Genomic Selection in Aquaculture

Sho HOSOYA[*1], Kiyoshi KIKUCHI[*1], Hiroshi NAGASHIMA[*2], Junichi ONODERA[*2],
Kouishi SUGIMOTO[*2], Kou SATOH[*2], Keisuke MATSUZAKI[*2], Masaki YASUGI[*3],
Atsushi J. NAGANO[*3,4], Akira KUMAGAYI[*2], Kenichi UEDA[*2], and Tadahide KUROKAWA[*5]

Abstract: Recent advances in genotyping and sequencing technology facilitate handling thousands of SNPs from hundreds of samples. This enables us genomic prediction of breeding values for economically important traits not only for livestock and plants but also for aquaculture species. In this article, we review six empirical studies on genomic prediction in aquaculture. Implementation of genomic selection for aquaculture is steadily progressing with no doubt, and we will see the tangible and practical outcomes of genomic selection within several years.

Key words: Genomic selection, NGS, selective breeding, SNP

Introduction

Genomic selection (GS) introduced by Meuwissen *et al.* (2001) is a form of marker-assisted selection where the genomic estimated breeding value (GEBV) is predicted as the sum of additive genetic effects of genome-wide dense genetic markers. All loci affecting the trait (e.g. quantitative trait loci; QTL) are supposed to be in linkage disequilibrium with at least one marker. Therefore, to predict marker effects for highly polygenic traits, such as human height and disease (Yang *et al.*, 2015; Abraham *et al.*, 2015), hundreds of thousands of genetic markers are required. When the idea was proposed for the first time, it was highly challenging to genotype thousands of markers from hundreds of specimens. However, recent advances in genotyping technology put it into practice.

Succinctly, GS consists of two steps. The first step is estimation of marker effects in a test (training) group, and the second step is prediction of GEBV of selection candidates (validation group), sibs and/or relatives of individuals of the training group. There are two major methods for marker effect estimation: GBLUP and BayesB. To use these methods, a prior distribution of marker effects is required, and one of the major differences between the two methods is the assumption of the distribution and variances of marker effects. Normal distribution with constant variance is assumed for GBLUP, and therefore the model is equivalent to a conventional BLUP (PBLUP) animal model; a numerator relationship matrix estimated from pedigree information is substituted by a realized relationship matrix estimated from genome-wide SNP information in GBLUP. On the other hand, non-normal distribution is assumed for BayesB, where only a subset of markers has effects and these effects follow a reflected exponential distribution. Simulation data revealed that the prediction accuracy of these two methods superior to that of traditional BLUP (Meuwissen *et al.*, 2001). This is mainly because genomic prediction takes the Mendelian sampling term into account (Daetwyler *et al.*, 2007). Because of its high prediction accuracy,

Sho HOSOYA, Kiyoshi KIKUCHI, Hiroshi NAGASHIMA, Junichi ONODERA, Kouishi SUGIMOTO, Kou SATOH, Keisuke MATSUZAKI, Masaki YASUGI, Atsushi J. NAGANO, Akira KUMAGAYI, Kenichi UEDA, and Tadahide KUROKAWA

Table 1. Summary of species, traits, genotyping platform and prediction method researches about genomic selection for aquaculture

Species	Trait	Genotyping platform	Prediction method	Citation
Atlantic salmon (*Salmo salar*)	parasite resistance fille color	SNP array	GBLUP	Ødegård *et al.* (2014)
	body size	SNP array	GBLUP	Tsai *et al.* (2015)
	parasite resistance	SNP array	GBLUP	Tsai *et al.* (2016)
Rainbow trout (*Onchorhynchus mykiss*)	disease resistance	SNP array, RAD	ssGBLUP, wssGBLUP, BayesB, BayesC	Vallejo *et al.* (2016)
Yellow croaker (*Larimichthys crocea*)	body size fatty acid composition	Genotyping-by-Sequencing (GBS)	GBLUP, emBayesB	Dong *et al.* (2106)
Yesso scallop (*Patinopecten yessoensis*)	body size	2d RAD-seq	GBLUP, LASSO, Bayesian LASSO, BayesA, BayesB, rrBLUP	Dou *et al.* (2016)
Coho salmon (*Onchorhynchus kisutch*)	body size	ddRAD-seq	GBLUP	Hosoya *et al.* (submitted)

* adjusted prediction accuracy: the correlation between the GEBV of the test population and the actual phenotypes divided by the square toot of the heritability

genomic selection is now widely used in livestock and plants, but there has been limited realization in aquaculture.

With the aids of new genotyping technologies, such as SNP array and restriction enzyme associated DNA sequencing (RAD-seq), genomic selection becomes applicable not only for the major livestock species and plants but also for aquatic species. So far, at least six studies have reported the possibility of genomic selection in aquatic species (Table 1). In this report, we review these studies. Five are about finfish, and the other is about scallop. Three of the studies were done by researchers from North Europe, one was from North America, and the other from Mainland China. All these studies have been published within the last couple of years, and more attempts of GS for a wide range of aquatic species will be published from all over the world.

Atlantic salmon

As mentioned above, there are three articles on Atlantic salmon, and one of them (Ødegård *et al.*, 2014) was the first report attempting to incorporate GS in aquaculture. Atlantic salmon is one of the most important aquaculture species as its estimated global economic values was $ 7.8 billion in 2010 (FAO, 2010). It is also known as a genomic model species for aquaculture as its chromosomal-level genome assembly (GCA_000233375.4) is available (Lien *et al.*,

2016). Additionally, dense SNP chips (Axiom® Salmon Genotyping Array, Affimetrix) are commercially available. All those three studies used SNP chips for genotyping. Analyzed traits for this species are salmon lice (*Lepeophtheirus salmonis*) resistance (LR: estimated heritabilities = 0.14) and fillet color (estimated heritabilities = 0.43) (Ødegård *et al.*, 2014); weight (estimated heritabilities = 0.5) and length (estimated heritabilities = 0.6) (Tsai *et al.*, 2015); LR (estimated heritabilities = 0.2 - 0.3) (Tsai *et al.*, 2016).

GEBV for these traits were predicted by GBLUP. Relatively high prediction accuracy was generated for traits with higher heritability. Interestingly, while over 50k SNPs are required to gain better prediction accuracy for livestock species, 5k SNPs are sufficient for Atlantic salmon within-family prediction. However, for prediction across populations or year groups, where genetic relationships are more distant between training and validation groups, accuracy was substantially low even when 30k SNPs were used for the prediction (Tsai et. al., 2016). In such cases, using larger population sample sizes and higher-density SNP genotypes will improve prediction accuracy. Thus, phenotyping training populations consisting of animals closely-related to the selection candidates and genotyping relatively lower-density of SNPs (5k) will gain better cost performance for commercial salmon breeding schemes. This strategy is supported by simulation

studies (e.g. Sonesson and Meuwissen, 2009) and generally works for populations from most aquaculture breeding programs since relatively limited numbers of broodstock are used in most of the aquaculture breeding programs.

Rainbow trout

Vallejo *et al*., (2016) reported an attempt to implement genomic selection for bacterial cold water disease (BCWD) resistance to the National Center for Cool and Cold Water Aquaculture (NCCCWA) BCWD resistance breeding line. Several major resistance QTL (including a QTL of PVE = 58%) have been detected for the trait using the same breeding line (Palti *et al*., 2015). However, because of complex genetic architecture and high genetic variation, the authors postulated that GS gives better performance for genetic improvement in BCWD resistance than the marker-assisted selection based on QTL analyses.

Genomic predictions for survival days and survival status were done using single step GBLUP (ssGBLUP), weighted ssGBLUP (wssGBLUP), BayesB and BayesC. ssGBLUP is an integrated version of GBLUP and PBLUP. While GBLUP and PBLUP use either genomic or pedigree information to construct a relationship matrix (i.e. G matrix and A matrix), ssGBLUP uses both types of information (H matrix) (Aguilar *et al*., 2010). wssGBLUP is an extended version of ssGBLUP, where QTL effects are weighted. Overall predictive abilities were similar among the GS models and PBLUP. ssGBLUP showed slightly better performance compared to the Bayesian methods. This will be partly because a large number of individuals was used for the training population in ssGBLUP. Construction of H matrix for ssGBLUP seems somewhat complicated since it includes "tuning" steps. However, this step can increase the performance of genomic prediction. Moreover, because individuals with phenotype, but without genotype, data can be included in the training population under ssGBLUP model, it is possible to increase the sample size of training population without increasing genotyping cost. Therefore, ssGBLUP will be one of the most powerful solutions for genomic prediction of GEBV. However, pedigree records are not typically maintained in aquaculture because of the difficulty in the tagging of individual larvae and the maintenance of separate families. Since the number of selection candidates is large and the value of the selection candidate is low for aquaculture species, the cost of pedigree recording may not be negligible, and this will be a major obstacle to implement ssGBLUP into selective breeding programs in aquaculture.

Large yellow croaker

Genomic selection is also attempted on non-salmonid fish. Dong *et al*. (2016) reported the feasibility of genomic selection in the traits of growth rate and meat quality (i.e. the percentage of n-3 highly unsaturated fatty acids (n-3HUFA) in muscle) of large yellow croaker (*Larimichthys crocea*). They first estimated heritability and then compared predictive ability between GBLUP and emBayesB. emBayesB is an alteration of BayesB. The Markov Chain Monte Carlo (MCMC) technique (i.e. Gibbs sampling) used in BayesB for the model fitting requires large computational time with dense SNP data. On the other hand, MCMC is replaced by the Expectation-Maximization (EM) algorithm in the emBayesB approach. This enables us fast but accurate GEBV prediction (Shepherd *et al*., 2010).

Estimated heritability for body weight, body length and n-3HUFA were 0.604, 0.586 and 0.438, respectively. GBLUP was superior to emBayesB in the predictive abilities for body weight and body length (GBLUP: BW = 0.41 and BL= 0.4; emBayesB: BW = 0.37 and BL= 0.37), but not for the n-3HUFA trait with relatively lower heritability (GBLUP: 0.30; emBayesB: 0.32). The differences are probably due to the number of QTLs affecting the traits; the number of QTLs is expected to be smaller for the n-3HUFA trait than that for body weight and body length. The results suggested the importance of testing algorithms on specific traits to gain the best prediction performance. The authors also estimated that at least 1000 individuals in the training population are required to get prediction accuracy of 0.8 by fitting the curve of prediction accuracy. The number will be affordable when GBS is used for genotyping.

Scallop

A species other than finfish tested for the

Sho HOSOYA, Kiyoshi KIKUCHI, Hiroshi NAGASHIMA, Junichi ONODERA, Kouishi SUGIMOTO, Kou SATOH, Keisuke MATSUZAKI, Masaki YASUGI, Atsushi J. NAGANO, Akira KUMAGAYI, Kenichi UEDA, and Tadahide KUROKAWA

possibility of implementation of GS is Yesso scallop (*Patinopecten yessoensis*) (Dou *et al.* 2016). In the study, the performance of 2b-RAD sequencing methods (a type of RAD-seq, or genotyped-by-sequencing GBS), where the uniform fragments produced by type IIB restriction endonucleases are sequenced: Wang *et al.*, 2012) was evaluated for shell length, shell width and shell height. The prediction accuracies calculated under models of GBLUP, LASSO, Bayesian LASSO, BayesA, BayesB and rrBLUP were compared.

The real dataset involved 349 individuals consisting of two full-sib families and three bi-parental families. A high-quality 2,364 putative SNPs with an average calling rate of 84% was obtained by 2b-RAD (minor allele frequency > 5%; SNP calling frequency > 70%). The estimated heritability of the three traits using the entire population were 0.36 - 0.48. Those values varied among families (0.28 - 0.61 for SH, 0.26 - 0.60 for SL, and 0.15 - 0.48 for SW). This implies large differences in genetic diversity among families. The (adjusted) prediction accuracies varied from 0.30 to 0.60 across the three traits, showing 2b-RAD to be a powerful and cost-effective genotyping method for GS for Yesso scallop breeding programs. The prediction accuracy of GBLUP, BayesA and BayesB outperformed the other methods across the three traits. This is partly because these three models can effectively capture the polygenic resemblance and genetic relationships (Neves *et al.*, 2012; Resende *et al.*, 2012; Moser *et al.*, 2009). However, the performance is largely depending on the number of samples, population structure and heritability of the traits, and the three methods may not always give better performance than the others in different populations and traits.

Conclusion

It seems a relatively small number of SNPs (~ 5k) is required for genomic prediction of aquaculture species compared to that of livestock (>50k) to obtain practical levels of prediction accuracy. Our result on coho salmon (Hosoya *et al.*, submitted) is also supportive of this idea. The reason is partly because the generation is still young and LD size is large in many aquaculture breeding populations compared to the livestock populations. Although GBLUP and Bayesian methods showed good performance in prediction accuracy, it will be better to compare the performance among between prediction models before deciding the model to use because the performance is largely depending on the number of samples, population structure and heritability of the traits.

References

Meuwissen T. H. E., Hayes B. J., and Goddard M. E., 2001: Prediction of total genetic valueusing genome-wide dense marker maps. *Genetics* **157**, 1819–1829.

Yang J., Bakshi A., Zhu Z., Hemani G., Vinkhuyzen A. A., Lee S. H., Robinson M. R., Perry J. R., Nolte I. M., van Vliet-Ostaptchouk J. V., Snieder H., LifeLines Cohort Study, Esko T., Milani L., Mägi R., Metspalu A., Hamsten A., Magnusson P. K., Pedersen N. L., Ingelsson E., Soranzo N., Keller M. C., Wray N. R., Goddard M. E., and Visscher P. M., 2015: Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120.

Abraham G. and Inouye M., 2015: Genomic risk prediction of complex human disease and its clinical application. *Curr. Opin. Genet. Dev.* **33**, 10–16.

Daetwyler H. D., Villanueva B., Bijma P., and Woolliams J. A., 2007: Inbreeding in genome-wide selection. *J. Anim. Breed. Genet.* **124**, 369–376.

Ødegård, J., Moen T., Santi N., Korsvoll S. A., Kjøglum S., and Meuwissen T. H. E. 2014: Genomic prediction in an admixed population of Atlantic salmon (*Salmo salar*). *Front. Genet.* **5**,402.

Food and Agriculture Organization 2010: FAO yearbook. Fishery and Aquaculture Statistics. F.A.O. Rome

Lien S., Koop B. F., Sandve S. R., Miller J. R., Kent M. P., Nome T., Hvidsten T. R., Leong J. S., Minkley D. R., Zimin A., Grammes F., Grove H., Gjuvsland A., Walenz B., Hermansen R.A., von Schalburg K., Rondeau E. B., Di Genova A., Samy J. K., Olav Vik J., Vigeland M.D., Caler L., Grimholt U., Jentoft S., Våge D. I., de Jong P., Moen T., Baranski M., Palti

Y., Smith D. R., Yorke J. A., Nederbragt A. J., Tooming-Klunderud A., Jakobsen K. S., Jiang X., Fan D., Hu Y., Liberles D. A., Vidal R., Iturra P., Jones S. J., Jonassen I., Maass A., Omholt S. W., and Davidson W. S. 2016: The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205.

Tsai H. Y., Hamilton A., Tinch A. E., Guy D. R., Gharbi K., Stear M. J., Matika O., Bishop S. C., and Houston R. D., 2015: Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. *BMC Genom.* **16**, 969.

Tsai H. Y., Hamilton A., Tinch A. E., Guy D. R., Bron J. E., Taggart J. B., Gharbi K., Stear M. J., Matika O., Pong-Wong R., Bishop S. C., and Houston R. D., 2016: Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. *Genet. Sel. Evol.* **48**, 47.

Sonesson A. K. and Meuwissen T. H. E., 2009: Testing strategies for genomic selection in aquaculture breeding programs. *Genet. Sel. Evol.* **41**, 37.

Vallejo R. L., Leeds T. D., Fragomeni B. O., Gao G., Hernandez A. G., Misztal I, Welch T. J., Wiens G. D., and Palti Y. 2016: Evaluation of genome-enabled selection for bacterial cold water disease resistance using progeny performance data in rainbow trout: insights on genotyping methods and genomic prediction models. *Front. Genet.* **7**, 96.

Palti Y., Vallejo R. L., Gao G., Liu S., Hernandez A. G., Rexroad C. E. 3rd., and Wiens G. D., 2015: Detection and validation of QTL affecting bacterial cold water disease resistance in rainbow trout using restriction-site associated DNA sequencing. *PLoS One* **10**, e0138435.

Aguilar I., Misztal I., Johnson D. L., Legarra A., Tsuruta S., and Lawlor T. J., 2010: Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* **93**, 743–752.

Dong L., Xiao S., Wang Q., and Wang Z., 2016: Comparative analysis of the GBLUP, emBayesB, and GWAS algorithms to predict genetic values in large yellow croaker (*Larimichthys crocea*). *BMC Genomics* **17**, 460.

Shepherd R. K., Meuwissen T. H. E., and Wooliams J. A., 2010: Genomic selection and complex trait prediction using a fast EM algorithm applied to genome-wide markers. *BMC Bioinfomatics* **11**, 529.

Dou J., Li X., Fu Q., Jiao W., Li Y., Li T., Wang Y., Hu X., Wang S., and Bao Z., 2016: Evaluation of the 2b-RAD method for genomic selection in scallop breeding. *Sci. Rep.* **6**, 19244.

Wang S., Meyer E., McKay J. K., and Matz M. V., 2012: 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nature Methods* **9**, 808–810.

Neves H. H., Carvalheiro R., and Queiroz S. A., 2012: A comparison of statistical methods for genomic selection in a mice population. *BMC Genet.* **13**, 100.

Resende M. F., Muñoz P., Resende M. D., Garrick D. J., Fernando R. L., Davis J. M., Jokela E. J., Martin T. A., Peter G. F., and Kirst M., 2012: Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda L.*). *Genetics* **190**, 1503–1510.

Moser G., Tier B., Crump R. E., Khatkar M. S., and Raadsma H. W., 2009: A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* **41**, 56.

Hosoya S., Kikuchi K., Nagashima H., Onodera J., Sugimoto K., Satoh K., Matsuzaki K., Yasugi M., Nagano A. J., Kumagayi A., Ueda K., and Kurokawa T., 2016: Genomic selection with ddRAD-seq for body size traits of Coho salmon (*Oncorhynchus kisutsh*). Submitted.