Original Paper

Human Heredity

Hum Hered 2008;66:87–98
DOI: 10.1159/000119108

Published online: March 31, 2008

Likelihood-Based Association Analysis for Nuclear Families and Unrelated Subjects with Missing Genotype Data

Frank Dudbridge

MRC Biostatistics Unit, Cambridge, UK

Key Words
Conditional likelihood · Family-based association tests · Missing data · Population stratification · Transmission/disequilibrium test · Unphased genotype data

Abstract
Missing data occur in genetic association studies for several reasons including missing family members and uncertain haplotype phase. Maximum likelihood is a commonly used approach to accommodate missing data, but it can be difficult to apply to family-based association studies, because of possible loss of robustness to confounding by population stratification. Here a novel likelihood for nuclear families is proposed, in which distinct sets of association parameters are used to model the parental genotypes and the offspring genotypes. This approach is robust to population structure when the data are complete, and has only minor loss of robustness when there are missing data. It also allows a novel conditioning step that gives valid analysis for multiple offspring in the presence of linkage. Unrelated subjects are included by regarding them as the children of two missing parents. Simulations and theory indicate similar operating characteristics to TRANSMIT, but with no bias with missing data in the presence of linkage. In comparison with FBAT and PCPH, the proposed model is slightly less robust to population structure but has greater power to detect strong effects. In comparison to APL and MITDT, the model is more robust to stratification and can accommodate sibships of any size. The methods are implemented for binary and continuous traits in software, UNPHASED, available from the author.

Copyright © 2008 S. Karger AG, Basel

Introduction

Association analysis is the preferred approach for identifying genes influencing complex traits, as linkage disequilibrium between genotyped markers and causal loci can now be detected at very fine scales [1]. Simple analyses can be performed in contingency tables [2], but genetic data have complicating features that have motivated a large number of novel methods. An important class of methods uses family-based designs, in which genotypes are measured on subjects of interest and also on their relatives, typically their parents or siblings [3]. These methods are robust to confounding by population stratification, as they consider the association within but not between families. This is potentially important because gene frequencies can vary randomly between sub-populations that have different trait distributions [4]. Furthermore, family-based designs allow the identification of parent-of-origin and maternal-fetal interaction effects [5]. These methods have been very popular in recent

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2008 S. Karger AG, Basel
0001–5652/08/0662–0087$24.50/0

Accessible online at:
www.karger.com/hhe

Frank Dudbridge
MRC Biostatistics Unit, Institute for Public Health
Robinson Way, Cambridge CB2 0SR (UK)
Tel. +44 1223 330 300, Fax +44 1223 330 388
E-Mail frank.dudbridge@mrc-bsu.cam.ac.uk

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

years, but the realization that large samples are needed to detect small effects, and the development of methods to adjust for population stratification, have led to a growing preference for population-based studies of unrelated subjects. In that case, standard epidemiological methods such as logistic and linear regression can be used with little modification [6].

Missing data occur in genetic association studies for a number of reasons. In family-based designs, it is not always possible to recruit all the required family members: for example, in late-onset disease it is often difficult to obtain both parents of an affected case. Another problem arises in haplotype analysis, as genotype data is usually observed without phase, leading to ambiguity in the haplotypes of multiple loci [7, 8]. Also, sporadic genotype failures may occur, or there is uncertainty in the genotype call. Although it is possible to restrict analysis to the complete data only [9], such an approach is likely to lose power in comparison with approaches that accommodate missing, or ambiguous, data. A number of such approaches are available for unrelated subjects, based on maximum likelihood using the EM algorithm, and are reviewed elsewhere [10, 11].

In applying maximum likelihood to family-based designs, there is a problem that a model for the missing data may be mis-specified, in particular by failing to allow for unobserved population stratification. For example, a common approach is to assume Hardy-Weinberg equilibrium (HWE) in order to reduce the number of model parameters, but this may not hold in a stratified population. Although this problem applies also to studies of unrelated subjects, robustness to a population model is a motivating property of family-based designs. Two approaches to this problem are in common use. Clayton's method, implemented in TRANSMIT [12], uses a weighted score function in which the weights are calculated from an estimated distribution of the missing data. Although not completely robust to population stratification, this approach has good operating characteristics [13] and reduces to the transmission/disequilibrium test [14] when the data are complete. The second approach uses a working distribution for the missing data to construct a score function that has expectation zero under the null hypothesis, whatever the true distribution [15–17]. This semi-parametric approach has recently been improved to allow the locally most efficient analysis, with an improved computational algorithm implemented in the PCPH software [18]. This method is always robust to population stratification, and its only potential disadvantage is that its optimality properties only hold locally to the null hypothesis. In addition to these approaches, some other methods have been implemented based on full-likelihood models for nuclear families, including TDTPHASE [16], FAMHAP [19], LAMP [20] and WHAP [21]. These methods are not robust to population stratification even when the data are complete, owing to the forms of their likelihood functions, and will not be considered further here.

Another problem in family-based association is that transmissions to multiple siblings cannot be treated as independent observations in the presence of linkage [22]. Both FBAT [9] and TRANSMIT [12] approach this problem by treating siblings as independent and then using a robust variance estimate that treats each family as a cluster of observations. However in this case TRANSMIT has been shown to be biased when there are missing data [23]. Some methods have been proposed to jointly estimate parameters for linkage and association, allowing valid association analysis in the presence of linkage, but none are entirely satistfactory. The APL [23] and QTDT [24] methods are limited by a rapidly increasing number of identity-by-descent parameters, as the sibship size increases, whereas the PSEUDOMARKER [25] and LAMP [26] methods do not address the issue of population stratification. Furthermore, multilocus analysis across unlinked regions seems problematical.

This paper describes a general-purpose model for association that can be used with nuclear families, unrelated subjects and combinations of the two, and addresses some of the limitations of current methods. The analysis is based on a retrospective likelihood that models the probability of all the genotypes in a nuclear family, given the traits of the children. By defining separate association parameters in the parental and offspring components of the likelihood, similar operating characteristics to TRANSMIT are achieved within an ordinary likelihood framework. Furthermore, this approach allows a procedure, called conditioning on the inheritance vector, which maintains robustness to linkage when there are multiple offspring, even when there are missing data. Unlike existing methods, this procedure does not require estimation of linkage parameters, and easily accommodates multilocus analysis. Unrelated subjects are regarded as the children of two missing parents and are then readily included in the same formulation. The methods are developed for binary and continuous traits and are implemented in software, UNPHASED, which is available from the author (see Electronic-Database Information).

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

Methods

Binary Traits

In nuclear families the data are often ascertained through the trait values of one child (the proband) and perhaps its siblings. Often too, unrelated subjects are selected through their trait values: examples include the standard case/control design and extreme sampling of continuous traits. For general purposes a retrospective model is appropriate, in which the likelihood reflects the probability of the genotypes given the traits of all children. Furthermore, if there are additional covariates whose main effects are not of interest, it is convenient to further condition on the covariate values [27]. As the ascertainment may not actually depend on the covariate values, this approach may not give efficient estimation, and likelihoods that reflect the true mechanism, when known, are more appropriate [28, 29]. Nevertheless, conditioning on both trait and covariate values is convenient as a general purpose strategy and often leads to estimation of genotype effects that is robust both to the ascertainment mechanism and to departures from the assumed trait distribution.

Consider a nuclear family having k children, with paternal genotype F, maternal genotype M, child genotypes $C = (C_1, ..., C_k)$, child traits $Y \in \{0, 1\}^k$ and child covariates Z. Genotypes may encompass multiple loci. The probability of genotypes, conditional on trait and covariate values, is

$$\begin{aligned} &Pr\left(F = f, M = m, C = c \mid Y = y, Z = z\right) \\ &= Pr\left(f, m \mid y, z\right) Pr\left(c \mid f, m, y, z\right) \qquad (1) \\ &= \frac{Pr\left(f, m \mid z\right) Pr\left(y \mid f, m, z\right)}{Pr\left(y \mid z\right)} \cdot \frac{Pr\left(c \mid f, m, z\right) Pr\left(y \mid f, m, c, z\right)}{Pr\left(y \mid f, m, z\right)} \qquad (2) \end{aligned}$$

Assume that child traits are independent of parental genotypes, so $Pr(y \mid f, m, c, z) = Pr(y \mid c, z)$. Assume further that there is no transmission distortion and no linkage, so $Pr(c \mid f, m, z)$ is either zero or a constant. Then

$$\begin{aligned} Pr\left(f, m, c \mid y, z\right) = &\frac{Pr\left(f, m \mid z\right) \sum_{c^* \in S(f,m)} Pr\left(y \mid c^*, z\right)}{\sum_{f^*, m^*} \sum_{c^* \in S(f^*, m^*)} Pr\left(f^*, m^* \mid z\right) Pr\left(y \mid f^*, m^*, c^*, z\right)} \\ &\cdot \frac{Pr\left(y \mid c, z\right)}{\sum_{c^* \in S(f,m)} Pr\left(y \mid c^*, z\right)} \qquad (3) \end{aligned}$$

where $S(f, m) = \{c: Pr(c \mid f, m, z) > 0\}$ is the set of child genotypes consistent with parents f, m. The second term in (3) contributes to the conditional on parental genotypes likelihood [30], from which the transmission/disequilibrium test [14] and extensions are derived. However it is unclear how to form that likelihood when the parental genotypes are missing. Here the full likelihood contribution (3) is used, with distinct sets of parameters used in each of the two terms. The parameters of interest are those in the second, conditional, term, with those in the first, parental, term regarded as nuisance parameters.

Following standard arguments for conditional logistic regression, the conditional term can be written

$$Pr\left(c \mid f, m, y, z\right) = \frac{\exp\left[y'\left(X_c'\beta + X_{c,z}'\gamma\right)\right]}{\sum_{c^* \in S(f,m)} \exp\left[y'\left(X_{c^*}'\beta + X_{c^*,z}'\gamma\right)\right]} \qquad (4)$$

where $X_c = (X_{c_1}, ..., X_{c_k})$ and X_{c_j} is a vector of numerical codes for genotype c_j, $X_{c,z}$ is similarly a matrix that codes for interactions between c and covariates z, and β and γ are vectors of fixed effects. The parameters β are the log odds ratios for the main genotype effects, whereas γ are those for gene-covariate interactions.

In the parental term, the same form can be used to model $Pr(y \mid c^*, z)$. For $Pr(f, m, \mid z)$, a multinomial logistic model is used since the parental mating type is a nominal categorical variable. This gives a parental term of the form

$$Pr\left(f, m \mid y, z\right) = \frac{\sum_{c^* \in S(f,m)} \exp\left[X_{f,m,z}'\lambda + y'\left(X_{c^*}'\tilde{\beta} + X_{c^*,z}'\tilde{\gamma}\right)\right]}{\sum_{f^*, m^*} \sum_{c^* \in S(f^*, m^*)} \exp\left[X_{f^*,m^*,z}'\lambda + y'\left(X_{c^*}'\tilde{\beta} + X_{c^*,z}'\tilde{\gamma}\right)\right]} \qquad (5)$$

where $X_{f,m,z}$ denotes a vector of codes for the mating type (f, m) including interactions with child covariates z, and λ, $\tilde{\beta}$ and $\tilde{\gamma}$ are vectors of fixed effects. When $y = 0$, the total likelihood contribution is that for a multinomial logistic model with predictor $X'_{f,m,z}\lambda$, so that λ should be regarded as a parameterization of the mating type distribution in the parents of unaffecteds. Similarly to the conditional term, $\tilde{\beta}$ are the log odds ratios for the main genotype effects and $\tilde{\gamma}$ are those for gene-covariate interactions.

When there are missing genotype or uncertain haplotype data, the likelihood contribution is the sum of the probabilities of each possible completion. That is, defining the set of possible completions as

$$\mathcal{C} = \{f, m, c: Pr(F = f, M = m, C = c \mid \text{observed } F, M, C) > 0\} \qquad (6)$$

the likelihood contribution becomes

$$\sum_{f, m, c \in \mathcal{C}} Pr\left(f, m, c \mid y, z\right) \qquad (7)$$

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

$$l\left(\beta, \gamma, \tilde{\beta}, \tilde{\gamma}, \lambda\right) = \sum_{i=1}^{N} \log \sum_{f, m, c \in \mathcal{C}_i} Pr\left(f, m, c \mid y_i, z_i\right) \qquad (8)$$

with the probability term given by (4, 5).

Note that for the true population parameters $\tilde{\beta} = \beta$ and $\tilde{\gamma} = \gamma$, if the mating type model is correct, but it need not hold for estimates. For a test of $\beta = 0$ it is valid to set $\tilde{\beta} = 0$: it is shown in the Appendix that this gives the same score function as TRANSMIT [12]. When there are no missing data, the likelihood factorizes completely into parental and conditional components, so that estimation of β is independent of the mating type model and inference is equivalent to that based on the conditional likelihood. This approach is therefore no less efficient than conditional inference, despite the additional nuisance parameters. When there are missing data, the mating type model is used to weight the possible conditional likelihoods, but without confounding the weights with the parameters of interest, which occurs when constraining the parameters by $\tilde{\beta} = \beta$. (It can be shown that this constraint is equivalent to the unconditional likelihood model implemented in TDTPHASE [16] and FAMHAP [19].) Because the mating type

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

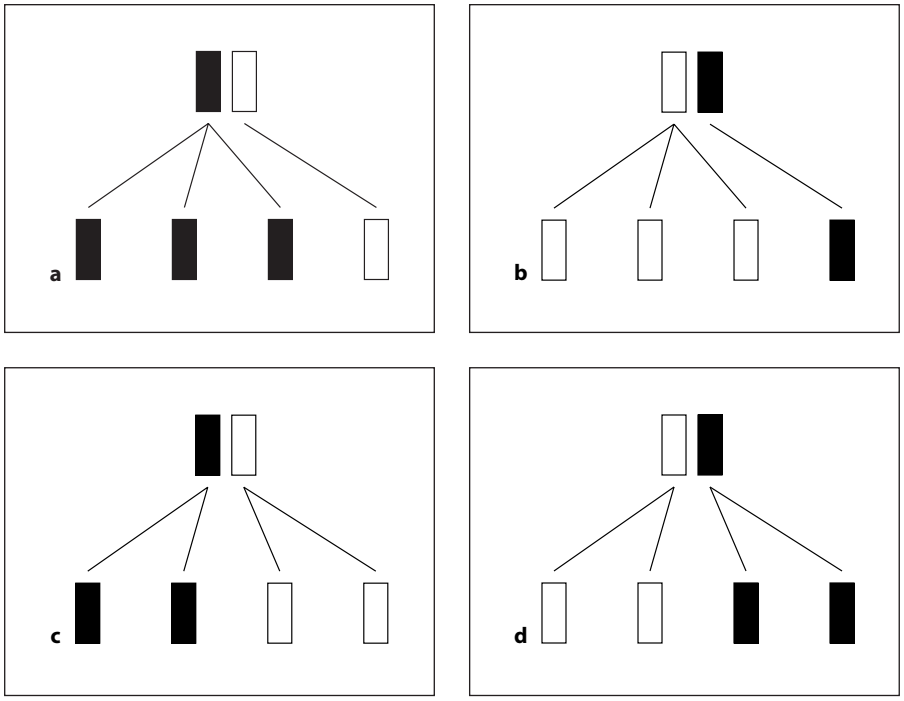

Fig. 1. a Transmissions from one parent to four siblings. Parental genotype is unphased but siblings are ordered. **b** Alternative transmissions from the same class of inheritance vector as **a**. **c**, **d** A second example of two inheritance vectors from the same equivalence class.

model affects inference on β only through a weighting role, this approach is expected to be robust to moderate mis-specification of the mating type model.

In the presence of linkage, $Pr(c \mid f, m, z)$ is not constant for $c \in S(f, m)$, as assumed in (3), since it depends upon the allele-sharing probabilities induced by the traits and the genetic model. In order to maintain a similar form to (3), a novel conditioning step is now introduced. A sufficient statistic for linkage is the inheritance vector, which specifies the grandparental origin of each allele transmitted to each sibling. Because nuclear families do not include the grandparents, there are two inheritance vectors for each parent that are consistent with the observed transmissions, and these form an equivalence class [31]. Conversely, for each parent there are two sets of possible child haplotypes consistent with the true inheritance vector, being the transmitted haplotypes themselves, and the set of untransmitted haplotypes (fig. 1). Replacing $S(f, m)$ by the combination of these haplotypes corresponds to conditioning on the equivalence class of the inheritance vector, in other words conditioning on the sufficient statistic for linkage.

More precisely, for child j with genotype c_j, define three *virtual genotypes* as follows:

u_j^m : haplotype transmitted by father/haplotype not transmitted by mother

u_j^f: haplotype not transmitted by father/haplotype transmitted by mother

u_j^{fm} : haplotype not transmitted by father/haplotype not transmitted by mother.

Then the virtual genotype vectors u^m, u^f and u^{fm} have the same inheritance vector as the observed child genotypes c, up to an equivalence class. Then, still assuming no transmission distortion, $Pr(c^* \mid f, m, z)$ is constant for $c^* \in \{c, u^f, u^m, u^{fm}\}$. Denoting the equivalence class by $\mathfrak{E}$, the likelihood contribution is now

$$Pr\left(f,m,c \mid y,z,\mathfrak{E}\right) = \frac{\sum\limits_{c^* \in \left\{c,u^m,u^f,u^{fm}\right\}} \exp\left[X'_{f,m,z}\lambda + y'\left(X'_{c^*}\tilde{\beta} + X'_{c^*,z}\tilde{\gamma}\right)\right]}{\sum\limits_{f^*,m^*} \sum\limits_{c^* \in S(f^*,m^*)} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde{\beta} + X'_{c^*,z}\tilde{\gamma}\right)\right]} \cdot \frac{\exp\left[y'\left(X'_c\beta + X'_{c,z}\gamma\right)\right]}{\sum\limits_{c^* \in \left\{c,u^m,u^f,u^{fm}\right\}} \exp\left[y'\left(X'_{c^*}\beta + X'_{c^*,z}\gamma\right)\right]} \tag{9}$$

This step has the effect of treating the whole family as a single sampling unit, whereas the unconditional form (4, 5) includes independent terms from siblings. Importantly, this is only possible after distinguishing the parameters $(\tilde{\beta}, \tilde{\gamma})$ in the parental term from (β, γ) in the conditional term. If $(\tilde{\beta}, \tilde{\gamma})$ is set equal to (β, γ), then cancellation of terms means that the inheritance vector is not identifiable. Furthermore, if $(\tilde{\beta}, \tilde{\gamma})$ is set to 0 then the inheritance vector cannot be identified in the parental term, leading to misspecification of the mating type distribution. Following arguments given in the Appendix, this suggests a reason for an observed bias in Clayton's method [23] when there are missing parents and multiple siblings.

Continuous Traits

The same approach is applied to continuous traits, except in the specification of the trait distribution $Pr(y \mid c, z)$. Assume that child traits are distributed as multivariate normal with variance-covariance matrix $\sigma^2 I$. This assumes no covariance between siblings, but the retrospective model gives some robustness to this assumption as well as to non-normality. Conditional on the child genotypes and covariates, the mean vector is specified by

$$\theta_{c,z} = \alpha + X'_c\beta + X'_{c,z}\gamma \tag{10}$$

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

where α is a fixed intercept vector and X_c and $X_{c,z}$ are as before. The association parameters β specify additive effects on the mean, relative to a baseline parameter. The likelihood contribution is

$$Pr(f,m,c \mid y,z) = \frac{\sum_{c^* \in S(f,m)} \exp\left[X'_{f,m,z}\lambda + \left(y'\tilde{\theta}_{c^*,z} - \frac{1}{2}\tilde{\theta}'_{c^*,z}\tilde{\theta}_{c^*,z}\right)/\sigma^2\right]}{\sum_{f^*,m^*}\sum_{c^* \in S(f^*,m^*)} \exp\left[X'_{f^*,m^*,z}\lambda + \left(y'\tilde{\theta}_{c^*,z} - \frac{1}{2}\tilde{\theta}'_{c^*,z}\tilde{\theta}_{c^*,z}\right)/\sigma^2\right]} \cdot \frac{\exp\left[\left(y\theta_{c,z} - \frac{1}{2}\theta'_{c,z}\theta_{c,z}\right)/\sigma^2\right]}{\sum_{c^* \in S(f,m)} \exp\left[\left(y\theta_{c^*,z} - \frac{1}{2}\theta'_{c^*,z}\theta_{c^*,z}\right)/\sigma^2\right]} \quad (11)$$

with a similar modification to condition on the inheritance vector in the presence of linkage.

A difficulty with this form is that when there are no genetic or covariate effects, so that $\theta = \alpha$, then α cancels from the likelihood and cannot be identified. When this is the null hypothesis of interest, this leads to problems with asymptotic theory as the intercepts are nuisance parameters that are present only under the alternative. Furthermore, when the effects are small the intercepts are technically identifiable but are difficult to estimate numerically. For these reasons, a practical solution is to subtract α from y, so the mean vector is $\theta_{c,z} = X'_c\beta + X'_{c,z}\gamma$. Now introduce new parameters ν_c to replace $\alpha\theta_{c,z}/\sigma^2$. The likelihood contribution then becomes

$$Pr(f,m,c \mid y,z) = \frac{\sum_{c^* \in S(f,m)} \exp\left[X'_{f,m,z}\lambda + \tilde{\nu}_{c^*} + \left(y'\tilde{\theta}_{c^*,z} - \frac{1}{2}\tilde{\theta}'_{c^*,z}\tilde{\theta}_{c^*,z}\right)/\sigma^2\right]}{\sum_{f^*,m^*}\sum_{c^* \in S(f^*,m^*)} \exp\left[X'_{f^*,m^*,z}\lambda + \tilde{\nu}_{c^*} + \left(y'\tilde{\theta}_{c^*,z} - \frac{1}{2}\tilde{\theta}'_{c^*,z}\tilde{\theta}_{c^*,z}\right)/\sigma^2\right]} \cdot \frac{\exp\left[\nu_{c^*} + \left(y\theta_{c,z} - \frac{1}{2}\theta'_{c,z}\theta_{c,z}\right)/\sigma^2\right]}{\sum_{c^* \in S(f,m)} \exp\left[\nu_{c^*} + \left(y\theta_{c^*,z} - \frac{1}{2}\theta'_{c^*,z}\theta_{c^*,z}\right)/\sigma^2\right]} \quad (12)$$

This approach introduces more nuisance parameters into the model, so there is a cost in power. However it ensures that the parameters are identifiable under all hypotheses about θ; furthermore it also represents a model for transmission distortion, which can be useful if subjects have been selected on the basis of a binary trait [32]. A disadvantage is that population stratification may result in ν_c being random rather than fixed: in this case, estimation of genetic effects will not be accurate, although testing of $\beta = 0$ remains valid.

The variance σ^2 is assumed known. Although may be estimable, the problem remains that it cannot be identified when $\theta = 0$. If no external estimate of the variance is available, the likelihood could be profiled over a range of values to give an indication of a realistic plug-in estimate.

Some simplification is possible by assuming that the genetic and covariate effects are sufficiently small that $\theta'\theta \approx 0$. This gives the likelihood contribution

$$Pr(f,m,c \mid y,z) = \frac{\sum_{c^* \in S(f,m)} \exp\left[X'_{f,m,z}\lambda + \tilde{\nu}_{c^*} + y\tilde{\theta}_{c^*,z}/\sigma^2\right]}{\sum_{f^*,m^*}\sum_{c^* \in S(f^*,m^*)} \exp\left[X'_{f^*,m^*,z}\lambda + \tilde{\nu}_{c^*} + y\tilde{\theta}_{c^*,z}/\sigma^2\right]} \cdot \frac{\exp\left[\nu_c + y\theta_{c,z}/\sigma^2\right]}{\sum_{c^* \in S(f,m)} \exp\left[\nu_{c^*} + y\theta_{c^*,z}/\sigma^2\right]} \quad (13)$$

This is closely related to previous approaches [32, 33] which used multinomial logistic regression models to test association to a quantitative trait. Here these approaches have been extended to allow for multiple siblings and missing data. The advantage of the multinomial models is that they are more robust to non-normality than the retrospective normal likelihood, and can allow more rapid computation via factorization of the denominator in (13). However there is no simple interpretation of the association parameters, and there may be considerable loss of power to detect effects on normally distributed traits.

Unrelated Subjects

A distinction should be made between nuclear families with one child and two missing parents, and unrelated subjects ascertained as such. The former could occur sporadically in a sample of families, and can be treated with the nuclear family models described above. In the latter case, it is more desirable to apply a model designed specifically for unrelated subjects. This can be obtained from the nuclear family model by equating the association parameters in the parental and conditional terms, so that $(\tilde{\beta}, \tilde{\gamma}) = (\beta, \gamma)$. Let g be the genotype of a singleton subject and let u denote the genotype composed of the two haplotypes not transmitted by its parents. Assuming Hardy-Weinberg equilibrium in the parents, the mating type model may be written

$$X_{f,m,z} = X_{g,z} + X_{u,z} \quad (14)$$

giving for binary traits, from (4, 5)

$$Pr(g \mid y,z) = \frac{\sum_{u^*} \exp\left[X'_{u^*,z}\lambda + X'_{g,z}\lambda + y\left(X'_g\beta + X'_{g,z}\gamma\right)\right]}{\sum_{u^*,g^*} \exp\left[X'_{u^*,z}\lambda + X'_{g^*,z}\lambda + y\left(X'_{g^*}\beta + X'_{g^*,z}\gamma\right)\right]} = \frac{\exp\left[X'_{g,z}\lambda + y\left(X'_g\beta + X'_{g,z}\gamma\right)\right]}{\sum_{g^*} \exp\left[X'_{g^*,z}\lambda + y\left(X'_{g^*}\beta + X'_{g^*,z}\gamma\right)\right]} \quad (15)$$

When $y = 0$, the likelihood only depends on $X'_{g,z}\lambda$, so that λ should be regarded as a model for the genotype frequency in controls. This is equivalent to the models for case/control data proposed by Epstein and Satten [34] and Kwee et al. [27]. Following arguments of those authors, β can be regarded as log odds ratios when the rare disease assumption applies.

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

For continuous traits, the likelihood contribution is

$$Pr(g \mid y,z) = \frac{\exp\left[X'_{g,z}\lambda + \nu_g + y\left(\theta_{g,z} - \frac{1}{2}\theta^2_{g,z}\right)/\sigma^2\right]}{\sum_{g^*}\exp\left[X'_{g^*,z}\lambda + \nu_{g^*} + y\left(\theta_{g^*,z} - \frac{1}{2}\theta^2_{g,z}\right)/\sigma^2\right]} \quad (16)$$

When the sample consists only of unrelated subjects, the parameters ν are absorbed into the frequency predictor $X'_{g,z}\lambda$, so they may be omitted, but this predictor is not now a direct model for the genotype frequency. The multinomial logistic approximation for small effects is

$$Pr(g \mid y,z) = \frac{\exp\left[X'_{g,z}\lambda + \nu_g + y\theta_{g,z}/\sigma^2\right]}{\sum_{g^*}\exp\left[X'_{g^*,z}\lambda + \nu_{g^*} + \theta_{g^*,z}/\sigma^2\right]} \quad (17)$$

In a combined sample of nuclear families and unrelated subjects, it is possible to form a total likelihood with shared parameters between the two samples. Such an approach should be treated with care, since the true parameter values may differ between samples. If the genotype frequencies differ, but common frequencies are assumed, then population stratification is effectively introduced into the sample. The model for unrelateds has no protection against stratification, and any type of population structure may bias the analysis. However, frequency differences between families and unrelateds may be accommodated by introducing a covariate into the frequency model indicating whether a subject is a singleton. This covariate may be tested to infer whether a frequency difference exists.

Heterogeneity of effects may exist between families and unrelateds, and indeed is likely since the family-based effects are conditional on shared environmental and genetic factors. This is not necessarily a problem for detecting an effect, as testing the pooled effect may have good power, but assuming a common effect is clearly inaccurate for estimation. Again, separate effects may be accommodated by introducing an indicator covariate into the association model. Epstein et al. [35] have noted that unmodelled heterogeneity in genotype frequencies can result in heterogeneous estimated effects, and suggested testing homogeneity of effects to assess whether samples can be combined. However, this approach would detect genuine heterogeneity of effects even when frequencies are homogeneous, and in that case it might still be useful to test a pooled effect. Using an indicator covariate in the frequency model may be more accurate for assessing whether samples can be combined, and when appropriate this indicator can be used in the association model to allow for effect heterogeneity.

Chromosome X

The proposed model may be applied to X-linked loci with minor modification. As fathers carry one copy of X, there are only two possible children of given parents, so when conditioning on the inheritance vector only one set of virtual genotypes should be considered, say u^m. If males and females are combined in the same analysis, there is a problem of how to model the genetic effect. Males often function as homozygous for the deleterious allele, but if this is not known in advance, the design vector is not easily specified. In general it is probably better to conduct separate analyses of males and females; when they occur in the same sample, this can be achieved through use of an indicator covariate. Note that by conditioning on the trait values, there is no need to account for differing prevalence or trait mean between males and females.

Estimation and Testing

Because the proposed model is an ordinary likelihood, standard methods can be used for parameter estimation and testing [36]. For any design specified in the linear predictors, the likelihood can be maximized to give estimates of the model parameters. Recalling the form of the total log-likelihood (8)

$$\ell(\Theta) = \sum_{i=1}^{N}\log\sum_{f,m,c\in C_i} Pr(f,m,c \mid y_i,z_i) \quad (18)$$

where Θ is the vector of all fixed effects, write the contribution for family i as

$$\ell_i = \log\sum_{f,m,c\in C_i} Pr(f,m,c \mid y_i,z_i) \quad (19)$$

The score vector for family i is

$$U_i = \frac{\partial \ell_i}{\partial \Theta},$$

whose form is given in the Appendix. The variance-covariance matrix of the maximum likelihood parameter estimates can be approximated by the outer product, or empirical, estimator $\{\Sigma U_i U'_i\}^{-1}$. This can be used to construct confidence intervals for the parameters, based on its diagonal elements. Likelihood ratio tests can be used to test nested hypotheses about the model parameters, and Wald tests are also possible for linear contrasts.

The most common test is an omnibus test that detects association to at least one genotype. At the null hypothesis, $\beta = 0$ whereas the alternative allows the elements of β to vary freely. The maximum likelihood estimate $\hat{\beta}$ gives the genetic effects relative to one baseline parameter.

An alternative is a test of individual effects. Here a genetic effect could be compared to the baseline effect, or to a pooled estimate of the other genetic effects. The baseline comparison can be easily performed by a Wald test using the maximum likelihood estimates, as indicated above. For a comparison to other pooled effects, one approach is to define the null hypothesis as $\beta = 0$, and the alternative as all $\beta_j = 0$ except for the effect of interest, which is freely estimated. However, this would require separate estimation for each tested effect. A more efficient approach is a score test based on the first derivatives of the log-likelihood at $\beta = 0$ [37]. Write the nuisance parameter vector as Λ, the score vector for the effects of interest as U_β and its variance-covariance matrix as V_β. Standard theory gives $V_\beta = V_{\beta\beta} - V_{\beta\Lambda}V_{\Lambda\Lambda}^{-1}V_{\Lambda\beta}$, where V_{ij} are the appropriate submatrices of var(U), which can be estimated by $\Sigma U_i U'_i$ evaluated at the maximum likelihood estimate for Λ at $\beta = 0$. Individual score statistics are $U^2_{\beta_i}/U_{\beta_{ii}}$, which are asymptotically χ^2 with one degree of freedom. Because the score test is based on the partial derivatives at $\beta = 0$, the same scores and variances can be used for all individual tests and need only be computed once, in contrast to the likelihood-ratio or Wald tests.

Implementation

Implementation of the proposed model requires specification of the design vectors $X_{f,m,z}$, X_c, $X_{c,z}$. Saturated models allow a parameter for every mating type and genotype, for each combina-

Downloaded by: 198.143.37.65 - 11/30/2015 8:58:47 AM

tion of covariate levels, but these are generally too high-dimensional to be useful. More practical are haplotype coding schemes, which define the genotype design vector as the sum of two haplotype designs, and locus coding schemes, which define multilocus designs as the sum of single-locus designs [38]. Under haplotype coding, the likelihood contributions can be factorized into independent contributions from haplotypes, so that the model assumes HWE. Similarly, locus coding assumes linkage equilibrium. In general it is convenient to adopt a haplotype coding scheme for the mating type design, since HWE is often a realistic assumption but linkage equilibrium is not, and the nuclear family model is expected to be robust to moderate deviations from HWE. The association model may be specified by haplotype or locus coding, or combinations of the two, depending on the desired inference [39]. The haplotype coding scheme is the default in the UNPHASED software and is used for the simulations reported below, although UNPHASED also allows a genotype coding scheme.

It is common to use the EM algorithm to maximize a missing-data likelihood. This is useful when maximum likelihood estimates are easily obtained from complete data, and particularly when they exist in closed form. However the likelihood developed here does not have a standard form: it has a hierarchical structure and allows for combinations of different types of data. Even when the data are complete, iterative numerical algorithms are needed to obtain maximum likelihood estimates, and it is not clear whether the EM algorithm is more efficient than direct maximization of the missing-data likelihood. The current implementation uses a quasi-Newton algorithm [40] based on the score function of the full missing-data likelihood, given in the Appendix, together with the outer product variance estimator.

Results

This section compares the main features of the proposed model to related work, using illustrative simulations. UNPHASED is used both with free estimation of $\tilde{\beta}$, which is the most complete model, and with $\tilde{\beta} = 0$, which gives valid tests of $\beta = 0$, is quicker to compute and is expected to perform similarly to TRANSMIT. The other methods compared are robust to population stratification when the data are complete, and allow for missing data or uncertain haplotype phase. They include TRANSMIT [12], FBAT [17], PCPH [18] and APL [23], whose relative advantages were briefly summarized in the Introduction. In addition, an implementation of multiple imputation is included,MITDT [41], which applies conditional logistic regression to a small number of randomly imputed data sets. With complete data, this approach is equivalent to the transmission/disequilibrium test, whereas when data are incomplete, it is similar to the present method in that the conditional analysis is separated from the missing data model. A disadvantage is that the imputation distribution is estimated under the assumption of association, which can lead to an increase in type-1 error [41]. In contrast, the other methods all have the correct type-1 error under their respective assumptions.

Table 1. Power to detect effect of a three-marker haplotype with odds ratio 1.35 in 600 families

Test	Design			
	Trio		AU1	
	haplotype	global	haplotype	global
UNPHASED free $\tilde{\beta}$	93.4	73.8	86.5	60.0
UNPHASED $\tilde{\beta} = 0$	93.3	73.6	87.1	60.4
TRANSMIT	93.3	73.6	86.8	58.4
PCPH	93.4	74.1	n/a	n/a
APL	93.5	74.1	76.5	62.8
MITDT	92.2	70.8	n/a	n/a
FBAT	93.3	72.6	74.3	42.5

Power (%) is shown for haplotype-specific and global tests, estimated from 1000 simulated samples. 95% confidence interval is approximately ±1.6% at 93%, ±3% at 60%.

Haplotype Analysis

Methods were compared for haplotype analysis under similar conditions to Horvath et al. [17]. Three biallelic markers were simulated with haplotype frequencies (as %) 35, 20, 20, 10, 5, 5, 4, 1. The common haplotype had multiplicative odds ratio 1.35 and the sample size was fixed at 600 families. Two family structures were considered: case parent trios, and families with one affected and one unaffected sibling and one parent genotyped (called AU1 by Horvath et al.). The power was estimated from 1000 samples. Table 1 shows that all methods had similar power in trio families. In AU1 families, UNPHASED, TRANSMIT and APL had similar power, which was somewhat greater than that of FBAT. This occurs because FBAT includes additional conditioning steps to ensure complete robustness to population stratification. APL was less powerful for a specific test of the risk haplotype. Furthermore, the χ^2 statistics from UNPHASED and TRANSMIT were strongly correlated ($\tilde{\beta} = 0$, $r > 0.99$; free $\tilde{\beta}$, $r > 0.98$), confirming the close relationship between their methods. The estimation of additional nuisance parameters by UNPHASED had very minor effects on power. PCPH and MITDT are currently unable to analyze families with more than one child.

The same family structures were then used to estimate type-1 error in a stratified population. All haplotypes now had odds ratio 1 and in half the families, the haplotypes with frequencies 35 and 10% were switched, again

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

Table 2. Type-1 error in the presence of population stratification

Test	Design			
	Trio		AU1	
	haplotype	global	haplotype	global
UNPHASED free $\tilde{\beta}$	4.8	6.0	7.8	7.9
UNPHASED $\tilde{\beta} = 0$	4.8	6.0	8.0	8.6
TRANSMIT	4.8	5.6	7.9	9.4
PCPH	4.4	5.8	n/a	n/a
APL	4.6	5.8	8.1	10.1
MITDT	5.6	5.7	n/a	n/a
FBAT	4.0	5.2	4.0	3.5

Two sub-populations are simulated with different 3-marker haplotype frequencies (see text). Error (%) is shown for haplotype-specific and global tests, estimated from 1000 simulated samples. 95% confidence interval is 3.65–6.35%.

Table 3. Power to detect SNP with odds ratio 1.8 in 200 families

	Power
UNPHASED free $\tilde{\beta}$	90.0
UNPHASED $\tilde{\beta} = 0$	89.8
TRANSMIT	89.8
PCPH	66.4
APL	89.8
MITDT	88.6
FBAT	66.9

Risk allele has frequency 0.3 and parental genotypes are missing at random with probability 0.4. Power (%) is estimated from 1000 simulated samples. 95% confidence interval is ±1.9% at 90%.

following Horvath et al. Type-1 error rates were close to the nominal level in trios, but were slightly inflated in AU1 families for both UNPHASED and TRANSMIT (table 2). Again, the two programs were strongly correlated ($\tilde{\beta} = 0$, $r > 0.99$; free $\tilde{\beta}$, $r > 0.97$). The type-1 error was further inflated for APL. FBAT had the correct error rate as expected.

The power was compared in this stratified population, using the same disease model as above. Absolute power was slightly lower than in table 1, which was expected as the marginal frequency of the risk haplotype was reduced, but differences between methods were similar (data not shown).

Table 4. Type-1 error when testing association in the presence of linkage

	Design	
	0.5AA+0.5AAU	AAAA
UNPHASED free $\tilde{\beta}$	5.0	6.0
UNPHASED $\tilde{\beta} = 0$	9.2	6.2
TRANSMIT	10.3	6.5
APL	5.8	n/a

Marker locus is completely linked to a disease locus with multiplicative relative risk 2.74. Error (%) is estimated from 1000 simulated samples. 95% confidence interval is 3.65–6.35%.

Strong Effect

The previous simulation concerned a weak effect (OR = 1.35 in one haplotype). To illustrate a stronger effect, a single SNP was simulated with risk allele frequency 0.3 and odds ratio 1.8 in 200 trios. Parental genotypes were missing at random with probability 0.4. Table 3 shows that all methods had similar power, except FBAT and PCPH whose power was considerably reduced. This shows that although PCPH is the locally efficient robust method, this property has a cost in power against non-local alternatives when there is a high proportion of missing data. As seen in the haplotype analysis results, there are smaller differences between methods when the effect is weak. Also, as the proportion of missing data is reduced, all methods approach the original TDT, so that the differences again are reduced.

Association in the Presence of Linkage

Following Martin et al. [23], data were simulated consisting of 250 families with two affected siblings (called AA) and 250 families with two affected and one unaffected siblings (called AAU). All parental genotypes were missing. Their MultA model was used, consisting of a SNP with risk allele frequency 0.15 and genotypic penetrances 0.004, 0.011 and 0.030. A marker SNP was simulated, also with minor frequency 0.15, that was completely linked to, but not associated with, the disease SNP. Type-1 error was estimated for TRANSMIT and APL, and for UNPHASED conditioning on the inheritance vector with $\tilde{\beta} = 0$ and with free estimation of $\tilde{\beta}$ (table 4). Similar to results of Martin et al., APL had the correct error rate but TRANSMIT had an inflated error rate. UNPHASED with $\tilde{\beta} = 0$ also had an inflated error rate, and again was strongly correlated to TRANSMIT ($r > 0.99$). However, UNPHASED with free estimation of $\tilde{\beta}$ did have

Downloaded by: 198.143.37.65 - 11/30/2015 8:58:47 AM

the correct error rate, with much weaker correlation to TRANSMIT ($r = 0.73$).

This model was then simulated on 500 families with four affected siblings (AAAA) and each parent having probability 0.5 of being genotyped. These families cannot currently be analysed by APL. Table 4 shows that UNPHASED again had the correct error rate with free estimation of $\tilde{\beta}$, whereas TRANSMIT and UNPHASED with $\tilde{\beta} = 0$ both had slightly increased error rates.

The power of UNPHASED was compared to that of APL, this time for the MultD model of Martin et al. [23], which has genotypic penetrances 0.004, 0.006 and 0.009. For 250 AA families and 250 AAU families, the power of APL was 79.5% and that of UNPHASED 79.3%. The two methods give highly correlated, but not identical results ($r = 0.94$).

Combining Families with Unrelated Subjects

Heterogeneity in genotype frequencies between family and singleton samples may lead to incorrect inference on the odds ratio. Let $\beta^{(t)}$ and $\beta^{(u)}$ denote the log odds ratios in families and singletons respectively. To test whether the samples could be combined, Epstein et al. [35] proposed first testing $\tilde{\beta}^{(t)} = \beta^{(t)}$, and if this is not rejected then testing $\beta^{(t)} = \beta^{(u)}$. Their motivation was to avoid the assumption of HWE in the population. If that assumption is made, however, a more direct approach is to test for homogeneity of frequencies between samples, using an indicator covariate in the frequency model. To illustrate this, one SNP was simulated with minor allele frequency 0.3 in 200 trios and frequency 0.4 in 100 cases and 100 controls. The SNP had no association with disease. In 1000 random samples, the power for the approach of Epstein et al. was 63.2%, whereas the power for testing homogeneity of frequencies was 85%. Furthermore, the indicator covariate can be used to test for a pooled effect in a combined sample with heterogeneous frequencies. Under these simulation conditions the estimated type-1 error for this test was 5.6%, within expectation.

Discussion

The proposed likelihood model is sufficiently flexible for general purpose usage. It accommodates nuclear families of any size, unrelated singletons and combinations of the two. As special cases it reduces to the conditional on parental genotypes model [30, 39] in nuclear families with complete data, and to retrospective likelihood analysis of unrelated subjects [34]. It allows for missing data and uncertain haplotype phase using standard likelihood methods. It has similar operating characteristics to TRANSMIT [12], owing to the relation between their score functions given in the Appendix, but does so within an ordinary likelihood framework.

The main innovations are separation of association parameters in the parental and conditional terms in the likelihood, and conditioning on the inheritance vector. The former has been implicitly done by previous authors [5, 33, 35] who fit a saturated model to the parental mating type. Here, a distinction is made between genotype frequencies and association effects in the parents, which allows more parsimonious models to be fit, including haplotype coding under the HWE assumption. When the mating type model is saturated, all families are the same size and all sibships have the same trait vector, then the association effects cannot be identified in the parental terms and the present model is equivalent to previous work. A related approach is the decomposition of total assocation into between- and within-family components [20, 21, 24]. In a prospective design, the within-family association is a valid estimate in the presence of population stratification [24]. However in the retrospective design used here, the frequency model does not factor out of the likelihood when the data are complete, and so must be correctly specified. This approach is therefore never valid under population stratification, unlike the present model.

Conditioning on the inheritance vector was previously proposed in the context of conditioning on sufficient statistics for missing genotype data [9]. Those authors found a noticable loss of power, owing to a large number of uninformative families, and preferred to use a cluster variance estimate in a test without conditioning. Here, the conditioning has been set into a missing data likelihood framework, in which all families are informative for association. Comparison with the APL program, which estimates the haplotype-sharing probabilities without conditioning on linkage, indicate that the cost in power from the additional conditioning is very small. Further simulations (data not shown) compared power with and without conditioning, when no linkage was assumed, and again found the loss in power to be small.

For combining family samples with unrelated subjects, the proposed approach is similar to that of Epstein et al. [35]. The main difference is that here, HWE is assumed in the singletons, which allows a simple adjustment for population heterogeneity, and reduction to a standard retrospective analysis when there are only singletons. The rationale is that HWE is a common working assumption for unrelated subjects, being somewhat en-

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

sured by standard quality control measures, and heterogeneity is quite likely between samples ascertained under different criteria. Adjustment for heterogeneous genotype frequencies is done through an indicator covariate, and this approach can also be used to combine samples of the same type coming from multiple populations.

In the simulations reported here, the UNPHASED implementation performed as well as the best available methods over a range of situations. In families with a single affected child, the operating characteristics were very similar to those of TRANSMIT. Indeed, when the parental association parameters are set to zero, the results of UNPHASED and TRANSMIT are nearly perfectly correlated, for reasons suggested in the Appendix. When the parameters are freely estimated, the correlation is weaker but the type-1 error and power are still similar. Estimation of the parental parameters is desirable for testing hypotheses in which some effects are nonzero, for estimating effect sizes and allowing for prior linkage in sibships. The additional estimation incurs a small cost in power.

The power of UNPHASED is generally higher than FBAT, at a cost of a small increase in type-1 error when there are missing genotypes and population stratification. The APL program had similar power to UNPHASED, but higher type-1 error under population stratification. PCPH and MITDT had similar power to UNPHASED in trios but cannot currently handle larger sibships.

PCPH is the locally optimal test among those making no assumption on the missing data, but when there is a strong effect and a high proportion of missing data, it loses power in comparison with UNPHASED. Its main advantage is that it is always robust to population stratification, but the compromise approach adopted here appears to incur only small increases in type-1 error. Of course, situations may be constructed in which the increase is much more severe, but in practice careful ascertainment and quality control measures such as HWE testing should ensure that undetected population stratification has only a minor effect on the proposed approach.

The methods may be adapted to categorical, time-to-onset and other traits, through appropriate specification of the trait distribution. The generalized linear model is a convenient representation for many distributions [37], although the retrospective formulation may lead to identifiability issues needing special treatment, as in the proposed model for normal traits. In general the multinomial regression approximation gives a valid test of $\beta = 0$, although it may not be powerful against strong effects and is less appropriate for testing other hypotheses.

In general pedigrees, a simple approach is to extract nuclear families and treat them as independent sampling units. This may ignore correlations between nuclear families in the presence of linkage, although the effect is likely to be small. It is possible to condition on the entire linkage information in a pedigree [42], although the impact of this conditioning may be more severe than in the case of sib pairs considered here. Likelihood models for association in general pedigrees remain an interesting subject for further work. A special case is a sample of sibships without parents. If more than one trait value is present, then the methods described here can be applied, but the analysis may be time-consuming. An alternative is to incorporate conditioning on the sufficient statistic for the missing parental genotypes [42] into a likelihood for the sibships only. This approach has not been pursued here, but offers potential for extending this work to situations in which it is currently not computationally efficient.

Acknowledgments

This work was supported by the EU 6th framework programme (LSHM-CT-2004-503485). Thanks to Pascal Croiseau for early access to the MITDT program.

Appendix: Score Function

Assume a sample of nuclear families measured for a binary trait; the development is readily extended to unrelated subjects and continuous traits using the arguments in Methods. The log-likelihood contribution for family i is (8)

$$\ell_i = \log \sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y,z) \tag{20}$$

with score contribution

$$U_i = \frac{\partial \ell_i}{\partial \Theta} = \frac{\sum_{(f,m,c)\in C_i} \frac{\partial Pr(f,m,c \mid y,z)}{\partial \Theta}}{\sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y,z)} \tag{21}$$

where $\Theta = (\beta, \gamma, \tilde{\beta}, \tilde{\gamma}, \lambda)$. Recall (4, 5)

$$Pr(f,m,c \mid y,z) = \frac{\sum_{c^*\in S(f,m)} \exp\left[X'_{f,m,z}\lambda + y'\left(X'_{c^*}\tilde{\beta} + X'_{c^*,z}\tilde{\gamma}\right)\right]}{\sum_{(f^*,m^*,c^*)\in F} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde{\beta} + X'_{c^*,z}\tilde{\gamma}\right)\right]} \cdot \frac{\exp\left[y'\left(X'_c\beta + X'_{c,z}\gamma\right)\right]}{\sum_{c^*\in S(f,m)} \exp\left[y'\left(X'_{c^*}\beta + X'_{c^*,z}\gamma\right)\right]} \tag{22}$$

Downloaded by: 198.143.37.65 - 11/30/2015 8:58:47 AM

Then the score contribution is given by

$$U_i = \frac{\sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y,z)(D_\beta,\ldots,D_\lambda)}{\sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y,z)} \tag{23}$$

where

$$D_\beta = y'X'_c - \frac{\sum_{c^*\in S(f,m)} \exp\left[y'\left(X'_{c^*}\beta + X'_{c^*,z}\gamma\right)\right] y'X'_{c^*}}{\sum_{c^*\in S(f,m)} \exp\left[y'\left(X'_{c^*}\beta + X'_{c^*,z}\gamma\right)\right]} \tag{24}$$

$$D_{\tilde\beta} = \sum_{c^*\in S(f,m)} y'X'_{c^*} - \frac{\sum_{(f^*,m^*,c^*)\in F} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde\beta + X'_{c^*,z}\tilde\gamma\right)\right] y'X'_{c^*}}{\sum_{(f^*,m^*,c^*)\in F} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde\beta + X'_{c^*,z}\tilde\gamma\right)\right]}$$

$$D_\lambda = X_{f,m,z} - \frac{\sum_{(f^*,m^*,c^*)\in F} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde\beta + X'_{c^*,z}\tilde\gamma\right)\right] X'_{f,m,z}}{\sum_{(f^*,m^*,c^*)\in F} \exp\left[X'_{f^*,m^*,z}\lambda + y'\left(X'_{c^*}\tilde\beta + X'_{c^*,z}\tilde\gamma\right)\right]}$$

and D_γ, $D_{\tilde\gamma}$ are defined similarly.

To see the relation with TRANSMIT, note that it uses a score function of the form

$$U_i = \frac{\sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y)\left(\frac{\partial \log Pr(c \mid f,m,y)}{\partial\beta}, \frac{\partial \log Pr(f,m \mid y)}{\partial\lambda}\right)}{\sum_{(f,m,c)\in C_i} Pr(f,m,c \mid y)} \tag{27}$$

with $\tilde\beta = \beta$ throughout. From (24),

$$\frac{\partial \log Pr(c \mid f,m,y)}{\partial\beta} = D_\beta \tag{28}$$

The difference from the present score function (23) is only in the specification of $Pr(f, m, c \mid y)$, which here serves as a weight, and the use of $(D_{\tilde\beta}, D_\lambda)$ in place of $\partial \log Pr(f, m \mid y)/\partial\lambda$. The score terms for β are identical for each possible (f, m, c). If the mating type model is correct and the effects are small, it is expected that $\hat{\tilde\beta} \approx \beta$ and the results for tests of $\beta = 0$ should be similar. If $\tilde\beta$ is set to zero, then furthermore

$$\frac{\partial \log Pr(f,m \mid y)}{\partial\lambda} = D_\lambda \tag{29}$$

and the two score functions are equal when evaluated at $\beta = 0$. The differences between UNPHASED and TRANSMIT in this case arise only from the specification of the variance estimators.

Electronic-Database Information

UNPHASED software, http://www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased/

References

1 Risch NJ: Searching for genetic determinants in the new millenium. Nature 2000; 405:847–856.
2 Sasieni P: From genotypes to genes: doubling the sample size. Biometrics 1997;53:1253–1261.
3 Laird NM, Lange C: Family-based designs in the age of large-scale gene association studies. Nat Rev Genet 2006;7:385–394.
4 Palmer LJ, Cardon LR: Population stratification and spurious allelic association. Lancet 2003;361:598–604.
5 Weinberg CR,Wilcox AJ, Lie RT: A log-linear approach to case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 1998;62:969–978.
6 Balding DJ: A tutorial on statistical methods for population association studies. Nat Rev Genet 2006;7:781–791.
7 Excoffier L, Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921–927.
8 Dudbridge F, Koeleman BP, Clayton DG, Todd JA: Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 2000;66:2009–2012.
9 Lake SL, Blacker D, Laird NM: Family-based tests of association in the presence of linkage. Am J Hum Genet 2000;67:1515–1525.
10 Satten GA, Epstein MP: Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol 2004;27:192–201.
11 Cordell HJ: Estimation and testing of genotype and haplotype effects in case-control studies: Comparison of weighted regression and multiple imputation procedures. Genet Epidemiol 2006;30:259–275.
12 Clayton D: A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 1999; 65:1170–1177.
13 Nicodemus KK, Luna A, Shugart YY: An evaluation of power and type I error of single-nucleotide polymorphism transmission/disequilibrium-based statistical methods under different family structures, missing parental data, and population stratification. Am J Hum Genet 2007;80:178–185.
14 Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506–516.
15 Rabinowitz D: Adjusting for population heterogeneity and misspecified haplotype frequencies when testing nonparametric null hypotheses in statistical genetics. J Am Stat Assoc 2002;97:742–751.
16 Dudbridge F: Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 2003;25:115–121.
17 Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM: Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol 2004;26:61–69.

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM

18 Allen AS, Satten GA: Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. Genet Epidemiol 2007;31:211–223.
19 Becker T, Knapp M: Maximum-likelihood estimation of haplotype frequencies in nuclear families. Genet Epidemiol 2004;27:21–32.
20 Li M, Boehnke M, Abecasis GR: Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am J Hum Genet 2006;78:778–792.
21 Purcell S, Daly MJ, Sham PC: WHAP: Haplotype-based association analysis. Bioinformatics 2007;23:255–256.
22 Spielman RS, Ewens WJ: The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 1996;59:938–989.
23 Martin ER, Bass MP, Hauser ER, Kaplan NL: Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 2003;73:1016–1026.
24 Abecasis GR, Cardon LR, Cookson WO: A general test of association for quantitative traits in nuclear families. Am J Hum Genet 2000;66:279–292.
25 Göring HH, Terwilliger JD: Linkage analysis in the presence of errors IV: Joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. Am J Hum Genet 2000;66:1310–1327.
26 Li M, Boehnke M, Abecasis GR: Joint modeling of linkage and association: Identifying SNPs responsible for a linkage signal. Am J Hum Genet 2005;76:934–949.
27 Kwee LC, Epstein MP, Manatunga AK, Duncan R, Allen AS, Satten GA: Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. Genet Epidemiol 2007;31:75–90.
28 Lin DY, Zeng D: Likelihood-based inference on haplotype effects in genetic association studies. J Am Stat Assoc 2006;101:89–104.
29 Huang BE, Lin DY: Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet 2007;80:567–576.
30 Schaid DJ, Sommer SS: Genotype relative risks: methods for design and analysis of candidate-gene association studies. Am J Hum Genet 1993;53:1114–1126.
31 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analysis: A unified multipoint approach. Am J Hum Genet 1996;58:1347–1363.
32 Waldman ID, Robinson BF, Rowe DC: A logistic regression based extension of the TDT for continuous and categorical traits. Ann Hum Genet 1999;63:320–340.
33 Kistner EO, Weinberg CR: Method for using complete and incomplete trios to identify genes related to a quantitative trait. Genet Epidemiol 2004;27:33–42.
34 Epstein MP, Satten GA: Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 2003;73:1316–1329.
35 Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA: Genetic association analysis using data from triads and unrelated subjects. Am J Hum Genet 2005;76:592–608.
36 Gould W, Pitblado J, Sribney W: Maximum-likelihood estimation with Stata. College Station, Stata Press, 2005.
37 Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA: Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–434.
38 Clayton D, Chapman J, Cooper J: Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 2004;27:415–427.
39 Cordell HJ, Clayton DG: 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 2002;70:124–141.
40 Press WH, Teukolsky SA, Vetterling WT, Flannery BP: Numerical Recipes in C, ed 2. Cambridge, Cambridge University Press, 1992.
41 Croiseau P, Génin E, Cordell HJ: Dealing with missing data in family-based association studies: a mulitple imputation approach. Hum Hered 2007;63:229–238.
42 Rabinowitz D, Laird NM: A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 2000;50:211–223.

Downloaded by:
198.143.37.65 - 11/30/2015 8:58:47 AM