Genomic Selection In French Dairy Cattle

D. Boichard[*], F. Guillaume[*†], A. Baur[‡], P. Croiseau[*], M.N. Rossignol[§], M.Y. Boscher[§], T. Druet[**], L. Genestout[§], A. Eggen[*], L. Journaux[‡], V. Ducrocq[*], and S. Fritz[‡]

Introduction

In 2000, a large scale program of marker-assisted selection (MAS) was implemented in French dairy cattle (Boichard et al, 2002, 2006). It was carried out in the main three French dairy breeds (Holstein, Normande, and Montbéliarde) by a consortium of three partners, INRA (Research), LABOGENA (genotyping lab) and UNCEIA, on behalf of eight breeding companies. Fourteen chromosome regions were chosen from an initial QTL detection experiment (Boichard et al, 2003) and additional subsequent information. Each region was 5-30 cM long and was traced by 3-4 microsatellite markers, and animals were genotyped for 45 markers. No population wide linkage disequilibrium was assumed and only within family information was used. Many relatives (50% of all animals genotyped) with phenotypes had to be genotyped in order to accurately evaluate young candidates to selection. After 7 years of activity and more than 70,000 animals genotyped, the efficiency of this program was shown to be close to its expectation (Guillaume et al, 2008a,b), *i.e.* rather limited but large enough to reduce the number of bulls entering progeny test by 15% through a better choice of the young candidates and, therefore, to generate a positive return.

However, since 2005, it had been anticipated that high-throughput SNP would be rapidly available, and would open the way to MAS based on linkage disequilibrium or to Genomic Selection. At that time, probably because of our previous experience in MAS, we trusted MAS more than Genomic Selection. In 2008, the first generation MAS program was replaced by a new version based on high-throughput SNP with a much larger expected efficiency, with the same partners and management. In this paper, we present the ideas, the philosophy and the evolution of this new program.

Fine mapping with a large reference population

A fine-mapping project called "CartoFine" was launched in 2005, funded by the French National Research Agency (ANR) and the industry (ApisGene). The initial idea was to develop a set of SNP, to produce a dedicated chip and to genotype a reference population of 3,200 bulls. A virtual chip was developed from an *in silico* analysis of all sequences present in the public data bases. At the beginning of the study, the bovine sequence was not yet

[*] INRA, UMR1313 Gabi, 78350 Jouy-en-Josas, France
[†] Institut de l'Elevage, 78350 Jouy-en-Josas, France
[‡] UNCEIA, 149 rue de Bercy, 75595 Paris, France
[§] Labogena, 78350 Jouy-en-Josas, France
[**] Liège University, Belgium

published and almost 100,000 SNP were mapped on the bovine genome through comparative mapping with the human genome. A pilot study was then conducted on a selection of 1,536 SNP located mainly on chromosome 3, genotyped with the Illumina Golden Gate methodology on 1,800 bulls, in collaboration with the National Genotyping Centre (CNG, Evry, France). This first study convinced us that a careful choice of SNP found in public data bases led to highly informative markers in a large range of breeds. It gave us some first results on the extent of linkage disequilibrium (LD) within and across breeds (Gautier et al, 2007) and on the potential of the Linkage Disequilibrium and Linkage Analysis approach (LDLA, Meuwissen et al, 2000) for QTL fine mapping (Druet et al, 2008). From this experiment, it became clear that small QTL confidence intervals close to 1 cM were realistic and that efficient MAS based on Linkage Disequilibrium was feasible.

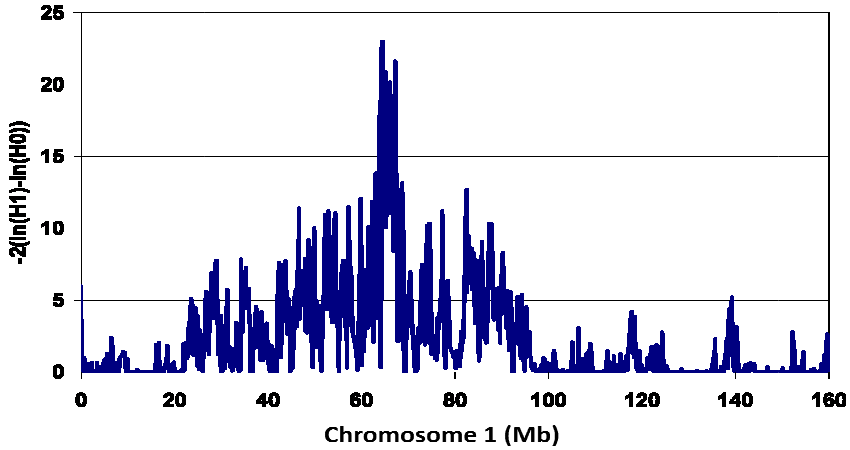

Figure 1: Detection of a QTL affecting milk yield on chromosome 1 in Montbéliarde breed

For strategic and financial reasons, this project of a private chip, although well advanced, was stopped and reoriented in 2007 when we decided to join the North American consortium and to use the publicly available Illumina BovineSNP50^{TM} beadchip. A posteriori, this difficult choice was the right one, although we often regret that some particular markers are not present on this chip.

A first set of 3,200 dairy bulls was genotyped in 2007-08 by the National Genotyping Centre (CNG, Evry, France) and analyzed by LDLA analysis with the same software as in Druet et al (2008) and Tarres et al (2009). This population was distributed in 3 breeds (1800 Holstein, 700 Normande, and 700 Montbéliarde bulls) and 69 sire families. Each AI bull was evaluated for 25 traits after progeny test. LDLA is a powerful, accurate and robust approach. It is robust as it accounts for family information, mainly sire families, and it is powerful and accurate as it accounts for population LD, mainly from the numerous dams. SNP with unknown location on the map or with a minor allele frequency (MAF) lower than 0.05 were deleted. Finally, from 38,885 to 40,757 SNP were used in each breed. Several hundreds of

QTL were detected, on average from 20 to 40 per trait, with a stringent significance threshold. Most already known QTL were confirmed, including all those used in first generation MAS, but many new ones were detected. Not more than one third of the detected QTL were shared across breeds, although most major already known QTL did. 80% of the detected QTL explained a proportion of genetic variance ranging from 1.5 to 6% (with a mode at 3%), i.e. much lower values than initially estimated (Boichard et al, 2003) and even revised (Druet et al, 2006). Only few QTL explained more than 10% of the total genetic variance. As a whole, at least 50% of the genetic variance was explained for each trait (and sometimes up to nearly 100% for some traits such as fat content). Due to the large number of dams and LD information, the location confidence intervals were generally very small (figure 1) and multi-QTL analyses were able to clearly distinguish linked QTL (for instance, on chromosome 20, figure 2).

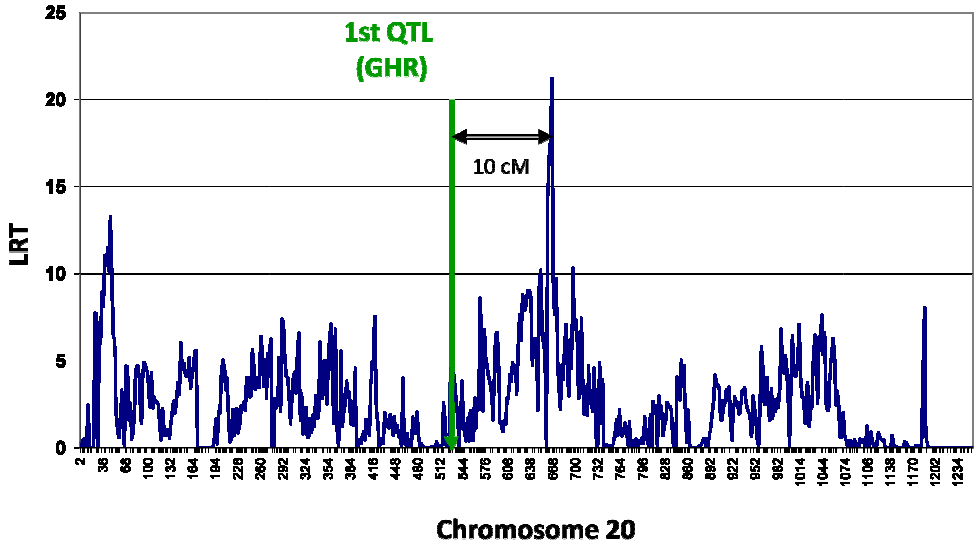

Figure 2: Detection of a second QTL (PRLR ?) affecting protein content on chromosome 20 in Holstein breed (bi-QTL analysis including GHR).

Use of these QTL in a first Genetic evaluation

In contrast with most evaluation teams throughout the world who oriented their developments towards genomic selection (GS), we maintained our initial choice of MAS. In its usual definition, genomic selection (GS) does not rely on information on particular QTL. Several implementations have been proposed but all of them estimate marker effects in a reference population and use these estimates to evaluate candidates. On the other hand, we had a strong experience in MAS and the switch from microsatellites to SNP was easier to manage just by extending MAS rather than by changing the complete approach. Moreover, we assumed that MAS based on many fine-mapped QTL could be as efficient as GS. The rationale behind this assumption was the following:

- Fine-mapped QTL could be traced by a haplotype of flanking markers with very limited risk of loss over time. Consequently, its use is easier over several generations without updating the reference population. When a QTL is confirmed across breeds, its use is easier without large reference population in the other breeds.
- Although this property is still debated, we believe that haplotypes are more efficient than single SNP to catch QTL information. There is an optimal number of SNP to consider. On the one hand, the number of different haplotypes should be large enough to likely generate a complete LD with the QTL ; on the other hand, it should be small enough to limit the number of effects to estimate. In practice, 15 to 25 haplotypes are usually found as the best solution. This usually corresponds to 4-6 markers per haplotype, depending on the informativity of each SNP and the population size.

Of course, MAS is limited in efficiency by the proportion of genetic variance explained by the individual QTL. Good results of MAS could be expected only if the major part of the genetic variance is explained by QTLs. With 50 to 100% of genetic variance explained and around 60% on average, MAS efficiency is expected to be good but still incomplete. From the beginning, it appeared very clear that MAS should be improved by accounting for the remaining part of genetic variance.

The first genetic evaluation based on these QTL was rather simple. It included all genotyped animals and three generations of ungenotyped ancestors. As for the previous MAS program, the phenotypes were derived from the national evaluation, as well as corresponding weights. They were daughter yield deviations (restricted to non genotyped daughters) for bulls and yield deviations for females. It should be emphasized that MAS relies on a good and unbiased classical evaluation (at both the national and international levels). The MAS evaluation model included the QTL effects and the residual polygenic breeding value, all effects being random and uncorrelated to each other. The variance of the QTL effect was that estimated in LDLA analysis and the variance of polygenic effect was the difference between the total genetic variance and the sum of the QTL variances. For some traits and breeds, it was arbitrarily set to 40%. The original model included 20 to 40 QTL per trait, according to the trait and breed. Each QTL was traced by haplotypes of 5 successive markers. For each QTL, one effect was estimated for each of the marker haplotypes present in the population.

Such an approach required to infer the missing genotypes and the different marker haplotypes at each QTL for each evaluated animal. In practice, all the phases were computed for the entire genome of each animal and this step represented the time-consuming part of the computations. The total breeding value of each animal was computed as the sum of all QTL effects and the residual breeding value. Reliabilities were computed by direct inversion of the Mixed Model Equations.

The results from this evaluation have been made available to the breeding companies since October 2008 for internal purpose. More than 20,000 young candidates were genotyped in the first year, 40% being females. To allow some young bulls to be marketed, some MAS breeding values were made official in June 2009. In practice, only the subset of those

marketed bulls received official genomic breeding values. The genomic evaluation service is expected to be fully opened in late 2010 to all users, the transition period being used to improve and stabilize the model and to set up the management system.

Comparison of MAS with Genomic Selection

MAS uses information of a given number of QTL with locations and variances assumed to be known. Conversely, GS usually does not make any strong assumption on the number and the parameters of the QTL and lets the data find informative SNP and estimate their variance. GS is generally believed to be superior to MAS because it theoretically uses all the genetic variance and because the number of SNP is large enough to catch all the genetic variance.

The ideal model could be defined as the model accounting for all (unknown) causal mutations. For a well characterized QTL, MAS uses marker haplotypes surrounding the causal mutation(s) and this haplotype is probably the best proxy of the QTL. But it is less efficient for small QTL (with poorly estimated location) and inefficient for polygenes. On the other hand, GS extracts more information from the whole genome (particularly with the G-Blup approach) but it does not take into account the knowledge about the individual QTL. Of course, one can argue that many SNP could be as efficient as the best haplotypes: because of their high number, there are always enough SNP to catch the QTL variance. But GS generally needs several linked SNP to catch all the information of a QTL, and the best SNP could be rather far from the QTL, provided that it is in strong LD with it. Such a model with individual SNP is likely to lack stability. It is particularly the case when several linked (not independent) SNP are needed to explain the same QTL and when the SNP in highest LD are not the closest to the QTL. In practice, GS is expected to be efficient to predict the breeding values of close relatives of animals with phenotypes and to gradually loose efficiency with genetic distance.

One can imagine that MAS and GS would converge to the same ideal model when: (1) MAS uses many QTL and a large proportion of the genetic variance, (2) MAS is extended to properly account for the non detected QTL (possibly with a GS approach applied to the rest of the genome), (3) GS fully uses LD between markers and QTL and, ideally, uses marker haplotype information. The last point is rather critical.

The true (additive) genetic model includes all the QTL, each of them being characterized by its variance. A model with all these QTL is equivalent to a genomic model where each chromosome region is weighted by its variance in the relationship matrix. A genomic BLUP is a good approximation of this true model when there are many QTL all with small variances. BayesB and derived approaches properly account for large QTL but are less efficient for small QTL th at are difficult to detect and regressed to zero.

Improvement of MAS

In this framework, MAS could firstly be improved by increasing the number of QTL accounted for. In the first version, only highly significant QTL were used. In a second

approach, many more QTL could be considered. With a less stringent threshold (LRT equal to 5), several hundreds of QTL (from 200 to 300 according to traits) were selected, of course with a larger proportion of false positive but with a much higher proportion of genetic variance explained. The location of each QTL was assumed to be the local LRT peak and each 4-Mb window was assumed to contain at most one QTL. The efficiency of MAS was strongly improved and is equivalent to all alternative approaches tested up to now. An optimum in the prediction ability in the validation sets was found with about 200 QTL, and no additional gain was observed with 250 or 300 QTL. However, assuming that these small QTL are true and fully characterized is excessive and one can imagine that a more proper treatment would be more appropriate.

A second and parallel improvement was achieved by increasing the size of the reference population by the genotyping of 3000 additional bulls (2000 Holstein, 500 Normande and Montbéliarde) with the Illumina BovineSNP50^{TM} beadchip.

But another way to increase the reference population is to exchange data with other partners. In a first step, exchanges have been organized with three other European organizations (Viking from Denmark, CRV from the Netherlands, and German AI industry). Each member of the EuroGenomics consortium exchanged data from 4000 bulls, leading to a large common Holstein reference population of 16,000 AI bulls (Lund et al, this meeting). One can assume that this consortium will grow in the near future and be opened to additional interested breeders and/or countries. Along the same idea, the French Brown joined the Brown Swiss consortium to benefit from a larger reference population in this breed.

For France, the two concurrent increases in the Holstein reference population led to an even finer mapping of the QTL and to the discovery of new QTL due to the increased power of the design. This first exchange of data with new partners also gave us an experience in genomic data standardization and on genotype imputation. Indeed, shared genotypes in Eurogenomics were obtained from two different chips and it was necessary for each bull to infer all missing genotypes of the other chip. Imputation is a very important tool for the future as it will enable to mix data obtained from different densities.

MAS improves its efficiency with more QTL included, even with a non-zero proportion of false positive. Although results are still preliminary, the optimum seemed to be found around 200 QTL per trait. These additional QTL were selected on the basis of their LRT value which exceeded 5, a very low threshold for such a high number of tests. Considering that they are true and accurately mapped is a strong assumption. Instead of including them directly in the MAS evaluation, alternative approaches were also used. The markers of these regions were preselected as input data in an Elastic Net selection procedure. This pre-selection of markers appeared to be successful in helping the procedure to select the most informative one. Croiseau et al (this meeting) present this study.

Another approach has been proposed by Legarra et al. (this meeting). The idea is to build a genomic BLUP with the appropriate variance. The most informative markers are selected by a LASSO procedure and their estimated variance is used to weigh them in the relationship matrix, whereas the other markers are given the same low variance. This method is very

appealing as it tries to reflect the biology, with some large individual QTL (spanning up to 20% of the genome) and the rest of the genome with a polygenic effect and a relationship matrix estimated from the markers.

Multi-breed evaluation

The current Illumina BovineSNP50™ beadchip includes one marker every 45kb on average, *i.e.* one informative marker every 70 kb. Compared to the length of the conserved segments within breed (several hundreds of kb for many breeds with a limited effective size), this density is theoretically large enough to catch all the genetic variability. It is however too small for an analysis across breeds. Indeed, because of their ancient divergence and an overall larger effective size, the segments conserved across breeds are much shorter, likely around 10 to 20 kb. Therefore, the density of the conventional chip is not suited to this genetic structure. A much larger chip (~800k) has been recently designed by Illumina in the framework of an international collaboration, from a large sample of breeds. With one marker every 4kb, this chip should be dense enough to detect the conserved segments across breeds.

Because of its higher cost, a systematic use of this chip is not realistic. But it is not necessary. A multi-breed genomic selection could be implemented by considering three populations: a reference population and a population of candidates, as for within breed genomic selection, and an additional population for efficient imputation. The reference population is usually genotyped with the 50k chip, candidates could be genotyped with the same chip or a low-cost chip of lower density. The imputation population is a mixture of animals of all breeds considered and is genotyped with the high-density chip. In practice, the imputation population should include several hundreds of animals per breed, it could be all or part of the reference populations and could also by composed of key ancestors and representative animals. Within each breed, this population makes it possible to impute all missing genotypes in the reference population and in the candidates. Consequently, all animals could be virtually genotyped with the high density chip. Across breeds, a probability of identity could be estimated between haplotypes, based on the shared information on conserved segments.

This approach, particularly appealing in beef cattle, is presently used to connect the three main French breeds but also to extend genomic selection to breeds of more limited size such as the Abondance, Tarentaise, Simmental, Brown, or Vosgienne breeds. This project will complement the international Brown Swiss initiative to build a large reference population for this breed. A main advantage of the multi-breed approach is that all breeds could share their reference populations, as long as traits have a similar definition. Within breed, the genomic evaluation relies on the reference population of the given breed, but also on those of the other breeds provided that QTL are still segregating in the different breeds.

Combining all these approaches, a realistic scenario could be a multi-breed evaluation with an imputation population genotyped with a very-high density chip (and, in the near future, by sequencing, providing several million polymorphisms), breed specific reference populations with accurate phenotypes and a high density genotyping, and candidates genotyped at high or low density, depending on the cost and the imputation quality.

Management of Genomic Selection in France

Most French results have been obtained in the framework of a consortium gathering public research, a genotyping lab and the French AI industry. This consortium decided to open the GS service to any user while paying back the investment. After the present transition period, the genomic evaluation computed by INRA will become fully official at the end of 2010. The extension to all French dairy breeds is a goal for late 2011. A company, Valogene, has been created to offer the service, to contract with genotyping laboratories, and to transfer genotype data to INRA. With a high GBV accuracy for both males and females, a large development is expected in the French population.

Characterization of QTL

Large reference populations provide a unique resource for QTL fine mapping and, therefore, QTL characterization. Use of linkage and linkage disequilibrium, applied to several thousands of animals, provides QTL location estimates with an accuracy never reached in the past. Merging reference populations from different breeding schemes of the same breed increases this accuracy. Assuming a common origin of the QTL alleles (which is a strong assumption), the merging of reference populations from different breeds could provide a very small location interval (smaller than a gene) and an excellent opportunity to find the underlying causal mutation. Associated with new generation sequencing, one can imagine large projects to characterize many QTL simultaneously. Although practical genomic selection does not require identifying the causal mutations, this exceptional information would clearly help selection and transposition of results to other populations. It will allow understanding the phenotypes determinism and to study the interactions between genes, one of the great challenges of the next years in genetics.

Acknowlegments

Cartofine (for fine mapping) and Amasgen (for GS development) projects were funded by the National Research Agency (ANR) and ApisGene.

References

Boichard, D., Fritz, S., Rossignol, M.N., Boscher, M.Y., Malafosse, A., Colleau, J.J. (2002). *In Proc 7th WCGALP*, 22-03.

Boichard, D., Fritz, S., Rossignol, M.N., Guillaume, F., Colleau, J.J. Druet, T. (2006) *In Proc 8th WCGALP,* 22-11.

Boichard, D., Grohs, C., Bourgeois, F., et al (2003) *Genet. Sel. Evol.* 35, 77-101.

Druet, T., Fritz, S., Boichard, D., et al. (2006) *J. Dairy Sci.*, 89, 4070-4076.

Druet, T., Fritz, S., Boussaha, M., et al. (2008). *Genetics*, 178, 2227-2235.

Gautier, M., Faraut, T., Moazami-Goudarzi, K, et al. (2007) *Genetics*, 177, 1059-1070.

Guillaume, F., Fritz, S., Boichard, D., et al. (2008). *Genet. Sel. Evol.,* 40, 91-102

Guillaume, F., Fritz, S., Boichard, D., et al. (2008). *J. Dairy Sci.,* 91, 2520-2522

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

Meuwissen, T.H.E., Hayes B.J., Goddard, M.E. (2001). *Genetics* 157: 1819–1829.

Tarres, J., Guillaume, F., Fritz, S. (2009). *BMC Proceedings,* 3(Suppl 1): S3.