博士論文(要約)

Genetic basis and genetic improvement of heterobothriosis resistance and growth performance in the tiger pufferfish, *Takifugu rubripes*, fed with low fishmeal diet

(低魚粉飼料給餌下におけるトラフグのヘテロボツリウム症耐性と成長能力に関する遺伝基盤と選抜育種に関する研究)

リン　コケツ

林 子杰

Lin Zijie

Contents

Introduction

Fishmeal shortage for aquafeeds and its possible solutions

Fishmeal produced from wild-captured fish is an important component of diets used in the aquaculture industry. However, the rising price of fishmeal and the necessity for sustainable fishing have encouraged the aquaculture industry to explore alternative protein resources (Hua et al., 2019). To reduce the amount of fishmeal, a wide range of potential replacements has been investigated, such as plant meal, animal byproducts, fishery and aquaculture byproducts, and insect meal. These have been tested on a variety of aquaculture species, e.g., Atlantic salmon (*Salmo salar*) (Belghit et al., 2019; Caballero-Solares et al., 2018a; Davidson et al., 2016; Egerton et al., 2020), Japanese flounder (*Paralichthys olivaceus*) (Kikuchi et al., 1993), rainbow trout (*Oncorhynchus mykiss*) (Adelizi et al., 1998; Rimoldi et al., 2021; Satoh et al., 2003; Yoshitomi et al., 2007), yellowtail (*Seriola quinqueradiata*) (Ido et al., 2021; Murashita et al., 2019), red sea bream (*Pagrus major*) (Seong et al., 2019), European sea bass (*Dicentrarchus labrax*) (Kaushik et al., 2004; Magalhães et al., 2017; Rimoldi et al., 2020; Serradell et al., 2020), and white shrimp (*Litopenaeus vannamei*) (Hernández et al., 2008; Motte et al., 2019; Tan et al., 2005; Xie et al., 2016, 2014).

To further reduce the environmental footprint of the aquaculture industry, single cell protein biomass from microorganisms such as bacteria, yeasts, and microalgae have been suggested as sustainable and cost effective replacements for fishmeal (Cottrell et al., 2020; Gamboa-Delgado and Márquez-Reyes, 2018; Matassa et al., 2016). Bacteria and yeasts have high protein contents and excellent nutritional profiles, which can be additionally improved by adjustment of culture methods or genetic engineering (Agboola et al., 2021; Wang et al., 2020). Recently, the use of bacteria and yeast-based diets as major protein ingredients has been explored in Atlantic salmon (Couture et al., 2019), rainbow trout (Roques et al., 2018), red sea bream (Takii et al., 2004), and white shrimp (Chen et al., 2021; Zhao et al., 2017). The availability of low fishmeal (LFM) diet would be largely depending on species specific characteristics of digestive system (Hua et al., 2019). In the worst case, LFM diets may diminish production traits (e.g., growth performance and disease resistance) of farmed fish due to the gene-environment interaction, that is, genes may have a different response under a different environment (Zhang and Belsky, 2020). On the other hand, it is highly expected that the LFM tolerance (i.e., production traits under the dietary treatment of LFM diet) can be genetically improved by means of selective breeding, as a complementary strategy for the LFM diet formulation to solve the problem of fishmeal shortage (Hua et al., 2019). Some preliminary studies documented success of genetic improvement for growth-related traits under the plant-based LFM diet, such as rainbow trout (Callet et al., 2017; Miura et al., 2020) and Amago salmon (*Oncorhynchus masou ishikawae*) (Yamamoto et al., 2015, 2016). However, the feasibility of selective breeding for the LFM tolerance has not been systematically examined.

Selective breeding in aquaculture

Selective breeding is the process to selectively improve particular traits through recurrently mating high-potential individuals and producing genetically superior progenies. This cumulative genetic gains bring about high economic returns (Oldenbroek and van der Waaij, 2015). So far, selective breeding significantly

accelerates the food production of farm animals and major crops, while its application and progress in aquaculture lags far behind, i.e., only 10% of aquaculture production is derived from selective breeding programs in 2012 (Gjedrem et al., 2012). Fortunately, for aquaculture species, the genetic improvement of economically important traits usually have higher genetic gain due to the high fecundity and genetic diversity compared to the terrestrial animals (Gjedrem and Baranski, 2009). Moreover, the production traits (i.e., growth performance, disease resistance, etc.) of aquaculture species commonly show moderate or high heritability, suggesting high potential for genetic improvement (Elaswad and Dunham, 2018; Gjedrem et al., 2012; Gjedrem and Rye, 2018; Hosoya et al., 2017). Consequently, the benefits of selective breeding have been widely recognized and it is routinely practiced in several species, such as, Atlantic salmon, rainbow trout, and Nile tilapia (*Oreochromis niloticus*) (reviewed in Houston et al., 2020). However, systematic selective breeding programs is still under development or even not existing for most farmed fish diminishing the farming efficiency (Gjedrem and Baranski, 2009). Considering the rapidly growth of aquaculture industry, cost-efficient selective breeding programs are highly demanded to establish the elite fish breeds for the most of fish species (Gjedrem et al., 2012).

Selection methods

To initiate a cost effective selective breeding program, it is essential to choose an optimal selection method, including mass selection, pedigree-based selection, marker-assisted selection and genomic selection (Oldenbroek and van der Waaij, 2015). While mass selection (also known as phenotypic selection) has been practiced since early prehistory, the first scientific attempt of mass selection was done by Robert Bakewell for terrestrial animals in the 18th century (Frana, 2003). Mass selection is based solely on phenotypes of the target traits (e.g., body size). Phenotypic values are not only the measurements of traits but also the consequences of gene-environment interaction (Oldenbroek and van der Waaij, 2015). Thus, phenotype can be further considered as an approximation of the genetic performance of the progeny. Although mass selection outperforms in cost efficiency, this method is ineffective for traits with low heritability because the genetic variance and environmental variance of the targeted traits are not differentiated. And worse still, mass selection lacks in detailed information for inbreeding control (Bentsen and Olesen, 2002).

With the developments of population genetics in the early 20th century, the pedigree-based selection was pioneered by Sewall Wright and Jay Laurence Lush (Gjedrem and Baranski, 2009), and further sophisticated by Charles Roy Henderson who developed the linear mixed model equations to solve best linear unbiased predictions (BLUP) of breeding values (Henderson, 1976, 1953). Pedigree-based selection trains a linear mixed model using phenotypes and pedigree information to estimate breeding values of each individual, enabling breeders to rank the candidate animals for selection. Compared to mass selection, pedigree-based selection is advantageous in inbreeding control and selection accuracy as this method can explicitly separate the phenotype into environmental component and genetic component (breeding value). Pedigree-based breeding methods have contributed to aquaculture development by improving economically important traits, as seen in the salmonids and tilapias (K. Janssen et al., 2017; Neira, 2010; Rye et al., 2010). However, pedigree-based methods have innate drawbacks where it is assumed that estimated breeding values of target traits for candidate individuals are the average breeding values of parents, ignoring stochastic Mendelian segregation (Mendelian sampling) within families (Gjedrem and Baranski, 2009). Thus, pedigree-based methods can not differentiate

estimated breeding values among full sibs. Prediction using large-scale pedigrees including many full- and half-sibs can solve this problem (Walsh and Lynch, 2000). However, collecting large-scale pedigree information is time-consuming and laborious especially in the practice of aquaculture because larvae are too small for tagging and thus it is necessary to keep each family in separate until fish reaches a body size large enough for tagging.

With the advent and development of DNA-based genetic markers, the association between traits of interest and those genetic material are detectable, and thus marker-assisted selection (MAS) become feasible (Wakchaure and Ganguly, 2015). This method is effective when the target trait is determined by a few loci (e. g. quantitative trait loci (QTL)) with large effects (Lande and Thompson, 1990); at least two strains have been established by means of MAS, i.e., lymphocystis resistant Japanese flounder (*Paralichthys olivaceus*) (Fuji et al., 2007) and infectious pancreatic necrosis resistant Atlantic salmon (Moen et al., 2015). However, it is often hard to find DNA markers that can explain a high proportion of genetic variance, since most of economic traits are polygenic and have complex genetic architecture in aquaculture species (Goddard and Hayes, 2009).

In 2001, Meuwissen *et al* proposed genomic selection (GS) to estimate the genomic estimated breeding values (GEBVs) of selection candidates by harnessing whole-genome high-density markers and advanced regression methods (Meuwissen et al., 2001). In GS, those markers are effective in handling errors due to Mendelian sampling by capturing genetic variance at DNA levels. Thanks to the recent advances in the next-generation sequencing (NGS) technologies, it is now affordable to genotype genome-wide single nucleotide polymorphisms (SNPs) for GS in aquaculture breeding programs (Robledo et al., 2018b). As expected, the greater performance of GS over the pedigree-based method in prediction and inbreeding control has been demonstrated by empirical studies (Tsai et al., 2015; Vallejo et al., 2017). Recently, the potential of GS for disease resistance has been seen in amoebic gill disease in Atlantic salmon (*Salmo salar*) (Robledo et al., 2018a), bacterial cold water disease resistance in rainbow trout (*Oncorhynchus mykiss*) (Vallejo et al., 2017), viral nervous necrosis in European sea bass (*Dicentrarchus labrax*) (Palaiokostas et al., 2018), and photobacteriosis in gilthead seabream (*Sparus aurata*) (Palaiokostas et al., 2016). Although the majority of the current breeding programs in aquaculture is still in its infancy, GS is highly expected to accelerate the establishment of high-performance strains (Hosoya et al., 2017).

Demand for genetic improvement in the farmed tiger pufferfish

The tiger pufferfish, *Takifugu rubripes*, is a delicacy and one of the most valuable marine fish species in Japanese aquaculture, ranking fourth in production value among cultured finfish (Hosoya et al., 2014; Ogawa, 2016). The farming efficiency of this species was largely improved by the technology for artificial fertilization developed since 1990s (Miyaki et al., 1998). However, systematic selective breeding program has not been practiced and high-performance fish breeds are not available until recently (at the start of my Ph.D. program). As the current standard feed of the tiger pufferfish containing high level of animal protein (70~80%) mainly from the fishmeal, genetic improvement of production traits under a LFM diet is highly expected to further improve farming efficiency and sustainability of the tiger pufferfish farming (Kikuchi, 2006).

Apart from growth performance, the disease resistance is an important production trait, as disease outbreaks easily hamper the aquaculture industry. One of the most serious diseases in the tiger pufferfish farming is heterobothriosis, which is a gill disease caused by a monogenean parasite *Heterobothrium okamotoi*.

The most severe infection occurs at early phases of production, just after transfer from land-based hatcheries to sea cages (Ogawa, 2002; Ogawa and Inouye, 1997). These naïve juveniles are afflicted by the parasite, persistently present at oceanic aquaculture sites, resulting in retarded growth and high mortality rate (Shirakashi et al., 2010). Therefore, frequent drug treatments are needed to control the parasitosis, which leads to high financial costs, environmental contamination and emergence of drug-resistant parasites (Ogawa, 2016). Instead of the drug treatments, genetic improvement of heterobothriosis resistance in farmed fish is considered as a cost-efficient solution. To establish selective breeding program for the resistance trait, it is essential to know the genetic basis of the trait to choose selection methods while mechanism of host immune response to *H. okamotoi* are still unclear (Igarashi et al., 2017; Matsui et al., 2020). Although the QTL analysis using a full-sib family of F_2 hybrids between a wild female grass pufferfish, *T. niphobles* and a wild male tiger pufferfish suggested that host resistance to heterobothriosis is polygenic (Hosoya et al., 2013), the genetic basis of heterobothriosis resistance in the tiger pufferfish is still unclear. Moreover, the diet change may reshape the genetic basis of these production traits because genes related to the production traits may have different response under different dietary treatments, i.e., genotype–environment interaction. As most of the disease resistance traits have moderate or high heritability in fish species (Ødegård et al., 2011; Robledo et al., 2018a) as well as the growth-related traits (Houston et al., 2020), it is expected that GS can also be applied with these traits in the species. In addition, the genetic resources of this species give a huge advantage to apply GS breeding program: its compact genome (around 400 Mbp) allows GS breeding program with fewer SNPs to reach the same density across the whole-genome compared to the other species; a high-quality genome assembly (FUGU5) simplify SNP panel construction and NGS analysis (Kai et al., 2011; Sato et al., 2019).

Objectives

In this study, my objective is to investigate the genetic basis of disease resistance and body size and the feasibility of selective breeding by means of GS under the dietary treatment of LFM diet using the tiger pufferfish.

In Chapter 1, I first investigated the genetic basis of the two traits and possibility of simultaneous improvements using a small-scale experiment. Then, in Chapter 2, I have enlarged the population size and increased SNP density to examine the generality of previous results and optimal SNP density for accurate GS for these traits. In Chapter 3, I have evaluated the effects of four types of LFM diets on growth performance, blood chemistry, transcriptomic responses in the liver, and resistance to heterobothriosis and determined the best LFM diet. In Chapter 4, I have investigated the genetic basis and feasibility of GS for these traits under the short-tern dietary treatment of the best diet. And finally, in Chapter 5, I have studied the impact of long-tern dietary treatment of LFM diet on genetic basis and feasibility of GS for body size.

Chapter 1. Availability of genomic selection for heterobothriosis resistance and body size under a standard feed

In this chapter, I have investigated heritability, genetic architecture, and the availability of genomic selection (GS) for heterobothriosis resistance and body size in the tiger pufferfish reared with a standard diet using a small population. Heritability (narrow sense) is the ratio of the additive genetic variance to the phenotypic variance, and if the trait is largely determined by the genetic factors, it means the trait have high heritability (de los Campos et al., 2015). Therefore, heritability has been widely recognized as an indicative of the potential for genetic improvements of the target trait (Mathew et al., 2018; Visscher et al., 2008). High heritability, however, does not indicate the existence of large effect genes underlying inter-individual phenotypic differences. Moreover, estimation of heritability does not refer to the underlying genetic architecture, i.e., the number of the genes, the location of the genes on the genome, and the effects of the genes affecting the trait.

The genetic architecture of the target trait is an important factor for the choice of selection method. As the extent of heritability and the number of genes affecting the trait are uncorrelated, genetic architecture should be studied separately from heritability. The first study which reveals genetic architecture of a trait is Mendel's work on pea genetics. The traits which Mendel treated (e.g., pea shape and colors) were binary traits and controlled by single (or a few) large-effect gene(s). It is relatively easy to determine the genetic architecture or detect the genomic position of the causative loci (quantitative/qualitative trait loci, QTLs) of such monogenic traits (also known as Mendelian traits) and oligogenic traits by means of forward genetic approaches, such as QTL analysis and genome-wide association study (GWAS), using experimental crosses (Uffelmann et al., 2021). If large effect QTLs are detected, we can assume that the target trait is monogenic or oligogenic. In such case, marker-assisted selection (MAS), in which breeders select broodstock according to the genotypes at the QTL, has economic advantages. However, it is often the case that large or even medium effect QTLs are not detectable for economic traits of plants, livestock, and aquaculture species as these are underpinned by large number of small-effect genes (i.e., polygenetic traits) (Crossa et al., 2017; Desta and Ortiz, 2014; Goddard and Hayes, 2009; Hosoya et al., 2017; Houston et al., 2020; Xu et al., 2020). In this case, GS has better selection response rather than MAS (Arruda et al., 2016).

Heritability and genetic architecture affect accuracy of genomic prediction (GP) of genomic estimated breeding values (GEBVs). Genomic prediction is the ranking process in GS for the selection of broodfish, and high accuracy of GP suggests high efficiency of GS breeding program (Meuwissen et al., 2001). In the process of GP, breeder creates a training group from a base population and then uses the phenotypes and genotypes of individuals from the group to train a prediction (regression) model. Next, the prediction model is applied to the selection candidates from the base population by substituting genotypes to estimate GEBV, which is a measure of the genetic potential for the trait of the candidate. Choose of a prediction model (i.e., linear mixed models, Bayesian models, machine learning models and deep learning models) also affects performance of GP as each model assumes different genetic architecture (Azodi et al., 2019) (detailed information is referred to in Section 1.2). Therefore, it is important to select the optimal model for each trait to achieve better genetic gain.

As described in Introduction, both heterobothriosis resistance and body size are important production traits for aquaculture of the tiger pufferfish. However, selecting one quantitative trait may improve or diminish others due to the genetic pleiotropy and/or linkage disequilibrium (Lynch and Walsh, 1998). For example, the breeding program which improves the resistance to sea lice possibly diminishes growth-related traits in farmed Atlantic salmon (Gjerde et al., 2011). Likewise, improving resistance to *H. okamotoi* may negatively affect growth-related traits in the tiger pufferfish. Thus, simultaneous genetic improvement of disease resistance and body size is most desirable, although it is complicated by the antagonistic genetic correlation. An index score, namely linear genomic selection index (LGSI), is highly expected to solve this problem (Ceron-Rojas et al., 2015). LGSI is calculated by a linear combination of GEBVs and corresponding weights (i.e., importance of the trait). Therefore, selection based on LGSI can take the importance of each trait into concern, rather than only one of the traits. In this chapter, I have examined the availability of LGSI for simultaneous genetic improvement of the two traits using simulation data.

1.1 Genetic dissection of heterobothriosis resistance and body size traits

To investigate the heritability and the genetic architectures of heterobothriosis resistance and standard length (SL), I have raised a test population derived from wild parents. These fish were subjected to an artificial infection trial to collect phenotypes and genotyped for genome-wide SNPs by means of target amplicon sequencing. These phenotype and genotype information were applied to genetic parameter estimation. Genome-wide association studies was also performed to clarify the genetic architecture of these traits.

Materials and methods

Tested population and artificial infection

The empirical experiments were performed in the Fisheries Laboratory, University of Tokyo (Hamamatsu, Shizuoka Prefecture, Japan). All samples (n = 240) were generated by a full-factorial mating among 10 wild males and 11 wild females, which are commercially caught from Wakasa Bay (Fukui Prefecture, Japan). For the mating, artificial fertilization was applied following the previous study (Kim et al., 2019) with minor modification. In brief, females were anesthetized with 200 mg/l of 2-phenoxyethanol and then ripened by injection of 150 μg/kg of luteinizing hormone-releasing hormone (LHRH, Sigma-Aldrich, St. Louis, MP, USA). Gametes were stripped from each individual and fertilized per male-female pair (110 pairs in total). Rearing and feeding conditions were set as previously described in Hosoya et al., 2014. In brief, fertilized eggs of each maternal half-sib family were mixed and kept in a hatching jar. After hatching, each maternal half-sib was kept in a one-ton tank for one month and then all families were mixed and cultured in a three-ton communal tank. All tanks were supplied with flow-through water and aeration. Fish larvae were fed nutrient-enriched live S-type rotifers, nutrient-enriched *Artemia* nauplii, and commercial pellets according to the developmental stage. At four months old, 240 fish were randomly collected and subjected to an artificial infection test.

To collect the phenotypes of tested fish, artificial infection was done following previous studies (Chigasaki et al., 2000; Kim et al., 2019). A day before the infection, fish were equally distributed into three identical one-ton experimental tanks (80 individuals/tank) supplied with *H. okamotoi*-free fresh seawater (UV treated and filtered). Meanwhile, eggs of *H. okamotoi* were collected from tanks containing infected fish and kept in a glass jar containing fresh seawater until infection. Hatching was induced by physical stimulation (shaking at 140 rpm for 10 min) and the density of oncomiracidia, the free-living larvae of *H. okamotoi*, in the suspensions was determined under the microscope just before the infection. At infection, the water depth of experimental tanks was adjusted to 15 cm, and approximately 4,000 oncomiracidia was introduced into each tank. At 3 h post-exposure, fish were transferred into three newly-setup one-ton holding tanks and reared for 32 days, when *H. okamotoi* reaches maturation and moves to the branchial cavity walls (BCW) (Ogawa, 2016). At the 32-day mark, fish were euthanized, measured for SL and the BCWs dissected from both sides. For each fish, the caudal fin was clipped and kept in 600 μl TNE8U buffer (Asahida et al., 1996) (10 mM Tris-HCl (pH7.5), 125 mM NaCl, 10 mM EDTA, 1% SDS, 8M urea) at room temperature to extract genomic DNA for genotyping. Collected BCWs tissues were kept in 10% formalin until counting the number of parasites under the microscope.

The parasites attached to the whole BCWs were counted under the stereoscopic microscope. The host resistance against *H. okamotoi* is assessed by parasite count on the whole BCWs (HC) (Supplementary Table S1-1).

Genotyping

I applied AmpliSeq technology (Sato et al., 2019), which uses polymerase chain reaction (PCR) to amplify the targeted regions for next generation sequencing. This approach has high repeatability of data and more robustness against *de novo* SNPs since only targeted regions can be consistently amplified by PCR (Sato et al., 2019). Specifically, genomic DNA was extracted using a Gentra Puregene Tissue Kit (QIAGEN, Hilden, Germany) following the manufacture's instruction and applied for AmpliSeq genotyping as previously described (Sato et al., 2019). In short, 3,187 genome-wide target regions were amplified by the first PCR with the custom AmpliSeq primer pools. After the adapter ligation and purification steps, PCR products were barcoded by a second PCR with 8-base index oligo sequences (Supplementary Table S1-1) for individual identification. The libraries of 240 individuals were pooled and sequenced on Illumina MiSeq System using the MiSeq reagent kit v2 (300 cycles) from both ends. The raw FASTQ reads were quality-trimmed using trimmomatic-0.36 (Bolger et al., 2014) with the following parameters: *ILLUMINACLIP TruSeq3-PE-2. fa:2:30:10, LEADING:19, TRAILING:19, CROP:146, HEADCROP:5, SLIDINGWINDOW:30:20, AVGQUAL:20, and MINLEN:60*. Then, trimmed reads were mapped onto the target regions of the reference genome, FUGU5/fr3 (Kai et al., 2011) using BWA-MEM (vv0.7.12) (Li, 2013). Reads with mapping quality values (MAPQ) less than 10 were removed by samtools (v1.7-2) (Li et al., 2009). Genotype calling of each sample was done using GATK-4.1.6.0 (Poplin et al., 2017) *HaplotypeCaller* with the following options: *--output-mode EMIT_ALL_CONFIDENT_SITES -ERC GVCF --stand-call-conf 30*. Obtained gVCF files were combined using GATK *CombineGVCFs* and then joint genotyping was performed using GATK *GenotypeGVCFs*. Variant filtering was done using vcftools (v0.1.5) (Danecek et al., 2011) with the following parameters: *--min-meanDP 15 --max-meanDP 500 --max-missing 0.7 --hwe 0.05 --minDP 10 --remove-indels*. The missing values of genotypes were imputed by LinkImputeR-1.2.1 (Money et al., 2017). First, the accuracy of 2 filters (i.e., the maximum missingness allowed per SNP and sample of 0.9 and 0.95) were tested with the parameters: *numbermaseked = 500*. The default setting was used for the other parameters. Subsequently, missing genotypes were imputed with a better filter (0.9). At first, SNPs which did not fulfill the maximum missingness per SNP and sample of 0.9 were filtered out, and then the missing genotypes were imputed. All samples were retained but 11 out of 6,718 SNPs were discarded. Subsequently, the imputation accuracy was accessed with *numbermasked* option (set as 500). PLINK (v1.0.7) (Purcell et al., 2007) *–recodeA* option was used to generate the allele coding matrix from the imputed VCF files. Command line scripts are supplied as Appendix.

Population structure

Population stratification gives biases to the results of heritability estimations, genome-wide association study (GWAS), and GP (Dandine-Roulland et al., 2016; Liu et al., 2015; Price et al., 2010). This is because the

regression model gives higher weight to the family-specific SNPs rather than SNPs in tight linkage disequilibrium with causative genes when stratification exists. Population structure can be easily grasped by visualization through dimensionality reduction methods, converting high-dimensional genomic data into low-dimensional maps without losing significant structure of the high-dimensional data. Traditionally, it is done by Principal Component Analysis, which reduces high-dimensional dimensions into a few to several principal components that explain the main patterns (Reich et al., 2008). Recently, a nonlinear dimension reduction technique, namely *t*-distributed stochastic neighbor embedding (*t*-SNE), gradually becomes popular in single-cell RNA-seq data analysis (Amir et al., 2013; Van Der Maaten and Hinton, 2008), and also shows good result in genetic studies (Li et al., 2017).

In this study, *t*-SNE was used to visualize population structure of the specimen. First, *t*-SNE transforms the genomic data into conditional probabilities that represent pairwise similarity in the high-dimensional space. Then, transformed data were applied to a heavy-tailed Student's *t*-distribution that measures pairwise similarities of corresponding samples in the low-dimensional embedding space. Finally, *t*-SNE minimized the sum of the Kullback–Leibler divergence between those two probability distributions (Kullback–Leibler divergence is the measure of the difference between two probability distributions.). The *t*-SNE analysis was implemented in *sklearn.manifold.TSNE* class of Python/scikit-learn-0.20.3. The perplexity was set as 20 and default values were used for the other parameters. Command line scripts are supplied as Appendix.

Heritability and genetic correlation

Heritability and genetic correlation were calculated by a multivariate linear mixed model as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i,$$

where $\boldsymbol{y}_i$ is a vector of phenotypes for trait *i* (*i* = 1 for transformed HC since HC is not normally distributed and 2 for SL); $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are incidence matrices for fixed effects $\boldsymbol{\beta}_i$ and random effects $\boldsymbol{u}_i$, respectively. The model assumes the random effects ($\boldsymbol{u}$) follow a multivariate normal distribution as $\boldsymbol{u} = [\boldsymbol{u}_1'\boldsymbol{u}_2']' \sim MVN(0, \boldsymbol{G} \otimes \boldsymbol{A})$ and the residuals ($\boldsymbol{e}$) follow $\boldsymbol{e} = [\boldsymbol{e}_1'\boldsymbol{e}_2']' \sim MVN(0, \boldsymbol{R} \otimes \boldsymbol{I})$; where $\boldsymbol{G}$ and $\boldsymbol{R}$ are the variance-covariance matrices of random effects and residuals for the two traits, respectively; $\boldsymbol{A}$ is the additive genetic relationship matrix constructed by *A.mat* function in R/rrBLUP-4.6 (Endelman, 2011) with the default settings; $\boldsymbol{I}$ is the identity matrix; $\otimes$ means the operation of Kronecker product. The model was solved by *mmer* function in R/sommer-4.0.1(Covarrubias-Pazaran, 2018, 2016) to solve the equations. The heritability (h_i^2) was computed as:

$$h_i^2 = \frac{\sigma_{g_i}^2}{\sigma_{g_i}^2 + \sigma_{e_i}^2},$$

where $\sigma_{g_i}^2$ and $\sigma_{e_i}^2$ are the genetic variance and the residual variance for trait *i*, respectively. Then, the genetic correlation (r_g) was computed as:

$$r_g = \frac{\sigma_{g_1,g_2}}{\sqrt{\sigma_{g_1}^2\sigma_{g_2}^2}},$$

where σ_{g_1,g_2} is the genetic covariance between two traits.

The genetic correlation estimated by means of the multivariate model could be biased from the phenotypic correlation. Therefore, I further tested correlation between GEBV for each trait using GBLUP (genomic best linear unbiased prediction) model. In the prediction model for HC, SL was included as the covariate to minimize non-genetic effects from SL. The prediction models are described as following:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is a vector of phenotypes; $\boldsymbol{X}$ is an incidence matrix for the fixed effect $\boldsymbol{\beta}$ (for the prediction of HC, SL was added as a covariate); $\boldsymbol{Z}$ is an identity matrix for the random effects $\boldsymbol{u}$, which models the breeding values; $\boldsymbol{\varepsilon}$ is a vector of residuals. The normality was assumed for random effects ($\boldsymbol{u}$) and residuals ($\boldsymbol{\varepsilon}$) as $\boldsymbol{u} \sim N(0, \boldsymbol{K}\sigma_u^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$, respectively; where $\boldsymbol{K}$ is a marker-based relationship matrix (Endelman, 2011); $\boldsymbol{I}$ is an identity matrix. $\boldsymbol{K}$ was constructed as described above; GEBVs were estimated by restricted maximum likelihood (REML) algorithm using *kin.blup* function in R/rrBLUP-4.6. Command line scripts are supplied as Appendix.

Genome-wide association study (GWAS)

To investigate the associated markers with transformed HC and SL, GWAS was performed based on the linear mixed model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{S\tau} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the vector of the phenotypes; $\boldsymbol{\beta}$ is a vector of fixed effects other than SNP effects; $\boldsymbol{g}$ is the vector of random effects that models the polygene background; $\boldsymbol{\tau}$ is a vector of fixed effects which represent the additive SNP effects; $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{S}$ are incidence matrices relating to $\boldsymbol{\beta}$, $\boldsymbol{g}$, and $\boldsymbol{\tau}$, respectively. $\boldsymbol{\varepsilon}$ is a vector of normal residuals. $\boldsymbol{g}$ and $\boldsymbol{\varepsilon}$ follow the normal distributions as $\boldsymbol{g} \sim N\left(0, \boldsymbol{K}\sigma_g^2\right)$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$, respectively; where $\boldsymbol{K}$ is the realized relationship matrix described above. A restricted maximum likelihood (REML) algorithm was performed to solve the linear mixed model using *GWAS* function of R/rrBLUP-4.6 with the parameter: n.PC = 10. The *p*-values were calculated for each SNP marker. The Bonferroni-corrected significant threshold was set as $\alpha = 7.45 \times 10^{-6}$ (0.05 divided by the number of SNPs: 6,707).

To further examine the effects of SNP markers, association analysis assuming the following Bayes C model was performed:

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{X\beta} + \sum_{i=1}^{p} \boldsymbol{z}_i\boldsymbol{g}_i + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are same as GBLUP; μ is an intercept; $\mathbf{1}_n$ is a vector of one; p is the total number of SNP loci for each individual; $\boldsymbol{z}_i$ is a scalar of genotypes at SNP i; $\boldsymbol{g}_i$ is a scalar of random effects that represent the genetic effects for SNP i following a mixture of scaled-t distribution and a point of mass at zero. The model was

solved using R/BGLR-1.0.8 (Pérez and De Los Campos, 2014) with nIter = 10,000 and burnIn = 2,000. Command line scripts are supplied as Appendix.

Results

Tested population and artificial infection

Specimens (n = 240) produced by artificially crossing 11 wild males and 10 wild females were subjected to an artificial infection for 37 days at four months old. The phenotypic mean was 15.85 (± 9.15 S.D.) for HC and 9.83 (± 0.78 S.D.) for SL (Figure 1-1 and Supplementary Table S1-1). As the plot shows, the distribution of HC was non-normal (Shapiro–Wilk test: $p = 3.79\times10^{-6}$, alpha level = 0.05) while SL approximated a normal distribution (Shapiro–Wilk test: p = 0.406, alpha level = 0.05). Therefore, I applied a square-root transformation on (HC +1), approximating a normal distribution (Shapiro–Wilk test: p = 0.235, alpha level = 0.05). Transformed HC was used in the following genetic analysis. Weak but significant phenotypic correlation was observed between HC and SL (Pearson's r analysis: r = 0.157, p = 0.015; 95% confidence interval: $0.031 \leq r \leq 0.278$).

Genotyping

The MiSeq sequencing generated an average of 174,870 (± 83,576 S.D.) raw reads per fish. Amplicon sequence reads have been deposited in the DDBJ Sequence Read Archive (DRA Accession: DRA010341) (Supplementary table S1-2). After the quality-trimming step, the mean number of reads for each fish was 161,426 (± 83,576 S.D.) with the mean read length of 124 bp. The survived reads were mapped onto a reference fugu genome (FUGU5/fr3) for SNP calling. Following the quality filtration of SNPs, 6,718 putative SNPs were yielded. Missing SNPs were imputed using LinkImputeR (Money et al., 2017). At this imputation step, 11 SNPs were discarded, and 6,707 imputed SNPs were called for each individual with the imputation accuracy of 0.888.

Population structure

Population structure, which can bias the genetic parameter estimation, was examined by t-SNE analysis (Van Der Maaten and Hinton, 2008) based on SNP data. As seen in Figure 1-2, I did not observe clear clusters or strong stratification within the tested samples. The distribution of HC showed weak linear relationship between both x and y coordinates (Pearson's r = 0.056 and 0.210, respectively). This population structure ensured limited biases to the results of heritability estimations, GWAS, and GP.

Heritability and genetic correlation

To investigate the extent of genetic effects on the phenotypic variation, heritability was estimated by a multivariate linear mixed model. Moderate heritability was observed for each trait (transformed HC: h_2 = 0.308

± 0.123 S.E.; SL: h_2 = 0.405 ± 0.131). With the same model, the strength of the genetic correlation between the transformed HC and SL was also estimated. A moderate antagonistic genetic correlation (r_g = 0.228) was observed between the traits, where large individuals were suffering from higher parasitic loads. This genetic correlation could be, at least partly, due to the phenotypic correlation, although phenotypic correlation between HC and SL was weak as described above. Therefore, I tested correlation between GEBV for each trait using a univariate linear model (i.e., GBLUP); SL was included as the covariate in the prediction model for HC to minimize non-genetic effects from SL. If an antagonistic genetic correlation exists between the two triats, the GEBVs would also show a positive correlation. As the results, I found moderate but significant positive correlation (Pearson's $r = 0.252$, $p = 7.67 \times 10^{-5}$).

Genome-wide association study (GWAS)

GWAS was applied to detect loci highly associated with transformed HC and SL (Figure 1-3). None of these loci exceeded the significance threshold of 7.45 (= $-\log_{10}$ (0.05/6,707)). Bayes C model supported the results that each SNP has minimal effects (effect absolute value < 0.1) and the two traits are polygenic.

1.2 Model comparison of genomic prediction

In the previous section, moderate heritability was observed for both heterobothriosis resistance and SL (transformed HC: $h^2 = 0.308 \pm 0.123$ S.E.; SL: $h^2 = 0.405 \pm 0.131$). In addition, the polygenetic nature of these traits was confirmed and high potential for genetic improvement by GS was indicated. In GS breeding program, genetic gain for a target trait depends on the accuracy of GP. Thus, many advanced regression models have been proposed to achieve higher accuracy of GP. For instance, a linear mixed model, GBLUP uses the marker-based realized relationship matrix and the best linear unbiased prediction (BLUP) to estimate the GEBVs. The Bayesian models (e.g., Bayes A, B, C, Ridge, and LASSO) assume different priors to manipulate the variance of genetic values under Bayesian rules and use stochastic methods, namely Markov chain Monte Carlo (MCMC), to solve the linear mixed model (Gianola et al., 2013). While these approaches assume parametric additive models, non- (or semi-) parametric model, such as a reproducing kernel Hilbert space (RKHS) regression, can model multiple and complex interactions among loci, potentially arising over whole genome with nonparametric treatments by introducing smoothing parameters as variance components (Gianola et al 2006; Gianola and van Kaam, 2008). On the other hand, deep learning models were developed under the inspiration of information processing processes in the biological nervous system. These models are built from lots of non-linear sub-models (neurons) connected by the neuron circuits (mimicking how neurons are connected each other) to capture non-linear interactions between the phenotypes and genotypes by feature extraction and transformation (Pérez-Enciso and Zingaretti, 2019). The potential of each prediction model depends on the genetic architecture of the trait and should be tested per trait. In this section, model comparison of GP among 12 models was done for the two traits. Since prediction accuracy was not available for non-linear models, predictive ability was used as the evaluation metric for model comparison among all models

Materials and methods

Predictive abilities of GP

Predictive abilities were obtained under 12 regression models described below. The tenfold cross-validation scheme was applied for predictive ability and accuracy calculation following the procedure described in Hosoya et al. (2018). Samples were randomly and equally divided into ten subsets: one for testing and the remaining for training. The phenotypic values of the test set were masked, and the regression model was trained using the training set. GEBVs of the test set were predicted and predictive ability was calculated as the correlation (Pearson's *r*) between GEBVs and observed values of the test set. Then, prediction accuracy was calculated for the GBLUP and Bayesian models as predictive ability divided by the square root of heritability, which was calculated as described previously. This step was repeated with rotating the test sets among the five subsets, and the average of Pearson's *r* was obtained. This cross-validation process was repeated 10 times to obtain the mean and the standard error of the mean (S.E.) for the predictive ability and accuracy. Transformed HC instead of the original phenotype was used in GBLUP and Bayesian models. To generate the consistent random state for sampling among 12 models, I fixed seeds for random sampling among the models.

Genomic best linear unbiased prediction (GBLUP)

GBLUP was implemented as Chapter 1, Section 1.

Bayesian models

The models of Bayes A, B, C, Ridge, and LASSO (Habier et al., 2011; Meuwissen et al., 2001; Park and Casella, 2008) are expressed as follows:

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{X\beta} + \sum_{i=1}^{p} \boldsymbol{z}_i\boldsymbol{g}_i + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are same as GBLUP; μ is an intercept; $\mathbf{1}_n$ is a vector of one; p is the total number of genotypes for one individual; $\boldsymbol{z}_i$ is a vector of genotypes at SNP *i*; $\boldsymbol{g}_i$ is a vector of random effects that represent the genetic effects for SNP *i*. following a specific prior distribution. Bayes A assumes a scaled-*t* distribution for the prior while Bayes B assumes a mixture of gaussian distribution and a point mass at zero. The prior of Bayes C is a mixture of scaled-*t* distribution and a point of mass at zero. The prior of Bayes Ridge and Bayes LASSO is a normal distribution and a double exponential distribution, respectively. These models were solved using R/BGLR-1.0.8 (Pérez and De Los Campos, 2014) with nIter = 10,000 and burnIn = 2,000.

Bayesian reproducing kernel Hilbert spaces regressions (Bayesian RKHS)

Bayesian RKHS is a Bayesian approach of semi-parametric regression (De Los Campos et al., 2010) structured as:

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{u} + \boldsymbol{\varepsilon},$$

where each parameter is the same as the Bayesian models, while $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ follow the normal distribution as $\boldsymbol{u} \sim N(0,\ \boldsymbol{K}\sigma_u^2)$ and $\boldsymbol{\varepsilon} \sim N(0,\ \boldsymbol{I}\sigma_\varepsilon^2)$, respectively; where $\boldsymbol{K}$ is a Gaussian reproducing kernel (bandwidth parameter = 1) which approximates the marker-based relationship matrix and $\boldsymbol{I}$ is an identity matrix. The model was solved using R/BGLR-1.0.8 with nIter = 10,000 and burnIn = 2,000.

Support vector machine regression (SVR)

SVR method can be viewed as a convex optimization problem that finds a function from observed values to estimated values at most ε-deviation from all observed values while balancing the model complexity and prediction error (Awad et al., 2015; Vapnik, 1995). The method of Lagrange multipliers is used to solve the optimization problem, and the derived approximate function follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} (a_i^* - a_i) k(\boldsymbol{x}, \boldsymbol{x}_i) + b,$$

where the input $\boldsymbol{x}$ is a vector of all genotypes for a single sample; N is the sample size; a_i^* and a_i are Lagrange multipliers; $k(\boldsymbol{x}, \boldsymbol{x}_i)$ is a kernel function; $\boldsymbol{x}_i$ is a vector of genotypes for individual i; b is a residual. SVR-linear, SVR-poly, and SVR-rbf using linear, polynomial, and radial basis for kernel function, respectively. The SVR models were implemented by the *sklearn.svm.SVR* function in Python/scikit-learn-0.20.31 (Pedregosa et al., 2011). The *gamma* parameter was set to '*auto*' and the default setting was used for the other parameters.

Neural networks

Feedforward neural networks (FNN), inspired by the biological neural network, can model genotype-phenotype regression (Gianola et al., 2011). Neural cells were modeled by non-linear functions (or activation function) and the network was mimicked by the chain structure. My FNN had one input layer, two hidden layers, and a regression output. The number of input units was 6,707, equivalent to the number of SNPs. The first hidden layer has 200 hidden units and the second 20. The rectified linear unit (ReLU) was used as an activation function in hidden layers. FNN was trained by minimizing the loss function, that is, the mean squared error in this case:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where *n* is the sample size of the training group**;** y_i and $\hat{y}_i$ are observed value and the predicted value of individual i, respectively.

Multi-task deep neural network (Multi-task FNN) is based on an assumption where HC and SL share underlying genetic architecture to some extent. These models can improve the accuracy of estimation of the main output using the related auxiliary task as an inductive bias to the main task in reproducing kernel Hilbert space (Caruana, 1997; Widmer and Rätsch, 2011). The model has one input layer, two hidden layers, one main regression output, and one auxiliary regression output. The first hidden layer has 200 units, which is a sharing layer for both tasks. Both outputs have separated second hidden layers that differ in the number of the hidden units (20 for the main trait and 100 units for the auxiliary trait). For the estimation of HC, the main regression output is for HC and auxiliary regression output is for SL. The model setting of the main regression and auxiliary regression output was reversed for SL estimation. The activation function and the loss function were the same as the FNN model described above. FNN and multi-task FNN were implemented in Python/keras-2.4.3 package (Chollet and others, 2015) with tensorflow-gpu-2.2.1 backend (Abadi et al., 2016). FNN used "Adam" optimizer, and multi-task FNN used "RMSprop" optimizer, both with the default parameters. Both models were trained by *model.fit* method in Python/keras with the parameters as epochs = 30, batch_size = 128, and others followed the default. Many combinations of model architecture, loss function, activate the function, and optimizer was arbitrarily chosen and tested to find the models here that have a high accuracy of GP for HC.

Results

Predictive abilities for transformed HC ranged from 0.248 to 0.344 under 12 models (Table 1-1). Among them, SVR-poly and SVR-rbf models were inferior, while two deep learning models were slightly better. Predictive abilities for SL ranged from 0.340 to 0.481 under 12 models (Table 1-1). In contrast to the case of HC, the two SVR based models ranked at the top for prediction of SL, and deep learning models were inferior. Bayes RKHS and GBLUP models showed good performance in both traits. Moderate predictive ability of GP for both of the traits suggested the availability of GS since these results ensure the selection accuracy of broodfish.

1.3 Breeding strategy for simultaneous improvements of both heterobothriosis resistance and body size

In the previous sections, the availability of GS for both heterobothriosis resistance and body size was shown from the moderate heritability and predictive ability. However, the results also showed an antagonistic genetic correlation between these traits (r_g = 0.228); larger individuals are possibly suffering from higher parasitic loads. This undesired correlation complicates simultaneous genetic improvements of the two traits. One of the conventional methods for multiple-trait improvement is the linear selection index (LSI) method developed by Smith and Hazel (Hazel, 1943; Smith, 1936). Net genetic merit (i.e., LSI) of each animal is calculated from each target trait and used for ranking breeding candidates. To maximize the selection response, a general LSI is computed by a linear combination of phenotypes or EBVs (estimated breeding values) and the corresponding coefficients. Extensive LSI methods have been proposed (Cerón-Rojas and Crossa, 2018), as determined by the method of coefficient calculation. For instance, the desired gain selection index allows breeders to restrict traits according to the expected change of genetic gain values of traits (Itoh and Yamada, 1987). In the era of GS, those LSI methods can be directly applied to compute the linear genomic selection index (LGSI), which showed higher efficiency in both simulation and real data, compared to pedigree-based LSI (Ceron-Rojas et al., 2015). Although LGSI showed great advantages, successful applications of LGSI still largely depend on the accurate estimation of GEBVs and genetic parameters (Togashi et al., 2011), which are sensitive to many factors, including the genetic architecture of target traits, population structure, genotyping technologies, etc. (Daetwyler et al., 2010; Guo et al., 2014; Solberg et al., 2008). Consequently, an LGSI method might have different performances in different cases. Therefore, it is essential to find the optimal strategy incorporating LGSI and examine its performance in each breeding program. The GS breeding simulator will be a practical tool that approximates the real genetic progress by sophisticated modeling of the meiosis and GS procedure at the DNA level (Daetwyler et al., 2013). Further, as regards selection targeting disease resistance traits in aquaculture, a recent simulation study of acute hepatopancreatic necrosis disease (AHPND) in shrimp (*Litopenaeus vannamei*) showed that GS was superior to pedigree-based methods (Wang et al., 2019). Therefore, with the assistance of simulation, the breeding strategies incorporating LGSI are expected to greatly accelerate the simultaneous genetic improvement of disease resistance and growth-related traits.

Materials and methods

Simulation

To investigate the breeding strategy that can improve SL and HC simultaneously, six scenarios different in recurrent selection schemes were simulated for ten generations with 50 replicates using R/AlphaSimR-0.11.0 package (Faux et al., 2016). The tested scenarios were named for the selection scheme applied: random mating (RAND), GS on HC only (GS_{HC}), GS on SL only (GS_{SL}), Smith-Hazel index ($S1_{SHI}$ and $S2_{SHI}$), and Desired gains index (S_{DGI}). The workflow of the simulation study is illustrated in Figure 1-4. In short, RAND was based on random mating while GS_{HC} and GS_{SL} were based on GS on either of the traits. GEBV was estimated by GBLUP. In $S1_{SHI}$, selection candidates were ranked based on the Smith-Hazel index. Since economic importance for each trait has not been evaluated in the tiger pufferfish aquaculture industry, I assume both traits have equal economic weights, which is w = [-1, 1] for HC and SL for $S1_{SHI}$ (HC is expected to

decrease by selection). For S_{DGI}, *d* was set as [-3, 0.3] for HC and SL, so that SL can be improved preferentially while HC can be reduced by 30% after 10 generations (-3*10/100 = -30%). To compare the two selection index methods, I also run an additional scenario ($S2_{SHI}$) based on Smith-Hazel index, where the economic weight for each trait was set as same as the designed weights of S_{DGI} (*w* = [-3, 0.3]).

The simulated population was generated by *runMaCS2* function in R/AlphaSimR package (Gaynor et al., 2020). First, all scenarios began with a founder population of 10,000 individuals was simulated assuming: the effective population size of 1,000, mutation rate of 2.5×10^{-8}, no inbreeding in founder individuals. The ploidy (*n* = 2), the number of chromosomes (*n* = 22), and genetic and physical size of each chromosome (Morgans and base pairs, respectively) were set according to the reference genome sequence integrated with the genetic map of the tiger pufferfish, FUGU5/fr3 (Kai et al., 2011). The relative ratio of recombination in females compared to males was set as 1.82 according to FUGU5/fr3. The phenotypic mean of SL was set in accordance with the phenotyping result and that of HC was set as 100 to avoid minus values of phenotypes after genetic improvement. Phenotypic variance, genetic variance, heritability, and genetic correlation were simulated referring to the analysis result using empirical data obtained in this study. The gender of each individual was randomly assigned. For each trait, 500 QTLs were placed per chromosome. The number of SNP markers per chromosome was equal to that detected in the Ampliseq custom panel (*n* = 6,707 in total) and randomly placed over each chromosome to form an SNP chip *in silico*. To initiate the GS breeding program, 20 sires and 20 dams were randomly sampled from the founder population to perform a full-factorial mating, and then each parent pair generates 20 progenies and in total 8,000 progenies were produced. From the progeny pool, 2,000 fish were randomly picked up as the broodstock population (F_0). The relatively small number of parents were used in this simulation study compared with the practical breeding programs due to the limited computer resources, but it will be enough for a test study.

Subsequently, the recurrent selection schemes were performed for ten generations with 50 replicates independently among six scenarios (RAND, GS_{HC}, GS_{SL}, $S1_{SHI}$, $S2_{SHI}$, and S_{DGI}). In each generation, according to the scenario-specific criteria, 20 sires and 20 dams were selected and crossed with a full-factorial mating system to create next-generation where each mating cross generated 20 progenies (total 8,000 progenies). Only 2,000 fish out of 8,000 progenies remained as the broodstock candidates for the next generation. This process was performed for a total of ten generations with 50 replicates. The broodstock population produced in the i-th generation was noted as F_i (i = 1,2,3…10). The scenario-specific criteria were as following. In the RAND scenario, parental individuals were randomly selected from the candidates (n = 2,000) in each generation. The individuals with high GEBVs for only a single trait were chosen in the GS_{HC} and GS_{SL} scenario, while the ones with high LGSIs in $S1_{SHI}$, $S2_{SHI}$ and S_{DGI} Scenario. For each generation, in GS_{SL} scenario, GBLUP model was trained using all candidates (n = 2,000) and broodfish were directly selected from these individuals, whilst, in GS_{HC}, S_{SHI} and S_{DGI} scenario, GBLUP model was trained using half of the candidates (n = 1,000) and broodfish were selected from the remaining (n = 1,000) since fish should be sacrificed to obtain HC phenotype. The GBLUP model was implemented by *RRBLUP* function in R/AlphaSimR package.

Linear genomic selection index

LGSI for $S1_{SHI}$, $S2_{SHI}$ and S_{DGI} was constructed as:

$$\boldsymbol{LGSI} = \boldsymbol{b}'\widehat{\boldsymbol{y}},$$

where $\boldsymbol{b}$ is a vector of index coefficients; $\widehat{\boldsymbol{y}}$ is a vector of GEBVs. $\boldsymbol{b}$ for $S1_{SHI}$ and $S2_{SHI}$ was computed as:

$$\boldsymbol{b} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{w},$$

where $\boldsymbol{P}$ and $\boldsymbol{A}$ are phenotypic and genetic variance-covariance matrices, respectively; $\boldsymbol{w}$ is the economic importance of both traits and set as [-1, 1] assuming both traits have equal economic weights (HC is expected to be decreased by selection) for $S1_{SHI}$. $\boldsymbol{P}$ was obtained by *varP* function of R/AlphaSimR, and $\boldsymbol{A}$ was obtained by *mmer* function of R/sommer-4.0.1 following the same procedure for the estimation of genetic correlation except that original HC value was used. On the other hand, $\boldsymbol{b}$ for S_{DGI} was computed as:

$$\boldsymbol{b} = \boldsymbol{P}^{-1}\boldsymbol{A}(\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{A})^{-1}\boldsymbol{d},$$

where $\boldsymbol{P}$ and $\boldsymbol{A}$ are the same as S_{SHI} while $\boldsymbol{d}$ is a vector of desired gains. The combination of $\boldsymbol{d}$ can be chosen arbitrary depending on the breeding goal of the program. In this study, I set $\boldsymbol{d}$ at [-3, 0.3] for HC and SL so that SL can be improved preferentially while HC can be reduced by 30% after 10 generations. To compare between different selection index methods, the vector of economic importance, [-3, 0.3], which is the same with $\boldsymbol{d}$ in S_{DGI} was assigned to $\boldsymbol{w}$ in $S2_{SHI}$.

Results

The availability of LGSI methods for simultaneous improvements of HC and SL was tested using simulation data. The scenario with only random selection, RAND, showed that true breeding values (TBVs) at F_{10} generation fluctuated up and down around the TBVs at F_0 for both of the traits, and no obvious genetic changes were observed (Figure 1-5. **a**). GS_{HC} and $GS_{SL,}$ the scenarios which performed selection on single trait, TBVs of the targeted trait was improved while hindered in the other (Figure 1-5. **b** and **c**). The scenario ($S1_{SHI}$ and $S2_{SHI}$) applying Smith-Hazel index showed average TBV for both SL and HC decreased, i.e., smaller fish with less parasite loads were selected (Figure 1-5. **d** and **e**). As expected, only S_{DGI} could improve the two traits simultaneously, where true breeding values (TBVs) of parasite load (HC) decreased while SL increased in each generation (Figure 1-5. **f**).

Discussion

In this chapter, the possibility of GS for heterobothriosis resistance and SL of the tiger pufferfish was tested from empirical data. In addition, a GS breeding strategy that can improve the resistance trait concurrently with SL was designed using a simulation study.

With 6,707 SNP makers, moderate estimated heritability of transformed HC (h^2 = 0.308, SE = 0.123) and SL (h^2 = 0.405, SE = 0.131) were obtained, indicating selective breeding for those traits is feasible. The estimated heritability was comparable to those estimated for resistance against sea lice in Atlantic salmon (h^2 = 0.22 to 0.33 with 35k SNPs) (Tsai et al., 2016) and bacterial cold water disease resistance (survival days) in farmed rainbow trout (h^2 = 0.33 with 35k SNPs) (Vallejo et al., 2017). Although the small SNP panel can capture the moderate heritability for HC and SL in the tiger pufferfish, the relatively large standard error suggested that a middle- or large-scale study was still needed to confirm the generality of this results. In this study, significant SNPs were not detected in GWAS. It has been shown that even with the small SNP panel and small sample size, strong effect SNPs (the sex-determining SNP) can be detected in a cultured population of the tiger pufferfish (Sato et al., 2019). Therefore, GWAS result suggests the parasitic resistance is controlled by a large number of quantitative trait locus (QTL) with small or moderate effects, and marker-assisted selection is not feasible, although larger sample size and more SNPs may increase the credibility of the GWAS result as same as the heritability estimation.

The predictive abilities for HC estimated under 12 models were moderate (0.248–0.344), and within the range observed for other disease resistance traits examined in other fish species (Odegård et al., 2014; Palaiokostas et al., 2018, 2016; Robledo et al., 2018a). The predictive abilities of Bayesian hierarchical linear models (i.e. Bayes A, B, C, LASSO, and Ridge) were similar (0.303–0.312) and scarcely higher than the GBLUP model (0.307 ± 0.018) for HC. This suggests that these linear models did not greatly differ regarding the predictive ability and the assumptions of the prior distribution of genetic effects have a limited impact on this trait. Bayes RKHS showed slightly better performance in HC compared to these linear models. For SVR-poly and SVR-rbf models, relatively low abilities for HC were observed, however, high abilities were found for SL. Since the default hyperparameters were used in the SVR models, hyperparameter tuning may aid achievement of better performance for HC as in the case of the previous study (Azodi et al., 2019). The architectures of FNN and multi-task FNN were tuned to achieve high predictive ability of GS for HC, however, the same architecture was applied to calculate the predictive ability of GS for SL. As expected, these models resulted in high predictive ability for HC but low for SL. This indicates that a deep learning model is task-specific and high accuracy can be obtained with careful optimization as described previously (Pérez-Enciso and Zingaretti, 2019). However, a great improvement in predictive ability was not achieved by FNN methods compared to GBLUP and Bayesian models even with the model complexity.

The simulation study showed the availability of DGI for simultaneous genetic improvement in HC and SL even when the unfavorable antagonistic genetic correlation was assumed. The two scenarios incorporated the Smith-Hazel index showed the undesired consequences, where the average TBV for both SL and HC increased (smaller HC is favored). This happened because the breeding scheme only selected the individuals with the top LGSI values, but the high LGSI calculated by the Smith-Hazel index method does not guarantee the selected

individuals are superior in both of the traits (Kempthorne and Nordskog, 1959), especially when target traits show a negative correlation. On the other hand, DGI, a variation of the selection index methods, allows selection with restrictions on multiple traits via the desired gains vector (*d*). In this study, the *d* vector was set with aiming to reduce HC by 30% during 10 generations while maximizing SL. The desired gains vector (*d*) can be further optimized by comparing simulation scenarios with various *d* to achieve the self-defined breeding goal. Unfavorable genetic correlation between body size and disease resistance is commonly observed in aquaculture species, e.g. vibriosis in Atlantic cod (Bangera et al., 2011), bacterial cold water disease in rainbow trout (Evenhuis et al., 2015), and piscirickettsiosis in coho salmon (Bangera et al., 2011; Evenhuis et al., 2015; Yáñez et al., 2016). Therefore, it is expected that DGI or the similar LGSI method can be widely applied for the simultaneous improvement of disease resistance trait and growth-related traits, which are the primary targets of most breeding programs.

In summary, the availability of GS for HC and SL in the tiger pufferfish under a standard diet was confirmed in this small-scale study. Moderate heritability for both traits suggest the genetic return from GS is high. GBLUP and Bayesian linear regression models showed similar prediction performance for these traits. Although an unfavorable antagonistic genetic correlation was suggested between the two traits, the GS breeding strategy incorporating DGI can be a solution for the simultaneous genetic improvement.

Chapter 2. The contents of this chapter cannot be published online since they are anticipated to be published in a paper in a scholarly journal. The paper is scheduled to be published within five years

Chapter 3. The contents of this chapter cannot be published online since they are anticipated to be published in a paper in a scholarly journal. The paper is scheduled to be published within five years

Chapter 4. The contents of this chapter cannot be published online since they are anticipated to be published in a paper in a scholarly journal. The paper is scheduled to be published within five years

Chapter 5. The contents of this chapter cannot be published online since they are anticipated to be published in a paper in a scholarly journal. The paper is scheduled to be published within five years

Acknowledgements

I would first like to thank Professor Dr. Kiyoshi Kikuchi, Fisheries Laboratory, Graduate School of Agricultural and Life Sciences, University of Tokyo, for his helpful guidance and continued encouragement. I would like to express my sincere appreciation to Dr. Sho Hosoya for his excellent teaching and I am gratefully indebted to his very valuable comments on this dissertation.

I am thankful to Mr. Naoki Mizuno for his efforts on management of fish. I am thankful to Dr. Mana Sato for the consultation of AmpliSeq. I thank to Dr. Takuya Itou and Dr. Yuki Kobayashi for their supports for running Illumina MiSeq sequencing. I thank Dr. Sota Yoshikawa and Dr. Masaomi Hamasaki, Nagasaki Prefectural Institute of Fisheries for their efforts on management of experimental specimens and experiment conduction. I am thankful to Dr. Takashi Koyama for the implementation of RNA-sequencing. I also thank all other members in my laboratory for their kind help.

Finally, I would like to thank my family: my parents, Qing Miao and Yufeng Lin, for their spiritual and financial support; my wife, Beibei Wang, for her love, companionship, and caring; my daughter, Yumu Lin, who always cheers me up.

Figure and Table

Table 1-1

Predictive ability (mean ± standard error) on *Heterobothrium okamotoi* count (HC) and standard length (SL) under 12 models: GBLUP, Bayes A, Bayes B, Bayes C, Bayes LASSO, Bayes reproducing kernel Hilbert space (Bayes RKHS), support vector machine with a linear kernel (SVR-linear), SVR with a poly kernel (SVR-poly), SVR with a radial basis function kernel (SVR-rbf), feedforward neural networks (FNN), and multi-task feedforward neural networks (multi-task FNN). The top three models for HC and SL are highlighted with bold font

Model	**HC**	**SL**
GBLUP	0.307 ± 0.018	0.463 ± 0.018
Bayes A	0.312 ± 0.018	0.461 ± 0.018
Bayes B	0.306 ± 0.018	0.460 ± 0.018
Bayes C	0.307 ± 0.018	0.460 ± 0.018
Bayes LASSO	0.303 ± 0.018	**0.464 ± 0.018**
Bayes ridge	0.304 ± 0.018	0.460 ± 0.018
Bayes RKHS	**0.325 ± 0.019**	0.463 ± 0.018
SVR-linear	0.322 ± 0.017	0.410 ± 0.019
SVR-poly	0.248 ± 0.019	**0.481 ± 0.018**
SVR-rbf	0.249 ± 0.019	**0.475 ± 0.018**
FNN	**0.330 ± 0.018**	0.405 ± 0.017
Multi-task FNN	**0.344 ± 0.019**	0.340 ± 0.022

Figure 1-1

Histograms with the estimated density of phenotypes: (**a**) *Heterobothrium okamotoi* count (HC), (**b**) transformed HC, and (**c**) standard length (SL).

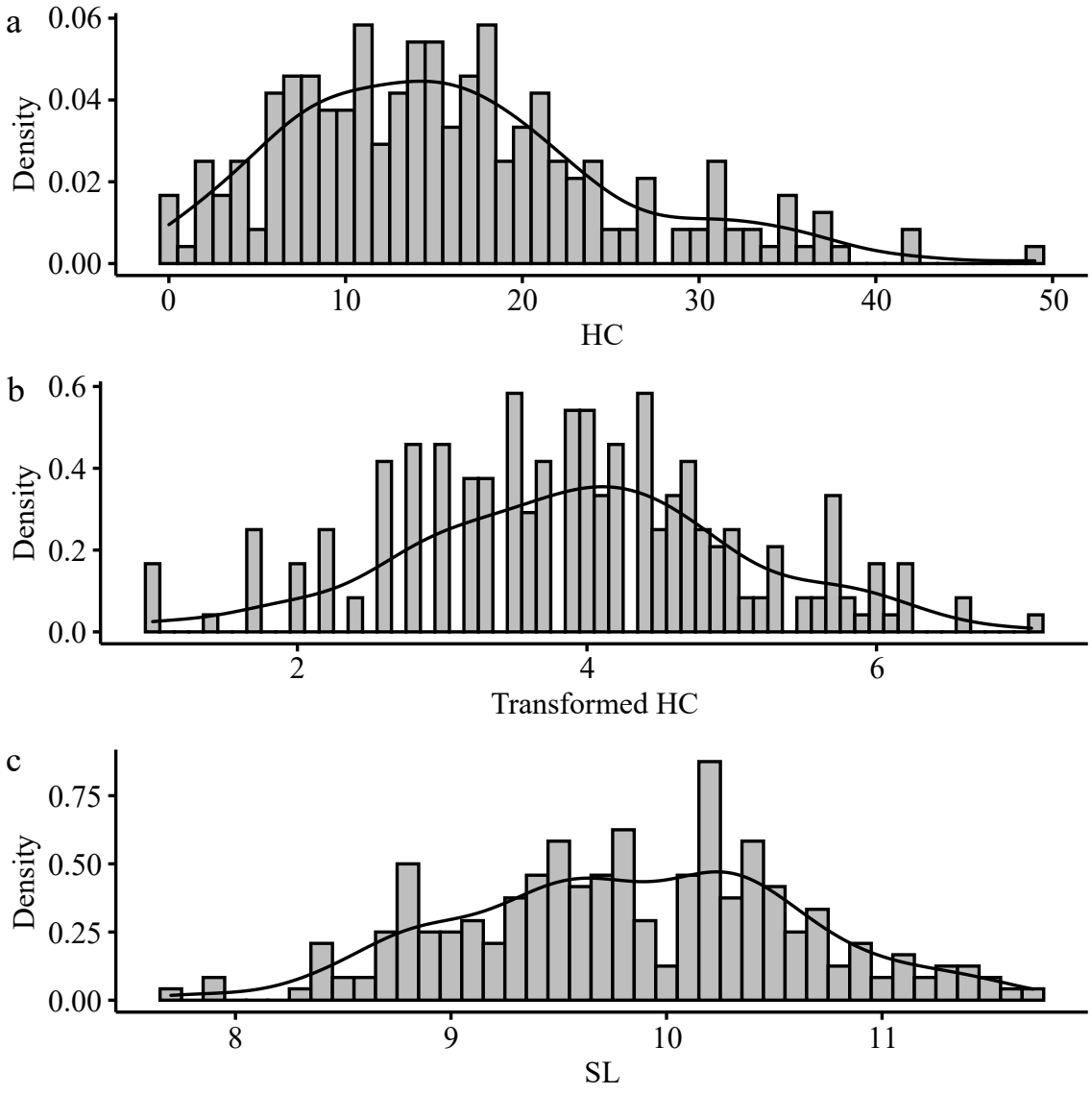

Figure 1-2

Population structure detected by *t*-SNE analysis based on the genomic SNP data of each individual (filled circle). Filled colors represent *Heterobothrium okamotoi* count of each individual based on the color bar (right panel).

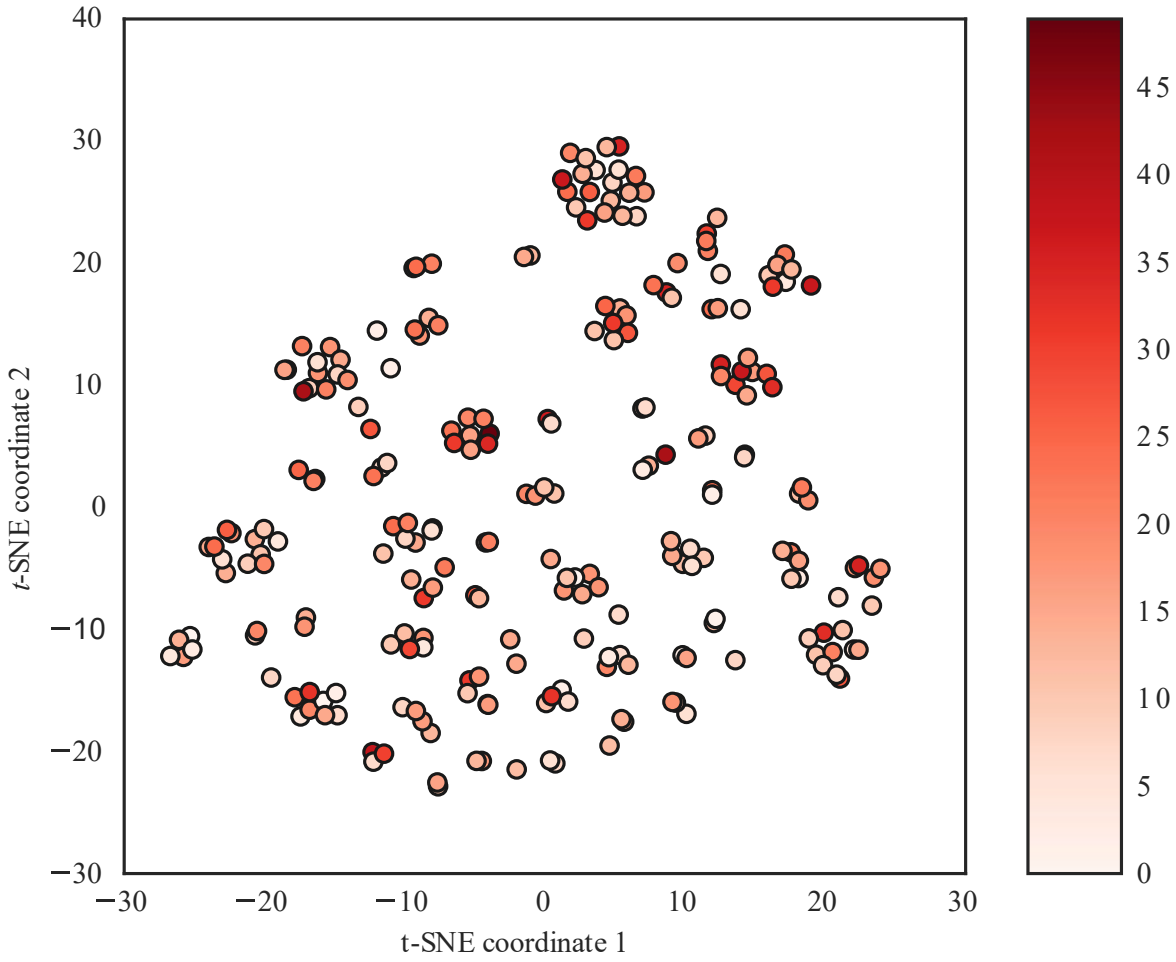

Figure 1-3

Manhattan plots from genome-wide association study: (**a**) transformed *Heterobothrium okamotoi* count (HC), and (**b**) standard length (SL). Adjacent chromosomes are distinguished by different colors. The X-axis is the physical order of the SNP markers across the 22 chromosomes of *Takifugu rubripes*. The Y-axis represents the negative logarithm of *p*-values (base: 10) for the target trait. Red dashed lines are Bonferroni-corrected significance thresholds of 5.128.

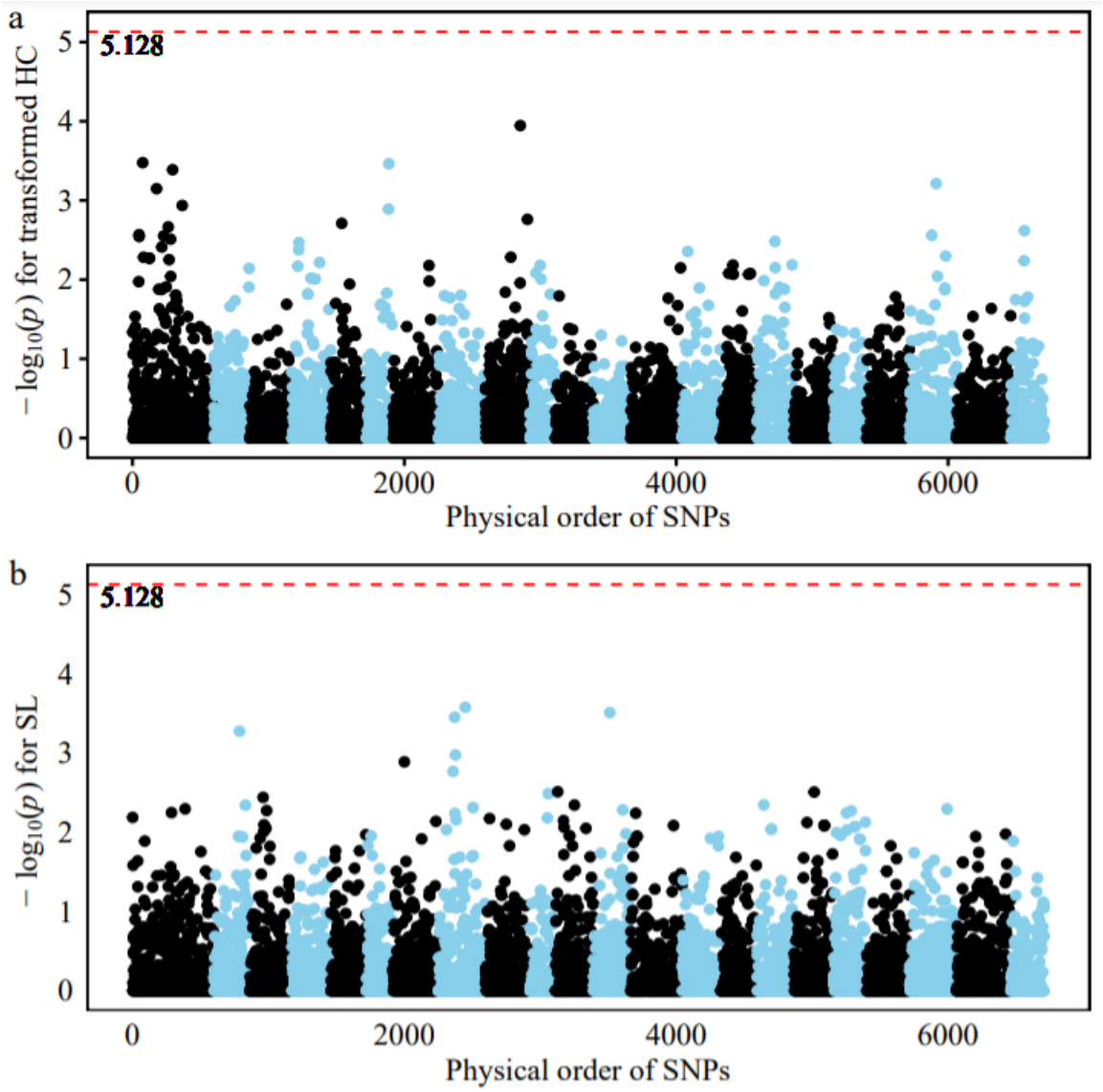

Figure 1-4

The diagrams of the simulation study. (**a**) The initiation of the breeding program shared among all scenarios. The founder population (n = 10,000) was constructed, and 20 sires and 20 dams were randomly sampled to produce 8000 progenies. Then, 2000 fish were randomly sampled from the progeny pool as the broodstock population (F_0). (**b**) The workflow of recurrent selection schemes. Parents (20 sires and 20 dams) were selected from F_0 according to the scenario-specific selection criteria and 8000 progenies were generated. The selection scenarios were: RAND, random selection; GS_{HC}, selection on *Heterobothrium okamotoi* counts (HC); GS_{SL}, selection on standard length (SL); $S1_{SHI}$ and $S2_{SHI}$, selection based on genomic Smith-Hazel index (SHI); S_{DGI}, selection based on the desired gains index (DGI). $S1_{SHI}$ has the same economic weights for both traits, and $S2_{SHI}$ uses the similar vector of economic weights as the vector of desired gains in S_{DGI}. Then, random sampling was applied to select 2000 progenies as the broodstock population for the next generation. A total of 10 generations (F_1 to F_{10}) of this process were replicated 50 times.

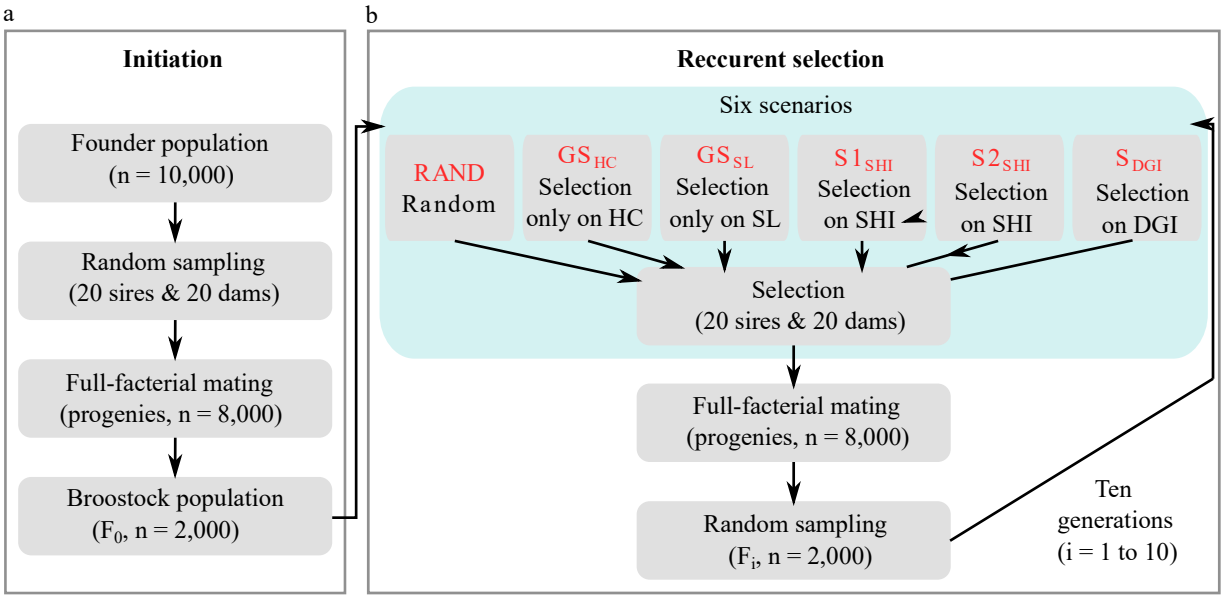

Figure 1-5

Genetic trends of average true breeding value (TBV) for *Heterobothrium okamotoi* count (HC, red lines) and standard length (SL, blue lines) of broodstock population in each generation (F_0 to F_{10}) among five different simulation scenarios with 50 replicates. (**a**) random mating (RAND), (**b**) GS on HC only (GS_{HC}), (**c**) GS on SL only (GS_{SL}), (**d**) Smith-Hazel index with the same economic weights $S1_{SHI}$), (**e**) Smith-Hazel index with the different economic weights ($S2_{SHI}$), and (**f**) desired gains index (S_{DGI}).