Selective Breeding Against Infectious Diseases In Atlantic Cod With Whole Genome Sequence Data

X. –J Yu[1], T.H.E. Meuwissen[1], M. Baranski[2], A.K. Sonesson[2],.
[1]Norwegian University of Life Sciences, [2]Nofima, Ås, Norway

ABSTRACT: Atlantic cod families from year 2009 of the Norwegian national cod breeding program were challenged for viral nervous necrosis (VNN) and vibriosis. Mortality was recorded. Around 1600 offspring and their parents were genotyped at 10,913 SNP loci, covering 2,285 scaffolds/contigs in the Atlantic cod reference genome, which accounts for ~71.3% of total sequence length. Genomic enabled breeding values (GEBV) were estimated. A 10-folds cross-validation shows that the correlations of the survival states and corresponding GEBV were 0.085 for vibriosis and 0.55 for VNN. Whole genome resequencing of 111 parents was performed to an approximately 12x coverage per individual. Variant calling in the sequence of a subset of parents showed that all 12K SNP array SNPs were called and had matching genotypes. Imputation with Beagle and LDMIP software will enable inference of sequence data for all the challenge tested fish and the resulting improvement in accuracy will be investigated.

Keywords: Atlantic cod; Disease breeding; genome sequence

Introduction

Vibriosis and viral nervous necrosis (VNN) are two major infectious diseases that are prevalent in marine fishes, e.g., Atlantic salmon and Atlantic cod. They cause significant losses in aquaculture worldwide. VNN in Atlantic cod has been reported to have a high heritability of 0.75 (Ødegård et al. (2010)). An earlier linkage analysis showed that five genome-wide significant QTL detected explain 68% of the VNN phenotypic variance for cod survival (Banranski et al. (2010)). The heritability of vibriosis in cod is low at 0.08-0.17 (Kettunen et al., (2007)).

Challenge testing is the standard procedure to evaluate family resistant performance to infectious pathogens. The challenged fish however can not be used as broodstock due to disease transmission risk. Genomic selection or marker assisted selection (MAS) provide a perfect approach to circumvent such difficulties for breeding of disease resistance traits.

Genomic selection (Meuwissen et al., (2001)) is now widely adopted in livestock breeding. Its ubiquitous use is mainly due to low cost high density SNP genotyping arrays. Up to 40% accuracy improvement can be expected with whole-genome sequence data (Meuwissen and Goddard, (2010a)). The objectives of this study are to 1) estimate genomic enabled breeding values (GEBV) of breeding candidates using SNP array data; and 2) examine the advantage of GEBV estimation with next generation sequence (NGS) data.

Materials and Methods

Genotyping, sequencing and disease resistance phenotypes. Atlantic cod families from the 2009 year class of the national cod breeding program in Norway were challenge tested for VNN and vibriosis. Day of death and binary dead/alive traits were recorded. Around 800 offspring from each challenge, consisting of about 10 animals per full-sib family, and their parents were genotyped using an Illumina 12K SNP array (containing 10,913 SNPs). In addition, whole genome resequencing of the parents of the families using paired-end 100bp reads at an approximate coverage of 12X per individual was performed by AROS Applied Biotechnology A/S in Aarhus, Denmark.

Variant calling. Raw sequence data were quality controlled and filtered (Hayes et al. (2013)). The reads with more than three 'N' in the sequence, or mean phred score of less than 20 were removed. Low quality bases (phred < 20) were removed from 3'-end. If after the trimming the read length was less than half of its original length, it was also discarded. Unpaired reads were stored separately as single ended reads.

The Atlantic cod reference genome is still at 'nonchromosomal' status (ftp://ftp.ensembl.org/pub/release-74/fasta/gadus_morhua/dna (2014)). It currently consists of 427,427 scaffolds and contigs. The 10,913 sparse SNPs mentioned above are distributed across 2,285 of the scaffolds and contigs, which account for 71.3% of available genome sequence. Sequence alignment and variant calling were then performed on this subset of the scaffolds and contigs. Using linkage information, the missing sequence data of sparsely genotyped individuals (offspring) will be determined.

The bwa mem algorithm (Li, (2013)) was used for pair ended alignment. After various BAM quality control and polishment with Picard tool set (http://picard.sourceforge.net (2014)), variants were screened with Samtools.

Imputation. Two software packages were chosen to be used for imputation, namely Beagle (Browning and Browning, (2009)) and LDMIP (Meuwissen and Goddard, (2010b)) to impute the offspring genotypes from sparse SNP data to sequence data. LDMIP can take the advantage of the pedigree information. A recent simulation (Yu et al,

(2014)) showed that LDMIP can significantly improve imputation rates if pedigree information is available.

GEBV estimation and cross-validation. Three methods were chosen to be used for GEBV estimation. They are GBLUP (Meuwissen et al., (2001)), MixP (Yu and Meuwissen, (2011)) and BayesB (Meuwissen et al., (2001)). GBLUP does not benefit much from higher marker density as it gives equal weights to all the markers, instead of giving extra weight to markers in high LD with genes. The MixP method fits a mixture of two normal distributions to the SNP effects, similar to BayesC (Meuwissen et al., (2009)), and thus attempts to give extra weight to important markers and little weight to the others, whilst keeping computation costs to a level comparable with GBLUP. The latter two methods were to be used for GEBV calculation with imputed dense genotypes.

The sequence data of the parents were obtained recently. Variant calling, results and its following imputation and subsequent GEBV calculation with imputed genotypes were being performed. The results presented herein were based on the 'sparse' 12K SNP array data. Two repeats of 10-fold cross-validation were conducted on both the vibriosis and VNN resistance datasets. In one repeat, the individuals were simply randomly grouped into 10-sets of similar size. The size difference of the 10 sets was at maximum 1. In the other cross-validation test, family effects were considered. Full sibs of each family were randomly ordered in a row. IDs in the same column were then assigned into the same set. These 10 sets then served as validation set one by one, with the remaining nine as training sets. The coefficients of determinant (r^2) between phenotypic observations and GEBV estimated were then calculated. The accuracy of the GEBV estimated was measured as r^2/h^2.

Results and Discussion

Table 1 and Table 2 show the cross validation results of the GEBV estimates for both disease resistance traits with two genomic selection methods. A much higher correlation for VNN resistance is obtained. With $h^2_{\text{vibriosis}} = 0.12$ and $h^2_{\text{VNN}} = 0.75$ the cross validation results show that the accuracy of the GEBV estimates were about 0.12 for vibriosis and 0.46 for VNN with the MixP method.

MixP is expected to have an advantage in GEBV estimation, especially when there are a few QTL of large effects segregating. On the other hand, if the accuracies between GBLUP and MixP show a significant difference, then there is a strong signal of segregating QTLs with large effects. These results are in agreement with previous studies (e.g., Baranski et al., (2010)). Different ways of fold assignments seem to have little effect on the correlation coefficients.

Table 1. Ten-fold cross validation of GEBV estimation of two disease resistance traits in Atlantic cod. Full sibs are allocated in different folds. Each fold served as validation set with the remaining as a training set. The coefficients of determinant (r^2) between GEBV and phenotypes were recorded.

Fold	Vibriosis		VNN	
	GBLUP	MixP	GBLUP	MixP
1	0.006	0.004	0.332	0.385
2	0.054	0.026	0.307	0.374
3	0.021	0.018	0.217	0.231
4	0.005	0.012	0.286	0.360
5	0.007	0.013	0.252	0.356
6	0.002	0.003	0.316	0.353
7	0.037	0.034	0.359	0.404
8	0.000	0.002	0.342	0.395
9	0.000	0.000	0.330	0.375
10	0.000	0.004	0.313	0.339
Mean	0.013	0.012	0.305	0.357

The variant calling procedure was applied to 111 sequenced individuals. All the 12K SNPs were identified, and the inferred genotypes were in agreement with the 12K SNP array genotypes. With a phred threshold of 70, about 2 million SNP were found for each individual. The total number of SNP loci will increase after the variant calling procedure is performed for all parents. This is because many homozygous loci in agreement with the reference are not detected within one individual, but will be detected when comparing more individuals.

According to Daetwyler et al., (2008); Meuwissen et al (2013), GEBV accuracy can be analytically determined with the equation below:

$$r^2 = \frac{\theta\beta}{\theta + 1 - h^2 r^2}$$

where, $\theta = \frac{Th^2\beta}{M_e}$, T is the size of the training set, and M_e is the effective number of segments in the genome.

$$M_e = \frac{2N_e G}{\log(2N_e)}$$

where N_e is the effective population size, G is the genome size in Morgan.

$$\beta = \frac{n}{n + M_e}$$

where n is the number of markers.

The current N_e is 133, with 50 sires and 100 dams. The cod genome size is 18.7M according to the current linkage map. The current sparse chips we used had 12K SNPs. The SNP calling procedure with the sequence data resulted in more than 10M SNPs.

The equation above containing item r^2 mentioned has the item on both sides. But it is handy to iterate to estimate r^2 with a program. It predicted that the r^2 between GEBV and phenotypes for VNN with 12K and 10M SNP markers were 0.45 and 0.53 respectively. The prediction from 12K SNP data fit the results in in Table 1 and 2 very well.

The predicted accuracies for vibriosis with 12k and 10M loci were 0.086 and 0.098.

Imputation of full sequence data and its use in genomic selection is currently under investigation. While NGS data provides a wealth of genetic information, it also presents major challenges with respect to data management and analysis. Huge digital storage is needed for these data, together with increased computation capacities. Advances of modern computer sciences have to address these challenges.

Table 2. Ten-fold cross validation of GEBV estimation of two disease resistance traits in Atlantic cod. Test sets were randomly assigned. Each fold served as validation set with the remaining as a training set. The coefficients of determinant (r^2) between GEBV and phenotypes were recorded.

Fold	Vibriosis		VNN	
	GBLUP	MixP	GBLUP	MixP
1	0.015	0.006	0.449	0.467
2	0.000	0.001	0.196	0.260
3	0.008	0.019	0.297	0.354
4	0.016	0.023	0.305	0.337
5	0.003	0.003	0.291	0.308
6	0.005	0.020	0.343	0.390
7	0.007	0.007	0.380	0.406
8	0.047	0.031	0.189	0.269
9	0.046	0.045	0.203	0.285
10	0.012	0.001	0.257	0.325
Mean	0.016	0.016	0.291	0.340

Conclusion

We have performed GEBV estimation in Atlantic cod for two disease resistance traits with data from a 12K SNP array, and found prediction accuracies of 0.29 and 0.69 for VNN and vibriosis, respectively. These accuracies were high relative to theoretical expectations. Further genomic evaluation of these traits is being performed with millions of additional loci derived from whole genome resequencing data. The genomic selection strategy is promising to predict GEBVs for selection candidates based on SNP effects estimated in challenge test experiments.

Literature Cited

Baranski M., Præbel A. K., Sommer A.-I. et al. (2010). WCGALP, Leipzig, Germany.

Browning B. L., Browning S. R. (2009). Am J Hum Genet 84:210-223.

Cod reference genome: ftp://ftp.ensembl.org/pub/release-74/fasta/gadus_morhua/dna (Accessed 2 April 2014).

Hayes B, Daetwyler H, Fries R, Stothard P, Pausch H, van Binsbergen R, Veerkamp R, Capitan A, Fritz S, Lund M, Boichard D, Van Tassell C, Guldbrandtsen B, Liao X, and the 1000 bull genomes consortium: The 1000 bull genomes project. 2013 http://www.1000bullgenomes.com (Accessed 2 April 2014)

Kettunen, A., Serenius, T., Fjalestad, K. T. (2007). J. Anim. Sci. 85:305-313.

Li H. (2013) [arXiv:1303.3997v2][12] [q-bio.GN].

Meuwissen T. H. E., Goddard M. (2010a). Genetics 185:623-631.

Meuwissen T. H. E., Goddard M. (2010b). Genetics 185:1441-1450.

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

Meuwissen T.H.E., Hayes B., Goddard M. (2013). Annu. Rev. Anim. Biosci. 1:221-237.

Meuwissen T. H. E., Solberg T. R., Shepherd R. K., Woolliams J. A. (2009) Genet Sel Evol 41:2.

Picard tools: http://picard.sourceforge.net (Accessed 2 April 2014)

Yu X. –J., Meuwissen T. H. E (2011). Genet. Sel. Evol., 43:35.

Yu X. –J., Woolliams J. A., Meuwissen T. H. E. (2014). Genet. Sel. Evol. (in press).

Ødegård J., Sommerb A. -I., and Præbelb A. K. (2010). Aquaculture 300:59-64.