Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Research Paper Ultra-low-depth whole-genome sequencing with reference panel-based imputation illuminates sex chromosome evolution and enables growth rate and disease-resistance GWAS and genomic prediction in sablefish (Anoplopoma fimbria)

Zhaoze Liu
University of Victoria
Briony Campbell
Golden Eagle Sable Fish
Lesperance Mary
University of Victoria
Isabel Ma
University of Victoria
Xuekui Zhang
University of Victoria
John Taylor
taylorjs@uvic.ca

University of Victoria

Research Article

Keywords: Sablefish (Anoplopoma fimbria), GWAS, Genome Prediction, SNPs, Growth-rate, imputation

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11167872/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Research Paper

Ultra-low-depth whole-genome sequencing with reference panel-based imputation illuminates sex chromosome evolution and enables growth rate and disease-resistance GWAS and genomic prediction in sablefish (*Anoplopoma fimbria*)

Zhaoze Liu[1], Briony Campbell[3], Lesperance Mary[1], Isabel Ma[2], Xuekui Zhang[1*] and John S. Taylor[2*]

1. Mathematics and Statistics, University of Victoria, PO Box 1700, STN CSC, Victoria, BC, V8W 2Y2, Canada
2. Department of Biology, University of Victoria, PO Box 1700, STN CSC, Victoria, BC, V8W 2Y2, Canada
3. Golden Eagle Sablefish, Walker's Hook, Salt Spring Island, BC, Canada

* Corresponding Author: Xuekui Zhang xuekui@uvic.ca John S. Taylor taylorjs@uvic.ca

Abstract

Genome-wide association studies and genome prediction (GWAS and GP) in non-model species are often limited by the lack of dense genotyping arrays or by high sequencing costs. Low-depth whole-genome sequencing (WGS) combined with genotype imputation can be a cost-effective alternative. However, it is not clear how low one can go with respect to target sample depth and reference panel size. In addition, these approaches have rarely been evaluated in populations with extensive family structure. We generated a sablefish (*Anoplopoma fimbria*) genome reference panel by sequencing 65 wild-caught individuals to a mean depth of 14x. Approximately 4.9 million SNPs were observed. SNP density varied among chromosomes and along chromosome arms and was lowest on chromosome 14, which contains the Master Sex-Determining gene, *gonodal somatic cell derived factor* (*gsdf*). We used this reference panel to impute genotypes in a target population of 1,380 farm-raised sablefish sequenced to a mean depth of 0.4x. Imputation accuracy was estimated by comparing imputed genotypes to those obtained from a PCR-based SNP genotyping panel and by comparing genotypes obtained from 14x alignments to those obtained from imputed down-sampled data. A total of 1.2 million SNPs (MAF ≥ 0.05) were utilized for GWAS and GP for weight at age 2.5 years (all target population individuals) and for sex and disease status (603 target population individuals sampled during an *Aeromonas salmonicida* outbreak). Significant associations with sex spanned chromosome 14. The lowest *p*-value SNP was 9 kb upstream of *gsdf*. Chromosome 14 SNPs were also associated with weight variation. This result was anticipated because females are known to grow faster than males. When only female data were analyzed, SNPs in *macrophage mannose receptor 1* (*mrc1*) were associated with weight variation. SNPs associated with disease status spanned a 46 Kb region of chromosome 20 with several candidate genes including peroxynitrite isomerase THAP4 (*thap4*). Genomic prediction models were developed using SNPs with low GWAS *p*-values. Sex and disease status were predicted with multi-locus models and had classification accuracies of 0.95 and 0.71, respectively. The best weight-prediction model achieved a Pearson correlation coefficient of $r = 0.65$ between observed and predicted weights. Ultra-low-depth WGS with imputation using a diverse but small reference panel provided an enormous set of reliable markers for GWAS and GP in this non-model aquaculture species.

Keywords:

Sablefish (*Anoplopoma fimbria*), GWAS, Genome Prediction, SNPs, Growth-rate, imputation

1. Introduction

Understanding the genetic basis of complex traits is central to both biological discovery and the development of effective breeding programs. Genome-wide association studies (GWAS) and genomic prediction (GP) have become standard tools for linking genetic variation to phenotypic trait variation and for predicting performance and breeding values from individual genotypes (Visscher et al., 2017; Meuwissen et al., 2001). Both approaches rely on dense genome-wide marker datasets generated from high-density SNP arrays or high-depth whole genome sequencing (WGS) that can capture linkage disequilibrium (LD) between markers and causal variants (Houston et al., 2020; Yáñez et al., 2015). Where such resources are available, genomic analyses have uncovered traits impacted by major-effect loci and traits associated with a large number of minor-effect loci (Gutierrez et al., 2015). A remarkable example is the identification of a locus affecting resistance to infectious pancreatic necrosis virus (IPNV) in Atlantic salmon (*Salmo salar*): This was followed by marker-assisted selection, which resulted in the number of IPNV outbreaks decreasing to near zero (Houston et al., 2010; Moen et al., 2009). More recently GWAS identified eight olive flounder (*Paralichthys olivaceus*) SNPs associated with body size and sex-specific growth patterns (Omeka et al., 2022). In common carp (*Cyprinus carpio*) 39 loci across four genomic regions were associated with growth at different developmental stages (Su et al., 2020). Genome prediction has also shown promise in aquaculture species. For example, Zhou et al. (2023) utilized genomic best linear unbiased prediction (GBLUP) to model resistance to white gill disease in large yellow croaker (*Larimichthys crocea*) and found that individuals with the highest genomic estimated breeding values (GEBVs) were more likely to survive a disease outbreak (Pearson correlation coefficient of $r = 0.54$). In Australasian snapper (*Chrysophrys auratus*), a linear classification approach (Chi-squared feature selection followed by distance-weighted discrimination) produced a multi-locus model that classified fish as slow- or fast-growers (with an accuracy of 0.60) (Chen et al., 2025). However, genomic resources are typically limited for non-model organisms, including most aquaculture species.

To expand GWAS and GP research to a larger number of species, low-depth WGS combined with reference panel-based genotype imputation is emerging as a cost-effective strategy (Li et al., 2010; Liu et al., 2023; Browning & Browning, 2016; Pasaniuc et al., 2012; Rubinacci et al., 2021; Daetwyler et al., 2014; VanRaden et al., 2017; Veerkamp et al., 2016; Koorevaar et al., 2025; Tsai et al., 2017). This approach is beginning to be utilized in aquaculture. For example, in golden pompan (Trach*inotus ovatus*) the number of SNP loci (and candidate genes) associated with body length was much higher after reference panel-based imputation of 6x data than it was without imputation (Sun et al., 2025). To test the limits of the low-depth WGS plus imputation approach, Zhang *et al*. (2025) down-sampled 10x – 16x data from spotted sea bass (*Lateolabrax maculatus*). They showed that 3x data imputed with a small but diverse reference panel produced GWAS and GP results similar to those obtained from the high-depth data.

Farmed fish populations often exhibit strong family structure, which can inflate GWAS and GP signals (Goddard et al., 2016; Ødegård et al., 2014). However, this structure can be difficult to predict a priori if standard protocols involve multi-parent breeding, stock size-sorting during grow-out, and/or among-family variation in survival. Evaluating the performance of low-depth WGS based imputation and downstream genomic analyses in such breeding populations is particularly important for aquaculture genetics.

Sablefish (*Anoplopoma fimbria*) is a long-lived marine teleost native to continental slope regions of the eastern and western Pacific Ocean, ranging from Mexico to Japan (Sasaki, 1985). Along the west coast of North America, sablefish are harvested primarily using long-line hook and long-line pot trap fisheries and represent an important economic resource in Alaska (USA) and British Columbia, Canada. The total allowable catch (TAC) in British Columbia has historically been valued at tens of millions of dollars annually (Fisheries and Oceans Canada, 2025), although landings and average body size have declined over the past decade (Lacko, 2021). Sablefish aquaculture has the potential to reduce fishing pressure on wild populations while supporting remote coastal economies. In British Columbia, Canada, expansion of sablefish farming may also help offset economic and social impacts associated with reductions in ocean-pen salmon aquaculture. Over the past two decades, substantial progress has been made in developing sablefish aquaculture protocols (Cook et al., 2015; Goetz et al., 2021). Nevertheless, sablefish aquaculture remains challenged by pronounced phenotypic variation among individuals and families. Traits such as growth rate, and disease susceptibility vary widely under commercial conditions. While this variability complicates production forecasting and farm management, it may present an opportunity for genetic improvement through selective breeding.

We previously used a 300-locus SNP panel to investigate sablefish body-weight variation in three cohorts (Liu et al., 2025). That work identified strong and reproducible associations between body weight and SNPs in the sex-determining region (SDR) on chromosome 14, reflecting pronounced sexually dimorphic growth: At 2.5 years of age, females had a mean weight of approximately 3.0 kg whereas males weighed approximately 2.0 kg. After accounting for sex, several non-SDR SNPs were associated with weight variation, and a 29-SNP model trained via elastic-net regression, predicted body weight in an independent cohort with a Pearson correlation coefficient of $r = 0.49$. We also used this genotyping panel to study the influence of individual SNPs and multi-locus genotypes on the occurrence of spine dysmorphology in sablefish (Gill et al. submitted). However, with approximately 15 SNPs per chromosome, this panel had limited resolution for detecting associations across diverse cohorts and for estimating LD, which makes the identification of candidate genes difficult. The availability of a reference genome assembly (Flores et al., 2023), new WGS data and advances in genotype imputation methodology, have the potential to enable dense genome-wide analyses in this species while maintaining cost efficiency.

Here we tested the potential of low-depth WGS data combined with haplotype data from a relatively small reference panel (sequenced at high-depth) to generate a dense genome-wide SNP dataset through genotype imputation. Then we looked for loci associated with sex, body weight, and disease status. We also used regularized regression models with a subset of SNPs selected from GWAS to generate genome prediction models for sex, growth rate, and disease (furunculosis) resistance.

2. Methods

The flowchart in Fig. 1 illustrates a four-stage bioinformatics framework designed to integrate low-depth and high-depth WGS data for GWAS and GP. Raw sequencing data from low-depth target population and high-depth reference panel were aligned to the reference genome to

produce standard vcf files. The high-depth WGS data was phased to reconstruct haplotypes, which were used to impute missing and inaccurate genotypes in the low-depth WGS target population dataset. Imputed data were analyzed using simple linear regression, multiple linear regression and linear mixed-model regression to identify SNPs and candidate genes associated with sex, weight at age 2.5 years, and susceptibility to *Aeromonas salmonicida* (a bacterial pathogen that causes furunculosis). Subsets of SNPs, selected based upon GWAS *p*-values, were used for Genome Prediction model testing.

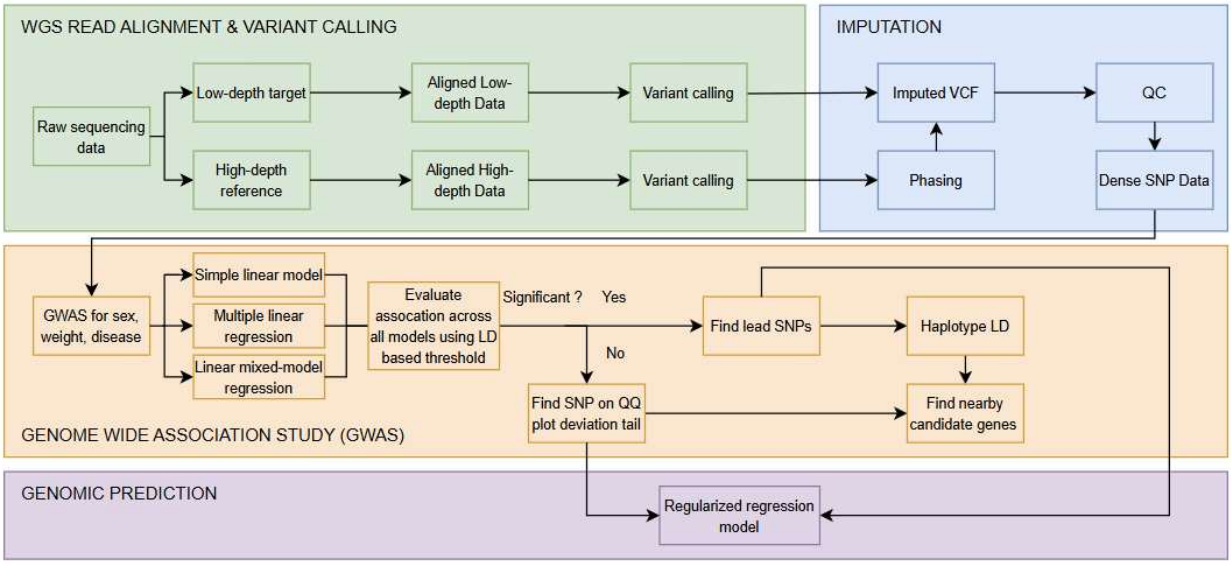


Figure 1 | Overview of the low-depth WGS, GWAS, and genomic prediction pipeline. Firstly, low-depth target population reads and high-depth reference panel reads were processed and aligned to the sablefish reference genome. Variants were called using HaplotypeCaller. Secondly, reference panel genotypes were phased using SHAPIT4 and target population genotypes were imputed using GLIMPSE. Thirdly do GWAS for sex, body weight, and disease status was performed using simple linear regression, multiple linear regression (with sex as a covariate) and linear mixed-model regression (to accounted for relatedness among individuals in the target population). This was followed by candidate gene identification. Finally Genome prediction models were then constructed using regularized regression and trait-specific SNP sets derived from GWAS.

2.1 Sampling, whole genome sequencing and variant calling.

The reference panel was generated from DNA isolated from 65 unknown-age, wild-caught adult sablefish. Low-depth WGS was carried out for 1380 individuals (the target population). These farm-raised fish spend six to nine months in a hatchery on Salt Spring Island, BC and up to two years in grow-out pens in Kyuquot Sound, BC. The target population included fish from several 'Implants'. This term was described by Liu *et al*. (2026). Briefly, gonadotropin-releasing hormone (GnRH) implants are used to promote egg production in broodstock females. A given 'Implant' describes a block of time (approximately two weeks) when eggs from one or more implanted females are fertilized with milt from one or more males and/or one or more 'neomales' (Luckenback et a., 2026). Offspring are then labeled by Implant number and year

(e.g., ‘Implant 1, 2019’). Fertilization success and hatchery survival vary among crosses within implants (Rubi et al., 2022). In addition, during the grow out period fish are sorted into different pens to generate populations of similar sized individuals. These factors make relatedness among harvest-size individuals sampled from any given pen difficult to estimate prior to genetic testing.

Most target population fish were weighed and photographed during processing at St. Jean’s Cannery in Nanaimo, BC (aged approximately 2.5 yrs). Fin clips were obtained from 400 fish in 2021 (200 individuals from Implant 1, 2019 and 200 from Implant 2, 2019) and from 380 fish in 2022 (Implant 5, 2019). Sex was assigned later by genotype (see Results). In 2024 sampling took place during a an *Aeromonas salmonicida* outbreak (this bacterial pathogen causes furunculosis). Moribund fish (N=220) were obtained by dip net at the farm site during the four weeks leading up to harvest and their healthy pen-mates (N= 380) were sampled at harvest. For the 2024 sample DNA was isolated from head kidney and sex was determined by examining gonads during the kidney biopsies. These fish were derived from a diversity of 2022 Implants

Sequencing depth was estimated from paired-end FASTQ files prior to read filtering, alignment, or duplicate read removal. For some fish the numbers of R1 and R2 reads were unequal so read counts from R1 and R2 were averaged. Mean genome-wide sequencing depth was then calculated as the total number of sequenced bases (each read assumed to be 150 bp long) divided by the sablefish reference genome size, 653,543,715 bp (Flores et al., 2023).

Reads were aligned to the sablefish reference genome (GCA_027596085.2_Afim_UVic_2022_genomic) using Bowtie2 v2.4.4 (Langmead & Salzberg, 2012), with default end-to-end alignment parameters. These paired-end reads were required to be aligned in the expected orientation and to have a specific insert-length range (0-500bp). Alignments were further required to exceed the default minimum alignment score defined as $-0.6 \times$ read length. Mismatch penalties were -6 for high-quality bases and -2 for low-quality bases, with gap opening and extension penalties of 5 and 3, respectively. Reads failing to meet any of these criteria were classified as unmapped. Duplicate reads were identified and marked using Picard to reduce biases arising from library amplification. Alignment files (duplicates excluded) were converted to binary format, sorted by genomic coordinates, and indexed using SAMtools (Li et al., 2009). Read group information (fish ID) was added to each alignment using Picard (Picard Toolkit, 2019) to track samples during downstream joint genotyping.

Variant discovery was performed using HaplotypeCaller in the Genome Analysis Toolkit (GATK; DePristo et al., 2011), run under a diploid model (--sample-ploidy 2) in GVCF mode. This mode first generates genotype likelihoods at variable positions in each sample, facilitating downstream multi-sample joint genotyping. At the read level, base qualities scores were incorporated into the genotype-likelihood calculation, with bases below the default minimum base quality threshold (--min-base-quality-score 10) excluded from calling and bases below the base-quality score threshold (--base-quality-score-threshold 18) contributing reduced evidence. To limit the over-representation of highly redundant reads at high-depth loci, reads sharing the same alignment start position were down-sampled above the default limit (--max-reads-per-alignment-start 50). The default activity profile score (0.002) was used. This is a way to retain variants in an individual observed in only a few reads, while rejecting variants likely to be sequencing errors (McKenna et al., 2010). Joint genotyping, using GenotypeGVCFs, was then

used to produce a multi-sample variant set. Final site-level quality metrics and genotype posterior probabilities were calculated by integrating individual sample genotype likelihoods with population-level prior expectations via a Bayesian model, using default heterozygosity priors (--heterozygosity 0.001) and indel heterozygosity priors (--indel-heterozygosity 1.25×10^{-4}). For the target population, joint genotyping was performed in 2Mb non-overlapping genomic windows for each chromosome to accommodate computational constraints imposed by sample number. These data were merged to generate one set of genotypes per chromosome. For this dataset, the default activity profile setting in HaplotypeCaller likely caused some true target population SNPs to be missed in individuals with lower-quality reads, resulting in incorrect calls at those positions. However, the subsequent imputation step (see below) provided an opportunity to recover these missed genotypes.

2.2. Imputation

The first steps in the imputation pipeline included the sorting of reference panel genotypes into chromosomes followed by genotype phasing using SHAPEIT4 (Delaneau et al., 2019). SHAPEIT4 infers reference panel haplotypes by leveraging linkage disequilibrium (LD) structure across samples. Reference panel SNP positions are then assigned to the target individuals. Many of these positions will not have been scored (no reads or low-quality calls) in some (or even all) target individuals. Conversely, variants present in target population individuals that were absent from the reference panel were excluded as imputation generates target population genotypes from reference panel haplotypes.

Imputation employed GLIMPSE/1.1.1 (Rubinacci et al., 2021), which was specifically designed for large-scale low-depth WGS data. For each target individual, genotype likelihoods were used to summarize the read-level data at each reference panel variant position. GLIMPSE then combines these single-site likelihoods with haplotype structure from the phased reference panel to infer phased genotypes across genomic windows, effectively modelling each target population 2Mb haplotype as a mosaic of reference haplotypes. These 2-Mb haplotypes were then ligated to generate chromosome-level variant call sets for each target population fish.

Twenty-nine reference panel fish had been previously genotyped using a 300-locus PCR-based SNP genotyping panel. The new 14x-WGS data provided an opportunity to evaluate the performance of this SNP panel. With confidence in the data derived from the panel (see Results) we were able to evaluate the accuracy of the 0.4x-derived genotypes (before and after imputation) because 800 target population fish had also been genotyped with the panel (Liu et al., 2026). A second assessment of the ability to impute genotypes from low-depth sequence data using the reference panel haplotypes involved down-sampling reads. For this, read data from five reference panel individuals were down sampled to 0.5x, 1.0x, 2.0x and 5x. Genotypes for each of these five fish were then imputed using a 64-individual (128 haplotypes) reference panel that excluded the focal fish and compared to the variant calls from the original 14x data.

2.3. Genome-Wide Association Study (GWAS)

2.3.1. GWAS *p*-value thresholds

In dense genotype datasets, neighboring markers may be correlated due to linkage disequilibrium (LD), meaning that the number of statistically independent tests can be substantially smaller than the total number of variants (Dudbridge and Gusnanto, 2008). To define the genome-wide (GW) significance threshold, α_{GW}, we first estimated the effective number of independent markers (M_{eff}) using LD pruning implemented using the --indep-pairwise command in PLINK 2.0 (Purcell et al., 2007). Pruning was performed using a sliding window of 1000 variants, advanced by 50 variants at each step, removing one SNP from pairs with squared correlation $r^2 > 0.8$. The retained set of SNPs was taken as an estimate of M_{eff}. As M_{eff} depends on the correlation structure of genetic markers and not on phenotype modeling, the same threshold was applied to all GWAS analyses:

$$\alpha_{\text{GW}} = \frac{0.05}{M_{\text{eff}}}.$$

2.3.2. GWAS by Simple Linear Model Regression

Imputed low-depth WGS genotype probabilities for target population individuals were converted into PLINK binary file format and each locus was assigned one of three genotypes, 0, 1, or 2, reflecting the number of alternative alleles. Alternate alleles are those not present in the reference genome and they may or may not be minor alleles. For the GWAS for weight at age 2.5 years, the following additive model was fitted to the data:

$$\text{weight}_i = \beta_0 + \beta_G G_i + \varepsilon_i,$$

where G_i denotes the genotype (0, 1, or 2) for individual i, $\beta_G$represents the per-allele effect estimate, and $\varepsilon_i$is the residual error term. For the binary traits (sex and disease status), PLINK's logistic regression framework was applied. For each SNP, the following model was fitted to the data:

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_G G_i,$$

where G_i denotes the genotype (0, 1, or 2) for individual *i,* $\beta_G$represents the per-allele effect estimate*,* y_i represents the binary phenotype (sex or disease status, scored as 0 or 1) and *P* = probability of male or moribund (y_i = 1).

2.3.3. GWAS by Multiple Linear Model Regression

Sex is a major determinant of growth in some teleost fishes and may also affect disease susceptibility (Piferrer, 2001). Accordingly, GWAS analyses were also conducted with sex as a covariate. The sex-adjusted model for body weight and the sex-adjusted logistic regression model for disease status are shown (in order) below:

$$\text{weight}_i = \beta_0 + \beta_G G_i + \beta_S \text{sex}_i + \varepsilon_i,$$

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_G G_i + \beta_S \text{sex}_i.$$

2.3.4. Linear Mixed-Model Regression

The target population was comprised of offspring from a limited number of brood stock crosses and consequently included an unknown number of full- and half-siblings. In this situation, standard GWAS test statistics may be inflated, and spurious associations discovered. To address this, we explicitly accounted for relatedness among individuals using a linear mixed-model GWAS framework that incorporated a genome relationship matrix (GRM) as a random effect. This approach separates locus-specific association signals from background genetic similarity across the genome. A GRM was estimated for the target population from the imputed SNP data using GEMMA (short for Genome-wide Efficient Mixed Model Association; Zhou & Stephens, 2012). As above, genotypes were coded under an additive model (0, 1, or 2 copies of the alternative (non-reference) allele) prior to matrix construction, then genotype dosages were centered and scaled according to allele frequency, so that each SNP contributed to the estimate of genome-wide relatedness on a comparable scale despite differences in minor allele frequency. The resulting GRM, was visualized as a heatmap in R using pheatmap. This GRM was used in all linear mixed-model analyses. For body weight the following model was fitted to the data:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{u} + \varepsilon, \qquad \mathrm{u} \sim \mathcal{N}\left(0, \sigma_g^2 \mathrm{K}\right), \qquad \varepsilon \sim \mathcal{N}(0, \sigma_e^2 \mathrm{I}),$$

where y is the vector of phenotypic values, Xβ represents fixed effects including the SNP genotype and sex, u is a random polygenic effect modeled using the GRM, and ε is the residual error term. Association statistics for weight variation were obtained using likelihood-ratio tests implemented in GEMMA. Sex and disease status were analyzed using GMMAT (short for Generalized linear Mixed Model Association Tests; Chen et al., 2016) with R (R Core Team, 2024). The following logistic mixed model was fitted to the data:

$$\mathrm{logit}[P(y_i = 1)] = \mathrm{x}_i^{\top}\beta + u_i, \qquad u \sim \mathcal{N}\left(0, \sigma_g^2 \mathrm{K}\right)$$

where y_i = binary phenotype (0 or 1), P = probability of y_i= 1 (male or moribund), x_i containing an intercept and sex when sex adjustment, was applied. A null model excluding G_i (SNP effects) was first fitted using restricted maximum likelihood. Genome-wide association testing was then performed using efficient score tests that fit the null mixed model only once. Next, for each SNP, a fast score statistic is calculated, instead of refitting the complete mixed model for each SNP test. This approach evaluates the association between each SNP and the phenotype without refitting the full model for every variant.

2.3.5 GWAS Interpretation

Simple linear model analyses (marginal single-SNP regression), multiple linear model analyses (sex-adjusted single-SNP regression) and the linear mixed-model regression, which accounted for genetic relatedness (GEMMA for continuous traits; GMMAT for binary traits) were used to identify loci associated with sex, weight at age 2.5 years and disease status (moribund or healthy). Imputation accuracy estimates were close to 50% (see Results). As a result, we focused our attention primarily on clusters of low *p*-value SNPs (GWAS peaks) as imputation errors are unlikely to be clustered in the genome.

2.4. Genomic Prediction

SNPs that were significantly associated with trait variation in all three regression analyses were used for Genome Prediction (GP). These SNPs were first filtered for LD to avoid 'feeding' the GP analyses with correlated (i.e., redundant) data. Pairwise LD (measured as r^2) was computed with LDBlockShow. GWAS-significant SNPs were grouped into LD clusters using a graph-based approach in which SNP pairs were connected if their $r^2 \geq 0.8$. Within each cluster, the SNP with the smallest GWAS *p*-value was utilized for GP. If no GWAS-significant SNPs were observed in all three analyses, a set of 'candidate' contributor-SNPs were selected from Quantile-Quantile (Q-Q) comparisons of expected and observed log-transformed GWAS *p*-values. Here a threshold was identified that separated SNPs into a group with little or no deviation from expectations and a set of SNPs with larger, albeit non-significant, deviations from expectations. No LD-filtering was applied to these candidate predictors.

Weight was modeled using penalized linear regression. Sex and disease status were modeled using penalized logistic regression. This distinction determines the likelihood function used in model fitting but does not affect SNP selection rules. For weight, we used elastic-net regularized regression (Zou & Hastie, 2005) implemented in glmnet with mixing parameter $\alpha = 0.5$, which balanced L1 (LASSO) (sparse variable selection) and L2 (ridge) (polygenic shrinkage) penalties. This approach is well suited for genomic prediction where predictors are numerous, and differ in effect magnitude. Elastic-net regularized regression is also designed to detect, down-weight and preserve correlated parameters. Thus, LD filtering (described above) was used as an unsupervised prefiltering step to reduce highly redundant GWAS *p*-value-derived markers, and elastic-net was applied as a supervised prediction method to handle the remaining correlation structure through penalized multivariable regression. The following model was fitted to the data:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k c_{ik} + \varepsilon_i,$$

where y_i = phenotype value of individual i, and x_{ij}= the genotype dosage of SNP j. c_{ik}denotes covariates (e.g., sex when applicable), $\beta_j$and $\gamma_k$are regression coefficients, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. For sex and disease status, an elastic-net regularized logistic regression model was used:

$$\log\left(\frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{k=1}^{q} \gamma_k c_{ik}.$$

This model estimates coefficients β by minimizing:

$$\text{Loss}(\beta) \; + \; \lambda[\alpha \sum_j |\beta_j| \; + \; (1-\alpha) \sum_j \beta_j^2]$$

For all three traits, 10-fold cross-validation was used to select the optimal regularization parameter (λ). Predictive performance was evaluated using out-of-fold predictions. For weight, performance was assessed using Pearson correlation and mean squared error between observed

and predicted phenotypes. For sex and disease status, classification accuracy and confusion matrices were used.

3. Results

3.1. Read sequencing and variant calling result

Read data have been deposited in the SRA database under BioProject 1300457. Ninety-nine percent of the reads mapped to reference genome. The reference panel individuals (N=65) had an average read depth of 14x and the target population individuals (N=1380) had an average read depth of 0.4x (Fig. 2). We identified 4,940,619 biallelic SNPs in the reference panel. SNP density ranged from approximately 4.8 to approximately 9.2 SNPs per Kb with variation distributed among chromosomes and along chromosome arms (Fig. 3). Chromosome 14, which contains the Major Sex Determining (MSD) gene, *gsdf*, had the lowest overall SNP density (Fig. 3). We identified 2,113,664 SNPs in the low-depth target panel. Of these, 307,054 (14.5%) were not found in the reference panel, leaving 1,806,610 error-prone SNP genotypes to guide imputation. After imputation 1,211,011 target population SNPs had MAF ≥ 0.05. These MAF-filtered data were used for GWAS and GP.

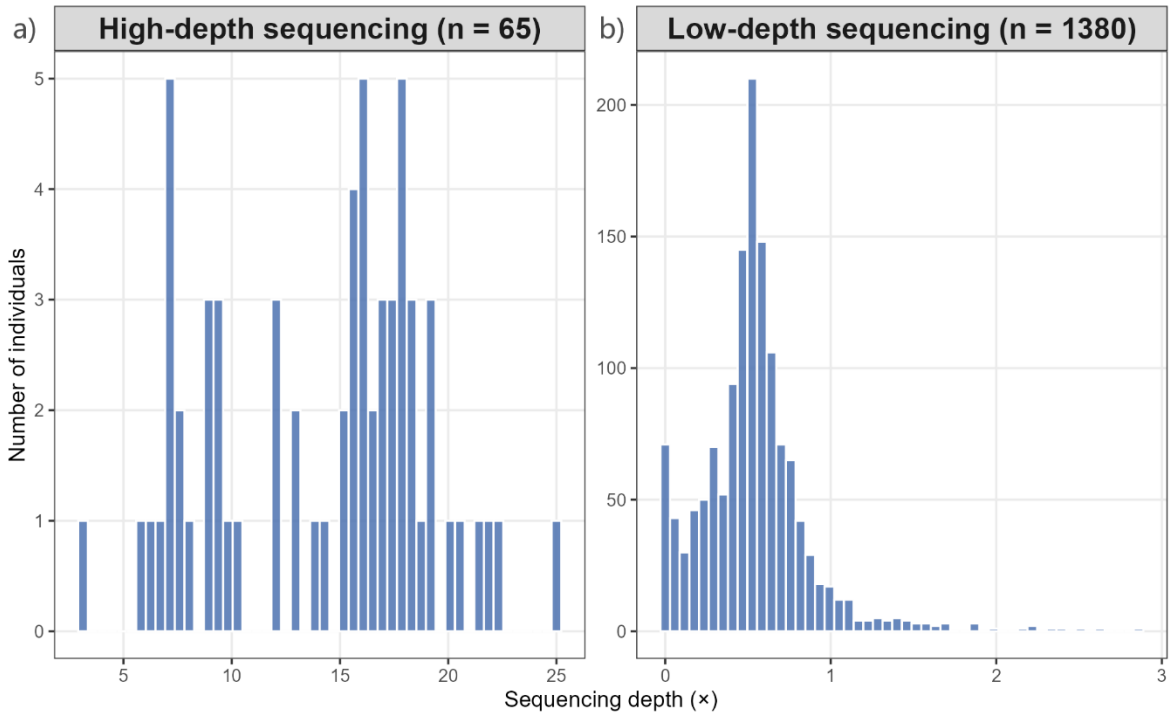


Figure 2 | Sequencing depth distributions of high- and ultra-low-depth sablefish genomes.

Sequencing depth was estimated from raw paired-end FASTQ files prior to read filtering and reference genome alignment. **a)** Sequencing depth distribution for the reference panel (n = 65), showing depth ranging from approximately 6x to 25x and a mean depth of approximately 14x. **b)** Sequencing depth distribution (mean depth ≈ 0.4x) for the target population (n = 1,380).

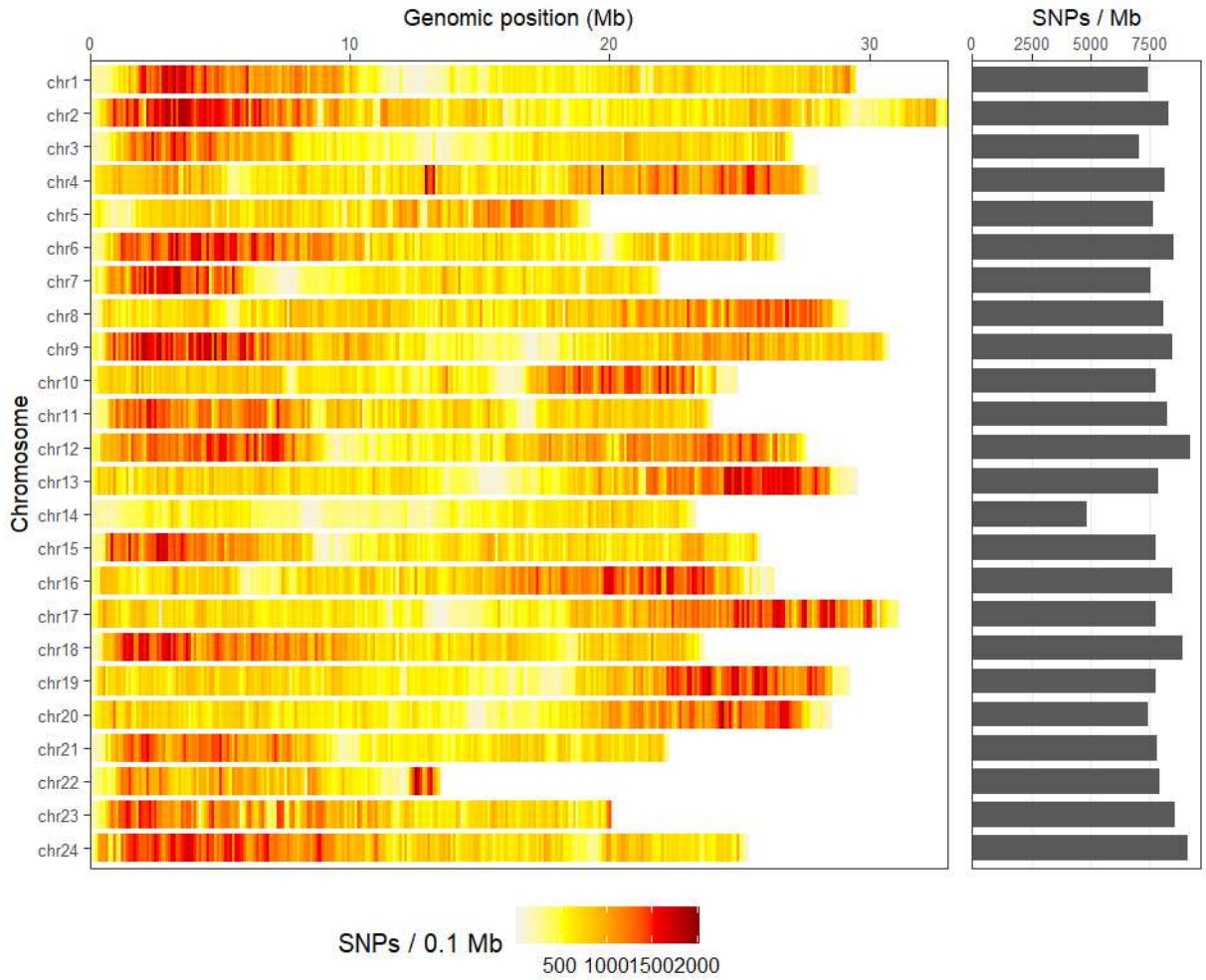


Figure 3 | Reference panel SNP density. A total of 4,940,619 reference-panel SNPs (no MAF filtering) were observed among 65 wild-caught individuals. SNP counts in 0.1 Mb windows mapped onto the 24 sablefish chromosomes while chromosome-level mean SNP density per Mb.

3.2. Imputation accuracy assessment

The first step in assessing imputation accuracy was to compare genotypes derived from high-depth WGS (14x) to those obtained using a 300-locus PCR-based SNP genotyping panel: 29 reference panel fish and 754 target population fish had been genotyped using this panel prior to this study. The 300-locus data from the 29 reference panel individuals (all wild caught) had been collected to provide sex and relatedness data to guide brood stock crosses. Data for 177 loci were present in both datasets. Concordance rate (matching SNPs/total SNPs) was 0.98 (Fig. 4a). Thus, the SNP panel provided an accurate benchmark for assessing SNP calls for the target

population. Genotypes for 161 loci (754 individuals) were present in the dataset derived from low-depth (0.4x) sequencing and the PCR-based SNP panel. Then down sample 29 reference individuals to 0.5x and using rest 36 reference impute, result compared with PCR panel, the accuracy is about 0.5 (Fig. 4b). For target panel, before imputation, concordance was low: At none of the 161 loci did genotypes match in more than 50% of individuals (Fig. 4c). After imputation concordance rate was 0.54 and many more than 50% of the SNP genotypes matched for 50% or more of the 754 individuals (Fig. 4d). These shows that the 65 and 36 reference panel sample size show similar imputation accuracy.

To expand this evaluation of imputation accuracy from sequence depth, five reference population individuals were independently down-sampled to 0.5x, 2.0x, 1.0x and 5x depth. The genotypes for each of these fish were then imputed using haplotype data from the remaining 64 reference panel individuals (Fig. 5). The aim here was to use the 14x data to determine how sequencing depth influenced concordance rates before and after imputation. The number of loci in each of these comparisons differed because only loci with direct read alignment evidence were utilized (i.e., fewer for 0.5x than for 5x). For the 0.5x data, genotype concordance ranged from 0.614 to 0.672 among the five down-sampled reference panel fish, with a mean concordance of 0.642. Concordance ranged from 0.66 to 0.76 (mean 0.708) for 1x data, from 0.66 to 0.80 (mean 0.726) for 2x data, and from 0.76 to 0.89 (mean 0.827) for 5x data (Fig. 5). The greatest imputation gains occurred at the lowest sequencing depth: An improvement from 233,707 matching SNPs (26.7% of 874,000) to 561,457 matching SNPs (64.2% of 874,000) or a gain of 327,750 loci, whereas concordance was 71.9% for 5x fish before imputation and 82.7% after imputation, increasing the total number of SNPs from 1,720,326 to 1,978423 (a gain of 258,097 SNPs).

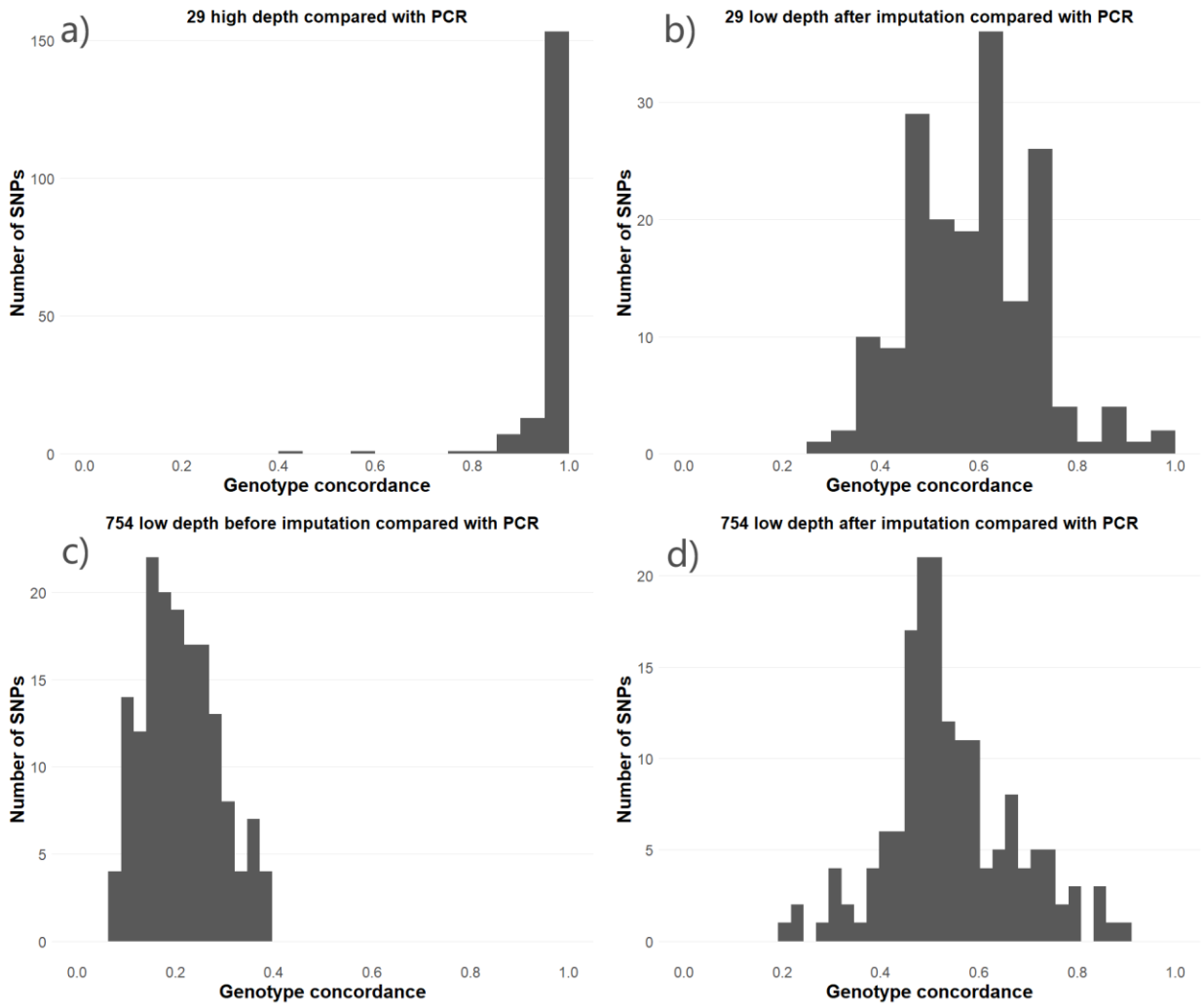


Figure 4 | Imputation accuracy estimated using data from a PCR-based 300-locus SNP panel. Each histogram summarizes locus-specific exact genotype concordance. The x-axis shows the proportion of individuals with matching hard-call genotypes and the y-axis shows the number of SNP loci. **a)** 177 loci scored using WGS (14x) and a PCR-based SNP genotyping panel for 29 reference panel fish. **b)** Concordance between imputed and down sampled 0.5x 29 reference individuals and PCR-panel genotypes. **c)** Concordance between 0.4x genotype calls and PCR-panel genotypes. **d)** Concordance between imputed 0.4x genotypes and PCR-panel genotypes.

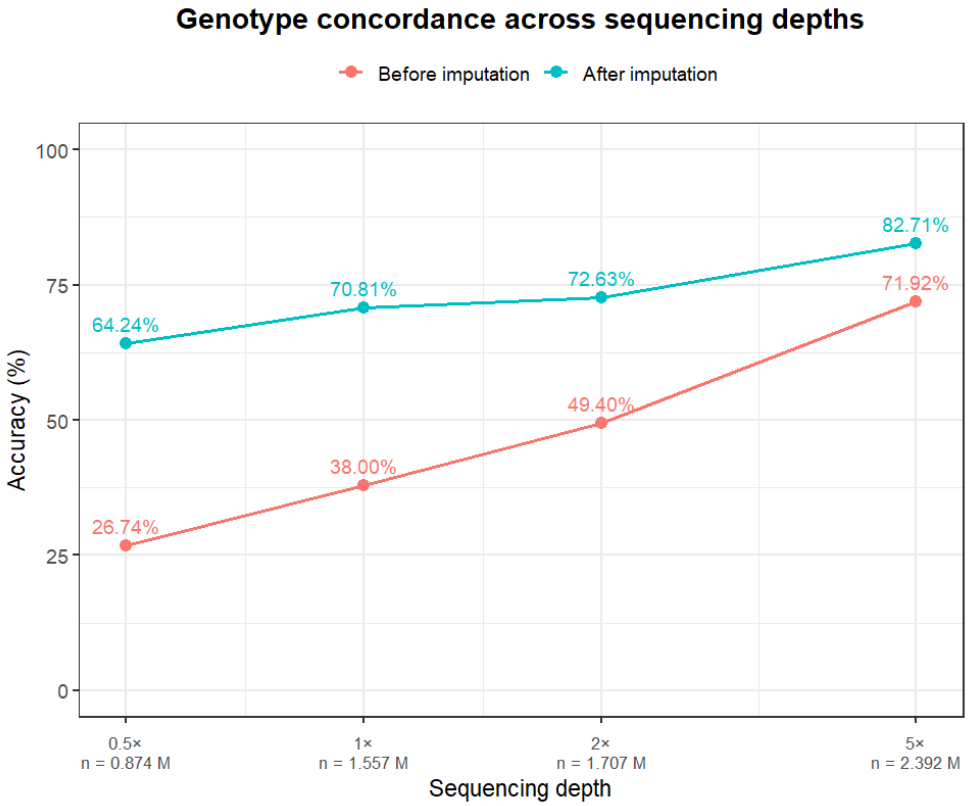


Figure 5 | Imputation accuracy estimated after down-sampling reference panel individuals. Concordance for imputed (blue line) and non-imputed (red line) variants. Points show mean values for five independently down-sample reference panel individuals. Imputation for each utilized haplotype data from a 64-sample reference panel that did not include the focal fish.

3.3. Target population relatedness

The target population Genome Relatedness Matrix (GRM) derived from the imputed SNP data showed a block-diagonal structure among the 1,380 target individuals (Fig. 6). Clusters of highly-related individuals (highly saturated red zones) represent sets of siblings while larger blocks with higher-than-average levels of relatedness (large less saturated red zones) show similarity consistent with half-sib relationships. Overall, the GRM visualization reveals substantial family structure in the target population.

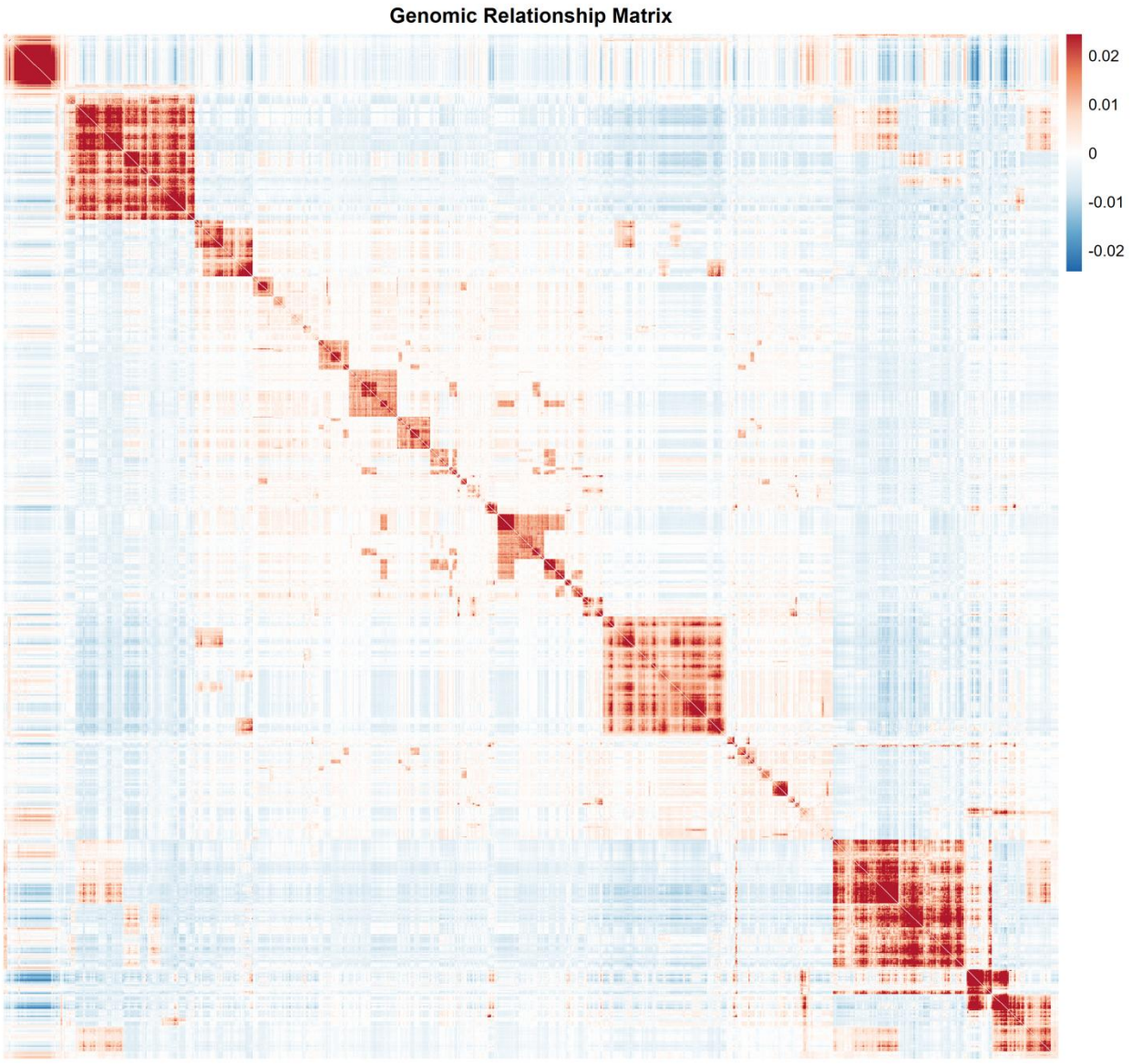


Figure 6 | Genome Relatedness Matrix (GRM) for target population (N=1,380). The centered GRM was calculated in GEMMA from 1,211,011 imputed SNPs and visualized as a heatmap in R using pheatmap. Each off-diagonal matrix element represents the average covariance of centered genotype values between two individuals across the imputed SNPs. Positive values (red) indicate greater genomic similarity than the population average, values near zero (white) indicate similarity close to the cohort average, and negative values (blue) indicate lower-than-average genomic similarity. Individuals were ordered using complete-linkage hierarchical clustering with identical row and column ordering. The prominent high-covariance blocks indicate groups of individuals with elevated genomic similarity, consistent with substantial family structure (i.e., the presence of full- and half-sibs).

3.4. GWAS and Genome Prediction Results

Genome-wide association analyses in the target population were conducted using 1,211,011 imputed SNPs. The effective number of independent tests (M_{eff}) was estimated to be 1,205,775 from the LD pruning (PLINK –indep-pairwise) analysis, corresponding to a genome-wide significance threshold of $\alpha_{GW} = 4.15 \times 10^{-8}$ or (-log $_{10}$P = 7.38). This threshold was used to identify significant associations in all GWAS tests. For Genome Prediction, significant SNPs, pruned using LD, were employed in the sex-prediction analyses and non-significant SNPs selected from QQ plots were employed in the weight-prediction and disease susceptibility prediction analyses.

3.4.1. Sex

GWAS: A kidney biopsy was performed on 603 sablefish sampled during the furunculosis outbreak in 2024 and at this time sex was determined by gonad inspection. The simple linear model analyses identified 21,459 SNPs, distributed throughout the genome, significantly associated with sex (Fig. 7a). This observation appears to be a consequence of the inclusion of neomale-derived individuals in the study. While collections were made with no knowledge of sex, the 2024 sample included 499 females and 104 males, with some all-female sets of siblings. Alleles unique to (or more common in) all-female families, no matter where in the genome they occur, appear to be driving this genome-wide SNP association with sex (and weight at age 2.5 yrs, see below). When the generalized linear mixed model was fitted to the data (i.e., the GMMAT analysis with the genomic relatedness matrix) the sex-associated signal was limited to 271 SNPs distributed over an approximately 13 Mb region of chromosome 14 (positions 2533498 bp to 12241604 bp) (Fig. 7b). After LD pruning at $r^2 = 0.8$, 204 SNPs were independently associated with sex.

A sub-threshold (non-significant) Manhattan Plot peak was observed on chromosome 8 after the GMMAT 'correction'. We tested the hypothesis that this was the result of genome miss-assembly (i.e., a fragment of chr 14 assigned to chr 8). Genes within, upstream of, and downstream of this chromosome 8 region are linked in other species (data not shown) indicating that an assembly error was not the explanation for this localized, sub-threshold, sex-associated peak. The QQ analysis of GMMAT-derived *p*-values (Fig. 7c) showed that the lower quantiles largely followed the null expectation. However, the upper quantiles showed a pronounced deviation from the null.

We used LDBlockShow to visualize LD in the vicinity of *gsdf*. These analyses (Fig. 7d) exposed a 10 Kb region with highly correlated genotypes that included the five lowest *p*-value SNPs (chr14:12111769, $P = 1.81 \times 10^{-28}$; chr14:12123203, $P = 5.69 \times 10^{-28}$; chr14:12091321, $P = 4.06 \times 10^{-26}$; chr14:12105310, $P = 6.28 \times 10^{-26}$; chr14:12121328, $P = 1.35 \times 10^{-25}$), and the sablefish SDR.

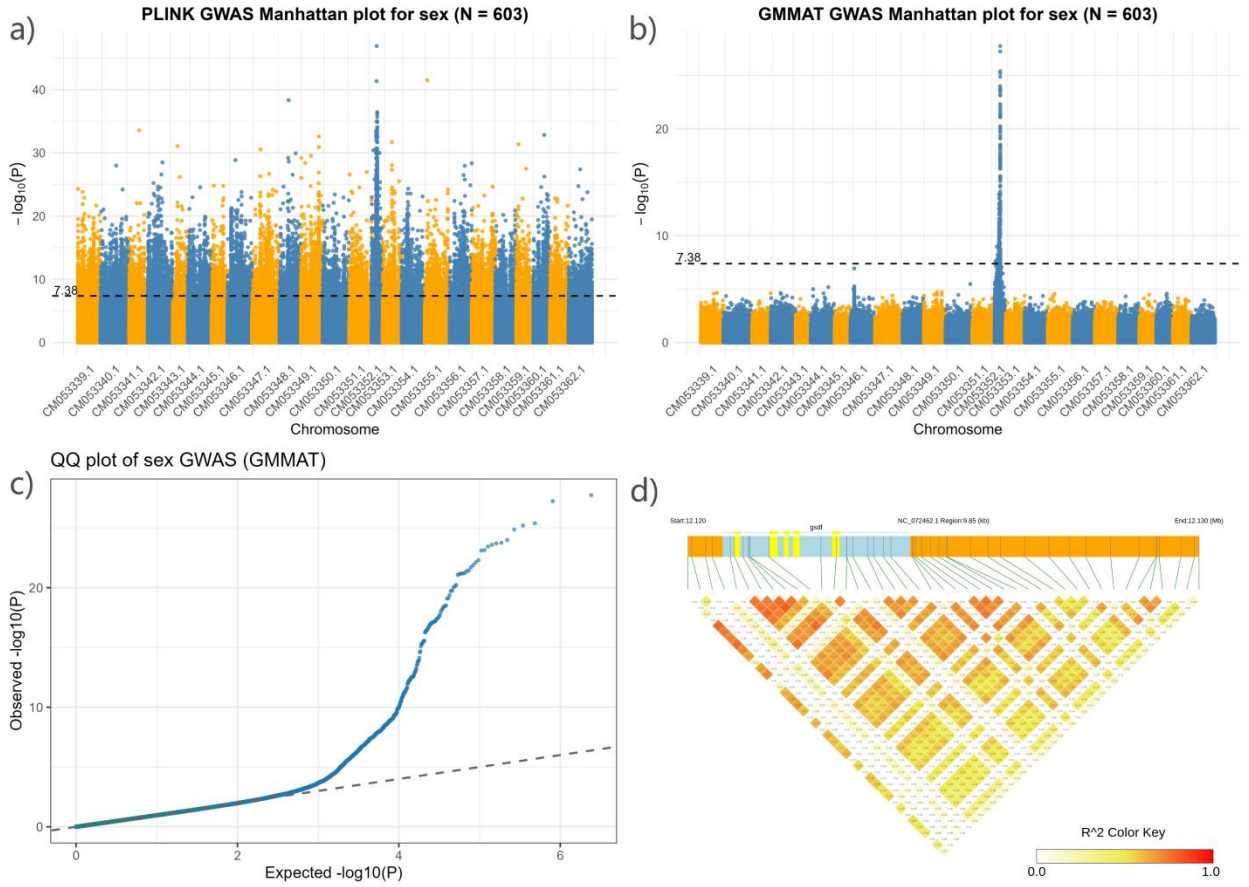


Figure 7 | GWAS for sex in 603 individuals whose sex was determined by visual inspection of gonads. a) Manhattan plot of sex GWAS analyses using logistic regression implemented in PLINK. **b)** Manhattan plot of sex GWAS using a generalized linear mixed model implemented in GMMAT. This analysis shows associated SNPs only on chromosome 14. The horizontal dashed line in panels **a** and **b** denote the genome-wide significance threshold ($-\log_{10}P = 7.38$). **c)** The QQ plot of GMMAT-derived *p*-values. **d)** Linkage Disequilibrium (LD) analysis for the region of chromosome 14 containing *gsdf* (blue exons, yellow introns) visualized using LDBlockShow. Heatmap reflects pairwise LD (r^2).

Because a few rare alleles can generate statistically significant GWAS associations, we used F_{ST} to attempt to delimit the size of the SDR. F_{ST} values reflect deviation from HWE expectations when data from two (or more) populations are combined (Wright, 1951; Holsinger and Weir, 2015). Here, F_{ST} analysis exposed loci with genotypes that were common in one sex and not in another. Fixation indices (F_{ST}) were estimated for chromosome 8 and chromosome 14 SNPs (Fig. 8). Loci with F_{ST} values > 0.5 were distributed across a 32 Kb region of chromosome 14. This is smaller than the region identified by GWAS but larger than the LD block that included *gsdf*.

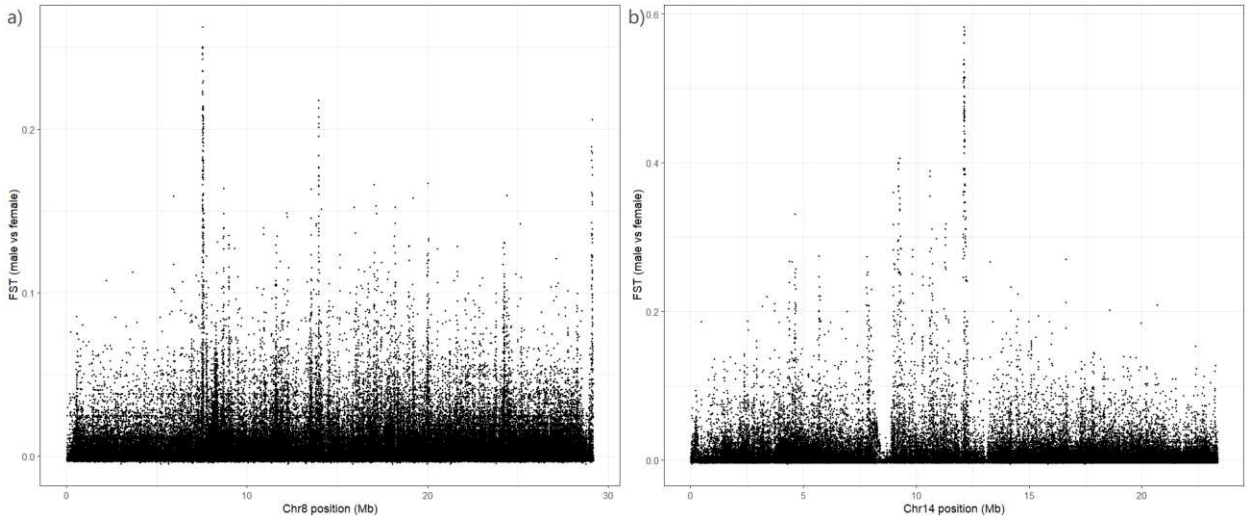


Figure 8 | F_{ST} values between males and females. a) Chromosome 8 F_{ST} values were generally low, with moderate localized elevations (max. F_{ST} = 0.25). **b)** Chromosome 14 F_{ST} values show a pronounced peak of genetic differentiation between males and females (max F_{ST} = 0.58). The strongest signal (F_{ST} >0.5) spanned a 32Kb region (chr14:12091321 to 12123914) that captured four annotated genes including *gsdf*, nucleoporin 54 (*nup54*), LOC129102360 (a leucine-rich, repeat-containing protein 2 isoform) and LOC129102793 (serine/threonine protein phosphatase with EF-hands 2-like).

Genome Prediction: LD pruning of GWAS-significant sex-associated SNPs generated a set of 204 chromosome 14 loci for GP analyses. The optimal elastic-net model (λ_{min} = 0.012) retained 61 loci. Collectively, these SNPs provide reliable sex inference without necessarily representing distinct biological sex-determining mechanisms. When applied to the full dataset the elastic-net model correctly classified 492 of 499 females and 83 of 104 males, achieving an overall classification accuracy of 0.97. Misclassifications were primarily observed among males. This reflects the fact that imputation errors are more likely to miss a heterozygote genotype (male) than a homozygote (female). In addition, rare phenotypic sex assignment errors are more likely to involve scoring a female as a male (immature ovaries) than the other way around and thus it is possible that a few genotypic females were incorrectly scored 'male' in the field. Using 10-fold cross-validation, the model achieved classification accuracy of 0.95, demonstrating strong generalization performance and minimal overfitting.

3.4.2. Weight

GWAS. When body weight GWAS was carried out without sex or relatedness as covariates we observed a distinct multi-locus low *p*-value association peak on chromosome 14 (Fig. 9a). Size-associated SNPs were also observed on autosomes. Females are known to grow approximately twice as fast as males during the time spent in grow-out pens (Liu et al., 2026), thus the association with SNPs on chromosome 14 was anticipated. When we fitted a multiple linear model to the data (which included a sex covariate) the chromosome 14 peak no-longer stood out from autosomal peaks and significant associations remained distributed throughout the genome (Fig. 9b). The presence of neo-male-derived all-female siblings in the sample appears to have had a similar impact as was observed in the sex GWAS. Alleles unique to (or more common in)

all-female families regardless of their position in the genome, would be associated with sex and weight. There were no significant SNP genotype-weight associations after the genomic relatedness matrix (GRM) was incorporated into the analysis (Fig. 9c).

Intriguingly, when males and females were analyzed separately significant SNPs were detected for both: For males, two SNPs were associated with weight variation: chr3:2915096 and chr8:8210647 (Fig10a). For females, one SNP on chromosome 9 and four SNPs on chromosome 14 were significantly associated with weight variation (Fig. 10b). Three of the four female-only chromosome-14 SNPs occurred within LOC129102365/macrophage mannose receptor 1 (*mrc1*). However, weight variation among females with different genotypes at the lowest *p*-value chromosome 14 SNP was small and the lowest weight genotype was rare (Fig. 11).

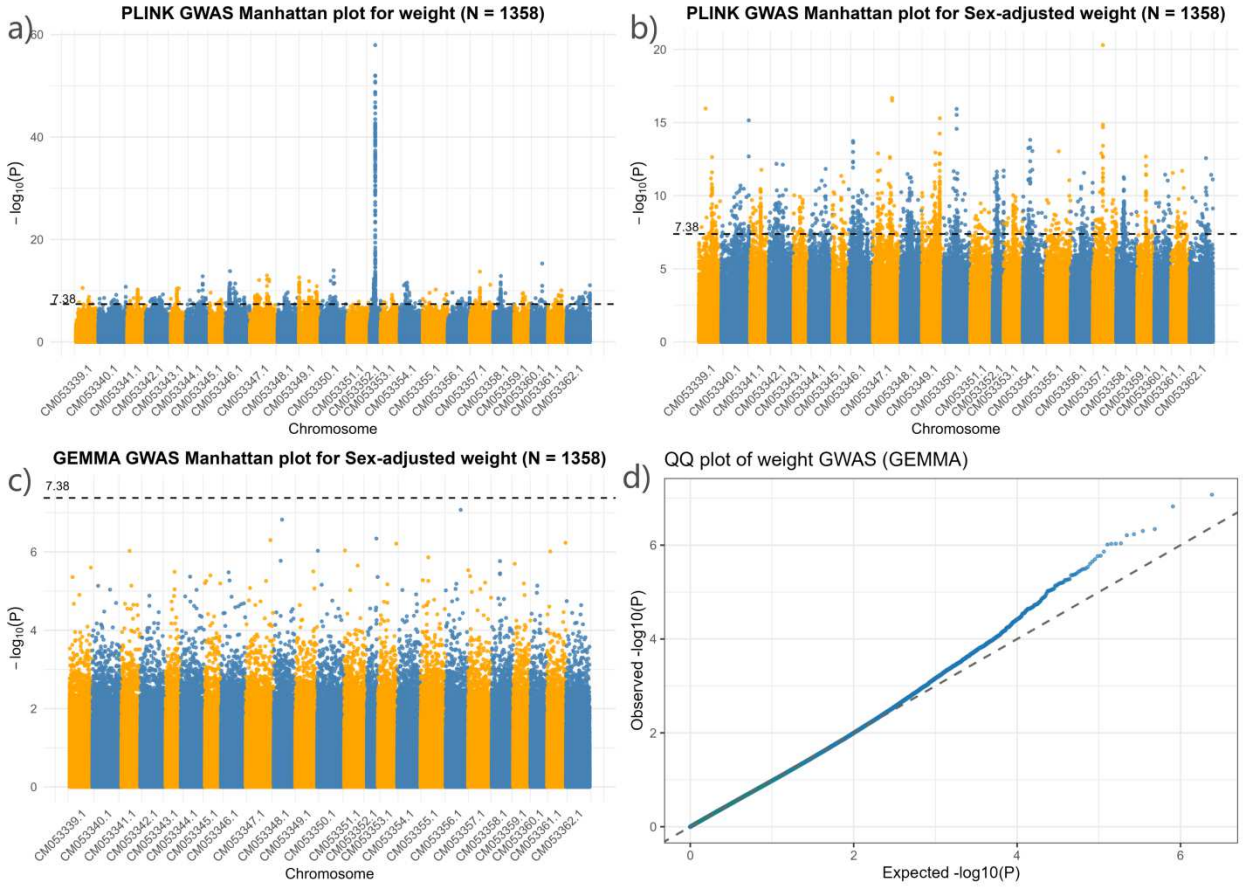


Figure 9 | Genome-wide association analysis and variance decomposition for body weight at 2.5 years in sablefish (N=1358). a) Manhattan plot of GWAS results for body weight without covariate adjustment (simple linear model). **b)** Manhattan plot for body weight after adjustment for sex (multiple linear model). **c)** Manhattan plot showing SNPs associated with weight identified using the linear mixed model with sex as a covariate and with *p*-values adjusted for relatedness among individuals. **d)** QQ plot of *p*-values from the linear mixed-model analyses.

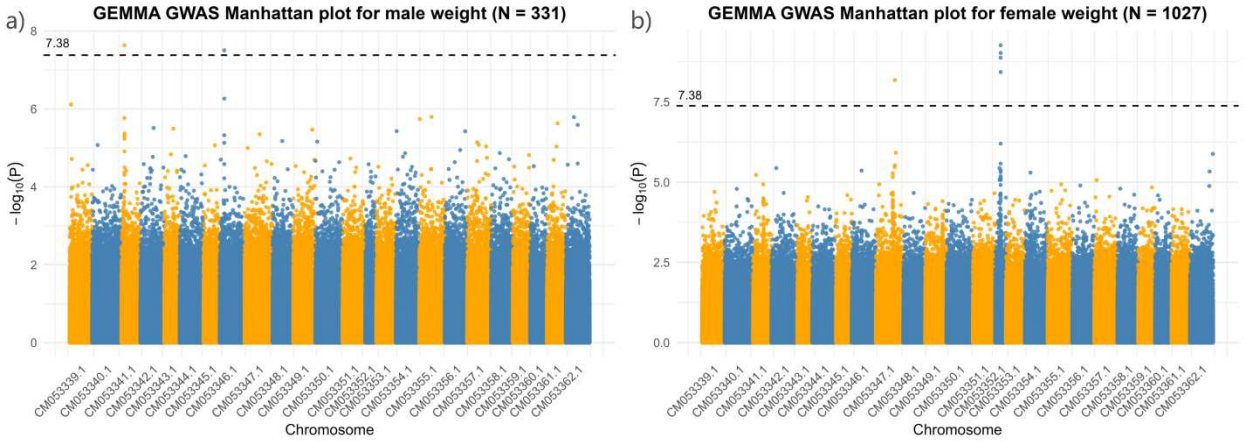


Figure 10 | GWAS for weight in sablefish sorted by sex. a) Manhattan plot of lonear mixed model (GEMMA) GWAS results for body weight among males (N = 331). Two SNPs, one on chromosome 3 and one on chromosome 7, and had significant associations. **b)** Manhattan plot of linear mixed model GWAS results for body weight among females (N = 1027). Significant associated observed on chromosome 9 and chromosome 14. Three of the chromosome 14 SNPs occur in LOC129102365/macrophage mannose receptor 1 (*mrc1*), which is approximately 26Kb downstream of the major sex determining gene, *gsdf*.

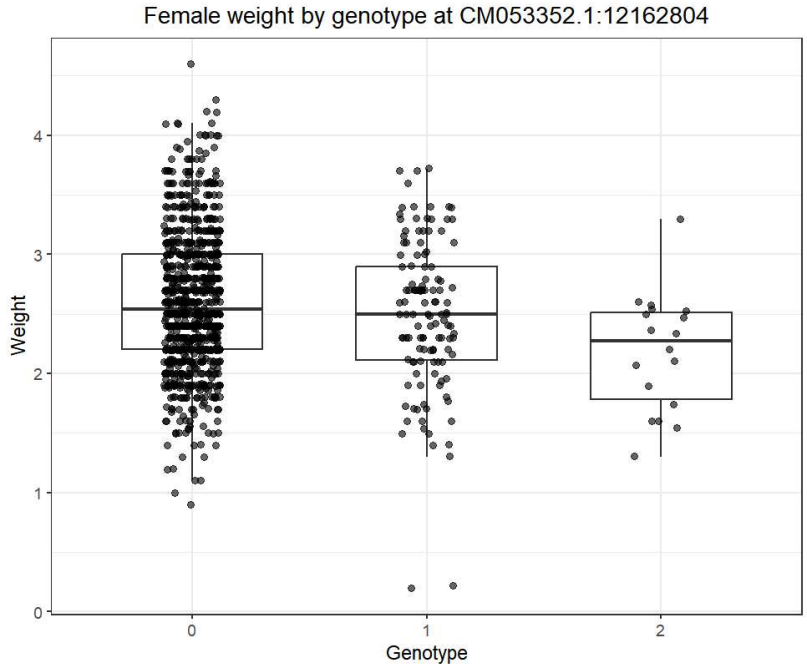


Figure 11 | Female weights sorted by genotype. Fish were grouped by imputed genotype at SNP, chr14:12162804, the locus most strongly associated with weight variation in target population females ($p = 1.11 \times 10^{-9}$). This is an intron 13 SNP in the macrophage mannose receptor 1 (*mrc1*) gene. Points represent individual fish, and boxes show the median weight in kilograms and interquartile range. Mean weights for genotypes CC, CA and AA were 2.62 Kg (N=885), 2.49 Kg (N=128), and 2.18 Kg (N=18) respectively.

Genome Prediction. A set of SNPs identified from the QQ plot (Fig. 9d) were employed to develop a weight prediction model. Using a $-\log_{10}(P)$ threshold of 2.5, we identified 3829 loci for elastic-net model testing. The model included sex as a fixed-effect covariate. Sex was directly observed for 603 individuals and assigned based upon the 61-SNP prediction (see above) model for the remaining 755 individuals.

The final prediction model ($\lambda_{min} = 0.016$) retained the sex covariate and 764 of the 3829 GWAS-selected SNPs. Predictive ability of the model was assessed using out-of-fold predictions from 10-fold cross-validation (see Methods) and achieved Pearson correlation coefficient of r = 0.654 (Fig. 12a) and mean squared error of 0.464 (Fig. 12b). The sex covariate (which represents the average difference between males and females and had a β coefficient ≈ −0.46). Sex had approximately five times the impact of any one of the hundreds of SNPs that helped predict growth rate. The two next largest β values were for chr14:12162804 (β ≈ −0.081), and one of the male size-associated SNPs, chr3:6767576 (β ≈ −0.089).

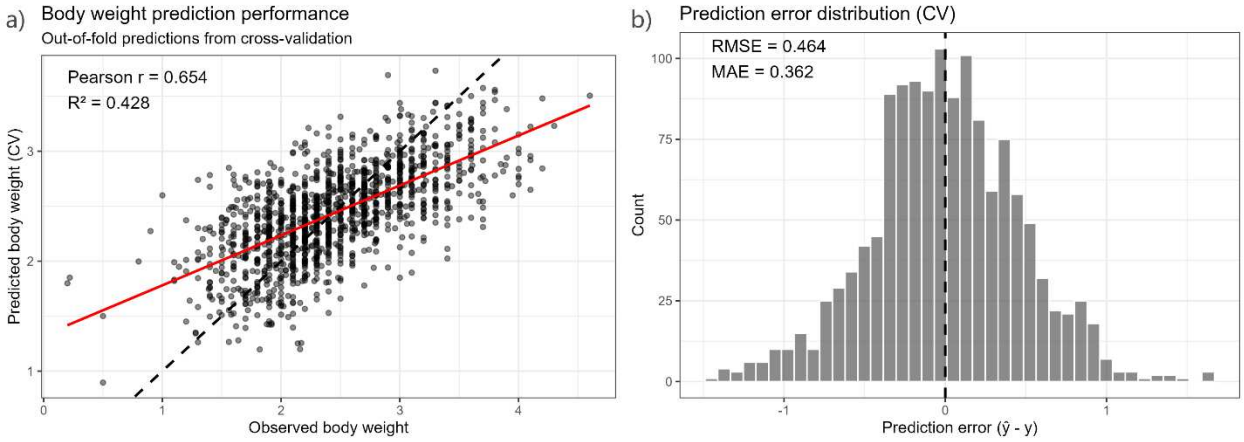


Figure 12 | Genomic prediction of body weight at 2.5 years of age. a) Relationship between observed body weight and out-of-fold predicted values obtained from cross-validation. Each point represents an individual, with predictions generated from models trained without that individual. The dashed black line corresponds to perfect prediction, whereas the solid red line represents the fitted linear regression between observed and predicted values. Prediction accuracy had a Pearson correlation coefficient of r = 0.654, corresponding to an explained variance of R^2 = 0.428, indicating that approximately 43% of phenotypic variation in body weight was captured by a set of markers (including sex) distributed across multiple chromosomes. **b)** Distribution of prediction errors ($\hat{y} - y$) across individuals from cross-validation. The dashed vertical line indicates zero prediction error. Errors were approximately symmetrically distributed around zero, suggesting little systematic bias in prediction.

3.4.3. Disease

GWAS. Moribund fish (N= 220) were sampled during an *Aeromonas salmonicida* outbreak in September and October, 2024 by dip net at the Kyuquot Sound farm site and healthy fish (N = 380) from the same pens were sampled one to four weeks later at harvest. The simple linear

regression model (Fig. 13a) showed individual SNPs and clustered SNPs significantly associated with disease status distributed among multiple chromosomes. When sex was included as a covariate in the multiple linear regression analysis most GWAS peaks remained (Fig. 13b). In the linear mixed-model analysis no SNPs exceeded the *p*-value threshold, however, clusters of low *p*-value SNPs (Manhattan plot peaks) remained on chromosomes 8, 11, and especially 20 (Fig. 13c).

The five lowest *p*-value variants from the linear mixed model analyses (chr20:13923794; chr20:13927614; chr20:13958888; chr20:13963124 and chr20:13970562) delimited a 46 Kb region on chromosome 20. Genotypes of the lowest *p*-value SNP, chr20:13927614 were correlated with genotypes of chr20:13923794, chr20:13958888, and chr20:13963124 (0.68 to 0.79), indicating that these markers collectively tag local haplotypic variation. There were no annotated genes in this region, but RNAseq data displayed by the NCBI genome data viewer provided evidence for unannotated exons: A BLASTx search showed that DNA upstream of the first SNP (chr20:13920529-13921960) encoded an ortholog of the Mango tilapia (*Sarotherodon galilaeus*) *rabankyrin-5* gene. A BLASTx search using putative exon-coding DNA spanning chr20:13925200-13926200 uncovered hits to a THAP4 domain-containing protein (peroxynitrite isomerase THAP-like, *thap4*). The amino acid translation was highly conserved across a great diversity of fish species. BLASTx using chr20:13926700-13929800, as a query sequence (includes the lowest *p*-value SNP position) uncovered hits to uncharacterized proteins that contain a 'sterile alpha-motif' (SAM) domain. This four exon gene (UCSC genome browser Augustus prediction g21441.t1 in sablefish) is found in a great diversity of fishes and in frogs but seemingly absent from other tetrapods. The best BLASTx hit to genomic exon-encoding DNA between the last two low *p*-value SNPs, (chr20:13965300-13968400) was KAO0271113.1 from the koaro, a 'climbing galaxia' (*Galaxias brevipinnis*). This gene is highly conserved in teleosts at the sequence level and has homologs with high identity in amphibians and reptiles but was described as 'uncharacterized' in all taxa.

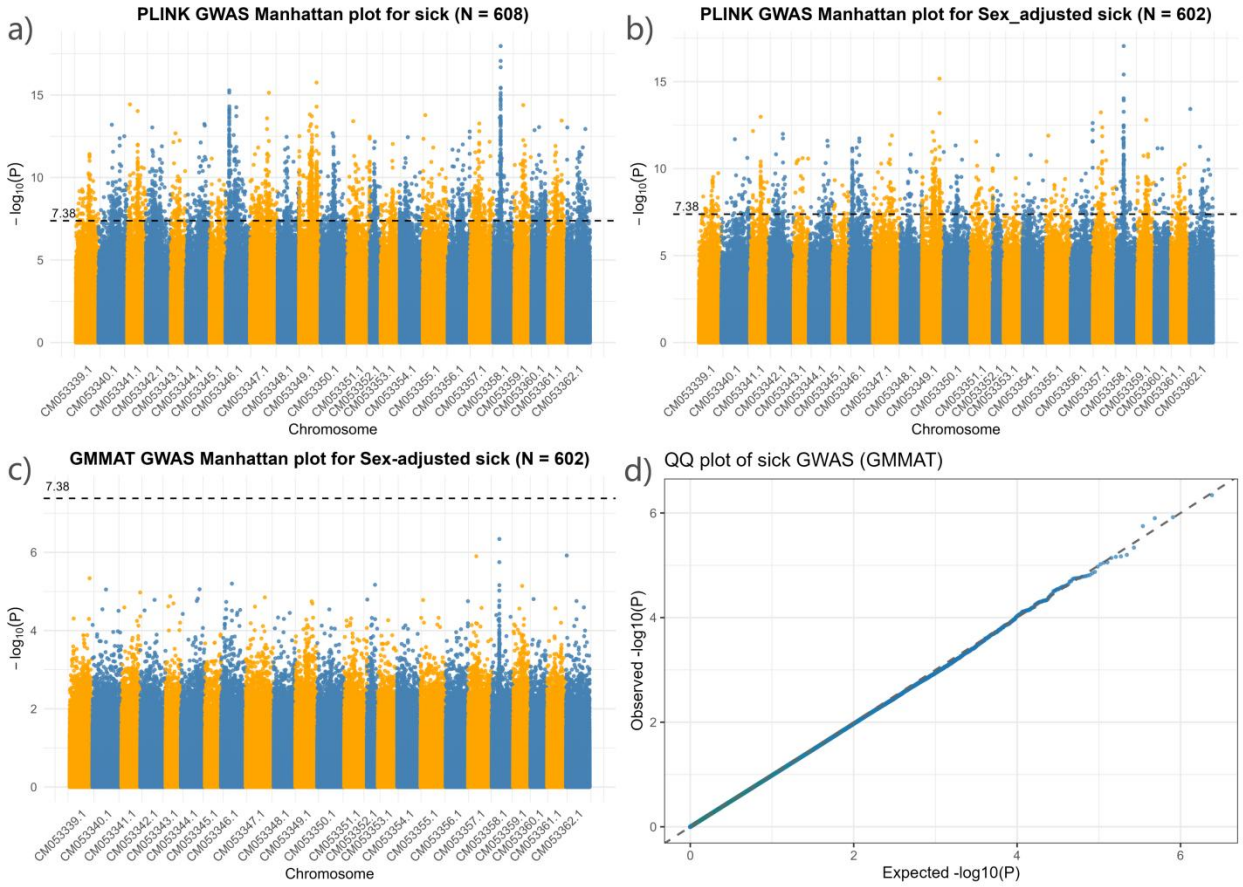


Figure 13 | Genome-wide association analysis of disease status in sablefish. **a)** Manhattan plot showing loci associated with disease status (i.e., moribund (N=220) vs healthy sablefish (N=380)). These data were derived from a simple logistic regression analysis implemented in PLINK. **b)** Manhattan plot showing the results of the multiple linear model analyses (with sex included as a covariate). **c)** Manhattan plot of disease status using a logistic mixed model implemented in GMMAT, which included the Genomic Relatedness Matrix as a random effect. **d)** Quantile–quantile (QQ) plot of the GMMAT disease GWAS showed that the *p*-value distribution largely follows the null expectation with 12 SNPs showing (qualitatively) lower or higher *p*-values than expected.

To provide an alternative perspective on these associations, we calculated odds ratios (ORs) for the five lowest *p*-value SNPs. Individuals with the homozygous reference genotype were considered 'unexposed' and they were compared to individuals 'exposed' to the heterozygous and the homozygous alternative genotypes. OR > 1 indicates higher odds of moribund status in the exposed individuals, whereas OR < 1 indicates that exposure to genotypes with one or two copies of the non-reference allele lowers the odds of moribund status (i.e., is protective). The genotype-specific ORs showed clear differences between moribund and healthy fish (Fig. 14). For chr20:13927614, chr20:13923794, chr20:13958888, and chr20:13963124, the heterozygous

and homozygous alternative genotypes had higher odds of moribund status than the homozygous reference genotype. By contrast, at chr20:13970562 the reference genotype A/A was more common in moribund fish, and the A/G and G/G genotypes had lower odds of moribund status (OR = 0.41 and 0.10, respectively). These OR data provide an interpretable summary of the genotype-status differences underlying the GWAS signals.

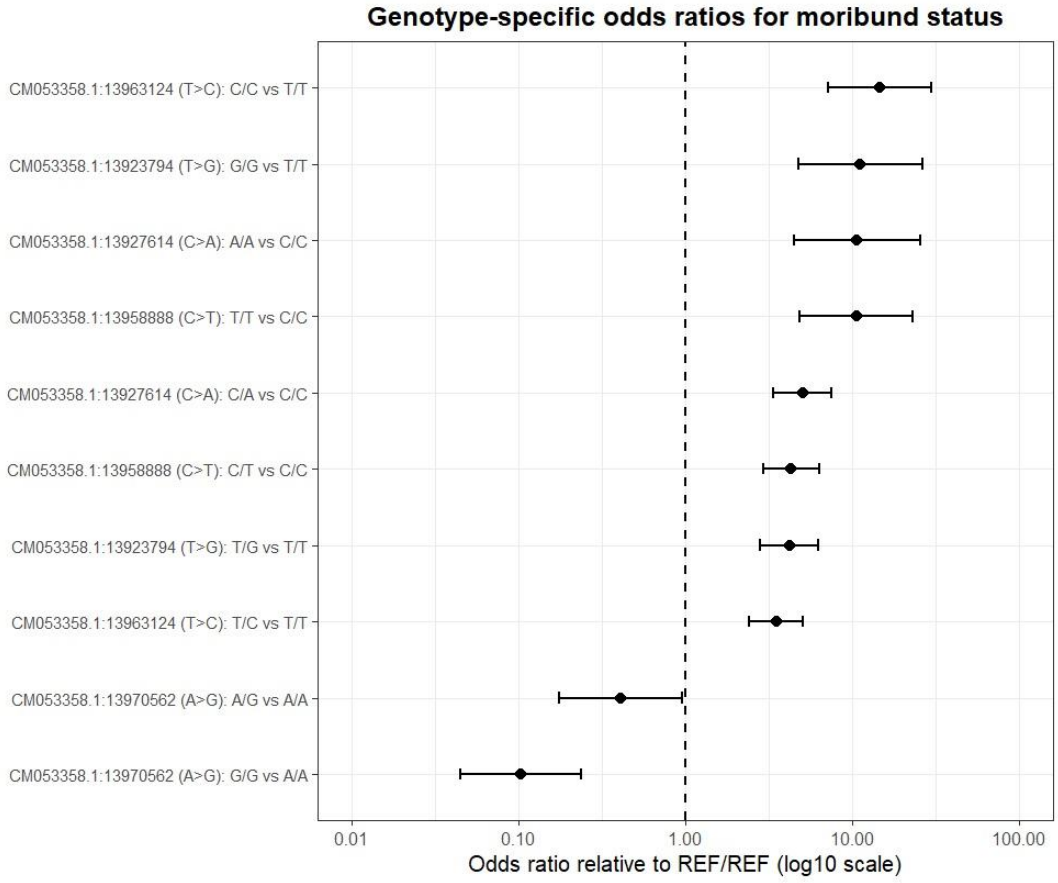


Figure 14 | Genotype-specific odds ratios for the five chromosome-20 SNPs with the lowest GWAS *p*-values. For each SNP, the allele notation in parentheses indicates the reference and alternative alleles, and each comparison contrasts the heterozygous or homozygous alternative genotype with the homozygous reference genotype. Black dots show unadjusted odds-ratio estimates calculated directly from 2 × 2 genotype-by-status (sick vs healthy), and horizontal bars show Wald-type 95% confidence intervals. The vertical dashed line indicates no difference in moribund odds relative to the homozygous reference genotype (OR = 1). The x-axis is shown on a logarithmic scale because ORs are multiplicative; for example, OR = 0.1 and OR = 10 represent reciprocal changes in odds and are displayed symmetrically around OR = 1. For four of the five loci, the homozygous alternative genotypes had the highest risk (as estimated using odd ratios): chr20:13963124 C/C vs T/T (OR = 14.59), chr20:13923794 G/G vs T/T (OR = 11.19), chr20:13927614 A/A vs C/C (OR = 10.68), and chr20:13958888 T/T vs C/C (OR = 10.58).

Genome Prediction. Next, we constructed a disease status prediction model using SNPs selected from the GWAS *p*-value QQ plot (Fig. 13d). The QQ curve showed a deviation beginning at $-\log_{10}(P) = 5$ (considered as a rank-based extreme tail) exposing 12 SNPs as candidate contributors to disease status for Elastic Net regularized regression. Disease status in 608 individuals was coded as 0 (healthy – sampled at processing) and 1 (moribund – dip netted at the farm site). Model tuning was conducted via 10-fold cross-validation using misclassification error as the optimization criterion. The optimal elastic-net model retained 10 predictors (SNPs) and had a $\lambda_{min} = 0.032$, with a slightly more conservative $\lambda_{1se} = 0.091$. When evaluated on the training data, the fitted model achieved a classification accuracy of 0.76. Model performance was also assessed using out-of-fold predictions from cross-validation. Under this more stringent evaluation, the elastic-net classifier retained 10 SNPs (distributed among chromosomes 8, 11, and 20) to correctly classified 354 of 388 healthy fish and 79 of 220 sick fish with 71% accuracy (total 433 of 608 individuals).

4. Discussion

Low-depth whole-genome sequencing combined with genotype imputation provides a cost-effective strategy for generating dense genome-wide marker datasets for GWAS and GP in a diversity of species where high-depth sequencing of large breeding populations remains prohibitively expensive (Auton et al., 2015; McCarthy et al., 2016). In aquaculture, genotype imputation has been applied to a diversity of species. In this study, 65 wild-caught sablefish were sequenced at high depth to generate a haplotype reference panel that was used to impute genotypes in a target population (N=1380) sequenced at low-depth. These reference panel data also provided insight into genetic variation in sablefish at a level of resolution not previously possible. We observed 4.9 million SNPs in the wild-caught reference panel population (65 individuals). With a 600 MB genome this equals an average of 8.17 SNPs/Kb. This SNP density estimate is likely to be a good proxy for the species. Broad geographic surveys have shown that sablefish along the west coast of North America are a panmictic population (Timm et al., 2024). In addition, despite the fact that the reference panel fish were all collected in British Columbia, sampling spanned several years. The level of variation observed in sablefish is almost exactly what was observed by Logsdon et al. (2025) in a set of 65 human genomes, but approximately half as many as observed by Zhang et al (2024) in a survey of 20 large scale loach (*Paramisgurnus dabryanus*) (17.6 SNPs/Kb).

SNP density in the reference population varied within and among chromosomes with more SNPs in telomeric regions and a much lower number of SNPs on chromosome 14 (which contains the sex-determining locus, *gsdf*) than on autosomes. Sex chromosomes have lower effective population sizes than autosomes, 0.75 for the X and 0.25 for the Y. In species where reproduction is restricted (e.g., to a few dominant males, for example) sex chromosome diversity is even lower than predicted by reduced effective population size (e.g., humans: Wilson Sayres et al., 2014; bighorn sheep: Kardos et al., 2015). The observation that SNP density for chromosome 14 was more than half that of the autosomes indicates that factors such as sex-specific variance in reproductive success and/or sex-bias migration are not key features of sablefish life history. There were fewer males than females in the reference panel (and thus fewer chr14Ys than chr14Xs) so the comparatively high value might be an under-estimate. Missing

and inaccurate SNP genotypes in the target population (sequenced at an average of 0.4x) were imputed using the reference panel haplotypes (derived from 14x WGS data). Imputation accuracy was assesses using two concordance tests. For 29 reference panel individuals genotype data were available from two sources, 14x WGS data and a PCR-based SNP genotyping panel. Concordance was high, with 98% of the loci scored the same. When 0.4x-derived genotypes were compared to those obtained from the same PCR-based genotyping panel (754 target population individuals) concordance was 60%. Where genotypes differed, imputation tended to score heterozygotes as homozygotes. These observations focused our attention on GWAS peaks so as not to be misled by erroneous single GWAS-significant SNPs. To evaluate reference panel imputation performance, read data from five individuals was sampled at 0.5x, 1x, 2x and 5x (one fish at a time) and imputed using a 64-individual reference panel that excluded that individual. The greatest improvement was realized at the lowest depth, with concordance changing from 27% before imputation to 64% after imputation. By contrast, concordance was much higher before imputation for the 5x data, 72%, and increased to 83% after imputation.

GWAS analyses using imputed 0.4x data from the target population individuals revealed genomic regions that differed between males and females, among fish that varied in weight (at age 2.5 years), and between fish that were moribund vs healthy (collected during an *Aeromonas salmonicida* outbreak). Imputation errors can lead to spurious GWAS results but are unlikely to generate erroneous clusters of significant SNPs (i.e., Manhattan plot peaks). Focusing on these peaks, GWAS exposed candidate genes and provided the raw material for multi-locus genome prediction models that reliably classified sablefish as males or females and as moribund or healthy and predicted weight.

The extent of relatedness among individuals in the target population was difficult to predict prior to sample collection. This is likely to be common in aquaculture, due to multi-parent batch fertilizations, differential offspring survival, and mixing of stock during grow-out (e.g., for size matching). The impact of relatedness on the GWAS analyses in this study was demonstrated in this study by comparing the results from simple linear regression and multiple linear regression with those from a linear mixed model, which accounted for genetic relatedness using a genome relatedness matrix (GRM) as a random effect. GWAS associations for weight and disease susceptibility fell below the *p*-value threshold in the linear mixed model analyses.

Sex

The simple linear regression GWAS for sex identified 21,000 SNP associations across the genome. Some of the families (sets of siblings) in this study were generated by female x neomale crosses and genotypes unique to, or more common in, these all-female families were artificially associated with sex. The linear mixed model analysis, which reduces family effects left only chromosome 14 loci associated with sex. SNPs in the vicinity of *gsdf* had the lowest *p*-values. This was expected, given the role this locus is known to play in sablefish sex determination (Herpin et al., 2021). The observation that SNPs throughout the chromosome were significantly associated sex in the linear mixed model analysis provided insight into the evolution of this chromosome 14 X/Y system. Most other cottoids have an *amhr* duplication-based sex determination system (Rondeau et al., 2016; Holborn et al., 2021). The TE insertion that spawned the derived *gsdf*-based sex-determination system in sablefish (Herpin et al., 2021)

would have been associated with rare alleles on the 'founder' or 'proto' Y that then increased in frequency rapidly via hitchhiking. These alleles do not show evidence of being part of a sex-associated tract of LD, however their frequencies appear not to have equilibrated between chromosome 14Y and chromosome 14X, i.e., there remains a stochastic imbalance in allele frequencies between the sexes. The apparent discrepancy between the broad GWAS signal across chromosome 14 and the narrow signals from F_{ST} analyses results from F_{ST} drawing attention to differences in expected and observed genotypes. For rare alleles (e.g., in one sex) heterozygosity is expected to be very low, that is, not much higher than it would be for a combined population of males and females. Both of these analyses deal with one locus at a time. LD analyses draw our attention to loci with correlated genotypes. A given locus may show have a high F_{ST} value (e.g., between males and females) yet not be in LD with any other locus. The data indicate that recombination outside the core sex-determining region (SDR) has erased rare-allele-LD but has not distributed rare alleles evenly between chromosome 14Y and chromosome 14X.

Weight

Seven SNPs were shown to be significantly associated with weight variation in a study that utilized the 300-locus PCR-based SNP panel (Liu et al., 2026). All were in the Sex Determining Region (SDR). Autosomal SNPs were not significantly associated with weight variation in that study but the success of a multi-locus prediction model (with autosomal SNPs) supported the hypothesis that growth rate was influenced by genetic factors beyond the SDR. By increasing the number of SNPs (to 1.2 million) and the number of fish surveyed (to 1380 individuals) the current study provided an opportunity to more thoroughly test the impacts of autosomal loci on weight variation at age 2.5 years.

When the entire target population was considered, no non-SDR regions were significant after considering relatedness. However, SNPs outside the SDRwere associated with weight in the linear mixed model analysis when the two sexes were analyzed separately. This included loci on chromosomes 3 and 8 in males and on chromosome 9 and 14 in females. As mentioned, the imputed genotypes are known to include errors (concordance was approximately 0.5 when imputed data were compared to reliable PCR-based genotyping panel data) so single-SNP associations have not been explored further. In females, four of the five size-associated SNPs were approximately 40 Kb downstream of *gsdf* in LOC129102365, which BLAST analysis showed was homologous to macrophage mannose receptor 1. The impact of these loci on weight was minor: Average weights of females with each of the three possible genotypes at the lowest *p*-value SNP were 2.62 Kg, 2.49 Kg and 2.18 Kg respectively and only 18 of the 1031 females had the lowest-weight genotype. Macrophage mannose receptor genes are involved in recognition and uptake processes in immune-related cell types, thus LOC129102365 is biologically interesting in the context of growth–immunity trade-offs. Elastic-net GP results showed that autosomal loci contribute to weight variation prediction. Weight was predicted with a Pearson correlation coefficient of $r = 0.65$ and a mean error of 0.46 in both sexes by a 764-SNP model that included loci distributed across all 24 chromosomes. Among the factors that impact growth rate in aquaculture are water temperature and fish density. Both are difficult to control in the ocean pen aquaculture setting. While temperature impacts all fish in a given pen and typically does vary among pens, juveniles from different implants are introduced to the pens at

different times in the season. Age-specific impacts of temperature on growth, if present, were not considered for in these analyses. From a genetic perspective, variation in growth rate among captive sablefish was largely predicted by sex. Autosomal SNPs improved weight prediction models, compared to the 300-SNP GP (the Pearson correlation coefficient improved from r = 0.49 to r = 0.65). This study also had a larger sample size, which might have provided clearer evidence of minor autosomal contributions.

Disease susceptibility

Sablefish sampled by dip net from the water surface during a furunculosis outbreak were labeled as moribund and healthy individuals from the same pens were sampled from one to four weeks later during processing. Loci with genotypes that differed between the two groups occurred in a GWAS peak on chromosome 20. While not significant in the linear mixed model analysis (which included a GRM as a random variable) a sub-threshold peak remained with the lowest *p*-value SNP occurring in an uncharacterized SAM domain-containing protein. That a SAM-domain containing protein appears to be associated with *Aeromonas* susceptibility is intriguing because *A. salmonicida* infects phagocytic macrophages and survives in ACVs (*Aeromonas*-containing vesicles) by disrupting lysosomal maturation (endosome/vesicle fusion) and lysosomal signaling (Vasquez et al., 2025). As 'scaffolds' that facilitate protein oligomerization, SAM domain-containing proteins have key roles in both lysosomal maturation and ubiquitination. Thus, further characterization of the chromosome 20, uncharacterized protein is warranted. The SNPs associated with increased risk appear to occur outside exons. Odds-ratio estimates for the five top-ranked chromosome 20 variants showed large differences in genotype frequencies between moribund and healthy fish. In four of the five loci the alternative allele increased risk (as estimated using OR) in an additive manner.

Limitations

This study shows that ultra-low-depth WGS when combined with imputation using a relatively small reference panel can be used to identify marker and candidate genes associated with important economical traits in aquaculture populations with high levels of relatedness. Associations and genome prediction models generated from these data exposed several candidate genes and may be used to select a subset of brood stock individuals for breeding.

However, several observations exposed deficiencies that could be moderated for future imputation-based GWAS and GP in sablefish. The variation among individuals in sequencing depth, for the reference panel and the target populations, was high. For some target populations individuals in particular, imputed genotypes were assembled from very little raw data. This will have resulted in imputed genotypes having common allele haplotypes and, indeed, likely explains the observation that some target population individuals appeared to be almost identical in the GRM heatmap.

In addition, the target population included a large number of related individuals. While predicting relatedness prior to sampling can be complicated by factors listed above, a sampling regime that is partitioned evenly over the course of a breeding cycle (maximizing the number of Implants surveyed) could be employed to enhance sample diversity. The study was carried out

as neo-male breeding was becoming standard practice at Golden Eagle Sablefish and as a result, the impacts of family structure were exaggerated by the fact that some families were comprised of females only. This clearly had an impact on GWAS for sex and size when standard linear models were employed.

The chromosome 20 signal for furunculosis resistance/susceptibility can be used to help identify female candidates for neo-male production and when the diversity in the neo-male pool is enhanced, GWAS for size can be employed to further identify autosomal contributors to female growth rate in a genetically diverse all-female target population.

Concordance experiments that employed down sampled data from five reference panel fish indicated provided insight into the ability of a small but diverse panel to correctly imputed low-depth data. The improvement from 38% to 71% correct genotypes represented the addition of __ SNPs into a 1x survey (for example) and is, therefore an important resource for future sablefish GWAS and GP studies.

Finally, using deflated *p*-value SNPs in a GP model would typically be avoided. Our methods focused on SNP that deviate from the diagonal, including the deflated SNPs (higher *p*-values than null, not lower). These loci appear to have improved prediction. A possible explanation for this is 'Suppression Re-emergence': These SNPs were suppressed in the linear mixed model analysis because they were associated with disease in a subset of related individuals and this led to higher *p*-values than predicted by chance. But when a model containing these SNPs was tested against the dataset (without considering family structure) they improved performance because they predicted status in a large enough proportion of individuals (albeit siblings). Thus, the disease susceptibility prediction model will need to be validated in additional populations.

Declarations

Funding Sources. The authors acknowledge generous financial support from NSERC Canada (Alliance program), MITACS, Golden Eagle Sablefish, DFO (Canada)/BCSRIF and the University of Victoria Faculty of Science.

Conflicts of Interest. The authors declare no conflicts of interest.

Ethics approval and consent to participate. Not applicable.

Data Statement. Raw sequencing SRA database for all sablefish genotyped in this study available on https://www.ncbi.nlm.nih.gov/bioproject/1300457

Consent for publication. Not applicable.

Acknowledgements. The authors thank staff at Golden Eagle Sablefish (hatchery and farm site), St. Jeans Cannery and Smokehouse, and The Center for Aquaculture Technologies for assistance with sablefish sampling, DNA isolation and SNP-panel genotyping.

Author Contributions. Study conceptualization: B. Campbell and J.S. Taylor. Sablefish sampling: B. Campbell and J.S. Taylor and Isabel Ma and Meara McIntosh. GWAS and Genome Prediction analyses carried out by Z. Liu, X. Zhang and J.S. Taylor. Juvenile sex test PCR carried out by N. Mokariasl, B. Raeisi, M. Inch and H.V. Barnes. Manuscript review and editing: Z. Liu, J.S. Taylor, X. Zhang, L. Mary.

References.

1. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. 101, 5–22 (2017).

2. Meuwissen, T. H. E. et al. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829 (2001).

3. Houston, R. D. et al. Harnessing genomics to fast-track genetic improvement in aquaculture. Nat. Rev. Genet. 21, 389–409 (2020).

4. Yáñez, J. M., Newman, S. & Houston, R. D. Genomics in aquaculture to better understand species biology and accelerate genetic progress. Front. Genet. 6, 128 (2015).

5. Li, Y. et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816–834 (2010).

6. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98, 116–126 (2016).

7. Pasaniuc, B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. 44, 631–635 (2012).

8. Rubinacci, S. et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. 53, 120–126 (2021).

9. Daetwyler, H. D. et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858–865 (2014).

10. VanRaden, P. M. et al. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol. 49, 32 (2017).

11. Veerkamp, R. F. et al. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein–Friesian cattle. Genet. Sel. Evol. 48, 95 (2016).

12. Liu, Y. et al. Increased accuracy of genomic prediction using preselected SNPs from GWAS with imputed whole-genome sequence data in pigs. Animals 13, 3871 (2023).

13. Zhao, H. et al. Using genotype imputation to integrate canola populations for genome-wide association and genomic prediction of blackleg resistance. BMC Genomics 26, 215 (2025).

14. Sun, Y. et al. Genotype imputation-based whole-genome association study for growth-related traits in golden pompano (Trachinotus ovatus). Aquaculture 596, 741893 (2025).

15. Zhang, C. et al. Optimizing genotype imputation pipeline for low-coverage whole genome sequencing data in spotted sea bass and its application in genomic prediction. Aquac. Rep. 45, 103088 (2025).

16. Goddard, M. E. et al. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. Proc. R. Soc. B 283, 20160569 (2016).

17. Ødegård, J. et al. Genomic prediction in an admixed population of Atlantic salmon (Salmo salar). Front. Genet. 5, 402 (2014).

18. Sasaki, T. Studies on the sablefish (Anoplopoma fimbria) resources of the North Pacific Ocean. Bull. Far Seas Fish. Res. Lab. (1985).

19. Lacko, L. C. Summary of the annual 2020 sablefish (Anoplopoma fimbria) trap survey, October 7–November 21, 2020 (Fisheries and Oceans Canada, 2021).

20. Fisheries and Oceans Canada. Pacific Region integrated fisheries management plan, groundfish, effective February 21, 2025 (Fisheries and Oceans Canada, 2025).

21. Cook, M. A. et al. Culture of sablefish (Anoplopoma fimbria) larvae in four experimental tank designs. Aquacult. Eng. 69, 43–49 (2015).

22. Goetz, F. W. et al. Status of sablefish, Anoplopoma fimbria, aquaculture. J. World Aquacult. Soc. 52, 607–646 (2021).

23. Gutierrez, A. P. et al. Genome-wide association study (GWAS) for growth rate and age at sexual maturation in Atlantic salmon (Salmo salar). PLoS ONE 10, e0119730 (2015).

24. Omeka, W. K. M. et al. Genome-wide association study (GWAS) of growth traits in olive flounder (Paralichthys olivaceus). Aquaculture 555, 738257 (2022).

25. Su, S. et al. Genome wide analysis for growth at two growth stages in a new fast-growing common carp strain (Cyprinus carpio L.). Sci. Rep. 10, 7259 (2020).

26. Zhou, M. et al. The study of the genomic selection of white gill disease resistance in large yellow croaker (Larimichthys crocea). Aquaculture 574, 739682 (2023).

27. Liu, Z. et al. Predicting growth rate in sablefish (Anoplopoma fimbria) using a 300-locus SNP panel. Aquaculture 743313 (2025).

28. Flores, A.-M. et al. Sablefish (Anoplopoma fimbria) chromosome-level genome assembly. G3 (Bethesda) 13, jkad089 (2023).

29. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

30. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).

31. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet. 98, 653–666 (2016).

32. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

33. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

34. Picard Toolkit. Picard toolkit. Broad Institute, GitHub repository (2019).

35. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).

36. Delaneau, O. et al. Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 16, 1161–1169 (2019).

37. Rubinacci, S. et al. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. Nat. Genet. 53, 120–126 (2021).

38. Piferrer, F. Endocrine sex control strategies for the feminization of teleost fish. Aquaculture 197, 229–281 (2001).

39. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2024).

40. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320 (2005).

41. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., et al., 2015. A global reference for human genetic variation. Nature 526 (7571), 68–74. https://doi.org/10.1038/nature15393.

42. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., et al., 2016. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. 48 (10), 1279–1283. https://doi.org/10.1038/ng.3643.

43. Herpin A, Schartl M, Depincé A, et al. Allelic diversification after transposable element exaptation promoted gsdf as the master sex determining gene of sablefish[J]. Genome research, 2021, 31(8): 1366.

44. Chen Z, Blommaert J, Mei Y, et al. Machine learning for genomic prediction of growth traits in aquaculture: a case study of the Australasian snapper (*Chrysophrys auratus*). BMC bioinformatics, 2025, 26(1): 278.