Genet. Sel. Evol. 34 (2002) 335–352

DOI: 10.1051/gse:2002011

Original article

Use of maternal information for QTL detection in a (grand)daughter design

Marc BOLARD*, Didier BOICHARD**

Station de génétique quantitative et appliquée,
Institut national de la recherche agronomique, 78352 Jouy-en-Josas cedex, France

(Received 17 September 2001; accepted 6 March 2002)

Abstract – In a (grand)daughter design, maternal information is often neglected because the number of progeny per dam is limited. The number of dams per maternal grandsire (MGS), however, could be large enough to contribute to QTL detection. But dams and MGS usually are not genotyped, there are two recombination opportunities between the MGS and the progeny, and at a given location, only half the progeny receive a MGS chromosomal segment. A 3-step procedure was developed to estimate: (1) the marker phenotypes probabilities of the MGS; (2) the probability of each possible MGS haplotype; (3) the probabilities that the progeny receives either the first, or second MGS segment, or a maternal grandam segment. These probabilities were used for QTL detection in a linear model including the effects of sire, MGS, paternal QTL, MGS QTL and maternal grandam QTL. Including the grandam QTL effect makes it possible to detect QTL in the grandam population, even when MGS are not informative. The detection power, studied by simulation, was rather high, provided that MGS family size was greater than 50. Using maternal information in the French dairy cattle granddaughter design made it possible to detect 23 additional QTL genomewise significant.

QTL detection / daughter design / granddaughter design / dairy cattle

1. INTRODUCTION

In laboratory animals and in plant, the most efficient QTL detection designs involve crosses (F2, back cross, recombinant lines, advanced intercross...) of inbred lines. Because parental lines are completely homozygous, the design is equivalent to one large family, whatever the number of reproducing animals. In most domestic animal species, however, no completely inbred line is available and QTL detection should be carried out within-family. Under these conditions, a key factor for detection power is the number of progeny per parent. In some species, as in cattle, female prolificacy is low and large dam families are very difficult to obtain, even with artificial reproduction tools. Consequently,

* Present address: URCEO, 69 rue de la Motte Brûlon, 35019 Rennes, France
** Correspondence and reprints
E-mail: boichard@dga.jouy.inra.fr

only sire families efficiently contribute to QTL detection, in the so-called daughter [7] or granddaughter design [9]. In these designs, maternal meioses, *i.e.* half the total number of meioses, are not used and this strongly affects their detection power. Consequently, these designs are typically 4 to 10 times larger than those involving inbred lines.

Due to the high cost of these designs, strategies should be developed to extract more information from the data. When dams are not or very little related, maternal information could be used in association analysis, by taking advantage of the linkage disequilibrium likely to be present in domestic populations due to recent bottlenecks and founder effects. This approach, however, requires dense maps and therefore additional genotyping work.

In other situations, dams are more related and are born from a limited number of maternal grandsires (MGS). Coppieters *et al.* [4] showed that this family structure could be used for QTL detection when MGS phenotypes and phases are known, even when dams are not genotyped. This situation typically occurs when MGS are also sires. Bink and Van Arendonk [1] proposed a Monte Carlo Markov Chain approach to use information of non genotyped individuals, and particularly information from the dams. The present paper proposes a deterministic methodology suited to the situation where dams and MGS are not genotyped for the markers and when MGS phases are unknown. It also presents a simulation study of the detection power which can be achieved by this additional information. The method is illustrated with the analysis of the granddaughter design (GDD) carried out in French dairy cattle.

2. METHODS

The method proposed involved the four following steps and this section is divided accordingly: the estimation of the marker phenotypes of the MGS, given the marker information of the grandprogeny and their sire; the estimation of the marker phases of the MGS; the estimation of the QTL transmission probabilities from the MGS to the grandprogeny; and QTL detection *sensu stricto* by linear regression. When MGS are already genotyped, step 1 could be skipped. And when MGS phases are already known (*e.g.* when MGS appear also as sires in the design), step 2 could also be skipped.

Conceptually, steps 1 and 2 could be carried out simultaneously, but the number of possible phases for one MGS with m markers and with n_i possible alleles per marker reaches $\frac{1}{2}\prod_{i=1}^{m} n_i^2 \left[\left(\prod_{j=1}^{m} n_j^2\right) + 1\right]$ and is usually too high (for instance 4×10^7 for $m = n = 5$) for a direct search.

As this analysis is supposed to be a by-product of a daughter or granddaughter design and of a within-sire analysis, we first summarise the most important concepts of this classical model to introduce some notations.

2.1. Within-sire QTL detection

The probability $p(hs_i|M_i)$ of each possible phase hs_i of each sire i, conditional on marker information, is determined, assuming linkage equilibrium in the population (*i.e.* equal prior probability of each phase) and that dams are unrelated, as follows [5]:

$$pr\,(hs_i/M_i) = \frac{\prod_j pr\left(mp_{ij}/hs_i\right)}{\sum_{hs_i}\prod_j pr\left(mp_{ij}/hs_i\right)} \tag{1}$$

where M_i denotes the whole marker phenotype information of the sire and their progeny, and mp_{ij} denotes the marker phenotype information of progeny j. $pr(mp_{ij}/hs_i)$ is obtained as follows. Let γ_{ij}^l be a variable indicating the grandparental origin of marker l transmitted by sire i to his progeny j. γ_{ij}^l is equal to 0, 1, or 2, when the origin is unknown or when the allele originates from the paternal grandsire or grandam, respectively. If only the informative markers are considered ($\gamma_{ij}^l > 0$), $pr(mp_{ij}/hs_i)$ is obtained by: $pr\left(mp_{ij}/hs_i\right) = \frac{1}{2}\prod_l pr(\gamma_{ij}^l, \gamma_{ij}^{l+1}/hs_i)$, where $pr(\gamma_{ij}^l, \gamma_{ij}^{l+1}/hs_i)$ is equal to $1 - r(l, l+1)$ if $\gamma_{ij}^l = \gamma_{ij}^{l+1}$ (no recombination) or $r(l, l+1)$ if $\gamma_{ij}^{l+1} \neq \gamma_{ij}^{l+1}$ (recombination), and $r(l, m)$ is the recombination rate between markers l and m.

Given the marker phase of sire i, the probability $pr\left(d_{ij}^x = q/hs_i, M_i\right)$ that the progeny received the paternal chromosomal segment at location x of origin q ($q = 1$ or 2) is derived from the two nearest flanking informative markers l^l and l^r, according to Table I.

The information content at location x could be measured by

$$I_x = \frac{\sum_i \sum_j \left(pr\left(d_{ij}^x = 1/hs_i, M_i\right) - pr\left(d_{ij}^x = 2/hs_i, M_i\right)\right)^2}{\sum_i n_i} \tag{2}$$

where n_i is the number of progeny of sire i.

In the simplest and most usual model, these transmission probabilities are used in a within-family regression analysis, as follows:

$$y_{ij} = \mu_i + \left(2pr\left(d_{ij}^x = 1/hs_i, M_i\right) - 1\right)\frac{\alpha_i^x}{2} + \varepsilon_{ij}^x \tag{3}$$

where y_{ij} is the phenotype of progeny j, μ_i is the family mean, α_i^x is the within-sire substitution effect of the QTL at location x, and ε_{ij}^x is the residual.

Table I. Probability of grandpaternal origin d_{ij}^{x} of a chromosomal segment at location x, given marker information M_i and sire haplotype hs_i.

Situation	$pr\left(d_{ij}^{x} = q/hs_i, M_i\right)$
$q = \gamma_{ij}^{l_l}$ and $\gamma_{ij}^{l_l} = \gamma_{ij}^{l_r}$	$\frac{\left(1 - r\left(l_l, x\right)\right)\left(1 - r\left(l_r, x\right)\right)}{1 - r\left(l_l, l_r\right)}$
$q \neq \gamma_{ij}^{l_l}$ and $\gamma_{ij}^{l_l} = \gamma_{ij}^{l_r}$	$\frac{r\left(l_l, x\right) r\left(l_r, x\right)}{1 - r\left(l_l, l_r\right)}$
$q = \gamma_{ij}^{l_l}$ and $\gamma_{ij}^{l_l} \neq \gamma_{ij}^{l_r}$	$\frac{\left(1 - r\left(l_l, x\right)\right) r\left(l_r, x\right)}{r\left(l_l, l_r\right)}$
$q = \gamma_{ij}^{l_r}$ and $\gamma_{ij}^{l_l} \neq \gamma_{ij}^{l_r}$	$\frac{r\left(l_l, x\right)\left(1 - r\left(l_r, x\right)\right)}{r\left(l_l, l_r\right)}$

l_l and l_r are the left and right flanking markers of position x, respectively; $r(a, b)$ is the recombination rate between positions a and b; $q = 1$ or 2 when the paternal QTL allele originates from grandpaternal or grandmaternal chromosomes, respectively; $\gamma_{ij}^{k} = 1$ or 2 when the paternal allele for marker k originates from grandpaternal or grandmaternal chromosomes, respectively.

2.2. Determination of the maternal chromosome transmitted to the progeny

The paternal chromosome could be deducted by comparing the sire genotype with the progeny phenotype marker by marker. There is some uncertainty, however, when both sire and progeny have the same heterozygous marker phenotype. Therefore, the probability of each paternal chromosome pc_{ijK} possibly transmitted to progeny j was determined by using the linkage information.

$$pr\left(pc_{ijk}\right) = \frac{\sum_{hs_i} pr\left(hs_i\right) * pr\left(pc_{ijk}/hs_i\right)}{\sum_{k}\sum_{hs_i} pr\left(hs_i\right) * pr\left(pc_{ijk}/hs_i\right)} \tag{4}$$

where pc_{ijk} was the k-th possible chromosome of paternal origin transmitted to progeny j by its sire i, $pr(hs_i)$ was already derived in equation (1), $pr\left(pc_{ijk}/hs_i\right) = \frac{1}{2}\prod_l pr\left(\gamma_{ij}^{l}, \gamma_{ij}^{l+1}/hs_i\right)$ and was computed as in 1.1. Note that the possible uncertainty in the sire phase determination was accounted for in this approach.

Once the paternal chromosome was determined (in probability), the maternal chromosome was obtained by the complement of the paternal chromosome to

the progeny phenotype. And of course, the probability of each possible maternal chromosome was equal to the probability of the corresponding paternal chromosome. In the following steps, two options are possible. In the simplest implementation, the most likely maternal chromosome is supposed to be the true one. Alternatively, all maternal chromosomes with a probability greater than a given threshold (0.05 for instance) can be considered, and not only the most likely. The latter option was used in the subsequent steps.

2.3. Estimation of MGS marker phenotypes

Each maternal allele transmitted to the progeny could originate from the MGS or from the maternal grandam (MGD). MGD were assumed to be unrelated to each other and unrelated with the MGS. Note that this assumption is similar to that used for dams in the usual daughter design. For each locus L, the likelihood was written as follows:

$$\begin{aligned}\Lambda^L &= \prod_i \Lambda_i^L \\ &= \prod_i \sum_p pr\left(mgs_{ip}^L\right) \prod_{j=1}^{n_i} \sum_k pr\left(pc_{ijk}\right) \sum_{o=1}^{2} pr\,(o) * pr\left(a_{ijk}^L/o, mgs_{ip}^L, \mathbf{f}\right) \quad (5)\end{aligned}$$

where Λ_i^L was the likelihood of MGS i and locus L, msg_{ip}^L was the p-th possible phenotype of MGS i at locus L, o represented the possible origins (1 = MGS, 2 = MGD) of the allele a_{ijk}^L at locus L of the maternal chromosome received by progeny j, and $\mathbf{f}$ was the vector of allele frequencies for locus L in the MGD population. The domain of possible phenotypes for MGS i was determined by the $na_{iL}(na_{iL}+1)/2$ possible pairs of the na_{iL} alleles found in his descendants. The *a priori* probability $pr(msg_{ip}^L)$ of each pair was supposed to be constant and equal to $2/\left(na_{iL}(na_{iL}+1)\right)$. The MGS and MGD origins were supposed to have the same probability, therefore $pr(o=1) = pr(o=2) = 1/2$. The last term, $pr\left(a_{ijk}^L/o, mgs_{ip}^L, \mathbf{f}\right)$, was obtained as presented in Table II.

The frequencies ($\mathbf{f}$) were estimated by maximising the likelihood and the probability of each possible MGS phenotype given the marker information was obtained by

$$pr\left(mgs_i^L/M_i\right) = \frac{\prod_{j=1}^{n_i} \sum_k pr\left(pc_{ijk}\right) \sum_{o=1}^{2} pr\,(o) * pr\left(a_{ijk}^L/o, mgs_i^L, \hat{\mathbf{f}}\right)}{\sum_{mgs_i} \prod_{j=1}^{n_i} \sum_k pr\left(pc_{ijk}\right) \sum_{o=1}^{2} pr\,(o) * pr\left(a_{ijk}^L/o, mgs_i^L, \hat{\mathbf{f}}\right)}. \quad (6)$$

Again, two options were possible in the following steps. In the simplest implementation, the most likely MGS marker phenotype is supposed to be the

Table II. Probability of the maternal allele a^L_{ijk} of marker L received by progeny j, given its grandparental origin o, the MGS phenotype $mgs^L_i = (mgs^{L,1}_i, mgs^{L,2}_i)$, and allelic frequencies $\mathbf{f}$ in the MGD population.

Origin o	Phenotype mgs^L_i of MGS i at locus L	Allele a^L_{ijk} of chromosome k of progeny j at locus L	$pr\left(a^L_{ijk}/o, mgs^L_{ip}, \mathbf{f}\right)$
MGS	$mgs^{L,1}_i = mgs^{L,2}_i$	$a^L_{ijk} = mgs^{L,1}_i$	1
		$a^L_{ijk} \neq mgs^{L,1}_i$	0
MGS	$mgs^{L,1}_i \neq mgs^{L,2}_i$	$a^L_{ijk} = mgs^{L,1}_i$ or $a^L_{ijk} = mgs^{L,2}_i$	$\frac{1}{2}$
		$a^L_{ijk} \neq mgs^{L,1}_i$ and $a^L_{ijk} \neq mgs^{L,2}_i$	0
MGD			$f\left(a^L_{ijk}\right)$

true one. Alternatively, all MGS marker phenotypes with a probability greater than a given threshold (0.05 for instance) can be considered, and not only the most likely. The latter option was used in the estimation of MGS phases.

2.4. Probability of MGS phases

As the allelic frequencies in the MGD population were already estimated in the preceding step, they were assumed to be known and the phase probabilities were estimated within-family. If the chromosome is marked with L locus and if MGS i has nt^l_i possible phenotypes for locus l with a non-zero probability, there are $2^{L-1}\prod_{l=1}^{L} nt^l_i$ possible phases.

The probability of a given phase (hgs_i) of MGS i was proportional to

$$pr\left(mgs_i/M_i\right)\frac{1}{2^{Lh_i-1}}\prod_j\sum_k pr\left(pc_{ijk}\right)\sum_{o=1}^{4^L} pr\left(o\right)pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right) \quad (7)$$

i.e. the product of the *a priori* probability of the phase with the likelihood of the progeny.

If the maternal chromosomes transmitted to the progeny and the MGS marker phenotypes were assumed to be known, this formula reduced to

$$\frac{1}{2^{Lh_i-1}}\prod_j\sum_{o=1}^{4^L} pr\left(o\right)pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right).$$

Let us develop the different terms involved in (7). The *a priori* probability of the phase was equal to the product of the *a priori* probability of the phase given the phenotype for all locus of the chromosome with the probability of

the phenotype $pr(mgs_i)$. Given the phenotype, all phases were considered equally probable, with probability $1/2^{Lh_i-1}$, where Lh_i was the number of heterozygous locus for MGS i. The probability of the phenotype at each locus was obtained independently of the other locus, therefore the probability of the whole phenotype was the product of the probabilities of the phenotype at each locus: $pr(msg_i) = \prod_{l=1}^{L} pr(msg_i^l)$.

Given the genotype hgs_i (*i.e.* the phenotype and the phase), the contribution of each progeny was the sum, over all possible maternal chromosomes (or, equivalently, over all possible paternal chromosome pc_{ijk}) transmitted to the progeny, of the *a priori* probability $pr(pc_{ijk})$ of this chromosome (*a priori*, *i.e.* given the sire information only) times the transmission probability $pr\left(t_{ijk}/pc_{ijk}, hgs_i, \mathbf{f}\right)$ of this maternal chromosome to the progeny.

Four different possible grandparental origins were defined for each marker locus. Origins 1 and 2 were MGS chromosome 1 and 2, respectively, given MGS was heterozygous for this marker; origin 3 was any of MGS chromosomes 1 or 2, given MGS was homozygous for this marker; and origin 4 was MGD origin. To compute this probability $pr\left(t_{ijk}/pc_{ijk}, hgs_i, \mathbf{f}\right)$, t_{ijk} was conditioned with these origins:

$$pr\left(t_{ijk}/pc_{ijk}, hgs_i, \mathbf{f}\right) = \sum_{o=1}^{4^L} pr\left(o\right) * pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right). \tag{8}$$

The probability of a combination of different origins, $pr(o)$, regardless of the different alleles, was obtained as the product $pr(o) = \pi_1 \cdot \pi_2$ of 2 probabilities: the probability π_1 that the grandparent transmitted the combination o, and the probability π_2 that the dam transmitted the combination of origins received by the progeny, given she received o. The probability π_1 that the MGD transmitted origin 4 was trivial and equal to 1. The probability π_1 that the MGS transmitted a given combination o of origins 1 and 2 was obtained by considering the successive couples of markers informative for MGS i: $\pi_1 = 0.5 \prod_{l=2}^{L} p(o_{l+1}/o_l)$, with $p(o_{l+1}/o_l) = 1 - r(l, l+1)$ if $o_l = o_{l+1}$ and $p(o_{l+1}/o_l) = r(l, l+1)$ otherwise. π_2 was computed in a similar way, by differentiating origins 4 and non-4. $\pi_2 = 0.5 \prod_{l=2}^{L} p(o_{l+1}/o_l)$, with $p(o_{l+1}/o_l) = 1 - r(l, l+1)$ if o_l and o_{l+1} were both 4 or both non-4, and $p(o_{l+1}/o_l) = r(l, l+1)$ otherwise.

Finally, the last term $pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right)$ was obtained by the product of the transmission probabilities at each locus, as the linkage between locus was already accounted for in $pr(o)$:

$$pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right) = \prod_{l=1}^{L} pr\left(t_{ijk}^l/o^l, pc_{ijk}^l, hgs_i^l, f^l\right).$$

And $pr\left(t_{ijk}^l/o^l, pc_{ijk}^l, hgs_i^l, f^l\right)$ is shown in Table III.

Table III. Probability of transmission of the maternal allele A of marker l to the progeny, given its grandparental origin o^l, the MGS genotype $hgs_i^l = (hgs_i^{l,1}, hgs_i^{l,2})$, and allelic frequencies $\mathbf{f}$ in the MGD population.

o^l	$hgs_i^{l,1}$	$hgs_i^{l,2}$	$pr\left(t_{ijk}^l/o^l, pc_{ijk}^l, hgs_i^l, \mathbf{f}^l\right)$
4			$f^l(A)$
1, 2, or 3	$\forall B \neq A$	$\forall C \neq A$	0
1 or 2	A	A	0
3	A	B	0
3	A	A	1
1	A	$\forall B \neq A$	1
1	$\forall B \neq A$	A	0
2	A	$\forall B \neq A$	0
2	$\forall B \neq A$	A	1

Origins o^l 1 and 2 were MGS chromosome 1 and 2, respectively, given MGS was heterozygous for this marker; origin 3 was any of MGS chromosomes 1 or 2, given MGS was homozygous for this marker; and origin 4 was MGD origin.

The most likely MGS genotype was the genotype $\hat{hgs}_i$ such that

$$\hat{hgs}_i = \text{Arg}\max_{hgs_i} \times \left[pr(mgs_i) \frac{1}{2^{Lh_i-1}} \prod_j \sum_k pr\left(pc_{ijk}\right) \sum_{o=1}^{4^L} pr(o)\, pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right) \right]. \tag{9}$$

Subsequently in this paper, only the most likely MGS genotype was retained and considered as the true one.

2.5. Probability of origin of the maternal chromosomal segment of the progeny at location x

A similar approach has been proposed by Coppieters *et al.* [4]. Three chromosomal origins were considered, the first and the second chromosomes of the MGS (origins $q = 1$ and 2), with an equal *a priori* probability of 0.25, and the MGD (origin $q = 3$), with an *a priori* probability of 0.5. The objective was to increase one of these probabilities to one, and to decrease the two others to zero, by using the marker information.

Let us denote $pr\left(e_{ij}^x = q/pc_{ijk}, hgs_i, \mathbf{f}\right)$ the probability that the chromosomal segment at location x has origin q ($q = 1, 2,$ or 3) given the marker information,

the genotype of the MGS and the marker allele frequencies in the MGD population.

Only the most likely maternal chromosome transmitted to the progeny was accounted for. However, this chromosome was more accurately defined than in formula (4) by using the information of MGS genotypes and allele frequencies in the MGD population, in addition to the sire genotypes. The maternal chromosome transmitted to the progeny was assumed to be known and was defined by

$$\hat{t}_{ij} = \text{Arg}\max_{pc_{ijk}} \left[pr\left(pc_{ijk}\right) \sum_{o=1}^{4^L} pr\left(o\right) * pr\left(t_{ijk}/o, pc_{ijk}, hgs_i, \mathbf{f}\right) \right].$$

As in 2.4, we considered the same four different possible origins (o) and

$$pr\left(e_{ij}^x = q/hgs_i, t_{ij}, \mathbf{f}\right) = \sum_{o=1}^{4^L} pr\left(o/hgs_i, t_{ij}, \mathbf{f}\right) pr\left(e_{ij}^x = q/o\right). \qquad (10)$$

According to Bayes theorem,

$$pr\left(o/hgs_i, t_{ij}, \mathbf{f}\right) = \frac{pr\left(t_{ij}/o, hgs_i, \mathbf{f}\right) * pr\left(o\right)}{\sum_{o=1}^{4^L} pr\left(t_{ij}/o, hgs_i, \mathbf{f}\right) * pr\left(o\right)},$$

which terms were already computed in 2.4.

The last term to develop is $pr\left(e_{ij}^x = q/o\right)$. As in 1.4, this probability was computed as the product of the probability π_1 that the grandparent transmitted segment q to the dam with the probability π_2 that the dam transmitted segment q to the progeny. The probability that the MGD transmitted a segment of origin $q = 4$ was 1. The probability that the MGS transmitted a segment of origin 1 or 2 was deducted from the nearest flanking informative markers l_l et l_r, in a similar way as in Table I. Obviously, when only one marker was informative (say l_l), π_1 reduced to $r(l_l, x)$ or $1 - r(l_l, x)$, and to 0.5 when no marker was informative.

π_2 was obtained in a similar way, but the two nearest flanking markers were considered, and not only the two nearest informative ones. A recombination was supposed to occur when the flanking marker allele origins were 4 and non-4 (1, 2, or 3).

With such a method, $pr\left(e_{ij}^x = q/pc_{ijk}, hgs_i, \mathbf{f}\right)$ could be computed at any point x of the genome. As for sires, the MGS information content was measured by

$$I_x = \frac{\sum_i \sum_{j=1}^{n_i} \left[pr\left(e_{ij}^x = 1/t_{ij}, hgs_i, \mathbf{f}\right) - pr\left(e_{ij}^x = 2/t_{ij}, hgs_i, \mathbf{f}\right)\right]^2}{\sum_i n_i}. \qquad (11)$$

Without distorsion of segregation in the dam population, this criterion theoretically varies from 0 to 0.5, as half the chromosomal segments originate from the MGD population.

2.6. QTL detection by regression

The linear model used in the daughter design was easily extended to take into account maternal information, as follows:

$$\begin{aligned} y_{ijk} = \mu_i + \nu_j + \left(2pr\left(d^x_{ijk} = 1/hs_i, M_i\right) - 1\right)\frac{\alpha^x_i}{2} \\ + \left(pr\left(e^x_{ijk} = 1/t_{ij}, hgs_i, \hat{\mathbf{f}}\right) - pr\left(e^x_{ijk} = 2/t_{ij}, hgs_i, \hat{\mathbf{f}}\right)\right)\frac{\alpha^x_j}{2} \\ + pr\left(e^x_{ijk} = 3/t_{ij}, hgs_i, \hat{\mathbf{f}}\right)\beta^x_j + \varepsilon^x_{ijk} \end{aligned} \quad (12)$$

where ν_j was the polygenic effect of MGS j, α^x_j was the within-MGS substitution effect of the QTL at location x, β^x_j was the average QTL effect transmitted by the MGD mated to MGS j, and the other terms were the same as in formula (3). It is noteworthy that QTL could be detected in the MGD population, even when MGS are homozygous, provided that the average MGD QTL effect differs from the MGS QTL effect.

2.7. Signification test

Because MGS are usually not nested within-sire, permutation tests [3] cannot easily be applied. The empirical distribution of test statistics under H0 was obtained alternatively by simulation. Phenotypes were simulated under H0, as the sum of the randomly sampled effects of the sire, the MGS and the residual, and, optionally, with a paternal QTL effect, but without any maternal QTL effect. Different tests could be considered to detect QTL specifically in sires, MGS, or MGD populations, or in the overall design.

3. MEASURE OF THE EFFICIENCY OF THE METHOD

The efficiency of the method was assessed by simulation, in a first step by measuring the quality of the MGS phenotype and genotype reconstruction, and in a second step by measuring the power of QTL detection in the MGS and MGD populations.

The simulated design included 1 000 progeny distributed in 10 sire families of size 100. The heritability of the trait was assumed to be 0.25. Sires and progeny were genotyped for 3, 5, 6, or 9 markers evenly spaced over a 100 cM long chromosome (*i.e.* with 50, 25, 20, or 12.5 cM intervals). Each marker had

5 alleles with equal frequencies 0.2. The number of MGS was assumed to be 5, 10, 20, or 50, corresponding to family size of 200, 100, 50, and 20 respectively. Sixteen sets of parameters were defined by combining the number of markers with the number of MGS. Factors sire and MGS were orthogonal. Markers and QTL were assumed to be in linkage equilibrium in the sire, MGS and MGD populations.

To estimate the quality of phenotype and genotype reconstruction, 100 replicates for each of the 16 sets of parameters were simulated. According to the number of MGS per simulation, the results were based on 1 500 to 45 000 phenotype reconstructions and 500 to 5 000 genotype reconstructions. A phenotype was classified as correct if both inferred alleles were those simulated. Only chromosomes with a correct phenotype reconstruction for all markers were used to measure the efficiency of the phase reconstruction, in order to specifically measure the efficiency of this step.

To estimate the power of QTL detection, the H1 hypothesis was simulated with one biallelic QTL at location $x = 35$ cM with an additive effect. The allele frequencies were equal to 0.5. The substitution effect was assumed to be 0.5, 0.7, or 1 phenotypic standard deviation. The test distribution under H0 was estimated from 10 000 replicates for each of the 16 sets of parameters. Under H1, 100 replicates were simulated for each of the 48 sets of parameters (defined by combining the number of markers, the number of MGS and the QTL effect). The detection power was estimated by the proportion of test values under H1 exceeding the 95% percentile of the corresponding H0 empirical distribution.

Several tests could be defined, based on any combination of sire, MGS, and MGD estimated effects. In QTL detection, the overall test combining all information is the best. In this simulation study to estimate the detection power, the statistical test used was based on MGS and MGD QTL contributions only.

4. APPLICATION TO THE FRENCH GRANDDAUGHTER DESIGN IN DAIRY CATTLE

A GDD design has been carried out in France [2]. It included 1 548 artificial insemination bulls distributed in 14 sire families, genotyped for 169 markers, and evaluated for up to 25 traits (milk production and composition, type, fertility, mastitis resistance) on the basis of their daughters information. The 14 largest MGS families were selected and included 939 grandsons. The MGS family size ranged from 24 to 206. The present method was applied to this subset for the 29 autosomal chromosomes and for the 17 traits with the most complete information. Empirical thresholds were obtained by simulating 10 000 H0 replicates per chromosome.

Table IV. Percentage of correct phenotype reconstruction for each MGS marker (100 replicates, 1 500 to 45 000 results per set of parameters).

MGS family size	Distance between markers (cM)			
	50	25	20	12.5
200	95.5	99.3	99.9	99.9
100	89.8	97.4	99.2	99.9
50	82.1	94.2	95.8	98.8
20	70.0	81.2	86.5	93.5

Table V. Percentage of correct MGS marker phase reconstruction (100 replicates, 500 to 5 000 results per set of parameters).

MGS family size	Distance between markers (cM)			
	50	25	20	12.5
200	99.9	100.0	99.9	100.0
100	99.8	99.9	99.8	99.8
50	96.7	98.3	98.6	99.0
20	76.7	82.0	84.5	89.6

5. RESULTS OF THE SIMULATION STUDY

5.1. Efficiency of phenotype and genotype reconstruction

Table IV presents the results of the phenotype reconstruction. The efficiency was found to be very high when the number of grandprogeny was high or when the marker map was dense. With less than 50 descendants, the reconstruction was more hazardous. It is worth noting that very high correct reconstruction rates have to be required here, because each chromosome has several markers and therefore the probability of the correct reconstruction of all phenotypes of the whole chromosome is the product of the probabilities for each marker. Moreover, this step is the first one and would hamper the next steps if it is not correct.

Table V presents the results of the phase reconstruction, assuming that the phenotypes are correct. The probability of correct reconstruction was of course lower than for sires. Mangin *et al.* [6] reported a probability of correct reconstruction of sire phase of 95% with 5 markers, 4 alleles per marker, and 20 progeny, against 82% in the present study for MGS with rather similar assumptions. But the loss of efficiency was close to expected: with 50 grandprogeny, *i.e.* 25 grandprogeny effectively contributing to the determination of MGS phase, the probability reached 98%. As for phenotype reconstruction, the efficiency was found to be very high when the number of grandprogeny was greater than 50. The effect of map density was more limited,

Table VI. Percentage of correct MGS marker phenotype and genotype reconstruction (100 replicates, 500 to 5 000 chromosome results per set of parameters).

MGS family size	Distance between markers (cM)			
	50	25	20	12.5
200	87.1	96.5	99.3	99.5
100	72.4	87.7	95.2	99.4
50	53.6	72.9	76.4	89.1
20	26.3	29.0	35.3	48.8

probably because when markers are more spaced, the number of markers and, therefore, the number of possible phases, is much lower.

When MGS were not genotyped and when both steps were combined (Tab. VI), reconstruction results were excellent for large families and dense maps but declined, with a cumulative effect, with family size and sparsity of the map.

5.2. Power of QTL detection with maternal information

Results of detection power are presented in Table VII. They clearly show that the key factor was the family size and that families with more than 50 descendants were the most informative. Detection power was of course much lower with the maternal information than with the paternal information. For instance, with 20 MGS with 50 descendants, the detection power was 26% for a QTL with a substitution effect of 0.5, whereas the corresponding value with sire information, obtained with 20 sires and 50 progeny per sire, was 60% (data not shown). But when the comparison was made with the same effectively contributing family size, *i.e.* with MGS sires families twice as large as sire families, results were more similar (52%), showing that only little information was lost with large families.

6. ILLUSTRATION WITH THE FRENCH GDD IN DAIRY CATTLE

The average information content over the 29 chromosomes, as defined in (2) and (11), reached 0.55 for paternal chromosomes and 0.18 for maternal chromosomes. This latter value could appear rather low, when compared to its maximum theoretical value (0.5). Figure 1 illustrates a situation (chromosome 6, $x = 10$ cM) corresponding to an information content of 0.22 and shows that such an apparently low value corresponds, however, to reasonably well estimated transmission probabilities.

The *a priori* probabilities are 0.25, 0.25, and 0.5 for $q = 1$, $q = 2$, and $q = 3$, respectively. On the figure, the frequency around this point was

Table VII. QTL Detection power with maternal information according to QTL substitution effect, MGS family size, and map density (chromosomewise type 1 error = 5%, 10 000 replicates under H0, 100 replicates under H1 per set of parameters).

Substitution effect (phenotypic standard deviation)	MGS family size	Distance between markers (cM)			
		50	25	20	12.5
0.5	200	0.25	0.42	0.49	0.70
	100	0.12	0.24	0.40	0.43
	50	0.13	0.15	0.26	0.26
	20	0.06	0.08	0.10	0.12
0.7	200	0.50	0.79	0.85	0.88
	100	0.40	0.61	0.73	0.87
	50	0.31	0.43	0.52	0.58
	20	0.08	0.13	0.14	0.19
1.0	200	0.87	0.99	1.00	0.99
	100	0.76	0.97	0.98	0.98
	50	0.47	0.80	0.93	0.95
	20	0.20	0.30	0.34	0.40

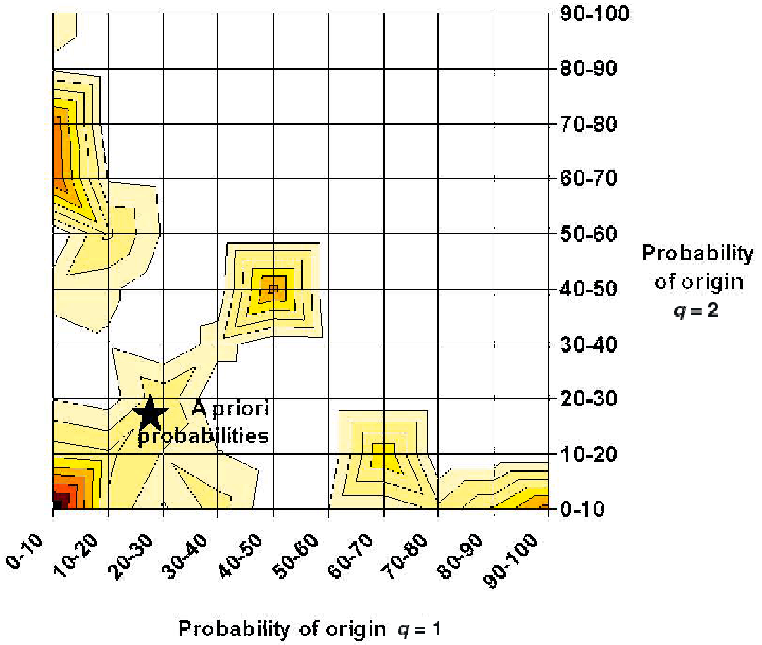

Figure 1. Distribution of the probabilities of origin of the maternal chromosomal segment for 939 progeny at location 10 cM on chromosome 6 (information content = 0.22).

strongly reduced, with about 10% remaining individuals, and it was replaced by 5 other peaks: $pr(q = 1) < 0.1$ and $pr(q = 2) < 0.1$, corresponding to the GDD origins, and gathering 33% of the individuals; $pr(q = 1) > 0.8$ and $pr(q = 2) < 0.2$, corresponding to the MGS chromosome 1 origin, with 15% of the individuals; $pr(q = 1) < 0.2$ and $pr(q = 2) > 0.8$, corresponding to the MGS chromosome 2 origin, with 20% of the individuals; $pr(q = 1)$ and $pr(q = 2)$ close to 0.5, corresponding to a MGS origin, but without differentiating chromosome 1 from chromosome 2 origins, with 10% of the individuals. The last concentration ($pr(q = 1) = 0.6$ and $pr(q = 2) < 0.1$) corresponded to an uncertainty between the MGS chromosome 1 origin and the MGD origin, whereas MGS chromosome 2 origin was excluded. Therefore, in spite of an apparently low information content (0.22), the grandparental origin (MGS or MGD) was known in 80% of cases, and when a MGS origin was determined, the chromosomal origin was deducted in more than 80% of the cases. This misleading behaviour could be attributed to the definition of the information content criterion, based on a sum of squares.

Table VIII presents the QTL detected, with the maternal information only, with a genomewise significance level of 10%. Note that this 10% genomewise threshold corresponds to a 0.34% chromosomewise significance level for 29 chromosomes. Twenty-two significant (or almost significant) results were tabulated, out of them 8 and 5 were also genomewise and chromosomewise significant, respectively, with within-sire analysis, while 9 were not detected at all with the sire information only.

Table IX presents 8 other QTL chromosomewise significant when only the paternal or the maternal information was used, but which became genomewise significant when both kinds of information were combined.

7. DISCUSSION AND CONCLUSION

In non-prolific outbred populations, QTL detection is usually carried out within-sire families and all the maternal meioses, *i.e.* half the potential information, are lost. Although the present approach does not retrieve such an amount of information, it takes advantage of the relationship between dams, without increasing the genotyping work. A similar method was already proposed by Coppieters *et al.* [4] but it was limited to the MGS already present as sires, *i.e.* already genotyped and with known phases. Moreover, these authors did not account for the grandmaternal origins and, therefore, lost some information. The goal of their approach was primarily to confirm results obtained in specific families with independent data already present in the same design. Our approach is more general and makes it possible to include new families, provided that their size is sufficiently large. In the case sires are also MGS, progeny and grandprogeny information are easily merged by

Table VIII. Genomewise significant QTL detected with the maternal information.

Trait	Chromosome	Chromosome wise *p*-value (%), maternal information	Chromosome wise *p*-value (%), paternal information
Milk	26	0.23	2.15
Fat	14	0.35	0.11
	19	0.09	0.13
Fat content	12	0.02	
	14	< 0.01	< 0.01
Protein content	14	< 0.01	< 0.01
Cell count	3	0.09	
	15	0.06	0.05
	24	0.14	
	25	0.19	
Ligament	1	0.14	
	28	0.37	0.27
Udder depth	11	0.12	3.18
Udder balance	18	0.13	0.05
	25	0.33	
Teat distance	17	0.37	1.97
	24	0.18	
	28	0.06	0.64
Teat placement	11	0.08	
	28	0.02	0.04
Height at sacrum	26	0.23	
Chest depth	11	0.34	2.74

Table IX. Previously putative QTL confirmed by the maternal information.

Trait	Chromosome	Chromosome wise *p*-value (%), maternal information 939 sons	Chromosome wise *p*-value (%), paternal information 1 548 sons
Fat content	22	1.50	2.31
	24	4.98	4.03
Protein content	3	1.18	0.98
Ligament	10	0.54	1.77
Udder depth	20	4.04	1.01
Udder balance	14	0.75	3.01
Teat distance	28	0.97	0.64
Teat placement	21	2.59	4.33
Rump width	29	4.48	2.35

estimating only one QTL effect in the linear model. This approach is rather specific to (grand)daughter designs and it significantly increases programming complexity in software development, but it is an efficient alternative to MCMC methods [1] in term of computing time requirements. Once implemented, it is easy to apply to the analysis of a complete design, with many traits and a genome scan.

In the present approach, only the most likely MGS marker phase is finally retained, which greatly simplifies the QTL detection step by linear regression, whereas the uncertainty about maternal marker allele transmission and MGS marker phenotype is accounted for in the first steps. This approach is a better guarantee of correct MGS phenotype and phase reconstruction, particularly when markers are little informative.

Because MGS QTL are compared with MGD QTL, all MGS are informative, even if they are homozygous for the QTL. Moreover, assuming that the expected MGD QTL effects are the same across MGS families, this provides a basis for MGS comparisons. This means that the absolute values of the QTL genotype carried by MGS could be predicted, regardless of their homo- or heterozygous status, and that MGS homozygous for favourable QTL alleles could be distinguished from MGS homozygous for unfavourable QTL alleles. With the additional assumption that the QTL is biallelic, allele frequencies could be estimated and the genotype of each homozygous MGS could be predicted, with only a minor modification of the QTL detection step. This frequency information is extremely valuable to assess the potential genetic gain due to one given QTL and achievable by marker-assisted selection.

The present method, however, is limited to large MGS families. The analysis of families with less than 50 descendants has limited power, mainly because many errors could arise and accumulate during the successive steps of the procedure. The efficiency of the phenotype reconstruction is also affected by the marker informativity and also by the number of MGS. With few MGS and biallelic markers, the efficiency of the phenotype reconstruction is a bit lower, because the likelihood (5) is quite flat. In the extreme situation, when all MGS have the same phenotype, the likelihood could be bimodal and two possible phenotypes are nearly equally likely. In large designs, however, this situation is rarely encountered and is less limiting than low family size.

Given the MGS and MGD, the dams were assumed to be unselected. Note that this assumption is always made when only sire information is used. This assumption is probably not critical in daughter designs, because it is likely to be fulfilled. But it is clearly not the case in a granddaughter design, because only the best females are selected as dam of sires. The effect of this selection process on the efficiency of the method is not known, but it could be important if this selection generates distortion of segregation, and favours or eliminates some QTL alleles which are analysed subsequently.

The definition of statistical tests is also a critical point. The appealing and popular permutation method [3] to estimate rejection threshold could be difficult to apply when MGS are not nested within-sire, because the sire by MGS cell could be very small. Thresholds should consequently be computed by simulation, which is known to be a bit less conservative [8] than permutation tests.

In the French granddaughter design, this method made it possible to analyse 14 additional bulls and this new information was very valuable, in marker-assisted selection as well as in the fine mapping of some QTL by providing new informative families. With this approach, 16 new QTL were detected, whereas they were either non detected or not significant enough with the paternal information only.

REFERENCES

[1] Bink M.C.A.M., Van Arendonk J.A.M., Detection of quantitative trait loci in outbred population with incomplete marker data, Genetics 151 (2000) 409–420.

[2] Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Milan D., Rupp R., Amigues Y., Boscher M.Y., Levéziel H., La recherche de QTL à l'aide de marqueurs : résultats chez les bovins laitiers, INRA Prod. Anim., Numéro spécial Génétique Moléculaire (2000) 217–222.

[3] Churchill G.A., Doerge R.W., Empirical threshold values for quantitative trait mapping, Genetics 138 (1994) 963–971.

[4] Coppieters W., Kvasz A., Arranz J.J., Grisard B., Farnir F., Riquet J., Georges M., The great-grand-daughter design: a simple strategy to increase the power of a grand-daughter design for QTL mapping, Genet. Res. Camb. 74 (1999) 189–199.

[5] Elsen J.M., Mangin B., Goffinet B., Boichard D., Le Roy P., Alternative models for QTL detection in livestock. I. General information, Genet. Sel. Evol. 31 (1999) 213–224.

[6] Mangin B., Goffinet B., Le Roy P., Boichard D., Elsen J.M., Alternative models for QTL detection in livestock. II. Likelihood approximations and sire marker genotype estimations, Genet. Sel. Evol. 31 (1999) 225–237.

[7] Soller M., Genizi A., The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations, Biometrics 34 (1978) 47–55.

[8] Spelman R., Coppieters W., Karim L., Van Arendonk J.A.M., Bovenhuis H., Quantitative trait locus analysis for five milk production traits on chromosome six in the Dutch Holstein Friesian population, Genetics 144 (1996) 1799–1808.

[9] Weller J.I., Kashi Y., Soller M., Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle, J. Dairy Sci. 73 (1990) 2525–2537.