Population-based Quantitative Trait Haplotype Analysis Using an EM algorithm to Resolve Phase.

A.P.Mander[1]

June 26, 2001

[1] MRC Biostatistics Unit, Institute of Public Health,
Forvie Site, Cambridge, UK.

email: adrian.mander@mrc-bsu.cam.ac.uk
phone: +44 (0)1223 330393
fax: +44 (0)1223 330388
website: http://www.mrc-bsu.cam.ac.uk/personal/adrian/welcome.shtml

Correspondence to A.P.Mander.

Abstract

An EM-algorithm is implemented to analyse the relationship between a normally distributed trait and a person's haplotype when phase is not known. Parameters are estimated by using two models, a log-linear regression model of haplotype frequencies and a linear regression model for the relationship between quantitative trait and haplotype. Simulation models are used to investigate the power of the method and to see the effects of the Hardy Weinberg Equilibrium assumption on power.

Keywords: Linear regression, missing data, phase resolution, haplotype analysis

1 Introduction

When a disease mutation occurs, it is in complete linkage disequilibrium (LD) with neighbouring loci. This LD is eroded over generations but can persist between tightly-linked loci for many generations. Association studies are used to detect trait loci by exploiting LD with adjacent markers and can be dichotomised into family-based and population-based studies. Family-based tests such as the Transmission Disequilibrium Test (Spielman *et al.* 1993) have become very popular, however family-based studies are not ideal for some situations: high-risk family data may not be available when there is a late-onset disease, or when the disease itself may interfere with reproduction. Population-based studies can be used in a more general setting although there is also the possibility that any significant association may be due to confounding by population stratification (Balding *et al.* 2001).

Population-based case-control association studies compare allele frequencies between cases and controls. In the absence of confounding significant differences indicate that different alleles carry different risks or that this locus is in LD with a causal locus. There are numerous diseases that are quantitative by nature (e.g. hypertension, diabetes, e.t.c.) and for these categorisation into two levels loses information. For quantitative association study a common approach is to investigate whether people carrying a particular allele have different mean traits.

Haplotype analysis is a means to detect chromosomal regions that either harbour, or that are in strong LD with, the disease-predisposing locus. With sufficient power, haplotype analysis should be able to isolate the locus of interest. Case-control haplotype analysis was described by (Chiano and Clayton 1998) and the algorithm described here extends this work to accommodate a quantitative outcome. The methods presented are most likely to apply to high-density SNP maps, even though there may be statistical

difficulties (Chapman and Wijsman 1998; Xiong and Jin 1999), for example high numbers of parameters; Intermediate models, ones that are not saturated, may alleviate the problems by having fewer parameters.

When phase is known, association analysis can be carried out using regression models in any standard statistical package. However, the basic regression-based approach can not be used directly for diploid genotypes when phase is unknown; Explanatory variables can not be constructed as the "true haplotypes" are not observed. Unknown phase can be defined in terms of a missing data problem and analysed accordingly either using Bayesian methods (Ayres and Balding 2001) or the likelihood-based EM algorithm approach (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Chiano and Clayton 1998; Mander 2000). Unlike Bayesian methods, likelihood-based approaches allow hypothesis testing and do not rely on prior distributions; testing rather than estimation is the focus of this paper.

Simulated data are used to demonstrate the power of the algorithm for testing association to a quantitative trait. Various levels of deviation from Hardy-Weinberg Equilibrium are investigated in order to determine the impact on power. It can be seen that excess heterozygosity leads to lower power.

2 Methods

The following section outlines a new implementation of the EM algorithm (Dempster *et al.* 1977) for detecting association to a quantitative trait. When the trait is normally distributed, and phase is known, linear regression may be used to test for association. The method can be extended to other types of statistical models but is beyond the scope of this paper.

The next sections illustrate the possible parameters in the regression linear predictors

when the phase is known. The same models are used in the EM algorithm for resolving phase.

For illustrative purposes two diallelic loci are used to show the possible parameters; the number of parameters are minimised with diallelic loci. Let the alleles at the first locus be a and A and at the second locus, b and B. The two-locus double homozygote genotype ab/ab is the reference genotype. The parameters of the linear predictor will therefore represent mean differences, derived from comparing a particular genotype to the reference genotype. Let the mean trait value for people with the genotype ab/ab be μ.

2.1 Singlepoint Additive Model (SAM)

The first stage of localising the region of association is a singlepoint additive analysis; this tests each individual marker and the strength of its association with the outcome. Additivity, in the context of haplotype analysis, is defined by any model that does not include between-chromosome interactions, in the statistical sense. A fuller explanation of interaction in quantitative models is given by (Cordell *et al.* 2001).

The SAM for the two diallelic loci system has two additive effects: the additive parameter α, for allele A; and the additive parameter β, for allele B. People with genotypes of the form ay/Ay and Ay/Ay, where y represents either the b or B allele, have mean traits $\mu + \alpha$ and $\mu + 2\alpha$, respectively. Similarly, people with genotypes of the form xb/xB and xB/xB, where x represents either the a or A allele, have mean traits $\mu + \beta$ and $\mu + 2\beta$, respectively. In this paper, parental imprinting is ignored; In other words the ordering of alleles in genotypes is not considered. The linear predictor can be extended to include parental imprinting by including two additional parameters, these represent the additional effect of inheriting the A or B allele from the mother.

Note that the SAM is better than just analysing each locus individually. The additive effect α is estimated adjusted for the additive effect, β, at locus 2. This counteracts the difficulty that if locus 1 is analysed without considering locus 2, then α may reflect the effect at locus 2 rather than any effect at locus 1.

There are two hypotheses of interest in the 2-loci SAM: is $\alpha = 0$, or is $\beta = 0$? A significant result indicates position but not inheritance, for example, if the alleles act in a dominant way.

2.1.1 Multipoint Additive Model (MAM)

The multipoint additive model is one that contains all within-chromosome between-loci interactions and the parameters have an additive dose response. In the two diallelic loci example, there is only one additional additive effect, Δ. The Δ parameter is the additive effect of the AB haplotype, and is added to the linear predictor of the 2-loci SAM. People with genotypes ab/AB and AB/AB will have mean trait values, under the 2-loci MAM, $\mu + \alpha + \beta + \Delta$ and $\mu + 2\alpha + 2\beta + 2\Delta$, respectively.

For the 2-loci MAM, two hypotheses are tested in order to isolate the locus that has the strongest association with the outcome of interest. The test $\beta = \Delta = 0$ determines whether locus 1 is associated, and similarly for locus 2, the test is $\alpha = \Delta = 0$. The model can be extended to three diallelic loci with the inclusion of all pair-wise within-chromosome interactions and one three-way interaction term. The model generalises to more loci, although the addition of each locus causes the number of extra parameters to increase exponentially.

For larger numbers of loci, the higher-order interaction terms are unlikely to be needed to describe the data and more parsimonious models could be used. One such model is a "Markov pattern" model (Balding *et al.* 2001), that includes all pairwise interactions of

adjacent loci but not higher order interactions.

2.2 Deviations from Additivity

The linear predictor for non-additive models includes between-chromosome interaction parameters. There is one between-chromosome within-loci interaction parameter, namely dominance, for each diallelic locus.

Returning to the two diallelic loci example, the dominance parameter for locus 1 is δ_α, and for locus 2, it is δ_β. In the 2-loci SAM, with the additional δ-parameters, people with genotypes ay/Ay and Ay/Ay (y is either allele b or B) have mean traits of $\mu+\alpha+\delta_\alpha$ and $\mu+2\alpha$, respectively. Similarly for locus 2, people with genotypes xb/xB and xB/xB (x is allele a or A) have mean traits $\mu+\beta+\delta_\beta$ and $\mu+2\beta$, respectively.

The dominance parameter has the following interpretations: when $\delta_\alpha = -\alpha$, allele a is dominant to A; when $\delta_\alpha = +\alpha$, allele a is recessive to A. The δ_β has the same interpretation in reference to the b and B alleles.

2.2.1 The full saturated model

Ignoring the effects of parental imprinting, there are 10 unique unordered genotypes, and the saturated model allows a different mean quantitative trait value for each genotype. Thus, the model contains 10 parameters as shown in table i.

There are four between-chromosome between-loci interactions. $\delta\Delta_\alpha$ and $\delta\Delta_\beta$ represent the interactions between the dominance parameter at one locus and the additive effect at the other, $\delta_{\alpha\beta}$ gives the interaction between the two dominance parameters, and $\Delta_{\alpha\beta}$ is the effect of having the haplotype of interest over both chromosomes, this is seen clearly for the Ab/aB genotype where the AB haplotype is "split" over the two chromosomes.

The saturated model can be used to test the hypothesis of epistasis if two tightly linked loci are both functional. Epistasis here is any between-loci interaction (Fisher 1918), and hence the test is of $\Delta_{\alpha\beta} = \delta_{\alpha\beta} = \delta\Delta_\alpha = \delta\Delta_\beta = \Delta = 0$. The second use of the saturated model is the general test for non-additivity: this compares the MAM to the saturated model. In other words, it tests $\delta\Delta_\alpha = \delta\Delta_\beta = \delta_\beta = \delta_\alpha = \Delta_{\alpha\beta} = \delta_{\alpha\beta} = 0$.

2.3 Resolving Phase

Often phase is not known and only unordered genotypic data is available. Using the two diallelic-loci example, the genotype for the double heterozygote may be represented as (a, A, b, B). Henceforth, this shall be referred to as the phenotype. For this phenotype the genotype is not uniquely defined. The linear predictor in the linear regression **is** uniquely defined for the SAM but **not** for the MAM.

For subject i, the probability density function of its normally distributed quantitative trait value, y_i, can be specified when the genotype, g_i, is known and the linear predictor, η_i, is known.

$$f_{g_i}(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(y_i - \eta_i)^2}{2\sigma^2} \tag{1}$$

In reality, when phase is unknown, the distribution of quantitative trait will follow a mixture distribution.

$$\sum_{\tilde{g}_i \in \tilde{G}_i} \pi_{\tilde{g}_i} f_{\tilde{g}_i}(y_i)$$

where $\tilde{G}_i$ is the set of possible phases and $\pi_{\tilde{g}_i}$ is the probability of genotype $\tilde{g}_i$. The genotype probabilities are estimated from the dataset under the assumption of Hardy-Weinberg Equilibrium (HWE) or, alternatively the probabilities can be fixed using ex-

ternal data sources. The latter is not explored here.

2.3.1 Set of Phases, $\tilde{G}_i$

The number of phase possibilities in the set $\tilde{G}_i$ depends on the linear model of interest.

The phenotype (a, b, A, B) has one phase for the 2-loci SAM and two phases for the 2-loci MAM ab/AB and aB/Ab. If the linear model included parental imprinting, there would be four phases to consider: AB/ab, Ab/aB, ab/AB and aB/Ab. If the four phases were taken as $\tilde{G}_i$ for each of the three models, the algorithm would converge to the same parameter estimates. This would introduce computational inefficiency and change the overall log-likelihood. It would not, however, alter the likelihood ratio test statistics for comparing two models, as long as the two models follow the same rules of phase resolution.

The set of phases can be enumerated when the available phenotype contains missing alleles, for example $(a, A, ., B)$, where "." represents a missing allele. Under the assumption of missing completely at random (MCAR) (Little and Rubin 1987), the two alleles b or B are equally likely to be the true allele, and so there are two possible phenotypes, (a, A, b, B) and (a, A, B, B). The first phenotype has two phases, and the second has only one, assuming that the 2-loci MAM is fitted. The three phases are ab/AB, aB/Ab and aB/AB.

For real data, all the alleles that could occur at a locus may not be fully known; the algorithm assumes that only the observed alleles are possible. If an allele is not present in the dataset then the MCAR assumption is violated, and this is likely to occur when a dataset is small or an allele is rare.

If the phase is known for a subset of the subjects, for example a control sample, then the set of phases is constructed for only those subjects with unknown phase.

2.4 The EM algorithm

The following section describes the steps of the j-th iteration of the EM algorithm.

The 'E-step' of the algorithm is the estimation of the posterior probability of each phase. The j-th estimate of the posterior probability of a particular phase, g_i^*, for subject i is $\hat{z}_{g_i^*}^{(j)}$. The formula for this, where $\hat{f_{g_i^*}}^{(j-1)}(y_i)$ (equation 1) and $\hat{\pi}_{g_i^*}^{(j-1)}$ are the $(j-1)$-th estimates of the normal p.d.f. and genotype probabilities, is given below,

$$\begin{aligned}
\hat{z}_{g_i^*}^{(j)} &= E(z_{g_i^*}|y_i, \hat{\eta}_i^{(j-1)}), \\
&= \frac{\hat{\pi}_{g_i^*}^{(j-1)} \hat{f_{g_i^*}}^{(j-1)}(y_i)}{\sum_{\tilde{g}_i \in \tilde{G}_i} \hat{\pi}_{\tilde{g}_i}^{(j-1)} \hat{f_{\tilde{g}_i}}^{(j-1)}(y_i)}.
\end{aligned}$$

Given these posterior probabilities, the full data likelihood, for n subjects, is given below,

$$\begin{aligned}
L(\underline{\pi}, \eta, \sigma|y_i, \hat{z}_{g_i^*}^{(j)}) &= \prod_{i=1}^{n} \prod_{\tilde{g}_i \in \tilde{G}_i} \{\pi_{\tilde{g}_i} f_{\tilde{g}_i}(y_i)\}^{\hat{z}_{\tilde{g}_i}^{(j)}}, \\
&= \prod_{i=1}^{n} \prod_{\tilde{g}_i \in \tilde{G}_i} \{\pi_{\tilde{g}_i}\}^{\hat{z}_{\tilde{g}_i}^{(j)}} \{f_{\tilde{g}_i}(y_i)\}^{\hat{z}_{\tilde{g}_i}^{(j)}}, \\
&= \prod_{i=1}^{n} \prod_{\tilde{g}_i \in \tilde{G}_i} \left\{\pi_{\tilde{h}_{1i}} \pi_{\tilde{h}_{2i}}\right\}^{\hat{z}_{\tilde{g}_i}^{(j)}} \{f_{\tilde{g}_i}(y_i)\}^{\hat{z}_{\tilde{g}_i}^{(j)}}.
\end{aligned}$$

In this way, the likelihood can be factorised into terms involving the genotype probabilities and the normal probability distribution function.

The 'M-step' of the algorithm is the estimation of the genotype probabilities and parameters of the normal distribution. When phase is unknown, there is no information to estimate the genotype probabilities directly and are replaced by the product of two haplotype probabilities using the HWE assumption, $\pi_{\tilde{g}_i} = \pi_{\tilde{h}_{1i}} \pi_{\tilde{h}_{2i}}$. It is possible to discard the HWE assumption by using known disequilibrium coefficients, but this is not implemented here.

The j-th estimate of the haplotype probabilities are estimated by a saturated log-linear model using the iterative proportional fitting algorithm (Agresti 1992) with the $\hat{z}^{(j)}$'s as weights. The log-linear model allows the investigation of intermediate models (Chiano and Clayton 1998; Mander 2000), in which the j-th estimates of the parameters in the linear predictor are estimated using linear regression with the $\hat{z}^{(j)}$'s as weights. This can be done in any standard statistical package, however the j-th estimate of the residual variance, $\hat{\sigma^2}^{(j)}$ can vary between packages. The residual variance estimator used here is a weighted average of the residuals (Ghosh and Majumder 1999),

$$\hat{\sigma^2}^{(j)} = \frac{\sum_i \sum_{g_i^* \in \tilde{G}_i} \hat{z}_{g_i^*}^{(j)} \left(y_i - \hat{\eta}_i^{(j)}\right)^2}{\sum_i \sum_{g_i^* \in \tilde{G}_i} \hat{z}_{g_i^*}^{(j)}}.$$

The first iteration of the EM algorithm involves the selection of the initial values for all the parameters to be estimated. By default, each haplotype is equally likely, the residual variance is 1 and the regression coefficients are 0. Then, the posterior probabilities are calculated using the data and these parameter estimates. The EM algorithm proceeds iteratively, until the change in the full log-likelihood is small (usually around 0.00001).

3 Results

Simulated data was used to demonstrate the power to detect association and to isolate regions of genetic association.

Phase-known genotypes were generated for three diallelic loci, then a normally distributed quantitative trait was generated conditional on the genotype. For the empirical power calculations, the phase information was ignored. As phase resolution is based on the assumption of HWE, power was also investigated allowing deviation from HWE.

3.1 Genotype Generation

The three simulated loci contained the alleles 1 or 2, and were generated with equal probability. This maximised the frequency of unknown phase and meant that the power results were based on a worst case scenario. The first haplotype for each subject was generated assuming that all the possible haplotypes were equally frequent. The second haplotype was constructed by simulating in the absence of HWE. For a given disequilibrium coefficient D, as defined by (Hernandez and Weir 1989), the second allele was generated with probability of heterozygosity of $(1/2 - 2D)$, assuming equally frequent alleles. The same disequilibrium coefficient was used for each locus and the expected number of homozygotes at each locus was the same. When D is 0, the alleles are expected to be in HWE, whereas an excess of heterozygotes occurs if $D < 0$ and an excess of homozygotes if $D > 0$. Power is expected to decrease with increasing levels of heterozygosity because phase is known for homozygotes.

3.2 Quantitative Trait Generation

For each person, a quantitative trait value was sampled from a $N(0, 1)$ distribution. This person's trait value was shifted by an amount, S, based on the presence of the 222 haplotype (model 1), or when either haplotypes 221 or 222 were present (model 2). An additive model was assumed and for model 1, a person with the genotype 222/222 had an expected trait of $2S$. Similarly, for model 2 a person with either genotype 221/221, 221/222 or 222/222 had an expected trait of $2S$.

3.3 Power

For the two models, 400 subjects were simulated. The genotypes were generated with D chosen from the set $\{-0.2, -0.1, 0, 0.1, 0.2\}$ and the quantitative trait was generated with

S chosen from the set $\{0, 0.25, 0.5, 0.75, 1\}$. For each combination of parameters 1000 simulated datasets were created and empirical power was calculated with the significance level 5%.

The test of interest for model 1 was detection of an association between the 3 loci and the quantitative trait. This test was a 7 df likelihood ratio test, comparing the 3-loci MAM and the one parameter (μ) constant model. The empirical power is shown in table ii and this is the proportion of the 1000 tests that showed a significant association.

It can be seen that there is over 90% power to detect a mean shift of 0.75 or more. This is smaller than the residual variance, which is 1. Also, there is slightly more power when there is an excess of homozygotes ($D > 0$) due to less phase ambiguity. Power decreases rapidly when there is an excess of heterozygotes.

For model 2, the association is present in the first two loci and it is clear that the third locus is not involved in the association. To test this hypothesis the 3-loci MAM is compared to the 2-loci MAM, a 4 df test. The 2-loci MAM has no parameters for the third locus, although 3-loci haplotype frequencies were estimated. It is possible to estimate 2-loci haplotype frequencies by collapsing the third locus for the 2-loci MAM. This would increase the degrees of freedom of the test, and would provide a joint test of LD between the third locus and the first two loci, as well as the quantitative trait genotype relationship.

The proportion of significant likelihood ratio test statistics are given in table iii.

A non-significant result means that the 2-loci MAM and the 3-loci MAM have similar likelihoods and so the third locus is not involved in the association. For most of the cells, about 95% of the tests correctly fail to indicate that the third locus is involved in the association, demonstrating the accuracy of the method in this framework.

With the same datasets, the 2-loci MAM is compared to the 1-locus SAM. Table iv

displays the proportion of tests that gave a significant result. Here, a significant result meant that the second locus could not be dropped from the model and when $S > 0.5$, over 90% of tests indicated that the second locus was involved in the association.

The results are very similar to those shown in table ii, except that power is greater, as there is now only 2 degrees of freedom. Again, when there is an excess of heterozygotes, the power decreases.

4 Discussion

Clearly, the EM-algorithm is able to detect association when phase is unknown in the case of a fairly small dataset, and small shifts in the mean quantitative trait for one particular haplotype. The power to detect association however depends on the extent of deviation from Hardy-Weinberg Equilibrium; this is the effect of changing phase uncertainty. When allele are not equally frequent the amount of phase uncertainty is much less and the power will be greater than the simulations suggest.

A by-product of having two models in the algorithm is that the same algorithm can be used to test for linkage disequilibrium between loci, conditional on the quantitative trait model. It is also possible to test for association in a case/control setting, conditional on the quantitative outcome. This joint modelling should increase the power to detect LD.

The most useful application of the methods described will be in analysing SNP genotypes within a candidate gene that is associated with some continuous outcome. Given the huge number of SNPs available and the number of parameters needed to model the relationship between haplotype and quantitative trait the future use of these methods depend on parsimonious SNP selection, or possibly on the generation of further hypotheses that incorporate fewer parameters.

The algorithm is implemented in the statistical package STATA (StataCorp. 1999) and is available within STATA by typing the commands

```
net from http://www.mrc-bsu.cam.ac.uk/personal/adrian
net install qhapipf
```

or by downloading from the web page `http://www.mrc-bsu.cam.ac.uk/personal/adrian/stata.shtml` .

The method can be easily programmed in any statistical package that allows weights in regression commands. The EM algorithm can be modified, relatively easily, to handle non-normal data and the command to estimate the parameters of the linear predictor allows weights. The linear regression could be replaced by a full multi-level model to handle family-based data and perform analyses described in (Cardon 2000; Burton *et al.* 1999).

5 Acknowledgements

I would like to thank Aruna Bansal and Peter Holmans for their help in preparing the manuscript.

References

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical methods in medical research*, **1**, 201–18.

Ayres, K. and Balding, D. (2001). Measuring gametic disequilibrium from multi-locus data. *Genetics*, **157**, 413–23.

Balding, D., Bishop, M., and Cannings, C. (ed.) (2001). *Handbook of statistical genetics.* John Wiley & sons,l.t.d. (chichester).

Burton, P., Tiller, K., Gurrin, L., William, O., Cookson, A., Musk, W., and Palmer, L. (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (glmms) and gibbs sampling. *Genetic Epidemiology*, **17**, 118–40.

Cardon, L. (2000). A sib-pair regression model of linkage disequilibrium for quantitative traits. *Hum. Hered.*, **50**, 350–8.

Chapman, N. and Wijsman, E. (1998). Genome screens using linkage disequilibrium tests: Optimal marker characteristics and feasibility. *Am. J. Hum. Genet.*, **63**, 1872–85.

Chiano, M. and Clayton, D. (1998). Fine genetic mapping using haplotype analysis and the missing data problem. *Ann. Hum. Genet.*, **62**, 55–60.

Cordell, H., Todd, J., Hill, N., Lord, C., Lyons, P., Peterson, L., Wicker, L., and Clayton, D. (2001). Statistical modelling of inter-locus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics*, **158**, 357–367.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *J.R.S.S. B*, **39**, 1–22.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–7.

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Trans. Roy. Soc. Edinb.*, **52**, 399–433.

Ghosh, S. and Majumder, P. (1999). mapping a quantitative trait locus via the em algorithm and bayesian classification. *Genetic Epidemiology*, **19**, 97–126.

Hawley, M. and Kidd, K. (1995). Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.*, **86**, 409–11.

Hernandez, J. and Weir, B. (1989). A disequilibrium coefficient approach to hardy-weinberg testing. *Biometrics*, **45**, 53–70.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data.* J. Wiley & Sons.

Mander, A. (2000). Haplotype frequency estimation using an em algorithm and log-linear modelling. *Stata Technical Bulletin*, **57**, 5–7.

Spielman, R., McGinnnis, R., and Ewens, W. (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus(iddm). *Am. J. Hum. Genet.*, **52**, 506–16.

StataCorp. (1999). *Stata statistical software: Release 6.0.* Stata Corporation, College Station, TX:Stata Corporation.

Xiong, M. and Jin, L. (1999). Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. *Am. J. Hum. Genet.*, **64**, 629–40.

Table i: The expected trait values for the 10 possible genotypes for two diallelic loci under the saturated model ignoring parental imprinting

Genotype	Additive parameters	Dominance parameters	Between-loci between-chromosome interactions
ab/ab	μ		
ab/Ab	$\mu+\alpha$	$+\delta_\alpha$	
ab/aB	$\mu+\beta$	$+\delta_\beta$	
ab/AB	$\mu+\beta+\alpha+\Delta$	$+\delta_\alpha+\delta_\beta$	$+\delta\Delta_\alpha+\delta\Delta_\beta+\delta_{\alpha\beta}$
Ab/Ab	$\mu+2\alpha$		
Ab/aB	$\mu+\beta+\alpha$	$+\delta_\alpha+\delta_\beta$	$+\delta\Delta_\alpha+\delta\Delta_\beta+\delta_{\alpha\beta}+\Delta_{\alpha\beta}$
Ab/AB	$\mu+\beta+2\alpha+\Delta$	$+\delta_\beta$	$+2\delta\Delta_\beta+\Delta_{\alpha\beta}$
aB/aB	$\mu+2\beta$		
aB/AB	$\mu+2\beta+\alpha+\Delta$	$+\delta_\alpha$	$+2\delta\Delta_\alpha+\Delta_{\alpha\beta}$
AB/AB	$\mu+2\beta+2\alpha+2\Delta$		$+2\Delta_{\alpha\beta}$

Table ii: The proportion of simulations with a significant association test statistic when comparing the 3-loci MAM to the constant model

	S				
D	0	0.25	0.5	0.75	1
-0.2	0.059	0.080	0.134	0.232	0.400
-0.15	0.065	0.111	0.275	0.498	0.739
-0.1	0.104	0.164	0.431	0.822	0.961
-0.05	0.074	0.181	0.549	0.900	0.998
0.0	0.068	0.175	0.604	0.949	0.999
0.05	0.062	0.193	0.682	0.976	1.000
0.1	0.059	0.215	0.682	0.976	0.999
0.15	0.053	0.212	0.717	0.977	1.000
0.2	0.055	0.180	0.679	0.981	1.000

Table iii: The proportion of simulations with a significant test statistic when comparing the 3-loci MAM to 2-loci MAM

	S				
D	0	0.25	0.5	0.75	1
-0.2	0.058	0.058	0.045	0.044	0.042
-0.15	0.080	0.060	0.045	0.065	0.035
-0.1	0.098	0.084	0.109	0.098	0.094
-0.05	0.077	0.060	0.065	0.082	0.085
0.0	0.063	0.067	0.059	0.053	0.050
0.05	0.062	0.050	0.062	0.037	0.052
0.1	0.047	0.052	0.051	0.053	0.055
0.15	0.043	0.045	0.072	0.043	0.052
0.2	0.045	0.053	0.059	0.048	0.050

Table iv: The proportion of simulations with a significant test statistic when comparing the 2-loci MAM to 1-locus SAM

	S				
D	0	0.25	0.5	0.75	1
-0.2	0.051	0.107	0.241	0.501	0.744
-0.15	0.063	0.139	0.496	0.830	0.970
-0.1	0.066	0.261	0.747	0.978	0.999
-0.05	0.052	0.415	0.928	0.997	1.000
0.0	0.051	0.497	0.983	1.000	1.000
0.05	0.056	0.624	0.992	1.000	1.000
0.1	0.057	0.648	0.997	1.000	1.000
0.15	0.059	0.770	1.000	1.000	1.000
0.2	0.059	0.846	1.000	1.000	1.000