A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps

Jean Michel Elsen, Olivier Filangi, Hélène Gilbert, Pascale Le Roy, Carole Moreno-Romieux

Genetics Selection Evolution

BioMed Central

Research

Open Access

A fast algorithm for estimating transmission probabilities in QTL detection designs with dense maps

Jean-Michel Elsen*[1], Olivier Filangi[2], Hélène Gilbert[3], Pascale Le Roy[2] and Carole Moreno[1]

Address: [1]INRA, SAGA, BP27, 31326 Castanet Tolosan cedex, France, [2]INRA, GARen, Agrocampus, 35000 Rennes, France and [3]INRA, GABI, 78352 Jouy en Josas cedex, France

Email: Jean-Michel Elsen* - Jean-Michel.Elsen@toulouse.inra.fr; Olivier Filangi - Olivier.Filangi@rennes.inra.fr; Hélène Gilbert - Helene.Gilbert@jouy.inra.fr; Pascale Le Roy - Pascale.Leroy@rennes.inra.fr; Carole Moreno - Carole.Moreno@toulouse.inra.fr

* Corresponding author

Published: 17 November 2009

Genetics Selection Evolution 2009, **41**:50 doi:10.1186/1297-9686-41-50

Received: 31 July 2009
Accepted: 17 November 2009

This article is available from: http://www.gsejournal.org/content/41/1/50

Abstract

Background: In the case of an autosomal locus, four transmission events from the parents to progeny are possible, specified by the grand parental origin of the alleles inherited by this individual. Computing the probabilities of these transmission events is essential to perform QTL detection methods.

Results: A fast algorithm for the estimation of these probabilities conditional to parental phases has been developed. It is adapted to classical QTL detection designs applied to outbred populations, in particular to designs composed of half and/or full sib families. It assumes the absence of interference.

Conclusion: The theory is fully developed and an example is given.

Background

Experimental designs used for mapping QTL in livestock based on linkage analysis techniques generally comprise two or three generations. The younger generation consists of large offsprings (either half sib only or mixture of half and full sib) measured on quantitative traits to be dissected. This generation and in most cases their parents are genotyped for a set of molecular markers. Genotyping an older generation (the grand parents) helps the determination of parents' phases, an information essential to linkage analysis. QTL detection is a multiple step procedure. First the parental phases must be determined from grand parental and/or progeny genotype information, either looking for their most probable phase, or building all possible phases and computing their probabilities. Then transmission probabilities of chromosomal segments from the parents to the progeny must be estimated conditional to the phases. Finally a test statistic (*e.g.* F or likelihood ratio test), based on a given model (*e.g.* regression, mixture model, variance component model...) is performed at each putative QTL position on the chromosomal segments traced. In crosses between inbred lines, the transmission probabilities are simply obtained, as described by [1], from the information given by markers flanking the QTL. In outbred populations, the computation is not straightforward, due to the variability of marker informativity between families and within families between progenies. In [2,3], the transmission probabilities were estimated conditionally to the sole flanking markers. [4-7] used a direct algorithm where all types of

gametes corresponding to a linkage group are successively considered: if L markers are heterozygous in the parent, 2^L gametes may be produced. This procedure is simple and computationnally fast for a small number of linked markers, but not feasible as soon as their number exceeds about 15. The difficulty can be circumvented in Bayesian approaches using MCMC techniques where these probabilities need not to be explicitly computed (*e.g.* [8]).

Nettelblad and colleagues [9] recently proposed a simple algorithm, which makes the transmission probabilities easily computable even for a large number of markers. In their approach the full length of the linkage group is still considered. A new algorithm, similar to the principle of [9] but exploring the minimum number of useful markers, was implemented in QTLMap software developed by INRA ([10]). Here, we describe and illustrate this algorithm.

Hypotheses. Notations. Objective

Progeny p was born from sire s and dam d. All were genotyped at L loci (M_l, $l = 1 \ldots L$). The location of M_l on the linkage group, *i.e.* its distance from one end of this group, is $x(M_l)$ centiMorgan, also denoted x_l. The hypothesis of absence of interference is made, allowing the Haldane distance function to be used.

The recombination rate between locus l_1 and l_2 will be noted r_{l_1, l_2}. Using the Haldane distance, $r_{l_1,l_2} = \frac{1}{2}(1 - exp\{-2(x_{l_2} - x_{l_1})\})$. When distances vary with sex, the superscript m (for males) or f (for females) will be used for x_l and r_{l_1, l_2}.

Let the l^{th} marker information be $P_{sl} = (P_{sl_1}, P_{sl_2})$ for the sire, $P_{dl} = (P_{dl_1}, P_{dl_2})$ for the dam, allele $P_{pl} = (P_{pl_1}, P_{pl_2})$ for the progeny. In P_{ilk}, $i = s$, d or p, the subscript k ($k = 1$, or 2) denotes the k^{th} allele read in the records file.

The probabilities of transmission of a chromosomal segment from the parents to the progeny are estimated conditional to parental phases. A phase of parent i (s or d) is characterised by a particular order of its marker phanotypes $P_i = \{P_{ilk}\}$, for loci $l = 1$ to L, giving $G_i = \{G_{ilk}\}$ where $k = 1$ means the grand sire allele and $k = 2$ the grand dam allele. If grand parental origins cannot be built, one of the alleles of the first heterozygous marker in the parent to be phased is arbitrary assigned the subscript $k = 1$.

Let $T(M_l)$ be the transmission event for marker l, and $T(M)$ the vector of transmission events on the linkage group: $T(M) = \{T(M_1), T(M_2) \cup T(M_L)\}$. $T(M_s)$ and $T(M_d)$ are respectively the transmission events from the sire and from the dam to the progeny. $T(M_{il}) = k$ if the progeny received G_{ilk}, $i = s$ or d. If the grand parental origins are known, progeny p may have received alleles from both its grand sires ($T(M_{sl}) = 1$ and $T(M_{dl}) = 1$, thus $T(M_l) = 11$), from its paternal grand sire and maternal grand dam ($T(M_l) = 12$), from its paternal grand dam and maternal grand sire ($T(M_l) = 21$), or from both its grand dams ($T(M_l) = 22$). The probabilities of the transmission events, given the marker phenotypes and parental phases are listed in Table 1 for a biallelic marker.

The 16 situations described in Table 1 belong to five types:

- Type '*ksd*' : Transmission fully known for both parents (cases 1 to 4),
- Type '*ks*0': Transmission known for the sire only (cases 5 to 8),
- Type '*k*0*d*': Transmission known for the dam only (cases 9 to 12),
- Type '*k*00': Unknown Transmission (cases 13 and 14),
- Type '*amb*': Ambiguous Transmission (case 15 and 16).

The *amb* type corresponds to fully heterozygous trios. It is essential to note that this is the only type of marker phenotypes for which the sire and dam transmissions are not independent (e.g. in situation 15, if sire transmits 1, dam transmits 2 and the reverse).

When the information about one or both parents is missing the conditionnal probability of $T(M_l)$ most often corresponds to the *k*00 type [1/4, 1/4, 1/4, 1/4]. However, when only one parent possesses a marker phanotype and is phased heterozygous (a, b), the probabilities are [1/2, 0, 1/2, 0] if $P_{pl} = (a, a)$ and [0, 1/2, 0, 1/2] if $P_{pl} = (b, b)$.

Two properties of the transmission probabilities must be underlined:

Property 1: Marginally to the marker phenotype, the sire and dam transmission events are independent: $P[T(M_l)] = P[T(M_{sl})].P[T(M_{dl})]$.

Property 2: Due to the no interference hypothesis, the transmission events follow a Markovian process described by:

$$P[T(M)] = P[T(M_1)].P[T(M_2) \mid T(M_1)].P[T(M_3) \mid T(M_2)] \cdots P[T(M_L) \mid T(M_{L-1}]$$

Table 1: $P[T(M_l) \mid G_{sl}, G_{dl}, P_{pl}]$: Probabilities of the transmission events, given the marker phenotypes and parental phases, in the case of a biallelic marker (*a*, *b* alleles)

						$P(T(M_l) \mid G_{sl}, G_{dl}, P_{pl})$ for $T(M_l)$ =			
Case	G_{sl_1}	G_{sl_2}	G_{dl_1}	G_{dl_2}	P_{pl}	**11**	**12**	**21**	**22**
1	a	b	a	b	(a, a)	1			
2	a	b	a	b	(b, b)				1
3	a	b	b	a	(a, a)		1		
4	a	b	b	a	(b, b)			1	
5	a	b	a	a	(a, a)	1/2	1/2		
6	a	b	a	a	(a, b) or (b, a)			1/2	1/2
7	b	a	a	a	(a, a)			1/2	1/2
8	b	a	a	a	(a, b) or (b, a)	1/2	1/2		
9	a	a	a	b	(a, a)	1/2		1/2	
10	a	a	a	b	(a, b) or (b, a)		1/2		1/2
11	a	a	b	a	(a, a)		1/2		1/2
12	a	a	b	a	(a, b) or (b, a)	1/2		1/2	
13	a	a	a	a	(a, a)	1/4	1/4	1/4	1/4
14	a	a	b	b	(a, b)	1/4	1/4	1/4	1/4
15	a	b	a	b	(a, b) or (b, a)		1/2	1/2	
16	a	b	b	a	(a, b) or (b, a)	1/2			1/2

G_{ilk} is the allele marker *l* the parent *i* is carrying on its k^{th} chromosome ((k = (1, 2)); P_{pl} is the marker *l* phenotype of the progeny; $T(M_l)$ = is the transmission event at marker *l*

Note that property 2 is also valid when considering subsets of M, M_b and M_a, allowing an independent estimation of probabilities before and after a given marker M_c. If $M = \{M_b, M_c, M_a\}$,

$$P[T(M)] = P[T(M_b) \mid T(M_c)].P[T(M_c)].P[T(M_a) \mid T(M_c)]$$

At any position x for a QTL, four grand parental origins are possible for the chromosomal segment Q_x inherited by the progeny. Let $q = (q_s, q_d)$, (q = (11), (12), (21) or (22)), the origin of Q_x.

The objective is to estimate $P_x(q) = P[T(Q_x) = q \mid G_s, G_d, P_p]$, the probability of q given the marker information.

To minimize the computation, two procedures are presented: the first one is an iterative exploration of the linkage group, the second a reduction of this group within bounds specific of the tested position x.

Iterative exploration of the linkage group

The observed marker phenotypes and parents' phases can be consistent with different transmission events $T(M)$. All these events must be considered in turn when evaluating the QTL transmission $T(Q_x)$. For a given marker transmission event, markers must be successively considered, the no interference hypothesis allowing an iterative estimation of the probability.

Proposition 1 : Let Ω be the domain, for the progeny p, of transmissions $T(M)$ consistent with the observations G_s, G_d and P_p. The transmission probability $P_x(q)$ is given by:

$$P[T(Q_x) = q \mid G_s, G_d, P_p] = \frac{\sum_{T(M)\in\Omega} P[T(Q_x)=q, T(M)]}{\sum_{T(M)\in\Omega} P[T(M)]} \quad (1)$$

This is obtained after very simple algebra (see appendix).

The domain Ω is obtained listing possible transmissions. If Ω_l is the consistent domain for marker l, the Ω domain is formed of nested domains $\Omega_1 \oplus \Omega_2 \oplus \cup \oplus \Omega_L \cdot \Omega_l$ is directly obtained from Table 1: it is formed of transmission events the probability of which are not nul. For instance, if $G_s = aa$, $G_d = ab$ and $P_p = aa$, then $\Omega_l = \{11, 12\}$.

In the following we shall note $S_\Omega = \sum_{T(M)\in\Omega} P[T(M)]$ and $T_\Omega = \sum_{T(M)\in\Omega} P[T(Q_x) = q, T(M)]$.

Proposition 2 : The summation $S_\Omega = \sum_{T(M)\in\Omega} P[T(M)]$ in (1) can be obtained recursively with the following algorithm:

$$\left.\begin{array}{ll} & S_\Omega = \sum_{T(M_L)\in\Omega_L} F[T(M_L)] \\ \text{With} & F[T(M_l)] = \sum_{T(M_{l-1})\in\Omega_{l-1}} P[T(M_l) \mid T(M_l - 1)].F[T(M_l - 1)] \\ \text{And} & F[T(M_l)] = P[T(M_1)] \end{array}\right\} \quad (2)$$

This is obtained under the hypothesis of absence of interference (see appendix).

Note 1: the numerator of (1) is obtained similarly, considering the extended domain $\Omega^* = \Omega_1 \oplus \Omega_2 \ldots \oplus\Omega_x \ldots \oplus\Omega_L$, with $\Omega_x = q$.

Note 2: The $P[T(M_l) \mid T(M_{l-1})]$ are simply obtained as given in Table 2, for $k = l - 1$.

They may be summarized by a single formulae. Let $\theta\langle r, i, j\rangle = 1 - r - (1 - 2r).(i - j)^2$,

$$P[T(M_l) \mid T(M_k)] = \theta\left\langle r_{l,k}^m, T(M_{sk}), T(M_{sl})\right\rangle.\theta\left\langle r_{l,k}^f, T(M_{dk}), T(M_{dl})\right\rangle$$

Note 3: System (2) may be generalized to any subdivision of the linkage group M, $M = \{M_1, M_2, \ldots M_G\}$, defining $T(M_g)$, $g = 1 \ldots G$, as the vector of $T(M_l)$, $l \in M_g$.

Reduction of the linkage group

The set of markers $M = \{M_l, l = 1 \ldots L\}$ may be sequenced as $M = \{M_a, M_\alpha, M_c, M_\beta, M_b\}$ where M_c is a subset of interest, M_β and M_α its flanking markers, and M_b and M_a all the remaining markers before and after the area (M_α, M_c, M_β). We now propose three simplifications of the summation $S_\Omega = \sum_{T(M)\in\Omega} P[T(M)]$.

Proposition 3 : In the summation S_Ω, the type $k00$ markers can be ignored, *i.e.* they may be bypassed in the iterative system (2).

Here M_c is a single $k00$ type marker. Proposition 3 means (see appendix for a demonstration) that, in (2), the sequence:

$$F[T(M_\beta)] = \sum_{T(M_c)\in\Omega_c} P[T(M_\beta) \mid T(M_c)]\left\{\sum_{T(M_\alpha)\in\Omega_\alpha} P[T(M_c) \mid T(M_\alpha)].F[T(M_\alpha)]\right\}$$

which corresponds to two iterations, may be replaced by:

$$F[T(M_\beta)] = \sum_{T(M_\alpha)\in\Omega(\alpha)} P[T(M_\beta) \mid T(M_\alpha)].F[T(M_\alpha)]$$

Proposition 4: In the summation S_Ω, the elements corresponding to the unknown parental transmission for types $k0d$ or $ks0$ markers can be ignored, *i.e.* they may be bypassed in the iterative system (2).

Here M_c is a single $ks0$ or $k0d$ type marker. Proposition 4 means (see appendix for a demonstration) that, in (2), the sequence

$$F[T(M_\beta)] = \sum_{T(M_c)\in\Omega_c} P[T(M_\beta) \mid T(M_c)]\left\{\sum_{T(M_\alpha)\in\Omega_\alpha} P[T(M_c) \mid T(M_\alpha)].F[T(M_\alpha)]\right\}$$

which corresponds to two iterations, may be replaced by (successively $k0d$ and $ks0$ markers):

$$F[T(M_\beta)] = P[T(M_{d\beta}) \mid T(M_{dc})] \sum_{T(M_\alpha)\in\Omega(\alpha)} P[T(M_{dc}) \mid T(M_{d\alpha})].P[T(M_{s\beta}) \mid T(M_{s\alpha})].F[T(M_\alpha)]$$

$$F[T(M_\beta)] = P[T(M_{s\beta}) \mid T(M_{sc})] \sum_{T(M_\alpha)\in\Omega(\alpha)} P[T(M_{sc}) \mid T(M_{s\alpha})].P[T(M_{d\beta}) \mid T(M_{d\alpha})].F[T(M_\alpha)]$$

Corollary 1: In the summation S_Ω, a sequence M_c of markers all belonging to "k" types (*i.e.* non *amb*) appears as a single element where only the certain transmissions are involved.

From propositions 3 and 4,

Table 2: Transmission probability at locus *l* given the transmission at locus *k*: $P[T(M_l) \mid T(M_k)]$

$T(M_k)$	11	12	21	22
$T(M_l)$				
11	$(1-r_{l,k}^m).(1-r_{l,k}^f)$	$(1-r_{l,k}^m)r_{l,k}^f$	$r_{l,k}^m(1-r_{l,k}^f)$	$r_{l,k}^m r_{l,k}^f$
12	$(1-r_{l,k}^m)r_{l,k}^f$	$(1-r_{l,k}^m)(1-r_{l,k}^f)$	$r_{l,k}^m r_{l,k}^f$	$r_{l,k}^m(1-r_{l,k}^f)$
21	$r_{l,k}^m(1-r_{l,k}^f)$	$r_{l,k}^m r_{l,k}^f$	$(1-r_{l,k}^m)(1-r_{l,k}^f)$	$(1-r_{l,k}^m)r_{l,k}^f$
22	$r_{l,k}^m r_{l,k}^f$	$r_{l,k}^m(1-r_{l,k}^f)$	$(1-r_{l,k}^m)r_{l,k}^f$	$(1-r_{l,k}^m)(1-r_{l,k}^f)$

$r_{l,k}^i$ is the recombination rate for sex *i*, between loci *l* and *k*.

$$F[T(M_\beta)] = P[T(M_{d\beta}) \mid T(M_{dc_1})].\left\{\prod_{j_d=1\cdots J_D} P[T(M_{dc_{j_d+1}}) \mid T(M_{dc_{j_d}})]\right\}$$
$$P[T(M_{s\beta}) \mid T(M_{sc_1})].\left\{\prod_{j_s=1\cdots J_S} P[T(M_{sc_{j_s+1}}) \mid T(M_{sc_{j_s}})]\right\}$$
$$\sum_{T(M_\alpha)\in\Omega_\alpha} P[T(M_{dc_{j_d}}) \mid T(M_{d\alpha})].P[T(M_{sc_{j_s}}) \mid T(M_{s\alpha})].F[T(M_\alpha)]$$

where the markers subscripted j_s (= 1 ∪ J_s) are successive markers belonging to *ksd* or *ks*0 types, and the markers subscripted j_d (= 1 ∪ J_d) to *ksd* or *k*0*d* types in the sequence M_c.

Definition : A series of markers $N = \{M_\alpha, M_c, M_\beta\}$ starting with a *ks*0 (*resp. k*0*d*) type marker $\{M_\alpha\}$, ending with a *k*0*d* (*resp. ks*0) type marker $\{M_\beta\}$, and only with *k*00 type markers between those bounds (in M_c) will be called a sd-node (*resp*. ds-node).

Proposition 5: If the sequence $N = \{M_\alpha, M_c, M_\beta\}$ of M is a sd-node, the summation S_Ω may be separated in three terms corresponding to $[M_b/M_{\beta s}, M_{\alpha d}]$, $[M_{\beta s}, M_{\alpha d}]$, and $[M_a/M_{\beta s}, M_{\alpha d}]$ Proposition 5 means (see appendix for a demonstration) that, in (2), S_Ω is obtained by

$$S_\Omega = \left\{\sum_{T(M_b)\in\Omega_b}\sum_{T(M_{d\beta})\in\Omega_{\beta b}} P[T(M_b), T(M_{d\beta}) \mid T(M_{s\beta}), T(M_{d\alpha})]\right\}.$$
$$P[T(M_{s\beta}), T(M_{d\alpha})].$$
$$\left\{\sum_{T(M_a)\in\Omega_\alpha}\sum_{T(M_{s\alpha})\in\Omega_{s\alpha}} P[T(M_a), T(M_{s\alpha}) \mid T(M_{s\beta}), T(M_{d\alpha})]\right\}$$

Note 4: The $\{M_\beta, M_c, M_\alpha\}$ sequence may be reduced to a single marker M_γ if it belongs to the *ksd* type. In this case,

$$S_\Omega = \left\{\sum_{T(M_b)\in\Omega_b} P[T(M_b) \mid T(M_\gamma)]\right\}.P[T(M_\gamma)].\left\{\sum_{T(M_a)\in\Omega_a} P[T(M_a) \mid T(M_\gamma)]\right\}$$

In general we shall note $T(N)$ the transmission event for a node, $\{T(M_{s\beta}), T(M_{d\alpha})\}$, $\{T(M_{d\beta}), T(M_{s\alpha})\}$ or $T(M_\gamma)$.

Corollary 2: If the tested QTL position x is located in segment M_c between two nodes N_1 and N_2, only the markers belonging to the interval $[N_1, N_2]$ have to be considered when computing the transmission probability $P[T(Q_x) = q \mid G_s, G_d, P_p]$, see appendix, giving:

$$P[T(Q_x) = q \mid G_s, G_d, P_p] = \frac{\sum_{T(M_C)\in\Omega_C} P[T(Q_x)=q, T(N_1), T(M_C), T(N_2)]}{\sum_{T(M_C)\in\Omega_C} P[T(N_1), T(M_C), T(N_2)]} \quad (3)$$

Algorithm

Based on the propositions and corollaries developed above, an algorithm for the computation of transmission probabilities of the chromosomic segment x can be given.

1. From the position x, the markers are explored towards the left until a node (a *ksd* type marker or a pair of markers one of *ks*0 and the other of *k*0*d* type, separated only by *k*00 type markers) or the extremity of the linkage group is found. Let $T(N_l)$ be the transmission events for the left node N_l. $P[T(N_l)] = 1/4$.

2. From the position x, the markers are explored towards the right until a node or the extremity of the linkage group is found. Let $T(N_r)$ be the transmission events for the right node N_r. $P[T(N_r)] = 1/4$. The only necessary informative segment for x in the full linkage group is $\{N_l, N_r\}$.

3. Let $M_{a_1}, M_{a_2}, \cdots, M_{a_n}$ the *amb* type markers in $\{N_l, N_r\}$. Together with N_l and N_r, the M_{a_k} delimit $n + 1$ intervals I_k, which may be empty or include *k*00, *ks*0 or *k*0*d* type markers. The reduced summation S^r_Ω, see (the part of S_Ω which differs from T_Ω and has to be used in $P_x(q) = \frac{S_\Omega}{T_\Omega} = \frac{S^r_\Omega}{T^r_\Omega}$ see appendix) is computed iteratively:

$$\left.\begin{array}{rrl} & S^r_\Omega & = F[T(N_r)] \\ \text{With} & F[T(N_r)] & = \sum_{T(M_{a_n})\in\Omega_{a_n}} P[T(N_r) \mid T(M_{a_n})].F[T(M_{a_n})] \\ \text{Then} & F[T(M_{a_l})] & = \sum_{T(M_{a_{l-1}})\in\Omega_{a_{l-1}}} P[T(M_{a_l}) \mid T(M_{a_{l-1}})].F[T(M_{a_{l-1}})] \\ & & \text{For } l = n, \cdots, 2 \\ \text{And} & F[T(M_{a_1})] & = P[T(M_{a_1}) \mid T(N_l)].P[T(N_l)] \end{array}\right\} \quad (4)$$

It must be underlined that there is no node between two adjacent *amb* type markers of the informative segment $\{N_l, N_r\}$, since this segment ends at the first node found on both sides. As a consequence, neither a *ksd* marker type nor a mixture of *ks*0 and *k*0*d* types markers could be found between the ambiguous markers $M(a_k)$ and $M(a_{k+1})$: the I_k interval may be classified as *K*00 (only *k*00 types markers), *Ks*0 (one or more *ks*0 type markers, no *k*0*d* type marker and any number of *k*00 type markers) or *K*0*d* (the reverse).

4. Let M_{a_k} and $M_{a_{k+1}}$ be two successive *amb* markers, in the iterative process (4), the probabilities $P[T(M_{a_{k+1}})/T(M_{a_k})]$ are given by

$$K00 \text{ interval} \quad \theta\left\langle r^m_{a_k,a_{k+1}}, T(M_{sa_k}), T(M_{sa_{k+1}})\right\rangle.\theta\left\langle r^f_{a_k,a_{k+1}}, T(M_{da_k}), T(M_{da_{k+1}})\right\rangle$$
$$Ks0 \text{ interval} \quad \left\{\prod_{i\in I_k}\theta\left\langle r^m_{i-1,i}, T(M_{si-1}), T(M_{si})\right\rangle\right\}.\theta\left\langle r^f_{a_k,a_{k+1}}, T(M_{da_k}), T(M_{da_{k+1}})\right\rangle$$
$$K0d \text{ interval} \quad \theta\left\langle r^m_{a_k,a_{k+1}}, T(M_{sa_k}), T(M_{sa_{k+1}})\right\rangle.\left\{\prod_{i\in I_k}\theta\left\langle r^f_{i-1,i}, T(M_{di-1}), T(M_{di})\right\rangle\right\}$$

where $\theta\langle r, i, j\rangle = 1 - r - (1 - 2r).(i - j)^2$.

5. The reduced summation T_Ω^r is computed iteratively adding the $T(Q_x)$ transmission in the list of transmission $\{T[N_l], T[M_{a_1}], \cup, T[M_{a_n}], T[N_r]\}$.

6. The transmission probability $P[T(Q_x) = q \mid G_s, G_d, P_p] = T_\Omega^r / S_\Omega^r$.

Note 5 : The algorithm can be organised scanning the interval $\{N_l, N_r\}$ from the left to the right rather than from the right to the left as described above.

Example

A linkage group of eight markers is available (Figure 1). Markers M_2 and M_6 are ambiguous, with types 15 and 16. Markers 1 and 8 are fully informative (types 1 and 2), the other markers are semi informative. The tested position for the QTL x is located between markers 4 and 5. The nodes are, on the left, marker 1 (*ksd* type) and on the right, the group M_7 - M_8. Thus the informative segment here is the full group. Steps of the proposed algorithm are detailed Table 3.

Discussion - Conclusion

The algorithm presented in this paper to estimate the transmission probability of QTL from parents to progeny needs only very limited computational resources, both in terms of time and space. Complementary to the algorithm presented by Nettleblad and colleagues (2009), it limits the exploration of the linkage group to the markers really informative for a given position to be traced, and thus performs faster. As [9], it deals with sex differences between recombination rates.

The QTL transmission probability is estimated conditionnaly to the observed transmission at the surrounding markers loci. The algorithm does not make use of possible

Table 3: Calculation of the marker transmission probability corresponding to the example in Figure 1

$T(N_l)$	11			
$P[T(N_l)]$	1/4			
$T(M_{a_1})$	12	21	12	21
$P[T(M_{a_1})\|T(N_l)]$	$(1-r_{12}^m)r_{12}^f$	$r_{12}^m(1-r_{12}^f)$	$(1-r_{12}^m)r_{12}^f$	$r_{12}^m(1-r_{12}^f)$
$F[T(M_{a_1})\|T(N_l)]$	$1/4(1-r_{12}^m)r_{12}^f$	$1/4r_{12}^m(1-r_{12}^f)$	$1/4(1-r_{12}^m)r_{12}^f$	$1/4r_{12}^m(1-r_{12}^f)$
$T(M_{a_2})$	11	11	22	22
$P[T(M_{a_2})\|T(M_{a_1})]$	$(1-r_{23}^m)r_{34}^m r_{46}^m r_{25}^f(1-r_{56}^f)$	$r_{23}^m r_{34}^m r_{46}^m(1-r_{25}^f)(1-r_{56}^f)$	$(1-r_{23}^m)r_{34}^m(1-r_{46}^m)r_{25}^f r_{56}^f$	$r_{23}^m r_{34}^m(1-r_{46}^m)(1-r_{25}^f)r_{56}^f$
$F[T(M_{a_2})]$	$1/4[(1-r_{12}^m)r_{12}^f(1-r_{23}^m)r_{25}^f + r_{12}^m(1-r_{12}^f)r_{23}^m(1-r_{25}^f)]r_{34}^m r_{46}^m(1-r_{56}^f) r_{67}^m r_{68}^f$		$1/4[(1-r_{12}^m)r_{12}^f(1-r_{23}^m)r_{25}^f + r_{12}^m(1-r_{12}^f)r_{23}^m(1-r_{25}^f)]r_{34}^m(1-r_{46}^m)r_{56}^f (1-r_{67}^m)(1-r_{68}^f)$	
$P[T(N_r)\|T(M_{a_2})]$				
$F[T(N_r)]$	$1/4[(1-r_{12}^m)r_{12}^f(1-r_{23}^m)r_{25}^f + r_{12}^m(1-r_{12}^f)r_{23}^m(1-r_{25}^f)].r_{34}^m.[r_{46}^m(1-r_{56}^f)r_{67}^m r_{68}^f + (1-r_{46}^m)r_{56}^f(1-r_{67}^m)(1-r_{68}^f)]$			

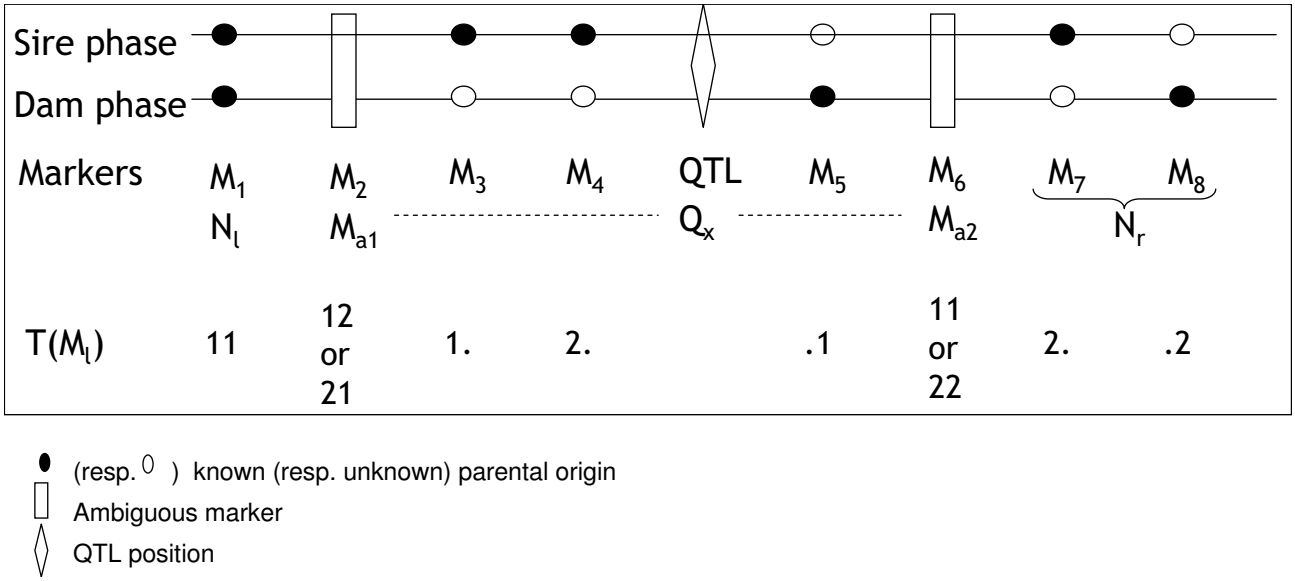

Figure 1

Example of a linkage group with 8 markers including 2 ambigous. The figure represents a chromosome with eight markers. Two (M_2 and M_6) are ambiguous (For M_2, the progeny received either the 1st allele of its sire and 2nd allele of its dam, or the 2nd of its sire and 1st of its dam. The nodes are, on the left, the first marker, and on the right, markers M_7 and M_8. The dark (respectively white) circles represent markers with a known (respectively unknown) grand parental origin.

information about the marker allele frequencies to fill potential information gaps.

The major difficulty addressed in this algorithm is the non independence of transmission events from the sire and the dam to the progeny in triple heterozygous trios. In the absence of such trios, the transmission from the parents are fully independent and may be treated separately simply by considering the flanking informative markers. This is the case for QTL located on the sex chromosome X or W.

The algorithm has been developed in the framework of QTL detection designs involving two or three generations in outbred populations. It has been implemented in QTLMap, a software for the analysis of such designs. QTLMap is available upon request to the authors.

In more complex pedigrees, the transmission probability should not be conditioned only on parents phases and progeny marker phanotypes. Information from the grand progeny (and the spouses lineages) may improve the estimation, since the progeny phase can be inferred, at least partially, from these data. A recursive process inspirated from [3] should possibly be implemented.

The transmission probabilities are estimated conditionally to parental phases. In linear approaches (*e.g.* the Haley Knott regression), if more than one phase is probable, the marginal transmission probability could be estimated considering all of them in a weighted sum of conditional probabilities. Alternatively, the only most probable phase could be considered [11].

The absence of interference hypothesis is central in the present algebra. If this is not true, then most of the propositions are not valid and the algorithm not applicable.

Finally, compared to the most common codominant markers, dominant markers will be characterized by a lower informativity, with an increase of the between nodes segment length and a concomitant decrease of the transmission probability.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JME drafted the manuscript. All authors participated in the development of the method and read and approved the final manuscript.

Appendix: Demonstration of the propositions and corollary

Proposition 1: $P[T(Q_x) = q \mid G_s, G_d, P_p] = \frac{\sum_{T(M)\in\Omega} P[T(Q_x)=q, T(M)]}{\sum_{T(M)\in\Omega} P[T(M)]}$

$$
\begin{aligned}
P[T(Q_x) = q \mid G_s, G_d, P_p] &= \frac{P[T(Q_x)=q, P_p \mid G_s, G_d]}{P[P_p \mid G_s, G_d]} \\
P[P_p \mid G_s, G_d] &= \sum_{T(M)} P[T(M), P_p \mid G_s, G_d] \\
\text{and } P[T(Q_x) = q, P_p \mid G_s, G_d] &= \sum_{T(M)} P[T(M), T(Q_x) = q, P_p \mid G_s, G_d] \\
P[T(M), P_p \mid G_s, G_d] &= P[P_p \mid T(M), G_s, G_d].P[T(M) \mid G_s, G_d] \\
P[P_p \mid T(M), G_s, G_d] &= 1 \text{ if } T(M) \in \Omega \\
&= 0 \text{ if not} \\
P[T(M) \mid G_s, G_d] &= P[T(M)]
\end{aligned}
$$

And, similarly, $P[T(Q_x) = q, P_p \mid G_s, G_d] = P[T(Q_x) = q, T(M)]$ if $T(M) \in \Omega$, = 0 if not

Proposition 2

Due to the no interference hypothesis, the transmission events follow a Markovian process described by:

$$P[T(M)] = P[T(M_1)].P[T(M_2) \mid T(M_1)].P[T(M_3) \mid T(M_2)] \cdots P[T(M_L) \mid T(M_{L-1})]$$

Thus

$$
\begin{aligned}
S_\Omega &= \sum_{T(M)\in\Omega} P[T(M)] \\
&= \sum_{T(M_1)\in\Omega_1} \sum_{T(M_2)\in\Omega_2} \cdots \sum_{T(M_L)\in\Omega_L} P[T(M_1)].\prod_{l=2\cdots L} P[T(M_l) \mid T(M_{l-1})]
\end{aligned}
$$

The summations may be inverted:

$$
\begin{aligned}
S_\Omega = &\sum_{T(M_L)\in\Omega_L} \sum_{T(M_{L-1})\in\Omega_{L-1}} P[T(M_L) \mid T(M_{L-1})] \{ \sum_{T(M_{L-2})\in\Omega_{L-2}} P[T(M_{L-1}) \mid T(M_{L-2})] \\
&\{ \cdots \{ \sum_{T(M_1)\in\Omega_1} P[T(M_2) \mid T(M_1)].P[T(M_1)] \} \} \cdots \}
\end{aligned}
$$

Consequently:

$$
\begin{aligned}
\text{If} \quad F[T(M_1)] &= P[T(M_1)] \\
\text{then} \quad F[T(M_2)] &= \sum_{T(M_1)\in\Omega_1} P[T(M_2) \mid T(M_1)].F[T(M_1)] \\
F[T(M_l)] &= \sum_{T(M_{l-1})\in\Omega_{l-1}} P[T(M_l) \mid T(M_{l-1})].F[T(M_{l-1})] \\
\text{And} \quad S_\Omega &= \sum_{T(M_L)\in\Omega_L} F[T(M_L)]
\end{aligned}
$$

Proposition 3

With an argument similar to the demonstration of proposition 2, the sum S_Ω may be expressed as:

$$S_\Omega = \sum_{T(M_b)\in\Omega_b} \sum_{T(M_\beta)\in\Omega_\beta} P[T(M_b) \mid T(M_\beta)].F[T(M_\beta)]$$

$$
\begin{aligned}
F[T(M_\beta)] &= \sum_{T(M_c)\in\Omega_c} P[T(M_\beta) \mid T(M_c)].F[T(M_c)] \\
&= \sum_{T(M_c)\in\Omega_c} P[T(M_\beta) \mid T(M_c)].\left\{ \sum_{T(M_\alpha)\in\Omega_\alpha} P[T(M_c) \mid T(M_\alpha)].F([T(M_\alpha)] \right\} \\
&= \sum_{T(M_\alpha)\in\Omega_\alpha} \left\{ \sum_{T(M_c)\in\Omega_c} P[T(M_\beta) \mid T(M_c)].P[T(M_c) \mid T(M_\alpha)] \right\}.F([T(M_\alpha)] \\
&= \sum_{T(M_\alpha)\in\Omega_\alpha} \left\{ \sum_{T(M_c)\in\Omega_c} P[T(M_\beta), T(M_c) \mid T(M_\alpha)] \right\}.F([T(M_\alpha)] \\
&= \sum_{T(M_\alpha)\in\Omega_\alpha} \left\{ \sum_{T(M_c)\in\Omega_c} P[T(M_c) \mid T(M_\beta), T(M_\alpha)] \right\}.P[T(M_\beta) \mid T(M_\alpha)].F([T(M_\alpha)]
\end{aligned}
$$

Thus

$$F[T(M_\beta)] = \sum_{T(M_\alpha)\in\Omega_\alpha} \{ \sum_{T(M_c)\in\Omega_c} P[T(M_c) \mid T(M_\beta), T(M_\alpha)] \}.P[T(M_\beta \mid T(M_\alpha)].F([T(M_\alpha)] \quad \text{(A1)}$$

As Ω_c forms a complete set of events, since all transmissions are possible,

$$\sum_{T(M_c)\in\Omega_c} P[T(M_c) \mid T(M_\beta), T(M_\alpha)] = 1$$

Thus

$$F[T(M_\beta)] \sum_{T(M_\alpha)\in\Omega_\alpha} P[T(M_\beta) \mid T(M_\alpha)].F([T(M_\alpha)]$$

Proposition 4

In the equation(A1), we have, from property 1,

$$P[T(M_c) \mid T(M_\beta), T(M_\alpha)] = P[T(M_{sc}) \mid T(M_{s\beta}), T(M_{s\alpha})].P[T(M_{dc}) \mid T(M_{d\beta}), T(M_{d\alpha})]$$

Without loss of generality, we assume that the parent with unknown transmission at M_c is the sire. There is a unique consistent $T(M_{dc})$, and the 2 possible $T(M_{sc})$ form a complete set of events, thus:

$$\sum_{T(M_c)\in\Omega_c} P[T(M_c) \mid T(M_\beta), T(M_\alpha)] = P[T(M_{dc}) \mid T(M_{d\beta}), T(M_{d\alpha})]$$

The simplification of $F[T(M_\beta)]$ follows:

$$
\begin{aligned}
F[T(M_\beta)] &= \sum_{T(M_u)\in\Omega_u} P[T(M_{dc}) \mid T(M_{d\beta}), T(M_{d\alpha})].P[T(M_\beta) \mid T(M_\alpha)].F[T(M_\alpha)] \\
&= \sum_{T(M_u)\in\Omega_u} P[T(M_{dc}) \mid T(M_{d\beta}), T(M_{d\alpha})].P[T(M_{d\beta}) \mid T(M_{d\alpha})].P[T(M_{s\beta}) \mid T(M_{s\alpha})].F[T(M_\alpha)] \\
&= \sum_{T(M_u)\in\Omega_u} P[T(M_{d\beta}) \mid T(M_{dc})].P[T(M_{dc}) \mid T(M_{d\alpha})].P[T(M_{s\beta}) \mid T(M_{s\alpha})].F[T(M_\alpha)]
\end{aligned}
$$

Proposition 5

When M_c contains markers of $k00$ type, they can be forgotten following proposition 3. We thus assume that the M_c group is empty, and the linkage group is described as $M = \{M_b, M_\beta, M_\alpha, M_a\}$

$$
\begin{aligned}
&S_\Omega = \sum_{T(M_b)\in\Omega_b} \sum_{T(M_\beta)\in\Omega_\beta} \sum_{T(M_\alpha)\in\Omega_\alpha} \sum_{T(M_a)\in\Omega_a} P[T(M_b), T(M_\beta), T(M_\alpha), T(M_a)] \\
&P[T(M_b), T(M_\beta), T(M_\alpha), T(M_a)] = \\
&P[T(M_b), T(M_{d\beta}) \mid T(M_{s\beta}), T(M_\alpha), T(M_a)].P[T(M_{sc}), T(M_a) \mid T(M_{s\beta}), T(M_{d\alpha})].P[T(M_{s\beta}) \mid T(M_{d\alpha})]
\end{aligned}
$$

But $P[T(M_b),\ T(M_{d\beta})\ |\ T(M_{s\beta}),\ T(M_\alpha),\ T(M_a)] = P[T(M_b),\ T(M_{d\beta})\ |\ T(M_{s\beta}),\ T\ (M_{d\alpha})]$

Thus

$$S_\Omega = \sum_{T(M_b)\in\Omega_b}\sum_{T(M_\beta)\in\Omega_\beta}\sum_{T(M_\alpha)\in\Omega_\alpha}\sum_{T(M_a)\in\Omega_a} P[T(M_b).T(M_{d\beta})\ |\ T(M_{s\beta}),T(M_{d\alpha})].$$
$$P[T(M_{s\alpha}),T(M_a)\ |\ T(M_{s\beta}),T(M_{d\alpha})].P[T(M_{s\beta})\ |\ T(M_{d\alpha})]$$
$$= \{\sum_{T(M_b)\in\Omega_b}\sum_{T(M_\beta)\in\Omega_\beta} P[T(M_b).T(M_{d\beta})\ |\ T(M_{s\beta}),T(M_{d\alpha})]\}.P[T(M_{s\beta})\ |\ T(M_{d\alpha})].$$
$$\{\sum_{T(M_\alpha)\in\Omega_\alpha}\sum_{T(M_a)\in\Omega_a} P[T(M_{s\alpha}).T(M_a)\ |\ T(M_{s\beta}),T(M_{d\alpha})]\}$$

Corollary 2

Let $M = \{M_b, N_l, M_c, N_r, M_a\}$, with $x(N_l) \le x \le x(N_r)$

From proposition 5, assuming both nodes N_l and N_r are sd-nodes,

$$S_\Omega = \left\{\sum_{T(M_b)\in\Omega_b} P[T(M_b)\ |\ T(N_l)]\right\}.P[T(N_l)].\left\{\sum_{T(M_c)\in\Omega_c}\sum_{T(M_a)\in\Omega_a} P[T(M_c),T(N_r),T(M_a)\ |\ T(N_l)]\right\}$$

From proposition 5 again,

$$\sum_{T(M_c)\in\Omega_c}\sum_{T(M_a)\in\Omega_a} P[T(M_c),T(N_r),T(M_a)\ |\ T(N_l)] =$$
$$\left\{\sum_{T(M_c)\in\Omega_c} P[T(M_c)\ |\ T(N_l),T(N_r)]\right\}.P[T(N_r)\ |\ T(N_l)].\left\{\sum_{T(M_a)\in\Omega_a} P[T(M_a)\ |\ T(N_l),T(N_r)]\right\}$$

The elements $\sum_{T(M_b)\in\Omega_b} P[T(M_b)\ |\ T(N_l)]$ and $\sum_{T(M_a)\in\Omega_a} P[T(M_b)\ |\ T(N_l),T(N_r)]$ being also present in the numerator T_Ω of (1) they can be forgotten.

The summation S_Ω may be reduced to S_Ω^r:

$$S_\Omega^r = P[T(N_l)].\left\{\sum_{T(M_c)\in\Omega_c} P[T(M_c)\ |\ T(N_l),T(N_r)]\right\}.P[T(N_r)\ |\ T(N_l)]$$
$$= \sum_{T(M_c)\in\Omega_c} P[T(M_c),T(N_l),T(N_r)]$$

Similarly

$$T_\Omega^r = \sum_{T(M_c)\in\Omega_c} P[T(Q_x),T(N_l),T(M_c),T(N_r)]$$

Acknowledgements

Financial support of this work was provided by the EC-funded FP6 Project "SABRE".

References

1. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121:**185-199.
2. Liu JM, Jansen GB, Lin CY: **The covariance between relatives conditional on genetic markers.** *Genet Sel Evol* 2002, **34:**657-678.
3. Pong-Wong R, George AW, Woolliams JA, Haley CS: **A simple and rapid method for calculating identity-by-descent matrices using multiple markers.** *Genet Sel Evol* 2002, **33:**453-471.
4. Haley CS, Knott SA, Elsen JM: **Mapping quantitative trait loci in crosses between outbred lines using least squares.** *Genetics* 1994, **136:**1195-1207.
5. Knott SA, Elsen JM, Haley CS: **Methods for multiple marker mapping of quantitative trait loci in half-sib populations.** *Theor Appl Genet* 1996, **93:**71-80.
6. Elsen JM, Mangin B, Goffinet B, Boichard D, Le Roy P: **Alternative models for QTL detection in livestock - I General introduction.** *Genet Sel Evol* 1999, **31:**213-224.
7. Le Roy P, Elsen JM, Boichard D, Mangin B, Bidanel JP, Goffinet B: **An algorithm for QTL detection in mixture of full and half sib families.** *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production: 12-16 January 1998; Armidale Australia* 1998.
8. Totir LR, Fernando RL, Dekkers JC, Fernández SA, Guldbrandtsen B: **A comparison of alternative methods to compute conditional genotype probabilities for genetic evaluation with finite locus models.** *Genet Sel Evol* 2003, **35:**585-604.
9. Nettelblad C, Holmgren S, Crooks L, Carlborg O: **cnF2freq: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations Using Markov Models.** *BICoB* 2009:307-319.
10. Elsen JM, Filangi O, Gilbert H, Legarra A, Le Roy P, Moreno C: **QTLMap: a software for the detection of QTL in full and half sib families.** *Proceedings of the EAAP Annual meeting 24-27 August 2009; Barcelona* 2009.
11. Windig JJ, Meuwissen THE: **Rapid haplotype reconstruction in pedigrees with dense marker maps.** *J Anim Breed Genet* 2004, **121:**2639.