Comparison of granddaughter design and general pedigree design analysis of QTL in dairy cattle: a simulation study

G. Freyer[1], N. Vukasinovic[2]

[1]Research Institute for the Biology of Farm Animals (FBN) Dummerstorf, Dummerstorf, Germany

[2]Monsanto Company Animal AG, St. Louis, Missouri, USA

ABSTRACT: Traditional methods for detection and mapping of quantitative trait loci (QTL) in dairy cattle populations are usually based on daughter design (DD) or granddaughter design (GDD). Although these designs are well established and usually successful in detecting QTL, they consider sire families independently of each other, thereby ignoring relationships among other animals in the population and consequently, reducing the power of QTL detection. In this study we compared a traditional GDD with a general pedigree design (GPD) and assessed the precision and power of both methods for detecting and locating QTL in a simulated complex pedigree. QTL analyses were performed under the variance component model containing a random QTL and a random polygenic effect. The covariance matrix of the polygenic effect was a standard additive relationship matrix. The (co)variance matrix of the random QTL effect contained probabilities that QTL alleles shared by two individuals were identical by descent (IBD). In the GDD analysis, IBD probabilities were calculated using sires' and daughters' marker genotypes. In the GPD analysis, IBD probabilities were obtained using a deterministic approach. The estimation of QTL position and variance components was conducted using REML algorithm. Although both methods were able to locate the region of the QTL properly, the GPD method showed better precision of QTL position estimates in most cases and significantly higher power than the GDD method.

Keywords: QTL mapping; granddaughter design; general pedigree design; computer simulation

Research on QTL mapping in dairy cattle breeding during the last decade has yielded a steadily growing success as Khatkar et al. (2004) showed in their review and meta analysis. This is true of adapting and developing new approaches and describing statistical-genetic tools for data analyses. However, the way from coarse QTL mapping to finally finding the causative gene is still a challenge. The more information is available in the pedigree, the better the results of QTL estimation, mainly in terms of precision of the QTL position. The ability to cope with different characteristics is of vital importance for adapting the right tool. Known simulation studies based on an ideal and well-defined family design with evenly spaced markers (e.g. Sørensen et al., 2003) are efficient for testing the methods in most cases, but they do not usually fit practical conditions. Data structure is important for a successful fine mapping study. As Lee and van der Werf (2004) reported on the basis of a simulation study, a design including many families of small size yielded a higher mapping resolution than a design with few families of large size. They showed that the difference was small in half sib designs and concluded that half sib designs might have sufficient information for fine mapping of QTL. Using fine mapping strategies based on combining linkage (LA) and linkage disequilibrium (LD) analysis that promise more powerful and precise QTL estimation within an interval of less than 5 cM, is becoming a focus of current research (e.g. Meuwissen et al., 2002). Such an approach uses information on the popula-

tion history along with pedigree and marker information. However, there is still a need to compare different methods with respect to specific practical conditions, e.g. how they deal with various complexity of a pedigree, inbreeding, missing marker and/or pedigree information, etc. So far we have not found any reports on approaches to QTL mapping that manage inbreeding.

In this study, we compare a traditional granddaughter design (GDD) with a general pedigree design (GPD) and assess the precision and power of both methods for detecting and locating QTL in simulated complex pedigrees. We focused on datasets reflecting practical conditions, which is usually neglected in simulation studies. We assumed a real dataset with various family structures, including inbreeding, size of datasets, overlapping generations, marker maps and missing marker information. Contemplating these aspects is a prerequisite for successful use of statistical-genetic approaches to QTL estimation under practical conditions.

MATERIAL AND METHODS

Data. We simulated several datasets that were different regarding size, family structure and length of the examined chromosomal segment (54 and 120 cM). We applied six different marker maps containing 11 unevenly distributed, moderately informative markers (Table 1). We also analyzed full marker information and different amounts of missing marker information (up to 30% random missing marker information). In GPD analysis, marker information was missing consequently for about 50% of dams.

The simulations were conducted using the approach described by Schelling et al. (1998). Basically, the pedigrees included ordered and overlapping generations, as we are usually confronted with in real dairy cattle breeding. Pedigree information was available for final offspring, sires, and two generations of their male ancestors, but was incomplete for dams. Phenotypic records were available for final offspring, sires and some of the dams. Maternal half sibs occurred in addition to design-dominating paternal half sibs. The families originated from two great-grand sires (GGS) that were assumed unrelated (Figure 1). We investigated different sizes of datasets with

– 200 individuals, 6 sires and 105 final offspring in ordered generations (family structure 0)
– 448 individuals, 6 sires and 240 final offspring in ordered generations, with inbreeding (half sib mating, family structure 0)
– 519 individuals, 9 sires and 283 final offspring in overlapping generations
– 850 individuals, 9 sires and 552 final offspring in overlapping generations, with and without inbreeding
– 1 155 individuals, 9 sires and 608 final offspring in overlapping generations
– 1 352 individuals, 9 sires and 1 046 final offspring in overlapping generations with and without inbreeding.

Only family structure 0 had a simple (ordered) generation structure. All other structures were more complicated with overlapping across generations.

For the purpose of this paper, we present the results of simulations based on a rather large QTL effect, contributing 60% to the variance, whereas

Table 1. Marker maps used in analyses: marker maps M1 to M4 cover 54 cM, marker maps M5 and M6 cover 120 cM. All maps include 11 markers (marker 1, 4 and 9 with four alleles, others with three alleles). Maps differ regarding the marker position and marker bracket containing the QTL. One single QTL is placed in the middle of the markers in bold script

Marker map	Map length (cM)	Marker position (cM) marker 1, marker 2, .., marker 11
M1	55	0, 13.7, 32.8, 35.7, **40.5, 42.5**, 43.5, 44.5, 45.5, 48.5, 53.2
M2	55	0, 10.9, 17.8, 25.9, 33.6, **39.8, 46.3**, 48.3, 49.3, 51.2, 54.4
M3	55	0, 13.7, 32.8, 35.5, 37.5, 39.3, **40.3 42.4**, 43.3, 44.3, 48.4
M4	55	0, 13.7, 32.8, 35.7, 37.7, **39.7, 43.5,** 44.5, 45.5, 46.5, 49.4
M5	120	0, 11, 22, 33, 44, **55, 67**, 78, 89, 110, 121
M6	120	0, 18, 36, 44, 52, **55, 57**, 63, 78, 93, 111

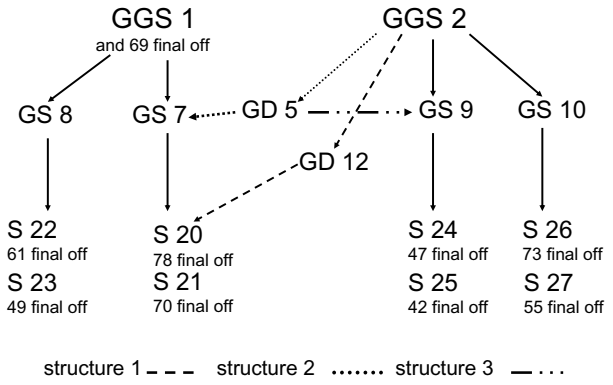

Figure 1. Basic family structure 0 (denoted by solid arrows) and resulting family structures 1, 2 and 3 for the datasets including a given number of final offspring (off) based on a sample size of 850 individuals. Shown are two great-grand-sire families (GGS 1 and GGS 2), descending grand sires and grand dam (GS7 to GS10 and GD), and sires (S). Sire 1 is great-grandsire with own final offspring. Structure 1, Structure 2 and Structure 3 differed by one path only (pattern of the arrow)

the proportions of polygenic and residual variance were 30 and 10%, respectively. This enabled us to clearly show different effects of applying GPD and GDD to various family settings on the particular results.

Analysis methods. The data was analyzed using a random model variance component approach. The QTL effect associated with each individual phenotype at a given genome position was assumed random, based on the variance-covariance structure which was the function of the proportion of identical-by-descent (IBD) alleles that two individuals shared at the QTL location. Therefore, the phenotypes of two individuals would be more similar if they shared a greater proportion of alleles inherited from the common ancestor at a locus affecting the trait.

Estimating the proportion of IBD alleles at the QTL position is one of the major issues in the application of the random model approach. The size of animal pedigrees (several thousand individuals) and complex relationships (sires mated to a large number of dams, dams mated to different sires) pose limitations on usability of standard software packages for calculation of IBD probabilities. To overcome this obstacle, Pong-Wong et al. (2001) developed a simple, rapid deterministic method to calculate IBD probabilities in large animal pedigrees. In order to avoid complicated calculations and integration over all possible phases in the entire pedigree when the haplotype phases are not known, the haplotype probabilities were calculated using the closest informative (phase known) marker bracket. Markers that were not informative (missing, homozygous, impossible to determine phases) were skipped.

In order to take full advantage of the recursive method, at least three generations in the pedigree must be genotyped because the "phase-known marker" means that it must be possible to determine the grandparental origin of marker alleles in the individual. To avoid the loss of information for individuals that were e.g. offspring of base animals, a method used to determine IBD probabilities in sibs (Knott and Haley, 1998) was used to determine IBD probabilities among sibs' gametes. An enhanced version of the algorithm (Vukasinovic and Martinez, 2002) used the same methodology. In addition, the (half) sib method was applied to the offspring of base individuals and to all full and half sib families for which the average number of informative markers (across both maternal and paternal gametes) in the family usable by the recursive method was smaller than the average number of informative markers usable by the (half) sib method. The loss of information, for example if a large proportion of marker genotypes was missing in grandparents, was prevented in this way.

Initially, the IBD probabilities were calculated for each pair of gametes independently to obtain matrix **G** of gametic IBD probabilities. Then, matrix **Q** ("IBD relationship matrix") containing IBD probabilities at the individual's level was obtained by joining together IBD probabilities of the paternal and maternal gametes for each individual by a linear transformation of the **G** matrix.

The following mixed animal model was used in our variance component analysis:

$$y = X\beta + Za + Zq + e$$

where: y = the $(n \times 1)$ vector of phenotypes
X = the $(n \times s)$ design matrix
Z = the $(n \times n_0)$ incidence matrix relating animals to phenotypes
β = the $(s \times 1)$ vector of fixed effects
a = the $(n_0 \times 1)$ vector of random polygenic effects
q = the $(n_0 \times 1)$ vector of random QTL effects
e = the $(n \times 1)$ residual vector

In our analyses, n refers to the number of animals with phenotypes and n_0 to the total number

of animals in the pedigree. X was equal to 1 because all phenotypes were assumed pre-adjusted for non-genetic effects, and thus $s = 1$ and $\beta = \mu$. The random genetic effects a and q were assumed to follow a normal distribution with mean zero and variances $A\sigma_a^2$ and $Q\sigma_q^2$, respectively. Matrix **A** was the standard additive relationship matrix. Matrix **Q** was the QTL relationship matrix, obtained as described above. The model was fitted for each position on the chromosome. The REML procedure implemented in the ASREML software (Gilmour et al., 2002) was used to maximize the likelihood for each testing position.

The GDD analysis was based on the same principles regarding the statistical model and variance component estimation. In the GDD analysis, however, IBD probabilities were calculated within each grandsire family independently, thereby ignoring "higher level" relationships among animals. Details regarding GDD analysis are described in Freyer et al. (2003).

Evaluation of results. The likelihood ratio (LR) test statistic was calculated as twice the difference between log likelihoods of the full model and the model without the QTL effect. The position with the highest LR statistic was declared the most likely position of the QTL resulting from the particular analysis. The estimates for QTL, polygenic, and total variance, and corresponding heritability at the most likely position were considered MLE for these parameters.

To obtain 95% confidence intervals (CI(95%)) for QTL position estimates, we used the bootstrapping method (via QTL express for GDD, from http://qtl.cap.ed.ac.uk/). We further applied the method by Darvasi and Soller (1997) for calculating the CI(95%) in cases of saturated marker maps, where it might be justified.

RESULTS AND DISCUSSION

Our focus was on comparing results from GDD and GPD under practical conditions with varying generation and family structures, comparing different sizes of datasets, and varying marker density and maps assuming a constrained budget for genotyping 11 markers, respectively. We abstain from presenting results of different sizes of simulated QTL effects because they persistently yielded the same shape of the test statistic profile, at least from the GDD analysis, with exactly the same peaks indicating the same estimated QTL position, when comparing 10, 25, 40 or 60% variance contributed by the QTL. Obviously, the levels of the test statistics which reflect the power of the design increased with increasing QTL variance, which was consistent with previous simulation studies (e.g. Mayer et al., 2004). The variance components estimated by GPD were much smaller than those obtained by GDD, but this did not cause a deterioration of the QTL position estimate (Table 2).

In most of the relevant cases, both the GDD analysis and the GPD analysis were able to locate the right QTL region within the confidence interval 95% and at a chromosome wise significance level of $P < 0.01$ (Table 2, datasets A, B, C, D, E, F, J). The main prerequisite for this was the "right pattern" of the marker map. Maps with more dense markers around the true QTL position (M1, M4, M6) were superior, regarding estimates of QTL position, to maps with evenly spaced markers. We analyzed datasets based on various marker maps, being slightly different at some marker positions, mainly mimicking a fine mapping like structure in order to investigate the impact of different marker positions and distances on the estimated QTL position within a 54 cM chromosomal segment (M1 to M4). This mimicked the situation that a QTL was suggested by a previous analysis using a coarse map and then trying to get closer to the position with varying marker constellations. Out of those four maps, only marker map M2 was not a sufficient precondition for the successful QTL estimation, even in very large data sets (dataset K, Table 2). Both GDD and GPD analyses yielded the highest MLE at 34 and 35 cM, respectively, whereas the true QTL position was 41 cM. The bootstrap CI(95%) was 54 cM, i.e. it covered the entire chromosomal segment. We analyzed dataset K additionally by QTL Express, obtaining the only insignificant result for the estimated QTL position within this group of marker maps. The density of markers around the actual QTL position (between markers 6 and 7) in M2 was not as high as for the other marker maps (i.e. M1 to M3 in Table 1). Conclusions from such comparisons are of vital importance for preparing fine mapping analyses. Marker maps M5 and M6 were related to a 120 cM chromosome, containing 11 markers as before. Results from analyses based on M5 and M6 showed that with M6 we were able to locate the correct QTL region. Using map M5 in a GDD analysis, we found two peaks with a gap around the true QTL position, and the CI(95%)

Table 2. Characteristics of the datasets, estimated map positions of QTL, confidence intervals CI(95%) based on Darvasi and Soller (1997) and relative QTL variance estimates from analyses of GDD and GPD

Dataset	Data characteristics			Analysis results				
	sample size	marker map	family structure, missing marker inf.	estimated QTL position (cM)		confidence interval CI(95%) (cM)		relative QTL variance component estimated by GDD and GPD[†]
				GDD	GPD	GDD	GPD	
A	200	M1	0, no	43	43	12	7	0.63 and 0.42
B	519	M1	1, yes	44	41*	4	3	0.83 and 0.50
C	1 155	M1	1, no	43	38	2	3	0.95 and 0.11
D	850	M1	1, no	43	42	2	6	0.64 and 0.14
E	519	M3	1, no	42*	42*	4	5	0.38 and 0.28
F	850	M4	1, no	42*	42*	3	5	0.48 and 0.18
H	448	M1	0, no	43	38	5	8	0.71 and 0.23
J	850	M1	3, yes	36	42	[36–44][a]		0.62 and 0.15
K	1 352	M2	1, yes	34[b]	35[b]	[0–54][a]		0.42 and 0.11
I	850	M4	1, yes	43	23	[38–45][a]		0.48 and 0.21
G	1 352	M6	2, yes	57*	51	[49–53][a,c]		0.64 and 0.55
L	1 352	M6	3, yes	56*	45	[48–52][a,c]		0.92 and 0.13

[†]true QTL variance was 60% of the total phenotypic variance; *indicates that the correct map position was found within a region of ± 1 cM

[a][ ] indicates the bootstrap CI(95%) because the formula by Darvasi and Soller (1997) is not suited here for at least one of both analyses

[b]results were not significant due to the *F*-test after permutation (not shown in detail), whereas all others were significant at $P < 0.01$

[c]the bootstrap CI(95%) did not contain the actual simulated QTL

did not cover one of them (results not shown in detail). An evenly spaced coarse map such as M5 is not apparently useful for fine mapping analyses. M6 was very useful in GDD with missing marker information. The simulated QTL region was found exactly at 56 cM with an extremely steep peak of the test statistic profile for family structure 3, and at 57 cM for family structure 2, respectively (Table 2). Compared to the short maps, M6 covers more than the double chromosomal length with the same number of markers. This result indicates that GDD-based linkage analysis can precisely locate QTL even with relatively sparse maps, based on large and informative offspring groups.

Comparing the sizes of the datasets, we obtained a more precise position estimate with large datasets from GDD analyses (dataset G, L), and with medium sized datasets from GPD analyses (datasets B and F in Table 2). Regarding the power of QTL detection, the GPD analysis was always superior to the GDD analysis, as it can be seen from the test statistic levels (Figure 2). This was not surprising because of the tremendous addition of pedigree information used in GPD in contrast to GDD. The advantage of GPD was especially visible in small datasets (Figure 2, datasets A and B). This is shown by a sharper peak of the test statistic profile, a more precise QTL position estimate, and a shorter confidence interval in several cases (datasets B, D, J, Table 2). Additionally, GPD shows sufficient robustness against missing marker information in a small dataset, as obtained from dataset B (Figure 2). Analyzing the same dataset by GDD led to a distance of 3 cM between simulated and estimated QTL position. In cases of higher numbers of individuals involved, the CI(95%) obtained from the GDD was shorter than those from GPD (Table 2, datasets C and D).

The size of the CI(95%) serves as an important criterion in QTL estimation, in addition to the cor-

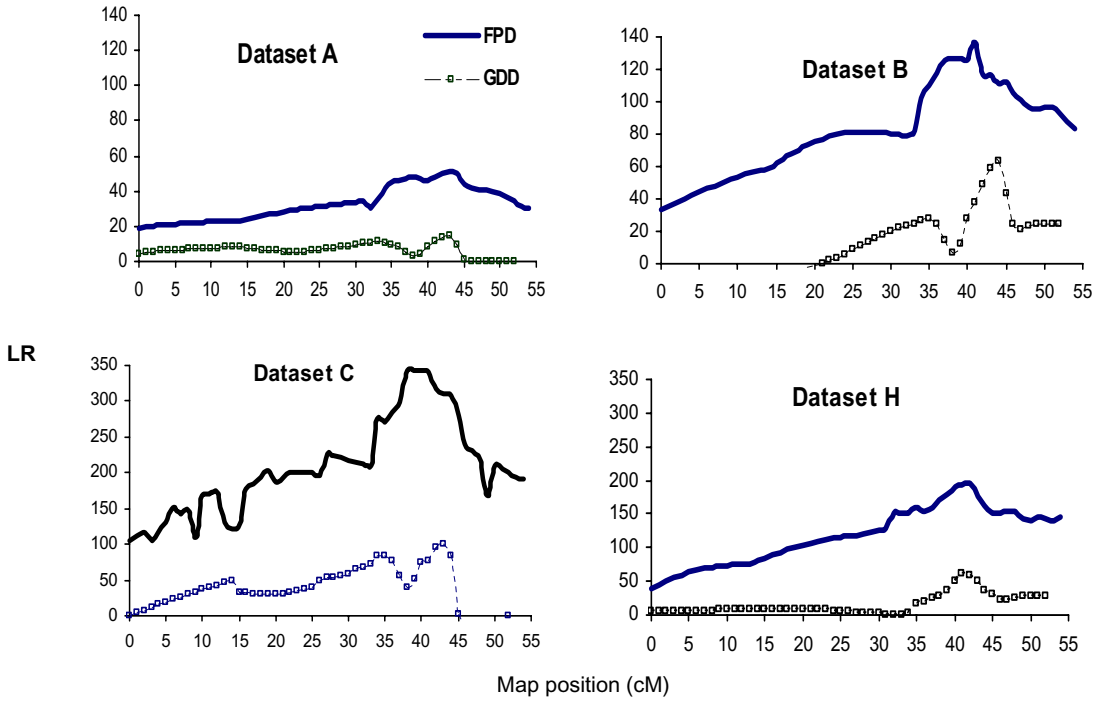

Figure 2. Comparison of test statistic profiles between analyses of granddaughter design (GDD) and general (or full) pedigree design (GPD or FPD) within four different datasets differing in pedigree size, family structure and marker information. The analyses were based on marker map M1 harbouring one QTL at 41 cM

rectly estimated QTL position. Because we used both saturated and relatively sparse marker maps, we could not rely in general on calculated CI(95%) as suggested by Davarsi and Soller (1997). When the QTL region was estimated correctly using a saturated marker map, the CI(95%) was reliable and in agreement with the confidence intervals obtained by bootstrapping in GDD analysis (for datasets A, B, C, D, E, F, J in Table 2).

Overlapping generations (Figure 1) did not cause any problems compared to ordered generations. But in family structure 2, when grand dam GD 12, a daughter of sire 2, was used to produce one half of the sires in family 1, compared to one quarter in structure 2 (Figure 1), we apparently got a bias in QTL position estimates obtained from the GPD analysis (e.g. dataset G, Table 2). The GDD analysis of family structure 2 led to an almost correct estimate of the QTL position, 1 cM apart from the simulated QTL position from dataset G (Table 2). The same result was obtained from structure 3, when the pedigree was based on close inbreeding (dataset L in Table 2). Actually, inbreeding did not affect the GDD analysis, but it did affect the GPD analysis. The correctness of results (from datasets with inbreeding) by GDD and GPD analyses varied a lot when based on smaller datasets and short marker maps (e.g. datasets F and J in Table 2).

It might be assumed that a nested pedigree structure, as shown by family structures 2 and 3, where we have in principle one big family instead of two independent great grandsire families, should increase the information content leading to more precise estimates of the QTL position. But this was not the case when inbreeding was included (structure 3). GPD analysis of large datasets with nested family structures seems to be inferior to GDD analysis with respect to the correctness of estimated QTL positions (datasets G and L, see Figure 3). The reason could be difficulties of the algorithm when handling large data sets and complicated family structures, which makes the matrix of IBD probabilities numerically unstable and causes problems in convergence. Another critical point could be the estimated QTL variance, being too small in all considered GPD analyses (Table 2). Family structures, as designed in Figure 1, are quite common within practical dairy cattle breeding and have to be investigated further in order to make the algorithm more effective. At the present time, the combined analysis based on LA and LD seems to be the method of choice (e.g. Meuwissen et al., 2002). Lee and van der Werf (2004) found that the

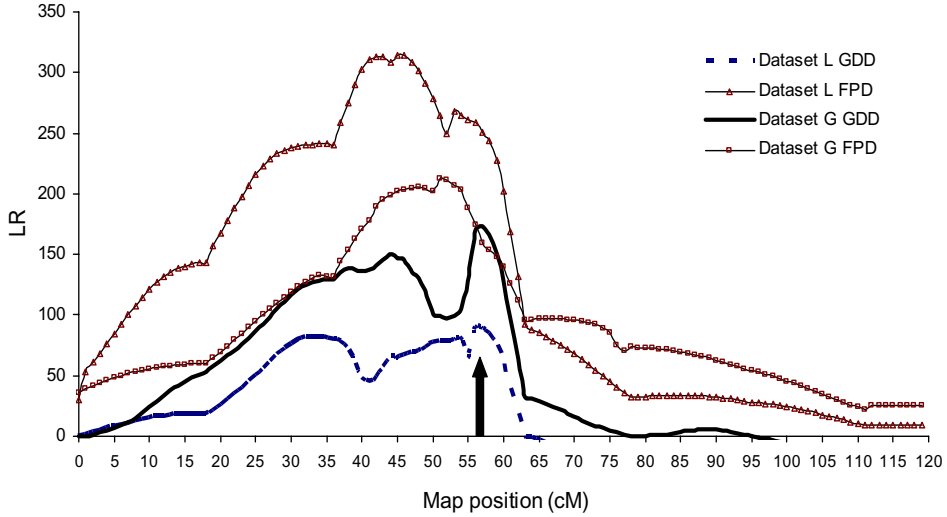

Figure 3. Test statistic profiles resulting from analyses of granddaughter design (GDD) and general (or full) pedigree analysis (GPD or FPD) based on family structure 2 (dataset G) and family structure 3 (dataset L) and marker map M6 covering a chromosomal length of 120 cM, where the QTL was simulated at 56 cM (indicated by an arrow)

superiority of the combined LA/LD analysis over linkage analysis was minor in datasets with few families of large sizes, as we investigated here in accordance with most of the international studies. The design of our simulated datasets, nine sires originating from two great grand sire families based on structure 1, is representative for the QTL experiments in dairy cattle breeding reported in the international literature so far.

CONCLUSIONS

We used simulated datasets based on family structures taken from real cattle breeding incorporating fine mapping like marker maps and treated them as unique real datasets in our GDD and GPD analyses of QTL mapping. Both the GDD analysis and the GPD analysis were able to find the QTL region within relatively small CI(95%) when analyzing more complicated family structures and using proper marker maps. This could also be shown by results from medium sized datasets with overlapping generations and missing marker information. Due to the additional amount of pedigree information, analyzing data by GPD is more powerful than by GDD in general. The GDD analysis was superior in terms of precision of QTL position estimates when analyzing large offspring groups. Further, GDD analysis was more successful with large offspring groups in combination with fine mapping like marker maps covering the whole chromosome instead of a chromosomal region. It can be concluded that it is not necessary to increase the number of markers to obtain reliable estimates of the QTL region, when the right pattern of marker map is used. The GPD is probably a sufficient tool for QTL estimation in medium sized pedigrees. It is likely much more efficient to use more information on ancestral generations and to spend genotyping efforts for smaller offspring groups. The GPD methodology should be investigated further for efficient processing of inbreeding information. Further research is also required in order to cope with inbreeding and large datasets simultaneously.

REFERENCES

Darvasi A., Soller M. (1997): A simple method to calculate resolving power and confidence interval of QTL map location. Behav. Genet., *27*, 125–132.

Freyer G., Sørensen P., Kühn C., Weikard R., Hoeschele I. (2003): Search for pleiotropic QTL on chromosome BTA6 affecting yield traits of milk production. J. Dairy Sci., *86*, 999–1008.

Gilmour A.R., Gogel B.J., Cullis B.R., Welham S.J.S., Thompson R. (2002): ASReml User Guide Release 1.0. VSN International Ltd., Hemel Hempstead, HP1 1ES, UK.

Khatkar M.S., Thomson P.C., Tammen I., Raadsma H.W. (2004): Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol., *36*, 163–191.

Knott S.A., Haley C.S. (1998): Simple multiple-marker sib-pair analysis for mapping quantitative trait loci. Heredity, *81*, 48–54.

Lee S.H., van der Werf J.H.J. (2004): The efficiency of designs for fine-mapping of quantitative trait loci using combined linkage disequilibrium and linkage. Genet. Sel. Evol., *36*, 145–161.

Mayer M., Liu Y., Freyer G. (2004): A simulation study on the accuracy of position and effect estimates of linked QTL and their asymptotic standard deviations using multiple interval mapping in an F2 scheme. Genet. Sel. Evol., *36*, 455–479.

Meuwissen T.H.E., Karlsen A., Lien S., Olsaker I., Goddard M.E. (2002): Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics, *161*, 373–379.

Pong-Wong R., George A.W., Wooliams J.A., Haley C.S. (2001): A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol., *33*, 453–471.

Schelling M., Stricker C., Fernando R.L., Künzi N. (1998): PEDSIM – A simulation program for pedgree data. In: Proc. 6th World Congr. Genetics Applied to Livestock Production, Armidale, Australia. Vol. 27, 475–476.

Sørensen P., Lund M.S., Guldbrandtsen B., Jensen J., Sorensen D. (2003): A comparison of bivariate and univariate QTL mapping in livestock populations. Genet. Sel. Evol., *35*, 605–622.

Vukasinovic N., Martinez M.L. (2002): Mapping quantitative trait loci in general pedigrees using the random model approach. In: Proc. Joint Stat. Meet. (CD-ROM), American Statistical Association, New York City, USA. Available at http://www.math.usu.edu/~vukasino/text/JSM2002_paper.pdf.

Received: 05–05–04
Accepted after corrections: 05–08–23

Corresponding Author

Dr. Natascha Vukasinovic, Monsanto Company Animal AG, St. Louis, 63167 Missouri, USA
Tel. +1 314 694 6898, fax +1 314 694 7120, e-mail: natasha.vukasinovic@monsanto.com