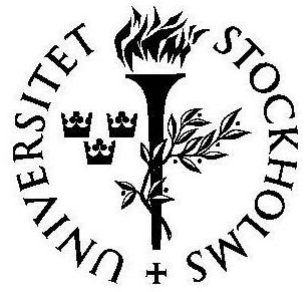

Mathematical Statistics
Stockholm University

Analysis of binary traits. Testing association in the presence of linkage.

Gudrun Jonasdottir
Juni Palmgren
Keith Humphreys

Research Report 2004:13

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2004:13**,
http://www.math.su.se/matstat

Analysis of binary traits. Testing association in the presence of linkage.

Gudrun Jonasdottir
Juni Palmgren
Keith Humphreys*

December 2004

Abstract

It has been shown that testing association in a region with confirmed linkage may increase the rate of false positives in family-based studies. If unaccounted for, the expected similarity between family members may be mistaken for association. Different remedies have been suggested, everything from using a robust variance estimator for the general test statistic FBAT (Family Based Association Tests) to a model-based approach where the linkage is modelled in the covariance structure, the VCM (Variance Components Model). Most methods for testing association in the presence of linkage have been developed for continuous traits. FBAT is one of few methods appropriate for discrete outcomes. In this article we describe a new test of association in the presence of linkage for binary traits. We use a gamma random effects model where association and linkage are modelled as fixed effects and random effects, respectively. We have compared the gamma random effects model to an FBAT and and a GEE-based alternative, in terms of their ability to pick up true signals and their associated false positive rates.

KEY WORDS: Association, Linkage, Association in the Presence of Linkage, Variance Components Model, FBAT.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: gudrunj@math.su.se

1 Background

Testing association in a region with confirmed linkage may increase the rate of false positives in family-based studies. In a linked region one expects similarity between related individuals. If unaccounted for, this similarity may be mistaken for association. Different remedies have been suggested, everything from using a robust variance estimator [1] for the general test statistic FBAT (Family Based Association Tests) [2] to a model-based approach where the linkage is modelled in the covariance structure [3] (VCM - Variance Components Model). The VCM was developed for continuous traits, while FBAT tests for association with both binary and continuous traits, but most methods for testing association in the presence of linkage have been developed for continuous traits. It is important to find new, more powerful, tests of association in the presence of linkage for binary traits.

We compare the program FBAT for binary traits to both the method described in section 2.1 and also a GEE (Generalised Estimating Equation) [4] approach. For the purpose of our comparisons we have used the simulated GAW14 data (Section 2.3). We have compared the three methods ability to pick up a true signal, as well as their rate of false positives.

2 Methods

We consider a random effects model for binary events which is similar in spirit to the multivariate survival model in [5], which models association and linkage as fixed effects and random effects respectively. We use a result for random effects models for binary outcomes which has been described in [6]. It is shown that for gamma distributed random effects, the unconditional distribution of the outcome using a log-log link can be written as a sum of easily calculated terms. Analytical results are only achievable for a few other random effects distributions, such as the beta distribution [6]. The random effects model in [5] assigns one random effect for each of the two parental allele, using the inheritance vector. The authors in [5] do not suggest a way to deal with unknown inheritance vectors. For bi-allelic loci, few parents are informative for transmission, leading to a missing data problem. We solve this crudely by taking a weighted sum over all possible inheritance vectors. We choose a simple mean for our weights, but note that more elaborate weights, such as the reciprocal of the prior probability of the inheritance vector, are possible. The method presented here works for all sizes of sibships, and may also be easily adapted to extended pedigrees.

2.1 A Gamma Random Effects (GRE) model

Let $(Y_{i1}, Y_{i2}, ..., Y_{iJ_i})$ be the binary trait vector for family i and let j denote offspring ($j = 1, 2, ..., J_i$). Let $P(Y_{ij}|\theta_{m_j}, \theta_{p_j})$ denote the conditional probability of trait, given the effect of the maternally and paternally transmitted alleles. We allow for different family sizes. We use a random effects model with a complimentary log(-log) link function.

$$\log(-\log(P(Y_{ij} = 1|\theta_{m_j}, \theta_{p_j}))) = \log(\theta_{m_j} + \theta_{p_j}) + X_j\beta \ . \tag{1}$$

The maternal allele is denoted by m_j $(= 1, 2)$ and the paternal allele by p_j $(= 3, 4)$. The θ's are gamma random effects with scale $\alpha/2$ and shape λ. The probability density function of θ_k is

$$f_\Theta(\theta_k) = \frac{\lambda^{\alpha/2}}{\Gamma(\alpha/2)}\theta_k^{\alpha/2-1}\exp(-\alpha/2 \cdot \theta_k) \ ,$$

$k = 1, 2, 3, 4$.

The unconditional probability of a sibship's trait outcome is not directly tractable. However, the probabilities for all possible ordered set of subsets $Y_{ij} = 1$ for $j \in T$, where T is a subset of the indexes $1, 2, ..., J_i$, can be written as a product of matrixes and scalars (derivation in appendix A1).

$$\pi^* = \prod_{k=1}^{4}\left(\frac{\lambda}{\lambda + \mathbf{B}\ \mathbf{diag}(X_j'\beta)\mathbf{a}}\right)^{\alpha/2} \ . \tag{2}$$

$\mathbf{B}$ is a matrix of indicators, indicating all subsets of subscripts for $\mathbf{Y} = 0$, i.e. $\{\emptyset\}$, $\{1\}$, $\{2\}$, $\{1, 2\}$, $\{3\}$, etc. The elements of matrix $\mathbf{a}$, a_{jk}, indicates if allele k has been transmitted to offspring j, $j = 1, 2, .., J_i$ and $k = 1, 2, 3, 4$. For example, if a sib inherits allele M maternally and allele m paternally allele , then that sib contributes with the row $(1, 0, 0, 1)$. It has been shown [6] that the unconditional probability for all possible outcomes of $\mathbf{Y}$ can be written,

$$\pi = \mathbf{A}^{-1}\ \pi^* \ . \tag{3}$$

The matrix $\mathbf{A}$ indicates all subsets of T. In order to get the probability of the observed Y_{ij} one only needs to pick a row in π. See Table 1 for an example of matrices A and B for three sibs. It is seldom the case that the mode of transmission is known. There are many possible remedies, but we have chosen to take a simple mean over all possible inheritance vectors. Other possibilities would be to weight by the reciprocal of the prior probability of the inheritance vector.

The likelihood for the observed data is

$$\log L(\beta, \alpha, \lambda) = \sum_{i=1}^{n} \pi_i \ . \tag{4}$$

2.2 FBAT and GEE

We compare the GRE with FBAT [2] and a GEE-based alternative [4]. For FBAT we assume a linear allele-dose model, and for the GEE-based alternative we assume a linear allele-dose on the logit scale and an exchangeable covariance structure.

2.3 The GAW14 simulated data

For details of the how the simulation was performed see http://www.gaworkshop.org/data.htm.

All analyses were performed with knowledge of the data simulation process. We chose to analyse the data with respect to trait A. Trait A is known to be associated with haplotypes in the region D3. For the purpose of our comparison we therefore chose to "purchase" markers in the D3 region (B05T4135-B05T4142) as well as markers from the D2 region (B03T3048-B03T3067). Markers in the D2 region are known to not be associated with trait A. Our aim was to use regions D2 and D3 to gain some insight into the performance of the different methods, in terms of both power and validity.

The Aipotu population (one of four simulated populations) only consists of nuclear families, although these are of different sizes. For simplicity, we chose to concentrate on the Aipotu population and to only include families of maximum size six (ie two parents and four offspring).

Power and validity were analysed using two two approaches. First, we merged 10 (out of 100) replicates, in order to get a realistic scenario. This provided us with a total number of 481 independent nuclear families. Secondly, we used a subset of the hundred replicates, each with approximately 48 families. In either scenarios, there was no missing data and we did not simulate any.

We then selected the markers described above and analysed them separately. The method we have described can, however, be easily extended to test multiple markers jointly.

3 Results

We analysed ten merged replicates in regions D2 and D3 and we were able to identify interesting markers in both regions (Figure 1 and Figure 2). In region D2, all three methods (FBAT, GEE and GRE) indicated marker B03T3056 as borderline significant with a p-value of around 0.01 (Figure 1). The peak was slightly less using FBAT. The result should, however, be adjusted for multiple testing. In region D3, where a haplotype based association was simulated, we were able to detect association with marker B05T4136. The detected association was more significant using GEE and GRE (p-value $\approx$ 0.0001), and slightly less significant using FBAT, see Figure 2.

4 Conclusions

In region D2, no association with trait A was simulated. Nevertheless all three methods (FBAT,GEE and GEE) indicated marker B03T3056. Although not significant when correcting for multiple testing, a p-value of 0.01 might raise attention. This illustrates the importance of adjusting for multiple testing when testing for association over an extended region. No overall conclusion of which method is most valid can be drawn from Figure 1, out of two reasons; (i) the results vary from marker to marker and (ii) we only investigate a very limited example from which no general conclusions can be made. The GRE seems at least comparable to both the FBAT and the GEE approach.

In region D3, association with trait A was simulated on a haplotype level. We still chose to test association on a marker level. Again, conclusions drawn should be cautious, given the limited example. However, it seems in this scenario that the GEE and the GRE are slightly better in detecting significant markers. There is of course a problem with power in testing association on markers in a region where the association is generated by haplotypes. However, this might be a real problem in most studies of complex traits where single nucleotide markers are analysed.

We continued analysing the regions of interest using subsets of the replicates. We analysed markers B03T3055, B03T3056 and B03T3057 in region D2, and markers B05T4135, B05T4136, C05R0380 and B05T4138 in region D3. However, due to the small number of families in each replicate, and due to the relatively few number of replicates, results were weak and inconsistent. Each replicate had around 48 families with a maximum of four children, giving a very small power to detect association. Therefore, no conclusions could be made out of the results.

The Gamma Random Effects model presented here seems to work well, compared to both GEE and FBAT. However, more rigid analysis of power and

validity needs to be performed in order to confirm these results. One problem with the GRE is computational time. It is time consuming to evaluate and to maximize the likelihood as many terms need to be calculated for each family. A major advantage is that the likelihood is tractable analytically. Approximate methods are therefore not necessary.

References

[1] Lake, SL, Blacker, D, Laird, NM: **Family-Based Tests of Association in the Presence of Linkage.** *Am J Hum Genet* 2000, **67**: 1515-1525.

[2] Rabinowitz, D, Laird, N: **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.

[3] Fulker, DW, Cherny, SS, Sham, PC, Hewitt, JK: **Combined Linkage and Association Sib-Pair Analysis for Quantitative Traits.** *Am J Hum Genet* 1999, **64**: 259-267.

[4] Liang, KY, Zeger, SL: **Longitudinal data analysis using generalized estimating equations.** *Biometrika* 1986, **73**: 13-22.

[5] Zhong, X, Li, H: **Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model.** *Biostatistics* 2004, **5(2)**: 307-327.

[6] Conaway, MR: **A random effects model for binary trait.** *Biometrics* 1990, **46(2)**: 317-328.

Appendix A1: Computation of π^* in the GRE model.

$$P(Y_{ij} = 1, \forall j \in T)$$

$$= \int_{\theta_4} \int_{\theta_3} \int_{\theta_2} \int_{\theta_1} \prod_{j \in T} P(Y_{ij} = 1 | \theta_{m_j}, \theta_{p_j}, X_j, \beta) \cdot$$

$$\cdot f(\theta_1) f(\theta_2) f(\theta_3) f(\theta_4) d\theta_1 d\theta_2 d\theta_3 d\theta_4$$

$$= E\left(\exp\left\{-\sum_{j \in T} \exp(X_j\beta) \cdot (\theta_{m_j} + \theta_{p_j})\right\}\right) \tag{5}$$

For simplicity of exposition, let $\theta_k = u_k$ and let $-\exp(X_j\beta) = c_j$. Then equation (5) equals

$$E\left(\exp\left\{\sum_{k=1}^{4} u_k \sum_{j \in T} c_j \cdot a_{jk}\right\}\right) = \prod_{k=1}^{4} E\left(\exp\left\{u_k \sum_{j \in T} c_j \cdot a_{jk}\right\}\right) \tag{6}$$

Equation (6) is the product of four gamma distributed mgf's, and therefore (6)

$$= \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + \sum_{j \in T} \exp(X_j'\beta) \cdot a_{jk}}\right)^{\alpha/2} \tag{7}$$

The probability for all possible ordered set of subsets T can be written as a product of matrixes and scalars.

$$\pi^* = \prod_{k=1}^{4} \left(\frac{\lambda}{\lambda + \mathbf{B}\ \mathbf{diag}(X_j'\beta)\mathbf{a}}\right)^{\alpha/2} \tag{8}$$