Genet. Sel. Evol. 38 (2006) 231–252

DOI: 10.1051/gse:2006001

Original article

Prediction of IBD based on population history for fine gene mapping

Jules HERNÁNDEZ-SÁNCHEZ*, Chris S. HALEY, John A. WOOLLIAMS

Roslin Institute, Midlothian, EH25 9PS, Scotland, UK

(Received 31 March 2005; accepted 19 December 2005)

Abstract – A novel multiple regression method (RM) is developed to predict identity-by-descent probabilities at a locus L (IBD_L), among individuals without pedigree, given information on surrounding markers and population history. These IBD_L probabilities are a function of the increase in linkage disequilibrium (LD) generated by drift in a homogeneous population over generations. Three parameters are sufficient to describe population history: effective population size (*Ne*), number of generations since foundation (T), and marker allele frequencies among founders (p). IBD_L are used in a simulation study to map a quantitative trait locus (QTL) *via* variance component estimation. RM is compared to a coalescent method (CM) in terms of power and robustness of QTL detection. Differences between RM and CM are small but significant. For example, RM is more powerful than CM in dioecious populations, but not in monoecious populations. Moreover, RM is more robust than CM when marker phases are unknown or when there is complete LD among founders or *Ne* is wrong, and less robust when p is wrong. CM utilises all marker haplotype information, whereas RM utilises information contained in each individual marker and all possible marker pairs but not in higher order interactions. RM consists of a family of models encompassing four different population structures, and two ways of using marker information, which contrasts with the single model that must cater for all possible evolutionary scenarios in CM.

QTL fine mapping/ identity-by-descent

1. INTRODUCTION

The concept of genetic relationship plays a key role in many areas of genetic research. Genetic relationships have been traditionally estimated from pedigrees, and also recently from marker information. Thus, if a pedigree is not available, relationships can still be inferred from marker data alone. Some methods use unlinked markers to calculate the probability of two individuals being, for example, full-sibs or parents and offspring [10, 20]. However, since

* Corresponding author: Jules.Hernandez@bbsrc.ac.uk

not all genetic relationships can be so clearly defined in a fixed, and usually small number of categories, it seems more practical to calculate a continuous quantity defined in terms of identity-by-descent (IBD) probabilities. These IBD probabilities are extensions of classical coefficients of kinship [11]. However, most of the available approaches (*e.g.* [9,10,20–22,24]) do not jointly use population history information and multiple linked markers to predict genetic relatedness, despite the fact that related individuals must share common ancestors, and therefore have a common history, and that multiple linked markers contain more information about IBD status at a particular chromosomal location L (IBD_L) than multiple unlinked ones.

Meuwissen and Goddard [13, 14] proposed a new estimator of relationship based on multiple linked markers and population history information. This method capitalises on the expected built-up of linkage disequilibrium (LD) over generations, accounting for the correlations between loci due to both linkage and genetic drift. However, they considered only the simplest population history, *i.e.* one in which initial allele frequencies (p) changed due to drift, in a monoecious population of constant effective size (Ne), over T discrete generations of random mating and selfing. Under this model, IBD_L probabilities are expected to increase with decreasing Ne, because parental alleles will be sampled from smaller pools, and with increasing T, because more alleles will have been lost by drift in the population. Meuwissen *et al*. [15], Meuwissen and Goddard [14] analysed a cattle pedigree using an IBD-based variance component analysis, and achieved a resolution of <1 centi-Morgan (cM) in mapping a QTL for twinning rate, and of ~0.04 cM for a QTL affecting milk traits. These examples highlight the potential benefits in terms of power and resolution of combining information on linkage and LD. Our prime interest is also to enable fine mapping of QTL using variance component estimation (*e.g.* [3]), without assuming unrelated and non-inbred pedigree founders (*e.g.* as in linkage analysis).

However, at the heart of the idea of inferring IBD_L from homozygosity, or identity-by-state (IBS), at linked marker loci lies the concept of expected IBD between k loci (θ_k). These parameters were not used in Meuwissen and Goddard's work, instead, the coalescent theory was applied [7, 8], and henceforth their method will be called the coalescent (-based) method (CM). Our method uses a regression approach, henceforth called regression (-based) method (RM), to predict IBD_L as a function of θ_k, p and IBS at markers (x).

Other differences between CM and RM are the following: (1) CM utilises all haplotype information and RM utilises information contained in single markers and marker pairs but not in higher order interactions, although it is

possible to accommodate these interactions; (2) CM infers IBD "en bloc" whereas RM infers loci IBD without interpolating IBD predictions between loci; (3) CM models a monoecious population with selfing (MS) whereas RM models that plus a monoecious population excluding selfing (ME), and dioecious populations with and without hierarchical mating (DH and D, respectively); (4) CM does not distinguish between inbreeding and coancestry whereas RM does; and (5) CM consists of a single model whereas RM consists of a family of models that can use genotype or (partial) haplotype information.

A weakness of both RM and CM methods is the assumption that population parameters Ne, T, p, and marker haplotypes are known without error. In practice, this information may be partially or totally unknown. Therefore, we use computer simulations to compare the robustness as well as the power of both methods in detecting QTL.

2. MATERIALS AND METHODS

2.1. General assumptions

Our objective was to calculate IBD_L probabilities between pairs of alleles sampled at a prospective QTL location L, using historical information and marker data, without pedigree information.

Under the same population model used by Meuwissen and Goddard, two completely homozygous (IBS) haplotypes such as [111L111] and [111L111] will have, on average, higher IBD_L than two completely heterozygous haplotypes such as [111L111] and [222L222] (the numbers denote observed marker alleles). The strength of this statement depends on recombination patterns, *i.e.* IBS information has greater weight with tight linkage than with loose linkage, and on population history, *i.e.* high IBD_L probabilities are more likely as Ne decreases, or as T increases. The value of IBS information to predict IBD also depends on initial levels of IBS, *i.e.* depends on marker allele frequencies p at $t = 0$, because an extreme p leads to high levels of homozygosity among random haplotypes.

We will develop a method (RM) that takes all this information into account, and accurately models four different mating schemes (populations). In MS populations, individuals within a generation are hermaphrodites and can mate with themselves. In ME populations, individuals are hermaphrodite but cannot mate with themselves. In D populations, individuals are single sex and can only mate with individuals of the opposite sex. In DH populations, every male is mated to a fixed number of females each generation.

In general, two marker haplotypes must be compared to predict IBD_L, so it is more useful to use the notation $IBD_{iu,jv}$, which stands for the probability that an allele at locus L carried by individual i on haplotype u is IBD with an allele at the same locus carried by individual j on haplotype v.

2.2. Constructing an RM model to predict inbreeding

2.2.1. A single marker model

The regression model to predict inbreeding of individual i at locus L ($IBD_{i1,i2}$) is

$$IBD_{i1,i2} = \theta_1 + bX_i \tag{1}$$

where 1 and 2 refer to the two different alleles at locus L ($IBD_{i1,i1} = 1$), θ_1 (usually denoted as $\bar{F}$) is the mean population inbreeding at one locus, b is the regression coefficient relating observed IBS at the marker with expected $IBD_{i1,i2}$, and $X_i = x_i - \bar{x}$ is the adjusted IBS at the marker. Moreover, x_i is an observed variable with expectation $\bar{x}$ (Eq. (A.1)), taking value 1 if individual i is homozygous at the marker, or 0 otherwise.

The regression coefficient is $b = \frac{\sigma(I_{IBD}, I_{IBS})}{\sigma^2(I_{IBS})}$, where I_{IBD} is an indicator variable taking value 1 if locus L is IBD, or 0 otherwise, and I_{IBS} is another indicator variable taking value 1 if a marker is IBS, or 0 otherwise. The variance $\sigma^2(I_{IBS})$ (Eq. (A.3)), and covariance $\sigma(I_{IBD}, I_{IBS})$ (Eq. (A.12)) are functions of parameters p and $\theta_{k=1...4}$, *i.e.* $b \propto (\theta_{k=1...4}, p)$. Parameter θ_k is the probability of sampling an individual inbred at k loci. θ_k is a function of T, Ne, recombination rates c between all loci (markers and locus L), and type of population, *i.e.* $\theta_k \propto (T, Ne, c, population)$, thus it contains information about historical recombinations (*i.e.* about linkage disequilibrium). Note that θ_k is different from $IBD_{iu,jv}$, the former is a population expectation independent of marker data (x_i), the latter is a variable which is a function of θ_k, p and x_i, *i.e.* $IBD_{iu,jv} \propto (\theta_{k=1...4}, p, x_i) \propto (T, Ne, c, population, p, x_i)$. Cockerham, Weir and co-authors [2,25–27] developed exact formulae for θ_2. We extended this theory to predict approximations for θ_3 and θ_4 in all populations [6], using a model similar to (1) but regressing IBD at neighbouring loci on IBD at a central locus. The average inbreeding at one locus θ_1 is equal to θ_2 given $c = 0$.

In the absence of marker data $IBD_{i1,i2} = \theta_1$, and all individuals would have the expected population inbreeding. Another situation in which different individuals will have the same predicted inbreeding at locus L is when they have identical marker data.

Table I. The vector of weights is $\mathbf{R} = \mathbf{G}\mathbf{V}^{-1}$. In RM_h, which includes markers interactions, **V** is the (co)variance of IBS among markers and marker pairs, and **G** the covariance of IBS between markers or marker pairs, and IBD at locus L. The elements of **V** and **G** for two markers A and B are given below. It was assumed that $\Pi_A = \Pi_B = \Pi$, and $\theta_A = \theta_B = \theta$. Moreover, x_i and y_i denote IBS and IBD, respectively, at locus i, and $\eta_{ij} = \theta_{ij} - \theta^2$.

V	x_A	x_B	x_{AB}
x_A	$(1-\theta)(1-\Pi)(\theta+(1-\theta)\Pi)$	$\eta_{AB}(1-\Pi)^2$	$\bar{x}_{AB}(1-\theta)(1-\Pi)$
x_B		$(1-\theta)(1-\Pi)(\theta+(1-\theta)\Pi)$	$\bar{x}_{AB}(1-\theta)(1-\Pi)$
x_{AB}			$\bar{x}_{AB}(1-\bar{x}_{AB})$

G	x_A	x_B	x_{AB}
y_L	$\eta_{AL}(1-\Pi)$	$\eta_{BL}(1-\Pi)$	$\overline{x_{AB}y_L} - \bar{x}_{AB}\bar{y}_L$

2.2.2. A multimarker model

Model (1) can be naturally extended to use M markers (vectors in bold)

$$IBD_{iu,iv} = \theta_1 + \mathbf{R}'\mathbf{X} \tag{2}$$

where **R** is a vector of weights, or partial regression coefficients, relating IBS at all single markers and marker pairs with $IBD_{iu,iv}$, and **X** is the vector of adjusted IBS observations at all markers and marker pairs. So, for M markers, the maximum size of **R** and **X** will be M(M+1)/2.

For example, assume markers A and B have been genotyped in individual i. The vector $\mathbf{R}'$ is $[b_A, b_B, b_{AB}]$, where $b_{A(B)}$ denotes the main effect of marker A(B) on $IBD_{iu,iv}$, and b_{AB} denotes the effect of the interaction between markers A and B on $IBD_{iu,iv}$. **R** can be obtained as $\mathbf{G}\mathbf{V}^{-1}$, where **G** is a vector of covariances between IBS and IBD, and **V** is a matrix of IBS (co)variances among all markers and marker pairs. Table I shows all the necessary elements to obtain **R** in the two marker case. Appendix A contains all the necessary information to expand Table I for any number of markers.

The vector of adjusted IBS observations is $\mathbf{X}' = [X_A, X_B, X_{AB}]$, where $X_{AB} = x_{AB} - \bar{x}_{AB}$, and where x_{AB} is an observed variable for IBS taking value 1 if both markers A and B are simultaneously homozygous, and 0 otherwise (for more details see App. A).

In theory, equation (2) can be extended to take into account higher order marker interactions, *e.g.* IBS at marker triplets, quadruplets, etc. However, in order to construct the appropriate model, parameters $\theta_{k>4}$ are essential. These multiloci parameters must be developed either by extending Weir and Cockerham's theory or by using our approximations [6].

2.2.3. Haplotype versus genotype RM models

The ability of RM to use IBS at single markers and also joint IBS at marker pairs, allows us to create a battery of models adaptable to particular experimental needs. Here, we call genotype RM (RM_g) the model that fits all markers singly, and we call haplotype RM (RM_h) the model that fits all possible marker pairs in addition to single markers. Strictly speaking RM_h, as implemented in this manuscript, should be called a partial haplotype RM since no interactions between 3 or more loci are fitted. RM_h should be preferentially used when full haplotype information is available otherwise RM_g would be more appropriate.

2.3. Constructing an RM model to predict coancestry

Coancestry measures IBD between individuals. In MS populations, there is no distinction between inbreeding and coancestry, and therefore equation (2) is appropriate to calculate $IBD_{iu,jv}$, for all $i \neq j$. Nevertheless, we have found that more accurate results can be obtained using $\theta_{k=1...4}$ at generation $t+1$ to predict coancestry at generation t, because inbreeding in offspring equals coancestry among parents.

In ME, D and DH populations, the probability that two haplotypes carry IBD alleles at locus L depends on whether these haplotypes were sampled within or between individuals. In Weir *et al.* [27], two θ_2 parameters are calculated, one within and the other between individuals. Thus, in equation (2), θ_2 within individuals should be used when $i = j$, and θ_2 between individuals when $i \neq j$.

All $IBD_{iu,jv}$ probabilities are arranged in a gametic matrix of rank $2N$. For the purpose of analysis, this matrix is reduced to an individual matrix of rank N where each cell (i, j) contains the probability $IBD_{ij} = \frac{1}{2} \sum_{u,v=1}^{2} IBD_{iu,jv}$, where u and v denote haplotypes within individuals i and j, respectively (App. B).

2.4. A coalescent-based method

Meuwissen and Goddard [13] inferred pair-wise IBD probabilities between haplotypes using the coalescent theory, which is an approximate representation of genealogies [7,8]. The coalescent recreates the phylogenetic tree among current population members proceeding backwards in time, until a single expected common ancestor is reached, and all lineages have merged.

Note that CM does not distinguish $IBD_{iu,jv}$ between a situation where $i = j$ (*i.e.* estimating inbreeding) from a situation where $i \neq j$ (*i.e.* estimating

coancestry), because it was derived only for MS populations. Another subtle difference between CM and RM is that, at its core, CM estimates *en bloc* IBD probabilities, *i.e.* the probability of IBD segments between two haplotypes that include locus L and a region on both sides of L. This is a feature common to other theories as well [17, 18]. On the contrary, RM makes point inferences, from locus to locus, without interpolating IBD statuses between loci, and therefore it is less restrictive in concept, particularly over large distances.

2.5. Simulations

A population is founded at generation zero ($t = 0$) with N unrelated and non-inbred individuals. When simulating D and DH populations, we chose a male to female ratio of 1:9, *e.g.* a cattle herd. All marker allele frequencies are 0.5, and the population is in linkage equilibrium (LE) and Hardy-Weinberg equilibrium (HWE). There are 11 biallelic markers in total, evenly spaced over 10 cM, and a QTL placed in the centre of that region, at 5 cM from each end (*i.e.* on top of marker 6). At $t = 0$, each marker allele is labelled either 0 or 1 with probability 1/2, and QTL alleles are labelled 1 to $2N$, so that we can easily distinguish what QTL alleles are IBD in later generations. The population evolves over T discrete generations, with a fixed size per generation. Mating is at random, although the specific mating pattern is one of four possible, *i.e.* MS, ME, D or DH (see earlier). These mating schemes are all specifically modelled in RM through parameters $\theta_{k=1\ldots4}$. The increase of IBD probabilities at the QTL and the build-up of LD between QTL and markers occur because of drift, *i.e.* no other evolutionary force is acting.

The QTL effect was simulated in the last generation ($T = 100$ for monoecious populations, and $T = 50$ for dioecious) by choosing a QTL allele at random among all surviving alleles, and giving an effect of 1 to all alleles identical to the chosen allele, or 0 otherwise. Thus, the frequency of the allele with effect 1 ranged from 0.005 to 1, with a heritability ranging from 0 to ~0.5, and averaging ~0.28. There was no polygenic variance, and the residual error was drawn from a standard normal distribution. There were no other fixed or random effects affecting the trait.

2.6. Comparing RM and CM in terms of power and robustness of QTL detection

Power is the probability of detecting a true QTL, and robustness is the independence between initial assumptions and final results. Both power

and robustness were measured as residual log-likelihood ratios, *i.e.* $LR = -2\ln(LH_0/LH_1)$, between a model for the null hypothesis of an unlinked QTL (H_0: $c_L = 0.5$), and a model for the alternative hypothesis of a linked QTL (H_1:$c_L < 0.5$), where L is now the location at which *LR* is evaluated, and c_L denotes the recombination rate between location L and the true QTL location.

The phenotypes are modelled as $\mathbf{y} = \mu + \mathbf{a} + \mathbf{e}$, where $\mathbf{y}$ is a vector of phenotypes, μ is a vector with the overall mean, $\mathbf{a}$ is a vector of additive QTL effects for each individual, and $\mathbf{e}$ is a vector of random, normally distributed residuals. The residual log likelihood of the previous model is

$$L\left(\mathbf{G}_{\mathrm{L}}, \sigma^2_{\mathrm{a}}, \sigma^2_{\mathrm{e}}\right) \propto -0.5\left[\ln|\mathbf{V}| + \ln\left|\mathbf{1}'\mathbf{V}^{-1}\mathbf{1}\right| + (\mathbf{y}-\boldsymbol{\mu})'\,\mathbf{V}^{-1}\,(\mathbf{y}-\boldsymbol{\mu})\right]$$

where σ^2_{a} is the variance of vector $\mathbf{a}$, which is equivalent to the QTL variance estimated at location L, σ^2_{e} is the variance of vector $\mathbf{e}$, and $\mathbf{G}_{\mathrm{L}}$ is the matrix of IBD probabilities among individuals at location L ($\mathbf{1}$ denotes a unit vector). The phenotypic variance is $\mathbf{V} = \sigma^2_{\mathrm{a}}\mathbf{G}_{\mathrm{L}} + \sigma^2_{\mathrm{e}}\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This likelihood is maximised with respect to σ^2_{a} and σ^2_{e} at each location L using ASREML [4].

The covariance between individuals due to the QTL was modelled with the individual IBD matrix $\mathbf{G}_{\mathrm{L}}$ obtained with CM, $\mathrm{RM_h}$ or $\mathrm{RM_g}$. Under H_0, *LR* is distributed with a probability mass of 1/2 at $LR = 0$ and continuously with a density equivalent to $1/2\,\chi^2_1$ for $LR \geq 0$ giving a critical value at a 5% error rate of 2.7 [17]. We calculated *LR* at the midpoint between every consecutive pair of markers (10 locations) plus marker 6 (the true QTL location), and averaged the results over 1000 replicates. The results were virtually the same as with 100 replicates (not shown).

We chose as a measure of power the average *LR* at marker 6 given correct population parameters, and as a measure of robustness the same *LR* given wrong population parameters. Robustness was studied only in the case of an MS population, to allow fair comparisons with CM. There were four different scenarios in which robustness was tested: (1) maximum LD among founders, (2) $p = 0.9$ was correct but a wrong 0.5 was used, (3) $Ne = 100$ was correct but a wrong 50 was used, and (4) only genotypes were available, *i.e.* marker phases were unknown.

3. RESULTS

3.1. The effect of mating scheme on power of QTL detection

Figure 1 shows the impact of different random mating schemes on the performance of $\mathrm{RM_h}$ and CM. The plots are average *LR* obtained at 11 positions,

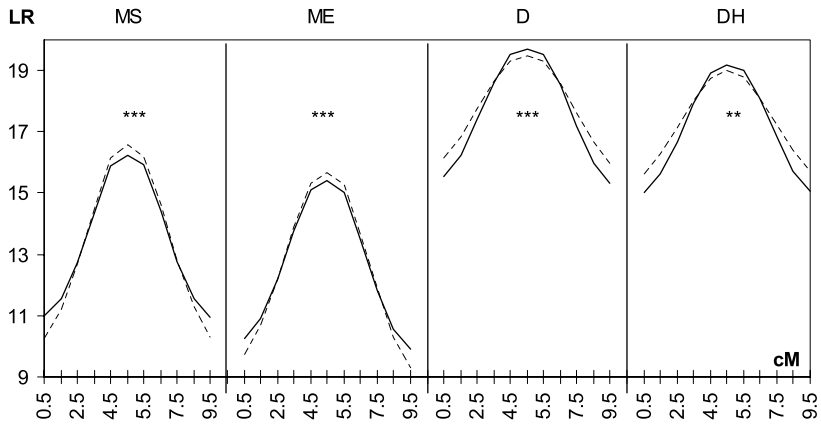

Figure 1. Power of QTL detection under four different random mating schemes using either CM (dotted line) or RM_h (continuous line) to predict IBD. MS: Monoecious with selfing, ME: Monoecious excluding selfing, D: Dioecious, DH: Dioecious with hierarchical mating, $LR = -2\ln(LH_0/LH_1)$, where LH_i is the likelihood under H_i hypothesis, cM: centi-Morgans. There were 1000 *LR* tests averaged at the centre of each marker interval, and at marker 6 (5 cM from either end and where the QTL is). Population parameters were: 10 males and 90 females; $T = 100$ for monoecious schemes, and 50 for dioecious schemes; $p = 0.5$; and LE at generation 0. Significance for a paired *t*-test of differences at marker 6: *** denotes $p < 0.001$, ** denotes $p < 0.01$.

i.e. at the centre of each marker interval plus at the true QTL position on marker 6, along a chromosome segment of 10 cM. One thousand different data sets were simulated within each of the mating schemes or populations (MS, ME, D, DH). Both RM_h and CM *LR* profiles are symmetric with respect to the maximum average value located on marker 6, *i.e.* the results are unbiased. The differences between RM_h and CM across populations were small but highly significant in most cases. The asterisks in Figures 1, 2 and 3 represent p-values associated with the H_0 hypothesis of no difference in mean *LR* between RM_h and CM at marker 6 (paired *t*-test with 999 degrees of freedom). Three asterisks denote $p < 0.001$, two denote $p < 0.01$, and ns denotes not significant. We tested *LR* differences between methods at all other positions on the chromosome. For example, the eleven *p*-values for the paired *t*-test in the MS population of Figure 1 are $<10^{-16}$, $<10^{-5}$, 0.4, 0.01, $<10^{-7}$, $<10^{-14}$, $<10^{-8}$, $<10^{-4}$, 0.5, $<10^{-4}$, $<10^{-14}$, indicating that unless the two profiles are very close (*e.g.* as in the LD population in Fig. 2), differences in mean *LR* are likely to be significant.

The average LR at marker six can be used as a measure of power of QTL detection. Thus, RM_h is more powerful than CM in dioecious populations,

and less powerful in monoecious populations ($p < 0.001$ for a paired *t*-test in ME, MS, and D, and < 0.01 in DH). This is unsurprising given that, on the one hand, CM can use all haplotype information, whereas RM_h uses partial haplotype information up to all pair-wise marker interactions, and on the other hand, RM_h can model four different populations, whereas CM models only an MS population.

In addition to the height, the sharpness of the profile also contains valuable information for mapping QTL because sharper profiles correspond to smaller confidence intervals around location estimates. RM_h produced sharper profiles than CM in dioecious populations, but not in monoecious populations.

Table II shows estimates of two alternative measures of power obtained at marker six, *i.e.* the probability of rejecting the H_0 of unlinked QTL with a 5% error rate, and the odds of RM_h rendering a higher *LR* than either CM or RM_g. The probabilities of rejecting H_0 were very similar across methods, and although CM rejected H_0 slightly more times than RH_h, it happened less than 2% of the time. The odds of RM_h/CM denote how many times is RM_h more likely to render a higher *LR* than CM for the same data set. These odds show clearer differences between methods than the proportion of times H_0 is rejected. For example, the odds of RM_h/CM in monoecious populations was ~0.9, whereas in dioecious populations was >1.6. In order to get the overall picture of power of these methods, all three measures of power (average *LR*, probability of rejecting H_0, and odds ratio) have to be considered together. For example, in D populations, RM_h rendered a larger *LR* than CM ~2/3 of the times (~1.8 odds ratio), its average *LR* was only slightly larger (19.7 *vs.* 19.5) although significant ($p < 0.001$), but there were no differences in the probability of rejecting H_0 (~0.9 for both methods). These power estimates reflect different features of the distribution of *LR*. For example, the distribution obtained with RM_h in D populations tends to be flatter, and with thicker and longer tails than the distribution obtained with CM (not shown), explaining why RM_h has a larger mean than CM and why it tends to render more extreme LR values. Even though, the proportion of the distribution greater than the threshold 2.7 seems to be approximately equal for both methods.

Table II also shows estimates of QTL (σ^2_{QTL}) and residual (σ^2_e) variances at marker six. Both methods overestimated σ^2_{QTL}, and slightly underestimated σ^2_e. These biases increased with distance between the tested position and the QTL (not shown). Nevertheless, CM was consistently less biased than RM_h. It is not clear what proportion of the bias was caused by the variance estimation procedure, *i.e.* constraining variances to be positive in REML analyses, and what by the method of IBD estimation *per se*.

Table II. Comparing RM_h *vs.* CM, and RM_h *vs.* RM_g regarding power, robustness, and variance components calculated at the true QTL location (marker 6).

	Reject H_0		Odds	σ^2_{QTL}		σ^2_e	
			Power				
	RM_h	CM	RM_h/CM	RM_h	CM	RM_h	CM
MS	87	88	0.9 ^{ns}	0.81	0.65	0.95	0.97
ME	87	88	0.92 ^{ns}	0.73	0.64	0.95	0.98
D	90	90	1.82 ***	0.82	0.63	0.99	0.99
DH	88	89	1.63 ***	0.78	0.59	0.99	0.98
			Robustness				
LD	80	83	0.85 *	1.03	0.98	0.94	0.96
P	53	55	1.15 *	1.56	0.82	1.06	1.14
N	83	84	1.21 **	0.71	0.64	1.03	1.04
H	87	86	1.61 ***	0.58	0.5	0.97	0.99
			RM_h *vs.* RM_g				
RM_h	RM_g	RM_h/RM_g	RM_h	RM_g	RM_h	RM_g	
C	86	85	1.22 **	0.73	1.18	0.96	0.86
LD	81	79	0.63 ^{ns}	0.83	0.9	1	0.9
P	56	56	1.11 ^{ns}	1.83	4.69	1.05	0.86
H	86	87	1.11 ***	0.59	0.9	0.97	0.9

$RM_{h(g)}$: regression method using haplotype (genotype) model;
CM: coancestry method;
Reject H_0: % times H_0 is rejected at a 5% error rate;
Odds RM_h/CM: number of times RM_h renders a higher *LR* than CM;
odds RM_h/RM_g: number of times RM_h renders a higher *LR* than RM_g;
σ^2_{QTL}: QTL variance (true average variance = 0.4);
σ^2_e: residual variance (true average variance = 1);
MS: monoecious with selfing ($Ne = T = 100$);
ME: monoecious excluding selfing ($Ne = T = 100$);
D: dioecious ($Ne = 100$, $T = 50$);
DH: dioecious with hierarchical mating ($Ne = 100$, $T = 50$);
LD: maximum LD among founders;
P: wrong p (true = 0.9, used =0.5);
N: wrong Ne (true = 100, used = 50);
H: wrong haplotypes (unknown marker phases);
C: correct parameters;
A non-parametric sign test was used to determine the significance of odds ratios: $p < 0.001^{(***)}$, $p < 0.01^{(**)}$, $p < 0.05^{(*)}$, or not significant $^{(ns)}$.
The comparison between RM_h and RM_g was for an MS population.

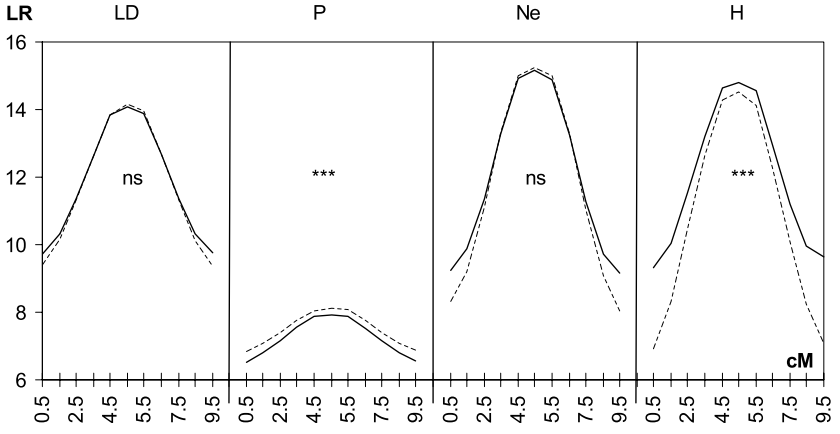

Figure 2. Robustness of QTL detection under a monoecious population with selfing using either CM (dotted line) or RM_h (continuous line) to predict IBD. Population parameters were: $N = 100$; $T = 100$; $p = 0.5$; LE at generation 0. The parameter changes were: LD denotes maximum LD among founders, P denotes p was 0.9 but assumed to be 0.5, Ne denotes Ne was 100 but assumed to be 50, and H denotes haplotype errors. $LR = -2\ln(LH_0/LH_1)$, where LH_i is the likelihood under H_i hypothesis, cM: centi-Morgans. There were 1000 *LR* tests averaged at the centre of each marker interval, and at marker 6 (5 cM from either end and where the QTL is). Significance for a paired t-test of differences at marker 6: *** denotes $p < 0.001$, ns denotes $p > 0.05$.

In summary, RM_h is slightly more powerful, and has sharper *LR* profiles than CM in dioecious populations but not in monoecious populations. Different measures of power help in understanding the performance of these methods, since each measure was more sensitive to different properties of the distribution of *LR*. The variance components were biased, especially σ^2_{QTL}, and more for RM_h than for CM.

3.2. The robustness of CM and RM to wrong assumptions

Figure 2 shows LR profiles when one of the population parameters used in the simulations was assumed to be inaccurately known, *i.e.* recreating a situation where there is poor historical information to predict IBD_L. We examined robustness in MS populations, because this is the only mating scheme modelled by CM. It is expected that RM_h will be more robust than CM in D and DH populations.

The use of inaccurate historical parameters did not bias the location estimates, but reduced the power of QTL detection in all cases, causing

power losses from ~1 to ~8 *LR* units. Both methods were fairly robust to misspecifications in all parameters but p. Underestimating homozygosity in founders, *i.e.* assuming $p = 0.5$ when the real value was 0.9, caused the largest observed power loss, probably due to an overoptimistic belief in the information content of homozygous markers coupled with higher than expected levels of homozygosity at any generation.

RM_h and CM are robust to wrong parameters *Ne* and *T* (not shown), and wrong LE assumption, *i.e.* when there is maximum LD among founders. In all these situations, both methods rendered similar *LR* profiles ($p > 0.05$ for a paired *t*-test everywhere but the most distant markers). Note that CM was more powerful than RM_h in MS populations under correct assumptions (Fig. 1). Thus, the fact that *LR* profiles became almost identical when some parameters were inaccurate, *e.g. Ne*, indicates that RM_h lost less power, and therefore it was more robust, than CM.

The two other scenarios in which robustness was tested were wrong p and unknown haplotypes. RM_h was less robust than CM in the former scenario but more in the latter. The fact that RM_h tends to give higher *LR* values than CM in both scenarios (odds RM_h/CM in Tab. II) can be explained by the shape of the distribution of *LR*, *i.e.* more skewed and flatter for CM than for RM_h (not shown). Hence, although in more replicates RM_h renders a higher *LR* than CM, in some replicates CM renders an *LR* considerably larger than RM_h.

Both methods still showed biased σ^2_{QTL} estimates at marker six, but this time, differences between methods were much smaller; especially when *Ne* was wrong, there was maximum LD among founders, or marker phases were unknown. The largest bias occurred in both methods when p was wrong, *i.e.* both σ^2_{QTL} and σ^2_e were overestimated.

In summary, RM_h lost less power relative to CM when *Ne* was wrong, there was maximum LD among founders, or marker phases were unknown. Moreover, in all these situations both methods rendered very similar σ^2_{QTL} estimates. Only when p was wrong, CM still showed a small advantage of 0.2 *LR* units over RM_h.

3.3. Comparing haplotype and genotype models of RM

RM_h was more powerful and robust than RM_g in all cases except, predictably, when marker phases were unknown (Fig. 3). This is because RM_g does not use the structure of haplotypes to predict IBD. Nevertheless, it seems that genotype information is sufficient to predict QTL location accurately, *e.g.*

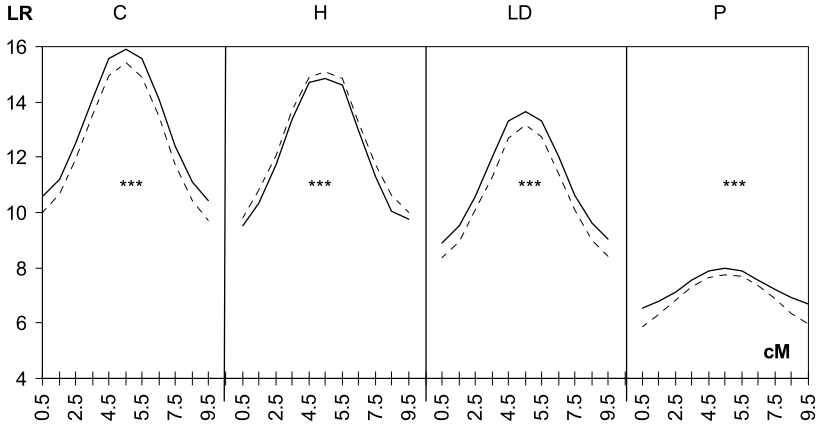

Figure 3. Differences between the genotype model RM_g (dotted line) and the haplotype model RM_h (continuous line) in terms of power and robustness of QTL detection, in a monoecious population with selfing. Population parameters: $N = 100$; $T = 100$; $p = 0.5$; LE at generation 0. C denotes correct information, H incorrect haplotypes, LD maximum LD among founders, and P wrong p (actual = 0.9, used = 0.5). $LR = -2\ln(LH_0/LH_1)$, where LH_i is the likelihood under H_i hypothesis, cM: centi-Morgans. There were 1000 LR tests averaged at the centre of each marker interval, and at marker 6 (5 cM from either end and where the QTL is). Significance for a paired t-test of differences at marker 6: *** denotes $p < 0.001$.

less than half *LR* unit difference between RM_h and RM_g at marker 6, despite variance components being consistently more biased with RM_g than with RM_h.

4. DISCUSSION

Genetic relatedness between individuals is a key parameter in QTL mapping *via* variance component estimation. In QTL mapping, individuals without known pedigree are commonly assumed unrelated and non-inbred. However, if they are sampled from the same population, then they are likely to be related to some degree. Our prime interest in this study was to explore a new method to estimate genetic relatedness among pedigree founders as IBD probabilities at a locus as functions of distant relationships, *i.e.* historical, and marker similarities. These IBD probabilities would then be used in mapping QTL *via* variance components estimation. Thus, the approach consists of a family of linear regression models that suits various QTL mapping scenarios, covering a range of availability of multilocus data from genotypes to haplotypes.

We compared the power and robustness of QTL detection of two of these models, RM_h and RM_g, against an alternative method, CM. This focus on CM

as a base for comparison was justified since other methods available for estimating IBD do not take into account population history, encounter difficulties in weighting information from several markers, or do not consider information from linked markers [10, 22, 24].

In general, differences in statistical power and robustness between RM and CM were small, although sometimes significant (*N.B*. Sample sizes of 1000 simulations were used in the comparison). RM_h tended to be more powerful than CM in dioecious populations, but not in monoecious populations. The main advantage of CM over RM_h is that it uses more LD information contained in haplotypes. RM_h captures LD information from single markers and all possible interactions between marker pairs, but not from marker triplets or higher-order interactions, whereas CM uses the full haplotype structure to infer IBD probabilities. It would be theoretically possible to increase the accuracy of RM_h by modelling these higher order marker interactions. However, most of the information used to predict IBD appears to come from genotypes rather than haplotypes, since differences between RM_h and RM_g are small (Fig. 3). Reassuringly, RM_g was more powerful than RM_h when marker phases (*i.e.* haplotypes) were unknown.

The RM and CM methods both require historical information that is poorly known in reality. Power losses occur when population parameters are wrong or model assumptions are not fulfilled. For this reason, these methods should ideally be robust in addition to powerful. Robustness was studied in MS populations, because this is the only mating scheme explicitly accounted for in CM. Thus, taken relative power loss with respect to the ideal situation as a measure of robustness (compare Figs. 1 and 2), RM_h is slightly more robust than CM in some scenarios but not all. For example, the difference in average LR at marker six between CM and RM_h was 0.4 under ideal conditions (Fig. 1, MS), however this difference dropped, vanished or favoured RM_h when population parameters or haplotypes were inaccurately known (Fig. 2). The exception was when p was wrong, where a difference of 0.4 *LR* units was maintained between methods.

Both methods rendered biased estimates of variances, especially of σ^2_{QTL}. In general, RM_h was more biased than CM, however differences diminished or disappeared when population history, or haplotypes, were inaccurately known. Nevertheless, we cannot rule out the possibility that part of that bias is due to the variance component estimation method, *i.e.* REML, because variances were constrained to be positive.

Differences in computational speed between methods were negligible in samples of 100 individuals and 11 markers, *i.e.* CM needed ~10 seconds

to calculate IBD at one location compared to ~0.1 seconds for RM_h, in a Dec-Alpha XP1000, with 667 MHz and 1152 Mb of memory. We do not yet know what would be the behaviour of these methods had they been challenged with thousands of markers in thousands of individuals, a likely future scenario where computational speed can be a serious limiting factor.

Notwithstanding its greater power in dioecious populations, there are further important advantages in RM related to its scope for expansion and improvement. RM is based on a multiple regression, and therefore one could fit a range of models easily in order to optimise for speed without reducing too much accuracy. For example, one could drop interactions between markers far from locus L, which may contain negligible information, whilst keeping all single marker effects and interactions between marker pairs close to L. In this way, one could construct less parameterised models without much loss in accuracy, *i.e.* achieving parsimony. Moreover, the key elements in RM are the multilocus IBD parameters $\theta_{k=1...4}$. Thus, exact $\theta_{k=1,2}$, and approximated $\theta_{k=3,4}$, allowed us to model four different randomly mating populations, without mutation, migration or selection. Choy and Weir [1] developed expressions for θ_2 under recurrent random selection, hence broadening the applicability of our method. However, the theory of long-term genetic contributions may offer a better alternative for modelling these parameters in non-randomly selected populations because, for example, it appropriately models selection as a non-markovian process [28]. We do not know of an equivalent parameter in the coalescent theory that could be used in CM. Furthermore, a recent publication expanded the two loci theory to include mutation and migration rates [23]. New developments of RM into this area could overcome the problem of hidden population structures in mapping QTL [5].

Finally, if a pedigree was available, then the RM method should be used to predict IBD among pedigree founders, provided they are genotyped, and then pedigree information should be applied to estimate IBD among all their descendants (*e.g.* [16]). Neither CM nor RM can take into account ungenotyped pedigree ancestors. Therefore, with this exception in mind, this joint procedure would ensure that all information available in a sample, *i.e.* markers, haplotypes, pedigree and population history, is used when mapping QTL.

In summary, given that methods to recapture historical information must deal in most circumstances with uncertainty in the key population parameters, and that populations may greatly differ in their mating schemes, the family of models presented in this study offers (RM) a robust, yet powerful, approach to map QTL using population history and multiple marker information.

ACKNOWLEDGEMENTS

We acknowledge the Biotechnology and Biological Sciences Research Council of UK for funding this work. We also thank Dr. Pong-Wong, and various anonymous reviewers that helped us to improve this manuscript.

REFERENCES

[1] Choy S.C., Weir B.S., Two-locus inbreeding measures for recurrent selection, Theor. Appl. Genet. 49 (1977) 63–77.
[2] Cockerham C.C., Weir B.S., Digenic descent measures for finite populations, Genet. Res. 30 (1977) 121–147.
[3] George A.W., Visscher P.M., Haley C.S., Mapping quantitative trait loci in complex pedigrees: a two step variance component approach, Genetics 156 (2000) 2081–2092.
[4] Gilmour A.R., Gogel B.J., Cullis B.R., Welham S.J., Thompson R., ASReml user guide release 1.0 VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, 2002.
[5] Grapes L., Dekkers J.C.M., Rothschild M.F., Fernando R.L., Comparing linkage disequilibrium-based methods for fine mapping quantitative trait loci, Genetics 166 (2004) 1561–1570.
[6] Hernández-Sánchez J., Haley C.S., Woolliams J.A., On the prediction of simultaneous inbreeding coefficients at multiple loci, Genet. Res. 83 (2004) 113–120.
[7] Kingman J.F.C., The Coalescent, Stoch. Proc. Appl. 13 (1982) 235–248.
[8] Kingman J.F.C., On the genealogy of large populations, J. Appl. Probab. 19 (1982) 27–43.
[9] Leutenegger A.-L., Prum B., Génin E., Verny C., Lemainque A., Clerget-Darpoux F., Thompson E., Estimation of the inbreeding coefficient through use of genomic data, Am. J. Hum. Genet. 73 (2003) 516–523.
[10] Lynch M., Ritland K., Estimation of pairwise relatedness with molecular markers, Genetics 152 (1999) 1753–1766.
[11] Lynch M., Walsh B., Genetics and analysis of quantitative traits, 1st edn., Sinauer Associates, Sunderland, 1998.
[12] Meuwissen T.H.E., Goddard M.E., Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci, Genetics 155 (2000) 421–430.
[13] Meuwissen T.H.E., Goddard M.E., Prediction of identity by descent probabilities from marker-haplotypes, Genet. Sel. Evol. 33 (2001) 605–634.
[14] Meuwissen T.H.E., Goddard M.E., Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data, Genet. Sel. Evol. 36 (2004) 261–279.
[15] Meuwissen T.H.E., Astrid K., Lien S., Olsaker I., Goddard M.E., Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping, Genetics 161 (2002) 373–379.

[16] Pong-Wong R., George A.W., Woolliams J.A., Haley C.S., A simple and rapid method for calculating identity-by-descent matrices using multiple markers, Genet. Sel. Evol. 33 (2001) 453–471.
[17] Self S.G., Liang K.Y., Asymptotic properties of maximum-likelihood estimators and likelihood ratio tests under non-standard conditions, J. Am. Stat. Assoc. 82 (1987) 605–610.
[18] Sved J.A., Linkage disequilibrium and homozygosity of chromosome segments in finite populations, Theor. Popul. Bio. 2 (1971) 125–141.
[19] Sved J.A., Feldman M.W., Correlation and probability methods for one and two loci, Theor. Popul. Bio. 4 (1973) 129–132.
[20] Thomas S.C., Hill W.G., Estimating quantitative genetic parameters using sibships reconstructed from marker data, Genetics 155 (2000) 1961–1972.
[21] Thompson E.A., The estimation of pairwise relationships, Ann. Hum. Genet. 39 (1975) 173–188.
[22] Van de Casteele T., Galbusera P., Matthysen E., A comparison of microsatellite-based pairwise relatedness estimates, Mol. Ecol. 10 (2001) 1539–1549.
[23] Vitalis R., Couvet D., Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population, Genet. Res. 77 (2001) 67–81.
[24] Wang J., An estimator for pairwise relatedness using molecular markers, Genetics 160 (2002) 1203–1215.
[25] Weir B.S., Cockerham C.C., Group inbreeding with two linked loci, Genetics 63 (1969) 711–742.
[26] Weir B.S., Cockerham C.C., Behavior of pairs of loci in finite monoecious populations, Theor. Popul. Bio. 6 (1974) 323–354.
[27] Weir B.S., Avery P.J., Hill W.G., Effect of mating structure on variation in inbreeding, Theor. Popul. Bio. 18 (1980) 396–429.
[28] Woolliams J.A., Bijma P., Villanueva B., Expected genetic contributions and their impact on gene flow and genetic gain, Genetics 153 (1999) 1009–1020.

APPENDIX A

We will specify the parameters used in equation (2). The theory of digenic descent measures provides most of the building blocks required to obtain all the parameters [2, 25–27]. This theory predicts θ_2 (IBD at two loci simultaneously) at generation t from its expected value in the previous generation. In fact, the theory calculates non-IBD probabilities iteratively, transforming them to IBD probabilities in the last generation. In ME, D and DH populations, θ_2 differs when haplotypes are sampled within or between individuals. The single locus θ_1 is equal to θ_2 given $c = 0$.

The weights in equation (2) are $\mathbf{R} = \mathbf{V}^{-1}\mathbf{G}$, where $\mathbf{V}^{-1}$ is the inverse of the IBS covariance matrix between markers, and $\mathbf{G}$ is a covariance vector between IBS at markers and $\mathrm{IBD_L}$. The corrections in $\mathbf{X}$ are the expected IBD at single

and pair of marker loci. We will proceed with the details of how to calculate **X** and **R**.

Remember that in what follows, IBS (homozygosity) statuses at marker loci are obtained comparing two different haplotypes, within or between individuals.

(i) Elements of X

The vector **X** contains IBS information from markers. For example, given markers A and B, then $\mathbf{X}' = [x_A - \bar{x}_A,\ x_B - \bar{x}_B,\ x_{AB} - \bar{x}_{AB}]$, where $x_i = 1$ if locus i is IBS (homozygous), or 0 otherwise, $x_{ij} = 1$ if loci i and j are both IBS, or 0 otherwise, and $\bar{x}_i$ and $\bar{x}_{ij}$ are the corresponding population expectations. The expected IBS at locus A in the population is

$$\bar{x}_A = \theta_A + (1 - \theta_A)\,\Pi_A, \tag{A.1}$$

where $\Pi_A = \sum p_A^2$ is the initial homozygosity at locus A, and θ_A is the probability of IBD at locus A. We need to specify each locus, therefore θ_A will denote θ_1 at locus A, θ_{AB} will denote θ_2 at loci A and B, etc.

The expected simultaneous homozygosity at markers A and B is

$$\bar{x}_{AB} = \theta_{AB} + (\theta_A - \theta_{AB})\,\Pi_B + (\theta_B - \theta_{AB})\,\Pi_A + (1 - \theta_A - \theta_B + \theta_{AB})\,\Pi_A\Pi_B. \tag{A.2}$$

For the special case of two unlinked loci, and assuming $\Pi_A = \Pi_B = \Pi$ and $\theta_A = \theta_B = \theta$ (and therefore $\theta_{AB} = \theta^2$), it can be shown that $\bar{x}_{AB} = \bar{x}_A\bar{x}_B = (\bar{x}_A)^2 = (\bar{x}_B)^2$, which is the square of (A.1).

(ii) Elements of V

The matrix **V** is a symmetric matrix containing the co-variances of IBS among marker loci. The rank of **V** is equal to the total number of terms fitted in the model. Let us assume a model with two main effects, from markers A and B, and their interaction. In this case, the diagonal elements of **V** are the variances $\boldsymbol{\sigma}^2(x_A)$, $\boldsymbol{\sigma}^2(x_B)$ and $\boldsymbol{\sigma}^2(x_{AB})$, and the (upper) off-diagonal elements of **V** are the covariances $\boldsymbol{\sigma}(x_A, x_B)$, $\boldsymbol{\sigma}(x_A, x_{AB})$, and $\boldsymbol{\sigma}(x_B, x_{AB})$ (Tab. II).

The variance of IBS at marker A is $\boldsymbol{\sigma}^2(x_A) = \overline{x_A^2} - (\bar{x}_A)^2$, where $\bar{x}_A$ is given in (A.1), and $\overline{x_A^2} = \sum_{i=0}^{1} i^2 P(x_A = i) = P(x_A = 1) = \bar{x}_A$. Hence, this variance is $\boldsymbol{\sigma}^2(x_A) = \bar{x}_A(1 - \bar{x}_A)$, which after using (A.1) and simplifying the algebra renders,

$$\boldsymbol{\sigma}^2(x_A) = (1 - \theta_A)(1 - \Pi_A)(\theta_A + (1 - \theta_A)\,\Pi_A). \tag{A.3}$$

The variance of simultaneous IBS at loci A and B is

$$\boldsymbol{\sigma}^2(x_{AB}) = \overline{x_{AB}^2} - (\bar{x}_{AB})^2 = \bar{x}_{AB}(1 - \bar{x}_{AB}), \tag{A.4}$$

since $\overline{x_{AB}^2} = \sum_{i=0}^{1} i^2 P(x_{AB} = i) = P(x_{AB} = 1) = \bar{x}_{AB}$, given in (A.2).

The covariance between IBS at markers A and B simultaneously is $\sigma(x_A, x_B) = \bar{x}_{AB} - \bar{x}_A \bar{x}_B$, thus

$$\boldsymbol{\sigma}(x_A, x_B) = (\theta_{AB} - \theta_A \theta_B)(1 - \Pi_A - \Pi_B + \Pi_A \Pi_B), \tag{A.5}$$

which simplifies to

$$\boldsymbol{\sigma}(x_A, x_B) = \eta_{AB}(1 - \Pi)^2$$

when $\Pi_A = \Pi_B = \Pi$, and $\theta_A = \theta_B = \theta$. The parameter $\eta_{AB} = \theta_{AB} - \theta^2$ has been called the *identity disequilibrium function* [2], and it is an estimate of the variation of IBD across loci due to the combined effects of drift and linkage.

The covariance between IBS at a single marker and simultaneous IBS at a pair of markers depends on whether the single marker is included in that pair or not. When only markers A and B have been genotyped, the appropriate covariance between locus A and loci A and B is

$$\boldsymbol{\sigma}(x_A, x_{AB}) = \overline{x_A x_{AB}} - \bar{x}_A \bar{x}_{AB} = \bar{x}_{AB}(1 - \bar{x}_A), \tag{A.6}$$

since $\overline{x_A x_{AB}} = \bar{x}_{AB}$, see (A.1) and (A.2).

There are additional covariances in $\mathbf{V}$ when 3 or more markers are used, for example, with markers A, B, C and D, the additional covariances are $\boldsymbol{\sigma}\left(x_i, x_{jk}\right)$, $\boldsymbol{\sigma}\left(x_{ij}, x_{ik}\right)$, and $\boldsymbol{\sigma}\left(x_{ij}, x_{kl}\right)$, where $i \neq j \neq k \neq l =$ A, B, C or D. The covariance between IBS at marker A and simultaneous IBS at markers B and C is

$$\boldsymbol{\sigma}(x_A, x_{BC}) = \bar{x}_{ABC} - \bar{x}_A \bar{x}_{BC}, \tag{A.7}$$

where the term $\bar{x}_{ABC}$ is

$$\theta_{ABC} + \sum_{i<j} \left(\theta_{ij} - \theta_{ABC}\right) \Pi_k + \sum_i \left(\theta_i - \sum_j \theta_{ij} + \theta_{ABC}\right) \Pi_j \Pi_k + \left(1 - \sum_i \theta_i + \sum_{i<j} \theta_{ij} - \theta_{ABC}\right) \Pi_A \Pi_B \Pi_C \tag{A.8}$$

where $i \neq j \neq k =$ A, B, or C.

Equation (A.8) involves 2 and 3-loci IBD parameters. A simple and accurate method for predicting simultaneous IBD at 3 and 4 loci has been proposed elsewhere [6].

The covariance between two pairs of markers with a common marker, for example between pairs AB and AC, is

$$\boldsymbol{\sigma}(x_{AB}, x_{AC}) = \overline{x_{AB}x_{AC}} - \bar{x}_{AB}\bar{x}_{AC} = \bar{x}_{ABC} - \bar{x}_{AB}\bar{x}_{AC}, \tag{A.9}$$

since $\overline{x_{AB}x_{AC}} = \bar{x}_{ABC}$. The covariance between two completely different pairs of markers, for example between pairs AB and CD, is

$$\boldsymbol{\sigma}(x_{AB}, x_{CD}) = \bar{x}_{ABCD} - \bar{x}_{AB}\bar{x}_{CD}, \tag{A.10}$$

where $\bar{x}_{ABCD}$ is

$$\begin{aligned}\theta_{ABCD} + \sum_{i<j<k}\left(\theta_{ijk} - \theta_{ABCD}\right)\Pi_l + \sum_{i<j}\left(\theta_{ij} - \sum_k \theta_{ijk} + \theta_{ABCD}\right)\Pi_k\Pi_l \\ + \sum_i\left(\theta_i - \sum_j \theta_{ij} + \sum_{j<k}\theta_{ijk} - \theta_{ABCD}\right)\Pi_j\Pi_k\Pi_l \\ + \left(1 - \sum_i \theta_i + \sum_{i<j}\theta_{ij} - \sum_{i<j<k}\theta_{ijk} + \theta_{ABCD}\right)\Pi_i\Pi_j\Pi_k\Pi_l,\end{aligned} \tag{A.11}$$

where $i \neq j \neq k \neq l$ = A, B, C or D. The covariances (A.5), (A.6), (A.7), (A.9) and (A.10) reflect the build-up of LD due to the combined action of linkage and drift. Hence, they are expected to be zero if there is neither linkage nor drift, and greater than zero in all other cases.

(iii) Elements of G

The vector **G** contains the covariances between IBS at single or pairs of markers and $\mathrm{IBD_L}$. For example, the covariance between $\mathrm{IBS_A}$ and $\mathrm{IBD_L}$ is

$$\boldsymbol{\sigma}(x_A, y_L) = \overline{x_A y_L} - \bar{x}_A \cdot \bar{y}_L = \eta_{AL}(1 - \Pi_A), \tag{A.12}$$

where $\bar{y}_L = \theta$, and $\eta_{AL} = \theta_{AL} - \theta^2$, and where the joint probability of $\mathrm{IBS_A}$ and $\mathrm{IBD_L}$, is

$$\overline{x_A y_L} = \theta_{AL} + (\theta - \theta_{AL})\Pi_A. \tag{A.13}$$

The covariance between simultaneous $\mathrm{IBS_{AB}}$ and $\mathrm{IBD_L}$ is

$$\sigma(x_{AB}, y_L) = \overline{x_{AB}y_L} - \bar{x}_{AB}\bar{y}_L, \tag{A.14}$$

where

$$\overline{x_{AB}y_L} = \theta_{ABL} + (\theta_{AL} - \theta_{ABL})\,\Pi_B + (\theta_{BL} - \theta_{ABL})\,\Pi_A + (\theta - \theta_{AL} - \theta_{BL} + \theta_{ABL})\,\Pi_A\Pi_B.$$

APPENDIX B

The matrix calculated with CM and RM provides IBD relationships for every gametic pair, and therefore it has rank $2N$, where N is the number of individuals in the sample. Let us call this matrix **G**. It is possible to transform **G** into an N rank matrix **Q** with $\mathbf{Q} = \frac{1}{2}\mathbf{KGK'}$, where $\mathbf{K} = \mathbf{I}^*[1, 1]$, and I is an identity matrix with rank N, and * denotes the Kronecker product between matrices [16].