Genetics: Published Articles Ahead of Print, published on June 11, 2007 as 10.1534/genetics.107.070953

Multipoint IBD prediction using

dense markers to map QTL and

estimate effective population size

Theo H. E. Meuwissen[1] and Mike E. Goddard[2,3]

[1] IHA - Norwegian University of Life Sciences, Ås, Norway

[2] Department of Primary Industry, Attwood, Australia

[3] University of Melbourne, Melbourne, Australia

Running Head: Multipoint IBD prediction

Keywords: Identity-by-descent; linkage disequilibrium; Quantitative Trait Locus mapping; effective population size; coalescence process

Corresponding author:

Theo Meuwissen

Mailing address:

IHA – Norwegian University of Life Sciences

Box 5003, 1432 Ås

Norway

Phone: +4764965107

Fax: +4764965101

Email: theo.meuwissen@umb.no

ABSTRACT

A novel multipoint method, based on an approximate coalescence approach, to analyse multiple linked markers is presented. Unlike other approximate coalescence methods, it considers all markers simultaneously but only two haplotypes at a time. We demonstrate the use of this method for LD mapping of QTL and estimation of effective population size. The method estimates Identity-by-Descent (IBD) probabilities between pairs of marker haplotypes. Both linkage disequilibrium (LD) and combined linkage and LD mapping rely on such IBD probabilities. The method is approximate in that it only considers the information on a pair of haplotypes, whereas a full modelling of the coalescence process would simultaneously consider all haplotypes. However, full coalescence modelling is only computationally feasible for few linked markers. Using simulations of the coalescence process, the method is shown to give almost unbiased estimates of the effective population size. Compared to direct marker and haplotype association analyses, IBD based QTL mapping showed clearly a higher power to detect a QTL and a more realistic confidence interval for its position. The modelling of LD could be extended to estimate other LD related parameters such as recombination rates.

Extensive genotyping of individuals for tens to hundreds of thousands of SNP markers is becoming common as automated high throughput techniques are established (Wang et al., 2005). This detailed genotype data provides important information about linkage disequilibrium (LD) between the genes or markers. LD, in turn, can be used to test hypotheses about the evolutionary history of the population (Hayes et al., 2003), to map QTL (Carlson et al., 2004) and to estimate the recombination rate at each position along a chromosome (Li and Stephens, 2003). Thus, extracting maximum information from the pattern of LD observed is very important.

LD, as pointed out by Chapman and Thompson (2003), occurs because multiple gametes inherit a chromosome segment from a common ancestor that is IBD ie. inherited without any recombination. Hayes et al. (2003) used this understanding of LD to define a measure of LD called chromosome segment homozygosity or CSH as the probability that random chromosome segments sampled from a population are IBD.

According to coalescence theory, the mutations at all loci (markers and QTL) are independent given the coalescence tree(s), i.e. given the IBD structure of the haplotypes (Hudson, 1993). Hence, markers only provide information about QTL alleles through their information on the underlying coalescence tree or IBD structure. Therefore the logical approach to using LD is to use the markers to infer the coalescent tree or properties of it and then to use the coalescent tree to infer properties of the population (e.g. effective size), or to map QTL or to discover recombination hotspots. In line with this approach, all QTL mapping methods can be described conceptually as following three steps: (1) Calculate the probability G_{ij} that two individuals carry chromosomes that are identical by descent at the putative QTL position. (2)

Compare the similarity in phenotype to G_{ij}. (3) The position of the QTL that maximizes the likelihood of the phenotypes given G_{ij} is the estimated position. Linkage mapping of QTL also follows this approach, except that G_{ij} is here solely due to within family IBD.

Unfortunately, a full coalescent analysis of many linked markers is not computationally feasible. Early multipoint IBD estimation methods assumed that the markers provided independent information about the IBD status (e.g. Terwilliger, 1995). In the case of dense SNP markers this assumption is clearly invalid. More recent multipoint methods approximate the coalescence, for instance composite likelihood methods consider markers only two at a time (Hudson, 2001). An alternative approach (Meuwissen and Goddard, 2001) is to model the coalescence of all markers but only a pair of chromosomes at a time, which fits with the definition of CSH, and is computationally much more tractable (can deal with 1000s of markers). This approach has been very useful for mapping QTL (e.g. Olsen et al., 2005) but it assumed that genes in the current population derived, without any mutation, from a base population that contained a single copy of the QTL mutant a known number of generations ago. This is satisfactory provided the mutation rate is negligible relative to the recombination rate, but as the density of markers increases, this assumption is less justified. Furthermore, they assumed that the effective population size and time since the most recent QTL mutation were known. Here, we will extend their approach to overcome these limitations.

We will present a method that predicts IBD probabilities at putative QTL positions using information from many dense markers. As part of the predictions, the effective population size will be estimated directly from the linked marker data. Predictions will be compared to true IBD states, and to direct association analyses using single markers and marker haplotypes. The approach is general and can be extended to effective population sizes that varied in the past, and to estimate recombination rates that vary at different points in the genome.

The Model

For each pair of gametes, *i* and *j*, define a vector ($\boldsymbol{y}_{ij}$) summarising the observed haplotypes where y_{ijk} = 1 if the alleles are alike-in-state at position *k* and 0 if they are not, where *k*=1,..,*l* and *l* denotes the number of marker loci. To simplify the notation, for the moment, we will suppress the *ij* subscript in what follows. That is, all data and parameters are defined for the pair of gametes *i* and *j*. The parameters for pair *i* and *j* will be connected with the parameters for other pairs in a hierarchical model. Underlying this observed $\boldsymbol{y}$, is a pattern of IBD relationships described by the vector $\boldsymbol{\pi}$. If the *b*th marker bracket is IBD, i.e. inherited as an IBD chromosome segment from a common ancestor without any recombination, the *b*th element of $\boldsymbol{\pi}$ is 1, and otherwise it is 0, where a marker bracket denotes the chromosome segment between two adjacent markers (including the marker positions themselves). E.g., the vector $\boldsymbol{\pi}$ = [1 1 1 0]' denotes that the first three brackets are inherited IBD from a common ancestor and the fourth bracket is not entirely inherited from one common ancestor. The probability of observing $\boldsymbol{y}$ is thus:

$$P(\boldsymbol{y}) = \Sigma_{\text{all } \boldsymbol{\pi}} \; P(\boldsymbol{y}|\boldsymbol{\pi})*P(\boldsymbol{\pi}), \qquad [1]$$

where the summation is over all possible $\boldsymbol{\pi}$ vectors, P($\boldsymbol{y}$ | $\boldsymbol{\pi}$) states the conditional probability of observing $\boldsymbol{y}$ given the pattern of IBD and nonIBD segments denoted by $\boldsymbol{\pi}$, and P($\boldsymbol{\pi}$) is the prior probability of observing this pattern of IBD and nonIBD segments. P($\boldsymbol{\pi}$) can be factored because, once a recombination occurs, chromosomes segments on either side of the recombination are assumed to evolve independently in coalescence theory. The latter assumes an unstructured population, i.e. no subpopulations or lineages. Therefore, we group elements in $\boldsymbol{\pi}$ into IBD segments (continuous sequences of 1's) and others. Eg., $\boldsymbol{\pi}$ = [0 1 1 0 1 0]' consists of two IBD segments, brackets 2, 3, and bracket 5. Therefore,

$$P(\boldsymbol{\pi}=[0\ 1\ 1\ 0\ 1\ 0]') = P(\boldsymbol{\pi}=[0\ 1\ 1\ 0\ .\ .]') * P(\boldsymbol{\pi}=[.\ .\ .\ 0\ 1\ 0]' \mid \boldsymbol{\pi}=[.\ .\ .\ 0\ .\ .]'),$$

where a dot [.] denotes that the IBD status for this bracket is not specified, i.e. it is not accounted for in the probability calculation, and could be either 0 or 1. This allows the probability of a long sequence of data on a chromosome to be factored into manageable pieces. The prior probability of an IBD chromosome segment which extends over *n* marker brackets is approximately (Hayes et al., 2003):

$$P(\boldsymbol{\pi}=\mathbf{1}_n) = 1 / (4Nc + 1), \qquad [2]$$

where $\mathbf{1}_n$ is a vector of *n* ones, *c* is the size of the IBD segment in Morgans, and *N* is the effective population size. The approximation in [2] assumes that the size of the segment, *c*, is small relative to 1. Terms like the above P($\boldsymbol{\pi}$=[0 1 1 0]'), where the IBD segment is bounded by nonIBD segments, can be rewritten involving only unbounded IBD segments as in Equation [2] (Meuwissen and Goddard, 2001):

$$P(\boldsymbol{\pi}=[0\ 1\ 1\ 0]') = P(\boldsymbol{\pi}=[.\ 1\ 1\ .]) - P(\boldsymbol{\pi}=[1\ 1\ 1\ .]') - P(\boldsymbol{\pi}=[.\ 1\ 1\ 1]') + P(\boldsymbol{\pi}=[1\ 1\ 1\ 1]'),$$

which follows from rearranging the equations:

$$P(\boldsymbol{\pi}=[\mathbf{.}\ 1\ 1\ \mathbf{.}]')=P(\boldsymbol{\pi}=[0\ 1\ 1\ 0]')+P(\boldsymbol{\pi}=[1\ 1\ 1\ 0]')+P(\boldsymbol{\pi}=[0\ 1\ 1\ 1]')+P(\boldsymbol{\pi}=[1\ 1\ 1\ 0]')$$

and $P(\boldsymbol{\pi}=[1\ 1\ 1\ \mathbf{.}]')=P(\boldsymbol{\pi}=[1\ 1\ 1\ 0]')+P(\boldsymbol{\pi}=[1\ 1\ 1\ 1]')$.

Also, the conditional probability $P(\boldsymbol{\pi}=[0\ 1\ 0]' \mid \boldsymbol{\pi}=[0\ \mathbf{.}\ \mathbf{.}]')$ can be rewritten in terms of unbounded probabilities of IBD segments as in Equation [2] using:

$$\begin{aligned}P(\boldsymbol{\pi}=[0\ 1\ 0]' \mid \boldsymbol{\pi}=[0\ \mathbf{.}\ \mathbf{.}]') &= P(\boldsymbol{\pi}=[0\ 1\ 0]')\ /\ P(\boldsymbol{\pi}=[0\ \mathbf{.}\ \mathbf{.}]')\\ &= P(\boldsymbol{\pi}=[0\ 1\ 0]')\ /\ [1-P(\boldsymbol{\pi}=[1\ \mathbf{.}\ \mathbf{.}]')].\end{aligned}$$

Thus, all these terms can be calculated by using [2] repeatedly. A factorisation to computationally speed up the summation in Equation [1] is described by Meuwissen and Goddard (2001).

Conditional Marker Homozygosity

Let $P(y_k \mid \boldsymbol{\pi})$ denote the probability that marker locus *k* is observed alike-in-state at a pair of gametes *i* and *j*, or not, given the pattern of IBD and nonIBD segments, $\boldsymbol{\pi}$. For instance, if $\boldsymbol{\pi}$ denotes that marker locus *k* is on an IBD segment of size *c*, then locus *k* is alike-in-state if there was no mutation before the gametes coalesce into their common ancestor (looking back in time). If the next locus, *k*+1, is on another IBD segment, its evolution is assumed independent in coalescence theory, as mentioned before. If locus *k* and *k*+1 are on the same IBD segment, locus *k*+1 is alike-in-state if no mutation occurred at locus *k*+1. The probability of a mutation at locus *k* and locus *k*+1 are independent given the coalescence tree, i.e. given the time when the common ancestor occurs. When the common ancestor occurs is unknown, but the vector $\boldsymbol{\pi}$ contains information on the size of the segment, c, which yields a prediction of the time since the common ancestor, i.e. the gametes are expected to have coalesced $1/(2c)$ generations ago (Hayes et al., 2003). We will assume that conditional on the size of the IBD segment, the probability of a mutation at any two (or more) loci on this IBD segment are independent, i.e. we assume that any remaining auto-correlation between the mutation probabilities at adjacent marker loci will have a negligible effect on our predictions. Thus, conditional on the pattern of IBD and nonIBD segments, $\boldsymbol{\pi}$, the (non) alike-in-state probabilities of the marker alleles are independent, i.e.:

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{\pi}) = \Pi_{\mathrm{k}}\, \mathrm{P}(y_k \mid \boldsymbol{\pi}).$$

The conditional alike-in-state probabilities $\mathrm{P}(y_k \mid \boldsymbol{\pi})$ are given in Table 1, and derived in the Appendix. The probability of a mutation is smaller for larger segments, since larger segments coalesce earlier giving less time for a mutation. If the marker is on a segment that recombined, the common ancestor was probably more distant in the past, because the segment had time to recombine, and the probability of a mutation is increased. The values of $\mathrm{P}(y_k \mid \boldsymbol{\pi})$ depend on the mutation rate, u, or more precisely on $4Nu$. Fortunately the method is not very sensitive to the value of u used because the probability that two alleles are not alike is proportional to $4Nu$ in many situations (see table 1), and therefore we use an artificially high value of u ie 10^{-5}. An alternative is to estimate $4Nu$ by the marker heterozygosity of the k-th marker, H_k. Use of H_k may also account for differences in information content between the markers (e.g. SNPs vs. microsatellites). However, since markers are only used if they are polymorphic, marker heterozygosity does not correctly predict $4Nu$. We will call the methods using mutation rate IBDMUT and the method using marker heterozygosity IBDHET.

Estimation of IBD at a putative QTL position

The above model is adapted to estimate IBD between the chromosomes at any position, which will be called ‘the putative QTL position’, q, here. Firstly, let y_q^*=1 denote that the QTL position is IBD, and the ‘*’ denotes that this is an auxiliary record, which is not actually observed but should be accounted for when evaluating the probability of the pair of gametes. Secondly, using Equation [1] we calculate P(y_q^*=1, $\boldsymbol{y}$), i.e. the probability that the QTL is on an IBD segment, and the marker records $\boldsymbol{y}$ occur. At the QTL position, q, we do not use Table 1 to obtain P($y_q^*|\boldsymbol{\pi}$) values, but we set P($y_q^*=1|\boldsymbol{\pi}$)=1, if $\boldsymbol{\pi}$ indicates that the QTL is on a IBD segment, and P($y_q^*=1|\boldsymbol{\pi}$)=0 if q is surrounded by 2 nonIBD segments. Note, that if $\boldsymbol{\pi}$ indicates that there is an IBD segment to the left of the QTL and a nonIBD segment to the right, the QTL is still on an IBD segment since the marker brackets are defined to include the loci that border them (thus the IBD segment to the right was partly but not entirely IBD, which is denoted by a 0 in $\boldsymbol{\pi}$). Thirdly, we calculate the probability that the putative QTL position is IBD given the marker data at the chromosome as:

$$\mathrm{P}(y_q^*=1|\,\boldsymbol{y}) = \mathrm{P}(y_q^*=1, \boldsymbol{y}) \,/\, \mathrm{P}(\boldsymbol{y}) \qquad [3]$$

where P($\boldsymbol{y}$) is calculated using equation [1] without considering the QTL locus (as in the previous sections).

Estimation of Effective Population Size

It is well known from coalescence theory (Hudson, 2001) that it is possible to estimate the product N*c, but not N separately. We will assume here that c is known, which enables us to estimate N. If c is unknown, but the relative distances between the markers are known (from their physical map positions), we may scale the marker distances such that they sum to one, and the method described below will estimate *Nc* instead of *N*, where *c* is the size of the chromosome.

We use the EM algorithm (Dempster et al., 1977) to estimate *N*, where the IBD status between each pair of markers *k* and *l* ($k<l$), is considered as missing data. Assuming a starting value for *N*, the steps are:

(1) Estimate the probability that the entire segment between markers *k* and *l* is IBD, which is denoted by a vector $\boldsymbol{y}_{k,l}^{*}$ being the unity vector, i.e. $\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{\boldsymbol{n}}$, where $n=l-k$. We calculate for each gamete pair and for each pair of markers *k,l* the probability that the entire segment between markers *k* and *l* is IBD, $\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{\boldsymbol{n}}$, given the marker data and the current estimate of *N*, similar to Equation [3]:

$$\mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{n}|\boldsymbol{y}) = \mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{\boldsymbol{n}}, \boldsymbol{y}) / \mathrm{P}(\boldsymbol{y})$$

where $\mathrm{P}(\boldsymbol{y})$ is obtained from Equation [1], and $\mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{\boldsymbol{n}}, \boldsymbol{y})$ is also obtained from [1], using $\mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{n}|\boldsymbol{\pi})=1$ if the segment between marker *k* and *l* as denoted by $\boldsymbol{\pi}$ is entirely IBD, and $\mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{n}|\boldsymbol{\pi})=0$ otherwise. The probabilities $\mathrm{P}(\boldsymbol{y}_{k,l}^{*}=\mathbf{1}_{n}|\boldsymbol{y})$ are averaged over all chromosome pairs to get population wide estimates for them, which will be denoted by $CSH_{k,l}$.

(2) Non-linear regression is used to obtain an updated estimate of *N*, using the statistical model:

$$CSH_{k,l} = 1/(4N^{*}c_{k,l}+1) + e_{k,l} \qquad [4]$$

where $c_{k,l}$ is the distance between the markers *k* and *l* (in Morgans), and $e_{k,l}$ is a random sampling error, whose variance is assumed constant. If the updated estimate of *N* deviates less than a factor 0.001 from the old value, the algorithm stops, otherwise go to step (1) and keep on iterating. The estimation of N was only adopted in combination with the IBDHET method, and the combined method is called IBDHETne.

Computer simulations to test the model

The ms program (Hudson, 2002) was used to generate SNP marker data, which uses the standard coalescence approach (Hudson, 1993) in which the random genealogy of a sample is first generated assuming a neutral model of inheritance, and then mutations are randomly placed on the genealogy. An infinite-sites model of mutation is assumed, and a finite sites model of recombination, although this number of sites was very large here (2,000,000 base pairs). The size of the simulated segment was 2 cM, $N = 1000$, the per base pair mutation rate was 10^{-8}, and 200 haplotypes were simulated. Only markers with a Minor Allele Frequency, MAF > 0.1, were retained, and the 20 most equidistant markers were used to span the 2 cM region. Since the ms simulations resulted in an abundance of markers, the 20 markers were close to equidistant. Another marker with MAF > 0.1, and which was as close as possible to the midpoint of the segment was appointed to act as the QTL. The QTL was thus approximately in the middle between markers 10 and 11. A phenotypic record was simulated for each haplotype using the equation: $p_i = Q_i + e_i$, where Q_i is the allele at the QTL position (0 or 1) and e_i is an environmental effect sampled from N(0,0.5). Compared to real life situations, the QTL effect of 1 was perhaps large relative to the environmental variance, but most real life QTL mapping experiments in outbreeding populations collect more than 200 phenotypes. We kept the number of phenotypes relatively small in order to be able to analyse 100 replicated data sets in a reasonable time.

The 100 data sets were analysed using the QTL mapping by variance components approach (George et al., 2000), and using the statistical model:

$$\boldsymbol{p} = \mu * 1_{\mathbf{200}} + \boldsymbol{q} + \boldsymbol{e},$$

where μ is overall mean; $\boldsymbol{q}$ is a (200x1) vector of QTL effects (assumed random with $\boldsymbol{q} \sim \text{MVN}(\mathbf{0}, \boldsymbol{G}\,\sigma_q^2)$), and $\boldsymbol{e}$ is a (200x1) vector of environmental effects ($\boldsymbol{e} \sim \text{MVN}(0, \boldsymbol{I}\,\sigma_e^2)$; $\boldsymbol{I}$ = identity matrix). The (200x200) matrix $\boldsymbol{G}$ contains the IBD probabilities estimated at the putative QTL position using the above methods, and the variance components σ_q^2 and σ_e^2 were estimated by Residual Maximum Likelihood (REML) using the computer package ASREML (Gilmour et al, 2000). ASREML also computes the REML likelihood of the data given the $\boldsymbol{G}$ matrix, and a likelihood ratio test-statistic (LRT) was calculated as twice the difference in log-likelihood between the model including the QTL effect, $\boldsymbol{q}$, and the model excluding the QTL effect. This LRT was calculated for every midpoint of 19 marker brackets, i.e. assuming a putative QTL at each of the midpoints, and the analysis that gave the highest LRT was denoted the most likely QTL position.

In addition, some direct association analyses were conducted where the phenotypes were directly regressed on the marker effects (MARK1), i.e. for marker k the model is $\boldsymbol{p} = \mu * 1_{\mathbf{200}} + \boldsymbol{m_k} + \boldsymbol{e}$, where $\boldsymbol{m_k}$ is (2x1) vector of random marker effects ($\boldsymbol{m_k} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}\sigma_m^2)$). A second direct association analysis fits the effects of 2-marker-haplotypes on the phenotypes (MARK2), i.e. for the k-th marker bracket $\boldsymbol{p} = \mu * 1_{\mathbf{200}} + \boldsymbol{h_k} + \boldsymbol{e}$ and $\boldsymbol{h_k}$ is the effect of the haplotype constituted by markers k and k+1 ($\boldsymbol{h_k} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}\sigma_h^2)$). In these analyses $\boldsymbol{m_k}$ and $\boldsymbol{h_k}$ was included as a random effect in order to compare their LRTs to that of the variance component QTL analysis. Including $\boldsymbol{m_k}$ and $\boldsymbol{h_k}$ as random effects also has a Bayesian interpretation in that they are assumed to have a MVN prior distribution. It is well known from Bayesian statistics, that the influence of the prior is small if there is a lot of information in the data, which is the case here since there are relatively many records to estimate the 2 effects of $\boldsymbol{m_k}$ and the 4 effects of $\boldsymbol{h_k}$. Thus, an analysis that treats $\boldsymbol{m_k}$ and $\boldsymbol{h_k}$ as random effects, a Bayesian analysis with MVN priors, and a conventional analysis that treats $\boldsymbol{m_k}$ and $\boldsymbol{h_k}$ as fixed effects (no use of prior information) are all expected to give very similar results.

RESULTS

Figure 1 shows the average LRT profiles for all methods, except IBDHETne (because the LRT profile of IBDHETne was almost indistinguishable from that of IBDHET). All the LRT curves are nicely centred around the true QTL position at the middle of the chromosome segment, i.e. there was no bias in the position estimates. IBDMUT and IBDHET had the highest LRT at the QTL position, and thus had approximately equal power to detect the QTL. However, the LRT drops more quickly for the IBDMUT analysis than for the IBDHET analysis when moving away from the true QTL position, which suggests that IBDMUT has a higher precision to map the QTL. A reason for this may be that IBDHET estimates the mutation rate from the marker data, which reduces their information content, and makes them less informative for positioning the QTL. This result also suggests that accounting for differences in heterozygosity (information content) between markers does not improve mapping precision, if the true underlying mutation rate was the same for all the markers. The MARK1 and MARK2 methods are also unbiased but have substantially less power to detect the QTL.

Although, the distributions of the LRT values under the null-hypothesis of no QTL effect are expected to be approximately the same, i.e. a Chi-squared distribution with the same degrees of freedom, we tested this assumption by analysing 100 replicated data sets where no QTL effect was simulated. The fifth highest LRT value from these analyses yielded an estimate of the chromosome segment wise critical significance threshold at P=0.05, and these values were 3.48, 3.03, 3.34, and 3.34 for MARK1, MARK2, IBDMUT and IBDHET, respectively. Hence, the critical significance thresholds of the 4 methods are similar, and the differences in LRT values in Figure 1 do translate into differences in power of detecting the QTL.

Table 2 shows the power of detecting the QTL and mapping precision of the methods expressed as the mean of the squared difference between estimated and true position. As expected from Figure 1, IBDMUT has the best precision, whilst IBDHET, IBDHETne, and MARK2 have similar precision. Power was calculated as the fraction of the replicates where LRT exceeded 11, which corresponds to a nominal P-value of .001 assuming that LRT is approximately chi-squared distributed with 1 degree of freedom. The IBD based methods all had substantially more power than direct regression on marker (haplotypes).

For the analyses where a significant QTL was detected (LRT>11), a 2-LOD-dropoff support interval was constructed for the position of the QTL, i.e. the interval surrounding the QTL peak where the likelihood exceeds $\text{LogLik}_{\text{max}} - 2*\ln(10)$, where $\text{LogLik}_{\text{max}}$ is the natural logarithm of the maximum likelihood. If the log likelihood was quadratic in the QTL position, this support interval is expected to contain the QTL in approximately 99.8% of the cases. Visscher and Goddard (2004) show that the log likelihood is not quadratic, and hence we expect somewhat fewer than 99.8% of the estimates to lie within this interval. The number of replicates in which the true QTL position was within the 2-LOD-dropoff support interval were counted (Table 2). The number of cases where the QTL was contained in the support interval of the IBD based analyses was about 3% less than expected. The average size of the support interval shows again that IBDMUT is more precise than IBDHET and IBDHETne. For the methods based on direct regression on the marker or haplotype, the 2-LOD-droffoff interval does not seem to provide reliable support intervals. The intervals seem much too short, such that the fraction of cases where the true QTL is within the interval is very low. These too short intervals are probably due to the spiky LRT profiles that are obtained by these analyses. Figure 2 shows an example replicate where all analyses erroneously place the QTL at around 0.6 cM, but the correct position of 1.0 cM is well within the LOD-2-support interval of the IBD based methods, whereas the spikiness of MARK1 and MARK2 makes their LRT signal drop quickly and rise again in an irregular manner. In case of MARK2, one might even conclude that there are 2 QTL in this region.

Another criterion by which to judge the different methods of analysis is the correlation between the IBD probability calculated from the markers and the 'true' probability that the QTL alleles are IBD. The true probability depends on the length of the coalescence tree joining the QTL alleles in the two gametes (τ) and this is provided by the ms simulation program of Hudson (2002). This allowed us to calculate the probability of QTL alleles being IBD, i.e. no mutation since coalescence, given the coalescence tree as P(IBD|tree) = exp(-2θ*τ), where $\theta = 4Nu_{QTL}$, with u_{QTL} being the mutation rate at the QTL position, and τ is time till the two haplotypes coalesce. Choice of θ is somewhat arbitrary since it does not change the ordering of the probabilities and so has only a small effect on the correlation with the predicted IBD probability. We assumed θ=2, since this seemed to give a spread in IBD probabilities. These P(IBD|tree) were considered as the 'gold standard', since they are based on the true simulated length of the tree which, of course, is not known in a real life situation, but is estimated using the information of linked markers. The IBD probabilities given the tree were correlated with the estimated IBD probabilities (Table 2), and the correlations mirror the differences in power between the methods. That is, the IBD methods have higher correlations and higher power to detect the QTL than the methods that use regression on the marker. The precision with which the methods position the QTL is probably more affected by how quickly the IBD probabilities change when moving from one position to the next than by the correlation between P(IBD|tree) and estimated IBD probabilities.

The estimate of *N* obtained by IBDHETne was on average 1048 with a standard deviation of 289. A histogram of the estimates is shown in Figure 3. It shows that, although the distribution of the estimates of *N* is centred around 1000, occasional estimates can be quite far away from *N*=1000. However, in most applications *N* is multiplied by another entity, e.g. mutation or recombination rate, which implies that the relative error of the estimate of *N* is more important than its absolute error. Also, when judging the effective size of a population, the relative error is more important than the absolute error (e.g. 100 vs. 200 is an important difference whereas 1000 vs. 1100 is not). The relative error is obtained by transforming *N* to the log-scale (log-10-base was used here), which gave an average of the estimates of $\log_{10}(N)$ of 3.00 with a standard deviation of 0.123. Using the normal distribution as an approximation, this implies that approximately 60-70% of the estimates have a relative error less than 33% (=($10^{0.123}$ -1)*100%).

DISCUSSION

This paper describes a novel method that predicts IBD probabilities between pairs of haplotypes at predefined positions based on the similarity of the marker alleles carried by the haplotypes. It is a coalescence based method but differs from other coalescence methods in that it is computationally feasible for many linked loci, but only considers a pair of gametes at a time. The method assumes that the haplotypes are known, i.e. that the genotypes have been phased. In situations with high marker density and/or large family sizes estimation of phase is quite accurate. If the phase of a marker is uncertain in haplotype *i*, this marker may be denoted as missing which implies that it is skipped in all $\boldsymbol{y}_{ij}$ vectors involving haplotype *i*. The method extends our previous method (Meuwissen and Goddard, 2001) by allowing for mutation at the markers and by requiring no assumptions about the effective population size or the number of generations since a 'base' population. The method can be used for many purposes that require an analysis of LD because it models the process that causes LD i.e. the inheritance of chromosome segments without recombination from a common ancestor. For instance, the method can be used to map QTL, to estimate effective population sizes and, could be extended, to estimate recombination rate.

A commonly used strategy for QTL mapping is to perform a whole genome linkage or association analysis, followed by fine mapping using association methods. However, both linkage and association mapping alone have limited power to detect QTL. Linkage mapping does not use the increased power due to association, and genome-wide association mapping suffers from multiple testing problems and false positive results (Carlson et al., 2004). The genome wide combined use of linkage and association mapping is expected to relieve these problems, because it combines both sources of information. The method presented is easy to extend to combined linkage and linkage disequilibrium (LLD) mapping by: (i) applying the described method to estimate IBD probabilities between the founder haplotypes of the genotyped pedigree; and (ii) use linkage analysis information to estimate within family based IBD probabilities between founder and offspring haplotypes and among offspring haplotypes (e.g. Meuwissen et al., 2002, Perez-Enciso, 2003). Prediction of IBD based on both LD and linkage information, as described by (i) and (ii), is a convenient way to combine both information sources for the mapping of QTL, and is expected to detect more QTL and map them more precisely than LD or linkage analysis alone would do.

Although the described methodology does not directly estimate coalescence trees and times, it is based on deterministic approximations of the coalescence process. Since the method does not build a coalescence tree for all the haplotypes, it neglects information from other haplotypes when estimating IBD probabilities for the haplotype pair *i,j*. This shortcoming is expected to become less important as marker density increases, since, at high density, the markers will be sufficiently informative to directly indicate IBD regions when comparing pairs of haplotypes. Furthermore, multi-haplotype identities are also ignored by QTL mapping by variance components methods, since it only uses the IBD matrix of the haplotype pairs, **G**, to position the QTL (George et al., 2000). An alternative approach to ours is to leave the coalescent-with-recombination model and calculate an ensemble of 'likely' Ancestral Recombination Graphs using Minichiello and Durbin's (2006) approach, which can handle hundreds of markers simultaneously. Further research is needed to compare this approach to ours, with respect to power, mapping precision and computational requirements, but since the method of Minichiello and Durbin does not require an estimate of the effective population size, it will not be able to estimate this parameter from the data.

Compared to direct association methods, i.e. that directly regress the phenotypes onto the marker (haplotype), the presented methods seem to have a higher signal to noise ratio in that there is a higher power of detecting the QTL, and the LRT profiles are more smooth (Figure 2). Since the regression methods use the same marker information, their peak LRT is often at a similar position. Because two locus LD is very variable (Hill and Weir, 1994), the LRT profile of MARK1 is much more erratic than that of multi-point IBD methods presented here, which use the LD with all markers simultaneously. This is, however, at the cost of a much more complicated QTL mapping methodology.

The method presented can also be used to estimate effective population size, *N*, or in case recombination rates, *c*, are unknown $\rho = 4Nc$. We obtained an estimate of *N* that was almost unbiased, and all 100 estimates of *N* are within a factor of 2 from the true value. Perhaps the best competitor to our estimator of ρ is that of Li and Stephens (2003) which was within a factor of 2 of the truth in 68% of the replicates and was biased. As Li and Stephens (2003) remark, a factor 2 of the truth may not sound very impressive in many statistical applications, but in this setting this accuracy is hard to achieve (Wall, 2000). Hence, the current method is competitive to all other methods that estimate *N* or ρ from linked marker data, which makes it attractive to extend the methodology to estimate variations in recombination rates and recombination hotspots. In a follow up paper we will extend the nonlinear model [3] to a model that estimates the recombination rate per marker pair as a function of the distance between the markers in kilobases.

In summary, we believe this method is a useful alternative to other coalescence based methods for analysing data on many dense polymorphic loci, because (i) it can handle large numbers of closely linked markers; (ii) at high marker density, its estimates of IBD probabilities are expected to be similar to those of methods that account for all marker haplotypes simultaneously, and (iii) it can be used to estimate parameters that affect LD since it is based on the modelling of the process that generates LD.

Literature Cited

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004). Mapping complex disease loci in whole-genome association studies.

Chapman NH, Thompson EA (2003) A model for the length of tracts of identity by descent in finite random mating populations. Theor Popul Biol. 64:141-50.

Dempster, AP, Laird NM & Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 39:1–38

George AW, Visscher PM, Haley CS (2000) Mapping Quantitative Trait Loci in Complex Pedigrees: A Two-Step Variance Component Approach. *Genetics* 156: 2081-2092.

Gilmour, AR, BR Cullis, S J Welham and R Thompson, 2000. ASREML reference manual. ftp.res.bbsrc.ac.uk/pub/aar.

Hayes BJ, Visscher PM, McPartlan HC, Goddard ME, (2003) Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. 4:635-43.

Hill WG, Weir BS (1994) Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet. 54:705-714.

Hudson RR (1993) The how and why of generating gene genealogies. In: Mechanisms of Molecular Evolution, N Takahata and AG Clark, eds, Japan Scientific Societies Press, Tokyo, pp. 23-36.

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159: 1805-1817.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. Biometrics 18: 337-338.

Li N, Stephens M (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data.
Genetics. 165:2213-33

Meuwissen THE, Goddard ME (2001) Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol. 33:605-634

Meuwissen THE, Karlsen A, Lien S, Olsaker I, Goddard ME (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373-379.

Minichiello MJ, Durbin R (2006). Mapping trait loci by use of inferred Ancestral Recombination Graphs. Am. J. Hum. Genet. 79:910–922.

Olsen HG, Lien S, Gautier M, Nilsen H, Roseth A, Berg PR, Sundsaasen KK, Svendsen M, Meuwissen THE (2005) Mapping of a milk production quantitative trait locus to a 420-kb region on bovine chromosome 6. Genetics169:275-283.

Perez-Enciso M (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163:1497-510.

Terwilliger JD (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet. 56:777-87.

Visscher PM, Goddard ME (2004) Prediction of the confidence interval of quantitative trait Loci location. Behav Genet. 34:477-482.

Wall JD (2000) A comparison of estimators of the population recombination rate. Mol. Biol. Evol. 17: 156-163.

Wang WYS, Baratt BJ, Clayton DG, Todd JA (2005). Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. 6: 109-118.

Table 1. Probability of (un)equal marker alleles given the IBD pattern.

Probability[1]	Value[2]	Appendix Equation
$P(y_k=0 \mid \boldsymbol{\pi}=\mathbf{1}_n)$ [3]	$\frac{4Nu}{4Nc+1}$	[A1]
$P(y_k=0 \mid \boldsymbol{\pi}=[0 \mid 0]')$ [4]	$\frac{4Nu(3+24Nc+32N^2c^2)}{(8Nc+1)(4Nc+1)}$	[A3]
$P(y_1=0 \mid \boldsymbol{\pi}=[\ \mid 0]')$ [5]	$\frac{4Nu(2+4Nc)}{4Nc+1}$	[A2]
$P(y_k=1 \mid \boldsymbol{\pi})$	$1- P(y_k=0 \mid \boldsymbol{\pi})$	

[1] Where possible, the position of the marker *k* is denoted by '|' within the vector of IBD and nonIBD segments $\boldsymbol{\pi}$.

[2] *N* = effective population size; *u* = mutation rate. Note: 4N*u* may be approximated by marker heterozygosity, H_k, if *u* is unknown.

[3] *k* is on an IBD segment of size *c* (note *k* may be right at the edge of this segment).

[4] *k* is in between two nonIBD segments, each of size *c*.

[5] The marker is at the start of the chromosome, and next to a nonIBD region of size *c*. Due to symmetry, the same Equation applies at the end of the chromosome.

Table 2. The mean square error of the position estimate (MSE), the fraction of replicates with a QTL significant effect (Power), the fraction of the significant QTL, where the true QTL position is within the 2-LOD-dropoff support interval (P(QTL in Interval)), correlation between estimated IBD probabilities and the IBD probability given the tree (Corr)[1].

Analysis	MSE	Power[2]	P(QTL in Interval)	Size of Interval	Corr
IBDHET	0.072 cM^2	0.94	0.97	0.82 cM	0.576
IBDMUT	0.049	0.95	0.97	0.58	0.537
IBDHETne	0.071	0.94	0.97	0.78	0.576
MARK1	0.088	0.83	0.11	0.11	0.357
MARK2	0.074	0.88	0.26	0.14	0.478

[1] The IBD probability given the tree is used as the ‘gold standard’ (see main text).

[2] A nominal P value of 0.001 was used.

Figure 1. Average LRT profiles.

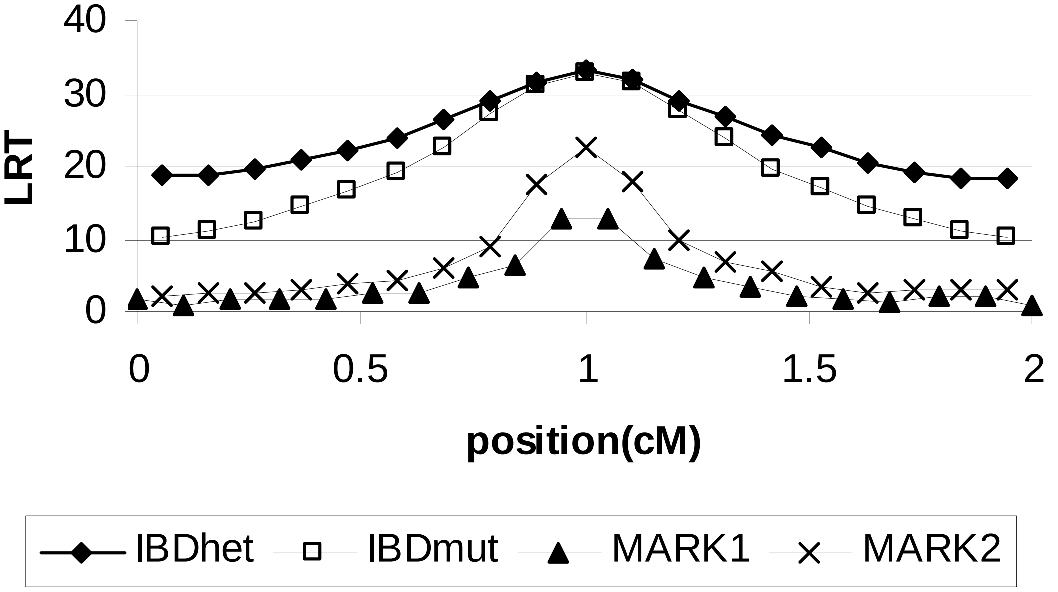


Figure 2. Example of LRT profiles in a single replicate.

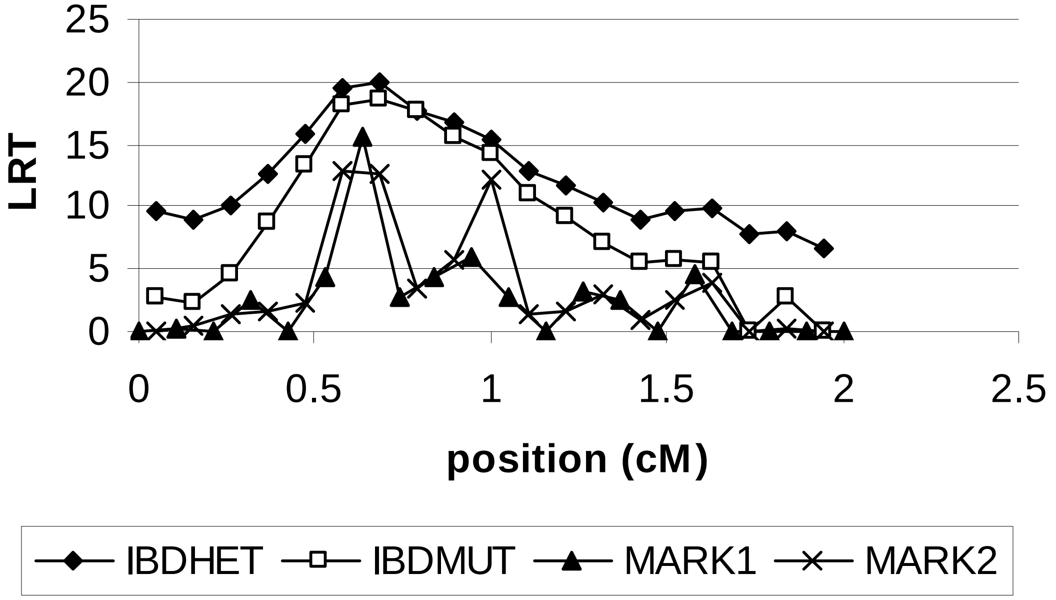
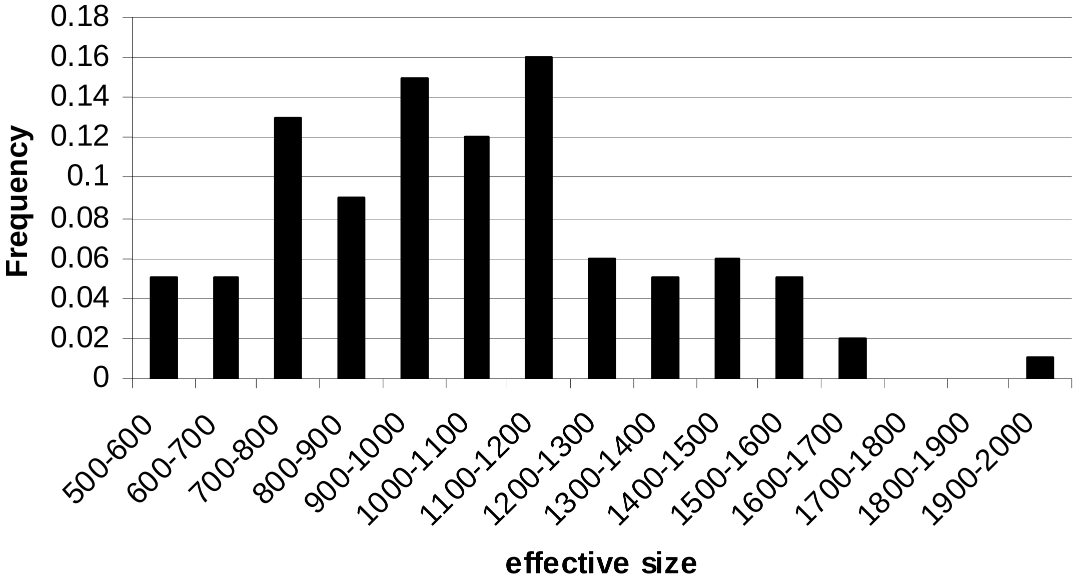


Figure 3. Histogram of estimates of the effective population size.



Appendix: Probability of marker homozygosity conditional on the IBD pattern.

This appendix derives the probabilities of mutation occurring at a locus given in table 1 of the main text. In a genealogical tree, eventually, all lineages coalesce to the most recent common ancestor. Whether a haplotype pair has identical marker alleles (homozygosity) or not depends on whether there was a mutation prior to the coalescence at the marker locus or not. Note: in the following we are always looking back in time, so if the mutation occurred 10 generations ago and the coalescence 20, the mutation was prior to the coalescence. We will consider two 'events' (E) namely 'recombination' (R) and 'coalescence' (C) and calculate probabilities of a 'mutation' relative to when these events occur, and $E(t)$, $R(t)$, and $C(t)$ specify also the time of the event. If there was no mutation will be denoted by M=0, no mutation prior to the occurrence of an event will be denoted by $M_{<E}$=0, and no mutation after the event by $M_{>E}$=0.

The derivation uses the following probabilities for looking back in time one generation for two gametes:

P(no mutation in either gamete) = P($M(1)$=0) = $(1\text{-}u)^2$
P(coalescence occurs) = P($E(1)$=$C(1)$) = 1/(2N)
P(recombination occurs in one of the gametes) = P(E(1)=R(1)) = 1 - $(1\text{-}c)^2 \approx$ 2c
P(coalescence occurs | E occurs) = P(E=C | E) = 1/(2N) / (1/(2N) + 2c) = 1/(1+4Nc)
P(E(1)=0) = P(no coalescence and no recombination) = (1-1/(2N)) * $(1\text{-}c)^2 = \lambda$
$\approx$ 1-1/(2N) – 2c
P(M(1)=0 & E(1)=0) = $(1\text{-}u)^2 * \lambda = \alpha$
where u is the mutation rate, c is the size of the segment (in Morgans), and N is the effective population size.

Therefore the probability of the first event occurring in generation t and no mutation prior to then is the probability of t-1 generations with no event and no mutation times the probability of an event in generation t. That is

$$P(E(t)\ \&\ M_{<E}=0) = \alpha^{t-1} * (1-\lambda)$$

and summation over all generations t gives the probability of no mutation prior to the event:

$$P(E\ \&\ M_{<E}=0) = P(M_{<E}=0) = \Sigma_t\ \alpha^{t-1} * (1-\lambda)$$
$$= (1-\lambda)/(1-\alpha) = [4Nc+1] / [4Nc+4Nu+1],$$

where the first equality is because an event will inevitably occur if we look an infinite number of generations back in time.

There are three conditions under which the probability of a mutation are required:

1. **The locus is located on a chromosome segment of size *c* Morgans that is IBD**

In the notation of the main text, this is P(y_k | $\boldsymbol{\pi}$=$\mathbf{1_n}$). That is, the first event that occurs to this chromosome segment in gametes i and j is a coalescence (E=C). Therefore the probability of no mutation conditional on the chromosome segments coalescing is, in our notation, P($M_{<E}$=0 | E=C) = P($M_{<E}$ =0 & E=C) / P(E=C)

And

P($M_{<E}$=0 & E=C) = Σ_t P($E(t)$ = $C(t)$ & $M_{<E}$=0)
= Σ_t P($E(t)$ & $M_{<E}$=0)* P($E(t)$=$C(t)$ | $E(t)$)
= P(E=C | E) Σ_t P($E(t)$ & $M_{<E}$=0)

because P($E(t)$=$C(t)$ | $E(t)$) is independent of t. It follows that:

P($M_{<E}$=0 & E=C) = 1/(1+4Nc) * Σ_t α^{t-1} * (1-λ)
= 1/(1+4Nc) * (1+4Nc)/ (1+4Nu + 4Nc)
= 1/(1+4Nu + 4Nc)

And since P(E=C) = 1/(1+4Nc),
P($M_{<E}$=0 | E=C) = (1+4Nc) / (1+ 4Nu+4Nc).

The probability of non-alike-in-state markers on a segment that coalesces is:

P($M_{<E}$=1 | E=C) = 1 – P($M_{<E}$=0 | E=C)
= 4Nu / (1+ 4Nu + 4Nc) ≈ 4Nu / (1+4Nc), [A1]

assuming 4Nu is small.

2. **The locus is located at the end of a chromosome segment of size c Morgans and there is a recombination in the segment to the right (or left) of the locus.**

In the notation of the main text, this is P(y_k | $\boldsymbol{\pi}$= [0]). That is, the first event that occurs to this segment is a recombination (E=R). In the notation of the appendix, the probability that there is no mutation conditional on a recombination is

P(M=0 | E=R) = P(M=0 & E=R) / P(E=R).

But, in case the event is a recombination, we must consider that a mutation can occur either before or after the event, i.e.

P(M=0 | E=R) = P($M_{<E}$=0 | E=R) * P($M_{>E}$=0 | E=R).

P($M_{<E}$=0 & E=R) = Σ_t P($E(t)$ = $R(t)$ & $M_{<E}$ =0)
= Σ_t P($E(t)$ & $M_{<E}$ =0)* P($E(t)$=$R(t)$ | $E(t)$)
= P(E=R | E) Σ_t P($E(t)$ & $M_{<E}$=0)

because P($E(t)$=$R(t)$ | $E(t)$) is independent of t. Also,

P($M_{<E}$=0 & E=R) = 4Nc/(1+4Nc) * Σ_t α^{t-1} * (1-λ)
= 4Nc/(1+4Nc) * (1+4Nc)/ (1+4Nu + 4Nc)
= 4Nc/(1+4Nu + 4Nc)

So

P($M_{<E}$ =0 | E=R) = (4Nc/(1+4Nu+4Nc) / (4Nc/(1+4Nc)) = (1+4Nc)/(1+4Nu+4Nc).

And, after a recombination, the coalescence of the two gametes is independent of what occurred before hand, so P($M_{>E}$=0 | E=R) is simply the unconditional probability of no mutation ie.

P($M_{>E}$=0 | E=R) = 1/(1+4Nu)

Therefore P(M=0 | E=R) = (1+4Nc)/(1+4Nu+4Nc) * 1/(1+4Nu).
And the probability of non-alike-in-state marker alleles given that the marker is on a segment that recombined is:

P(M=1 | E=R) = 1- P(M=0 | E=R) ≈ 4Nu(4Nc+2)/(4Nc+1) [A2]

assuming 4Nu is small, and as given in table 1 for P(y_k=1 | $\boldsymbol{\pi}$=[0]).

3. **The locus is located between two chromosome segments of size c Morgans and there are recombinations in both the segment to the right and left of the locus.**

In the notation of the main text this is P(y_k | $\boldsymbol{\pi}$=[0 0]). That is, the first event (E_1) that occurs in a recombination in one of the two segments and the second event (E_2) is a recombination in the other segment. So the probability of no mutation conditional on these two events is

$$P(M=0|E_1=R\&E_2=R)=P(M_{<E_1}=0|E_1=R)*P(M_{E_1:E_2}=0|E_1=R\&E_2=R)$$
$$*P(M_{>E_2}=0|E_2=R)$$

where $M_{E_1:E_2}=0$ denotes no mutation between events E_1 and E_2.

Now $P(M_{<E_1}=0|E_1=R)$ is very similar to the probability derived in case 2 but now the recombination can occur anywhere in a segment of size 2c Morgan, so

$$P(M_{<E_1}=0|E_1=R)=\frac{1+8Nc}{1+4Nu+8Nc}$$

In the coalescence, probabilities are not affected by what has happened previously so

$$P(M_{E_1:E_2}=0|E_1=R\&E_2=R)=P(M_{<E}=0|E=R)$$
$$=\frac{1+4Nc}{1+4Nu+4Nc}$$

as in case 2 because now the second recombination must occur within a segment of size c Morgans.

And, after the second recombination, the probability of coalescence without a mutation is:

P($M_{>E2}$=0 | E_2=R) = 1/(1+4Nu),

as in case 2.

Therefore:

$$P(M=0\,|\,E_1=R\,\&\,E_2=R)=\frac{(1+8Nc)*(1+4Nc)}{(1+4Nu+8Nc)*(1+4Nu+4Nc)*(1+4Nu)}$$

The probability of non-alike-in-state markers given that there was a recombination to the right and to the left is:

$$P(M=1\,|\,E_1=R\,\&\,E_2=R)=1-P(M=1\,|\,E_1=R\,\&\,E_2=R)$$

$$\approx\frac{4Nu*(3+24Nc+32N^2c^2)}{(1+4Nc)*(1+8Nc)} \qquad \text{[A3]}$$

for small 4*Nu*, as given in table 1 for P(y_k=0 | $\boldsymbol{\pi}$=[0 0]).

The above assumed that the segments to the left and right of *k* have the same size, *c*. If this is not the case *c* can be set equal to the mean of the two recombination rates, which is a good approximation as long as the harmonic mean of (1+4Nc_i) approximates its usual arithmetic mean, i.e. as long as 4Nc_i is small or the c_i values are not very different. It may be noted that each of the P(y_k=0|…) equations has a 4*Nu* term in the numerator, which means that these probabilities are proportional to the mutation rate, *u*. In case the mutation rate is unknown, 4*Nu* may be approximated by the marker heterozygosity.