

Robust tests of association for multilocus haplotypes in nuclear families

Frank Dudbridge
MRC Biostatistics Unit
Institute for Public Health
Robinson Way
Cambridge CB2 2SR UK
frank.dudbridge@mrc-bsu.cam.ac.uk

1. Introduction

Genetic epidemiology relies heavily on tests of association between genetic variants and disease outcomes or quantitative traits. Associations imply either that a genetic variant has a direct causal influence on the trait, or that the variant is a marker that is physically close on a chromosome to a causal variant. Despite the popularity of association tests, there is a wide and sometimes confusing variety of available methods, owing partly to the nature of genetic data. For example, different methods are available for discrete or quantitative traits, to deal with missing genotypic data, and to analyse family-based designs or samples of unrelated subjects. One consequence of this situation is that it is difficult to combine samples from different studies into a pooled analysis, without resorting to meta-analysis approaches. Furthermore, different methods often carry different assumptions about the data. It is desirable to have a general approach for studying genetic association that includes the widest range of study designs with the fewest number of assumptions.

Here I propose a general and flexible model for genetic association, motivated by the problem of uncertain multilocus haplotypes in nuclear families. Family-based designs, consisting of a proband, its two parents and some number of full siblings, are popular because they control for population stratification and provide well-matched controls (Cardon and Palmer 2003). A haplotype is a set of variants observed along the same chromosome, and is useful in association studies because it carries more information than a single locus, potentially giving greater power to detect association. However, haplotypes are usually not directly observed: it is more common to observe genotypic data at the individual loci, which gives the set of alleles carried by a subject but gives no information on how the alleles are jointly allocated to the two chromosomes.

Uncertain haplotypes are a form of missing data, and methods are available to deal with the uncertainty, either by some forms of weighted analysis (eg. Zaykin et al. 2002, Dudbridge 2003) or by using missing-data likelihood methods (eg. Clayton 1999, Epstein and Satten 2003). Here I focus on the missing-data likelihood approach in a generalised linear model, which offers a flexible method for testing a wide range of hypotheses together with a natural way to include additional covariates. Previous applications of these models to uncertain haplotypes have all required, at some point, some assumption of independence for the chromosomes within a family. Here I propose a model that allows testing of uncertain-haplotype and missing-genotype data under much more general conditions. The main result is a linear model based on additive effects of haplotypes that includes nuisance parameters for homozygosity in parents and offspring.

2. Methods

To approach the model for a full nuclear family, first consider the case of one single subject. Let Y denote a trait of interest, common examples being a 0/1 binary outcome or real-valued trait. Let \mathcal{H} denote the set of all haplotypes and $(h_1, h_2) \in \mathcal{H}^2$ denote a haplotype pair, or phased genotype. I shall use a retrospective likelihood which regards the genotype as a categorical response conditional on the trait. The genotype category has a multinomial distribution that is modelled by main effects of the two haplotypes, representing their population frequencies,

interactions between the haplotypes, representing the deviation of the genotype frequency from Hardy-Weinberg expectation, and interactions between the haplotypes and the trait value. A saturated model includes all the main effect and interaction terms, suggesting a generalised logistic likelihood of the form

$$\Pr((h_1, h_2) | Y) = \frac{\exp(\gamma_{h_1} + \gamma_{h_2} + \gamma_{h_1 h_2} + Y(\beta_{h_1} + \beta_{h_2} + \beta_{h_1 h_2}))}{\sum_{(h_1^*, h_2^*) \in \mathcal{H}^2} \exp(\gamma_{h_1^*} + \gamma_{h_2^*} + \gamma_{h_1^* h_2^*} + Y(\beta_{h_1^*} + \beta_{h_2^*} + \beta_{h_1^* h_2^*}))} \quad (1)$$

Association between haplotypes and trait is tested by inference on the ‘‘association’’ parameters β , with the ‘‘frequency’’ parameters γ as nuisance parameters, but the saturated model can have a very large number of parameters. A more parsimonious approach includes only the main haplotype effects and trait interactions, setting all $\gamma_{h_1 h_2} = 0$ and $\beta_{h_1 h_2} = 0$. This however gives a likelihood that is identical to that which would be obtained if chromosomes were regarded as independent sampling units, and thus imposes an assumption of Hardy-Weinberg equilibrium on the model. In order to avoid this assumption while maintaining a parsimonious model, I propose a compromise in which all $\beta_{h_1 h_2} = 0$, and $\gamma_{h_1 h_2} = 0$ when $h_1 \neq h_2$ (ie for heterozygous genotypes) and $\gamma_{h_1 h_2}$ is constrained to equal some common value γ_{HZ} when $h_1 = h_2$ (homozygous genotypes). The additional nuisance parameter γ_{HZ} characterises the excess homozygosity of the genotype distribution. The likelihood then has the form

$$\Pr((h_1, h_2) | Y) = \frac{\exp(\gamma_{h_1} + \gamma_{h_2} + Y(\beta_{h_1} + \beta_{h_2}) + \gamma_{HZ} I(h_1 = h_2))}{\sum_{(h_1^*, h_2^*) \in \mathcal{H}^2} \exp(\gamma_{h_1^*} + \gamma_{h_2^*} + Y(\beta_{h_1^*} + \beta_{h_2^*}) + \gamma_{HZ} I(h_1^* = h_2^*))} \quad (2)$$

Under this model, the sampling units are pairs of haplotypes, but the theory of linear models allows for the corresponding predictors to be correlated, provided that the residual distribution is normal. Thus, while this model is most accurate in Hardy-Weinberg populations, the assumption is not required for valid inference.

An advantage of the retrospective model (2) is that when haplotypes are uncertain, the likelihood is obtained simply by summing over the compatible haplotype solutions. In contrast, methods based on a prospective linear model require calculation of conditional probabilities for each of the possible solutions, leading to missing-data likelihoods that make explicit population assumptions (Clayton 1999, Schaid et al. 2002, Epstein and Satten 2003). Let G denote the unphased multilocus genotype of the subject, and let $\mathcal{S}(G)$ be the set of haplotype pairs compatible with G . Then the likelihood is

$$\Pr(G | Y) = \sum_{(h_1, h_2) \in \mathcal{S}(G)} \Pr((h_1, h_2) | Y) \quad (3)$$

To extend this construction to nuclear families, the categorical response will now be the ensemble of genotypes in the parents and offspring, with a multinomial distribution modelled by main effects and interactions of the parental haplotypes and interactions between the traits of the children and their haplotypes. Saturated models of this type have been described for family data (Kistner and Weinberg 2004), but require parameters for each possible mating type. Similarly to the above, I propose a more parsimonious model based on additive main effects of the four parental haplotypes, and main interaction terms between child haplotypes and traits. Without haplotype interaction terms, this is equivalent to assuming independent chromosomes in parents, random mating, random union of gametes, and independent transmissions to siblings given their trait values.

The latter is a well-known problem in association studies of genes showing prior evidence of linkage (Abecasis et al. 2000, Martin et al. 2003).

The model is made robust to these assumptions by including suitable interaction terms. Specifically, non-independence of parental haplotypes is controlled by an interaction parameter that has a common value γ_{HP} for homozygous parents, zero otherwise. Non-random union of gametes (and hence non-random mating) is controlled by an interaction between the two haplotypes transmitted to a child, and also between the two non-transmitted haplotypes. In both cases this interaction takes a common value γ_{HZ} for homozygous genotypes, zero otherwise. Non-independent transmissions to siblings are controlled by parameters that model the possible inheritance vectors in the family. In sib-pair data, the most common situation, this parameter takes a common value γ_{I1} when the siblings have one haplotype identical-by-descent (IBD), and another γ_{I2} when the siblings have both haplotypes IBD. Thus, in this case, there are four additional nuisance parameters in the model. For families with one child, the IBD parameters can be omitted leaving two nuisance parameters. Of course, any of these parameters can be omitted by making an appropriate assumption.

The likelihood follows the form of (2) and (3), with the additional parameters described. For a sib pair, let h_{FT1} denote the haplotype transmitted by the father to the first sib, h_{MU2} denote the haplotype not transmitted by the mother to the second sib, and so on. Let π be the proportion of haplotypes shared IBD by the two sibs (when this is uncertain, similar comments apply as to uncertain haplotypes). Let Y_i be the trait value in sib i . Then, assuming no recombination,

$$\begin{aligned} \Pr((h_{FT1}, h_{FU1}, h_{MT1}, h_{MU1}, h_{FT2}, h_{FU2}, h_{MT2}, h_{MU2}) | Y) \propto \\ \exp(\gamma_{h_{FT1}} + \gamma_{h_{FU1}} + \gamma_{h_{MT1}} + \gamma_{h_{MU1}} + Y_1(\beta_{h_{FT1}} + \beta_{h_{MT1}}) + Y_2(\beta_{h_{FT2}} + \beta_{h_{MT2}})) \\ + \gamma_{HZ}(I(h_{FT1} = h_{MT1}) + I(h_{FT2} = h_{MT2}) + I(h_{FU1} = h_{MU1}) + I(h_{FU2} = h_{MU2})) \\ + \gamma_{HP}(I(h_{FT1} = h_{FU1}) + I(h_{MT1} = h_{MU1})) + \gamma_{I1}I(\pi = \frac{1}{2}) + \gamma_{I2}I(\pi = 1) \end{aligned} \quad (4)$$

This model allows parsimonious inference on uncertain haplotypes in nuclear families, with no assumptions on the joint distribution of haplotypes beyond normally distributed residuals. Furthermore, by regarding unrelated subjects as single children with missing parents, case-control samples can be combined with family-based studies. Additional covariates including environmental risks and modifying disease loci can be naturally included by adding additional parameters to the linear model.

Haplotype association is tested by likelihood-ratio tests comparing the likelihood maximised over all parameters to the maximum likelihood under the constraint of all $\beta_h = 0$, for a global test, or a specific $\beta_h = 0$ for a test of one haplotype. More generally, any nested models can be compared to give a wide range of possible tests.

For continuous traits, this model is not necessarily robust to population stratification. The problem can be overcome by contrasting transmitted to untransmitted haplotypes within families. For example, in a family with one child the likelihood is given by

$$\begin{aligned} \Pr((h_{FT}, h_{FU}, h_{MT}, h_{MU}) | Y) \propto \exp(\gamma_{h_{FT}} + \gamma_{h_{FU}} + \gamma_{h_{MT}} + \gamma_{h_{MU}} + \frac{1}{2}Y(\beta_{h_{FT}} + \beta_{h_{MT}} - \beta_{h_{FU}} - \beta_{h_{MU}})) \\ + \gamma_{HP}(I(h_{FT} = h_{FU}) + I(h_{MT} = h_{MU})) + \gamma_{HZ}(I(h_{FT} = h_{MT}) + I(h_{FU} = h_{MU})) \end{aligned} \quad (5)$$

The factor $\frac{1}{2}$ ensures that, in the case of a diallelic marker, the haplotype association terms are equivalent to conditional logistic regression for a 0/1 binary outcome and gives a similar model to that of Abecasis et al. (2000) for continuous traits.

3. Conclusions

By including nuisance parameters for homozygosity in parents and offspring, regression models for haplotype association can be constructed without assuming independent chromosomes in parents or offspring. By modelling the genotypes as a multinomial response, the robustness is maintained in the presence of haplotype uncertainty and missing data. Genetic correlation among multiple offspring is controlled by nuisance parameters for the possible inheritance vectors. Furthermore, by conditioning on trait values, the model is robust to deviation from trait normality, as well as latent correlation among siblings. This model allows a wide range of association tests, including combination of case-control with family data, but with minimal assumptions about the genetic structure of the study population. The approach is generally applicable to any situation in which the observations are sampled in blocks within which the units of interest may be correlated.

REFERENCES

- Abecasis GR, Cardon LR, Cookson WO (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279-92.
- Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. *Lancet* 361:598-604.
- Clayton D (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 65:1170-7.
- Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. *Genet Epidemiol* 25:115-21.
- Epstein MP, Satten GA (2003) Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 73:1316-29.
- Kistner EO, Weinberg CR (2004) Method for using complete and incomplete trios to identify genes related to a quantitative trait. *Genet Epidemiol* 27:33-42.
- Martin ER, Bass MP, Hauser ER, Kaplan NL (2003) Accounting for linkage in family-based tests of association with missing parental genotypes. *Am J Hum Genet* 73:1016-26.
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70:425-34.
- Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered* 53:79-91.

RÉSUMÉ

Methods for testing genetic association are broadly divided into those for family-based studies or population samples, and for quantitative or discrete traits. When multilocus haplotypes are of interest, there is often an additional assumption of Hardy-Weinberg equilibrium (HWE). Here I describe a general and robust model for genetic association, motivated by the problem of uncertain haplotypes in nuclear families. A multinomial logistic regression is proposed in which the genotype is regarded as a categorical response, conditional on trait values. A parsimonious model includes additive main effects of haplotypes and first-order interactions with the trait, together with interaction terms that are constrained to be equal for homozygous genotypes and zero for heterozygous genotypes. The model is robust to violations of HWE in parents, random mating, random union of gametes, and independent transmissions to multiple siblings. For incomplete data such as uncertain haplotypes, the likelihood contribution is the sum of contributions of the possible solutions. By regarding unrelated subjects as the children of missing parents, this approach allows case-control samples to be combined with family-based studies. A within-family variation ensures robustness to population stratification. The approach is generally applicable to situations in which observations are sampled in blocks containing correlated units of interest.