Genomic Predictions in Aquaculture: Reliabilities in an Admixed Atlantic Salmon Population

J. Ødegård[1], T. Moen[1], N. Santi[1], S.A. Korsvoll[1], S. Kjøglum[1] and T.H.E. Meuwissen[2]
[1]AquaGen AS, Trondheim, Norway, [2]Norwegian University of Life Sciences, Ås, Norway,

ABSTRACT: Reliability of genomic selection (GS) models was tested in an admixed population of Atlantic salmon, originating from crossing of several wild subpopulations. The GS models included ordinary genomic BLUP models (IBS-GS), using varying marker densities (1 to 220K) and a genomic IBD model (IBD-GS) using genomic relationships estimated through linkage analysis of sparse markers (ignoring LD). The models were compared based on 5-fold cross-validation. The traits studied were log density of salmon lice on skin (logDL) and fillet color (FC), with respective estimated heritabilities of 0.14 and 0.43. IBD-GS and IBS-GS (220K) had similar reliabilities' for FC, while IBS-GS was superior for logDL. The IBS-GS model was remarkable robust to marker density, especially for logDL, and outperformed pedigree-based models at all densities, which may be explained by admixture of the population, introducing long-range LD. Increasing SNP densities beyond 22K had limited effect for both traits.
Keywords: Atlantic salmon; genomic selection; admixture

Introduction

Aquaculture populations are characterized by high male and female fecundity, typically resulting in large full-sib families. For invasive traits, traditional aquaculture selection programs involve sib-testing, which has limited reliability under classical selection schemes. The large family sizes of aquaculture populations imply a substantial potential for within-family selection, given that EBVs can be calculated individually rather than family-wise. Thus, genomic selection (GS) models have a great potential in aquaculture breeding, as they provide individual EBVs even for non-phenotyped (albeit genotyped) selection candidates, based on genotypes and phenotypes of training animals usually from the same full-sib families. Superior performance of GS models for aquaculture populations has been documented in simulation studies (e.g., Nielsen et al. (2009, Ødegård et al. (2009), Ødegård and Meuwissen (2014)), while documentation on its suitability in real aquaculture data has been largely absent so far.

The original idea behind GS was that it would capture linkage disequilibrium between marker loci and QTL (Meuwissen et al. (2001)). However, accuracy of GS has been shown to be non-zero even in absence of LD (Habier et al. (2007)), and the actual reliability of GS models can thus be explained by three types of quantitative-genetic information contained in the genomic data (Habier et al., 2013):

1) Pedigree (classical relationships)
2) Linkage analysis (co-segregation)
3) Population-wide linkage disequilibrium (LD)

The classical pedigree relationships are reflected in inheritance of marker loci and thus implicitly included in GS using dense marker information, although pedigree is not used directly. Linkage analysis information is the deviation from independent segregation of alleles as a result of linkage (i.e., deviations between pedigree-based and linkage-analysis-based relationships), and LD is the statistical dependency between alleles at different loci in the base generation (i.e. the generation with unknown parents). Information on 1) and 2) can thus explain the non-zero reliability of GS even in absence of LD. Furthermore, in populations of strong relationship structure (e.g., livestock and aquaculture populations) LD may not even be the most important of these factors under GS: Wientjes et al. (2013) showed that the level of family relationship between selection candidates and the reference population had a higher effect on reliability of GS than LD *per se*.

There are currently numerous available GS methodologies. The most widely used methods are GS models using identity-by-state (IBS) information on dense genome-wide SNP markers, including the so-called genomic BLUP and Bayesian methods (e.g., BayesA, BayesB, BayesC, BayesD) (Meuwissen et al. (2001), Habier et al. (2011)). Other methods involve use of SNP haplotypes (combining multiple SNPs), that also take identity-by-descent (IBD) information into account (Calus et al. (2008)). Finally, GS may be performed based on linkage analysis of genome-wide markers, producing an IBD genomic relationship matrix (IBD-GS), completely ignoring LD information (Villanueva et al. (2005), Luan et al. (2012)).

In the following, we will focus on two of these methodologies for use in aquaculture breeding: Ordinary genomic BLUP (called IBS-GS) and IBD-GS. IBS-GS can be implemented either by ridge-regression on genome-wide marker genotypes (Meuwissen et al. (2001)) or alternatively based on an animal model using a realized genomic relationship matrix estimated from marker genotype similarities across the genome (Hayes et al. (2009)). The latter method will be used here.

The advantage of the IBD-GS model lies in its ability to model realized IBD relationships more accurately than the pedigree alone, e.g., full-sibs (which are numerous in aquaculture) are no longer necessarily related by a coefficient of ½, but their relationships depend on the actual length of shared IBD chromosome segments, traced by the markers through linkage analysis. Compared with other GS methods IBD-GS has the advantage that it can be successfully implemented even at extremely low marker densities. This is due to the fact that number of recombinations from parent to offspring is usually low (i.e., on average one per Morgan), and inheritance of long chromosomal blocks can thus be traced accurately even with a few genome-wide markers. A simulation study on an aquaculture-like population has shown that IBD-GS works effectively at densities were IBS-based methods is expected to fail, e.g., with 10-20 SNPs/Morgan (Vela-Avitúa et al

(2014)). Thus, there is no need for dense marker panels, making IBD-GS attractive for cost-effective GS implementation. For dairy cattle, IBD-GS models have been shown to give similar reliability as ordinary IBS-GS models with dense markers (Luan et al., 2012). Hence, for livestock populations with large family sizes, realized close relationships (factor 1) and 2) above) are essential for the reliability of any GS model, and GS methodology may thus have large potential even in absence of strong LD structures. Aquaculture populations typically have strong relationships structures, with selection candidates having numerous full-sibs and potentially both maternal and paternal half-sib groups.

The Norwegian AquaGen Atlantic salmon population originates from the first family-based selective breeding program on Atlantic salmon, going back to the 1970'ies, based on crossing of wild founders from numerous Norwegian river strains (http://aquagen.no). Originally, four parallel populations were created, one for each year class in a four year generation interval. Although as much as 42 river strains were originally included, contributions of the different rivers varied considerably, both between the original base populations of the four year classes and as result of subsequent selection. Hence, the original farmed populations were indeed heavily admixed. The year-class strains were selected for a common breeding goal, but kept largely separate for 7 generations until 2005, when they were merged into a single population. Hence, the AquaGen population can be regarded as an admixed/synthetic population comprised of genetic material from many wild subpopulations, which probably have been separated for a long time in nature.

Admixture between genetically distinct populations increases LD between all loci (linked and unlinked) that have different allele frequencies in the founding populations (Pfaff et al., 2001). However, LD between unlinked loci will quickly be removed through recombinations, while LD between linked loci will be more persistent (Figure 1), e.g., for loci separated by 1 or 10 cM, respectively 90% and 35% of the admixture-induced LD (ALD) still remains after 10 generations, while, respectively 82% and 12% remains after 20 generations. However, admixture will not only introduce long-range ALD, it will also reduce the short range LD, i.e., LD existing in the founder populations prior to admixture. The short-range LD will decrease as phase associations between marker and QTL can differ depending of the origin of the chromosome segments (Thomasen et al. (2013)), and haplotype segments with strong LD are thus shorter in admixed populations (Toosi et al. (2010)). This can be illustrated by the following example: The frequency of a M1N1 haplotype is $(p+\kappa)(q+\lambda) + D_I$ in population I, where $p+\kappa$ ($q+\lambda$) is the frequency of allele M1 (N1), expressed as a deviation from the across population frequencies p (q), with frequency deviations κ and λ, and D_I is the LD in population I. Similarly, $(p-\kappa)(q-\lambda) + D_{II}$ is the haplotype frequency in population II, and that in their crossbred-offspring (F1) is $pq + \kappa\lambda + \bar{D}$, where $\bar{D}$ is the average of D_I and D_{II}. Thus the LD in F1 is $\kappa\lambda + \bar{D}$, which comprises a ALD term $\kappa\lambda$ due to the crossbreeding (depends on frequency differences and is independent of distances between the loci), and the average of the old LD coefficients between the loci $\bar{D}$. $\bar{D}$ is probably smaller than either D_I or D_{II} since they may have opposite signs in the two populations, resulting in a reduced short-range LD in the admixed population.

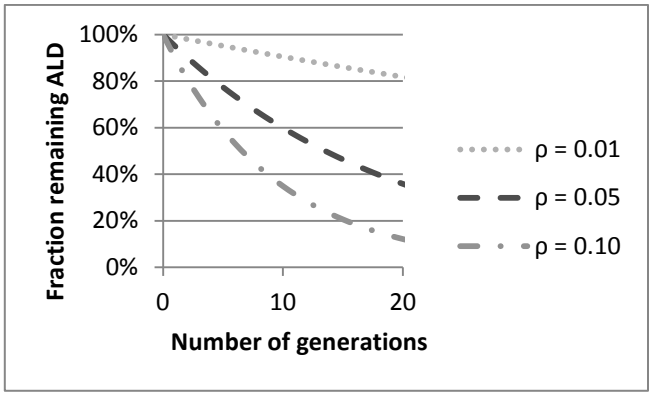

Figure 1: Decay of admixture-induced LD (ALD) between loci over generations as a function of the recombination rate ρ(distance in Morgans)

The reduced short-range LD originating from founder populations may challenge accurate genomic prediction. Still, long-range ALD (the $\kappa\lambda$ term) can be effectively captured even by sparse markers, but may explain a limited fraction of the genetic variance, depending on the degree of differentiation between the founding populations. Hence, effectiveness of GS in admixed populations depends on several layers of information: remaining LD from the founder populations, long-range ALD, and the relationship structure within the existing population. Furthermore, the relative importance of these factors likely depends on genetic architecture, marker density, heritability and the GS methodology used.

The aim of the study was to quantify the importance of marker density on the reliability of ordinary IBS-GS models and to compare these estimates with IBD-based models completely ignoring LD, i.e., classical pedigree-based models and IBD-GS models. To this end, two traits measured on Atlantic salmon, with high and low heritability, using alternative GS models and marker densities were studied.

Materials and Methods

Data.

The fish material used in the current study consists of second generation offspring after the merge of the AquaGen year-class populations. The phenotypes consisted of salmon lice (*Lepeophtheirus salmonis*) counts (LC) on the skin surface and fillet color (FC), machine-recorded on a continuous scale after slaughter. The LC was recorded in two separate tests on live fish during the grow-out period (July and October, 2012). Different fish were recorded in the two tests. LC phenotypes were generally highly skewed and counts increased with body size (i.e., with skin

surface). For these reasons, the LC was transformed to log density of lice, calculated as:

$$\text{logDL} = \log\left(\frac{\text{LC}+1}{\text{BW}^{\frac{2}{3}}}\right)$$

where BW is the body weight of the fish at recording, and $\text{BW}^{\frac{2}{3}}$ is a measure for the approximate surface of the fish as a function of body weight. In a preliminary analysis, the logDL of the two tests showed high, albeit imperfect, genetic correlation (~0.7). For simplicity, the current study only involves logDL phenotypes from the first test, which constitute the majority of the genotyped fish. In total, 2850 (1444 genotyped) and 1936 (1869 genotyped) fish were phenotyped for the traits logDL and FC, respectively. The fish came from 157 full-sib families. These families resulted from mating of 97 sires of 99 dams (i.e., 1-2 offspring groups per parent). A custom-made 220k Affymetrix axiom SNP-chip was used for genotyping.

The fish genotyped in this study were part of a selection experiment aimed at selecting for high/low logDL. For this purpose, fish from high and low logDL families were more likely to be genotyped than intermediate families with respect to this trait.

Model. The data were analyzed using univariate animal models, with the following general characteristics:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}$$

Where $\mathbf{y}$ is a vector of phenotypes (logDL or FC), $\mathbf{a}\sim\text{N}(\mathbf{0}, \mathbf{G}\sigma_g^2)$ is a vector of random additive genetic effects, where **G** is a given relationship matrix (model dependent), and $\mathbf{e}\sim\text{N}(\mathbf{0}, \mathbf{I}\sigma_e^2)$ is a vector of random residuals. The fixed effects (**b**) included person (responsible for counting) by day for logDL, and gender of fish for FC. Common environmental effects of family were also tested, but these effects were small and not significantly different from zero (P>0.20) for both traits, and were thus dropped in the final model.

The different models differed solely with respect to their specification of the relationship matrix **G**:

PED: Classical pedigree-based analysis, i.e., **G** = **A** (numerator relationship matrix)

IBD-GS: Identity-by-descent GS, using a linkage-based IBD relationship matrix for the genotyped animals. The matrix was calculated from a sparse marker set containing 5590 mapped genome-wide SNP markers, using the LDMIP software (Meuwissen and Goddard, 2010). The number of mapped SNPs per chromosome varied from 52 to 396, and relationship matrices were thus computed for each chromosome separately and subsequently averaged over chromosomes to produce **G**.

IBS-GS: Identity-by-state GS (ordinary GBLUP), calculating the **G** directly from genome-wide SNP markers using the second method by VanRaden (2007). Alternative **G** matrices were tested by extracting random sub-sets from the complete marker data set, including either a) 1100 (1K), b) 2200 (2K), c) 4400 (4K), d) 22 000 (22K), e) 55 000 (55K), or f) all 220 000 (220K) SNP markers, respectively. For a) to d) at total of 10 non-overlapping replicates were generated, while e) was replicated 4 times.

All models utilizing genomic information (IBD-GS and IBS-GS) used one-step estimation of EBVs (Legarra et al., 2009, Christensen and Lund, 2010), combining relationships from genotyped and ungenotyped individuals into a unified relationship matrix **H**. The one-step method allows mixing of genomic and polygenic relationship matrices as $\mathbf{G_\omega} = (1-w)\mathbf{G} + w\mathbf{A}$. Here, w = 0, implying that relationships among genotyped animals were solely based on the respective genomic relationship matrices. However, the **G** matrices were adjusted to the same average rate of inbreeding and relationship as the numerator relationship matrix, using the ADJUST option in DMU (Madsen and Jensen (2014)). Identical variance components were used in all models, which were estimated with the PED model using all phenotypic data.

Reliabilities of the different models were assessed through predictive ability, using five-fold cross-validation, i.e., individuals being both phenotyped and genotyped were randomly sampled into five validation sets, which were predicted one at a time, masking the phenotypes of the validation animals and using all the remaining phenotypes and genotypes as training data. Reliability was estimated as:

$$R^2_{EBV,BV} = \frac{R^2_{EBV,y}}{h^2}$$

where $R^2_{EBV,y}$ is the squared correlation between EBVs of a given model (predicted from the training data, without the phenotype of the animal itself) and the recorded phenotype (y), while h^2 is the estimated heritability of the trait.

Results and Discussion

Estimated heritability for the two traits differed considerably, with logDL having a relatively low heritability (0.14±0.03), while FC had a relatively high heritability (0.43±0.06).

Reliability of the PED model was slightly higher for FC (0.36) than for logDL (0.34). Reliability is expected to increase with heritability (although not linearly), but the relative difference in reliability between these traits was likely reduced, as the genotyped validation animals were more likely to come from selected high/low families with respect to logDL. Hence, between-family variation for logDL is expected to be inflated among the genotyped animals, giving an apparently higher reliability of the PED model based on genotyped animals for this trait. Relative reliabilities of the five-fold cross validation of the different GS models (relative to PED) are presented in Figure 2 for logDL and Figure 3 for FC. In general, all GS models outperformed the classical PED model, but the relative increase in reliability varied considerably between models and traits. For logDL, the relative increase in reliability using GS was substantial (up to 52% for IBS-GS with 220K), but moderate for FC (21% for IBD-GS and 22% for IBS-GS with 220K). Still, the relative advantage of GS (compared with PED) for logDL may be underestimated due to the inflation of between-family variation for genotyped animals of this trait, as this is expected to

especially increase the apparent performance of the PED model.

Using IBS-GS, higher marker density was always favorable, but the relative advantage was considerably more expressed in FC than in logDL. For example, the relative increase in reliability of IBS-GS for FC was 39% when going from 4K to 220K, while the corresponding increase for logDL was only 11%. Nevertheless, IBS-GS was superior to PED for both traits, even at the lowest marker densities (1K). For both traits, going from 22K to 220K SNPs increased reliability by only ~1%. Hence, increasing SNP density beyond 22K would have little practical effect on selection.

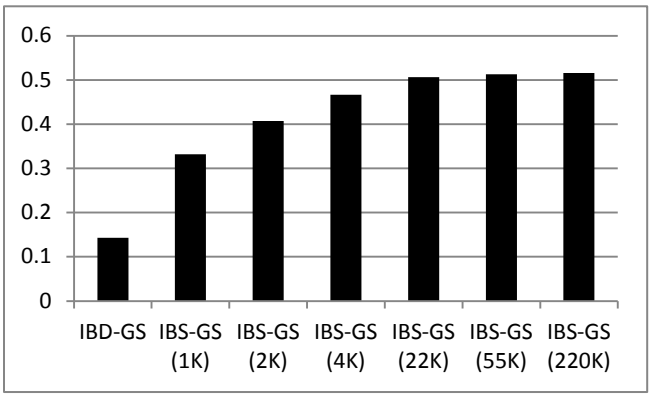

Figure 2: Relative increase in reliability of logDL through genomic selection models compared with the PED model

Another striking result was the enormous difference between the traits with respect to relative reliability of the IBD-GS model. This model does not utilize LD, but utilizes more exact IBD relationships between close relatives through linkage analysis. For the lowly heritable logDL the IBD-GS only slightly improved reliability compared with PED, while for the more highly heritable FC, IBD-GS with sparse mapped SNP markers was about as reliable as IBS-GS with 220K SNP markers.

The models used in this study utilize the sources of information contained in genomic data differently: IBS-GS utilizes (potentially) pedigree, linkage analysis and LD; IBD-GS utilizes pedigree and linkage analysis, while PED, by definition, utilizes the pedigree relationships only. For the IBS-GS model, high marker density would be needed to capture both short-range LD and (tiny) variations in co-segregation. In contrast, the IBD-GS model will utilize linkage analysis information accurately, even at very low marker densities. Furthermore, the relative importance of the different types of information depends on several factors such as; structure of the dataset (i.e., number of close relatives in the population), historical *Ne* (amount of LD), as well as the heritability of the traits involved. In general, it is expected that for a lowly heritable trait, genetic effects estimated over larger groups of individuals, such as LD-associated effects (general association between marker genotypes and phenotypes) and pedigree relationships (i.e., mid-parent means) would be relatively more important for the reliability, while linkage-analysis based deviations from pedigree relationships (i.e., largely minor individual deviations) would be relatively more important at higher heritabilities. Thus, the relative advantage of the IBS-GS model may be largest at low heritability (e.g., logDL), while IBD-GS would be expected to perform relatively better at higher heritability (e.g., FC), which is consistent with results of this study.

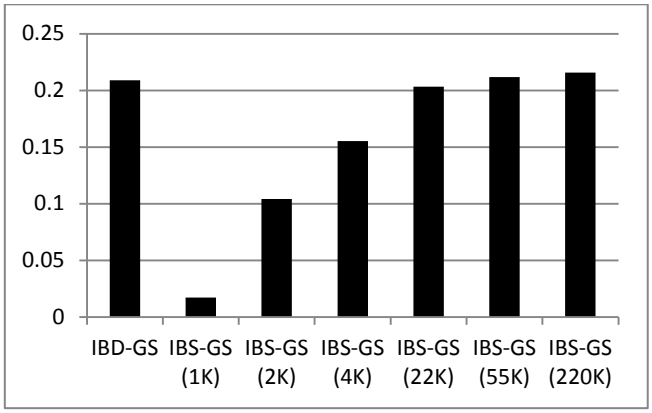

Figure 3: Relative increase in reliability of FC through genomic selection models compared with the PED model

Still, the factors discussed above do not explain the favorable performance of IBS-GS for logDL at extremely low marker densities (e.g., 4K), for which limited LD is expected (in absence of admixture). The explanation may thus lie in the selection history of farmed Atlantic salmon; as described in the introduction, admixture from several distinct wild strains is likely to have introduced long-range LD, while simultaneously reducing the short-range LD in the population. Simultaneous increase in long-range LD and decrease in short-range LD will likely reduce the relative advantage of dense SNP data, as a relatively larger fraction of the available LD may be captured even by sparse marker panels. This may explain the good performance of IBS-GS for logDL at extremely low marker densities. Current terrestrial livestock populations may also be formed by admixtures of old populations, but these admixture events occurred longer ago and may have been less extreme than in Atlantic salmon. Still, some admixture effects on the LD structure, as reported here, may also be seen in terrestrial livestock species, and may contribute to the rather small increases observed in accuracy of GS as marker density increases (VanRaden et al., 2011). Still, a high marker density in IBS-GS will be favorable for utilization of linkage analysis information, which is mainly an advantage at higher heritabilities and strong relationship structures (Ødegård and Meuwissen (2014)), i.e., as seen with FC.

The number of genotyped animals was rather limited in the current study. Genotyping larger fractions of the population would be expected to increase the reliability of GS models even further.

Conclusion

There is a substantial potential for more accurate selection for sib-evaluated traits in Aquaculture species through GS. The GS models outperformed pedigree-based

models even at extremely low marker densities (1K). The latter result may be explained by the admixed origin of the AquaGen Atlantic salmon population, likely reducing short-range and increasing long-range LD in the population. Long-range LD may be effectively utilized even at low marker densities. Still, increasing SNP density was indeed favorable for IBS-GS, but densities beyond 22K had limited additional effect on the reliability for both traits. As expected, linkage analysis information (IBD-GS) was relatively more important at high heritability (FC), while LD and classical additive-genetic relationships (mid-parent means) were relatively more important at low heritability (logDL).

Acknowledgements

The study was partly funded by the Norwegian Research Council through projects no. 200551/s40 and 225181.

Literature Cited

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos et al. (2008). Genetics 178, 553-561.

Christensen, O. and M. Lund. (2010). Gen. Sel. Evol. 42, 2.

Habier, D., R. Fernando, K. Kizilkaya et al. (2011). BMC Bioinformatics 12, 186.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. (2007). Genetics 177, 2389-2397.

Habier, D., R.L. Fernando, and D.J. Garrick. (2013). Genetics 194, 597-607.

Hayes, B.J., P.M. Visscher, and M.E. Goddard. (2009). Gen. Res. 91, 47-60.

Legarra, A., I. Aguilar, and I. Misztal. (2009). J. Dairy Sci. 92, 4656-4663.

Luan, T., J.A. Wooliams, J. Ødegård, et al. (2012). Gen. Sel. Evol. 44, 28.

Madsen, P. and J. Jensen. (2014). DMU. Version 6, release 5.2.

Meuwissen, T.H.E. and M.E. Goddard. (2010). Genetics 185, 1441-1449.

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

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

Ødegård, J. and T.H.E. Meuwissen. (2014). Gen. Sel. Evol. 46, 3.

Ødegård, J., M.H. Yazdi, A.K. Sonesson et al. (2009). Genetics 181, 737-745.

Pfaff, C.L., E.J. Parra, C. Bonilla, et al. (2001). Am. J. Hum. Gen. 68, 198-207.

Thomasen, J.R., A.C. Sorensen, G. Su, et al. (2013). J. Anim. Sci. 91, 3105-3112.

Toosi, A., R. L. Fernando, and J. C. M. Dekkers. (2010). J.

VanRaden, P. M. (2007). J. Dairy Sci. 90, 374-375.

VanRaden, P.M., J. O'Connell, G. Wiggans, et al. (2011). Gen. Sel. Evol. 43, 10

Vela-Avitúa, S., T.H.E. Meuwissen, T. Luan (2014). Gen. Sel. Evol. (submitted).

Wientjes, Y.C.J., R.F. Veerkamp, and M.P.L. Calus. (2013). Genetics 193, 621-631.

Villanueva, B., R. Pong-Wong, J. Fernández et al. (2005). J. Anim. Sci. 83, 1747-1752.