Performance Of Genomic Selection Depending On The Age Of New Mutations

J. Casellas[*] and L. Varona[†]

Introduction

Current availability of a plethora of single nucleotide polymorphism (SNP) markers in livestock species has led to multiple initiatives within the context of genomic selection programs. They rely on the main idea that the genetic variability originated by several dozens or few hundreds of quantitative trait loci (QTL) spread in the whole genome (Hayes and Goddard (2001)) would be captured by a dense panel of SNP markers by applying appropriate statistical models. This scenario was introduced by Meuwissen et al. (2001) and originated an authentic revolution in the animal breeding research field.

Genomic selection tries to exploit statistical dependencies existing in the joint distribution of SNP and QTL, i.e. linkage disequilibrium. It is demonstrated that linkage disequilibrium between a pair of genetic markers must reduce with successive generations (Bink and Meuwissen (2004); de Roos et al. (2008)), although little is known about its consequences in genomic selection models. Given the continuous uploading of new additive mutations in livestock and experimental species (Casellas and Medrano (2008); Casellas et al. (2010)), genomic selection models must deal with mutations originated from a wide range of generations ago. The objective of this research was to elucidate the performance of genomic selection depending on the age of new mutations, also characterizing the evolution of other relevant parameters such as linkage disequilibrium or allele frequencies.

Material and methods

Simulation procedure. 5,000 data sets were simulated following in part Meuwissen's et al. (2001) design. Each population started and evolved during 1,000 non-overlapping generations with an effective population size of $N_e = 100$, expanding to 5,000 individuals in generation 1,001. In all simulations, the genome consisted of a 50 cM chromosome with 1,001 SNP (one SNP each 0.05 cM) and a unique QTL located in cM 0. All SNP were homozygous in the founder generation, although a mutation rate of 2.5×10^{-5} per SNP was applied in the following generations, where mutations switched the allele state from 1 to 2 or vice versa. New alleles for the QTL marker were originated under a mutation rate of 5×10^{-4} per zygote between generations 1 and 990, whereas new mutations were prevented during the last 11 generations. Linkage disequilibrium between subsequent loci was generated on

[*]Grup de Recerca en Remugants, Departament de Ciència Animal i dels Aliments, Unitersitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

[†]Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, 50013 Zaragoza, Spain.

the basis of Kosambi's mapping function (Kosambi (1944)). In order to elucidate the statistical performance of genomic selection on the basis of the age of new mutations, only those populations where the QTL showed a biallelic conformation involving the founder allele in generation 1,001 were used for further analyses (25.7%). In addition, the generation where the other allele (i.e. mutant allele) arose by mutation was also stored. The founder QTL allele (*Q*) was assumed with null effect on phenotype whereas all mutant alleles (*q*) were assumed with an additive allelic effect of +0.5. A unique phenotypic record was simulated for each individual in generation 1,001 as the sum of the additive allelic effects from the QTL (genotype *QQ*: +0; genotypes *Qq* or *qQ*: +0.5; genotype *qq*: +1) plus a residual term sampled from a standard normal distribution with mean 0 and variance 1.

Each data set was analyzed by applying the methodology developed by Gianola et al. (2003) and within a Bayesian context. The Bayesian likelihood was defined as multivariate normal, i.e. $N(\boldsymbol{\mu}+\mathbf{M}\boldsymbol{\gamma},\mathbf{I}_e\sigma_e^2)$, where $\boldsymbol{\mu}$ was the 5,000×1 vector of population means, $\mathbf{M}$ was the 1,001×1,001 incidence matrix of known SNP genotypes, $\boldsymbol{\gamma}$ was the 1,001×1 vector of random SNP effects, $\mathbf{I}_e$ was an identity matrix with dimensions 5,000×5,000 and σ_e^2 was the residual variance. The a priori distribution for $\boldsymbol{\gamma}$ was assumed multivariate normal, i.e. $N(\mathbf{0},\mathbf{I}_\gamma\sigma_\gamma^2)$, and uniform a priori distributions were stated for $\boldsymbol{\mu}$, σ_e^2 and σ_γ^2. Note that $\mathbf{0}$ was a 1,001×1 vector of zeros, $\mathbf{I}_\gamma$ was and identity matrix with dimensions 1,001×1,001 and σ_γ^2 was the variance of SNP effects. A unique Monte Carlo Markov chain (MCMC) with 21,000 elements was launched for each analysis, discarding the first 1,000 iterations as burn-in (Raftery and Lewis (1992)). All unknown parameters in the model were updated by Gibbs sampling (Gelfand and Smith (1990)).

Calculations in the last generation. The mean square error (MSE) between real (i.e. simulated) and predicted genetic effects (i.e. QTL effects) was calculated by using average estimates of $\boldsymbol{\gamma}$ at the end of the MCMC sampling process. The number of polymorphic SNP markers was registered and their linkage disequilibrium with the biallelic QTL marker were calculated as $[(f_{QS}f_{qs}-f_{Qs}f_{qS})^2]/(f_Qf_qf_Sf_s)$, where *Q* and *q* were the alleles of the QTL marker, *S* and *s* were the alleles of the SNP marker, f_{QS} was the frequency of haplotype *QS* and f_Q was the frequency of allele *Q*.

Results and discussion

The QTL was biallelic and maintained the founder allele in almost 26% of the simulated populations. Within this subset of populations, the mutant allele had arose during the last 100 generations in 55.4% of the cases, and this percentage rose up to more than 70% when considering the last 200 generations (Table 1). Nevertheless, older mutations were provided by almost 30% of the simulated data sets, allowing to characterize MSE from genomic selection analyses at early generations.

Although the average number of polymorphic SNP markers did not vary with the age of the mutant QTL allele (~39 over 1,001 SNP, 3.9%; Table 1), average MSE drastically reduced from 0.358 ± 0.044 (mutant alleles from generations 1 to 299) to 0.021 ± 0.001 (generations 980 to 989). MSE estimates showed a consistent negative trend when the age of new

mutations decreased (Table 1), suggesting that the performance of genomic selection increased for young mutations and reduced for the older ones. This suggested that variability from new mutations must be mostly captured by genomic selection models, whereas a substantial fraction of genetic variance coming from old mutations could be ignored without further notice, reducing the potential response to selection. Note that this impaired ability to account for old mutations becomes a clear disadvantage for genomic selection when comparing with standard BLUP models.

Table 1: Summary of the statistical performance of simulated data sets under genomic selection analyses

Generation$^{\alpha}$	n$^{\beta}$	SNP$^{\chi}$ ± s.e.	MSE$^{\delta}$ ± s.e.
0 to 299	20	39.5 ± 4.6	0.358^{a} ± 0.044
300 to 499	61	39.5 ± 2.5	0.301^{a,b} ± 0.025
500 to 699	129	39.5 ± 1.8	0.234^{b} ± 0.013
700 to 800	163	39.3 ± 1.6	0.276^{a} ± 0.014
800 to 849	81	39.1 ± 2.1	0.233^{b} ± 0.018
850 to 899	118	39.2 ± 1.8	0.140^{c} ± 0.010
900 to 919	78	39.1 ± 2.2	0.118^{c,d} ± 0.010
920 to 939	115	39.0 ± 1.8	0.092^{d,e} ± 0.006
940 to 959	92	38.9 ± 2.0	0.078^{e} ± 0.006
960 to 989	224	38.9 ± 1.2	0.039^{f} ± 0.002
980 to 989	202	40.0 ± 1.4	0.021^{g} ± 0.001

$^{\alpha}$Generation where the mutant allele arose; $^{\beta}$Number of simulated data sets; $^{\chi}$Average number (± s.e.) of polymorphic single nucleotide polymorphisms; $^{\delta}$Average mean square error (± s.e.) between simulated and predicted genetic effects.

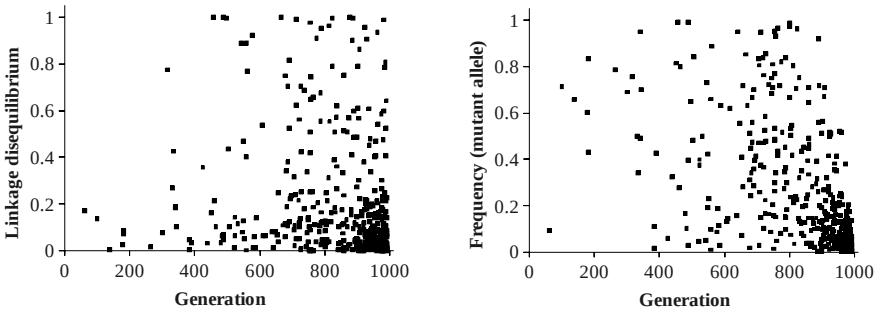

Figure 1: Distribution of maximum linkage disequilibrium between the QTL and a SNP marker locus (left) and frequency of the mutant QTL allele (right) depending on the generation where the mutant QTL allele arose.

As expected (Bink and Meuwissen (2004); de Roos et al. (2008)), linkage disequilibrium increased for younger mutations (Figure 1), optimizing the information content of SNP markers about their neighbor QTL. On the other hand, the average frequency of the mutant

allele decreased for younger mutations (Figure 1), although this evidence must not be considered more than an anticipateable consequence of the limited time for new mutations to spread in the simulated population, and the lack of selection.

Conclusion

These results clearly show that the statistical performance of genomic selection procedures are highly-dependent on the age of QTL mutations and, by extension, on the linkage disequilibrium between the QTL and neighbor SNP markers. Far from invalidating current genomic selection studies, our results must be viewed as a warning about the relevant amount of genetic variability originated by old mutations that cannot be properly accounted for in current genomic selection models. Further studies and statistical developments are required to recover this source of additive genetic variance for animal breeding purposes.

References

Bink, M. C. A. M., and T. H. E. Meuwissen. (2004). *Euphytica*, 137:95-99.

Casellas, J., Caja, G., and Piedrafita, J. (2010). *J. Anim. Sci.* (in press).

Casellas, J., and Medrano, J. F. (2008). *Genetics*, 179:2147-2155.

de Roos, A. P., Hayes, B. J., Spelman, R. J., and Goddard, M. E. (2008). *Genetics*, 179:1503-1512.

Gelfand, A., and Smith, A. F. M. (1990). *J. Am. Stat. Assoc.*, 85:398-409.

Gianola, D., Perez-Enciso, M., and Toro, M. A. (2003). *Genetics*, 163:347-365.

Hayes, B., and Goddard, M. E. (2001). *Genet. Sel. Evol.*, 33:209-229.

Kosambi, D. D. (1944). *Ann. Eugen.*, 12:172-175.

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

Raftery, A. E., and Lewis, S. M. (1992). In *Bayesian Statistics IV*, Oxford University Press, Oxford, UK, pages 949-953.