Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 126-129

APPLYING MACHINE LEARNING TO UNLOCK GENOMIC INSIGHTS IN SNAPPER BREEDING

J. Blommaert[1] and M. Wellenreuther[1,2]

[1] The Institute of Plant and Food Research, Nelson, New Zealand
[2] School of Biological Sciences, University of Auckland, Auckland, New Zealand

SUMMARY

The Australasian snapper (*Chrysophrys auratus*) is a promising candidate for aquaculture. In the snapper breeding programme at Plant and Food Research, identifying genetic variants associated with key economic traits is essential for optimising selective breeding. Using machine learning for phenotyping and to analyse genomic data, our work focuses on detecting single nucleotide polymorphisms (SNPs) that influence growth traits within the breeding programme. Our approach and resulting knowledge has the potential to accelerate breeding programmes and provide more efficient ways to improve broodstock quality and sustainability in aquaculture.

INTRODUCTION

Traditionally, genomic-informed breeding programmes have focused on using approaches such as GWAS to identify single nucleotide polymorphisms (SNPs) as markers for trait-associated genetic variation (Dekkers 2012). While these approaches have been invaluable for advancing selective breeding, relying on these approaches can miss the more complex genetic architecture of traits such as growth (Chafai *et al*. 2023). Machine learning approaches accounting for interactions between variants have shown promise in other breeding programmes, despite the problem of high-dimensionality of genomic data (Feldner-Busztin *et al*. 2023). This kind of approach can facilitate the development of more robust and resilient aquaculture populations (Gill *et al*. 2022).

To diversify and build resilience in the aquaculture sector in New Zealand, the Australasian snapper (*Chrysophrys auratus*) has been the focus of a two decade long selective breeding program (Samuels *et al*. 2024). This breeding program has demonstrated substantial advancements in growth rates, survival, and feed conversion ratios, marking significant progress toward optimising snapper for aquaculture (Moran *et al*. 2023; Samuels *et al*. 2024).

Here, we apply computer vision and machine learning techniques to unlock genomic insights in snapper breeding. By cataloguing and integrating SNP variant data with extensive phenotypic data from 1,011 snapper in the F_4 cohort of a long-term selective breeding programme, we conducted genome-wide association studies (GWAS) to identify variants associated with growth traits. We also evaluated whether we could also identify genetic variants for genomic prediction, employing XGBoost as a predictive model. Our findings provide a valuable framework for enhancing genomic selection in aquaculture and offer insights into the potential applications of machine learning in selective breeding programmes, with implications for improving resilience and sustainability in aquaculture and beyond.

MATERIALS AND METHODS

Genotype and phenotype data. At 3 months post-hatch, fish were manually measured for weight and fork length and imaged. The images were used to measure a further 13 phenotypes via an in-house computer vision pipeline (Figure 1a). Genetic data was generated using a SNP chip designed for use in snapper (Montanari et al. 2023). Overall, 1,011 F_4 fish were included. After quality filtering, 11,006 SNPs were retained for analyses. For both the genotype and phenotype datasets, PCAs were performed using prcomp in R. PC1 of the phenotype data was also included as

a trait in the GWAS and genomic prediction. To infer family relationships, kmeans clustering was performed on a genetic relatedness matrix for all fish.

GWAS and genomic prediction. FarmCPU was used for GWAS within rMVP. For genomic prediction, we used XGBoost trained on 80% of the fish, using the relatedness cluster assignments in the stratification step. We included the first two genetic PCs as covariates in the models. ML analyses were applied to three phenotypes- condition factor, weight, and the distance from the top lip to the eye. We included condition factor since it is a derived trait which considers weight and length. We included weight because it is a common target for breeding programmes, and distance between eye and top lip since this yielded the highest number of significant SNPs from the GWAS. Each model was tuned using tidymodels and Root Mean Squared Error (rmse) and R^2 were recorded for both training and test data sets.

RESULTS AND DISCUSSION

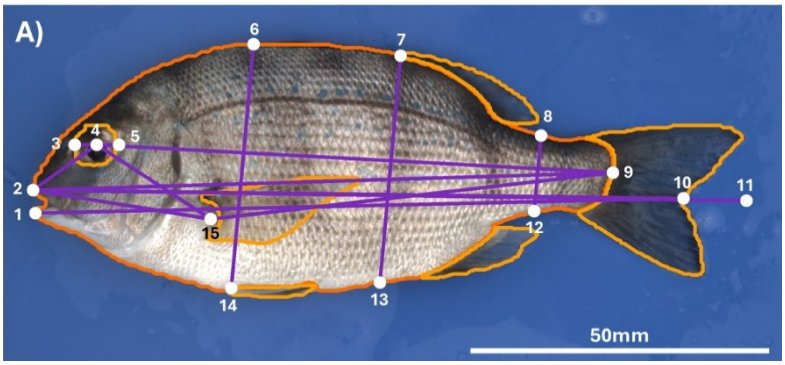
Across all fish included, weight ranged from 7.46g to 45.19g, while fork length ranged from 74.47mm to 123.03mm. Across the height measurements, the biggest variation was at the 75% of the length of the fish point, with 24.25mm difference between the minimum and maximum. Overall, traits were highly correlated with each other (Figure 1b), correlation coefficients ranged from 0.27 (eye width vs. distance between the caudal peduncle and pectoral joint) to 1 (fork length vs distances between each lip and the tail fork, and distance between top lip and tail fork vs. distance between bottom lip and tail fork). For the phenotypes, 96.96% of the variance was explained by PC1 of the phenotypes. For this reason, PC1 was used also included as predictive trait for the GWAS and XGBoost model.

Overall, 24 SNPs were identified by the GWAS (e.g. Figure 2) as being involved in the 15 phenotypes and phenotype PC1. Of these, 8 were shared between at least two traits, and 16 were unique to the trait they were identified in. Significant SNPs were across 14 chromosomes in the snapper genome and found to overlap with genes potentially involved in growth via metabolic pathways and even appetite signalling.

The machine learning models were found to be moderately accurate on the training data, but less so on the testing data (Table 1), suggesting that the high dimensionality of the SNP data is leading to overfitting. This was not addressed by reducing the number of input SNPs by random selection or by ranking SNPs by GWAS p-value (data not shown). Together, this suggests that the dataset is prone to overfitting and would benefit from the addition of additional fish samples.

Table 1. Residual mean squared error and R^2 statistics for XGBoost models on the test and train datasets for three selected traits

Trait	rmse train	R^2 train	rmse test	R^2 test
Condition factor	0.1	0.671	0.132	0.241
Weight	3.94	0.71	5.97	0.146
Top lip to eye (mm)	0.221	0.975	0.965	0.0675



b)

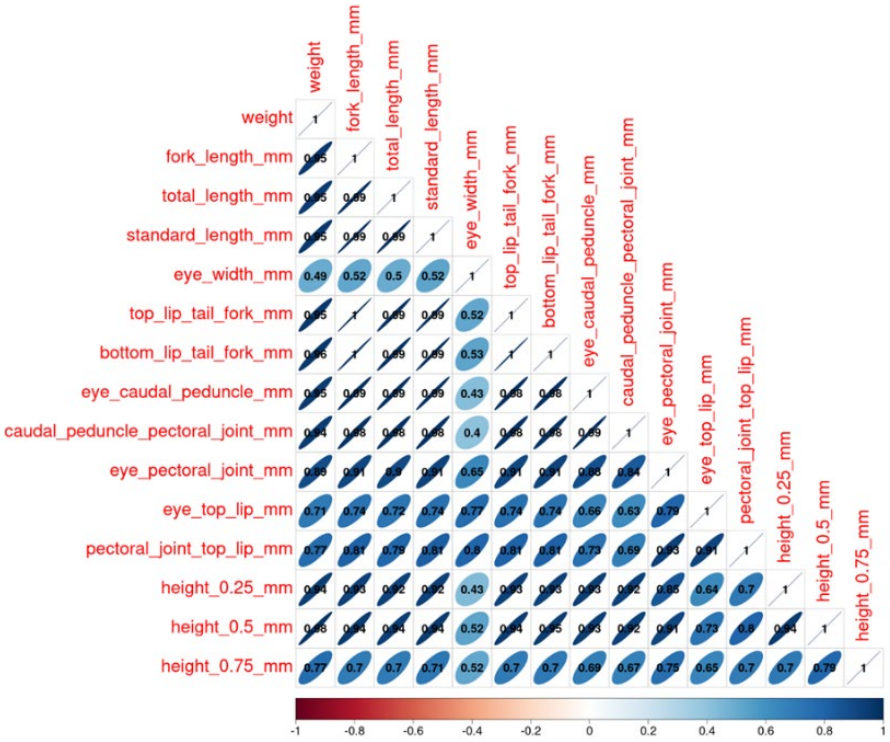


Figure 1. A) An output of the computer vision phenotyping pipeline showing the contours of fish body parts (orange) and landmarks (white points) used for each of the 13 measurements gathered by this pipeline (purple). The labelled points indicate landmarks used for measurements from the computer vision pipeline. Landmarks are 1- bottom lip, 2- top lip, 3 and 5- the left and right edges of the eye respectively, 4- centre of the eye, 6 and 14- top and bottom of the fish at 25% of its total length respectively, 7 and 13- top and bottom of the fish at 50% of its total length respectively, 8 and 12- top and bottom of the fish at 75% of its total length respectively, 9- peduncle, 10- tail fork, 11- total length end point. B) Correlation matrix of all 15 measured traits in this study. The colour and shape of each ellipse represents the R^2 value for that correlation (written in the corresponding box for each correlation). R^2 values were only included where p-values were < 0.05.

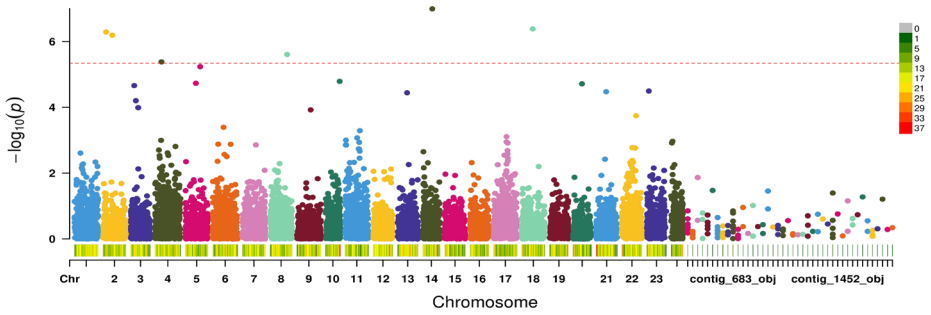


Figure 2. Manhattan plot showing SNPs significantly associated with weight in snapper. The heatmap along the bottom of the figure shows SNP density across the genome.

.

CONCLUSION

Including traditional genomic selection methods in our breeding program has already seen large gains in growth phenotypes, and while high dimensionality still hampers the machine learning approach for genomic prediction, applying more sophisticated statistical methods would allow us to consider the genetic, and eventually environmental interactions in complex traits. Future work will investigate similar analyses on other traits, such as internal organ measurements and look to increase the size of the data set used for the machine learning approach.

ACKNOWLEDGEMENTS

We thank the staff at Plant & Food Research for their contributions to breeding and rearing the snapper populations in this programme. This project was supported by the MBIE Endeavour Fund (C11X1603) under the 'Accelerated Breeding for Enhanced Seafood Production' programme and a Technology Development Fund, both awarded to MW. We also extend our thanks to Philipp Bayer from the Minderoo Foundation.

REFERENCES

Chafai N., Hayah I., Houaga I. and Badaoui B. (2023) *Front. Genet.* **14**: 1150596.

Dekkers J.C.M. (2012) *Curr. Genom.* **13**: 207.

Feldner-Busztin D., Nisantzis P.F., Edmunds S.J., Boza G., Racimo F., Gopalakrishnan S., Limborg M.T., Lahti L. and de Polavieja G.G. (2023). *Bioinformatics* **39**: btad021.

Gill M., Anderson R., Hu H., Bennamoun M., Petereit J., Valliyodan B., Nguyen H.T., Batley J., Bayer P.E. and Edwards D. (2022) *BMC Plant Biol.* **22**: 180.

Montanari S., Deng C., Koot E., *et al*. (2023). *G3-Genes Genom. Genet.* **13**: jkad170.

Moran, D., Schleyken J., Flammensbeck C., Fantham W., Ashton D. and Wellenreuther M. (2023) *Aquaculture* **563**: 738970.

Samuels G., Hegarty L., Fantham W., Ashton D., Blommaert J., Wylie M.J., Moran D. and Wellenreuther M. (2024) *Aquaculture* **586**: 740782.