frontiers in Genetics

ORIGINAL RESEARCH
published: 06 August 2019
doi: 10.3389/fgene.2019.00675

Scanning of Genetic Variants and Genetic Mapping of Phenotypic Traits in Gilthead Sea Bream Through ddRAD Sequencing

*Dimitrios Kyriakis[1,2,3], Alexandros Kanterakis[2], Tereza Manousaki[3], Alexandros Tsakogiannis[3], Michalis Tsagris[4], Ioannis Tsamardinos[5], Leonidas Papaharisis[6], Dimitris Chatziplis[7], George Potamias[2] and Costas S. Tsigenopoulos[3]**

[1] School of Medicine, University of Crete, Heraklion, Greece, [2] Foundation for Research and Technology–Hellas (FORTH), Heraklion, Greece, [3] Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Center for Marine Research (HCMR) Crete, Greece, [4] Deparment of Economics, University of Crete, Gallos Campus, Rethymnon, Greece, [5] Department of Computer Science, University of Crete, Voutes Campus, Heraklion, Greece, [6] Nireus Aquaculture SA, Koropi, Greece, [7] Department of Agriculture Technology, Alexander Technological Education Institute of Thessaloniki, Thessaloniki, Greece

OPEN ACCESS

Edited by:
Lior David,
Hebrew University of Jerusalem, Israel

Reviewed by:
Gen Hua Yue,
Temasek Life Sciences Laboratory, Singapore
Diego Robledo,
University of Edinburgh, United Kingdom
Gonzalo Martinez-Rodriguez,
Institute of Marine Sciences of Andalusia (ICMAN), Spain

***Correspondence:**
Costas S. Tsigenopoulos
tsigeno@hcmr.gr

Specialty section:
This article was submitted to Livestock Genomics, a section of the journal Frontiers in Genetics

Received: 19 October 2018
Accepted: 27 June 2019
Published: 06 August 2019

Citation:
Kyriakis D, Kanterakis A, Manousaki T, Tsakogiannis A, Tsagris M, Tsamardinos I, Papaharisis L, Chatziplis D, Potamias G and Tsigenopoulos CS (2019) Scanning of Genetic Variants and Genetic Mapping of Phenotypic Traits in Gilthead Sea Bream Through ddRAD Sequencing. Front. Genet. 10:675. doi: 10.3389/fgene.2019.00675

Gilthead sea bream (Sparus aurata) is a teleost of considerable economic importance in Southern European aquaculture. The aquaculture industry shows a growing interest in the application of genetic methods that can locate phenotype–genotype associations with high economic impact. Through selective breeding, the aquaculture industry can exploit this information to maximize the financial yield. Here, we present a Genome Wide Association Study (GWAS) of 112 samples belonging to seven different sea bream families collected from a Greek commercial aquaculture company. Through double digest Random Amplified DNA (ddRAD) Sequencing, we generated a per-sample genetic profile consisting of 2,258 high-quality Single Nucleotide Polymorphisms (SNPs). These profiles were tested for association with four phenotypes of major financial importance: Fat, Weight, Tag Weight, and the Length to Width ratio. We applied two methods of association analysis. The first is the typical single-SNP to phenotype test, and the second is a feature selection (FS) method through two novel algorithms that are employed for the first time in aquaculture genomics and produce groups with multiple SNPs associated to a phenotype. In total, we identified 9 single SNPs and 6 groups of SNPs associated with weight-related phenotypes (Weight and Tag Weight), 2 groups associated with Fat, and 16 groups associated with the Length to Width ratio. Six identified loci (Chr4:23265532, Chr6:12617755, Chr:8:11613979, Chr13:1098152, Chr15:3260819, and Chr22:14483563) were present in genes associated with growth in other teleosts or even mammals, such as semaphorin-3A and neurotrophin-3. These loci are strong candidates for future studies that will help us unveil the genetic mechanisms underlying growth and improve the sea bream aquaculture productivity by providing genomic anchors for selection programs.

Keywords: aquaculture, *Sparus aurata*, double digest random amplified DNA, Genome Wide Association Study, feature selection

Abbreviations: GWAS, Genome Wide Association Study; ddRAD, double digest Random Amplified DNA; SNPs, Single Nucleotide Polymorphisms; FS, Feature Selection; MAS, Marker Assisted Selection; GS, Genomic Selection; QTL, Quantitative trait locus; BIC, Bayesian Information Criterion; LD, Linkage Disequilibrium; QQ plot, Quantile–Quantile plot; SES, Statistically Equivalent Signature; OMP, Orthogonal Matching Pursuit; MB, Markov Blanket; CV, Cross-Validation; PCA, Principal Component Analysis.

INTRODUCTION

The gilthead sea bream, Sparus aurata (Linnaeus, 1758), is a teleost fish of great economic importance for the Mediterranean aquaculture industry (Tsigenopoulos et al., 2014). It ranks first among other aquacultured species in South Mediterranean with total production of 160,563 tons for 2016 (FEAP, 2017). One of the top interests of the aquaculture industry is the genetic improvement of the stocks to maximize the efficiency of the production and the product quality (Fernandes et al., 2017). Coupled with this concern, various areas of sea bream biology are being explored, such as nutrition requirements (Silva-Merrero et al., 2017; Guardiola et al., 2018), immune responses (Antonopoulou et al., 2017; Bahi et al., 2018; Tapia-Paniagua et al., 2018), skeletal development (Negrín Báez et al., 2015; Vélez et al., 2018), reproduction, and broodstock management (Loukovitis et al., 2011). Recently, the genome of sea bream has been sequenced and analysed offering a backbone for conducting genomic analyses on the species (Pauletto et al., 2018).

One of the main avenues to genetically improve the cultured stock is to identify associations between genetic variants and traits of interest, such as growth, disease resistance, and fat content. Genome Wide Association Studies (GWAS) offer the way to accomplish this by comparing the genotypes of individuals having varying phenotypes for a specific trait of interest. GWAS have boosted the field of human genetics as well as plant and livestock breeding (Geng et al., 2017), leading to improved higher selection accuracies of the animal breeding programmes, which in turn leads to lower costs and greater yield (Geng et al., 2017). To conduct a GWAS experiment in non-model species, genome-wide sampling of genetic variants is required. Application of double digest Random Amplified DNA (ddRAD) leads to thousands of polymorphic loci that require sophisticated strategies for data analysis (Catchen, 2013) and is widely used for GWAS studies (Baird et al., 2008; Etter et al., 2011; Anderson et al., 2012; Palaiokostas et al., 2013). It is well known that biological datasets are susceptible to the curse of dimensionality (Lie, 2014; Stephens et al., 2015). Various methods have been developed to solve such complicated problems, such as feature selection (Tsagris et al., 2018a). Feature selection (FS) is used to identify the important, predictive genetic variants by removing the noise propagated by redundant features, i.e., markers that have the same genotypic profile across all samples. Several FS algorithms have been developed like (Fontanarosa and Dai, 2011), Orthogonal Matching Pursuit (OMP) (Cai and Wang, 2011), and Statistically Equivalent Signature (SES) (Lagani et al., 2017), differing mainly in the approach to discover associations and the computational efficiency.

In aquaculture breeding programs, these features-markers can be used for marker assisted selection (MAS) (Yue, 2014). However, genome-wide variants can also be used to directly evaluate breeders, the so-called genomic selection (GS) method (Yue, 2014). Genomic selection is a breeding value estimation methodology that aims to increase the rate of genetic gain, leading to improvement of certain phenotypes *via* genetic marker utilization (Heffner et al., 2011; Lorenz et al., 2011; Yue, 2014; Khatkar, 2017). Genetic markers associated with production traits are used to predict breeding values with high accuracy (Goddard and Hayes, 2007; Sonesson and Meuwissen, 2009; Wang et al., 2017, Gutierrez et al., 2015). Although high availability of genetic markers (i.e., SNP markers) could be used for the improvement of the accuracy of breeding value estimation through the use of a Genomic Relationship matrix (i.e., GBLUP), some genetic markers that are also associated with production traits could further increase the accuracy of breeding value estimation and, moreover, allow for the inclusion of alternative models of inheritance, rather than only additive, in the genetic evaluation procedures. Genomic selection based on specific traits such as fat, weight, and disease resistance can have great effects on the productivity and profitability of several aquaculture species (Yue, 2014).

In this study, we sought to identify genetic markers associated with important phenotypes in sea bream. We used ddRAD sequencing to identify and genotype genome-wide single nucleotide polymorphisms (SNPs) in multiple sea bream families. We performed both GWAS and FS to test the association among a combination of loci and the phenotypes of fat, weight, tag weight, and length/width. Finally, genomic prediction of the phenotypes was tested using the selected polymorphisms to evaluate its potential in selection for improved phenotypic traits like weight in sea bream. Our ultimate goal was to construct a signature—a combination of genetic markers—that will lead to maximizing the sea bream aquaculture efficiency, by improving the selected phenotypic traits.

MATERIALS AND METHODS

Sample Collection

The fish used in this study were a subset of a larger experiment with progeny from 66 male and 35 female brooders constituting 73 different full sib families from the breeding program of a commercial aquaculture company (Nireus Aquaculture S.A.). From those 73 full sib families, 14 families originating from 13 males and 11 females were selected (selective genotyping), based on their within-family variation of bodyweight at harvest, for genotyping with microsatellite markers in order to perform a QTL confirmation experiment (Chatziplis et al. 2018, in preparation). Seven male and six female brooders with 105 progeny in total, constituting six full sib families and one maternal half sib family (10 progeny on average per family), were used for ddRAD library preparation and sequencing. These seven families were those exhibiting the greatest family variation of bodyweight at harvest out of 14 total families included in the QTL verification experiment (Chatziplis et al. 2018, in preparation). All progeny were reared in commercial conditions, and after PIT tagging, they were transferred to sea cages at 220 Days Post Hatching (DPH) for the growth period. For all progeny, the weight at tagging (g) (205 DPH), weight at harvest (g) (750 DPH), percentage (%) of fat at harvest (as measured in terms of body electrical conductivity, 692 Distell) as described by Besson et al. (2019), the total length at harvest (cm) (750 DPH), and the width at harvest (cm) (750 DPH) were measured.

Library Preparation and Sequencing

Individual DNA library preparation and sequencing of the samples, which were extracted using a modified salt-based extraction protocol based on Miller et al. (1988) and treated with RNase to remove residual RNA, were performed. Genomic DNA was eluted in 5 mmol/L Tris, pH 8.5, and stored in 4°C. Each sample was quantified by spectrophotometry (Nanodrop 1000–Thermo Fisher Scientific) and quality assessed by 0.7% agarose gel electrophoresis. To build the ddRAD library, we used the protocol described by Manousaki et al. (2016), with some minor modifications. Briefly, each of 144 DNA samples (13 parents in triplicates and 105 offspring; 21 ng DNA per sample) was separately but simultaneously digested by two high-fidelity restriction enzymes (RE): SbfI (CCTGCA|GG recognition site) and SphI (GCATG|C recognition site), both sourced from New England Biolabs (NEB), UK. Digestions were incubated at 37°C for 90 min, using 10 U of each enzyme per microgram DNA in 1 CutSmart Buffer (NEB), in a 6 μl total reaction volume. The reactions were slowly cooled to room temperature, and 3 μl of a premade adapter mix was added to the digested DNA and incubated at room temperature for 10 min. This adapter mix contained individual-specific combinations of P1 (SbfI-compatible) and P2 (SphI-compatible) adapters at 6 and 72 nM concentrations, respectively, in 1· reaction buffer 2 (NEB). The ratio of P1 to P2 adapter (1:12) was selected to reflect the relative abundance of SbfI and SphI cut sites present. P1 and P2 adapter included an inline five- or seven-base barcode for sample identification. Ligations were implemented over 3 h at 22°C by addition of a further 3 μl of a ligation mix comprising 4 mM rATP (Promega, UK) and 2000 cohesive-end units of T4 ligase (NEB) in 1· CutSmart buffer (NEB). The ligated samples were pooled together, and the single pool was column-purified (MinElute PCR PurificationKit, Qiagen, UK) and eluted in 70 μl EB buffer (Qiagen, UK). The size selection was performed by agarose gel separation, keeping the fragments between 400 and 700 bp. Following gel purification (MinElute Gel Extraction Kit, Qiagen, UK), the eluted size-selected template DNA (68 μl in EB buffer) was PCR amplified (15 cycles PCR; 32 separate 12.5 μl reactions, each with 1 μl template DNA) using a high-fidelity Taq polymerase (Q5 Hot Start High-Fidelity DNA Polymerase, NEB). The PCR reactions were combined (400 μl total) and column-purified (MinElute PCR Purification Kit). The 57 μl eluate, in EB buffer, was then subjected to a further size-selection clean-up using an equal volume of AMPure magnetic beads (Perkin-Elmer, UK) to maximize removal of small fragments. The final libraries were eluted in 24 μl EB buffer. Lastly, the ddRAD libraries were sequenced in one HiSeq 2500 lane (2x125 bp reads).

Raw Read Quality Control and Demultiplexing

We used FastQC v.0.11.5 software to check the quality control of the raw sequence data (Andrews and Babraham Bioinformatics Group, 2010). To recover the reads belonging to each individual, we then cleaned and demultiplexed the raw data using Process radtags program from STACKS v.1.46 software (Catchen, 2013). In this step, -c parameter was used to remove reads with an uncalled base, -q parameter was used to discard sequencing reads of low quality (below 20) using the Phred scores provided from the FASTQ files (Catchen, 2013), and -t parameter was set to 100 to truncate final reads length to 100 bp.

Data Alignment Against Sea Bream Reference Genome

The annotated reference genome of gilthead sea bream has been provided by Hellenic Centre for Marine Research (H.C.M.R.) (Accession Numbers: SRR6244977-SRR6244982) (Pauletto et al., 2018). To align our samples to the reference genome, we used Bowtie2 v.2.3.0 (Langmead and Salzberg, 2012) with the following parameters: {end-to-end {sensitive {no-unal. Then, we removed multi-aligned reads, reads with >3 mismatches, and reads with mapping quality lower than 20 with Samtools (Li et al., 2009).

Genotyping RAD Alleles

Genotypes of each sample were constructed using STACKS pipeline (Catchen, 2013). For each individual, pstacks program was used to build the rad loci based on the alignment on the reference genome, setting the minimum depth of coverage to create a stack (-m) equal to 3 (default) (Paris et al., 2017). Then, a catalogue of loci was constructed using only the parental reads on cstacks program, using default parameters. To match the data of each offspring separately against the respective catalogue, we used sstacks program with ––aligned parameter. Finally, to retrieve the vcf file with the genotypes, we used populations program.

Kinship

To check family relationship and indicate possible pedigree errors, we used KING v.2.1 software (Manichaikul et al., 2010). Kinship coefficients have been estimated by KING, setting the ––degree parameter equal to 10. Kinship coefficient is a measurement of kinship between two individuals; 1 means homozygous twins, 0 means unrelated (Manichaikul et al., 2010). Finally, to see the genetic distances of studied individuals, we performed a Principal Components analysis (PCA) and Hierarchical clustering, using Euclidean distance. Both PCA and Hierarchical clustering were implemented in R using prcomp and hclust functions, respectively.

Linear Mixed Models

To fit the mixed model for every phenotype, we used the command lmer from the lme4 R package (Bates et al., 2014). Random effects were fitted for each family to control for the correlation within the families. In mathematical notation, the linear mixed model is written as

$$y_i = a + \tau_i + \sum_{i=1}^{p} \beta_j X_j + e_i \quad (1)$$

where i = 1,…,K, with K denoting the number of families and y_i is the vector of measurements of the i-th family containing n_i

measurements with $\sum_{(i=1)}^{K} n_i = n$, the overall sample size. The term τ_i is the overall constant term. The τ_i is the random effect of the i-th family, the deviation of the i-th family from the overall constant a. The term β_j is the fixed regression coefficient of the variable X_j, and e_i is the vector of residuals of the i-th family. The model has two sources of variation: one stemming from the residuals and one stemming from the repeated measurements, $e_{ij}\ N\left(0,\sigma_e^2\right)$ and $\tau_i\ N\left(0,\sigma_\tau^2\right)$, respectively. Residuals represent elements of variation unexplained by the fitted model. Since this is a form of error, the same general assumptions apply to the group of residuals that we typically use for errors in general: one expects them to be normal and approximately independently distributed with a mean of 0 with some constant variance (Bates et al., 2014). To compare two linear mixed models, we used the Bayesian information criterion (BIC). BIC is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based on the log-likelihood function and takes into account the number of estimated parameters. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over-fitting (Vrieze, 2012). BIC attempts to resolve this problem by introducing a penalty term for the number of parameters in the model.

Genome Wide Association Study

A typical GWAS analysis tests for variant significance in a set of independent samples. The most common source of sample dependence is family relationships. Yet, our study is based on a family designed cohort. For this reason, we applied a family-based test for variant significance. To perform this, we used lmer in order to create a linear mixed model for each phenotype. This model includes family id as a random effect. To correct for multiple testing, we set the significance threshold to 10^{-4}, which is the typical significance level a = 0.05 divided to the number of independent SNPs (497) based on linkage disequilibrium (LD) (Johnson et al., 2010; Clarke et al., 2011). We used the plink tool v.1.90 in order to calculate the LD score (––indep-pairwise 50 5 0.05) (Purcell et al., 2007). Finally, we presented the distribution of the p-values across the genome in Manhattan plots, and we tested for possible p-value inflation through Quantile–quantile (QQ) plots. For these plots, we used the GWASTools (Gogarten et al., 2012) library in R (scripts available upon request).

Feature Selection

The typical GWAS pipeline reveals individual SNPs that are associated with a specific phenotype. One limitation of this pipeline is that it cannot produce signatures that contain combinations of variants. This problem is commonly referred as SNP to SNP interaction induction (Balliu and Zaitlen, 2016). The large number of tested genotypes in a typical GWAS experiment makes prohibitive the efficient computation of variant combinations. Also, the burden of multiple testing increases linearly to the number of combined variants. This means that a SNP–SNP interaction should be of extreme significance in order to be detected by a method that tests all possible combinations of variants. To tackle this problem, we employed a different approach. We considered SNPs as variables that describe a certain phenotype. We then applied methods that seek the optimum subset of variables with which we can construct a predictive model for a trait of interest (e.g., Weight). This approach is called Variable selection, or Feature Selection (FS). Solving the FS problem has numerous advantages (Tsamardinos and Aliferis, 2003). Features in biology (e.g., SNPs and gene expressions) are commonly found to be expensive to measure, store, and process (Stephens et al., 2015). By reducing the number of measurable markers-loci via FS, one can reduce this cost. A high-quality FS algorithm improves the predictive performance of the resulting model by removing the noise propagated by redundant features. For our study, we used two different FS algorithms: The first is the statistically equivalent signature (SES) algorithm, and the second is the Orthogonal Matching Pursuit (OMP) algorithm.

The Statistically Equivalent Signature Algorithm

Commonly FS algorithms aim to find a single group of features that has the highest predictive power. On the contrary, SES algorithm introduced by Lagani et al. (2017) attempts to identify multiple signatures (feature subsets) whose performances are statistically equivalent. SES produces several signatures of the same size and predictive power regardless of the limited sample size or high collinearity of the data (Statnikov and Aliferis, 2010). It performs multiple hypothesis tests for each feature, conditioning on subsets of the selected features. For each feature, the maximum p-value of these tests is retained and the feature with the minimum p-value is selected. This heuristic has been proved to control the False Discovery Rate (Tsamardinos and Brown, 2008). SES is specially engineered for small sample sizes and eliminates the need for Bonferroni correction and/or FDR filtering (Lagani et al., 2017). Here, we used an adaptation of the SES algorithm that accommodates repeated measurements (Tsagris et al., 2018a). SES algorithm is influenced by the principles of constraint-based learning of Bayesian networks (Lagani et al., 2017). Bayesian networks are directed acyclic graphs that represent the dependency relationships between variables in a dataset. An edge A → B in a Bayesian graph represents the conditional dependence of variable B from variable A. There is a theoretical connection between S and the Bayesian (causal) network that describes best the data at hand (Tsamardinos and Aliferis, 2003). Following the Bayesian networks terminology, the Markov Blanket (MB) of a variable or node A in a Bayesian network is the set of nodes ∂A composed of A's parents (direct causes), its children (direct effects), and its children's other parents (other direct causes of the A's direct effects). Every set of nodes in the network is conditionally independent of A when conditioned on the Markov blanket of the node A (∂A as described in formula 2). Thus, the Markov blanket of a node contains the only knowledge needed to predict the behavior of that node.

$$\Pr(A \mid \partial A, B) = \Pr(A \mid \partial A) \tag{2}$$

Orthogonal Matching Pursuit Algorithm

Orthogonal Matching Pursuit is an iterative algorithm. At each iteration, it selects the column-marker of the SNP data matrix,

which has the greatest correlation with the current residuals (Cai and Wang, 2011). OMP updates the residuals by projecting the observation onto the linear subspace spanned by the columns that have already been selected and then proceeds to the next iteration. No column is selected twice because the residuals are orthogonal to all the selected columns. The algorithm stops when a criterion is satisfied. We have used its generalized form, gOMP, whose stopping criterion is based upon the difference of the BIC score between two successive models. If the difference is lower than a predefined threshold, the algorithm stops. The major advantage of OMP compared with other alternative methods is its simplicity and fast implementation (Cai and Wang, 2011).

Model Selection Through Cross Validation

The selection of the appropriate algorithm for each dataset is a challenging task. Commonly, a k-fold cross-validation (CV) is used in order to end up with the algorithm with the best fit in the examined dataset. Cross-validation is a model validation technique for assessing the results of a model. It is commonly used for estimating how precisely a predictive model performs in unknown data samples. The standard method of a prediction problem, where a dataset of known data is given, is to split data samples in folds and every time use the n-1 folds as training dataset and the one fold that is left, as test dataset ("unknown data"). The goal of cross validation is to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model. This approach limits problems like over-fitting and gives an insight on how the model will generalize to an independent dataset (Tibshirani and Tibshirani, 2009). To compare the algorithms and select the best model (including algorithm and parameters), we performed cross validation by using all but one sample as training set and the remaining sample as test set iterating over all samples, the so-called Leave-One-Out cross validation method.

The different models were assessed based on the sum or errors when assuming that the "unknown data" belong to each family (Equation 3). The model with the lowest mean sum of errors is selected as best model (Equation 4).

$$ErrOB = \sum_{i=1}^{m} E(y_{i(n_i+1)} - x^T_{i(n_i+1)}\hat{\beta} - z^T_{i(n_i+1)}\hat{b}_i)^2 / m, \quad (3)$$

where $y_i(n_i + 1)$, $x_i(n_i + 1)$ and z_i $(n_i + 1)$ are, respectively, the outcome and predictors of the new observation in cluster i, and $\hat{\beta}$ and $\hat{b}_i$ are, respectively, the estimates of β and b_i based on all the training data. This can be estimated by the leave-one-out cross validation,

$$LOOCV = \sum_{i=1}^{m}\sum_{j=1}^{n_i} (y_{ij} - x^T_{ij}\hat{\beta}^{[i,j]} - z^T_{ij}\hat{b}^{[i,j]}_i)^2 / N, \quad (4)$$

where $\hat{\beta}^{[i,j]}$ and $\hat{b}^{[i,j]}_i$ are, respectively, the estimates of β and b_i based on the training data without subject j in cluster i (Fang 2011).

Selected SNP Annotation

To identify potential genes that might be affected by the retrieved SNPs, we searched the reference genome and classified the SNPs to those falling within a genic region (located within or in a window of 10Kb upstream or downstream of an annotated gene) and those that do not. If these regions were described as conserved at the genome browser of Gilthead sea bream (http://biocluster.her.hcmr.gr/myGenomeBrowser?search=1&portalname=Saurata_v1) in any of the following species: Stickleback, Asian sea bass, Medaka, Asian swamp eel and Amazon molly, they were considered as conserved.

RESULTS

Genotyping RAD Alleles

Illumina sequencing yielded 559,191,588 raw reads. Following quality control, we filtered out ~ 15.2% due to ambiguous barcodes, ~ 2.9% due to low quality, and 1% due to the lack of restriction sites. The rest were successfully assigned to individuals (**Supplementary Table 1** with number of reads per individual). After the demultiplexing, the high-quality reads of each sample were aligned against the reference genome. In total, 93% of the reads were mapped. Downstream filtering resulted in further discarding of multi-aligned reads (~ 8%) and those with more than three mismatches (~ 2.96%), keeping finally 351,781,485 reads for analysis. This resulted in an average coverage of 188.25. Although we did not experiment with greater values of m and used the default value proposed by STACKS, the sequencing effort was enough to have 188.25 coverage on average (s.e +/− 9.68) for the loci in our study. However it has been suggested that moderate values of m (3–6) (Paris et al., 2017) might not have any effect on the mean coverage of the reconstructed loci on a teleost species. The ddRAD catalogue built from all parental samples consisted of 15,233 SNPs. The used ddRAD protocol has been applied in other sparids as well (Manousaki et al., 2016; Manousaki et al. unpublished data). In all cases, the number of produced SNPs was in the range of 5,000–10,000 per individual (Manousaki et al., 2016; Palaiokostas et al., 2018). In this study and in accordance to this protocol, the SNP catalogue was built using solely parental data. Thus, the discovered SNPs are within the expected range given the following ddRAD protocol. Variants with allele frequency lower than 0.05 (n = 2,065) were filtered out. From the remaining 13,168, we filtered out the SNPs with call rate lower than 90% (n = 7,882). From the remaining 5,286 SNPs, 3,028 had at least one missing value and 2,258 had no missing values.

Kinship Assignment

To verify the family identity of the studied individual, we used three different methods: King kinship, Principal Component Analysis (PCA), and Hierarchical clustering (**Supplementary Figure 1**). All three resulted in similar results, and they confirmed the tagging family id, except for two samples, one placed in different family (sample 133 that was identified as a member of Family 2 instead of Family 3) and one that was

not placed in any family (sample 882). These two samples were discarded and not included in downstream analyses.

Association Analysis Through GWAS

The results from the GWAS test among all SNPs and the four phenotypes are shown in **Table 1**. In total, we found five SNPs associated with Weight, four SNPs with Tag Weight, and none for Fat and Length/Width. In **Figure 1**, we show the phenotype distribution, Manhattan plot, and QQ-plot for each phenotype. For illustration purposes, the Manhattan plot depicted was built with variants of known ordered positions on the reference genome. The Manhattan plot for the variants in scaffolds that we do not know the exact position in the genome is given in the **Supplementary Figure 2**. The QQ-plot of Weight revealed a systemic inflation of the observed p-values possibly attributed to the fact that families were selected in such a way as to maximize the weight variation within the cohort. Regarding the loci associated with weight and tag weight, we identified nine SNPs in total (**Table 1**). Five SNPs associated with weight at harvest have been retrieved from the typical GWAS analysis. The first was found in chromosome 1 (chr1:16636968) on "ethanolamine phosphate cytidylyltransferase-like" gene and the second (chr6:12617755) in a conserved region upstream of "myosin-7-like" gene. The third (chr16:2232897) was located on two overlapping genes acetylserotonin O-methyltransferase-like and LBH-like isoform X1. Another two SNPs were found in chromosome 1. The first (chr1.6970078) located downstream of "lymphoid enhancer-binding factor 1" and the second (chr1:20827142) located upstream of "mucin-5AC-like isoform X1" (**Table 1**). Finally, four SNPs (in chromosomes 2, 13, and 22) were associated with weight at tagging. Two were found at "RNA-binding 27 isoform X1" gene (chr13:20975921,chr13:20975924), the third upstream from "Tetratricopeptide repeat 36" gene (Chr2:2623351), and the fourth upstream from "tectonin beta-propeller repeat-containing 2" gene (chr22:18343985).

Association Analysis Through FS

Feature selection methods generate groups of SNPs that are associated with a phenotype en masse. Therefore, FS is a valuable family of methods for association analysis. We performed FS with 10 models (8 variants of SES and 2 variants of OMP), and from each model, we extracted the median squared error as an evaluation metric (**Figure 2**). All OMP models were inferior to SES. The best models for Fat and Weight have been constructed by SES algorithm (significance threshold equal to 0.01; number of condition set equal to three). The best model for Tag weight and Length/Width ratio prediction was the model constructed by variables retrieved from SES with size of condition set equal to two. The selected features of the best model, for each phenotype, are presented in **Tables 2–5**. SES produced different combination of SNPs (signatures) that have the same predictive strength on each one of the examined traits. In **Tables 2–5**, we illustrate one of these combinations, while the rest are illustrated in **Supplementary Tables 2–5**. Finally, the effects of all selected SES SNPs (17 in total, out of which 6 were also found in GWAS) from all traits are presented in **Figures 3–6**.

Selected SNPs for Fat Content

The selected variables/SNPs associated with Fat content (%) at harvest, retrieved from SES algorithm (threshold 0.01), recovered three SNPs, out of which two were located within or proximal to an annotated gene (**Table 2**). The first annotated SNP is located within "telomeres 1 (POT1)" gene (chromosome 8), a region found conserved in other species as well (Medaka, Asian swamp, Asian sea bass). The second SNP was located within the "Rho family GTP-binding" gene (chr13:1098152). However, when lowering the significance threshold to 0.05, the number of SNPs increased to six (**Table 2**).

Selected SNPs for Weight at Harvest

Four selected variables associated with weight at harvest (800 g average weight at harvest) have been retrieved from SES algorithm

TABLE 1 | Selected SNPs from GWAs analysis using linear mixed models, with significance threshold equal to 10^{-4}.

Position	Gene	P-value	Beta coefficient	Conserved	Position
Weight					
Chr1:6970078	Lymphoid enhancer-binding factor 1 isoform X1	3.265E–5	174.721	–	Downstream
Chr1:16636968	Ethanolamine-phosphate cytidylyltransferase-like	5.059E–5	189.556	✓	3'UTR
Chr1:20827142	Mucin-5AC-like isoform X1	4.976E–5	–161.835	✓	Upstream
Chr16:2232897	Acetylserotonin O-methyltransferase-like, LBH-like isoform X1	4.648E–5	–338.149	✓	3'UTR
Chr6:12617755	Transmembrane 199 myosin-7 like	3.838E–5	205.210	✓	Upstream Downstream
Tag Weight					
Chr13:20975921	RNA-binding 27 isoform X1	3.168E–5	4.748	–	Intron
Chr13:20975924	RNA-binding 27 isoform X1	3.168E–5	4.748	–	Intron
Chr2:2623351	Tetratricopeptide repeat 36	2.823E–5	6.183	–	Upstream
Chr22:18343985	tectonin beta-propeller repeat-containing 2	5.405E–5	–5.139	–	Upstream

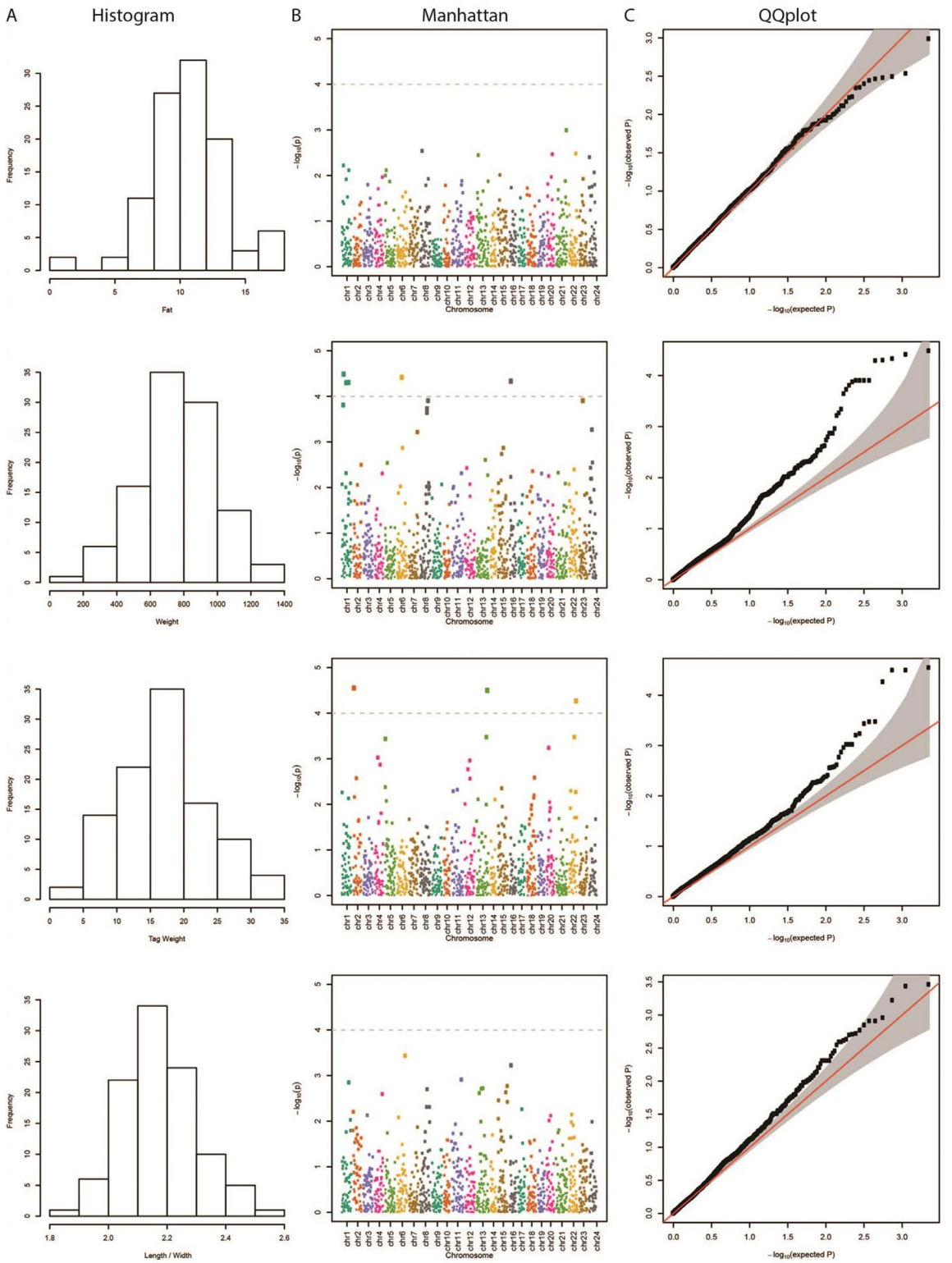

FIGURE 1 | (A) Distribution of each examined trait in our samples. **(B)** Manhattan plot demonstrating the locations across the chromosomes of the sea bream genome (horizontal axis) versus the –log (p-values) of the association between the genetic variants and phenotype (vertical axis). The higher the dots, the stronger the genetic association. The significance threshold was set to 10^{-4}, in order to correct for multiple testing (dashed line). The different colors represent the different chromosomes. **(C)** Quantile–quantile (QQ) plot of the data shown in the Manhattan plot. The grey area represents the 95% simultaneous confidence bands. Red line is the diagonal (Y = X) or else how the observed data should be placed if they were normally distributed.

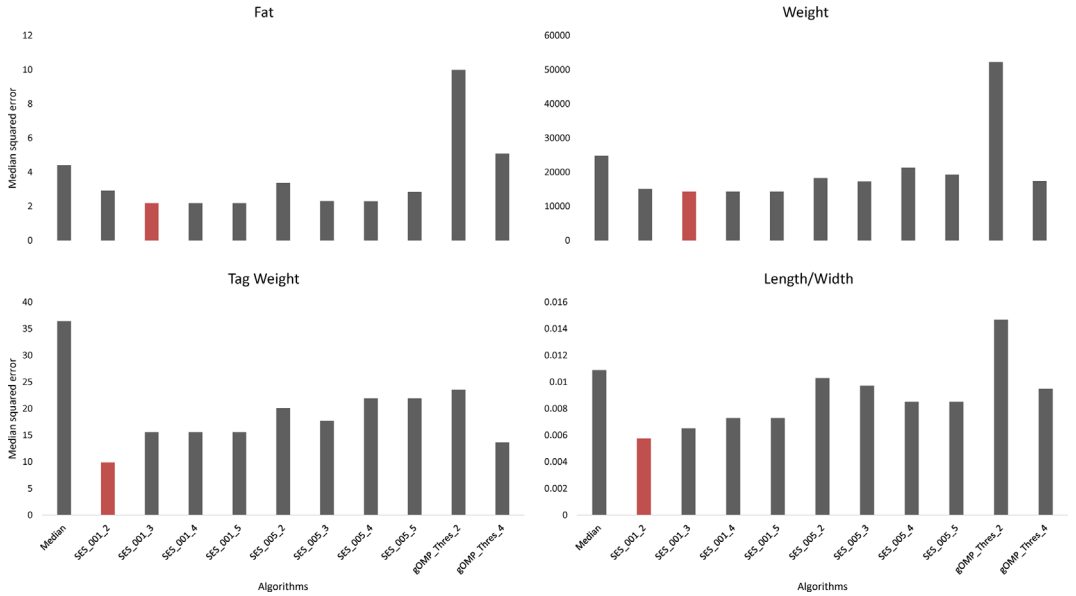

FIGURE 2 | Comparison of different algorithms predicting the traits of interest, based on median squared error, after leave one out cross validation. SES algorithm tested for different thresholds (Threshold equal to 0.01 or 0.05) and for different numbers of SNPs as condition set (k = 2, 3, 4, 5). OMP algorithm tested for different thresholds as stop criterion (Threshold = 2 or 4 units in BIC score).

TABLE 2 | Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Variables	Locus	P-value	Beta coefficient	Threshold	GWAS	Conserved	Position
			Fat				
Chr13:1098152	Rho-related GTP-binding -like	0.007	1.60	0.01	–	–	3' UTR
Chr21:19924408	–	0.006	–1.238	0.01	–	–	–
Chr8: 1385781	Protection of telomeres 1	0.0024	1.55	0.01	–	✓	Intron
Scaffol8147:18634	Death-associated kinase 3-like	0.015	0.7	0.05	–	✓	Intron
Chr7:2453106	Solute carrier family 41 member 1-like isoform X1-2	0.046	0.86	0.05	–	–	Intron
Chr4:23265532	NT-3 growth factor receptor isoform X1	0.017	–2.25	0.05	–	–	Upstream

TABLE 3 | Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error).

Variables	Locus	P-value	Beta coefficient	Threshold	GWAS	Conserved	Position
			Weight				
Chr1:16636968	Ethanolamine-phosphate, cytidylyltransferase-like	0.0006	121.84	0.01	✓	✓	3' UTR
Chr6:12617755	Myosin-7-like isoform X1,short transient receptor potential channel 4-associated	0.0024	138.07	0.01	✓	✓	Upstream
Chr8:11613979	Semaphorin-3A	0.0114	99	0.01	–	✓	Intron
Chr16:2232897	Acetylserotonin O-methyltransferase-like, LBH-like isoform X1	0.0022	–193	0.01	✓	–	3' UTR
Scaffold29:195838	Mitogen-activated kinase-binding 1-like	0.0285	64.669	0.05	–	–	Intron
Chr24:8282385	STE20-related kinase adapter beta Trafficking kinesin-binding 2 isoform X1	0.0022	160.80	0.05	–	✓	Downstream Upstream

TABLE 4 | Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Variables	Locus	P-value	Beta coefficient	Threshold	GWAS	Conserved	Position
			Tag Weight				
Chr2:2623351	Tetratricopeptide repeat 36	0.0019	4.577	0.01	✓	–	Upstream
Chr13:20883924	DNA repair RAD50	0.0127	2.678	0.01	–	✓	Intron
Chr13:20975921	RNA-binding 27 isoform X1	0.0073	1.810	0.01	✓	–	Intron
Chr22:18343985	Zinc finger BED domain-containing 4-like Midasin isoform X2	0.0117	–1.967	0.01	✓	–	Upstream Downstream
Scaffold4139:36071	Predicted uncharacterized protein LOC106518831, partial	0.033	–0.634	0.01	–	✓	Upstream
Chr15:3260819	Follistatin-related 1-like	0.0124	3.106	0.05	–	–	Downstream
Chr20:6671436	UBA-like domain-containing 1	0.021	2.665	0.05	–	✓	2nd
Chr22:14483563	Exostosin-1-like	0.0448	4.898	0.05	–	–	Intron
Scaffold14083:12192	–	0.042	–1.349	0.05	–	–	-

TABLE 5 | Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Variables	Locus	P-value	Beta coefficient	Threshold	GWAS	Conserved	Position
			Length/Width				
Chr6:23799286	Phosphatase 1 regulatory subunit 3D-like	0.0052	0.0397	0.01	–	✓	3d
Chr1:20827142	Upstream: mucin-5AC-like isoform X1	0.049	0.026	0.01	–	✓	Upstream
Chr13:9665394	ATP-dependent RNA helicase DHX33	0.0211	0.048	0.01	–	✓	3d
Chr3:9671223	A-kinase anchor 9 isoform X3	0.0144	–0.0597	0.01	–	✓	2nd
Scaffold13177:8369	Phosphatase 1 regulatory subunit 3C	0.015	0.057	0.01	–	✓	Downstream
Chr8:11613979	Semaphorin-3A	0.0193	–0.025	0.05	–	✓	Intron
Chr22:2545133	Neurexin-3b isoform X3	0.049	–0.029	0.05	–	–	Intron
Scaffold5661:35982	–	0.049	0.031	0.05	–	✓	–

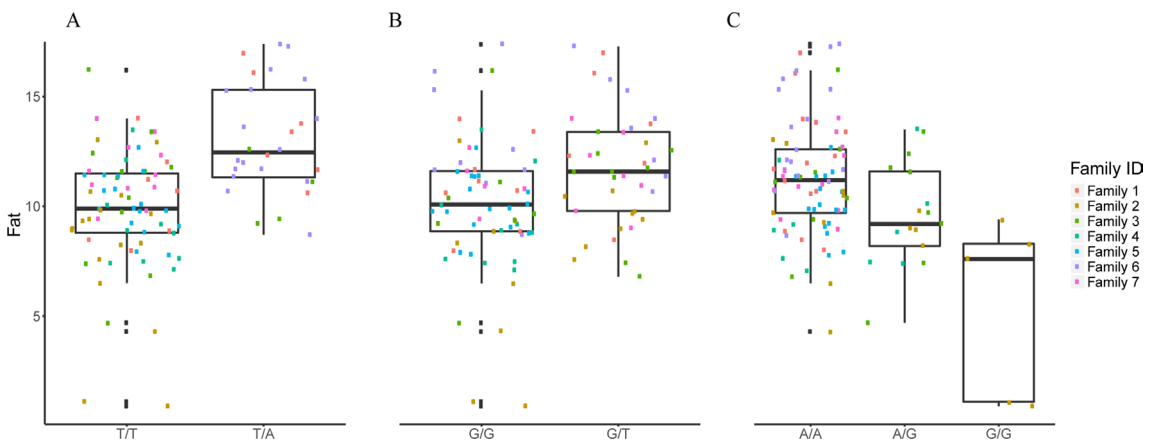

FIGURE 3 | The effect of each of the selected SES SNPs associated with fat content. (A-C) Boxplots of selected SNPs. **(A)** chr8:1385781, **(B)** chr13:1098152, **(C)** chr21:19924408.

with number of condition set equal to three. The first was found in chromosome 1 (chr1:16636968) on "ethanolamine phosphate cytidylyltransferase-like" gene, the second (chr6:12617755) in a conserved region upstream of "myosin-7-like" gene, the third (chr8:11613979) was located in "semaphorin-3A" gene (Conserved in Asian sea bass, Asian swamp eel) and upstream of 'Piccolo' gene, and another one (chr16:2232897) and the fourth on two overlapping genes acetylserotonin O-methyltransferase-like and LBH-like isoform X1. When lowering the significance threshold to 0.05, four SNPs were added to the signatures, retrieving two more annotated genes **(Table 3).**

Selected SNPs for Weight at Tagging

Five SNPs were associated with Tag Weight, as retrieved from SES algorithm (**Table 4**). The first was found at "RNA-binding 27 isoform X1" gene (chr13:20975921), the second upstream

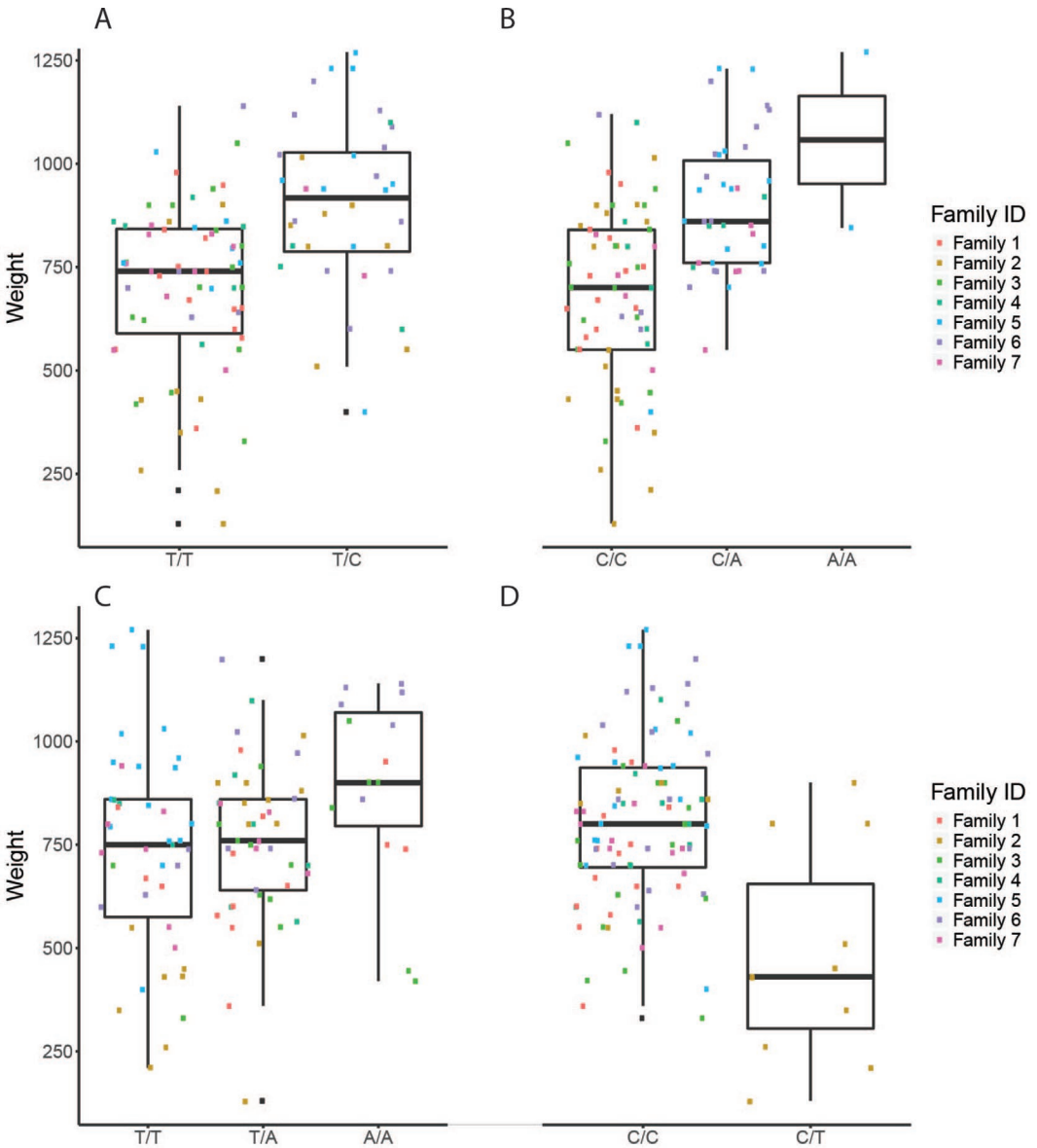

FIGURE 4 | The effect of each of the selected SES SNPs associated with weight at harvest. (A-D) Boxplots of selected SNPs. **(A)** chr1:16636968, **(B)** chr6:12617755, **(C)** chr8:11613979, **(D)** chr16:2232897

from "Tetratricopeptide repeat 36" gene (Chr2:2623351), the third at "DNA repair RAD50" gene (chr13:20883924), the fourth upstream from "tectonin beta-propeller repeat-containing 2" gene (chr22:18343985), and the fifth (scaffold4139:36071) was not in an annotated region. Lowering the significance threshold to 0.05, four annotated SNPs were added to the discovered signatures (**Table 4**).

Selected SNPs for Length/Width Phenotype

Finally, five SNPs were associated with Length/Width ratio (at 750 DPH) as retrieved from SES algorithm (**Table 5**). The first SNP (chr6:23799286) was located on the "phosphatase 1 regulatory subunit 3D-like." The second SNP (chr16:2232897) was located in two genes "acetylserotonin O-methyltransferase-like" and LBH-like isoform X1. The third SNP (chr13:9665394) was located in "ATP-dependent RNA helicase DHX33," the next one in "A-kinase anchor 9 isoform X3," and the last one (scaffold13177:8369) downstream of phosphatase 1 regulatory subunit 3C.

DISCUSSION

Here, we present a family-based approach for the discovery of genetic variants that are significantly associated with a set of phenotypes with economic importance for the farmed gilthead sea bream. The application of these methods on seven families, each measured on four phenotypes, revealed several genetic

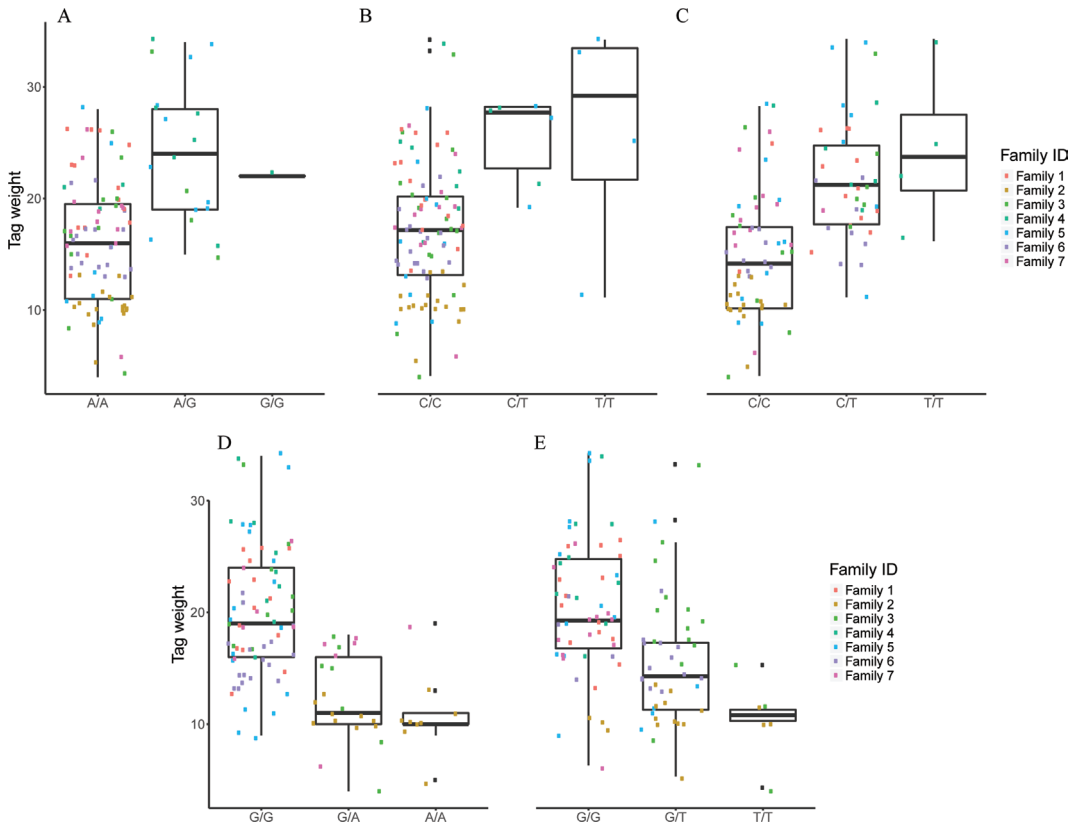

FIGURE 5 | The effect of each of the selected SES SNPs associated with tag weight. (A-E) Boxplots of selected SNPs. **(A)** chr2:2623351, **(B)** chr13:20883924, **(C)** chr13:20975921, **(D)** chr22:18343985, **(E)** scaffold4139:36071.

signatures that may be used for genomic selection. Various QTL affecting growth, morphology, and stress-related traits have been detected using microsatellite markers in gilthead sea bream (Boulton et al., 2011; Loukovitis et al., 2011; Loukovitis et al., 2012; Loukovitis et al., 2013). Some of those QTL have been verified in genetically unrelated populations (Loukovitis et al., 2016). However, no association study using SNP markers was available for production traits in sea bream except this by Palaiokostas et al. (2016) on pasteurelosis. Our study fills this gap enabling for the first time a genomic scan for SNPs that are linked to important traits. We applied two intrinsically different methods. The first is a typical GWA study that examines variants independently, and the second is a family of methods (SES and OMP) that generates signatures with multiple variants.

The sample size of our study (N = 103) might indeed produce some artifacts of this kind. Nevertheless, the analysis pipeline that we apply (SES) is specially tailored for small or moderate sample sizes in order to detect statistically significant QTLs. We anticipate that a future study with greater sample size will refine our findings and might locate additional important QTLs.

In GWA analysis after the LD-pruning, we found 497 independent SNPs. It expected the LD-pruning to reduce drastically the number of SNPs. Studies has shown that a strict LD filters like the one that we applied has minimal effect on the predictive accuracy of the remaining SNPs (Palaiokostas et al., 2019). In general, we noticed a concordance between the SNPs discovered by GWAS and SES. Both methods include tests for SNP–phenotype statistical association, whereas OMP conducts residual-based tests for SNP association. SES algorithm attempts to identify specific sets of SNPs that model a specific phenotype, whereas the typical GWAS pipeline reveals statistical associations. An interpretation of the significance of the SNPs that were located from GWAS but not from SES is that these SNPs do not have a direct effect. Or else, the effect of these SNPs can be eliminated by conditioning on the SNPs that SES revealed. For example, two SNPs that were identified from the typical GWAS, to be associated with weight at tagging (chr13:20975921, chr13:20975924), were marked by SES as equivalents. SES was built upon MMPC algorithm (Tsamardinos et al., 2003). The difference between these two algorithms is that MMPC does not return multiple solutions. MMPC was shown to achieve excellent false positive rates (Aliferis et al., 2010). Seen from the biological perspective, multiple equivalent signatures may arise from redundant mechanisms, for example, genes performing identical tasks within the cell. For example, Ein-Dor et al. (2005) demonstrated that multiple, equivalent prognostic signatures for breast cancer can be extracted just by analyzing the same dataset with a different partition in training and test set, showing the existence of several loci that are practically interchangeable in terms of predictive power. SES was tested against LASSO (Lagani et al., 2017) with continuous, binary, and survival target variables, resulting in SES outperforming the LASSO algorithm (Groll and Tutz, 2014) both in predictive

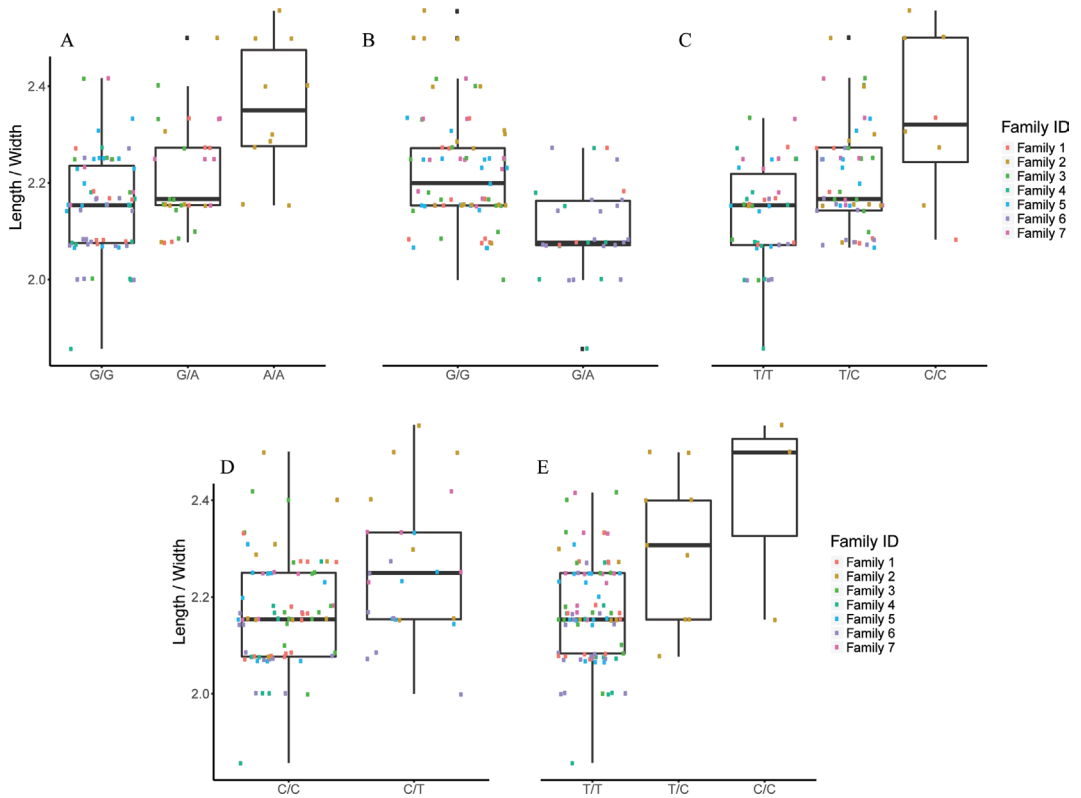

FIGURE 6 | The effect of each of the selected SES SNPs associated with length/width. **(A-E)** Boxplots of selected SNPs. **(A)** chr1:20827142, **(B)** chr3:9671223, **(C)** chr6:23799286, **(D)** chr13:9665394, **(E)** scaffold5661:35982.

performance and computational efficiency. Overall, SES seems to be performing well in smaller datasets, while OMP is known to perform better in larger datasets (Tsagris et al., 2018b). A known limitation in every GWA study is that the power to detect small QTL effects is limited by the number of samples. An under-powered GWA study may fail to detect some associations, whereas the detected signals might be inaccurate in terms of location and/or biological interpretation. The sample size of our study (N = 103) might indeed produce some artifacts of this kind. Nevertheless, the analysis pipeline that we applied (SES) is specially tailored for small or moderate sample sizes in order to detect statistically significant QTLs. We anticipate that a future study with greater sample size will refine our findings and might locate additional important QTLs. Our findings highlight novel SNPs found within or close to coding genes that are significantly associated with our focal traits of interest in sea bream. However, multiple of those genes have been linked with such traits in other species as well. Multiple interesting genes were associated with fat content. For example, one SNP locus is linked with the gene Rho-GTP binding, which is involved in adipogenesis in mice, (Sordella et al., 2003). This gene and its regulator (p190-B RhoGAP) seem to have a key role in the outcome of the differentiation of mesenchymal stem cells to either adipocytes or myocytes (Sordella et al., 2003). Another SNP associated with fat was located on neurotrophin-3 (NT-3), a gene with well-recognized effects on peripheral nerve and Schwann cells, promoting axonal regeneration and associated myelination (Yalvac et al., 2018). NT-3 increases muscle fiber diameter in the neurogenic muscle through direct activation of mTOR pathway and that the fiber size increase is more prominent for fast twitch glycolytic fibers. Thus, fat content seems to be influenced greatly by few genes with well-known role in adipogenesis.

Regarding the loci associated with weight and tag weight, we identified 15 genes in total. Interestingly, although those two traits represent the same trait at different stages, we found no gene associated with both. There are many reasons for such result. One reason may be due to the low power of the experiment and the differences in variation in the weight of the fish at different ages. Another reason may be because different genes are affecting growth at different stages of development. A third reason is that may be the gene action is not only additive and epistatic effects exist. In any case, all these scenarios should be further investigated in a more powerful experiment, which would be necessary in any case. The outcome of our analysis revealed SNPs close to very important genes with a well-known role in weight gain–loss, such as Follistatin, myosin-7, and semaphorin (SEMA3A) genes. Follistatin binds and inhibits the activity of several TGF-family members in mice (Lee and McPherron, 2001). Strikingly, follistatin knockout mice have reduced muscle mass at birth underlying the importance of this gene in muscle growth (Lee and McPherron, 2001). Apart from Follistatin, the significant association with Myosin, an actin-based motor molecule with ATPase activity essential for muscle contraction, shows the

importance of regulation of muscle growth-related genes in weight. The third gene, semaphorin, is significantly associated with both weight and length/width. SEMA3A gene is involved in synapse development underlying the importance of genes in regulating the nervous system in length. Also, the same SNP, which is located on SEMA3A, was direct upstream of Piccolo gene. Piccolo play roles in regulating the pool of neurotransmitter-filled synaptic vesicles present at synapses. Mice lacking Piccolo are viable; nevertheless, each mutant displays abnormalities. Piccolo mutants reduced postnatal viability and body weight (Mukherjee et al., 2010). Another associated gene, ethanolamine phosphate cytidylyltransferase, plays a role in lipid metabolism and finally EXT1, a gene regulating important developmental pathways such as hedgehog (Siekmann and Brand, 2005).

The compilation of an annotated reference genome for this species has been recently published by the Hellenic Centre for Marine Research (H.C.M.R). (Pauletto et al., 2018) and is also available on the Genome Browser[1]. To our knowledge, this analysis is the first to use this genome as a reference for read alignment and variant calling. Moreover, a literature review did not reveal any study examining the same collection of traits on this species. As an effect, for the moment, we cannot provide a comparative analysis with other studies. Studies on related species include those of Yoshida et al. (2019), which examines weight paper on Nile tilapia, Nguyen et al. (2018), which examines weight on Yellowtail Kingfish, and Yu et al. (2018), which examines weight and total length on Epinephelus coioides. Although our study does not have any common gene with these studies, it is interesting that among these studies, there are also no common genes. This suggests the high genetic variability on these traits across different species and also the need for future studies with higher sample sizes and better coverage that can provide additional insights on the common genetic content of aquacultured species.

CONCLUSION

In this study, we employed two different approaches to identify variants associated with growth-related phenotypic traits. Our chosen selected panel combined with the vigorous bioinformatic analyses revealed the most significant SNP loci on the sea bream genome. The discovered candidates are located in the proximity of genes with known involvement in processes related to growth. The combination of these novel loci may lead to the selection of brooders based on specific genetic signatures and can have a great effect on the efficiency of the aquaculture. Moreover, these results could be used to verify or not putative QTL identified in previous studies and could also be used in order to fine map identified QTL in the same population using other types of genetic markers (Chatziplis et al, 2018, in preparation). Following this step, the use of these variants independently as individual SNP (or SNP haplotypes) and/or in combination with other marker information in a MAS program could be a form of direct application in the aquaculture breeding industry. When more dense SNP markers would be available (i.e., SNPchip) for the species and more families from more populations are genotyped (i.e., increase LD), then the application of Genomic Selection will be more feasible and cost effective in terms of any selection accuracy benefits. Nevertheless, our study presents, in a small scale example, the feasibility of GS application as well as the availability of the tools necessary before its application (i.e., GWAS using SNP markers) in an important Mediterranean aquaculture species such as gilthead sea bream.

[1] http://biocluster.her.hcmr.gr/myGenomeBrowser?portalname=Saurata_v 1

ETHICS STATEMENT

Animal welfare was achieved according to the "Guidelines for the treatment of animals in behavioural research and teaching" (Guidelines for the treatment of animals in behavioural research and teaching, 1997) (see also Tsakogiannis et al., 2018). All fish utilized in the study were kept in registered and authorized facilities to maintain and perform animal experiments; rearing and sampling followed the guidelines of the Directive 2010/63/EU for the protection of animals used for experimental and other scientific purposes (Official Journal L276/33) (EU, 2010. Directive 2010/63/EU of the European Parliament and the Council of 22 September 2010 on the protection of animals used for scientific purposes. Official Journal of the European Union L 276/33, Animal protection.). In addition, experimental sampling protocols were approved by the IMBBC's aquaculture department committee and methods were in accordance with relevant guidelines and regulations approved by the Hellenic Ministry of Rural Development and Food and the Regional Directorate of Veterinary Medicine for certified experimental installations (EL 91-BIO-04) and experimental animal breeding (AQUALABS, EL 91-BIO-03). Laboratory personnel include accredited technicians by the Federation for Laboratory Animal Science Associations (FELASA).

AUTHOR CONTRIBUTIONS

CT, GP, TM, AK, and DK conceived and designed the study. LP, DC, and CT designed and performed the family selection. AT performed the DNA extraction and ddRAD library preparation. DK performed the bioinformatic analyses with guidance from AK and TM. DK performed the statistical analyses with guidance from MT and IT. DK wrote the first draft of the manuscript. MT, AT, DC, AK, and TM wrote sections of the manuscript. All authors contributed to manuscript revision and read and approved the submitted version.

FUNDING

Financial support for this study has been provided by the General Secretariat for Research and Technology (GSRT), Ministry of Education and Religious Affairs, under the National Programme for Competitiveness & Entrepreneurship (EPAN II) funded by National sources and the European Regional Development Fund for the gilthead sea bream. This research was supported in part through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology, and Aquaculture

of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI, and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00675/full#supplementary-material

REFERENCES

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010). Local causal and Markov Blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *J. Mach. Res.* 11, 235–284.

Anderson, J. L., Marí, A., Braasch, I., Amores, A., Hohenlohe, P., Batzel, P., et al. (2012). Multiple sex-associated regions and a putative sex chromosome in zebrafish revealed by RAD mapping and population genomics. *PLoS One* 7, 1–14. doi: 10.1371/journal.pone.0040701

Andrews, S., and Babraham Bioinformatics Group. (2010). Fastqc: a quality control tool for high throughput sequence data.

Antonopoulou, E., Kaitetzidou, E., Castellana, B., Panteli, N., Kyriakis, D., Vraskou, Y., et al. (2017). In vivo effects of lipopolysaccharide on peroxisome proliferator-activated receptor expression in juvenile gilthead seabream (Sparus Aurata). *Biology* 6, 36. doi: 10.3390/biology6040036

Bahi, A., Guardiola, F., and Esteban, M. (2018). A time course study of glucose levels and innate immune response in gilthead sea bream (sparus aurata l). after exposure to clove oil eugenol derived anaesthetic. *Fish Shellfish Immunol.* 77, 280–285. doi: 10.1016/j.fsi.2018.03.057

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3, 1–7. doi: 10.1371/journal.pone.0003376

Balliu, B., and Zaitlen, N. (2016). A novel test for detecting SNP-SNP interactions in case-only trio studies. *Genetics* 202, 1289–1297. doi: 10.1534/genetics.115.179846

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *ArXiv e-prints*. doi: 10.18637/jss.v067.i01

Besson, M., Allal, F., Chatain, B., Vergnet, A., Clota, F., and Vandeputte, M. (2019). Combining individual phenotypes of feed intake with genomic data to improve feed efficiency in sea bass. *Front. Genet*. 10, 219. doi: 10.3389/fgene.2019.00219

Boulton, K., Massault, C., Houston, R. D., de Koning, D. J., Haley, C. S., Bovenhuis, H., et al. (2011). QTL affecting morphometric traits and stress response in the gilthead seabream (Sparus aurata). *Aquaculture* 319, 58–66. doi: 10.1016/j.aquaculture.2011.06.044

Cai, T. T., and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inf. Theory* 57, 4680–4688. doi: 10.1109/TIT.2011.2146090

Catchen, J. M. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol*. 22, 3124–3140. doi: 10.1111/mec.12354

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., and Andrew, P. (2011). Europe PMC Funders Group Basic statistical analysis in genetic case-control studies. *Nat. Protoc.* 6, 121–133. doi: 10.1038/nprot.2010.182

Ein-Dor, L., Kela, I., Getz, G., Givol, D., and Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 21, 171–178. doi: 10.1093/bioinformatics/bth469

Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A., Cresko, W. A. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol. Biol.* 772, 157–178. doi: 10.1007/978-1-61779-228-1_9

Fang, Y. (2011). Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. *J. Data Sci*. 9, 15–21.

FEAP (2017). Federation of european aquaculture producers. *Annual report 2017*.

Fernandes, T., Herlin, M., Belluga, M. D. L., Ballón, G., Martinez, P., Toro, M. A., et al. (2017). Estimation of genetic parameters for growth traits in a hatchery population of gilthead sea bream (Sparus aurata L). *Aquacult. Int*. 25, 499–514. doi: 10.1007/s10499-016-0046-5

Fontanarosa, J. B., and Dai, Y. (2011). Using lasso regression to detect predictive aggregate effects in genetic studies. *BMC Proc*. 5, S69. doi: 10.1186/1753-6561-5-S9-S69

Geng, X., Zhi, D., and Liu, Z., (2017). "Genome-wide association studies of performance traits," in *Bioinformatics in Aquaculture: Principles and Methods*. John Wiley & Sons, Ltd. 415–433. doi: 10.1002/9781118782392.ch23

Goddard, M., and Hayes, B. (2007). Genomic selection. *J. Anim. Breed. Genet*. 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., et al. (2012). Gwastools: an r/bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 28, 3329–3331. doi: 10.1093/bioinformatics/bts610

Groll, A., and Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Stat. Comput*. 24, 137–154. doi: 10.1007/s11222-012-9359-z

Guardiola, F., Bahi, A., Jiménez-Monreal, A., Martínez-Tomé, M., Murcia, M., and Esteban, M. (2018). Dietary administration effects of fenugreek seeds on skin mucosal antioxidant and immunity status of gilthead seabream (sparus aurata l). *Fish Shellfish Immunol*. 75, 357– 364. doi: 10.1016/j.fsi.2018.02.025

Gutierrez, A. P., Yáñez, J. M., Fukui, S., Swift, B., and Davidson, W. S. (2015). Genome-wide association study (gwas) for growth rate and age at sexual maturation in atlantic salmon (salmo salar). *PLoS One* 10, 1–15. doi: 10.1371/journal.pone.0119730

Heffner, E. L., Jannink, J.-L., and Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4, 65–75. doi: 10.3835/plantgenome2010.12.0029

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., et al. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* 11, 724. doi: 10.1186/1471-2164-11-724

Khatkar, M. S. (2017). "Genomic selection in aquaculture breeding programs," in *Bioinformatics in Aquaculture: Principles and Methods*. John Wiley & Sons, Ltd. 380–391. doi: 10.1002/9781118782392

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature selection with the R package MXM: discovering statistically-equivalent feature subsets. *J. Stat. Softw.* 80. doi: 10.18637/jss.v080.i07

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lee, S.-J., and McPherron, A. C. (2001). Regulation of myostatin activity and muscle growth. *Proc. Natl. Acad. Sci*. 98, 9306–9311. doi: 10.1073/pnas.151270098

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Lie, H. C. (2014). Towards breaking the curse of dimensionality in computational methods for the conformational analysis of molecules. *BMC Bioinf*. 15, A2. doi: 10.1186/1471-2105-15-S3-A2

Linnaeus, C. (1758). *Systema Nature.* Ed. Tomus I. 10. Holmiae: Laurentii Salvii, (1–4), 1–824.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H. et al., (2011). "Chapter two - genomic selection in plant breeding: knowledge and prospects," in *Advances in Agronomy*, vol. 110. Ed. D. L. Sparks (Academic Press), 77– 123. doi: 10.1016/B978-0-12-385531-2.00002-5

Loukovitis, D., Batargias, C., Sarropoulou, E., Apostolidis, A. P., Kotoulas, G., Magoulas, A., et al. (2013). Quantitative trait loci affecting morphology traits in gilthead seabream (sparus aurata l). *Anim. Genet*. 44, 480–483. doi: 10.1111/age.12027

Loukovitis, D., Sarropoulou, E., Tsigenopoulos, C. S., Batargias, C., Magoulas, A., Apostolidis, A. P., et al. (2011). Quantitative Trait Loci involved in sex determination and body growth in the gilthead sea bream (Sparus aurata L). through targeted genome scan. *PLoS One* 6, 1–9. doi: 10.1371/journal.pone.0016599

Loukovitis, D., Sarropoulou, E., Vogiatzi, E., Tsigenopoulos, C. S., Kotoulas, G., Magoulas, A., et al. (2012). Genetic variation in farmed populations of the gilthead sea bream sparus aurata in greece using microsatellite dna markers. *Aquacult. Res*. 43, 239–246. doi: 10.1111/j.1365-2109.2011.02821.x

Loukovitis, D., Siasiou, A., Mitsopoulos, I., Lymberopoulos, A. G., Laga, V., and Chatziplis, D. (2016). Genetic diversity of Greek sheep breeds and transhumant populations utilizing microsatellite markers. *Small Rumin. Res.* 136, 238–242. doi: 10.1016/j.smallrumres.2016.02.008

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi: 10.1093/bioinformatics/btq559

Manousaki, T., Tsakogiannis, A., Taggart, J. B., Palaiokostas, C., Tsaparis, D., Lagnel, J., et al. (2016). Exploring a nonmodel teleost genome through rad sequencing—linkage mapping in common pandora, pagellus erythrinus and comparative genomic analysis. *G3: Genes, Genomes, Genet.* 6, 509–519. doi: 10.1534/g3.115.023432

Miller, S. A., Dykes, D. D., and Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 16, 1215. doi: 10.1093/nar/16.3.1215

Mukherjee, K., Yang, X., Gerber, S. H., Kwon, H.-B., Ho, A., Castillo, P. E., et al. (2010). Piccolo and bassoon maintain synaptic vesicle clustering without directly participating in vesicle exocytosis. *Proc. Natl. Acad. Sci*. 107, 6504–6509. doi: 10.1073/pnas.1002307107

Negrín-Báez, D., Navarro, A., Lee-Montero, I., Soula, M., Afonso, J. M., and Zamorano, M. J. (2015). Inheritance of skeletal deformities in gilthead seabream (sparus aurata) – lack of operculum, lordosis, vertebral fusion and lsk complex1. *J. Anim. Sci*. 93, 53–61. doi: 10.2527/jas.2014-7968

Nguyen, N., Rastas, P., Premachandra, H., and Knibb, W. (2018). First high-density linkage map and single nucleotide polymorphisms significantly associated with traits of economic importance in yellowtail kingfish seriola lalandi. *Front. Genet*. 9, 127. doi: 10.3389/fgene.2018.00127

Palaiokostas, C., Bekaert, M., Khan, M. G., Taggart, J. B., Gharbi, K., McAndrew, B. J., et al. (2013). Mapping and validation of the major sex-determining region in Nile tilapia (Oreochromis niloticus L). using RAD sequencing. *PLoS One* 8, 1–9. doi: 10.1371/journal.pone.0068389

Palaiokostas, C., Ferraresso, S., Franch, R., Houston, R. D., and Bargelloni, L. (2016). Genomic prediction of resistance to pasteurellosis in gilthead sea bream (Sparus aurata) Using 2b-RAD Sequencing. *G3: Genes–Genomes–Genet.* 6 (11), 3693–3700. doi: 10.1534/g3.116.035220

Palaiokostas, C., Kocour, M., Prchal, M., and Houston, R. D. (2018). Accuracy of genomic evaluations of juvenile growth rate in common carp (cyprinus carpio) using genotyping by sequencing. *Front. Genet*. 9, 82. doi: 10.3389/fgene.2018.00082

Palaiokostas, C., Vesely, T., Kocour, M., Prchal, M., Pokorova, D., Piackova, V., et al. (2019). Optimizing genomic prediction of host resistance to koi herpesvirus disease in carp. *BioRxiv* 609784. doi: 10.3389/fgene.2019.00543

Paris, J. R., Stevens, J. R., and Catchen, J. M. (2017). Lost in parameter space: a road map for stacks. *Methods Ecol. Evol.* 8, 1360–1373. doi: 10.1111/2041-210X.12775

Pauletto, M., Manousaki, T., Ferraresso, S., Babbucci, M., Tsakogiannis, A., Louro, B., et al. (2018). Genomic analysis of Sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish. *Commun. Biol.* 1, 119. doi: 10.1038/s42003-018-0122-7

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 81, 559–575. doi: 10.1086/519795

Siekmann, A. F., and Brand, M. (2005). Distinct tissue-specificity of three zebrafish ext1 genes encoding proteoglycan modifying enzymes and their relationship to semitic Sonic Hedgehog signaling. *Dev. Dyn*. 232, 498–505. doi: 10.1002/dvdy.20248

Silva-Marrero, J. I., Sáez, A., Caballero-Solares, A., Viegas, I., Almajano, M. P., Fernández, F., et al. (2017). A transcriptomic approach to study the effect of long-term starvation and diet composition on the expression of mitochondrial oxidative phosphorylation genes in gilthead sea bream (Sparus aurata). *BMC Genomics* 18, 1–16. doi: 10.1186/s12864-017-4148-x

Sonesson, A. K., and Meuwissen, T. H. (2009). Testing strategies for genomic selection in aquaculture breeding programs. *Genet. Sel. Evol.* 41, 37. doi: 10.1186/1297-9686-41-37

Sordella, R., Jiang, W., Chen, G. C., Curto, M., and Settleman, J. (2003). Modulation of Rho GTPase signaling regulates a switch between adipogenesis and myogenesis. *Cell* 113, 147–158. doi: 10.1016/S0092-8674(03)00271-X

Statnikov, A., and Aliferis, C. F. (2010). Analysis and computational dissection of molecular signature multiplicity. *PLoS Comput. Biol*. 6, 1–9. doi: 10.1371/journal.pcbi.1000790

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol*. 13, 1–11. doi: 10.1371/journal.pbio.1002195

Tapia-Paniagua, S. T., Ceballos-Francisco, D., Balebona, M. C., Ángeles Esteban, M., and Ángel Moriñigo, M. (2018). Mucus glycosylation, immunity and bacterial microbiota associated to the skin of experimentally ulcered gilthead seabream (sparus aurata). *Fish Shellfish Immunol*. 75, 381– 390. doi: 10.1016/j.fsi.2018.02.006

Tibshirani, R. J., and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.* 3, 822–829. doi: 10.1214/08-AOAS224

Tsagris, M., Lagani, V., and Tsamardinos, I. (2018a). Feature selection for high-dimensional temporal data. *BMC Bioinf.* 19, 17. doi: 10.1186/s12859-018-2023-7

Tsagris, M., Papadovasilakis, Z., Lakiotaki, K., and Tsamardinos, I. (2018b). Efficient feature selection on gene expression data: which algorithm to use? *bioRxiv*. doi: 10.1101/431734

Tsamardinos, I., and Aliferis, C. F. (2003). Towards principled feature selection: relevancy, filters and wrappers In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, AISTATS 2003, Key West, Florida, USA, January 3-6, 2003

Tsamardinos, I., Aliferis, C. F., and Statnikov, A., (2003). Time and sample efficient discovery of Markov Blankets and direct causal relations In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM) 673–678. doi: 10.1145/956804.956838

Tsamardinos, I., and Brown, L. E. (2008). "Bounding the false discovery rate in local bayesian network learning," in *AAAI*, 1100–1105.

Tsigenopoulos, C. S., Louro, B., Chatziplis, D., Lagnel, J., Vogiatzi, E., Loukovitis, D., et al. (2014). Second generation genetic linkage map for the gilthead sea bream Sparus aurata L. *Mar. Genomics* 18, 77–82. doi: 10.1016/j.margen.2014.09.008

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychol. Methods* 17, 228. doi: 10.1037/a0027127

Vélez, E. J., Azizi, S., Lutfi, E., Capilla, E., Moya, A., Navarro, I., et al. (2017). Moderate and sustained exercise modulates muscle proteolytic and myogenic markers in gilthead sea bream (sparus aurata). *Am. J. Physiol. Regul. Integr. Comp. Physiol*. 312, R643–R653. doi: 10.1152/ajpregu.00308.2016

Vélez, E. J., Perelló, M., Azizi, S., Moya, A., Lutfi, E., Pérez-Sánchez, J., et al. (2018). Recombinant bovine growth hormone (rbgh) enhances somatic growth by regulating the gh-igf axis in fingerlings of gilthead sea bream (sparus aurata). *Gen. Comp. Endocrinol*. 257, 192– 202. doi: 10.1016/j.ygcen.2017.06.019

Wang, L., Liu, P., Huang, S., Ye, B., Chua, E., Wan, Z. Y., et al. (2017). Genome-wide association study identifies loci associated with resistance to viral nervous necrosis disease in asian seabass. *Mar. Biotechnol*. 19, 255–265. doi: 10.1007/s10126-017-9747-7

Yalvac, M. E., Amornvit, J., Chen, L., Shontz, K. M., Lewis, S., and Sahenk, Z. (2018). AAV1.NT-3 gene therapy increases muscle fiber diameter through activation of mTOR pathway and metabolic remodeling in a CMT mouse model. *Gene Ther*. 25 (2), 129–138. doi: 10.1038/s41434-018-0009-8

Yoshida, G., Paul Lhorente, J., Correa, K., Soto, J., and Yáñez, J. (2019). Genome-wide association study and low-cost genomic predictions for growth and fillet yield in nile tilapia (oreochromis niloticus). *bioRxiv*. doi: 10.1101/573022

Yu, H., You, X., Li, J., Zhang, X., Zhang, S., Jiang, S., et al. (2018). A genome-wide association study on growth traits in orangespotted grouper (epinephelus coioides) with rad-seq genotyping. *Sci. China Life Sci.* 61, 1–13. doi: 10.1007/s11427-017-9161-4

Yue, G. H. (2014). Recent advances of genome mapping and marker-assisted selection in aquaculture. *Fish Fish*. 15, 376–396. doi: 10.1111/faf.12020

Conflict of Interest Statement: Author LP was employed by company Nireus Aquaculture SA, Greece. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.