum over all the possible haplotypes that are not transmitted to any sib. Intuitively, there is little information to be gained from a parental haplotype that is completely missing, and our simulations confirm that our proposed method has comparable power to the previous approach, with substantial gains in speed.

In comparison to the FBAT software, our model had the same power for single SNP analysis and slightly but consistently increased power for haplotype analysis in sibships. Greater improvements were seen in a mixture of sibships and nuclear families with missing parents, but this was due to families with missing parents that were not treated as sibships, such as those with only affected sibs. In addition to higher power, our model gives proper estimates of effect size and easily accommodates covariates. While estimating equations have been derived for the FBAT approach (Vansteelandt et al., 2008) its semiparametric nature tends to require piecemeal development of specific applications. In contrast, we immediately have access to the full flexibility of linear models within a likelihood based analysis, with modest dependence on parametric assumptions.

Nevertheless, our approach remains noticeably slower than that of FBAT and similar programs. This is due to the estimation of nuisance parameters including allele frequency and two sets of effect sizes. We have seen that the programs have similar power in samples of sibs only or complete nuclear families, in which case it seems sensible to use faster methods for large-scale screening and our models for more refined analyses including estimation of effect sizes. However when there is a sizeable number of families with only affected sibs or only one parent, the power gain from our approach suggests that it should be preferred despite the increased running time.

In principle our approach could be applied to families with only one parent, again by conditioning on the set of transmitted haplotypes, but it turns out that there is much less reduction in the number of possible scenarios. Furthermore, single-parent families are rarely the sampling unit of design; more usually they occur sporadically through incomplete ascertainment of a nuclear family. In contrast, sibships without parents are often the intended sampling unit, such as in studies of twins or of late-onset disease. We were motivated to develop an improved method for sibships both by the computational cost of analysis and by the large number seen in current studies.

A recent study compared FBAT to a conditional logistic regression analysis of sibships, using a cluster variance estimator to account for linkage (Nsengimana & Barrett, 2008). They found that the two approaches were broadly similar in power; FBAT had the advantage of an integrated haplotype analysis, whereas conditional logistic regression is a standard tool that easily allows estimation of effects and inclusion of covariates. Our approach is also comparable in power, and combines the advantages of both these alternatives. By setting the problem in a missing-data framework for families, haplotype analysis is immediately possible, as are other missing-data applications such as testing untyped SNPs (Lin et al., 2008). Because our likelihood arises as a special case of the likelihood for a nuclear family, as does a retrospective model for case/control data (Epstein & Satten, 2003; Dudbridge, 2008), it is simple to analyse a combined sample of sibships, families, and unrelated subjects by defining their respective likelihoods in terms of common parameters and working with the summed log-likelihood. We have described an extension to quantitative traits and further extensions to, say, survival or categorical data would seem to be straightforward.

Electronic Database Information

UNPHASED is available from http://homepages.lshtm.ac.uk/frankdudbridge/

Acknowledgements

We thank Doug Levinson, Richard Prince, and Tim Spector for providing data. Funding was provided by the UK Medical Research Council (U.1052.00.012.00001.01) and the National Health and Medical Research Council of Australia (project grants: 294402, 343603). This paper was greatly improved by the thoughtful comments of the reviewers and associate editor.

References

Abecasis, G. R., Cardon, L. R. & Cookson, W. O. (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**, 279–292.

Abecasis, G. R., Cookson, W. O. & Cardon, L. R. (2001) The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am J Hum Genet* **68**, 1463–1474.

Allison, D. B., Heo, M., Kaplan, N. & Martin, E. R. (1999) Sibling-based tests of linkage and association for quantitative traits. *Am J Hum Genet* **64**, 1754–1763.

Aulchenko, Y. S., de Koning, D. J. & Haley, C. (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**, 577–585.

Chen, W. M. & Abecasis, G. R. (2007) Family-based association tests for genomewide association scans. *Am J Hum Genet* **81**, 913–926.

Clayton, D. (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**, 1170–1177.

Cordell, H. J. (2006) Estimation and testing of genotype and haplotype effects in case-control studies: comparison of weighted regression and multiple imputation procedures. *Genet Epidemiol* **30**, 259–275.

Cordell, H. J. & Clayton, D. G. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet* **70**, 124–141.

Croiseau, P., Genin, E. & Cordell, H. J. (2007) Dealing with missing data in family-based association studies: a multiple imputation approach. *Hum Hered* **63**, 229–238.

Dudbridge, F. (2006) UNPHASED technical report 2006/5 (MRC Biostatistics Unit).

Dudbridge, F. (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* **66**, 87–98.

Dudbridge, F., Koeleman, B. P., Todd, J. A. & Clayton, D.G. (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* **66**, 2009–2012.

Epstein, M. P. & Satten, G. A. (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* **73**, 1316–1329.

Gauderman, W. J. (2003) Candidate gene association analysis for a quantitative trait, using parent-offspring trios. *Genet Epidemiol* **25**, 327–338.

Horvath, S. & Laird, N. M. (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. *Am J Hum Genet* **63**, 1886–1897.

Horvath, S., Xu, X. & Laird, N. M. (2001) The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet* **9**, 301–306.

Horvath, S., Xu, X., Lake, S. L., Silverman, E. K., Weiss, S. T. & Laird, N. M. (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* **26**, 61–69.

Jonasdottir, G., Humphreys, K. & Palmgren, J. (2007) Testing association in the presence of linkage–a powerful score for binary traits. *Genet Epidemiol* **31**, 528–540.

Jonasdottir, G., Becker, T., Humphreys, K. & Palmgren, J. (2008) Testing association in the presence of linkage using the GRE and multiple markers. *Genet Epidemiol* **32**, 425–433.

Kistner, E. O. & Weinberg, C. R. (2004) Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genet Epidemiol* **27**, 33–42.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. & Lander, E. S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**, 1347–1363.

Lake, S. L., Blacker, D. & Laird, N. M. (2000) Family-based tests of association in the presence of linkage. *Am J Hum Genet* **67**, 1515–1525.

Li, M., Boehnke, M. & Abecasis, G. R. (2006) Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* **78**, 778–792.

Lin, D. Y., Hu, Y. & Huang, B. E. (2008) Simple and efficient analysis of disease association with missing genotype data. *Am J Hum Genet* **82**, 444–452.

MacGregor, A. J., Snieder, H., Schork, N. J. & Spector, T. D. (2000) Twins. Novel uses to study complex traits and genetic diseases. *Trends Genet* **16**, 131–134.

Mullin, B. H., Prince, R. L., Dick, I. M., Hart, D. J., Spector, T. D., Dudbridge, F. & Wilson, S. G. (2008) Identification of a role for the ARHGEF3 gene in postmenopausal osteoporosis. *Am J Hum Genet* **82**, 1262–1269.

Mullin, B. H., Prince, R. L., Mamotte, C., Spector, T. D., Hart, D. J., Dudbridge, F. & Wilson, S. G. (2009) Further genetic evidence suggesting a role for the RhoGTPase-RhoGEF pathway in osteoporosis. *Bone* **45**, 387–391.

Nsengimana, J. & Barrett, J. H. (2008) Power, validity, bias and robustness of family-based association analysis methods in the presence of linkage for late onset diseases. *Ann Hum Genet* **72**, 793–800.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. & Sham, P. C. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575.

Rabinowitz, D. (2002) Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. *J Am Stat Assoc* **97**, 742–751.

Rabinowitz, D. & 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.

Siegmund, K. D., Langholz, B., Kraft, P. & Thomas, D. C. (2000) Testing linkage disequilibrium in sibships. *Am J Hum Genet* **67**, 244–248.

Spielman, R. S. & Ewens, W. J. (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62**, 450–458.

Stone, J., Gurrin, L. C., Hayes, V. M., Southey, M. C., Hopper, J. L. & Byrnes, G. B. (2010) Sibship analysis of associations between SNP haplotypes and a continuous trait with application to mammographic density. *Genet Epidemiol* **34**, 309–318.

Tiwari, H. K., Barnholtz-Sloan, J., Wineinger, N., Padilla, M. A., Vaughan, L. K. & Allison, D. B. (2008) Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered* **66**, 67–86.

Vansteelandt, S., Demeo, D. L., Lasky-Su, J., Smoller, J. W., Murphy, A. J., McQueen, M., Schneiter, K., Celedon, J. C., Weiss, S. T., Silverman, E. K. & Lange, C. (2008) Testing and estimating gene-environment interactions in family-based association studies. *Biometrics* **64**, 458–467.

Waldman, I. D., Robinson, B. F. & Rowe, D. C. (1999) A logistic regression based extension of the TDT for continuous and categorical traits. *Ann Hum Genet* **63**, 329–340.

Wheeler, E. & Cordell, H. J. (2007) Quantitative trait association in parent offspring trios: extension of case/pseudocontrol method and comparison of prospective and retrospective approaches. *Genet Epidemiol* **31**, 813–833.

Whittaker, J. C. & Morris, A. P. (2001) Family-based tests of association and/or linkage. *Ann Hum Genet* **65**, 407–419.

Received: 4 February 2009
Accepted: 4 October 2010