Accuracy Of Genome-Wide Evaluation For Disease Resistance In Aquaculture Breeding Programmes

B. Villanueva [*], J. Fernández[*], L.A. García-Cortés[*], L. Varona [†], H.D. Daetwyler[‡] and M.A. Toro[§]

Introduction

Disease resistance is an important breeding objective in aquaculture breeding programmes. This trait presents a discrete (dichotomous) distribution of phenotypes (diseased or non-diseased) and is difficult to improve by traditional methods. Recording is performed in controlled challenge tests and, given that tested fish can not be used as breeding candidates, selection is based on sib performance and applied only between families.

In contrast with traditional genetic evaluation in sib based aquaculture breeding schemes, genome-wide evaluation (GWE) allows the use of both between and within family variation and leads thus to higher accuracy and response to selection (Sonesson and Meuwissen 2009). Nielsen *et al*. (2009) and Sonesson and Meuwissen (2009) showed benefits of GWE as high as 30% for these schemes but considered only continuous traits. The objectives of this study were i) to extend the BayesB method of GWE to include dichotomous traits; and ii) to quantify, through computer simulation, the accuracy of GWE for disease resistance in aquaculture breeding programmes, using the methodology developed.

Methods

Extension of BayesB method to include dichotomous traits. For the analysis of dichotomous disease resistance traits, we follow the threshold liability model which assumes an underlying variable (liability) with a continuous distribution. The continuous variability results in a binary response which depends on whether the value of liability exceeds or not a fixed threshold. A BayesB model (Meuwissen *et al*. 2001) was assumed for the liabilities.

The conventional Bayesian model for dichotomous records can be expressed as $f(\mu, \mathbf{g}, \mathbf{v}, \mathbf{k} \mid \mathbf{y})$, that is, the posterior distribution of the mean μ, the SNP effects (**g**), the SNP variances (**v**) and the vector indicating which SNP have zero effect (**k**) can be obtained after the dichotomous records (**y**). The threshold t is considered fixed to a value that depends on the prevalence of the disease (q) and the residual variance is fixed to 1 (Sorensen *et al*. 1995). The Bayesian inference can be greatly simplified by augmenting the conventional

[*] INIA, Departamento de Mejora Genética Animal, Carretera de La Coruña km 7,5, 28040 Madrid, Spain
[†] Universidad de Zaragoza, Facultad de Veterinaria, , Miguel Servet 177, 50013 Zaragoza, Spain
[‡] Department of Primary Industries, 1 Park Drive, Bundoora, Victoria 3083, Australia
[§] Universidad Politécnica de Madrid, ETS Ingenieros Agrónomos, Ciudad Universitaria, 28040 Madrid, Spain

dichotomous model to include the unobservable underlying liabilities, **u** (Albert and Chib, 1993). The augmented model is $f(\mathbf{u}, \mu, \mathbf{g}, \mathbf{v}, \mathbf{k} \mid \mathbf{y})$ and can be usually implemented easily after a hierarchical model such as $f(\mathbf{u}, \mu, \mathbf{g}, \mathbf{v}, \mathbf{k} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \mathbf{u})\ f(\mathbf{u} \mid \mu, \mathbf{g}, \mathbf{v}, \mathbf{k})\ f(\mu, \mathbf{g}, \mathbf{v}, \mathbf{k})$. Note that $f(\mathbf{y} \mid \mathbf{u}) = f(\mathbf{y} \mid \mathbf{u}, \mu, \mathbf{g}, \mathbf{v}, \mathbf{k})$, the likelihood, is an indicator function taking only two values (0 and 1) as described in Sorensen et al. (1995). The model implemented for a particular liability i, $f(u_i \mid \mu, \mathbf{g}, \mathbf{v}, \mathbf{k}) = f(u_i \mid \mu, \mathbf{g})$, is a conventional univariate normal distribution, $f(u_i \mid \mu, \mathbf{g}) \propto \exp\{-0.5(u_i - \mu - \mathbf{x}_i\mathbf{g})'(u_i - \mu - \mathbf{x}_i\mathbf{g})\}$, which is transformed by the indicator function $f(\mathbf{y} \mid \mathbf{u})$ in a truncated normal distribution, defined within the interval $(-\infty,t)$ when $y_i = 0$ and within the interval (t,∞) when $y_i = 1$. Finally, $f(\mu, \mathbf{g}, \mathbf{v}, \mathbf{k})$ is a prior distribution as defined by Meuwissen *et al*. (2001) for the BayesB method. The model was implemented with a MCMC scheme as in Meuwissen *et al*. (2001) but with a data augmentation step from truncated Gaussian distributions for the liabilities.

Population and genetic model. A population at mutation-drift equilibrium was generated by simulating 5,000 generations of random mating with mutation and recombination and constant effective population size (N_e = 100). The genome consisted of 10 chromosomes of 1 Morgan each. At generation 0, the number of SNP and QTL (n_q) per chromosome were 900 and 100, respectively. After the 5,000 generations, the population size was increased to 6,000 individuals. The expansion of the population from 100 to 6,000 individuals was done in one single generation (schemes A) or over three extra generations (schemes B). In schemes A, the 50 males and the 50 females were mated at random (each male with a single female and each female with a single male) to produce 50 full sib families of 120 offspring each. Half of the offspring (60 individuals) from each family were assigned to the training set that was genotyped and measured for disease resistance in a challenge test, and half to the testing set that was only genotyped. In schemes B, three generations were run to expand the population. Sires and dams were randomly sampled with replacement to produce 500, 2,500 and 6,000 offspring at generations 5,001, 5,002 and 5,003, respectively. From the 6,000 individuals at the last generation, a random sample of 3,000 individuals constituted the training set and the other 3,000 individuals constituted the testing set. After discarding loci with MAF < 0.05, the number of SNP and QTL still segregating in the last generation was about 8,000 and 30, respectively. Additive effects for the QTL were sampled from a N(0,1). Phenotypes in the underlying scale for individuals in the training set were generated by adding to the genotypic values a random environmental deviation normally sampled from a N(0,V_e), and V_e was chosen such as the underlying heritability (h^2) was 0.1 or 0.3. Individuals with phenotypes in the underlying scale exceededing the threshold were assumed to be affected by the disease. The thresholds chosen corresponded to q = 0.1 or 0.5.

Genetic evaluation models. Two GWE were performed on the dichotomous data: i) GWE using the threshold model described above on dichotomous phenotypes; and ii) GWE using a linear model on dichotomous phenotypes. In addition, and for comparison purposes, GWE was performed on continuous phenotypes (liabilities) using a linear model. In most simulations, SNP genotypes were used to estimate associations with phenotypes as described above. However, some simulations were carried assuming that QTL were known, genotyped and used in the evaluation instead of the SNP. These scenarios provide an upper limit for the accuracy of genomic EBV. For the threshold model the number of cycles run was 50,000 and the first 5,000 were discarded. For the linear models, 10,000 cycles were run and the first

1,000 were discarded. Models were compared in terms of accuracy of evaluation (ρ), defined as the correlation between true and estimated breeding values in the underlying scale. Each scenario was replicated 25 times and results presented are averages over all replicates.

Results and discussion

Table 1 shows ρ for schemes A from the different GWE performed. For continuous traits, results agree with those obtained by Nielsen et al. (2009) and Sonesson and Meuwissen (2009) who, for similar schemes, found a ρ of around 0.7 for a trait with h^2 = 0.4 using genomic BLUP. However, there was a clear loss in ρ for dichotomous traits when compared with continuous traits, particularly with low h^2 and low q. This loss was higher in the testing set than in the training set and ranged from 9 to 27% across scenarios. The value of ρ in the testing set was from 2 to 5% lower than in the training set across scenarios. The lowest loss in ρ in testing individuals was for the linear model using continuous phenotypes and the highest was for the linear model using dichotomous phenotypes.

Table 1: Accuracy of GWE in training and testing sets for schemes A with different disease prevalences (q) and heritabilities (h^2), using linear or threshold models with continuous or dichotomous phenotypes, and SNP or QTL genotypes in the evaluation

			h^2 = 0.1		h^2 = 0.3	
q	Model	Phenotypes	Training	Testing	Training	Testing
			Using SNP in evaluation			
	Linear	Continuous	0.674	0.648	0.823	0.794
0.1	Linear	Dichotomous	0.429	0.395	0.620	0.584
0.1	Threshold	Dichotomous	0.485	0.453	0.672	0.643
0.5	Linear	Dichotomous	0.527	0.489	0.739	0.703
0.5	Threshold	Dichotomous	0.554	0.519	0.758	0.725
			Using QTL in evaluation			
	Linear	Continuous	0.899	0.897	0.923	0.922
0.1	Linear	Dichotomous	0.724	0.724	0.755	0.755
0.1	Threshold	Dichotomous	0.750	0.750	0.789	0.788
0.5	Linear	Dichotomous	0.824	0.821	0.807	0.806
0.5	Threshold	Dichotomous	0.839	0.836	0.832	0.830

The advantage of using the threshold model for GWE was evident when the phenotype is of discrete nature. For dichotomous phenotypes, ρ was from 3 to 16% higher with the threshold model than with the linear model. The highest gain was for the lowest h^2 and the lowest q and for the testing set.

When QTL genotypes were used in the evaluation, ρ was substantially higher than when using SNP and it was practically the same in training and testing sets. The difference in ρ using SNP or QTL was highest for dichotomous phenotypes, the testing set, the lowest q, and

particularly, for the lowest h^2. For h^2 = 0.1, the loss in ρ when using SNP was as high as 35% for the dichotomous trait analysed with the threshold model, indicating that the SNP markers only captured a proportion of the total genetic variation.

Results presented above correspond to schemes where families need to be kept separately until individual tagging as training and testing individuals are full sibs. Schemes B would avoid this requirement but at the expense of ρ which, for the same marker density and number of phenotypes, is decreased when relationships are more distant (Table 2).

Overall, genomic ρ for dichotomous phenotypes was high even when a linear model was used in the analysis. For instance, for $q = h^2 = 0.1$, ρ was more than double the ρ expected from mass selection (0.17). When data was correctly modelled by using the threshold model, ρ was even higher.

Table 2: Accuracy of GWE in training and testing sets for schemes B with different disease prevalences (q) and heritabilities (h^2), using linear or threshold models with dichotomous phenotypes, and SNP genotypes in the evaluation

			$h^2 = 0.1$		$h^2 = 0.3$	
q	Model	Phenotypes	Training	Testing	Training	Testing
0.1	Linear	Dichotomous	0.371	0.331	0.573	0.523
	Threshold	Dichotomous	0.383	0.338	0.587	0.531
0.5	Linear	Dichotomous	0.444	0.398	0.666	0.603
	Threshold	Dichotomous	0.479	0.431	0.690	0.630

Conclusion

This study shows that the threshold model fits very well with the BayesB model of GWE and that the advantage of using the former is clear when dealing with dichotomous traits, particularly under the conditions where linear models are less appropriate for analysing this type of traits (i.e., low h^2 and low q). Although the method is illustrated in an aquaculture context, its application is general and it could be easily extended to include more categories.

Acknowledgements

This work was funded by the Ministerio de Ciencia e Innovación (CGL2009-13278-C02-02).

References

Albert, J.H. and Chib, S. (1993). *J. Am. Stat. Assoc*., 88:669-679.

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

Nielsen, H.M., Sonesson, A.K., Yazdi, H. et al. (2009). *Aquaculture*, 289:259-264.

Sonesson, A.K. and Meuwissen, T.H.E. (2009). *Genet. Sel. Evol.*, 41: 37.

Sorensen, D.A., Andersen, S., Gianola, D. et al. (1995). *Genet. Sel. Evol*., 27:229-249.