GENOMIC SELECTION: THE FUTURE OF MARKER ASSISTED SELECTION AND ANIMAL BREEDING

Theo Meuwissen
Institute for Animal Science and Aquaculture, Box 5025, 1432 Ås, Norway,
theo.meuwissen@ihf.nlh.no

Summary
Marker-Assisted-Selection (MAS) is mainly important in situations, where the current accuracy of selection is low, e.g. traits with low heritability, limited, late-in-life, or after-slaughter recording. In velo- and whizzo genetics schemes, the number of selection cycles per time period is minimized, which may increase rates of genetic gain dramatically. Genomic selection is used to obtain a high accuracy of selection of in-uturo calves or early embryos.

Introduction
Currently, we have genetic marker maps available for many species, and Quantitative Trait Loci (QTL) - regions have been identified. The question arises: how are we going to use all this new information into animal breeding? I.e. how are we going to apply Marker Assisted Selection (MAS). The genetic gain (ΔG) in animal breeding programs can be calculated as:

$$\Delta G = intensity\_of\_selection * accuracy\_of\_selection * genetic\_standard\_deviation / generation\_interval$$

The new information from genetic markers mainly affects the *accuracy_of_selection*. Hence, in MAS schemes, genetic gain is mainly increased by increasing the *accuracy of selection*. However, most current animal breeding schemes are designed such that the *accuracy of selection* is already high (e.g. in progeny testing schemes the accuracy of selection may well be 90%). This implies that MAS will be especially useful for traits where the accuracy of conventional selection is low, such as:
- traits with low heritability;
- traits with few recordings (e.g. due to expensive recording);
- traits that are measured late in life, such that trait recordings are not available at the time of selection;
- slaughter quality traits (only available after slaughtering of the animals);
- disease resistance traits (requires expensive and risky challenge testing).

The fact that conventional selection schemes are designed to have high accuracy of selection also implies that we should reconsider the design of the breeding scheme, such that it fits the use of marker information better. For example, progeny testing schemes have a high *accuracy of selection*, but the *generation_interval* is also high (i.e. it takes a long time to perform a cycle of selection), which decreases the rate of genetic gain. MAS schemes could be designed that shorten the *generation intervals* considerably and still maintain a high *accuracy of selection* due to the use of marker information. If we could half the *generation interval* while maintaining the same *accuracy of selection*, genetic gain would double (see above equation for ΔG). This idea is taken to the extreme in the velogenetics (Georges and Massey, 1991) and whizzo genetics schemes (Haley and Visscher, 1998).

Current MAS schemes consist of the following steps:
1) find the biggest, statistically significant QTL(s) in a genome wide scan for QTL;
2) select for these big QTL next to selecting for polygenes (the remaining (often smaller) genes that have not been identified).
The polygenes can not be ignored because they constitute a large fraction of the total genetic variance. The velo- and whizzo genetics schemes with extreme short generation intervals can not be applied here, because an important fraction of the genetic variability, i.e. that due to the polygenes, is still selected for by trait recording (we still have to await trait records before turning over the generation).

The future of MAS schemes lies in the prediction of total genetic value (total value for large and small genes). Genomic selection aims at making those predictions of total genetic value, and its combination with velo- and/or whizzo genetics schemes seems especially fruitful. The aim of this paper is to describe current MAS schemes, and combinations of velo- and whizzo genetics schemes with genomic selection. Comparisons of the genetic gains of these schemes will be made.

State of the art in MAS
The state of the art in many species is that QTL regions have been identified, but that the size of the confidence interval is often rather large (10 cM or more). We will have markers in and surrounding this QTL region. We can include the information from these markers in our BLUP breeding value estimation models (Fernando and Grossman, 1989):

$$\mathbf{y} = \mathbf{Zu} + \mathbf{Qq} + \mathbf{e}$$

where **y** = data vector; **u** = vector of polygenic effects (to be estimated); **q** = vector of QTL effects (to be estimated). In stead of assuming that there are 2 QTL alleles, and estimating probabilities for each animal having one of these 2 alleles, the Fernando and Grossman model presumes that every animal has two unique QTL alleles, and then estimates the effects of all these alleles (infinite alleles model). If two alleles are quite likely the same, e.g. a parent and an offspring allele, where the flanking markers indicate that the offspring allele is probably a copy of the parent allele, this information is accounted for by a high correlation between the effects of the parent and offspring allele. In fact, the IBD (Identity-by-Descent) probability between any two alleles (which can be calculated from pedigree and marker data) equals the correlation between their effects. Thus, $\mathrm{Var}(\mathbf{q}) = \mathbf{G}\,\sigma_q^2$, where **G** = IBD matrix between all the QTL alleles and σ_q^2 is the variance due to the QTL. Further, as usual in animal model - BLUP estimation, $\mathrm{Var}(\mathbf{u}) = \mathbf{A}\,\sigma_u^2$, where **A** = relationship matrix between the animals and σ_u^2 is the polygenic variance. The marker assisted - estimated breeding value (MA-EBV) of animal i is now:

$$g_i = u_i + q_{ip} + q_{im},$$

where q_{ip} (q_{im}) = the effect of the paternal (maternal) QTL allele of animal i. Selection for g_i will utilize both the polygenic and the QTL variance.

Table 1 shows genetic gains for a MAS nucleus schemes when the trait recording was before selection (e.g., growth rate in pigs), or after selection (e.g., fertility traits or milk production in dairy cattle). When recording was before selection, MAS increased genetic gains by 8% in the first generation after the start of the MAS scheme, and by only 2% after 5 generations. This

shows that a) the extra response due to MAS is limited when selection is for traits that are also easily addressed in conventional selection schemes; b) the extra response due to MAS is reduced as the time period of the selection program becomes longer. The latter is because, both the MAS and the non-MAS scheme, will fix the positive QTL allele in the long term, and thus the differences between the two schemes will become small in the long term. MAS is thus a way to increase short term gains, not long term gains. This may still yield continuous benefits in an ongoing MAS scheme, because new QTL will be detected continuously.

Table 1
Genetic gain in a MAS scheme (the corresponding genetic gain in the non-MAS scheme is set to 100; Meuwissen and Goddard, 1996).

	Availability of records	
Generation	Before selection	After selection
1	108	138
2	106	131
3	104	125
5	102	116

($\square_q^2$ = .125; $\square_u^2$=.25; $\square_e^2$ =1; probability that inheritance of QTL allele could be traced from parent to offspring by markers was 90%)

Table 1 shows that the benefits of MAS are substantial when trait recording is after that the animals have been selected. In this situation, conventional selection has to rely on a pedigree index, which is not very accurate. Hence, the large increase in selection response of up to 38% when the MAS is used. The situation becomes even more extreme in the case of carcass traits, where genetic gains of MAS schemes are up to 65% higher than those due to conventional selection schemes (Meuwissen and Goddard, 1996). When selecting for carcass traits using conventional selection, some of the selection candidates are slaughtered in order to provide sib-information for the selection of their sibs. In the MAS scheme, all animals are candidates for the MAS programme, and only the non-selected animals are slaughtered in order to re-estimate the marker effects.

In conclusion, these 'conventional' MAS schemes yield 8-38% extra genetic gain, where the higher figures apply to traits that are difficult to address by conventional selection, e.g. because they are recorded after the selection step or after slaughtering of the animal. Also, MAS yields more extra gains for traits with low heritability. The extra genetic gain is only achieved in the short term, because the variance at the QTL decreases rapidly, but this may not be a problem if new QTL are detected continuously. It should be noted, however, that these situations in which MAS is most beneficial are also the situations in which QTL detection is most difficult, and often requires the use of special experiments instead of using field data.

Velo- and whizzo genetics schemes

The idea of shortening the generation intervals in MAS schemes was taken to the extreme in the velogenetics schemes of Georges and Massey (1991). They reduced the generation interval of cattle by harvesting oocytes from calves while still in utero. The harvested oocytes are matured and fertilized in vitro before being transferred to a recipient female. This process is repeated by harvesting oocytes from this second generation animals with the generation interval being reduced to 3 – 6 months. A few cells from the embryos can be used to

determine their marker genotypes, and these marker genotypes are used for selecting the animals.

The whizzo genetics schemes of Haley and Visscher (1998) reduce the generation interval even further. Cell cultures derived from fertilized oocytes will be selected based on markers. In the selected cultures, meiosis will be induced followed by fertilization. The resulting cultures could again be selected on marker information, and the cycle could be repeated. The complete breeding scheme could be performed in the lab, and the generation interval depends on the time needed to perform the required lab techniques. If these techniques reduce the generation interval by a factor X, than the increase in genetic gain is also by a factor X, *if* the same accuracy of selection can be maintained. The latter is not possible in phenotypic selection schemes, because the animals have not been phenotyped, and it is also not possible in schemes where only a part of the total genetic variance is explained by genetic markers. In the following, genomic selection will be described, which attempts to explain all genetic variation by genetic markers.

Genomic selection

The reason why a limited fraction of the genetic variation is explained by the identified QTL is that, in order to identify a QTL, we have to perform very stringent tests for statistical significance. These tests are stringent because we are testing many positions for the presence of a QTL, and if our tests were not stringent, we would find many false positives. The idea of genomic selection is to omit the significance testing, and simply estimate the effects of all genes or chromosomal positions simultaneously (Meuwissen, Hayes and Goddard, 2001). Genes with small (big) effects are expected have small (big) estimates, such that we can directly select for the estimates of the effects.

The entire human genome has been sequenced, and all $\pm$ 30,000 genes have been identified. In cattle and pigs, similar sequencing projects are underway, and probably also about 30,000 genes will be identified. We can search for polymorphisms within all these genes, and say on average one polymorphism will be found in each gene, leading to a total of 60,000 alleles. We would like to know the effect of all these alleles on our breeding goal traits. We could set up an experiment to estimate all these effects, where a typical experiment would comprise 1,000 – 2,000 phenotyped and genotyped animals. Classical statistics will tell us we have a problem: we want to estimate 60,000 effects from 1,000-2,000 records, i.e. we have not enough degrees of freedom (and this shortage of degrees of freedom is huge). A similar problem arises if we do not know all the genes but we use a dense marker map to identify all chromosomal segments (e.g. segments of 1 cM), and want to estimate the effects of all these segments simultaneously.

There are three ways to get around this huge shortage of degrees of freedom problem:
1) Least Squares (LS). Test all the genes one by one for their statistical significance, and set the effects of the non-significant genes to zero, while estimating the effects of the significant genes simultaneously using LS. Other stepwise testing approaches may be applied, but due to the degrees of freedom shortage not all genes can be tested simultaneously. Note that this resembles the QTL mapping / conventional MAS approach.
2) Best Linear Unbiased Prediction (BLUP). Fit the allelic effects as random effects instead of as fixed effects. The fitting of random effects does not require degrees of freedom, and thus all allelic effects can be estimated simultaneously. Random effects require however an estimate of the variance of the allelic effects. Obtaining such an estimate is no problem, but

the fact that for every gene the same variance is used is problematic, since the majority of the genes will have very little effect on the trait and they will dominate the estimate of the variance of the allelic effects, i.e. this estimate will be close to zero.
3) Bayesian estimation (Bayes). This is similar to BLUP, except that the variance of the allelic effects is assumed different for every gene, and is estimated by using a prior distribution for this variance. The prior distribution of the variance of gene i (V_{ai}) is assumed here:

$V_{ai} = 0$ with probability p
$V_{ai} \sim \chi^{-2}(v,S)$ with probability (1-p)

where p depends on the mutation rate at the gene, and $\chi^{-2}(v,S)$ denotes the inverse – chi squared distribution with v degrees of freedom and scale parameter S. The parameters v and S depend on the distribution of the mutational effects, and would need to be estimated in practice.
These three methods were tested in a simulation study, where the effects of 1 cM –large chromosome segments were estimated in genome of 1000 cM. The chromosomal segments were identified by dense marker haplotypes, and there were on average 50 different haplotypes per cM, i.e. a total of about 50,000 effects needed to be estimated. Estimation was performed in an experiment with 200 parental animals having a total of 2,000 offspring, where parental and offspring animals were phenotyped and genotyped for 1010 marker loci. The 2,000 offspring obtained again 2,000 offspring which were only marker genotyped, and their marker estimated breeding value (MEBV) was predicted by:

$MEBV_i = \Sigma_j M_{ij}$

where M_{ij} = the estimate of the effect of the j-th haplotype of animal i (the effect of M_{ij} was estimated in the 2,200 parents and grand parents of the animals i).

Table 2 shows the accuracy of selection, when selection is for $MEBV_i$. The accuracy was very low for the LS method, probably because LS picked up only a fraction of the total genetic variance, and because the stringent significance testing leads to overprediction of allelic effects. BLUP obtained a reasonably high accuracy of selection, which is comparable to that of animals with phenotypic records. Bayes obtained an even higher accuracy of selection, which is comparable to that of a progeny test. Thus, genomic selection can achieve an accuracy of selection of marker-genotyped embryos, e.g. in a whizzo genetics scheme, that is comparable to that of a progeny test. The combination of whizzo-genetics and genomic selection makes it possible to reduce the generation intervals many folds while maintaining a high accuracy of selection. The result is an many fold increase in the rate of genetic gain.

Table 2
Accuracy of genomic selection for marker genotyped, non-phenotyped, non-progeny tested animals (Meuwissen et al., 2001).

Method	Accuracy
LS	.36
BLUP	.74
Bayes	.84

References

Fernando, R.L. & Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* **21**: 467-477.

Georges, M., & Massey, J.M., 1991. Velogenetics, or the synergistic use of marker assisted selection and germline manipulation. *Theriogenology* 35: 151-159.

Haley, C.S., & Visscher P.M., 1998. Strategies to utilize marker – quantitative trait loci associations. *J. Dairy Sci.* 81(2): 85-97.

Meuwissen, T.H.E. & Goddard, M.E. (1996). The use of marker haplotypes in animal breeding schemes. *Genetics Selection Evolution* **28**: 161-176.

Meuwissen, T.H.E., Hayes B.J. & Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.