

JOURNAL OF ANIMAL SCIENCE

The Premier Journal and Leading Source of New Knowledge and Perspective in Animal Science

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

J ANIM SCI published online July 8, 2011

The online version of this article, along with updated information and services, is located on the World Wide Web at:

<http://jas.fass.org/content/early/2011/07/08/jas.2010-3814>



American Society of Animal Science

www.asas.org

**Accuracy of genome-wide evaluation for disease resistance in aquaculture
breeding programmes¹**

**B. Villanueva^{*2}, J. Fernández^{*}, L.A. García-Cortés^{*}, L. Varona[†], H.D. Daetwyler[‡] and
M.A. Toro[§]**

^{*} Departamento de Mejora Genética Animal, INIA (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria), Carretera de La Coruña km 7,5, 28040 Madrid, Spain.

[†] Unidad de Genética Cuantitativa y Mejora Animal, Facultad de Veterinaria, Universidad de Zaragoza, Miguel Servet 177, 50013 Zaragoza, Spain.

[‡] Biosciences Research Division, Department of Primary Industries, 1 Park Drive, Bundoora, 3083, Victoria, Australia.

[§] Departamento de Producción Animal, ETS Ingenieros Agrónomos, Universidad Politécnica de Madrid, Ciudad Universitaria, 28040 Madrid, Spain

¹ This work was funded by the Ministerio de Ciencia e Innovación (CGL2009-13278-C02-02). We thank two anonymous reviewers for their constructive comments.

² Corresponding author: villanueva.beatriz@inia.es

ABSTRACT

Current aquaculture breeding programs aimed at improving resistance to diseases are based on challenge tests, where performance is recorded on sibs of candidates to selection, and on selection between families. Genome-wide evaluation (GWE) of breeding values offers new opportunities for using variation within families when dealing with such traits. However, up to date studies on GWE in aquaculture programs have only considered continuous traits. The objectives of this study were i) to extend GWE methodology, in particular the Bayes B method, to analyze dichotomous traits such as resistance to disease; and ii) to quantify, through computer simulation, the accuracy of GWE for disease resistance in aquaculture sib based programs, using the methodology developed. Two heritabilities (0.1 and 0.3) and two disease prevalences (0.1 and 0.5) were assumed in the simulations. We followed the threshold liability model which assumes that there is an underlying variable (liability) with a continuous distribution and assumed a BayesB model for the liabilities. It was shown that the threshold liability model used fits very well with the BayesB model of GWE. The advantage of using the threshold model was clear when dealing with disease resistance dichotomous phenotypes, particularly under the conditions where linear models are less appropriate (low heritability and disease prevalence). In the testing set (where individuals are genotyped but not measured), the increase in accuracy for the simulated schemes when using the threshold model ranged from 4 (for heritability equal to 0.3 and prevalence equal to 0.5) to 16% (for heritability and prevalence equal to 0.1) when compared with the linear model.

Key Words: accuracy prediction, aquaculture, BayesB, disease resistance, genome-wide evaluation, threshold model

INTRODUCTION

It is now widely recognised that the accuracy of EBV from genome wide evaluation (GWE) can be substantially higher than that of traditional BLUP evaluation, and that selection based on genomic EBV has a high potential, particularly for improving traits difficult to target by traditional selection. This is the case with disease resistance, an important selection objective in aquaculture breeding programmes. Disease resistance presents a discrete (dichotomous) distribution of phenotypes (diseased or non-diseased) and is difficult to improve. In current schemes, recording is usually performed in controlled challenge tests and tested fish can not be used as breeding candidates. Thus, selection is based on sib performance and applied only between families unless multi-trait genetic evaluations involving other traits genetically correlated with disease resistance (and measured in the candidates) are performed (Vehviläinen et al., 2010).

In contrast with traditional genetic evaluation using univariate models in sib based aquaculture breeding schemes, GWE directly allows the use of both between and within family variation and leads thus to higher accuracy and response to selection (Sonesson and Meuwissen, 2009). Under the GWE scheme, sibs of candidates would be genotyped and measured (i.e., they would constitute the training set) and the candidates would be only genotyped. Genomic EBV would then be obtained for the candidates and both within and between family variation could be used in selection. Using computer simulation, Nielsen et al. (2009) and Sonesson and Meuwissen (2009) showed benefits of GWE as high as 30% for these schemes. However, they only considered traits showing a continuous distribution of phenotypes. The objectives of this study were i) to extend the BayesB method of GWE (Meuwissen et al., 2001) to include dichotomous

traits; and ii) to quantify, through computer simulation, the accuracy of GWE for disease resistance in aquaculture breeding programs, using the methodology developed.

MATERIAL AND METHODS

BayesB is a Bayesian method that uses prior knowledge on variances of marker effects and proportions of markers with zero effect and it is currently one of the most used methods in GWE. When using BayesB, there are two steps involved in the estimation of breeding values (EBV) for animals without phenotypes. In the first step, the effects of the markers are estimated in a group of individuals (the training set) that are genotyped and measured for the trait of interest. In the second step, the estimates of marker effects are used to obtain EBV of animals that are genotyped but not measured (the testing set). Meuwissen et al. (2001) developed the method for continuous traits but to our knowledge, there have been no extensions of this method to obtain genomic predictions for traits with a dichotomous distribution of phenotypes.

Extension of BayesB method to include dichotomous traits

For the analysis of dichotomous disease resistance traits (diseased or non-diseased), we follow the threshold liability model of Wright (1934). This model assumes that there is an underlying variable (liability) which has a continuous distribution. The continuous variability however results in a binary response which depends on whether or not the liability exceeds a fixed threshold (t). We assume a BayesB model (Meuwissen et al., 2001) for the liabilities. Thus,

the liability for individual i (u_i) is $u_i = \mu + \sum_{j=1}^{n_m} x_{ij} g_j + e_i$, where μ is the mean, n_m is the number of SNP, x_{ij} are observable covariates, g_j is the additive value of the j th SNP, and e_i is the

environmental deviation for individual i . We fixed covariates at values $\sqrt{2}$, 0 and $-\sqrt{2}$ for individuals homozygous for SNP allele 1, heterozygous or homozygous for allele 2, respectively.

We express the conventional Bayesian model for dichotomous records as $f(\mu, \mathbf{g}, \mathbf{v}, \mathbf{k} \mid \mathbf{y})$, that is, the posterior distribution of μ , the SNP effects (\mathbf{g}), the SNP variances (\mathbf{v}) and the vector indicating which SNP have zero effect (\mathbf{k}) can be obtained based on the dichotomous records (\mathbf{y}). Here, μ , \mathbf{g} , \mathbf{v} and \mathbf{k} refer to the underlying scale. The threshold t was fixed to 0 and the residual variance, σ_e^2 , was fixed to 1 as in Sorensen et al. (1995).

The Bayesian inference to analyze the threshold model was greatly simplified by Albert and Chib (1993) by using a data augmentation scheme (Tanner and Wong, 1987). They augmented the conventional dichotomous model by including the unobservable underlying liabilities, \mathbf{u} . The augmented model is $f(\mathbf{u}, \mu, \mathbf{g}, \mathbf{v}, \mathbf{k} \mid \mathbf{y})$ and, in genetics, was first applied by Sorensen et al. (1995) to analyze mixed models for categorical traits and by Yi and Xu (2000) for QTL mapping. The augmented model is usually implemented as 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 $f(\mathbf{y} \mid \mathbf{u}) = f(\mathbf{y} \mid \mathbf{u}, \mu, \mathbf{g}, \mathbf{v}, \mathbf{k})$ is an indicator function taking only two values (0 and 1) as described in Sorensen et al. (1995).

The model that we implemented for a particular liability i , $f(u_i \mid \mu, \mathbf{u}_{-i}, \mathbf{g}, \mathbf{v}, \mathbf{k}) = f(u_i \mid \mu, \mathbf{g})$, where \mathbf{u}_{-i} refers to “all u ’s except u_i ”, is a conventional univariate normal distribution, $f(u_i \mid \mu, \mathbf{g}) \propto \exp\left\{-0.5(u_i - \mu - \mathbf{x}_i' \mathbf{g})(u_i - \mu - \mathbf{x}_i' \mathbf{g})\right\}$, 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. All variables have a flat distribution except $f(g_i | v_i)$ for each individual, which are univariate centered normal distributions. An almost standard Gibbs sampler algorithm was implemented to draw successively for all variables in the model except v_i . Variances cannot be conditioned on g_i because this value will be null with a non negligible probability. It can be noticed that a null SNP value will provide a SNP null variance and *vice versa*. This problem is usually solved by using a composition sampling scheme, that is, by sampling v_i from $f(v_i | \mu, \mathbf{u}, \mathbf{g}_{-i}, \mathbf{k})$, a distribution marginal with respect to the variable g_i , where \mathbf{g}_{-i} is the conventional notation for “all g ’s except g_i ”. In this case, ignoring the prior part of the equation, the conditional for v_i is

$$f(v_i | \mu, \mathbf{u}, \mathbf{g}_{-i}) \propto |\mathbf{V}_i|^{-0.5} \exp\{-0.5 \mathbf{u}^* \mathbf{V}_i^{-1} \mathbf{u}^*\}, \quad [1]$$

where $\mathbf{V}_i = \mathbf{x}_i \mathbf{x}_i' v_i + \mathbf{I} \sigma_e^2$ and $\mathbf{u}^* = \mathbf{u} - \mathbf{1} \mu - \sum_{j \neq i} \mathbf{x}_j' \mathbf{g}_j$.

Formula [1] does not have any known closed form and we used a Metropolis-Hastings algorithm for drawing v_i from [1]. Also, we used the prior distribution as the proposal distribution. For comparing densities, only the evaluation of [1] is required because the proposal density and the prior density cancel out. In order to perform evaluations of the log-density of the conditional [1] required by the Metropolis Hastings algorithm, we can exploit the simple structure of \mathbf{V} . After some algebra (see Appendix), the logarithm of conditional [1] simplifies to

$$\log[f(v_i | \mu, \mathbf{u}, \mathbf{g}_{-i})] \propto -0.5 \log(2\eta_i v_i + \sigma_e^2) + \frac{(w_{11} - w_{22})^2}{\sigma_e^2 (2\eta_i + \sigma_e^2 / v_i)}$$

where w_{11} and w_{22} are the sums of the pre-adjusted liabilities \mathbf{u}^* for the corresponding homozygote individuals and η_i is the number of homozygous individuals for the i th SNP in the training set.

Population and genetic models

An initial population (generation 0) of 50 males and 50 females was generated. The genome consisted of 10 chromosomes of 1 Morgan each. The numbers of SNP and QTL affecting the trait per chromosome simulated at generation 0 were 900 and 100, respectively. All loci (SNP and QTL) were biallelic and homozygous for the same allele at generation 0. QTL were additive. In order to generate a population at mutation-drift equilibrium, 5,000 generations of random mating were simulated. The effective population size (N_e) was kept constant and equal to the initial size (100) across generations by sampling sires and dams with replacement. When generating gametes from the parents, we assumed the Haldane's model (e.g., Lynch and Walsh, 1998). Thus, the number of recombinations per chromosome was sampled from a Poisson distribution with mean equal to 1 given that chromosome length was assumed to be 1 Morgan. Recombinations occurred between homologues in random positions of a particular chromosome without interference. Mutation rates in the gametes at the marker loci and at the QTL were 2.5×10^{-3} and 2.5×10^{-5} , respectively (Meuwissen et al., 2001). The number of mutations per generation was sampled from a Poisson distribution with mean equal to $2N_e n_c n_l m$, where n_c is the number of chromosomes, n_l is the number of loci (markers or QTL) per chromosome and m is the corresponding mutation rate. Mutations were then randomly distributed across individuals, chromosomes and loci and they switched allele 1 to 2 and *vice versa*.

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 order to investigate how the family structure affects accuracy. A regular family structure of full sibs was simulated in Schemes A where only two degrees of relationships with respect to generation 5,000 (full sibs or ‘unrelated’) were represented in the training and testing sets. A much wider range of degrees of relationships were simulated in training and testing sets for Schemes B by simulating the extra generations of random mating. In schemes A, the 50 males and the 50 females were mated at random (matings were monogamous) to produce 50 full sib families of 120 offspring each in generation 5001. Half of the offspring (60 individuals) from each family were assigned to the training set that was genotyped and measured for the trait of interest, 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 and the offspring produced was 500, 2,500 and 6,000 offspring at generations 5001, 5002 and 5003, 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 minor allele frequency (MAF) less than 0.05, the number of SNP and QTL segregating in the last generation (generation 5,001 in schemes A or generation 5,003 in schemes B) was about 8,000 and 30, respectively.

Genotypic values due to the i th QTL were a_i , 0 and $-a_i$ for individuals homozygous for the favourable allele, heterozygous or homozygous for the unfavourable allele, respectively (Falconer and Mackay, 1996). Values for a_i (defined as half the difference between the two homozygotes) were sampled from a standard normal distribution. Genotypic values were calculated for each individual in both the training and the testing sets by accumulating values

over all QTL. Phenotypes in the underlying scale for individuals in the training set were generated by adding a random environmental deviation normally distributed with mean zero and variance V_e , and V_e was chosen such as the underlying heritability (h^2) was 0.1 or 0.3. Individuals whose phenotypic value in the underlying scale exceeded the threshold were assumed to be affected by the disease. Thresholds were chosen to achieve a prevalence of $q = 0.1$ or 0.5. Robertson and Lerner (1949) show that the relationship between h^2 and the observed heritability (h_o^2) is expressed as $h_o^2 \approx h^2 q^2 i_q^2 [q(1-q)]^{-1}$, where i_q is the mean liability of diseased individuals. Using this expression, the values of 0.1 and 0.3 for h^2 correspond, respectively, to $h_o^2 = 0.06$ and $h_o^2 = 0.19$ for $q = 0.5$, and to $h_o^2 = 0.03$ and $h_o^2 = 0.10$ for $q = 0.1$.

Genetic evaluation models

Two different GWE were performed on the simulated dichotomous data: i) GWE using the threshold model described above; and ii) GWE using a linear model. In addition, and for comparison purposes, GWE was performed on continuous phenotypes (i.e., we treated the liabilities as observed phenotypes) using a linear model. 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, defined as the correlation between true and estimated breeding values on the underlying scale. The true breeding value for individual i was computed as $TBV_i = \sum_{j=1}^{n_q} \beta_{ij}$, where n_q is the number of QTL affecting the trait, β_{ij} is the breeding value of individual i due to QTL j which equals to $2(1-p_j)\alpha$, $(1-2p_j)\alpha$ and $-2p_j\alpha$ if the individual is homozygous for the favourable allele, heterozygous or homozygous for

the unfavourable allele, respectively, α is the average effect of the gene substitution and p_j is the frequency of the favourable allele of QTL j (Falconer and Mackay, 1996). The estimated breeding value for individual i was $EBV_i = \sum_{j=1}^{n_m} x_{ij} \delta_{ij}$, where n_m and x_{ij} are defined as above and δ_{ij} is the estimated effect of SNP j . Each scenario was replicated 25 times and results presented are averages over all replicates.

In addition to simulations where SNP genotypes were used to estimate associations with phenotypes (as described above), we also carried out simulations assuming that QTL were known, genotyped and used in the evaluation instead of the SNP (i.e., only QTL genotypes were used in the evaluation). These scenarios provide an upper limit for the accuracy of genomic EBV when effects need to be estimated.

Comparison of simulated and predicted accuracy

The accuracy of GWE obtained from the simulations was compared with that predicted from equations developed by Daetwyler et al. (2008, 2010). For a continuous phenotype, Daetwyler et al. (2008) predicted the accuracy of GWE as

$$r_{g\hat{g}} = \sqrt{\frac{n_p h^2}{n_p h^2 + n_G}} \quad [2]$$

where h^2 is the heritability, n_p is the number of phenotypes recorded and n_G is the number of marker loci, some of them associated with QTL. Many of these loci may have zero effect and they are assumed to be independently segregating. For a dichotomous disease phenotype, Daetwyler et al. (2008) showed that accuracy can be also predicted from [2] provided the heritability used is that for the observed dichotomous scale.

The predictions of Daetwyler et al. (2008) were derived for a simple least-squares genome-wide evaluation approach where the effect of each allele is estimated by regressing phenotypes on the genotypes, one locus at a time because the loci were assumed to be independent. However, for continuous phenotypes, Daetwyler et al. (2010) have recently extended the original formulae for predicting the accuracy of BayesB. This is obtained by replacing, in [2], n_G with the number of QTL affecting the trait if n_q is smaller than the number of independent chromosome segments (M_e) or with M_e if $n_q \geq M_e$. The value for M_e can be obtained from $M_e = 2N_eL/\log(4N_eL)$, where L is the genome length in Morgans (Goddard, 2009). In our case, $n_q \approx 30$ and $M_e \approx 241$ and therefore, the accuracy of BayesB is predicted as

$$r_{gg} = \sqrt{\frac{n_p h^2}{n_p h^2 + n_q}} \quad [3]$$

We used this expression also for dichotomous phenotypes but in this case h^2 was replaced with h_o^2 .

RESULTS

Table 1 shows the accuracy for schemes A from the different GWE performed on the simulated data. As expected, accuracy was higher with the highest h^2 and the highest prevalence. There was a clear loss in accuracy for dichotomous traits when compared with continuous traits, particularly with low h^2 and low prevalence. This loss was higher in the testing set than in the training set and ranged approximately from 9 to 27% across scenarios when results from continuous phenotypes were compared with results for dichotomous phenotypes analysed with the threshold model.

The accuracy in the testing set was from 2 to 5% lower than in the training set across scenarios, and again, the loss in accuracy for the individuals only genotyped was higher with low h^2 and low prevalence. The lowest loss was for the linear model using continuous phenotypes and the highest was for the linear model using dichotomous phenotypes.

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

When QTL genotypes were used in the evaluation, the accuracy was substantially higher than when using SNP and it was practically the same in training and testing sets. The increase in accuracy observed when using QTL instead of SNP was highest for dichotomous phenotypes, the testing set, the lowest prevalence, and particularly, for the lowest h^2 . For $h^2 = 0.1$, the accuracy obtained for the dichotomous trait analysed with the threshold model when using QTL was 35% higher than that obtained when using SNP.

Results for schemes B are presented in Table 2. Comparisons for different h^2 , prevalences and data sets (training versus testing) followed the same trends as in schemes A but the accuracy of GWE in schemes B was decreased when compared to that in schemes A. For dichotomous phenotypes analysed with the threshold model, accuracy in schemes B was from 10 to 27% lower than in schemes A for $h^2 = 0.1$ and from 8 to 18% lower for $h^2 = 0.3$ (Tables 1 and 2). However, the opposite was observed in the scenarios where QTL genotypes were assumed to be known and used directly in the evaluations. In these scenarios, schemes B consistently outperformed schemes A. For instance, for dichotomous phenotypes analysed with the threshold model, accuracy in schemes B was from 7 to 13% higher than in schemes A. As with schemes A, the difference in accuracy between training and testing sets when using QTL was minimal.

A comparison of simulated and predicted accuracy for different models and phenotype types is shown in Table 3. Predictions based on [3] greatly overestimate the accuracy of GWE obtained in the simulations using SNP genotypes in all scenarios. The percentage prediction

error (i.e., $100(P - S1)/P$, where P is the predicted accuracy and $S1$ is the accuracy from simulation) was highest for dichotomous phenotypes, for the lowest h^2 and for the lowest prevalence. For $h^2 = 0.1$ and $q = 0.1$, the percentage prediction error was as high as 45%. This error decreased drastically when QTL genotypes were used in the evaluation instead of the SNP. In this situation (using QTL), the percentage error ranged from 8 to 9% for the continuous trait and from 14 to 21% for the dichotomous trait.

DISCUSSION

In this study we have combined the BayesB method of Meuwissen et al. (2001), described for continuous traits with the Bayesian threshold model of Sorensen et al. (1995) in order to obtain genomic predictions of breeding values for traits that show a discrete distribution of phenotypes. We have also applied this methodology to predict genomic accuracy for fish sib-based selection scheme designed to improve disease resistance. We show that the threshold model fit very well with the BayesB method and leads to improved accuracies of GWE when dealing with disease resistance dichotomous phenotypes, compared with accuracies obtained with the linear model. For the schemes simulated, the increase in accuracy obtained when using the threshold model was up to 16%. The highest advantage of the threshold model was under the conditions where the linear models are less appropriate; i.e., low heritability, low disease prevalence and individuals only genotyped but not measured (testing set).

Most studies on GWE have dealt with continuous traits although very recently González-Recio and Forni (2011) have developed versions of different methods (BayesA, Bayesian LASSO and two machine learning methods) for analyzing dichotomous traits. They have showed

that the differences between methods are small, particularly with a large number of QTL. The proposed methodology presented here considers another method, the BayesB method, which is widely used in GWE and thus represents an advance in the prediction of genomic EBV. It could be easily extended for traits with more than two discrete categories. For instance, in fish breeding programs, polychotomous traits considered in breeding goals can include flesh quality traits such as visual fat content and fillet gaping score. These traits are also improved through sib selection as recording on candidates is not possible. For polychotomous traits, one of the thresholds and the residual variance would be fixed and the other thresholds would be estimated following Sorensen et al. (1995). It is worth to mention that when dealing with threshold models, there is a potential advantage of GWE with respect to standard animal models. With GWE, SNP effects are estimated with a large number of records distributed across both discrete categories, avoiding thus the possibility that all records fall in one category. These extreme category problems are a main limitation of standard threshold models. In addition, another important limitation of the use of the animal model is the unfeasibility of disentangling between Mendelian sampling terms and residuals when there is only one phenotypic record for each individual. This fact may cause a severe estimation bias when the genetic variance is estimated based on all breeding values rather than based on parents only (Ødegård et al., 2010). In contrast, GWE provides information for each Mendelian sampling term as it uses individual genotyping and, thus, it is able to separate residual variation and Mendelian sampling in the liability.

The methodology described here would be also useful to analyse dichotomous data from human diseases. Although recent successes of genome-wide association studies (GWAS) with dense SNP chips have lead to the identification of a large number of loci that are significantly associated with disease traits (Hirschhorn and Daly, 2005; Wellcome Trust Case Control

Consortium, 2007), the proportion of the genetic variance explained by these loci remains small for any one trait. Recently, Yang et al. (2010) have shown that most “missing heritability” (the unexplained variability) is not missing but has not been previously detected because the individual effects are too small to be statistically significant when using the stringent thresholds of statistical significance applied in GWAS. Similarly, when predicting the genetic component of disease risk (the equivalent to breeding value in the animal breeding context), the use of large numbers of SNP simultaneously and the avoidance of significance testing through the application of GWE methodology could have benefits over the current approach focused on specific genes and based only on pedigree analysis (Daetwyler et al., 2008).

In fish breeding programs disease resistance is usually recorded as 0 or 1, depending on whether the fish is still alive or not at the end of the challenge test. Often the test is terminated before the survival curve has descended all the way to zero or has flattened to a level plateau (Moen et al., 2007). Therefore, survival analysis would be more suitable for analyzing challenge test data because it makes the most efficient use of all the information available (from dead animals and from animals still alive at the end of the test) rather than describing survival at a given point in time arbitrarily defined (e.g., Ducrocq et al., 2000). In the context of QTL detection in salmon, Moen et al. (2007) found that models based on the dichotomous trait had higher statistical power than survival models due, in part, to the large number of effects to be estimated with the latter. Further developments are however needed to incorporate survival analysis in the context of GWE. Approximate survival analyses could be carried out through survival score models (Veerkamp et al., 2001) where the survival time is measured through sequential binary scores that can be fitted with a threshold model. Our model should be easily

adapted to this type of models and even to a cure survival model where some animals are not susceptible at all (Ødegård et al., 2011).

For continuous traits, our results are consistent with those obtained by Nielsen et al. (2009) and Sonesson and Meuwissen (2009) who simulated family-based aquaculture schemes where the test population consists of sibs of candidates. For scenarios similar to schemes A (with training and testing sizes of 3000 individuals) but with a larger number of families (100 rather than 50), they found accuracies of around 0.7 for a trait with $h^2 = 0.4$, using genomic BLUP. The higher accuracy (about 0.8 for $h^2 = 0.3$; Table 1) found here is probably due to the evaluation method used (i.e., BayesB). We show that there was a clear loss in accuracy for dichotomous traits when compared with continuous traits, particularly with low h^2 and low prevalence. However, overall, genomic accuracy for dichotomous phenotypes was high even when a linear model was used in the analysis (from 0.41 to 0.70 in testing sets of Schemes A). When data were correctly modeled by using the threshold model, the accuracy was even higher.

Under schemes A, individuals simulated are full sibs and therefore they would share longer chromosome segments than individuals in Schemes B which have a more diverse set of relationships (Hayes et al., 2009). Also, some disequilibrium between SNP and QTL would have broken down when modeling the extra generations in schemes B. As a consequence, the accuracy of evaluation using SNP was consistently lower in schemes B than in schemes A. However, when QTL genotypes were used directly in the evaluation, the accuracy in schemes B was consistently higher than in schemes A. Here, the linkage disequilibrium between SNP and QTL is no longer an issue and schemes B benefit from having more independent data that permit a more precise estimation of QTL effects.

Schemes A produce higher accuracies of GWE but implicitly assume that families are kept separate until individual tagging is possible. Parental assignment using the SNP would permit communal rearing of families but would be very costly, particularly when contributions vary widely across families (Gjerde et al., 2002; Chatziplis et al., 2007). These schemes would thus imply costly multi-tank facilities with the consequent problem of having to estimate and correct for common environmental (tank) effects to avoid having these effects confounded with genetic effects. Schemes B would avoid this requirement but at the expense of accuracy.

Predictions available for accuracy of the BayesB method (Daetwyler et al. 2008, 2010) seem to overestimate at a large extent the accuracy obtained in the simulations, particularly for dichotomous phenotypes, low h^2 and low prevalence indicating that the SNP markers only captured a proportion of the total genetic variation (Daetwyler, 2009). This was confirmed by the high increase in accuracy observed when using QTL genotypes directly in the evaluation instead of the SNP. In this situation, the percentage prediction error decrease drastically. In the scenarios using QTL instead of SNP in the evaluation there was practically no loss in accuracy when moving from the training to the testing sets. This is not surprising as QTL effects were evaluated with high accuracy and the existence of linkage disequilibrium between QTL and SNP is not required to obtain accurate estimates in the testing set.

The wide application of GWE that has occurred during the last years in terrestrial farm animals, particularly in dairy cattle breeding, came as a consequence of having genome-wide dense SNP chips available at a reasonable cost. For instance, the bovine 50K chip has become a standard tool for dairy cattle breeding industries, and a higher density chip with approximately 800K SNP is now also available. Development of SNP chips for aquaculture species is far behind that for terrestrial species but dense genetic maps containing thousand of markers are

starting to be available. In Atlantic salmon a chip with over 15K SNP has been developed by the Center for Integrative Genetics, in Norway and manufactured by Illumina (Kent et al., 2009; Dominik et al., 2010). We can expect that continuous developments in this area will make possible GWE in aquaculture in the near future. This will be of great benefit for improving disease resistance traits where aquaculture species have some advantages such as the possibility of performing experimental challenge tests that can avoid some biases in genetic parameter estimates implicit when using field data; e.g., lack of exposure to the agent causing the disease (Bishop and Woolliams, 2010). Such tests are now a standard approach for obtaining phenotypic information for a wide variety of diseases with the overall aim of improving resistance.

LITERATURE CITED

- Albert, J.H., and S. Chib. 1993. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88:669-679.
- Bishop, S.C., and J.A. Woolliams. 2010. On the genetic interpretation of disease data. *PLoS ONE* 5(1): e8940.
- Chatziplis, D., C. Batargias, C.S. Tsigenopoulos, A. Magoulas, S. Kollias, G. Kotoulas, F.A.M. Volckaert, and C.S. Haley. 2007. Mapping quantitative trait loci in European sea bass (*Dicentrarchus labrax*): The BASSMAP pilot study. *Aquaculture* 272S1:S172-S182.
- Daetwyler, H.D. 2009. Genome-wide evaluation of populations. PhD. Diss. Wageningen Univ., Wageningen, The Netherlands.
- Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* 3(10): e3395.
- Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. *Genetics* 185:1021-1031.
- Dominik, S., J.M. Henshall, P.D. Kube, H. King, S., Lien, M.P. Kent, and N.G. Elliott. 2010. Evaluation of an Atlantic salmon SNP chip as a genomic tool for the application in a Tasmanian Atlantic salmon (*Salmo salar*) breeding population. *Aquaculture* 308:S56–S61.
- Ducrocq, V., B. Besbes, and M. Protais. 2000. Genetic improvement of laying hens viability using survival analysis. *Genet. Sel. Evol.* 32:23-40.
- Falconer, D.S., and T.F.C. Mackay. 1996. *Introduction to quantitative genetics*, 4th ed., Longman.

- Gjerde, B., B.Villanueva, and H.B. Bentsen. 2002. Opportunities and challenges in designing sustainable fish breeding programs. Proc. 7th World Congr. Genet. Appl. Livest. Prod., Montpellier, France, CD-ROM communication no. 06-01.
- Goddard, M.E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136:245-257.
- González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet. Sel. Evol.* 43:7.
- Hayes, B.J., P.M. Visscher, and M.E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res., Camb.* 91:47–60.
- Hirschhorn, J.N., and M.J. Daly. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95-108.
- Kent, M.P., B. Hayes, Q. Xiang, P.R. Berg, R.A. Gibbs, and S. Lien. 2009. Development of 16.5 K SNP chip for Atlantic salmon. Proc. 17th Plant Anim. Genome Conf., San Diego, USA. <http://www.intl-pag.org/> Accessed Nov.19, 2010.
- Lynch, M., and B. Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. 1st ed. Sinauer Associates, Sunderland, MA.
- Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome wide dense marker maps. *Genetics* 157:1819–1829.
- Moen, T., A.K. Sonesson, B. Hayes, S. Lien, H. Munck, and T.H.E. Meuwissen. 2007. Mapping of a quantitative trait locus for resistance against infectious salmon anaemia in Atlantic salmon (*Salmo Salar*): comparing survival analysis with analysis on affected/resistant data. *BMC Genetics* 8:53. doi:10.1186/1471-2156-8-53.

- Nielsen, H.M, A.K. Sonesson, H. Yazdi, and T.H.E. Meuwissen. 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. *Aquaculture* 289:259-264.
- Ødegård, J., T.H.E. Meuwissen, B. Heringstad, and P. Madsen. 2010. A simple algorithm to estimate genetic variance in an animal threshold model using Bayesian inference. *Genet. Sel. Evol.* 42:29.
- Ødegård, J., T. Gitterle, P. Madsen, T.H.E. Meuwissen, M.H. Yazdi, B. Gjerde, C. Pulgarin, and M. Rye. 2011. Quantitative genetics of taura syndrome resistance in pacific white shrimp (*Penaeus vannamei*): a cure model approach. *Genet. Sel. Evol.* 43:14.
- Robertson, A., and I.M. Lerner. 1949. The heritability of all-or-none traits - viability of poultry. *Genetics* 34:395-411.
- Sonesson, A.K., and T.H.E. Meuwissen. 2009. Testing strategies for genomic selection in aquaculture breeding programs. *Genet. Sel. Evol.* 41: 37; doi:10.1186/1297-9686-41-37.
- Sorensen, D.A., S. Andersen, D. Gianola, and I. Korsgaard. 1995. Bayesian inference in threshold models using Gibbs sampling. *Genet. Sel. Evol.* 27:229-249.
- Tanner, M.A., and W.H. Wong. 1987. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82:528-540.
- Veerkamp, R.F., S. Brotherstone, B. Engel, and T.H.E. Meuwissen. 2001. Analysis of censored survival data using random regression models. *Anim. Sci.* 72:1-10.
- Vehviläinen, H., A. Kaune, A. Kettunen, Præbel, H. Koskinen, and T. Paananen. 2010. Comparison of direct and indirect selection for rainbow trout sea grow-out survival. *Proc. 9th World Congr. Genet. Appl. Livest. Prod., Leipzig, Germany, paper 951.*

- Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661-678.
- Wright, S. 1934. An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* 19:506-536.
- Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, N.G. Martin, G.W. Montgomery, M.E. Goddard, and P.M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42:565–569.
- Yi, N., and S. Xu. 2000. Bayesian mapping of quantitative trait loci for complex binary trait. *Genetics* 155:1391-1403.

Appendix: Computation of the logarithm of $|\mathbf{V}_i|^{-0.5} \exp\{-0.5\mathbf{u}^*'\mathbf{V}_i^{-1}\mathbf{u}^*\}$

The key issue when computing $|\mathbf{V}_i|^{-0.5} \exp\{-0.5\mathbf{u}^*'\mathbf{V}_i^{-1}\mathbf{u}^*\}$ (or at least its proportionality) is the calculation of the determinant and the inverse of \mathbf{V}_i . Given the structure of \mathbf{V}_i , its inverse can be obtained easily. Matrix \mathbf{V}_i can be re-written as $\mathbf{V}_i = v_i(\mathbf{x}_i\mathbf{x}_i' + \mathbf{I}\phi_i)$, where $\phi_i = \sigma_e^2 / v_i$. Reordering rows and columns with respect to genotype,

$$\mathbf{V}_i = v_i \begin{bmatrix} 2 + \phi_i & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 \\ 2 & 2 + \phi_i & 2 & -2 & -2 & -2 & 0 & 0 & 0 \\ 2 & 2 & 2 + \phi_i & -2 & -2 & -2 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 + \phi_i & 2 & 2 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 + \phi_i & 2 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 & 2 + \phi_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_i & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_i \end{bmatrix}$$

where the first diagonal block refers to homozygous 11, the second block refers to homozygous 22 and the third block refers to heterozygous. Note that the dimension of each block depends on the number of animals and on the genotypic frequencies. Thus, the dimensions of the first, second and third blocks are respectively Nf_{11} , Nf_{22} and Nf_{12} , where N is the number of animals and f_{11} and f_{22} and f_{12} are respectively the frequencies of homozygous 11, homozygous 22 and heterozygous. The inverse of \mathbf{V}_i is

$$\mathbf{V}_i^{-1} = \frac{1}{\varepsilon \sigma_e^2} \begin{bmatrix} \varepsilon - 2 & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 \\ 2 & \varepsilon - 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 \\ 2 & 2 & \varepsilon - 2 & -2 & -2 & -2 & 0 & 0 & 0 \\ \\ 2 & 2 & 2 & \varepsilon - 2 & -2 & -2 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & \varepsilon - 2 & -2 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & -2 & \varepsilon - 2 & 0 & 0 & 0 \\ \\ 0 & 0 & 0 & 0 & 0 & 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \varepsilon \end{bmatrix} \quad [1]$$

where $\varepsilon = 2\eta_i + \phi_i$ and $\eta_i = N(f_{11} + f_{22})$.

The determinant of \mathbf{V} is $|\mathbf{V}| = (2\eta_i + \phi_i)(\sigma_e^2)^{N-1}v_i \propto 2\eta_i v_i + \sigma_e^2$, and by substituting this expression into the logarithm of $|\mathbf{V}_i|^{-0.5} \exp\{-0.5\mathbf{u}^* \mathbf{V}_i^{-1} \mathbf{u}^*\}$,

$$\begin{aligned} \log\left(|\mathbf{V}_i|^{-0.5} \exp\{-0.5\mathbf{u}^* \mathbf{V}_i^{-1} \mathbf{u}^*\}\right) = \\ -0.5 \log(2\eta_i v_i + \sigma_e^2) - 0.5 \sum_i \sum_j u_i^* u_j^* \mathbf{V}_{ij}^{-1} \end{aligned} \quad [2]$$

where \mathbf{V}_{ij}^{-1} is the corresponding element of \mathbf{V}^{-1} . We can simplify $\sum_i \sum_j u_i^* u_j^* \mathbf{V}_{ij}^{-1}$ and avoid the explicit inversion of \mathbf{V} . Let w_{11} and w_{22} be the sums of the pre-adjusted liabilities (u_i^*) for animals with genotypes 11 and 22, respectively, and let $s = \sum_N (u_i^*)^2$. Then, from [1],

$$\begin{aligned} \sum_i \sum_j u_i^* u_j^* \mathbf{V}_{ij}^{-1} = \frac{1}{\sigma_e^2 (2\eta_i + \phi_i)} \left[4w_{11}w_{22} - 2w_{11}^2 - 2w_{22}^2 + (2\eta_i + \phi_i)s \right] \\ \propto \frac{1}{\sigma_e^2 (2\eta_i + \phi_i)} \left[4w_{11}w_{22} - 2w_{11}^2 - 2w_{22}^2 \right] \end{aligned} \quad [3]$$

Finally, by substituting [3] into [2],

$$\log\left(\mathbf{V}_i^{-0.5} \exp\{-0.5\mathbf{u}^* \mathbf{V}_i^{-1} \mathbf{u}^*\}\right) \propto -0.5 \log(2\eta_i v_i + \sigma_e^2) + \frac{(w_{11} - w_{22})^2}{\sigma_e^2 (2\eta_i + \phi_i)}$$

Table 1. Accuracy of genome-wide evaluation 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 evaluation¹

q	Model	Phenotypes	$h^2 = 0.1$		$h^2 = 0.3$	
			Training	Testing	Training	Testing
<i>Using SNP in evaluation</i>						
	Linear	Continuous	0.662	0.647	0.821	0.796
0.1	Linear	Dichotomous	0.430	0.410	0.611	0.583
0.1	Threshold	Dichotomous	0.487	0.474	0.673	0.651
0.5	Linear	Dichotomous	0.507	0.486	0.732	0.699
0.5	Threshold	Dichotomous	0.532	0.513	0.751	0.724
<i>Using QTL in evaluation</i>						
	Linear	Continuous	0.899	0.898	0.916	0.915
0.1	Linear	Dichotomous	0.732	0.728	0.737	0.734
0.1	Threshold	Dichotomous	0.739	0.738	0.753	0.750
0.5	Linear	Dichotomous	0.781	0.778	0.780	0.776
0.5	Threshold	Dichotomous	0.799	0.796	0.806	0.802

¹Standard errors (over replicates) ranged from 0.0002 to 0.0023 when SNP were used in the evaluation and from 0.0004 to 0.0146 when QTL were used. SE were calculated as $(\sigma_r^2 / n_r)^{1/2}$, where σ_r^2 is the variance of the corresponding parameter over replicates and n_r is the number of replicates.

Table 2. Accuracy of genome-wide evaluation in training and testing sets for schemes B and 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 evaluation¹

q	Model	Phenotypes	$h^2 = 0.1$		$h^2 = 0.3$	
			Training	Testing	Training	Testing
<i>Using SNP in evaluation</i>						
	Linear	Continuous	0.553	0.501	0.755	0.691
0.1	Linear	Dichotomous	0.371	0.331	0.573	0.523
0.1	Threshold	Dichotomous	0.388	0.348	0.587	0.537
0.5	Linear	Dichotomous	0.444	0.398	0.666	0.603
0.5	Threshold	Dichotomous	0.480	0.435	0.690	0.631
<i>Using QTL in evaluation</i>						
	Linear	Continuous	0.907	0.905	0.932	0.931
0.1	Linear	Dichotomous	0.811	0.809	0.839	0.839
0.1	Threshold	Dichotomous	0.815	0.813	0.845	0.845
0.5	Linear	Dichotomous	0.853	0.852	0.871	0.870
0.5	Threshold	Dichotomous	0.858	0.856	0.879	0.878

¹Standard errors (over replicates) ranged from 0.0002 to 0.0018 when SNP were used in the evaluation and from 0.0004 to 0.0069 when QTL were used. SE were calculated as $(\sigma_r^2 / n_r)^{1/2}$, where σ_r^2 is the variance of the corresponding parameter over replicates and n_r is the number of replicates.

Table 3. Predicted accuracy of genome-wide evaluation (P) and percentage prediction error (E_{S1} and E_{S2}) assessed by simulation for continuous phenotypes analysed with a linear model and for dichotomous phenotypes analysed with a threshold model for different disease prevalences (q) and heritabilities (h^2). Simulated values are for the reference population where individuals are full-sibs (schemes A)¹

q	Phenotypes	$h^2 = 0.1$			$h^2 = 0.3$		
		P	E_{S1}	E_{S2}	P	E_{S1}	E_{S2}
	Continuous	0.954	30.6	5.8	0.984	16.6	6.9
0.1	Dichotomous	0.880	44.6	16.0	0.955	29.5	21.1
0.5	Dichotomous	0.930	42.8	14.1	0.975	22.9	17.3

¹ $E_{S1} = 100(P - S1)/P$, and S1 is accuracy from simulation when SNP are used in GWE; $E_{S2} = 100(P - S2)/P$, and S2 is accuracy from simulation when QTL are used instead of SNP in GWE