bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Integrating genomics, machine learning, and computer vision to understand growth traits in selectively bred snapper (*Chrysophrys auratus*)

Julie Blommaert[1]*, Philipp E. Bayer[2, 3], David Ashton[1], Georgia Samuels[1], Linley Jesson[1], Maren Wellenreuther[1,4]

[1] The New Zealand Institute for Plant and Food Research Limited, Nelson, New Zealand

[2]Minderoo Foundation, Perth, Australia

[3]Minderoo OceanOmics Centre at UWA, Oceans Institute, The University of Western Australia, Crawley, Western Australia, Australia

[4] School Biological Sciences, The University of Auckland, Auckland, New Zealand

*Corresponding author

Email addresses:

JB: Julie.blommaert@plantandfood.co.nz

PEB: pbayer@minderoo.org

DA: David.ashton@plantandfood.co.nz

GS: Georgia.samuels@plantandfood.co.nz

LJ: Linley.jesson@plantandfood.co.nz

MW: Maren.wellenreuther@plantandfood.co.nz

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background: Understanding the genetic basis of growth-related traits is essential for optimising selective breeding programmes in aquaculture species. In this study, we analysed phenotypic and genomic data from a selectively bred population of snapper (*Chrysophrys auratus*) to identify genetic variants associated with key growth traits. We used a high-throughput, image-based phenotyping pipeline to extract 13 measurements rapidly and with minimal impact on the fish. These phenotypic measurements, together with manually measured weight and fork length, were analysed for correlations and principal component structure. Additionally, heritabilities were estimated for each trait. Then, genome-wide association studies (GWAS) were performed to identify growth-associated SNPs. To trial genomic prediction, we implemented machine learning (ML) models in XGBoost trained on SNP genotypes, with relatedness-based clustering used to minimise data leakage.

Results: Via GWAS, we identifying 24 SNPs significantly associated with growth traits, with several mapping to genes involved in metabolic and developmental pathways. Despite the high-dimensionality of these data, the ML approach still achieved moderate levels of predictability. The top ML growth SNPs showed some congruency with the GWAS growth SNPs, and 75 % of the GWAS SNPs were used by the ML model to predict weight. Functional annotation identified putative gene-level effects, providing insights into potential biological mechanisms underlying growth variation.

Conclusions: Our findings contribute to the development of genomic selection tools for snapper breeding and highlight the utility of integrating computer vision-based phenotyping with GWAS and ML for trait prediction in aquaculture species.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

Aquaculture breeding programmes urgently need to be expanded to improve food production for a growing global population [1-3]. Breeding can be used not only to enhance economically important traits in existing commercial species, improving production efficiency, but also to help develop new species for aquaculture [4-8]. This latter point is particularly relevant, as new species may better utilize available ocean farming space and be cultivated in regions where aquaculture currently plays a limited role [9]. These advancements hold promise for adding resilience to the sector and creating new economic opportunities in underserved regions.

The Australasian snapper (*Chrysophrys auratus*) has been the focus of research to diversify the aquaculture sector in New Zealand [10-14], with a selective breeding programme initiated in the 1990s [12]. This programme has produced an F_5 generation of selectively enhanced fish that demonstrate superior growth rates, survival, and food conversion ratios compared with wild snapper offspring. The first two selection rounds relied on domestication selection [15], while the F_3, F_4, and F_5 generations were developed using genomics-informed selective breeding to improve growth. An ongoing challenge is improving the selection of highly polygenic traits, such as growth, which exhibit strong positive allometry with traits such as length and weight [16, 17].

With the decreasing costs of genome-wide markers and the establishment of robust workflows for managing genomic data and associated bioinformatics pipelines, genetic improvement in aquaculture has advanced from conventional breeding to marker-assisted selection, genomic selection, or a combination of both [16]. However, genomic data is no longer the limiting resource in genetic improvement. Instead, the primary bottleneck lies in the lack of high-throughput and accurate phenotyping methods for aquaculture species such

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as fish. While some phenomics platforms have been developed for the precise, rapid, and non-invasive measurement of growth traits, platforms for other critical traits—such as disease resistance, stress tolerance, and behaviour—remain significantly underdeveloped [18, 19].

Recent advances in novel technologies, including automated imaging and computer vision, diode frame measurements, and deep learning networks, hold promise for addressing this limitation [20-22]. These innovations can facilitate the development of comprehensive phenomics platforms, encompassing phenotyping, data acquisition, and processing, which are critical for better selection of individuals with desirable traits in aquaculture breeding programmes. These phenomics platforms have been integrated in some species [23], but these methods are still under development, particularly for new aquaculture species. Notable examples of automated phenotyping platforms that have been successfully integrated into breeding programmes include rice [24] and bivalves [25]

In this study, we employ novel approaches integrating whole-genome information with computer vision-assisted phenotyping to identify genomic markers that could predict growth in future generations. To achieve this, we first develop a high-throughput phenotyping pipeline using deep learning models trained on morphological traits captured via computer vision. This allows for precise, automated measurements of growth dynamics across multiple developmental stages. Next, we compare genome-wide association studies (GWAS) with machine learning approaches to assess the predictive power of individual loci versus polygenic models and identify important variants [26]. By combining these methodologies, we aim to refine the selection process and accelerate genetic gains in snapper breeding, ultimately improving aquaculture productivity and sustainability.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Fish rearing and genotyping

The F_4 snapper were generated in November 2021 from the third generation of selectively bred snapper. Fertilized eggs were collected and incubated over a period of 5 days. Larval rearing protocols followed standard methods developed for this species (Samuels et al. 2024). Land-based on-growing of Australasian snapper (*Chrysophrys auratus*) was conducted at The New Zealand Institute for Plant and Food Research Limited's (PFR) Research Facility, located in Nelson, New Zealand (41.2985° S, 173.2441° E). This facility is equipped with a flow-through tank system, where water from the Nelson Haven is withdrawn from an engineered aquifer in the intertidal zone filled with various hard substrates that provide filtration. All work was done under the conditions of animal ethics application AEC-2021-PFR-05, approved by the Animal Ethics Committee at the Nelson Marlborough Institute of Technology (Te Pūkenga), along with fish-farm licence FW208.03.

Previous work [13] has reported the SNP data for this cohort of fish, where 1103 F_4 individuals were genotyped.

Manual and image-based phenotyping

Prior to any measurements, snapper were anaesthetised using AQUI-S® (15–20 ppm), a dose that resulted in a loss of equilibrium and no reaction to net capture. Fish were manually weighed (FX-300i WP scales from A&D Company Limited) and measured at three months of age. Phenotypic measurements (Fig. 1) were made with an in-house computer vision pipeline [12].

[**Figure 1 here]**

Data processing and statistical analyses

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phenotypic data were not available for 98 of the genotyped fish. In total, 1011 individual fish were included in this study once the genotyping and phenotyping data were combined. The measured phenotypes were assessed for correlation using the R package corrplot v 0.92. In addition to the directly measured traits, derived phenotypes (condition factor (K) and principal components (PCs)) were computed for inclusion in downstream analyses. K was calculated using the Fulton formula (K = 100 * weight / length^3). PC analysis was completed using the measured phenotypes and prcomp in R.

SNP genotype data were obtained as described in Montanari et al. [13]. SNP positions were remapped from the original snapper reference genome to the updated snapper genome assembly [27] using snplift v 1.0.4. SNPs that could not be lifted over were further processed by extracting a genomic region of 70 bp surrounding each SNP from the original assembly and performing a BLAST search against the new genome assembly. SNPs with multiple mapping locations in the updated genome were excluded, as were SNPs that did not align to the expected chromosome.

In order to stratify test and training groups based on genetic relatedness, relatedness clusters were generated based on the similarity of SNP genotype data. Pairwise relatedness coefficients were calculated in R by first centering the SNP matrix by subtracting the column-wise mean genotype value for each SNP across all individuals. The resulting mean-centres genotypes were used to compute the genomic related matrix (GRM) as the dot product of the centred matrix and its transpose divided by the total number of SNPs. Clustering was then performed using k-means clustering with k = 19. The resulting clusters were visualized using principal component analysis (PCA) to confirm concordance between the *k*-means clusters and the genetic structure inferred from PCA. Heritability estimates were calculated for each measured trait using ASREML v4.2 (Butler, Cullis et al. 2023) using a mixed-effects model

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with fish id as a random effect. Heritability and its standard error were obtained using the vpredict function, based on the ratio of the variance components.

GWAS and machine learning

Three methods (general linear model or GLM, mixed linear model or MLM, and fixed and random model circulating probability unification or FarmCPU) to perform a GWAS were applied using rMVP v1.1.1 [28] for all traits, including condition factor (K) and the phenotypic PC1 (explaining 96.96% of phenotypic variability), and the first two genomic PCs were used as covariates. Downstream analyses were performed only using the SNPs identified by the FarmCPU method, which improves power and reduces false positives by iteratively applying a fixed-effect model to test SNPs while using a random-effect model to control for confounding factors.

Genomic prediction analyses were performed using the *XGBoost* [29] algorithm in the R package xgboost v1.7.8.1. Models were tuned using tidymodels v1.2.0 with 5-fold cross-validation. Performance was evaluated using root mean square error (rmse) as a metric. To minimise potential data leakage due to genetic relatedness, the dataset was partitioned into training (80%) and testing (20%) subsets based on the relatedness clusters identified above. Principal components 1 and 2 of the SNP dataset were included as covariates to account for genetic relatedness in the predictive models. To avoid overfitting with too many features, a model was also run for weight using a subset of the SNPs, namely the top 500 SNPs when ranked using the p values from the FarmCPU GWAS. The *XGBoost* importance matrix was extracted to assess the number of features used by each model and for comparison to the SNPs identified as significantly associated with each trait. The top 10 SNPs in each XGBoost model, as determined by gain, were extracted for further functional analyses.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SNP functional annotation was performed via snpEff v 5.2 [30] using the genome assembly and snapper specific annotations from previous work [27]. Functional profiling was performed with g:GOSt on the g:profiler website [31]. Other functional information was gathered from the Zebrafish Information Network (ZFIN) [32].

Results

Phenotypic and genotypic variation

Summary statistics for each of the 15 measured phenotypes across the 1011 fish can be found in Table 1. 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 75% of the length of the fish, with 24.25mm difference between the minimum and maximum. Overall, traits were highly positively correlated with each other (Fig. 2), and 96.96% of the phenotypic variance was explained by PC1 of the phenotypes. Correlation coefficients ranged from 0.27 (eye width vs. distance between the caudal peduncle and pectoral joint) to 1 (fork length versus distances between each lip and the tail fork, and distance between top lip and tail and distance between bottom lip and tail fork) (Fig. 2).

[Figure 2 here]

Table 1. Phenotypic measurements.

Measurement	Landmarks	min	mean	max	h^2	SE
weight_g	-	7.46	25.46	45.19	0.38	0.05
fork_length_mm	-	74.47	104.68	123.03	0.33	0.05
total_length_mm	1-11	80.10	112.61	132.73	0.34	0.05
standard_length_mm	2-9	66.93	95.31	112.27	0.33	0.05
eye_width_mm	3-5	5.83	7.66	9.16	0.20	0.05
top_lip_tail_fork_mm	2-10	73.50	103.60	121.70	0.34	0.05
bottom_lip_tail_fork_mm	1-10	73.40	103.55	122.03	0.34	0.05
eye_caudal_peduncle_mm	4-9	53.97	79.75	94.50	0.33	0.05
caudal_peduncle_pectoral_joint_mm	9-15	41.03	62.83	75.50	0.33	0.05

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

eye_pectoral_joint_mm	4-15	15.70	21.14	25.37	0.31	0.05
eye_top_lip_mm	2-4	8.87	12.04	15.77	0.24	0.05
pectoral_joint_top_lip_mm	2-15	20.83	27.60	33.87	0.25	0.05
height_0.25_mm	6-14	23.70	36.68	46.10	0.33	0.05
height_0.5_mm	7-13	25.53	37.56	46.77	0.38	0.05
height_0.75_mm	8-12	0.60	13.65	24.85	0.26	0.05

A summary of each measured phenotype including the landmarks the measurements were taken from in the images as indicated in Figure 1, minimum, mean, maximum, estimated heritability (h^2), and standard error of the estimated heritability (SE) of each measurement across the 1011 fish. Weight is in grams, and all other measurements in mm. Weight and fork length were measured manually, and all other measurements from the computer vision pipeline.

In total, after QC and SNP lift over, 11,006 SNPs were included in downstream analyses. Genetic relatedness corresponded to the clusters on the PCA plot, and PC1 explained 7.55% of the variance in the SNPs, while PC2 explained 5.63% (Fig. 3). Estimated heritabilities ranged from 0.201 (eye_width_mm) to 0.3792 (height_0.5_mm). Notably, the estimated heritability of weight was also moderately high at 0.3775 (Table 1).

[Figure 3 here]

Identifying growth-related SNPs using GWAS

[Figure 4 here]

Of the three GWAS methods included in rMVP, MLM identified no associated SNPs with any phenotypes, GLM showed spurious p-value inflation for almost all p-values [See Additional file 1, Figure S1], and the number of SNPs identified by FarmCPU (Fig. 4) as being significantly associated with a trait ranged from two (caudal_peduncle_pectoral_joint_mm) to eight (eye_top_lip_mm). For PC1 of the phenotype, seven significant SNPs were identified. Across all phenotypes (including phenotypic PC1 and K), 24 SNPs were found to be associated, with 16 being unique to the trait they were identified in, and eight shared across at least two phenotypes [See Additional file 2, Table S1]. Significant SNPs were distributed across 14 of the 24 chromosomes in the snapper genome with the highest number of SNPs (three each) on chromosomes 2 and 3 where each

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

chromosome harboured three significant SNPs (Fig. 5). The 24 SNPs were found to affect 10 genes with modifier effects [See Additional file 2, Table S1]. Of these genes, four (OSBPL3B, PDE4DIP, ZDHHC3A, LEAP2) were found to be targets of transcription factor *p53*, and six (CLSTN2A, OSBPL3B, PDE4DIP, SOAT1, TPST1, ZDHHC3A) were found to be targets of the transcription factor *Sox-10*. The two SNPs that were significant across the most phenotypes were found to affect genes interacting with *Sox-10*. The other genes affected by these SNPs were an ATP synthase (atp5pf), a glutamate receptor (grik5), and a potassium voltage-gated channel (kcna7).

[Figure 5 here]

Identifying growth-related SNPs using machine learning

The machine learning (ML) approach to identifying growth-related SNPs was limited to five traits (weight, fork length, condition factor, distance between eye and top lip, and PC1 of all measured phenotypes). Weight and fork length were included as common production targets for breeding programmes, condition factor was included as a derived trait combining weight and length information, and distance between the eye and top lip was also included since this had the most SNPs significantly associated in the GWAS, and PC1 of the phenotypes as a combined metric of all measured phenotypes. R^2 values of the training data (80%) ranged from 0.671 to 0.975, while the same measure in test data (20%) ranged between 0.0675 and 0.283 (Table 2). In total, 29 different SNPs were identified in the top 10 of each of the three directly measured traits, and one SNP was shared between the models for weight and condition factor. These 29 SNPs were distributed across 18 chromosomes and 1 scaffold, with the most SNPs per chromosome being four each on chromosomes 2 and 12 (Fig. 5). These 29 SNPs were found to impact 24 genes, but no significant Gene Ontology (GO) terms or other interactions were identified [See Additional file 2, Table S1]. There was no direct overlap

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between the SNPs identified via GWAS and those identified here, however one GWAS SNP and one machine learning SNP were within 0.5 Mbp of each other on chromosome 2, and another ML SNP on chromosome 3 was within 0.5 Mbp of a QTL identified as being involved in weight in previous work [27]. Additionally, another ML SNP on chromosome 6 was within 1 Mbp of a QTL identified in previous work. Additional clusters of 3 ML only SNPs were within 1Mbp of each other on chromosome 12 (Fig. 5). However, of the 24 growth-associated GWAS SNPs, four were used directly by the ML model for weight, nine were directly neighbouring SNPs used by the same ML model, and a further five were within five SNPs of one used by the weight ML model. The remainer (six SNPs) were further than five SNPs away from any SNP in the weight model feature list.

Table 2. Machine learning summary statistics.

Trait	Root mean squared error (test)	R^2 test	Root mean squared error (train)	R^2 train	Features used by model
K	0.132	0.241	0.1	0.671	2607
weight	5.97	0.146	3.94	0.71	2739
eye_top_lip	0.965	0.0675	0.221	0.975	5103
weight*	5.46	0.283	1.85	0.930	394

Summary statistics and number of features used by the XGBoost model in training (80%) and testing (20%) data for condition factor (K), weight, and distance between the eye and top lip. The weight* represents an XGBoost model using only the top 500 most significant SNPs as identified by a GWAS for weight.

Discussion

Selecting individuals for breeding programmes to enhance food production sustainability and improve animal welfare is a global priority [2, 33]. Traditionally, selection has relied solely on phenotypic data to identify desirable individuals carrying traits of interest. However, the advent of molecular tools, and more recently, genomic technologies, has significantly improved the ability to capture the genetic basis of selection traits [16]. This has enhanced the precision of selection and enabled genetic improvement before animals reach maturity or express the target traits [7, 34]. In parallel, advances in machine learning computer vision and

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

AI technology have facilitated the digital extraction of phenotypic data from animals and plants, improving the accuracy of measuring economically important production and welfare traits [18, 19]. Integrating these high-dimensional datasets remains a complex challenge, and recent efforts have focused on combining dense genomic datasets with multiple phenotypic traits to optimize breeding selection, with some notable examples [24, 35]. In this study, we apply computer vision techniques to extract individual trait information from images of an F_4-selected line of Australasian snapper (*Chrysophrys auratus*). We integrate this phenotypic data with genome-wide SNP chip data to identify genomic regions associated with growth, thereby improving selection accuracy in breeding programmes. Our study highlights the relatively high heritability of growth traits, strong correlations among measured phenotypes, and the significant genetic components underlying these traits. The high phenotypic variance explained by PC1 underscores the interconnectedness of growth-related traits, while the identification of 24 significant SNPs, many of which were significant for multiple traits, across multiple chromosomes emphasizes the polygenic nature of these growth-related traits. Notably, the overlap of SNPs with genes linked to metabolic pathways and appetite signalling suggests a biological basis for growth differences that could inform future breeding strategies.

This work highlights the predicted impact that high throughput phenotyping can have on breeding programmes [19]. If the GWAS could only focus on weight and fork length as measured manually, only six SNPs would be identified as being important in growth traits. However, when all traits were analysed, 22 SNPs could be identified. Additionally, when all traits were combined into a single trait via a PCA approach, two additional growth associated SNPs could be identified, bringing the total to 24 associated SNPs. Improved digital phenotyping via computer vision enables the inclusion of more traits, while also requiring less handling effort from staff and less stress for the fish being measured [18]. There is strong

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

potential for this method to be expanded to underwater phenotyping, rather than relying on benchtop images. Additionally, other approaches to reducing phenotypic dimensionality, such as clustering by overall body shape, could be explored in future. Doing so would lead to greater understanding of commercially important fish phenotypes beyond growth directly. This could be linked to other phenotypes, such as fillet size and fat composition, and would further increase the understanding of the genetic architecture underlying fish phenotypes independently of their relationship to fish size.

Growth-related traits are typically considered highly-polygenic, and even in studies with large sample size and high SNP numbers, these traits are often still poorly predicted [36]. The identification, however, of SNPs with impacts on genes involved in *p53* and *Sox10* pathways does suggest plausible biological mechanisms for these SNPs to influence growth traits in snapper. Although *p53* has mostly been examined in a stress response role in species relevant to aquaculture [37, 38], it has been successfully targeted for knockout in pig myoblasts for cultured meat [39] and has also been implicated in muscle development in quail [40]. Despite the broad cellular functions of *p53*, it is therefore plausible that it is involved in snapper growth via indirect routes related to stress response, or more directly involved in muscle cell growth. Meanwhile, Sox10 also has broad biological activity across the neural crest [41], but could influence snapper growth through its involvement in the enteric nervous system [42, 43]. Similarly, some of the genes impacted by growth-associated SNPs identified via ML approaches have plausible biological links to growth, despite the lack of significant GO terms. Of note, col1a1b encodes for a subunit of collagen, an essential component of animal tissues that has been found to be highly expressed in zebrafish muscle [44]. Additionally, mad2l1bp is involved in regulating mitosis and has also been found to be in the *p53* network as well [45]. Several genes (e.g. nos1, csf3r, cst, mad2l1bp, gimap4, mrps15) are implicated in stress and immune responses in fish and other species [46-52]. While the involvement of

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

these two pathways and other biological mechanisms in growth directly cannot be confirmed with our current data, this highlights the complex and polygenic nature of growth related traits [17] and provides further candidate genes and variants for functional studies investigating growth in snapper or related species. This complex and multi-pathway architecture of growth-related traits is consistent with studies of similar traits, with thousands of SNPs implicated to explain similar levels of heritability of height in humans in even the most powerful of studies [36].

This work has deepened our understanding of growth-related traits in snapper. The heritabilities reported here are comparable to, but slightly higher than those previously reported in this species [11]. The increased estimates could be explained by a number of factors, including the generation of the fish included, the age at which the measurements were taken, and the changes in phenotypic and genetic diversity through the breeding programme [12].

While machine learning approaches such as XGBoost demonstrated strong predictive ability in the training dataset, the lower R^2 values observed in the test dataset highlight challenges in accurately capturing complex phenotypic traits. These limitations may stem from factors such as sample size constraints and the inherent complexity of the traits under selection. Despite this, we did see concordance between the regions identified in machine learning approaches, the GWAS herein, and previous studies in snapper [10, 27, 53]. Additionally, other studies have noted marked improvement in understanding the genetic basis of complex traits by increasing genomic marker panels included in these types of analyses [54]. Future studies should aim to expand sample sizes, and the number of genetic markers analysed to enhance model robustness and improve predictive accuracy. Despite these challenges, our findings provide a foundation for refining genomic prediction models and incorporating novel phenotypic datasets to improve trait predictability and selection efficiency in breeding

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

programmes. We suggest that future work should also continue to focus on the integration of multiple approaches—including computer vision, traditional genetic analyses, and machine learning for genomic prediction—to explore additional avenues of improving selection strategies. However, in all of this, overcoming the high dimensionality of genomic and phenomic data remains a key obstacle.

Conclusions

In conclusion, our study demonstrates that even with relatively small sample sizes, meaningful insights can be gained to support decision-making in breeding programmes. Growth is a well-known polygenic trait, and even though genomic dissections have been challenging because of the low effect sizes of most genetic variants [17], our study is in line with other work showing that careful design allows the discovery of novel polygenic variants [55, 56]. Future research may explore deep learning methods, such as neural networks, which have shown improved predictive power. However, their application must be balanced against the need for interpretability, as these models often lack transparency in identifying causal genetic relationships.

Declarations

Ethics approval and consent to participate

All work was done under the conditions of animal ethics application AEC-2021-PFR-05, approved by the Animal Ethics Committee at the Nelson Marlborough Institute of Technology (Te Pūkenga), along with fish-farm licence FW208.03.

Consent for publication

Not applicable

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability of data and materials

The genotyping data used in this work are already published [13] as is the genome assembly [27]. Both these datasets are available with permission from representatives of Māori iwi (tribes). Guardianship of these datasets are managed by the Aotearoa Genomic Data Repository [57].

Competing Interests

The authors declare they have no competing interests.

Funding

This project was supported by the New Zealand Ministry of Business, Innovation & Employment (MBIE) Endeavour Fund (C11X1603) ‘Accelerated Breeding for Enhanced Seafood Production’ and a Technology Development Fund from The New Zealand Institute for Plant and Food Research Limited. It also benefited from funding through the MBIE Data Science Platform Grant, *A Data-Science Driven Evolution of Aquaculture for Building the Blue Economy*, and an Early Career Researcher Travel Grant from Genomics Aotearoa.

Authors’ contributions

JB analysed and interpreted the data and was a major contributor to writing the manuscript. PEB analysed and interpreted the data and contributed to writing the manuscript. DA and GS provided the pipeline for analysing the phenotypic data. LJ contributed to data analysis. MW was a major contributor to writing the manuscript. All authors read and approved the final manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

We thank all the staff of The New Zealand Institute for Plant and Food Research Limited who were involved in breeding and rearing the snapper populations that formed part of this breeding programme.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Lubchenco J, Haugan PM, Pangestu ME. Five priorities for a sustainable ocean economy. Nature. 2020;588:30-2.
2. Fong CR, Gonzales CM, Rennick M, Gardner LD, Halpern BS, Froehlich HE. Global yield from aquaculture systems. Rev Aquac. 2024;16:1021-9.
3. Garlock T, Asche F, Anderson J, Bjørndal T, Kumar G, Lorenzen K, et al. A global blue revolution: aquaculture growth across regions, species, and countries. Rev Fish Sci Aquac. 2020;28:107-16.
4. Gjedrem T, Robinson N, Rye M. The importance of selective breeding in aquaculture to meet future demands for animal protein: A review. Aquaculture. 2012;350–353:117-29.
5. Gjedrem T, Robinson N. Advances by selective breeding for aquatic species: a review. Agri Sci. 2014;5:1152.
6. Gjedrem T. Selection and breeding programs in aquaculture. 2005 ed: Springer; 2005.
7. Song H, Dong T, Yan X, Wang W, Tian Z, Sun A, et al. Genomic selection and its research progress in aquaculture breeding. Rev Aquac. 2023;15: 274-91.
8. Yáñez JM, Barría A, López ME, Moen T, Garcia BF, Yoshida GM, et al. Genome-wide association and genomic selection in aquaculture. Rev Aquac. 2023;15:645-75.
9. Cai J, Chan HL, Yan X, Leung P. A global assessment of species diversification in aquaculture. Aquaculture. 2023;576:739837.
10. Ashton DT, Ritchie PA, Wellenreuther M. High-density linkage map and QTLs for growth in snapper (*Chrysophrys auratus*). G3. 2019;9:1027-35.
11. Ashton DT, Hilario E, Jaksons P, Ritchie PA, Wellenreuther M. Genetic diversity and heritability of economically important traits in captive Australasian snapper (*Chrysophrys auratus*). Aquaculture. 2019;505:190-8.
12. Samuels G, Hegarty L, Fantham W, Ashton D, Blommaert J, Wylie MJ, et al. Generational breeding gains in a new species for aquaculture, the Australasian snapper (*Chrysophrys auratus*). Aquaculture. 2024:740782.
13. Montanari S, Jibran R, Deng C, David C, Koot E, Kirk C, et al. A multi-species SNP chip enables diverse breeding and managment application. G3. 2023;13:jkad170.
14. Moran D, Schleyken J, Flammensbeck C, Fantham W, Ashton D, Wellenreuther M. Enhanced survival and growth in the selectively bred *Chrysophrys auratus* (Australasian snapper, tāmure). Aquaculture. 2023;563:738970.
15. Baesjou JP, Wellenreuther M. Genetic signatures of domestication selection in the Australasian snapper (*Chrysophrys auratus*). Genes. 2021;12:1737.
16. Houston RD, Bean TP, Macqueen DJ, Gundappa MK, Jin YH, Jenkins TL, et al. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat Rev Genet. 2020;21:389-409.
17. Wellenreuther M, Hansson B. Detecting polygenic evolution: problems, pitfalls, and promises. TIG. 2016;32:155–64.
18. Fu G, Yuna Y. Phenotyping and phenomics in aquaculture breeding. Aquac Fish. 2022;7:140-6.
19. Houle D. Numbering the hairs on our heads: The shared challenge and promise of phenomics. PNAS. 2010;107:1793-9.
20. Babu KM, Ashton DT, Bentall D, Lin HT, Tuckey NPL, Wellenreuther M, et al. Computer vision in aquaculture: A case study of juvenile fish-counting. J Roy Soc New Zealand. 2022;53:52-68.
21. Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S. Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. TAG. 2013;126:867-87.
22. Zhao S, Zhang S, Liu J, Wang H, Zhu J, Li D, et al. Application of machine learning in intelligent fish aquaculture: A review. Aquaculture. 2021;540:736724.
23. Chafai N, Hayah I, Houaga I, Badaoui B. A review of machine learning models applied to genomic prediction in animal breeding. Front Gen. 2023;14:1150596.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24. Yang W, Guo Z, Huang C, Duan L, Chen G, Jiang N, et al. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. Nature Comm. 2014;5:5087.

25. Bassim S, Chapman RW, Tanguy A, Moraga D, Tremblay R. Predicting growth and mortality of bivalve larvae using gene expression and supervised machine learning. Comp Biochem Physiol Part D Genomics Proteomics. 2015;16:59-72.

26. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, et al. Machine learning in genome-wide association studies. Genet Epidemiol. 2009;33:S51-S7.

27. Blommaert J, Sandoval-Castillo J, Beheregaray L, Wellenreuther M. Peering into the gaps: Long-read sequencing illuminates structural variants and genomic evolution in the Australasian snapper. Genomics. 2024:110929.

28. Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, et al. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genom Proteom Bioinform. 2021;19:619-28.

29. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. R package version 04-2. 2015;1:1-4.

30. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6:80-92.

31. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51:W207-W12.

32. Bradford YM, Van Slyke CE, Ruzicka L, Singer A, Eagle A, Fashena D, et al. Zebrafish information network, the knowledgebase for Danio rerio research. Genet. 2022;220:iyac016.

33. Boyd CE, McNevin AA, Davis RP. The contribution of fisheries and aquaculture to the global protein supply. Food Sec. 2022;14:805-27.

34. Yáñez JM, Xu P, Carvalheiro R, Hayes B. Genomics applied to livestock and aquaculture breeding. Evol App. 2022;15:517-22.

35. SF C, PM V. Association mapping in outbred populations: power and efficiency when genotyping parents and phenotyping progeny. Genet. 2009;181:755.

36. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, et al. A saturated map of common genetic variants associated with human height. Nature. 2022;610:704-12.

37. Song Y, Salbu B, Heier LS, Teien H-C, Lind O-C, Oughton D, et al. Early stress responses in Atlantic salmon (*Salmo salar*) exposed to environmentally relevant concentrations of uranium. Aquat Toxicol. 2012;112-113:62-71.

38. Qian B, Qi X, Bai Y, Wu Y. The p53 signaling pathway of the large yellow croaker (Larimichthys crocea) responds to acute cold stress: evidence via spatiotemporal expression analysis of p53, p21, MDM2, IGF-1, Gadd45, Fas, and Akt. PeerJ. 2020;8:e10532.

39. Srila W, Pangjantuk A, Kunhorm P, Chaicharoenaudomrung N, Noisa P. Development of CRISRP/Cas9-based TP53-knockout pig muscle stem cells for use in the cultured meat industry. 3 Biotech. 2025;15:92.

40. Park J-W, Lee JH, Han JS, Shin SP, Park TS. Muscle differentiation induced by p53 signaling pathway-related genes in myostatin-knockout quail myoblasts. Mol Biol Rep. 2020;47:9531-40.

41. Hu Y, Wang B, Du H. A review on sox genes in fish. Rev Aquac. 2021;13:1986-2003.

42. Kunke M, Kaehler M, Boni S, Schröder K, Weier A, Chunder R, et al. SOX10-mediated regulation of enteric glial phenotype in vitro and its relevance for neuroinflammatory disorders. J Mol Neurosci. 2025;75:26.

43. Bondurand N, Sham MH. The role of SOX10 during enteric nervous system development. Dev Biol. 2013;382:330-43.

44. Lin F, Ye X, Lin J, Liu X, Yuan Y, Guo H, et al. Comparative transcriptome analysis between muscle and swim bladder reveals key genes regulating collagen deposition in zebrafish. Aquac Rep. 2022;23:101053.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

45. Le MTN, Shyh-Chang N, Khaw SL, Chin L, Teh C, Tay J, et al. Conserved regulation of p53 network dosage by microRNA–125b occurs through evolving miRNA–target gene pairs. PLos Genet. 2011;7:e1002242.

46. Dietrich MA, Irnazarow I, Inglot M, Adamek M, Jurecka P, Steinhagen D, et al. Hormonal stimulation of carp is accompanied by changes in seminal plasma proteins associated with the immune and stress responses. J Proteomics. 2019;202:103369.

47. Cioni C, Angiulli E, Toni M. Nitric oxide and the neuroendocrine control of the osmotic stress response in teleosts. Int J Mol Sci. 2019;20:489.

48. Faiza B, Rasighaemi P, Liongue C, Ward AC. Zebrafish granulocyte colony-stimulating factor receptor maintains neutrophil number and function throughout the life span. Infect Immun. 2019;87:10.1128/iai.00793-18.

49. Gerber L, Jensen FB, Madsen SS. Dynamic changes in nitric oxide synthase expression are involved in seawater acclimation of rainbow trout *Oncorhynchus mykiss*. Am J Physiol Regul Integr Comp Physiol. 2017;314:R552-R62.

50. Zhang Y, Wen H, Liu Y, Qi X, Sun D, Zhang C, et al. Gill histological and transcriptomic analysis provides insights into the response of spotted sea bass (*Lateolabrax maculatus*) to alkalinity stress. Aquaculture. 2023;563:738945.

51. Limoges M-A, Cloutier M, Nandi M, Ilangumaran S, Ramanathan S. The GIMAP family proteins: an incomplete puzzle. Front Immun. 2021;Volume 12 - 2021.

52. David F, Roussel E, Froment C, Draia-Nicolau T, Pujol F, Burlet-Schiltz O, et al. Mitochondrial ribosomal protein MRPS15 is a component of cytosolic ribosomes and regulates translation in stressed cardiomyocytes. Int J Mol Sci. 2024;25:3250.

53. Ruigrok M, Xue B, Catanach A, Zhang M, Jesson L, Davy M, et al. The relative power of structural genomic variation versus SNPs in explaining the quantitative trait growth in the marine teleost *Chrysophrys auratus*. Genes. 2022;13:1129.

54. Yuan C, Gillon A, Gualdrón Duarte JL, Takeda H, Coppieters W, Georges M, et al. Evaluation of genomic selection models using whole genome sequence data and functional annotation in Belgian Blue cattle. Genet Sel Evol. 2025;57:10.

55. Sinclair-Waters M, Ødegård J, Korsvoll SA, Moen T, Lien S, Primmer CR, et al. Beyond large-effect loci: large-scale GWAS reveals a mixed large-effect and polygenic architecture for age at maturity of Atlantic salmon. Genet Sel Evol. 2020;52:9.

56. Debes PV, Piavchenko N, Ruokolainen A, Ovaskainen O, Moustakas-Verho JE, Parre N, et al. Polygenic and major-locus contributions to sexual maturation timing in Atlantic salmon. Mol Ecol. 2021;30: 4505-19.

57. Te Aika B, Liggins L, Rye C, Perkins EO, Huh J, Brauning R, et al. Aotearoa genomic data repository: An āhuru mōwai for taonga species sequencing data. Mol Ecol Resour. 2025;25:e13866.

bioRxiv preprint doi: https://doi.org/10.1101/2025.05.29.656727; this version posted May 29, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

Figure 1 An example output of the computer vision phenotyping.

The contours of fish body parts (orange) and landmarks (white points) used for each of the 13 measurements gathered by this pipeline (purple). The white line indicates a scale bar of 50mm. 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.

Figure 2 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 3 Principal component (PC) clustering based on the SNP chip genotypes.

Each point represents an individual fish (n=1011). Clusters are coloured based on relatedness clusters determined by *k*-means clustering (k=19) of a genetic relatedness matrix.

Figure 4 GWAS results from FarmCPU within rMVP for weight at 3 months of age.

Chromosomes shown on the y-axis, with SNP density in a heat map below the Manhattan plot of log(p values). The significance threshold (log(p) = 5.34) is shown as a red dashed line.

Figure 5 The distribution of SNPs and QTLs across the snapper genome involved in growth phenotypes.

QTLs identified in previous work are shown in teal, and SNPs identified in this study via GWAS in purple and machine learning in red.

Additional files

Additional file 1 Figure S1

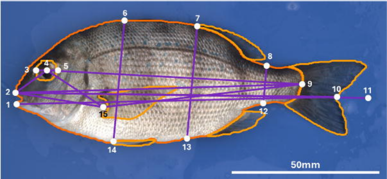
Format: DOCX

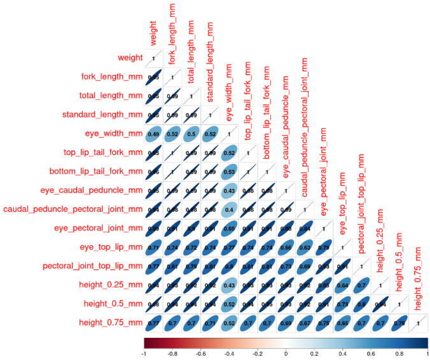
Title: A Quantile-quantile plot showing the observed versus expected $-\log_{10}(p)$ values for the GWAS for weight using three different methods

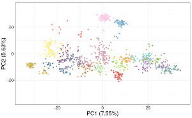
Additional file 2 Table S1

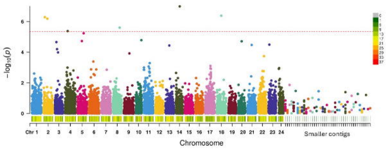
Format: XLSX

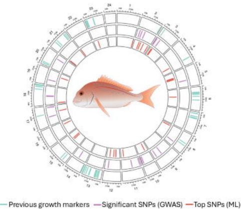
Title: SNPs identified as being significant across all GWAS tests performed


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
50mm


weight
fork_length_mm
total_length_mm
standard_length_mm
eye_width_mm
top_lip_tail_fork_mm
bottom_lip_tail_fork_mm
eye_caudal_peduncle_mm
caudal_peduncle_pectoral_joint_mm
eye_pectoral_joint_mm
eye_top_lip_mm
pectoral_joint_top_lip_mm
height_0.25_mm
height_0.5_mm
height_0.75_mm
weight 1
fork_length_mm 0.95 1
total_length_mm 0.95 0.99 1
standard_length_mm 0.95 0.99 0.99 1
eye_width_mm 0.49 0.52 0.5 0.52 1
top_lip_tail_fork_mm 0.95 1 0.99 0.99 0.52 1
bottom_lip_tail_fork_mm 0.96 1 0.99 0.99 0.53 1 1
eye_caudal_peduncle_mm 0.95 0.99 0.99 0.99 0.43 0.98 0.98 1
caudal_peduncle_pectoral_joint_mm 0.94 0.98 0.98 0.98 0.4 0.98 0.98 0.99 1
eye_pectoral_joint_mm 0.89 0.91 0.9 0.91 0.65 0.91 0.91 0.88 0.84 1
eye_top_lip_mm 0.71 0.74 0.72 0.74 0.77 0.74 0.74 0.66 0.63 0.79 1
pectoral_joint_top_lip_mm 0.77 0.81 0.79 0.81 0.8 0.81 0.81 0.73 0.69 0.93 0.91 1
height_0.25_mm 0.94 0.93 0.92 0.92 0.43 0.93 0.93 0.93 0.92 0.85 0.64 0.7 1
height_0.5_mm 0.98 0.94 0.94 0.94 0.52 0.94 0.95 0.93 0.92 0.91 0.73 0.8 0.94 1
height_0.75_mm 0.77 0.7 0.7 0.71 0.52 0.7 0.7 0.69 0.67 0.75 0.65 0.7 0.7 0.79 1
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1


PC2 (5.63%)
20
0
-20
-20
0
20
PC1 (7.55%)


−log10(p)
6
4
2
0
Chr 1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
Smaller contigs
Chromosome
0
1
5
9
13
17
21
25
29
33
37


24
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Previous growth markers
Significant SNPs (GWAS)
Top SNPs (ML)