Genet. Res., Camb. (1999) **74**, *pp.* 189–199. *With 3 figures.* Printed in the United Kingdom © 1999 Cambridge University Press

The great-grand-daughter design: a simple strategy to increase the power of a grand-daughter design for QTL mapping

WOUTER COPPIETERS, ALEXANDRE KVASZ, JUAN-JOSÉ ARRANZ, BERNARD GRISART, JULIETTE RIQUET, FRÉDÉRIC FARNIR AND MICHEL GEORGES*

Department of Genetics of Genetics, Faculty of Veterinary Medicine, University of Liège (B43), 20 Bd de Colonster, 4000 Liège, Belgium

(*Received 10 November 1998 and in revised form 8 March 1999*)

Summary

In dairy cattle, quantitative trait loci (QTL) are usually mapped using the grand-daughter design (GDD), i.e. sets of progeny-tested paternal half-brothers. Linkage information is typically extracted from the segregation of the sire chromosomes amongst their sons. We herein propose to increase the power of a GDD by exploiting the frequently occurring relationship between sires and grandsons which has so far been ignored in most methods of analysis. The proposed approach is a multipoint interval mapping method based on the Wilcoxon sum-of-rank test. Three alternative approaches to combine information from sons and grandsons are evaluated by simulation. In these either (i) sons and grandsons are ranked separately, (ii) sons and grandsons are ranked separately but the sign of the QTL effect is constrained to be the same in both generations, or (iii) sons and grandsons are ranked jointly. The proposed methods have been applied on a real data-set in which a GDD including 907 sons is analysed with a marker map comprising nine microsatellites spanning 46 cM on bovine chromosome 6.

1. Introduction

In 1990, Weller *et al.* proposed an experimental design that takes advantage of the progeny-testing procedure which is routinely applied to select elite dairy bulls, in order to efficiently map quantitative trait loci (QTL) influencing milk production. In this proposed 'grand-daughter design' (GDD), the analysed pedigree material consists of sets of half-brothers sharing a common founder sire. The records analysed to map QTL are the sons' breeding values estimated from the milking performances of their respective daughters (the progeny test). The data would typically be analysed by maximum likelihood, linear regression or rank-based methods measuring the contribution (nested within founder sires) of alternate paternal alleles to the trait variance (e.g. Georges *et al.*, 1995; Knott *et al.*, 1996; Coppieters *et al.*, 1998*b*). Inferences about which paternal allele is transmitted to a given son at a defined chromosome position are usually made by multipoint analysis, i.e. considering information from all linked markers jointly. Sons typically have 50 or more daughters, yielding breeding value estimates with reliabilities of the order of 85–95%. Squared reliabilities can be compared with heritabilities of the order of 35% for the actual phenotypes as expressed in the daughters. It can be shown that the concomitant reduction in environmental noise leads to a decrease in the required sample size by a factor of the order of 3 to 4 (Weller *et al.*, 1990; Georges *et al.*, 1995). Several studies have demonstrated convincingly that this design may indeed allow for the mapping of QTL in elite dairy cattle populations (e.g. Georges *et al.*, 1995; Spelman *et al.*, 1996; Gomez-Raya *et al.*, 1996; Kühn *et al.*, 1996; Ron *et al.*, 1998; Coppieters *et al.*, 1998*a*; Arranz *et al.*, 1998).

Despite the considerable gain in power that can be achieved using this approach when compared with a daughter design (DD), the size of most GDDs that have been assembled to date has been limited by sample availability and essentially provides inadequate power to detect QTL with moderate effects when performing whole genome scans and applying the commonly used statistical procedures. Much attention has therefore been devoted to devising strategies that

* Corresponding author. Tel: +32 (0)4 366.41.50. Fax: +32 (0) 366.41.22. e-mail: michel.georges@ulg.ac.be.

would increase the amount of information that is extracted from the available data set. One option is to exploit additional familial relationships that exist amongst members of the grand-daughter design and which are presently being ignored. The ultimate goal would be to account for all known pedigree relationships. While substantial progress has been made towards so-called full pedigree analysis (e.g. Hoeschele *et al.*, 1997; Bink & van Arendonk, 1998), these methods still face a number of computational issues which make them difficult to apply yet on very large data-sets, especially when one attempts multipoint linkage analyses.

In this paper, we propose a simple strategy that increases power of detection by taking advantage of a frequently occurring relationship found in most GDDs, i.e. the fact that many sons share a limited number of common grandsires. The resulting scheme is referred to as the 'great-grand-daughter design' (G^2DD). We show in this paper how a previously described non-parametric rank-based statistical method can conveniently be extended to analyse a G^2DD.

2. Materials and methods

2.1. General principles of the great-grand-daughter design

As illustrated in Fig. 1, the only relationships providing linkage information in the GDD are the connections between sires and their sons. Examination of the pedigree records, however, shows that many bulls have additional, potentially informative links. In particular, sons often share common grandsires which can be part of the GDD themselves as sires of sons. The objective of the G^2DD is to exploit the relationships between a founder sire and its grandsons present in the available pedigree material.

Assume that a sire is heterozygous Qq for a QTL at a given chromosome position. Its sons will fall into two QTL genotype classes – *Q?* and *q?* – with phenotypic difference, Δ, equal to the average effects of the $Q \rightarrow q$ allele substitution (Falconer & Mackay, 1996). Grandsons of the corresponding sire will fall in three QTL genotypic classes – *Q?*, *q?* and *??* – with frequency of 0·25, 0·25 and 0·50, respectively. Mean phenotypic values of *Q?* versus *q?* grandsons differ by the same amount, Δ, as in the sons.

In the GDD, the *Q?* versus *q?* QTL genotype probabilities of the sons are inferred from flanking marker genotypes and used to estimate Δ. In the G^2DD it is simply proposed to extend this to grandsons, i.e. to infer the QTL genotype probabilities (*Q?*, *q?* and *??*) from flanking marker genotypes and combine both sources of information (sons and grandsons) to increase the power of QTL mapping with the available marker genotype information. Note that only half the grandsons inherit either of the grandpaternal alleles and therefore contribute information in the proposed G^2DD.

2.2. G^2DD: QTL mapping procedure

(i) *Determination of the most likely linkage phase of the sires*

The marker linkage phase is determined for each sire from the marker genotypes of his sons. This is

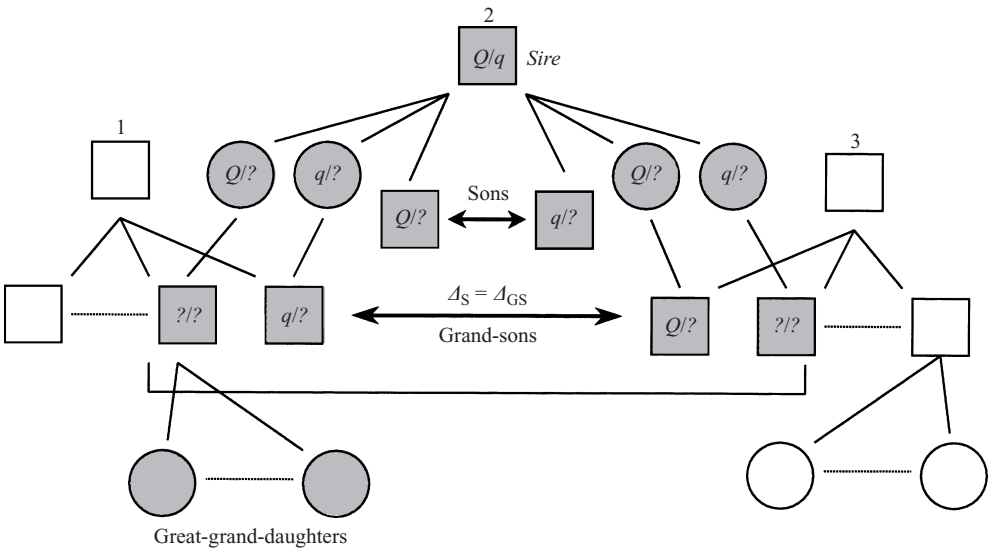

Fig. 1. Schematic representation of the G^2DD. Three sires are shown with sons. In the regular GDD (Weller *et al.*, 1990), sons are sorted according to the paternal allele inherited, yielding a phenotypic contrast Δ_S between *Q?* and *q?* sons. The G^2DD exploits the fact that many sons of sires are also grandsons of sires as shown for four sons of sires 1 or 3 which are also grandsons of sire 2. These grandsons can be sorted according to the grandpaternal allele inherited: *Q?*, *q?* or *??*. *Q?* and *q?* grandsons differ by a phenotypical contrast $\Delta_{GS} = \Delta_S$.

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

accomplished by calculating the likelihood of the corresponding half-sib pedigree data under the $2^x/2$ possible phases (assuming x informative markers) as follows (Georges *et al.*, 1995):

$$L_i = \prod_{j=1}^{n}\left[\sum_{k=1}^{2^x}[P(k\,|\,i)\times\prod_{m=1}^{x}AFM_m]\right],$$

where

L_i is the likelihood of the pedigree data for linkage phase i;

$\prod_{j=1}^{n}$ is the product over all n half-sibs;

$\sum_{k=1}^{2^x}$ is the sum over all possible sire's gametes k;

$P(k\,|\,i)$ is the probability of gamete k given Mendelian laws, phase i and known recombination rates between adjacent markers, θ_1 to θ_x;

$\prod_{m=1}^{x}$ is the product over all m markers within the synteny group;

AFM_m is the population frequency of the obliged maternal marker allele of marker m, given the paternal gamete k.

Table 1. *Illustration of the calculation of QTL genotype probabilities* (Q? *or* q?) *of sons given flanking marker genotypes*

Sire	Marker/QTL genotype	θ_1 θ_2 θ_3 1 Q 1 1 2 q 2 2		
Son	Marker phenotype	1 ? 1 2 3 ? 2 3		
	Compatible genotypes	1 Q 1 2 3 ? 2 3	$0{\cdot}5\times(1-\theta_1)(1-\theta_2)\theta_3$ $\times(f_{1,3}\,f_{2,2}\,f_{3,3})^{(a)}$	(a)
		1 Q 2 2 3 ? 1 3	$0{\cdot}5\times(1-\theta_1)\theta_2(1-\theta_3)$ $\times(f_{1,3}\,f_{2,1}\,f_{3,3})$	(b)
		1 q 1 2 3 ? 2 3	$0{\cdot}5\times\theta_1\theta_2\theta_3$ $\times(f_{1,3}\,f_{2,2}\,f_{3,3})$	(c)
		1 q 2 2 3 ? 1 3	$0{\cdot}5\times\theta_1(1-\theta_2)(1-\theta_3)$ $\times(f_{1,3}\,f_{2,2}\,f_{3,3})$	(d)
	QTL genotype probabilities	$P[Q?_{(p)}\,\vert\,g_L, g_R]$ $P[q?_{(p)}\,\vert\,g_L, g_R]$	$(a+b)/(a+b+c+d)$ $(c+d)/(a+b+c+d)$	

[(a)] $f_{x,y}$ = population frequency of allele x of marker y.

All marker phases are *a priori* considered to be equally likely, i.e. linkage equilibrium is assumed to have been reached between all markers. The marker phase maximizing the likelihood of the pedigree data is considered the true one and selected for further analysis.

(ii) *Calculation of QTL genotype probabilities* (Q?) *or* q?) *of sons given flanking marker genotypes*

Assuming that the marker phase of the sire is known, and that he is of genotype Qq for a hypothetical QTL with fixed position (p) on the corresponding marker map, the QTL genotype probabilities ($Q?$ or $q?$) of each son can easily be computed given its genotype on flanking markers (Coppieters *et al.*, 1998*b*). Table 1 illustrates how such a calculation is performed. $P_i[Q?_{(p)}\,|\,g_L, g_R]$ is defined as the probability that son i has inherited QTL allele Q from the founder sire at map position (p) given left (g_L) and right (g_R) flanking marker genotypes. Only markers for which the founder sire is heterozygous are considered when computing $P_i[Q?_{(p)}\,|\,g_L, g_R]$. Moreover, while the nearest flanking markers contain all the information needed to compute $P_i[Q?_{(p)}\,|\,g_L, g_R]$ in a given interval when dealing with experimental crosses, information from more distant markers is considered in the outbred half-sib situation, when closer markers are not fully informative. This occurs in the case of missing genotype or when the offspring has the same marker genotype as the sire and the dam is either not genotyped or has the same heterozygous genotype as well. In the former case, part of the information is recovered by considering marker allele frequencies in the population.

(iii) *Calculation of QTL genotype probabilities* (Q?, q? *or* ??) *of grandsons given flanking marker genotypes*

QTL genotype probabilities for the grandsons are determined along the same lines as for the sons. An example of such calculations is given in Table 2. It is assumed in this example that the marker linkage phase of the grandsire and sire are known (Section 1), and that the dams are not marker genotyped.

(iv) *Information content mapping*

Information content along the marker maps (Kruglyak & Lander, 1995*a*; Coppieters *et al.*, 1998*b*) was measured as:

$$\frac{\sum_{i=1}^{n}[P_i(Q?_p\,|\,g_L, g_R)-P_i(q?_p\,|\,g_L, g_R)]^2}{n-1}$$

for the sons and grandsons.

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

Table 2. *Illustration of the calculation of QTL genotype probabilities* (Q?, q? *or* ??) *of grandsons given flanking marker genotypes*

Sire	Son				Grand-sire
Phase-known genotype	Phase-unknown genotype 1 ? 2 2 ? 4				Phase-known genotype
		Paternal allele	Maternal allele		
	1 ? 4	$\frac{1}{2}[\theta_1(1-\theta_2)+(1-\theta_1)\theta_2]$ (a)	2* Q 2*	$\frac{1}{4}\theta_1(1-\theta_1)\theta_2(1-\theta_2)$ (c_1)	
			2* Q 2	$\frac{1}{4}\theta_1(1-\theta_1)\theta_2 f_{2,2}$ (c_2)	
			2 Q 2*	$\frac{1}{4}\theta_1\theta_2(1-\theta_2) f_{1,2}$ (c_3)	
			2 Q 2	$\frac{1}{4}\theta_1\theta_2 f_{1,2} f_{2,2}$ (c_4)	
			2* q 2*	$\frac{1}{4}(1-\theta_1)^2(1-\theta_2)^2$ (d_1)	
			2* q 2	$\frac{1}{4}(1-\theta_1)^2\theta_2 f_{2,2}$ (d_2)	
			2 q 2*	$\frac{1}{4}\theta_1(1-\theta_2)^2 f_{1,2}$ (d_3)	
			2 q 2	$\frac{1}{4}\theta_1\theta_2 f_{1,2} f_{2,2}$ (d_4)	
			2* ? 2*	$\frac{1}{4}[(1-\theta_1)(1-\theta_2)+\theta_1\theta_2]\theta_1\theta_2$ (e_1)	
			2* ? 2	$\frac{1}{4}\theta_1(1-\theta_1) f_{2,2}$ (e_2)	
			2 ? 2*	$\frac{1}{4}(1-\theta_1)\theta_2 f_{1,2}$ (e_3)	
			2 ? 2	$\frac{1}{2}(1-\theta_1)(1-\theta_2) f_{1,2} f_{2,2}$ (e_4)	
1 ? 3	2 ? 4	$\frac{1}{2}(1-\theta_1)(1-\theta_2)$ (b)	1* Q 2*	$\frac{1}{4}(1-\theta_1)^2\theta_2(1-\theta_2)$ (f_1)	1* Q 1*
2 ? 4			1* Q 2	$\frac{1}{4}(1-\theta_1)^2\theta_2 f_{2,2}$ (f_2)	2* q 2*
			1 Q 2*	$\frac{1}{4}\theta_1(1-\theta_2)\theta_2 f_{1,1}$ (f_3)	
			1 Q 2	$\frac{1}{4}\theta_1\theta_2 f_{1,1} f_{2,2}$ (f_4)	
			1* q 2*	$\frac{1}{4}\theta_1(1-\theta_1)(1-\theta_2)^2$ (g_1)	
			1* q 2	$\frac{1}{4}\theta_1(1-\theta_1)\theta_2 f_{2,2}$ (g_2)	
			1 q 2*	$\frac{1}{4}\theta_1(1-\theta_2)^2 f_{1,1}$ (g_3)	
			1 q 2	$\frac{1}{4}\theta_1\theta_2 f_{1,1} f_{2,2}$ (g_4)	
			1* ? 2*	$\frac{1}{4}[\theta_1(1-\theta_2)+(1-\theta_1)\theta_2]\theta_1\theta_2$ (h_1)	
			1* ? 2	$\frac{1}{4}\theta_1(1-\theta_2) f_{2,2}$ (h_2)	
			1 ? 2*	$\frac{1}{4}(1-\theta_1)\theta_2 f_{1,1}$ (h_3)	
			1 ? 2	$\frac{1}{2}(1-\theta_1)(1-\theta_2) f_{1,1} f_{2,2}$ (h_4)	

From these gametic probabilities, the QTL genotype probabilies of the son are calculated as follows:
$P[Q?_p \mid g_L, g_R] = [a\Sigma c_i + b\Sigma f_i]/\Sigma_{Tot}$ $P[Q?_p \mid g_L g_R] = [a\Sigma c_i + b\Sigma f_i]/\Sigma_{Tot}$ $P[q?_p \mid g_L g_R] = [a\Sigma d_i + b\Sigma g_i]/\Sigma_{Tot}$

where

$$\Sigma_{Tot} = a\left(\Sigma c_i + \Sigma d_i + \Sigma e_i\right) + b\left(\Sigma f_i + \Sigma g_i + \Sigma h_i\right).$$

(v) *QTL mapping*

The QTL genotype probabilities at a given map position (p), obtained as described in (ii) and (iii) for sons and grandsons respectively, can be used in conjunction with phenotype ranks to measure the evidence in favour of a segregating QTL at the corresponding map position using a variant of the Wilcoxon rank-sum test (Kruglyak & Lander, 1995*b*; Coppieters *et al*, 1998*b*). In addition to exploring the potential information provided by the analysis of the grandsons (G²DD Option I), we explored three different approaches to combine the information from sons and grandsons (G²DD Options II–IV). The

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

power of these approaches was compared with the conventional GDD exploiting information from sons only (GDD).

(*a*) *GDD*

The GDD was performed as described in Coppieters *et al.* (1998*b*). Briefly, for each founder sire we calculated the value of:

$$Z_S(p) = Y_S(p)/\sqrt{\langle Y_S(p)^2\rangle},$$

where

$$Y_S(p) = \sum_{i=1}^{n_S} [n_S + 1 - 2{\cdot}rank(i)] \times [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)],$$

in which n_S is the number of sons; $rank(i)$ is the rank by phenotype of son i; $P[Q?_p \mid g_L, g_R]$ is the probability that progeny i has genotype $Q?$ at map position (p) given genotypes at the left (g_L) and right (g_R) flanking markers; $P[q?_p \mid g_L, g_R]$ is the probability that progeny i has genotype $q?$ at map position (p) given genotypes at the left (g_L) and right (g_R) flanking markers; and

$$\sqrt{\langle Y_S(p)^2\rangle}$$

is the standard deviation of $Y_S(p)$ expected under the null hypothesis of no QTL over all possible sets of genotypes. $\langle Y_S(p)^2\rangle$ can be shown (Kruglyak & Lander, 1995*b*) to equal

$$\langle Y_S(p)^2\rangle = \left(\frac{n^3 - n}{3}\right)\langle [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2\rangle.$$

$\langle [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2\rangle$,

the expected value of

$$[P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2$$

over all possible genotypes was determined for each sire by generating all possible sons (Y_S) and calculating the mean of

$$[P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2$$

weighted by the expected frequencies of the respective offspring. Squared values of $Z_S(p)$ were summed over the n_f sires, yielding a χ^2 statistic with n_f degrees of freedom.

(*b*) *G²DD – Option I*

In the first option of the G²DD we explored the amount of information that could be extracted from the grandsons by generating the statistic $Z_{GS}(p)$ for each sire, calculated as

$$Z_{GS}(p) = Y_{GS}(p)/\sqrt{\langle Y_{GS}(p)^2\rangle},$$

where

$$Y_{GS}(p) = \sum_{i=1}^{n_{GS}} [n_{GS} + 1 - 2{\cdot}rank(i)] [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)],$$

in which n_{GS} is the number of grandsons; $rank(i)$ is the rank by phenotype of grandson i; $P[Q?_p \mid g_L, g_R]$ is the probability that grandson i has genotype $Q?$ at map position (p) given genotypes at the left (g_L) and right (g_R) flanking markers; $P[q?_p \mid g_L, g_R]$ is the probability that grandson i has genotype $q?$ at map position (p) given genotypes at the left (g_L) and right (g_R) flanking markers; and

$$\sqrt{\langle Y_{GS}(p)^2\rangle}$$

is the standard deviation of $Y_{GS}(p)$, expected under the null hypothesis of no QTL over all possible sets of genotypes. As for the sons, $\langle Y_{GS}(p)^2\rangle$ can be shown (Kruglyak & Lander, 1995) to equal

$$\langle Y_{GS}(p)^2\rangle = \left(\frac{n^3 - n}{3}\right)\langle [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2\rangle.$$

$\langle [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2\rangle$,

or the expected value of

$$[P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2$$

over all possible genotypes was determined by simulating 1000 grandsons (Y_{GS}) and calculating the mean of

$$[P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2.$$

Under the null hypothesis of no QTL, Z_{GS} is a standard normal variable that reduces to a Wilcoxon rank-sum test at the marker positions. Squared values of $Z_{GS}(p)$ were summed over the n_f sires, yielding a χ^2 statistic with n_f df.

(*c*) *G²DD – Option II*

In the second option of the G²DD, the information extracted from sons and grandsons was combined by generating the $Z_S(p)$ and $Z_{GS}(p)$ statistics for each sire as described above, and combining them as follows:

$$\sum_{f=1}^{n_f} [Z^2_{S,f} + Z^2_{GS,f}] = \chi^2_{2n_f}, \quad (1)$$

which yields a χ^2 statistic with $2 \times n_f$ df.

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

(*d*) *G²DD – Option III*

Squaring $Z_{S,f}$ and $Z_{GS,f}$ prior to their summation, as in (1), has the advantage that it yields a statistic with a known χ^2 distribution. The disadvantage of this approach, however, is that it does not constrain the sign of the QTL effect to be the same in the sons and grandsons of a given sire. We have therefore explored an alternative approach in which the pairs of Z values obtained by this procedure for the n_f founder sires are combined as follows:

$$\sum_{f=1}^{n_f} [Z_{S,f} + Z_{GS,f}]^2.$$

The resulting statistic is not distributed as a χ^2 any longer. However, statistical significance of the data can be estimated by phenotype permutation according to Churchill & Doerge (1995; see hereafter). Note that this approach does not constrain the magnitude of the QTL effect to be identical in sons and grandsons.

(*e*) *G²DD – Option IV*

In the fourth approach, a single statistic was generated for each founder sire:

$$Z_{S+GS}(p) = Y_{S+GS}(p)/\sqrt{\langle Y_{S+GS}(p)^2\rangle},$$

where

$$Y_{G+GS} = \sum_{i=1}^{n_S+n_{GS}} [n_S + n_{GS} + 1 - 2{\cdot}rank(i)] \; [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)],$$

in which n_S is the number of sons, and n_{GS} the number of grandsons; $rank(i)$ is the rank by phenotype of son or grandson i, within the pool of sons and grandsons of founder sire f; $P[Q?)_p \mid g_L, g_R]$ is the probability that progeny i has genotype $Q?$ (or $q?$) at map position (p) given genotypes at the left (g_L) and right (g_R) flanking markers; and

$$\sqrt{\langle Y_{S+GS}(p)^2\rangle}$$

is the standard deviation of $Y_{S+GS}(p)^2$, expected under the null hypothesis of no QTL over all possible sets of genotypes. $Y_{S+GS}(p)^2$ was calculated as

$$((n_S + n_{GS})^3 - (n_S + n_{GS}))/3,$$

multiplied by the weighted average of

$$\langle [P_i(Q?_p \mid g_L, g_R) - P_i(q?_p \mid g_L, g_R)]^2\rangle$$

of the sons and grandsons.

The resulting Z_{S+GS} values were combined across the n_f founder sires as follows:

$$\sum_{f=1}^{n_f} Z^2_{S+GS,f} = \chi^2_{n_f},$$

yielding a χ^2 statistic with n_f df.

(vi) *Significance thresholds*

Chromosome-wide significance thresholds were determined from the distribution of the test statistic over 10000 permutations (simulated data set) or 1000000 permutations (real data set) of the ranks in a manner similar to that suggested by Churchill & Doerge (1995). Permutations of the ranks (whether of sons, grandsons or the pools of sons plus grandsons) were performed *within* sires. For each permutation, the highest value of the test statistic over the entire chromosome was retained in order to yield 'chromosome-wide' distributions of the test statistic under the null hypothesis. For the real data set, a Bonferroni correction was applied to the chromosome-wide significance level, considering that chromosome 6 represents 1/29th of the bovine autosomes and that we analysed the equivalent of three independent traits (Spelman *et al*., 1996), in order to obtain 'experiment-wide' significance thresholds.

2.3. Pedigree material

The pedigree material used in this study was a subset (Dutch population) of a previously described Holstein–Friesian grand-daughter design comprising 907 sons distributed over 22 paternal half-sib families (Spelman *et al*., 1996; Coppieters *et al*., 1998*a*). The number of sons per sire-family ranged from 11 to 148. Analysis of the pedigree relationships showed that 904 of the sons were also grandsons of one of the founder sires in our pedigree material. Six hundred and thirty-one were maternal grandsons from 14 founder sires, while 273 were paternal grandsons from three founder sires. Two hundred and nineteen sons had both their paternal and maternal grandsire in the available pedigree material. The number of grandsons per maternal grandsire ranged from 2 to 219, and from 22 to 225 per paternal grandsire.

2.4. Simulated data-set

To evaluate the relative efficacy of the four G²DD options with respect to each other and the conventional GDD, we simulated the segregation of a biallelic QTL (Q, q) in the previously described pedigree material. The QTL was assumed to be in Hardy–Weinberg equilibrium in the general population with allelic frequencies of 0·75 and 0·25 for Q and q respectively. Founder-sires therefore had an *a priori* probability $2pq = 0{\cdot}375$ to be heterozygous Qq for the QTL. Following Falconer's notation (Falconer & MacKay, 1996), and assuming additively acting alleles, the average phenotypic values of the QQ, Qq and qq genotypic classes were set at $+a, d = 0$ and

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

$-a$, respectively. The residual variation σ^2_R, was assumed to be non-genetic and normally distributed. Values for a and σ_R were chosen such that the average effect of the Q to q allele substitution, $\alpha = a$, equalled $0{\cdot}5\sigma_P$. Therefore, the variance attributable to the segregation of the QTL ($\sigma^2_{QTL} = 2pqa^2$) corresponded to 9·4% of the total phenotypic variance ($\sigma^2_P = \sigma^2_{QTL} + \sigma^2_R$). To test the effect of marker density on detection power, we positioned the QTL within both a sparse and a dense marker map. The sparse map comprised three markers which were 15 recombination units apart, with the QTL positioned in the middle of the second marker interval: M_1–(15%)–M_2–(8%)–QTL–(8%)–M_3. Two additional markers flanking the QTL were added in the dense map: M_1–(15%)–M_2–(5·6%)–M_3–(2·7%)–QTL–(2·7%)–M_4–(5·6%)–M_5. Both maps therefore totalled 35·3 cM (Haldane).

Markers were assumed to be polyallic markers with frequencies randomly assigned from a uniform distribution and re-scaled to sum to unity, yielding a heterozygosity of

$$h = 1 - \int_0^1 \dots \int_0^1 \int_0^1 \frac{\sum_{i=1}^{b} p_i^2}{\left(\sum_{i=1}^{b} p_i\right)2} \cdot dp_1 \cdot dp_2 \dots dp_b,$$

where p_i is the frequency of the ith allele randomly chosen from the uniform distribution for the locus in question. The number of marker alleles was set at four yielding an expected heterozygosity of 67%, which is very comparable to what is observed in reality with microsatellite markers in cattle populations.

Three hundred different data-sets were simulated with both the dense and sparse map. These simulated data-sets were analysed using the conventional GDD approach as well as using the four options of the G²DD. For each data-set, marker allele frequencies to be used in the QTL mapping procedure were estimated from the data as previously described (Georges *et al.*, 1995). For each of the three hundred replicates we performed 10000 phenotype permutations, which were each analysed using the five models. For each permutation, the highest values of the χ^2 statistics along the chromosome map were stored for each model. These values were combined in order to yield a data-set- and model-specific distribution of the chromosome-wide test statistic under the null hypothesis of no QTL. The corresponding distributions were then utilized to measure the p value of the unpermutated data. The distribution of p values was compared across models though within maps using the Wilcoxon matched pair test (Hollander & Wolfe, 1973). The distribution of p values obtained with the sparse versus dense maps was compared within models using the Mann–Whitney U-test (Hollander & Wolfe 1973).

2.5. Real data-set

Performance of the G²DD was evaluated on a real data-set, i.e. the previously described pedigree material genotyped for nine microsatellite markers spanning 46 cM of bovine chromosome 6: URB016, BM1329, BM143, TGLA37, ILSTS097, BM4528, BM4621, RM028 and BM415. The corresponding microsatellites, marker order and recombination rates between adjacent markers as deduced from the corresponding data have been described previously (Coppieters *et al.*, 1998*b*).

The records that were used for linkage analysis were the sons' daughter yield deviations (DYD) for protein per cent, corrected for half the DYD of their sire (Van Raden & Wiggins, 1991), DYDs were directly obtained from Holland Genetics (Arnhem, The Netherlands), and Livestock Improvement Corporation (Hamilton, New Zealand).

3. Results

(i) *Simulated data*

Table 3 summarizes the results obtained with the simulated data. It can be seen that information from grandsons can indeed by extracted using the proposed approach as the simulated QTL could be detected 20–25% of the time using information from the grandsons only (G²DD-I). This value has, however, to be compared with a power of 46–58% when using the conventional GDD.

Moreover, analysis of Table 3 shows that information from sons and grandsons can be advantageously combined as average p value decreases and conversely power increases when comparing the GDD with the G²DD versions II–IV. This is particularly true for the G²DD–IV where sons and grandsons are ranked as a single group. Using the Wilcoxon matched pair test to compare the distribution of p values obtained with the G²DD–IV versus the other models shows that it is very significantly superior to all other approaches ($P < 0{\cdot}0001$), using both the sparse and the dense maps. There is some indication of the superiority of the G²DD–II and G²DD–III approaches above the conventional GDD as average P values decrease and power increases, respectively. For the trio GDD, G²DD–II and G²DD–III, however, distributions of P values as compared with the Wilcoxon matched pair test were essentially not significantly different ($0{\cdot}10 < P < 0{\cdot}60$).

The effect of marker density on power was assessed by comparing the distribution of P values obtained with the sparse versus the dense map for each of the five analysis models. Power increased by 26%, 25%, 25%, 32% and 27% for the G²DD–I, GDD, G²DD–I, G²DD–III and G²DD–IV respectively. Comparing the corresponding distributions of P values using the

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

Table 3. *Summary of simulations: results on power and mapping precision*

	Map	G²DD–I	GDD	G²DD–II	G²DD–III	G²DD–IV
Average *P* value	Sparse	0·333 (<)	0·164 (=)	0·146 (=)	0·144 (<)	0·124
	Dense	0·307 (<)	0·117 (=)	0·103 (=)	0·101 (<)	0·072
Power at $\alpha = 0{\cdot}05$	Sparse	0·20	0·46	0·48	0·49	0·55
	Dense	0·25	0·58	0·60	0·65	0·70
Mapping precision	Sparse	14·0	10·9	12·6	12·2	11·4
	Dense	12·3	9·8	10·7	9·3	8·4

Comparison of the average *P* value, detection power (% of simulations yielding a statistic significant at the 5% level) and mapping precision (standard deviation of the difference between real and estimated QTL position) obtained on 300 data-sets simulated as described and analysed according to five distinct methods labelled GDD and G²DD–I to –IV and genetic maps of two marker densities: sparse and dense. The (<) and (=) signs indicate that the corresponding model is either significantly inferior to (<) or not significantly different from (=) the models in the next columns.

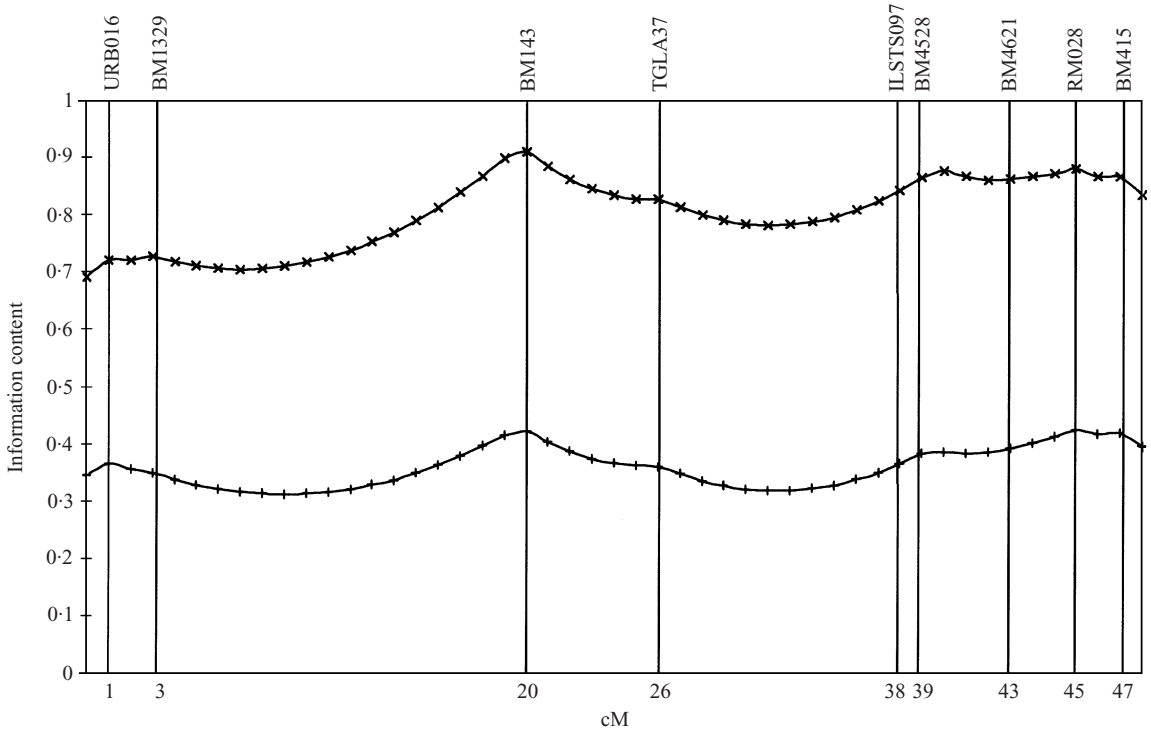

Fig. 2. Information content obtained for sons (upper line, crosses) and grandsons (lower line, plus signs) along the chromosome 6 microsatellite map.

Mann–Whitney *U*-test showed the high significance of this effect for all models ($P < 0{\cdot}000\,01$) except for the G²DD–I ($P = 0{\cdot}13$). We would have assumed *a priori* that an increase in marker density would have benefited particularly the four G²DD options as the chromosomes are to be traced via the ungenotyped dam generation in these schemes. This was, however, not clearly apparent from these results.

Estimates of the precision in the estimation of QTL positions were also compared. Table 3 shows the standard deviation of the difference between real and estimated position of the QTL for all simulations yielding a signal exceeding the 5% chromosome-wide significant threshold. Comparing the difference between real and estimated position using the Mann–Whitney *U*-test, we found that the G²DD–I was significantly less precise than all other schemes using both the sparse and the dense map. None of the other observed differences proved significant at the 5% level. Nevertheless, we note the following tendencies: (i) as expected, the mapping precision improves overall when increasing the marker density, (ii) GDD²–II to –IV seem to be slightly less precise than the GDD when using the sparse map, and (iii) GDD²–II and –III seems to be equally as precise as the GDD when using the dense map, while GDD²–IV might even be superior. As previously noted, for most QTL mapping experiments (e.g. Coppieters *et al.*, 1998*b*), the

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

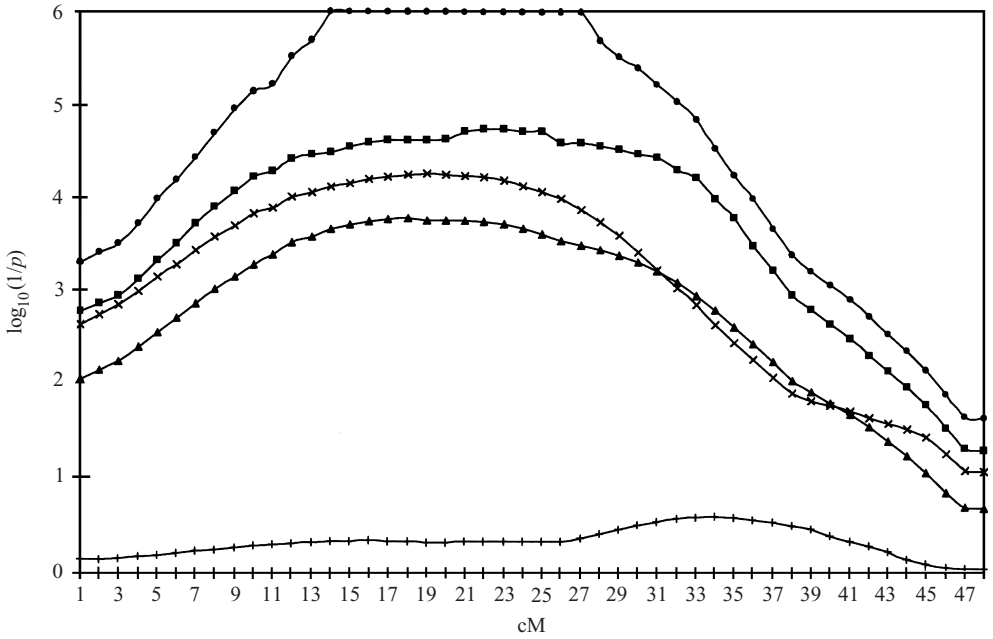

Fig. 3. Location scores, expressed as $\log(1/p)$, obtained along the chromosome 6 microsatellite map using the GDD (crosses), G²DD–I (plus signs), G²DD–II (triangles), G²DD–III (squares) and G²DD–IV (circles). Using the G²DD–IV option the location scores 'saturate' between positions 13 and 29 cM, reflecting the fact that none of the million permutations yielded chi-squared values as high as those obtained with the real data.

mapping precision is essentially poor under all envisaged scenarios.

(ii) *Real data set*

Fig. 2 illustrates the information content of the marker data when considering the sons and grandsons respectively. The information content is calculated (see Section 2) such that it cannot exceed 50% for the grandsons, as half of these would inherit neither of the grand-paternal chromosomes. It can, therefore, be seen that more than 70% of the theoretical maximum information can be extracted from the grandsons using this map. This is not very different from the information content obtained with the sons, showing that the multipoint approach used to trace the segregation through the ungenotyped dam generation is quite effective.

Fig. 3 shows the location scores ($\log_{10}(1/p)$) obtained along the marker map using the five approaches. It can be seen that when analysing the grandsons only, the analysed data-set provides very limited evidence for the segregation of a QTL. Analysing the sons only using the GDD provides very significant evidence for the presence of a QTL on this chromosome as previously reported (Spelman *et al.*, 1996; Coppieters *et al.*, 1998*b*). Combining the information from sons and grandsons seem to have a modest but deleterious effect when using the G²DD–II approach, though a favourable effect for the G²DD–III and G²DD–IV approaches. Particularly when using the G²DD–IV model the increase in significance is quite substantial, going from $P < 0{\cdot}005$ (GDD) to $P < 0{\cdot}0001$.

4. Discussion

We herein develop an approach to extracting linkage information from a frequently occurring relationship that is usually ignored when using the grand-daughter design (Weller *et al.*, 1990): the fact that many sons share a common grandsire. We demonstrate how a previously described sum-of-rank based method (Coppieters *et al.*, 1998*b*) can be extended to extract linkage information from the relationship between a sire and its grandsons either independently (GDD²–I) or combined with the conventional information from the relationship between the sire and its sons (GDD²–II to –IV). Besides the fact that it is easy to implement, this approach extends the scope of QTL mapping to a variety of traits not normally distributed, such as counts generated by a Poisson distribution, truncated data, probabilities and qualitative data. It is also perfectly applicable to normally distributed traits with minimal loss of power. The proposed approach is sufficiently fast to allow for the determination of chromosome-wide significance thresholds using phenotype permutation (Churchill & Doerge, 1995).

Disadvantages of the method are the fact that the rank-based approach does not yield an estimate of the QTL effect, and that it obviously does not exploit all available linkage information as would 'full pedigree

https://doi.org/10.1017/S0016672399004012 Published online by Cambridge University Press

analysis' (e.g. Hoeschele *et al*., 1997; Bink & van Arendonk, 1998). We believe, however, that it targets the most informative relationships in the context of the GDD and therefore represents a useful compromise between power increase and ease of implementation.

The method accounts for the missing genotypes of the dams, i.e. chromosome segregation is traced from grandsires to grandsons via their ungenotyped dams. By doing so, the method requires estimation of the probabilities that the dam has inherited either one or the other paternal QTL allele. This could be used to extract additional information from the dams' phenotypes. It would be fairly easy to implement this added feature. However, so far we have elected not to as the phenotypes of the dams are characterized by a much lower reliability than that of progeny-tested sires, and because bull-dams are known to represent a very biased sample.

The G²DD–I, in which information is extracted from grandsires only, has successfully been used as a validation tool to confirm the genuine nature of QTL detected using the GDD (Coppieters *et al*., 1998*a*; Arranz *et al*., 1998; Spelman & Bovenhuis, 1998). QTL are first mapped using the GDD in an across-family approach. Likely heterozygous '*Qq*' sires can then be identified on the basis of within-family statistics. The confirmation of a significant phenotypic contrast between the sire's homologues in their grandsons validates the putative QTL. Note that the number of grandsons needs to be sufficiently large to reach an acceptable validation power. However, as one can focus on a single trait and a limited chromosome segment in such validations, the penalty paid for multiple testing is much lower than in the initial whole-genome scan.

Alternatively, information from sons and grandsons can readily be combined (G²DD–II to –IV) to extract more information from the available data. QTL that did not reveal a significant signal when performing a conventional GDD analysis could be detected using the G²DD–IV model (W. Coppieters, unpublished). As expected, the G²DD–IV option, in which sons and grandsons of a given sire are treated and ranked as a single pool of observations, proved to be the most powerful approach when applied to the simulated data. To be as effective with real data, however, the expected average rank of sons and grandsons should be identical, allowing them to be pooled. A number of reasons could be invoked that could lead to a violation of this assumption when dealing with real data. Selection might cause an upward shift of the phenotypes when comparing grandsons and sons. Also, while the phenotypes used in this analysis were DYDs corrected for half of the paternal DYDs to account for sire effects, the quarter of the grandpaternal effect not being accounted for in the grandsons could lead to a shift between sons and grandsons. Evidence for a son versus grandson effect on the utilized phenotype was tested for in our material and shown not to depart from random expectation (data not shown). Not surprisingly, therefore, the G²DD–IV proved also to be the most effective approach when applying it on the real data. If there were evidence for significant shifts between sons and grandsons, however, the G²DD–III approach would be preferred despite the *a priori* potential loss of power.

This work was funded by grants from Holland Genetics, Livestock Improvement Corporation, the Vlaamse Rundvee Vereniging and the Ministère des Classes Moyennes et de l'Agriculture, Belgium. Continuous support from Nanke den Daas, Jeremy Hill, Brian Wickham, Denis Volckaert and Pascal Leroy is greatly appreciated. We thank Johan van Arendonk, Richard Spelman, Henk Bovenhuis, Marco Bink and Dorian Garrick for fruitful discussions.

References

Arranz, J.-J., Coppieters, W., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Riquet, J., Simon, P., Vanmanshoven, P., Wagenaar, D. & Georges, M. (1998). A QTL affecting milk yield and composition maps to bovine chromosome 20: a confirmation. *Animal Genetics* **29**, 107–115.

Bink, M. C. A. M. & van Arendonk, J. A. M. (1998). Detection of quantitative trait loci in outbred populations with incomplete marker data. (Submitted)

Churchill, G. A. & Doerge, R. W. (1995). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Coppieters, W., Riquet, J., Arranz, J.-J., Berzi, P., Cambisano, N., Grisart, B., Karim, L., Marcq, F., Simon, P., Vanmanshoven, P., Wagenaar, D. & Georges, M. (1998*a*). A QTL with major effect on milk yield and composition maps to bovine chromosome 14. *Mammalian Genome* **9**, 540–544.

Coppieters, W., Kvasz, A., Arranz, J.-J., Grisart, B., Mackinnon, M. & Georges, M. (1998*b*). A rank-based non-parametric method to map QTL in outbred half-sib pedigrees: application to milk production in a grand-daughter design. *Genetics* **149**, 1547–1555.

Falconer, D. S. & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics*, 4th edn. New York: Longman Scientific and Technical.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., Sargeant, L., Sorensen, A., Steele, M. R., Zhao, X., Womack, J. E. & Hoeschel, I. (1995). Mapping quantitative trait loci controlling milk production by exploiting progeny testing. *Genetics* **139**, 907–920.

Gomez-Raya, L., Våge, D. I., Olsaker, I., Klungland, H, Klemetsdal, G., Heringstad, B., Lie, O., Rønningen, K. & Lien, S. (1996). Mapping QTL affecting traits of economical importance in Norwegian cattle. In *Proceedings of 47th Annual Meeting of the European Association for Animal Production*, Lillehammer, 25–29 August 1996.

Hoeschele, I., Uimari, P., Grignola, F. E., Zhang, Q. & Gage, K. M. (1997). Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics* **147**, 1445–1457.

Hollander, M. & Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: Wiley.

Knott, S. A., Elsen, J. M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71–80.

Kruglyak, L. & Lander, E. S. (1995*a*). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Kruglyak, L. & Lander, E. S. (1995*b*). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

Kühn, C., Weikard, R., Goldammer, T., Grupe, S., Olsaker, I. & Schwerin, M. (1996). Isolation and application of chromosome 6 specific microsatellite markers for detection of QTL for milk-production traits in cattle. *Journal of Animal Breeding and Genetics* **113**, 355–362.

Ron, M., Heyen, D. W., Weller, J. I., Band, M., Feldmesser, E., Pasternak, H., Da, Y., Wiggans, G. R., Vanraden, P. M., Ezra, E. & Lewin, H. A. (1998). Detection and analysis of a locus affecting milk concentration in the US and Israeli dairy cattle populations. In *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production*, 11–16 January 1998, Armidale, NSW, Australia.

Spelman, R. J. & Bovenhuis, H. (1998). Moving from QTL experimental results to the utilization of QTL in breeding programmes. *Animal Genetics* **29**, 77–84.

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

Van Raden, P. M. & Wiggans, G. R. (1991). Derivation, calculation, and use of National Animal Model Information. *Journal of Dairy Science* **74**, 2737–2746.

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