IMPLEMENTATION OF MARKER-ASSISTED SELECTION: PRACTICAL LESSONS FROM DAIRY CATTLE

D. Boichard, S. Fritz [1], M.N. Rossignol [2], F. Guillaume, J.J. Colleau, and T.Druet

INRA, Station de Génétique Quantitative et Appliquée, 78352 Jouy en Josas, France
[1] UNCEIA, 149 Rue de Bercy, 75 595 Paris Cedex 12, France
[2] LABOGENA, 78352 Jouy en Josas, France

INTRODUCTION

In the last decade, advances in molecular genetics have made it possible to dissect the genetic variability of complex traits into quantitative trait loci (QTL). In dairy cattle, after the pioneering work of Georges *et al* (1995), many large scale QTL detection experiments were designed to exploit the population structure and the recording systems existing in large dairy breeds. Most of them used the so-called granddaughter design. A recent review (Khatkar *et al* 2004) summarized the results of 55 experiments and presented a meta-analysis. Most of these experiments were carried out by consortia gathering both research and industry and some of them clearly aimed to generate results to be used in selection. Traits analysed were restricted to those present in the national data bases, although some rare designs set up in experimental facilities addressed more original traits (Larroque *et al*, 2002). As these QTL were very poorly first located, most groups oriented their activity towards fine-mapping. To date, only few QTL have been fully characterized with a strong putative or well-confirmed causal mutation: *DGAT1* on chromosome 14 (Grisard *et al* 2002; Winter *et al* 2002; Kühn *et al* 2004), *GHR* on chromosome 20 (Blott *et al* 2003), *ABCG2* (Cohen-Zinder *et al* 2005) or *SPP1* (Schnabel *et al* 2005) on chromosome 6. They all affect milk production traits, whereas no QTL affecting functional traits such as fertility or mastitis resistance has been characterized so far. In the near future, however, many QTL would be characterized or, at least, fine-mapped following the international effort in the development of the genomic tools (bovine genome sequence, SNP validation, functional genomics) and in fine-mapping projects.

Even if the genes involved are still unknown, individual QTL information could enhance selection efficiency and dairy cattle provide the best technical and economic opportunity to implement marker-assisted selection (MAS). Although theoretical bases have been extensively studied since Fernando and Grossman (1989), to our knowledge only a few initiatives resulted in large-scale MAS programmes. Two consortia started in 2000 and 2003 in France (Boichard *et al* 2002) and Germany (Bennewitz *et al* 2003), respectively. MAS development still remains limited due to its cost, its high organisational demand, to the limited number of genes of importance fully characterised, and also to some lack of confidence of users. In this paper, we present the lessons from this MAS programme implemented in 2000 in French dairy cattle.

CONDITIONS OF APPLICATION OF MAS

Whatever the knowledge about the genes involved, QTL information could theoretically enhance selection efficiency by decreasing generation interval and/or increasing selection pressure and index accuracy. In most situations, however, identified or marked genes explain only a fraction of the total genetic variability of selected traits and, consequently, selection on molecular information only cannot simply replace conventional selection. The best results are usually obtained by adding new early selection steps in the breeding scheme and using all available information by combining molecular and phenotypic information.

In this paper, MAS refers to any selection procedure incorporating molecular information about QTL or known genes. MAS is easier to implement when the causal mutations are known, but one may need to implement MAS before they are discovered. Indeed, whereas a QTL detection experiment has a well defined time and budget framework, the identification of a gene is a less predictable task and waiting for it is not always reasonable.

MAS is known to be particularly beneficial when the traits of interest are difficult or expensive to measure (trait not expressed, sex-limited or expressed late in life, invasive measure such as disease challenge or recording after slaughtering), when each individual performance brings little information to breeding value prediction (trait with low heritability, recessive or low penetrance genetic determinism), or, more generally, when the polygenic approach has limited efficiency or a high cost. Therefore, it is believed that MAS could be particularly profitable in dairy cattle which concentrates conditions unfavourable to conventional selection and, therefore, favourable to MAS: most traits of interest are sex-limited; generation interval is long; progeny tested is a long and costly step; bull dams are often selected before their first lactation on pedigree information only; finally, functional traits, such as disease resistance or fertility, with a low heritability, have increasing weight in the breeding goal.

Three steps could be distinguished in the characterisation of a QTL:
1) A primo-localisation of the QTL is first obtained. Because the true location of the QTL in this region is unknown, the confidence intervals on QTL location is large (>10 cM) and the linkage disequilibrium with genetic markers present in a given generation (within family, or resulting from crossbreeding…) could rapidly decline over a few generations.
2) The fine-mapping step narrows the confidence interval and provides new markers in the vicinity of the QTL. Recombination events between these markers and the QTL are rare and any linkage disequilibrium could be maintained at the population level over many generations.
3) The gene involved is identified, as well as the DNA polymorphism responsible for the genetic variability of the phenotype.

These three situations provide opportunities for MAS, but with different cost and efficiency. Situation 3 (MAS3) is the most favourable and the simplest one to be implemented. Genetic merit of candidates is simply predicted from their genotyping results, at least for QTL with additive effects. In the simple case of two alleles, a single test considerably limits the lab cost. The total cost, however, should include possible intellectual property fees, which are a major challenge in the future. For a trait with a mixed inheritance, a total breeding value could be predicted by combining molecular and phenotypic information into a molecular score, as proposed by Lande and Thompson (1990). More generally, all the information can be combined in a BLUP statistical model where the phenotype is described by the fixed effect of the genotype at the locus, the random polygenic fraction and a residual. Such a model, however, requires inferring the genotype of all individuals, at least in probability, although there are some approximations when this is not the case (Tier and Bunter, 2003). This limitation probably explains why the molecular information is frequently used in a very crude way only in the early steps of selection, as additional and independent information.

The first generation MAS, denoted MAS1, takes advantage of the QTL primo-detection results. Linkage disequilibrium (LD), for instance after crossbreeding, could be used to select candidates during a few generations but its efficiency is rapidly decreased by recombination events. This is the most common way to apply MAS in plants. In an outbred population in linkage equilibrium, there is no preferential association between marker alleles and QTL alleles and it is not possible to select for given marker alleles. LD, however, is present within-family and could be used in selection. The marker information is used to estimate the relationship matrix between relatives at the QTL locus or, equivalently, the probability of identity-by-descent (PID) of chromosome segments of two related individuals (Wang *et al* 1995; Pong Wong *et al* 2002). In the case of a major gene, these PID could be used to infer the unknown genotype of the candidates from their relatives. This approach was used in France to eradicate the Achondroplasia (Bulldog) defect in Holstein cattle or to introgress the Polled gene in a

Charolais population. For a QTL, it is recommended to combine the QTL information with the phenotypic and pedigree information into an index estimated by marker-assisted BLUP (Fernando and Grossman 1989; Goddard 1992). The total breeding value is partitioned into a polygenic value and the additive values of the parental chromosome segments. Covariances between polygenic values are proportional to the conventional relationship coefficients, whereas the covariances between QTL values are assumed to be proportional to PID.

In contrast to MAS3, MAS1 presents severe limitations and a high cost. The information is not in the nature of the marker alleles but in the QTL parent to progeny transmission they reveal. Each QTL effect is estimated from the phenotype of the individual and of all relatives carrying this QTL in probability. Therefore, for each QTL, several markers have to be genotyped in order to accurately trace chromosomal segments and compensate for the incomplete marker informativity. Flanking markers are useful to assess recombination events. When the segment is large, recombinations are frequent after several generations and most information is lost. Markers should be genotyped not only for candidates but also for a number of relatives with phenotypic information. Mackinnon & Georges (1996) and Colleau (1999) proposed some possible designs. Last but not least, only a part of the QTL genetic variability can be used by MAS1. It is often argued that only Mendelian sampling within heterozygous parents is available for selection. It is not entirely true because the QTL genotype of homozygous parents could be inferred if the design is well connected over a number of generations, allowing for an increase in evaluation accuracy and selection across families. Most often, however, the latter condition is not fulfilled, and only within-family MAS is available.

As a summary, due to the multiple sources of information loss (recombinations, non-informativity of markers…), MAS1 technical efficiency is much lower than MAS3 and is superior to conventional selection only in some favourable conditions (Ruane & Colleau 1995; Spelman & Bovenhuis 1998). Moreover, because several markers are required to trace each QTL and many non-candidates should be genotyped, MAS1 is more expensive than MAS3. Combining a high cost and a limited technical efficiency, MAS1 is profitable only when the cost of the classical breeding scheme is high.

When the QTL are accurately mapped with very close markers, recombination events are rare and these markers are maintained in LD with the QTL at the population level. This information used in MAS2 can enhance MAS efficiency. In an outbred population, MAS2 could be implemented as follows. Within-family LD is accounted for as in MAS1 (but provides more information because markers are closer to the QTL). Whereas MAS1 assumes founder QTL effects are independent and with zero expectation, MAS2 uses LD at the population level to generate ties between founders. Several methods have been proposed to estimate PID of founders QTL from marker haplotype information. These methods were developed for fine-mapping but they can also be used for genetic evaluation. Farnir *et al* (2002) assumed a biallelic QTL with one allele appeared after a mutation. Meuwissen & Goddard's approaches (2000, 2001), based on genetic drift or coalescence, do not make any assumption on the number of QTL alleles. Computational difficulties arise from the large number of non-zero terms generated in the mixed model equations. A less demanding alternative is to cluster founder QTL effects according to their PID. A grouping strategy could also be used in the case of recent crossbreeding between populations, with groups being defined according to the known or inferred origin of the QTL alleles. The efficiency of MAS2 could be much higher than that of MAS1, for three different reasons: it uses much better basic information (QTL are accurately located), recombination events are rare (PIDs are close to one or zero), and across families information is used. As a result, MAS2 could be nearly as efficient as MAS3, without intellectual property limitation, but it remains more complex to use.

DESCRIPTION OF THE FRENCH DAIRY CATTLE MAS PROGRAMME.

In French dairy cattle, a large scale MAS1 programme was implemented in 2000 (Boichard *et al* 2002). It was carried out by a consortium of three partners, INRA (Research), LABOGENA (genotyping lab) and UNCEIA (artificial insemination (AI) industry federation) on behalf of eight breeding companies operating in the three main French dairy cattle breeds (Holstein, Normande, and Montbéliarde). During the first four years, 12 chromosome regions were chosen from a QTL detection experiment (Boichard *et al* 2003). These regions, 5-30 cM long, carry QTL affecting at least one of the following traits: milk, fat, or protein yield, fat or protein content, somatic cell score, female fertility. Regions affecting milk production or composition were located on chromosomes 3, 6, 7, 14, 19, 20, and 26. Those affecting mastitis resistance were on chromosomes 10, 15, and 21. Finally, those affecting fertility were on chromosomes 1, 7, and 21. Each region was found to affect 1 to 4 traits, and 4 regions on average contributed each trait. According to initial estimates, each QTL contributed from 8 to 20% of the genetic variance of the trait (except *DGAT1*, which contributed 40% of fat content). For each trait, several QTLs were considered, in order to contribute a high fraction of the genetic variance and also to decrease the probability of a zero or nearly zero Mendelian sampling prediction for the candidates (due to homozygous parents or inefficient QTL tracing), a situation making MAS poorly unacceptable by the breeders.

Each region was initially monitored by 2 to 4 microsatellite markers evenly spaced and each animal was genotyped for 33 markers. This design was rapidly extended to 43 markers. This number was a compromise between MAS efficiency and genotyping cost.

The genotyped population included two kinds of animals, referred to as 'candidates' and 'relatives' (RE) with phenotypic information. Candidates included young males before progeny test and young females of high pedigree value, before first breeding. The strategy mixed both bottom-up and top-down approaches. Accordingly, RE included sire and dam of candidates, all male AI ancestors, AI uncles of candidates, and daughters of bull sires (figure 1). DNA of AI males was readily available from the INRA semen bank created in 1992 and maintained with the help of AI companies. About 10,000 animals were genotyped each year. A genetic evaluation was performed each month for 7 traits. The genetic evaluation system was a single trait multi QTL BLUP. PID were estimated with a method similar to that of Pong Wong *et al* (2001). QTL variances were estimated by REML. A phenotype was defined to concentrate the information of the rest of the population onto the evaluated animals, as proposed by Meuwissen & Goddard (1999). In order to keep the equation system as small as possible, phenotypes were precorrected for non-genetic effects (Ducrocq et al, 2001). Genotyped females were characterized by their average performance (with the appropriate weight), whereas males were characterized by twice the yield deviation of their ungenotyped daughters. With such a strategy, only ~50,000 genotyped animals and ungenotyped connecting ancestors were included in the evaluation in 2005 (*vs* 15 million animals in the official evaluation).

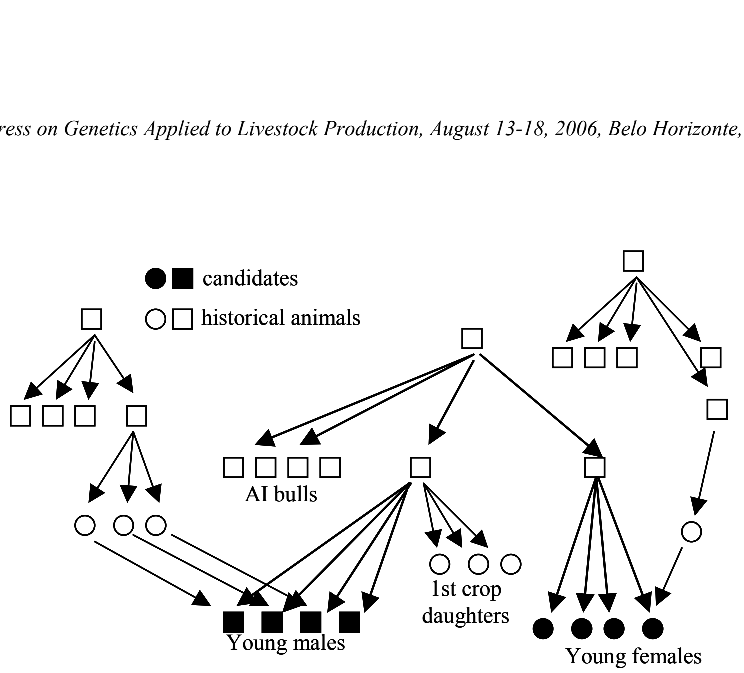

Figure 1. Representation of genotyped animals in a given family

The design was organized by candidate's paternal grandsire families. For each family chosen by the steering committee, "relatives" were genotyped in the next months. Then AI companies sent candidate samples and a constant genotyping work was maintained all along the year with a strict time schedule for sample reception, genotyping delivery, and genetic evaluation. With a monthly evaluation, results were sent back 4 to 6 weeks after sample reception.

The genotypings were carried out by LABOGENA and results were sent weekly to INRA. All pedigree and phenotypic information was retrieved from the national data base. After the evaluation, the estimated breeding values (EBV) of the candidates were sent back only to the relevant breeding company. The information sent back was the EBV for each trait and the gain in accuracy. All information pertaining to individual QTL remained confidential, as well as EBV of "relatives". The reason was to avoid any confusion between MAS and official EBV for older animals. For a breeding company, fees included a part proportional to the number of candidates, and a part proportional to the AI number to cover the RE genotyping cost.

LESSONS FROM THE MAS PROGRAMME.

Over the period, the design evolved slightly: udder depth was included in the list of traits; the set of markers was also adapted to account for the progress in fine mapping and included 45 markers in 2005. QTL detection was carried out regularly either with the whole sample, within breed, or even within strain. QTL variance components were re-estimated within breed from much larger data sets by REML (Druet *et al* 2006) and are presented in table 1. They were, as expected, consistently smaller than initial estimates of Boichard *et al* (2003). Accumulating data and experience made it possible to estimate MAS efficiency and to fine map some QTL.

Table 1. Proportion (%) of genetic variance of dairy traits explained by seven regions

Trait	Chromosome						
	3	6	7	14	19	20	26
Milk yield			6.2	14.7		7.1	4.1
Fat yield		4.9	4.7	10.3	3.3	5.3	6.5
Protein yield		3.1	4.7	9.1		4.2	5.8
Fat content		4.3		35.5	3.5	8.0	
Protein content	5.8	10.4		8.7		13.7	

QTL fine-mapping. The MAS programme, as projected initially, generated a unique resource for fine-mapping. The idea was to build a synergy between MAS and fine mapping. MAS provided a large number of animals with primary genotypes to fine-mapping work; on the other hand, progress in QTL characterisation directly benefited to MAS. A scan of 12 regions was performed on nearly all AI bulls progeny-tested in France. Nowadays, 50, 19, and 14 half sib families in Holstein, Normande and Montbéliarde breeds, respectively, have been added to the initial granddaughter design of 14 families. In total more than 10,000 AI bulls with progeny-test information entered the MAS programme and this number increases by 1000 per year.

Among these families, 3 within-breed samples (or "kits", Gautier *et al*, these proceedings) were defined for QTL fine mapping purpose. In a first step, three chromosomes carrying QTL for production, mastitis resistance and fertility, were specially targeted (7, 15, and 26). A major finding encountered was that fine-mapping detected several linked QTL in each region.

With large samples new QTL were detected, *e.g.* for production on chromosome 2, fertility on chromosome 3, or somatic cell counts on chromosome 6. In 2006 all regions were simultaneously studied in a large scale project based on SNP and new genotyping technologies.

MAS Validation. The gain in prediction accuracy over pedigree index (Table 2) reached 0.05 to 0.19 according to the traits. These values are expected to increase over time. A major challenge was to convince the AI industry to fund programme and to use MAS predictions in selection procedures. Therefore validations were produced when candidates had their first phenotypic information, in 2003 for females and 2005 for AI males. In addition, 800 females were genotyped especially to generate a data set for validation. All these results were in agreement and showed that MAS was able to predict a part of the Mendelian sampling (MS) further expressed through performances or progeny data: the correlation was positive and reached 0.15 to 0.30 according to traits between MAS prediction and MS after performance (females) or progeny test (males). A popular illustration was the comparison of full sibs. Within family, MAS was able to detect 72% of the sibs with the best economic index and to generate within family average differences of 0.4-0.5 genetic standard deviations.

Table 2. Average gain in accuracy (R^2) of MAS vs pedigree index

Breed	Milk	Fat content	Cell counts
Montbéliarde	0.08	0.11	0.06
Normande	0.09	0.09	0.05
Holstein	0.09	0.19	0.06

Moving to MAS2 (and MAS3). Moving to MAS2 requires: a) a very accurate estimate of the QTL location, which is already achieved internally or through the literature for 7 QTL; b) an update of the marker set and the genotyping of the past animals; c) an update of the evaluation software. We developed a version based on clusters of founder QTL, defined from relationships. To date, a MAS2 beta version is under test.

Table 3. Reduction in number of sampling bulls allowed by MAS without reduction in genetic gain

Selected Trait	Reduction (%)	Selected Trait	Reduction (%)	Selected Trait	Reduction (%)
Milk	**10**	Fat content	**> 20**	Cell count	**7**
Fat	**> 20**	Protein content	**20**	Fertility	**13**
Protein	**13**				

MAS and profitability. As 100 bulls are selected as elite sires and marketed each year, the cost of each elite sire should include the genotyping of 10,000/100=100 animals. According to Colleau (1999), the optimum design would be even larger and involve up to 160 genotyped animals per elite bull. This apparently large extra-cost, in fact, is rather small for an elite bull and could be easily balanced by a small reduction (<5%) of the number of bulls entering progeny-test. According to simulations, after 3 years of MAS1 activity, reducing size of batches by 10-20% would not decrease genetic gains (Table 3). Optimized selection procedures would also reduce by 15% co-ancestry between selected candidates. Therefore MAS, even in the first years, appears to be profitable in dairy cattle. Moreover, it is the straight way to prepare the implementation of MAS2 and MAS3. This favourable situation is rather specific to dairy cattle, due to the cost of its breeding scheme and the high value of the elite sires. It is usually believed that MAS1 cannot be profitable in any other species.

CONCLUSION

This MAS uses a data base and an evaluation system common to the populations involved. It used all French information to accurately estimate EBV and to share the genotyping cost of “relatives”, whereas results remained private and were distributed only to the appropriate breeding company. Few years after implementation, gain in accuracy (R^2) reaches 5 to 19% for marker-assisted pedigree index. Although the efficiency of the design is continuously increasing, efforts are oriented towards using linkage disequilibrium and fine mapping results, in order to strongly increase efficiency as well as to simplify the practical implementation.

REFERENCES

Bennewitz J. *et al* (2003). 54th Ann Meet EAAP, G1.9.

Blott, S., Kim, J.J., Moisio, S., *et al.* (2003). *Genetics* **163:** 253-266.

Boichard D. *et al* (2002). Proc 7th WCGALP, 22-03.

Boichard D. et al. (2003) *Genet. Sel. Evol.* **35**: 77-102

Cohen-Zinder M. *et al.* (2005) *Genome Res.* **15**:936-44.

Colleau J.J. (1999) 6th Rencontres Recherches Ruminants, Paris, 231-233.

Druet T., Fritz S., Boichard D., Colleau J.J. (2006) *J Dairy Sci* (in press)

Ducrocq V., Boichard D., Barbat A., Larroque H. (2001) *52nd Eur. Assoc. Anim. Prod.,***7**: 2

Farnir, F., Grisart B., Coppieters W., et al. (2002). *Genetics*, **161:** 275–287

Fernando R., Grossman M. (1989) *Genet. Sel. Evol,* **21**:467-477

Georges M. et al (1995). *Genetics,* **139**: 907-920

Goddard M.E. (1992). *Theor. Appl. Genet*., **83**: 878-886

Grisart, B. *et al.* (2002) *Genome Res.* **12**:222-31.

Khatkar, M.S. et al. (2004) *Genet Sel Evol*. **36**:163-90.

Kuhn, C. et al. (2004) *Genetics* **167**:1873-81.

Lande R., Thompson R. (1990). *Genetics*, **124**: 743-756

Larroque H., Gallard Y., Thaunat, L., et al (2002). Proc 7th WCGALP, 01-42.

Mackinnon M.J., Georges M. (1998). *Livest. Prod. Sci.,* **54**:229-250

Meuwissen T.H.E., Goddard M.E. (1999) *Genet Sel Evol,* 31: 375-394

Meuwissen T.H.E., Goddard M.E. (2000) *Genetics,* **155**: 421-430

Meuwissen T.H.E., Goddard M.E. (2001) *Genet Sel Evol,* **33**: 605-634

Pong-Wong R., George A.W., Woolliams J.A., Haley C.S. (2001) *Genet Sel Evol,* **33**: 453-472

Ruane J., Colleau J.J. (1995*). Genet. Res.*, **66**:71-83

Schnabel, R. *et al.* (2005) *PNAS* **102**:6896-6901.

Spelman R.J., Bovenhuis H. (1998). *Genetics*, 148:1389-1396

Tier, B., Bunter K. (2003) Proc. 15th AAABG, 214-217.

Wang,T., Fernando R.L., van der Beek S., *et al.* (1995). *Genet. Sel. Evol.* **27**:251–274.

Winter A. et al (2002). *PNAS*, **99**:9300-9305