UNIVERSITY
OF CRETE

Scanning of genetic variants and genetic mapping of phenotypic traits in gilthead seabream through ddRAD sequencing data analysis

Dimitrios Kyriakis

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Bioinformatics

University of Crete
School of Medicine
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Ioannis Tsamardinos*

This work has been performed at the University of Crete, School of Medicine.
The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).
The work has been supported by the Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR) Crete, Greece.
The work has been supported by the Department of Agricultural Technology, Alexander Technological Education Institute of Thessaloniki, Greece.
The work has been supported by the Nireus Aquaculture SA, Greece.

UNIVERSITY
OF CRETE

Γενετική χαρτογράφηση και αναζήτηση μεταλλάξεων που επηρεάζουν την απόδοση ιχθυοκαλλιέργειας τσιπούρας (*Sparus aurata*) μέσω ανάλυσης δεδομένων από ddRAD αλληλούχηση

Δημήτριος Κυριάκης

Πτυχιακή εργασία

Μεταπτυχιακές σπουδές στην Βιοπληροφορική

Πανεπηστίμιο Κρήτης

Τμήμα Ιατρικής

Βούτες, 700 13 Ηράκλειο, Κρήτη, Ελλάδα

Επιβλέποντας Διπλωματικής: Prof. Ιωάννης Τσαμαρδίνος

2 Οκτωβρίου 2018

UNIVERSITY OF CRETE
FACULTY OF MEDICINE

Scanning of genetic variants and genetic mapping of phenotypic traits in gilthead seabream through ddRAD sequencing data analysis

Thesis submitted by
Dimitrios Kyriakis
in partial fulfillment of the requirements for the
Masters' of Science degree in Bioinformatics

THESIS APPROVAL

Author: _____
Dimitrios Kyriakis

Committee approvals: _____
Ioannis Tsamardinos
Professor, Thesis Supervisor

_____
George Potamias
Doctor, Committee Member

_____
Pavlos Pavlidis
Doctor, Committee Member

Departmental approval: _____
Nektarios Tavernarakis
Professor, Director of Graduate Studies

Heraklion, 2/10/2018

Abstract

Gilthead seabream (*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 seabream 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 that produces groups with multiple-SNPs associated to a phenotype. In total, we identified nine single-SNPs and six groups of SNPs associated with weight related phenotypes (Weight and Tag Weight), two groups associated with Fat, and 16 groups associated with the Length to Width ratio. Six identified loci 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 seabream aquaculture productivity by providing genomic anchors for selection programs.

Abstract

Η τσιπούρα Gilthead (*Sparus aurata*) είναι ένας τελεόστεος, μεγάλης οικονομικής αξίας στη νοτιοευρωπαϊκή υδατοκαλλιέργεια. Οι υδατοκαλλιέργειες δείχνουν ένα αυξανόμενο ενδιαφέρον για την εφαρμογή γενετικών μεθόδων που μπορούν να εντοπίσουν συσχετίσεις φαινοτύπου-γονότυπου με υψηλό οικονομικό αντίκτυπο. Μέσω της εκλεκτικής αναπαραγωγής, ο κλάδος της υδατοκαλλιέργειας μπορεί να εκμεταλλευτεί αυτές τις πληροφορίες για να μεγιστοποιήσει τη χρηματοοικονομική του απόδοση. Στην μελέτη αυτή, παρουσιάζουμε μια Genome Wide Association analysis (GWAS) με 112 δείγματα που ανήκουν σε επτά διαφορετικές οικογένειες, οι οποίες συλλέχθηκαν από μια ελληνική εμπορική εταιρεία υδατοκαλλιέργειας. Μέσω double digest Random Amplified DNA (ddRAD), δημιουργήσαμε ένα γενετικό προφίλ ανά δείγμα που αποτελείται από 2.258 υψηλού επιπέδου πολυμορφισμούς νουκλεοτιδίων (SNPs). Αυτά τα προφίλ δοκιμάστηκαν για συσχέτιση με τέσσερις φαινοτύπους μείζονος οικονομικής σημασίας: Λίπος, Βάρος κατα την σήμανση, Βάρος κατα την αλιεία και το λόγο Μήκος προς Πλάτος. Εφαρμόσαμε δύο μεθόδους ανάλυσης συσχετισμού. Η πρώτη είναι η τυπική συσχέτιση φαινοτύπου με ένα SNP και η δεύτερη μέθοδος επιλογής χαρακτηριστικών (FS) που παράγει ομάδες απο πολλαπλά SNP που σχετίζονται με ένα φαινότυπο. Συνολικά, εντοπίσαμε εννέα SNP και έξι ομάδες SNP που σχετίζονται με το βάρος, δύο ομάδες που σχετίζονται με το λίπος και 16 ομάδες που σχετίζονται με το λόγο μήκος προς πλάτος. Έξι απο τους τόπους που εντοπίσθηκαν υπήρχαν σε γονίδια που σχετίζονταν με την ανάπτυξη σε άλλους τελεόστεους ή ακόμα και θηλαστικά, όπως η semaphorin-3A και η neurotrophin-3. Αυτοί οι τόποι είναι ισχυροί υποψήφιοι για μελλοντικές μελέτες που θα μας βοηθήσουν να αποκαλύψουμε τους γενετικούς μηχανισμούς που αποτελούν τη βάση για την ανάπτυξη και να βελτιώσουμε την παραγωγικότητα της υδατοκαλλιέργειας τσιπούρας, παρέχοντας γενομικούς δείκτες για προγράμματα επιλογής.

Acknowledgements

I would like to thank Dr. A.Kanterakis, Dr. T.Manousaki and Dr. M.Tsagris for their support to my work. Their contribution helped me get results of better quality. I am also grateful to the members of my supervising committee Prof. I.Tsamardinos, Dr. G.Potamias and Dr. P.Pavlidis for their guidance. Also, I want to thank the Hellenic Center of Marine Research (HCMR) and especially Dr. C.S. Tsigenopoulos for their trust. Finally, I want to thank Dr. A. Tsakogianni, Dr. D. Chatzipli and L. Papaharisi for their contribution to this project regarding the data preparation.
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 (BREAMIMPROVE).

Table of Contents

1 Introduction

1.1 *Genome Wide Association Studies*

Genome Wide Association Studies (GWAS) has accelerated the field of human, plant, and livestock genetics [1]. Using GWAS, plenty genetic risk factors for many common human diseases have been identified, and many genetic regions regulating crucial economical traits have been located in plants and livestock. GWAS could help us to understand the relations between traits and the underlying genetic architecture in aquaculture [1]. While genotyping technologies are evolving, GWAS could be widely used for the analysis of aquaculture traits to improve the brood stocks of aquaculture species, with lower costs in the long term [1].

1.2 *Sparus aurata*

The gilthead sea bream (*Sparus aurata*) is a fish species of great economic importance for the Mediterranean aquaculture industry [2]. It ranks first among other aquacultured species in South Mediterranean with total production of 160,563 tons for 2016 (FEAP Production Report 2017). The main producers are Turkey and Greece, representing 42% and 37% respectively of the total production.
The Marine aquaculture industry, especially in Mediterranean, shows a growing interest in genetic improvement in order to maximize the efficiency of its production [3]. Genomic selection is a breeding methodology that aims to increase the rate of genetic gain, leading to improved of certain phenotypes of various species. Selection design should take into consideration both the minimization of inbreeding and the maximization of the response to selection [3]. Multitrait selection is desired for two reasons (1) avoids changes in traits, (2) maximizes productivity resulting from genetic improvement programmes [3]. In order to make the genomic selection, we developed a prediction model for the trait of interest using a training population.

Country	Sea bream production (tons)
Turkey	67,612
Greece	59,000
Spain	13,740
Italy	7,600
Cyprus	5,136
Croatia	4,304
France	1,671
Portugal	1,500
Total	**160,563**

Table 1.1: Gilthead seabream production volume (FEAP Production Report 2017).

1.3 Double Digest Restriction Associated DNA (ddRAD) Sequencing

The determination of genotypes is a grounding technology in genetics. It is mainly used in research for genotype-phenotype association studies and in clinical diagnostics for variant discovery. The number of individuals or samples has crucial role for the reliability and statistical power of a comparative and population analyses. Several methods have occurred to increase the number of individuals keeping the resource investment at the same level. The basic method is reducing the fraction of each individual genome sequenced [4]. The main challenge of this method is to acquire an adequate and representative coverage of the genome of the population given that short regions are profiled for each individual. Double digest RAD-sequencing (ddRADseq), uses a two enzyme double digest followed by precise size selection (Figure: 1.1) [5], [6]. Only a very small fraction of the fragments are sequenced. These fragments are naturally selected to be from the same genomic regions across individuals. Representation, is expected to be inversely proportional to deviation from the size-selection target, thus read counts across regions are expected to be correlated between individuals [4].

The libraries, produced by ddRADseq, consist of fragments generated by cuts with both restriction enzymes which fall within the size-selection window. Some of the advantages are:

- **Reduces duplicate**
  Small fraction of restriction fragments will fall in the target size-selection regime, the probability of sampling both directions from the same restriction site is low.

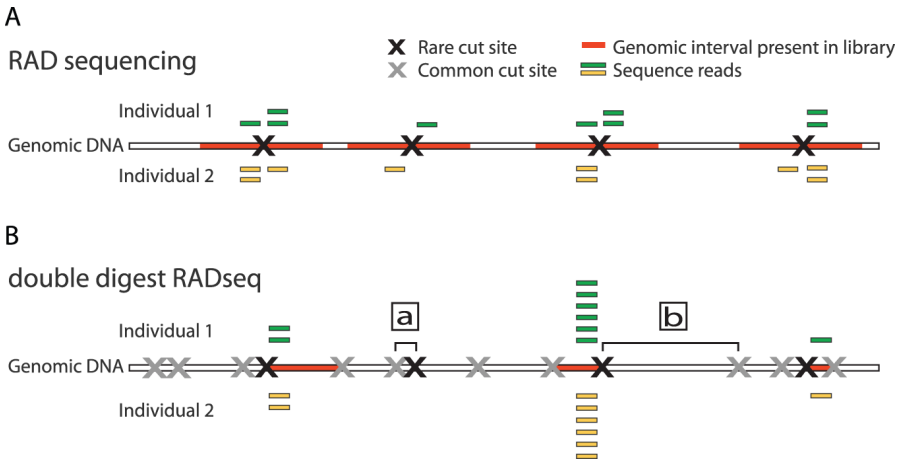


Figure 1.1: The outline of Double digest RAD sequencing (ddRADseq) technology. ddRAD improves efficiency and robustness while minimizing cost. (A) Single restriction enzyme (RE), (B) Double digest RAD-sequencing [4].

Effectively halves the number of reads that are required to reach high-confidence sampling of a SNP associated with a given restriction enzyme (RE) cut site [4].

- **Robust to under-sampling**
  Regions are "filled in" with reads in approximately the same order across all individual samples. Samples with read recovery counts below saturation will still share a significant number of well-covered regions [4].

The preparation of a ddRADseq library requires less time while is of low cost and, in addition, needs finite amounts of genomic material [4]. Furthermore, by removing the random shearing (and therefore random recovery), increases the coverage of regions across individuals and augments robustness and variability of read counts [7].

1.4 Stacks

Stacks is a software pipeline for variant discovery from short-read sequences, such as those generated on the Illumina platform. Stacks was developed to work with restriction enzyme-based data, such as RAD-seq, for the purpose of building genetic maps and conducting population genomics and phylogeography [8]. The stacks pipeline can be summarized as:

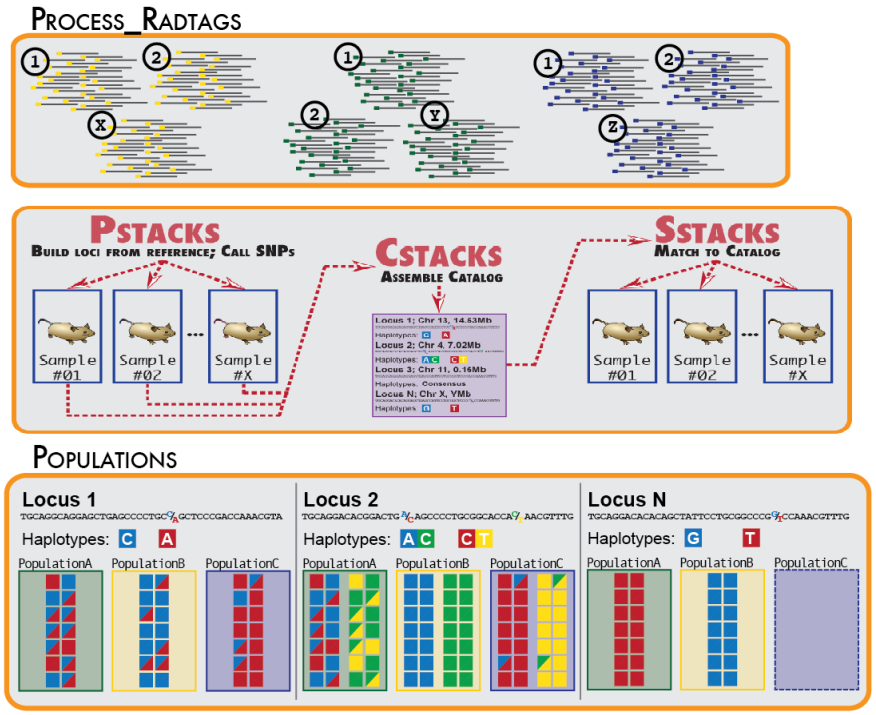


Figure 1.2: The outline of stacks pipeline. 1) `Process radtags`: Raw sequence reads are demultiplexed and apply filtering and quality control. 2) `Pstacks` extracts stacks of reads, that have been aligned to a reference genome, per individual and identifies single polymorphic nucleotide. 3) `Cstacks`: Loci are clustered together across parents and a loci catalogue is generated. 4) `Sstacks`: Loci from each individual are matched against the catalogue constructed by parental alleles. 5) `Populations` function of Stacks, used to produce a vcf file and several useful output files [8].

Also, a web based front end, backed by a MySQL database, is available to visualize the data [8].

1.5 Genotype Imputation

RAD sequencing typically has a large proportion of missing data. Through genotype imputation we can produce estimates of these missing values based on proximal SNPs with known genotypes. Imputing missing values can maximize the power of an analysis and reduce genotyping cost [9]. Imputation methods, model the correlation between SNPs that occur due to LD and use these models for missing value inference [10]. Imputation can be also used to augment the number of SNPs that have been typed by a genotyping platform, but this requires the availability of a densed genotyped reference panel [11]. Some of the most known tools for this task rely on a reference genome in order to construct probabilistic haplotypes which in turn are used for genotype inference [10].
Marker imputation algorithms were developed for species with a reference genome, were the markers are ordered [12]. Based on previous research [13], two parameters have been found to affect the accuracy of imputation of unordered markers. The first is linkage disequilibrium and the second is relatedness [13]. There are many general imputation methods that do not require any prior information about the variables to be imputed, but they have not been tested for imputation accuracy of genome-wide marker data. Some of those imputation strategies are K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest Imputation (RF).
Given that an imputation reference panel is not available for gilthead seabream we used imputation solely for the purpose of inferring missing values. In our case, the available reference genome of gilthead is in a drafting stage and contains numerous markers in unmapped sequencing scaffolds. For this reason, we apply general imputation methods that do not require any previous knowledge regarding the underlying LD structure of the studied genome [13]. Simulation of these imputation methods in genetic selection (GS) studies, resulted in predicted genotypes that increased the efficiency of GS regardless the rate of missing values [13].

1.5.1 Model Selection

Cross-validation, is a model validation technique for assessing the results of a classifier. It is commonly used, when we want to estimate how precisely a predictive model will perform in previously 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 we 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 give an insight on how the model will generalize to an independent dataset. Common types of cross-validation are:

- **Leave-p-out cross-validation (LpO CV)** : Leave-p-out cross-validation involves using p observations as the validation set and the remaining observations as the

training set. This is repeated until all data have been participated in the test and in the train dataset.

- **Leave-one-out cross-validation (LOOCV)**: Leave-one-out cross-validation is a particular case of leave-p-out cross-validation with p = 1.

A better way of using the holdout method for model selection is to separate the data into three parts: a training set, a validation set, and a test set. The training set is used to the different models, and the performance on the validation set is then used for the model selection. The advantage of having a test set that the model hasn't seen before during the training and model selection steps is that we can obtain a less biased estimate of its ability to generalize to new data. Figure 1.3 illustrates the concept of holdout cross-validation where we use a validation set to repeatedly evaluate the performance of the model after training with different parameter values. Once we are satisfied with the tuning of parameter values, we estimate the models' generalization error on the test dataset [14]. This method is called nested cross-validation. In figure 1.4 we have an outer k-fold cross-validation loop to split the data into training and test folds, and an inner loop is used to select the model using k-fold cross-validation on the training fold [14].

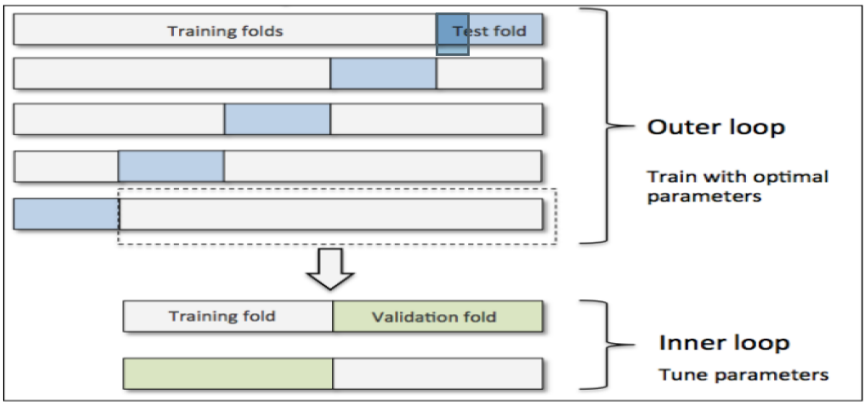


Figure 1.3: Nested cross-validation with five outer and two inner folds, which can be useful for large data sets where computational performance is important; this particular type of nested cross-validation is also known as 5x2 cross-validation [14]

According to some research [15], there is evidence in the machine learning, regarding whether N-fold cross-validation has better performance than LOOCV and vice-versa for binary classification .

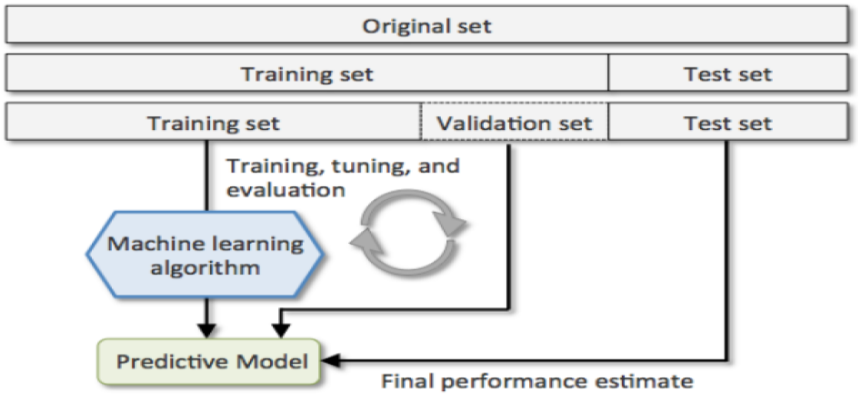


Figure 1.4: Nested cross validation method. First hide a test set. Then split the remain data to train and validation data in order to tune the hyperparameters and to choose the best model for our data [14]. In machine learning, a hyperparameter is a parameter whose value is set before the learning process begins.

1.5.2 Stratification

A better approach over the standard k-fold cross-validation is stratified k-fold cross-validation, which can yield better bias and variance estimates, especially in cases of unequal class proportions [14]. One common issue in data is the limited (inadequate) size of the data set. When this is the case, testing the model becomes an issue. Usually, 2/3 of the data are used for training and validation and 1/3 for final testing. Folding the dataset by chance, could lead to no representative subsamples of the initial (or complete) data set. For example in a data set of 100 samples and 5 classes, it is likely that one of these 5 classes may not be presented in the validation or test set. To avoid this problem, we should take care of the fact that each class should be correctly represented in both the training and testing sets. This process is called stratification. This process guarantee a correct class distribution among the training and validation sets. So we can select the data in every fold based on the probability of a class.

1.5.3 Classification

Support Vector Machines (SVM) is used in order to infer the missing data of a feature. SVMs map the data to a higher dimensional space via a kernel function and then identify the maximum-margin hyperplane in order to separate training instances [16]. The margin is defined as the distance between the separating hyperplane (decision boundary) and the training samples that are closest to this hyperplane, which are the so-called support vectors [17]. In SVMs, our optimization objective is to maximize the margin. More specific, the hyperplane is based on a set of boundary training instances, called support vectors. New samples are classified based on the side of the hyperplane they fall into. The optimization problem is most often formulated in a way that allows for non-separable data by penalizing misclassification [16]. Also, SVMs seems to be

insensitive in dimensionality and handle very large- scale classification in both sample and variables. SVMs in the beginning could only be applied to binary classification problems, but in the years, SVM were created that allowed classification of binary and multi-category data. We used one vs rest mode in order to handle multi-class and simple SVM for binary classification.

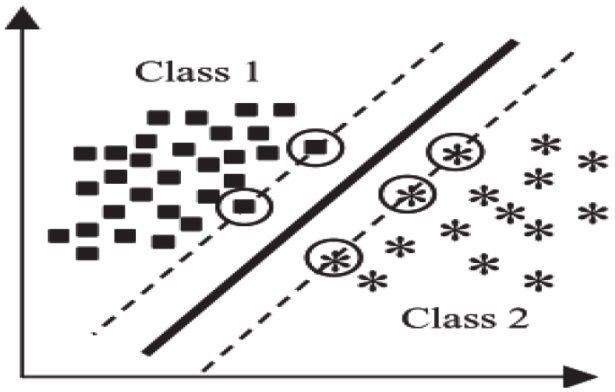


Figure 1.5: A binary SVM selects a hyperplane (bold line) that maximizes the width of the 'gap' (margin) between the two classes. The hyperplane is specified by 'boundary' training instances, called support vectors shown with circles. New cases are classified according to the side of the hyperplane they fall into [16].

1.6 Linear Mixed Models

In general, a mixed model is a statistical model containing both fixed effects and random effects. These models are proper in a wide variety of disciplines in the biological and physical sciences. They are appropriate in settings where measurements are made on clusters (Family) of related statistical units, or where repeated measurements are made on the same statistical units (longitudinal study) [18].
We used the R package `lme4` [18]. We used the command `lmer` to fit the mixed model for every phenotype. Random effects were fitted for every family to control for the correlation within the families. In mathematical notation the linear mixed model is written as

$$y_i = a + \tau_i + \sum_{j=1}^{p} \beta_j X_j + e_i$$

where i=1,...,K, with K denoting the number of families, y_i is the vector of measurements of the i-th family containing n_i measuments with $\sum_{(i=1)}^{K} n_i = n$, the overall sample size. The term a, 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} $N(0,\sigma_e^2)$ and τ_i $N(0,\sigma_\tau^2)$ respectively.
Examining residuals is a major part of all statistical modeling. Residuals can give us a hint whether our assumptions are reasonable and our choice of model is appropriate. 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.
The Bayesian information criterion (BIC), was used in order to compare two linear mixed models. 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. Both BIC and Akaike information criterion (AIC) attempt to resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

1.7 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 [19]. 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 [20]. Features in biology (e.g. SNPs, gene expressions) are commonly found to be expensive to measure, store and process [21]. 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.

1.7.1 The statistically equivalent signature (SES) algorithm

Commonly FS algorithms aim to find a single group of features, which has the highest predictive power. On the contrary, SES algorithm introduced in [22], 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 [23]. 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 [24]. Here, we used an adaptation of the SES algorithm that accommodates repeated measurements [25].

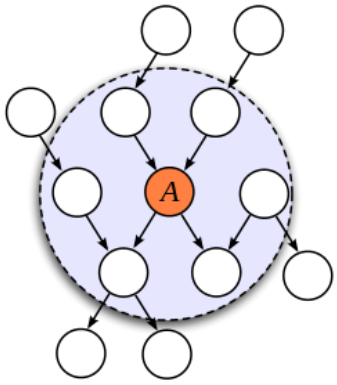


Figure 1.6: In a Bayesian network, the Markov blanket of node A includes its parents, children and the other parents of all of its children.

$$\Pr(A \mid \partial A, B) = \Pr(A \mid \partial A) \tag{1.1}$$

SES algorithm is influenced by the principles of constraint-based learning of Bayesian networks [26]. Bayesian networks are directed acyclic graphs that represent the dependency relationships between variables in a dataset. An edge A $\rightarrow$ B in a bayesian graph, represents the conditional dependence of variable B from variable A. There is a theoretical connection between FS and the Bayesian (causal) network that describes best the data at hand [20]. 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 1.1). Thus, the Markov blanket of a node contains the only knowledge needed to predict the behavior of that node.

1.7.2 Orthogonal Matching Pursuit (OMP) algorithm

Orthogonal Matching Pursuit is an iterative algorithm. At each iteration, it selects the column-marker of the SNP data matrix, that have the greatest correlation with the current residuals [27]. 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 [27].

The goal of this thesis was to signature of genetic markers that set the ground for understanding growth and other traits of interest in Gilthead seabream, in order to maximize the aquaculture efficiency, by improving phenotypic traits.

2 Methodology

2.1 Sample collection

The fish used in this study was 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, fourteen families originating from thirteen males and eleven 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 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), the total length at harvest (cm) (750 DPH) and the width at harvest (cm) (750 DPH) were measured.

2.2 Library preparation & Sequencing

Individual DNA samples were extracted using a modified salt-based extraction protocol based on [28] and treated with RNase to remove residual RNA. Genomic DNA was eluted in 5 mmol/L Tris, pH 8.5 and stored in 4oC. 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 in [29], 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 37oC 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, 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 nM 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 hrs at 22oC by addition of a further 3 μlof 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 Purification Kit, Qiagen, UK), and eluted in 70 μl EB buffer (Qiagen, UK). The size-selection, was performed by agarose gel separation, keeping the fragments between 400bp to 700bp. 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-ml 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 library was eluted in 24 μl EB buffer. Lastly, the ddRAD library was sequenced at the Norwegian Sequencing Centre in one HiSeq 2500 lane (2x125 bp reads).

2.3 Raw read quality control and demultiplexing

We have used `FastQC` software to do the quality control check on the raw sequence data retrieved from Illumina sequencing [30]. Demultiplexing the raw data was the next step in order to recover the reads belonging to each individual. `Process radtags` program from `STACKS` v.1.46 software [8] was used for this process. 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 [8], and `-t` parameter set to 100 to truncate final reads length to 100 bp. In particular, this step is depicted below:

Listing 2.1: Process RadTags

```
1
2 # Code
3 stacks process_radtags -P \
4                 -1 $DATADIR/Sample_Saurata-ddRAD-GR/*_R1_001.fastq.gz \
5                 -2 $DATADIR/Sample_Saurata-ddRAD-GR/*_R2_001.fastq.gz \
6                 -b $DATADIR/barcodes \
7                 -o $OUTDIR/ \
8                 -c -q -r \
9                 --inline_inline \
```

```
                --renz_1 sbfI \
                --renz_2 sphI \
                -i gzfastq -t 100 \
                -D
# -r  Rescue barcodes and RAD-Tags.
# -c  Clean data, remove any read with an uncalled base.
# -q  Discard reads with low quality scores.
# -t  Truncate final read length to this value.
# -D  Capture discarded reads to a file.
```

2.4 Data alignment against seabream reference genome

The annotated reference genome of gilthead seabream has been provided by Hellenic Centre for Marine Research (H.C.M.R.) (Pauletto et al. in press). To align our samples to the reference genome, we used Bowtie2 v.2.3.0 [31] with the following parameters: `-end-to-end -sensitive -no-unal`. Then, we removed multi-aligned reads, reads with > 3 mismatches and reads with map quality lower than 20 with Samtools [32].

Listing 2.2: Align ddRAD

```
# ------------------------------------ #
# ============  BOWTIE2 ================ #
# ------------------------------------ #
 bowtie-build /genome.fasta  sparus_aurata
 bowtie2 -p 20 --end-to-end --sensitive --no-unal
                 -x sparus_aurata_bowtie \
                 -1 /data/seabream/RAL357_1.fastq \
                 -2 /data/seabream/RAL357_2.fastq \
                 -U remain.1.fq.gz,remain.2.fq.gz,
         -S result_bowtie.sam \

# --no-unal Suppress SAM records for reads that failed to align.

# ======================== FILTERING ============================ #
# Save header
os.system("samtools view -H {1}/{0}.sam > {1}/header.sam".format(name,SAMDIR))
# Remove multi aligned | Remove >mismatches| Create Bam file
os.system("samtools view -F 4 -q 20 {1}/{0}.sam | grep -v 'XS:'| grep 'XM:i:[0-3]' | cat {1}/h
# Sort Bam file
os.system("samtools sort {1}/{0}.bam -o {1}/{0}.sortred.bam".format(name,SAMDIR))
# Delete header
os.system("rm {1}/header.sam {1}/{0}.bam {1}/{0}.sam ".format(name,SAMDIR))

# -q INT : Skip alignments with MAPQ smaller than INT [0].
# ------------------------------------------------------------------------------------------
```

2.5 Stacks Pipeline

The pstacks program will extract stacks that have been aligned to a reference genome by an aligner such as BWA or Bowtie2. Pstacks compares "stacks of reads" and forms putative sets of loci. These sets are used in order to detect SNPs at each locus using a maximum likelihood framework [8]

Listing 2.3: Detect SNPs

```
# ================================ PSTACKS ================================== #

command = "stacks pstacks -p 12  -o {0}/3_Pstacks -m 3 ".format(WORKDIR)
lista = []
counter = 0
for file in os.listdir(SAMDIR):
        if file.endswith("bam"):
                file = file [:file.index(".",file.index(".")+1)]
                if file not in lista:
                        counter+=1
                        lista.append(file)
                        command += " -f {0}/{1}.bam -i {2}".format(SAMDIR,file,counter)
                        os.system(command)
# ============================================================================ #
```

A SNP catalogue was built from the parents of the cross. Cstacks created a set of all possible alleles expected in the progeny of the cross.

Listing 2.4: Catalog of SNPs

```
# ================================== CSTACKS ================================= #

command = "stacks cstacks -b 1 -p 12 -o {0} --aligned ".format(CStaDIR)
lista = []
for file in os.listdir(RMAPDIR+"/Pstacks/Parents/"):
        if file.startswith("Br"):
                file = file [:file.index(".",file.index(".")+1)]
                if file not in lista:
                        lista.append(file)
                        command += " -s {1}/{0} ".format(file,PStaDIR)
command += " &>> {0}/4_Cstack_Log ".format(WORKDIR)
os.system(command )
# ============================================================================ #
```

The sets of stacks that constructed by the pstacks program searched against the catalog produced by cstacks. All samples in the population matched against the catalog with sstacks.

Listing 2.5: Matching against Catalog

```
# ================================= SSTACKS ================================== #
sstacks_com = "stacks sstacks -g -p 12 -b 1 -c {0}/batch_1 -o {1}/".format(CStaDIR,SStaDIR)

```

```
lista = []
for file in os.listdir(PStaDIR):
        if file.endswith("gz") :
                file = file [:file.index(".",file.index(".")+1)]
                if file not in lista:
                lista.append(file)
                file_com = sstacks_com + " -s {1}/{0}".format(file,PStaDIR)
                        os.system(file_com)
# ============================================================================= #
```

The `populations` program uses the population map to determine which groupings to use for calculating summary statistics, such as heterozygosity.

Listing 2.6: Produce VCF file

```
# ================================== Population ================================== #

Main       = "stacks populations "
Files      = " -P {0} -O {0}/Results -M {0}/popmap1 -b 1 -k ".format(POPStaDIR)
Params     = " -f p_value -t 12 --structure --vcf --vcf_haplotypes --plink "

Command = Main+Files+Params
os.system(Command)
============================================================================== #
```

2.6 Kinship

To check family relationship and indicate possible pedigree errors we used KING v.2.1 software [33]. 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 [33]. 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.

Listing 2.7: Kinship

```
# ======================== PLINK ========================== #
# PLINK v1.90p 64-bit (14 Nov 2017)             www.cog-genomics.org/plink/1.9/
# --allow-extra-chr : allow scaffolds

plink --vcf batch_1.vcf --allow-extra-chr --make-bed --out plink
plink --vcf batch_1.vcf --allow-extra-chr --recode oxford

# ======================== KING ========================== #
# (C) 2005-2017 Shaun Purcell, Christopher Chang   GNU General Public License v3
#KING 2.1 - (c) 2010-2018 Wei-Min Chen

king -b plink.bed --cluster --degree 10
```

```
king -b plink.bed --kinship --degree 10
================================================================================ #
```

2.7 Imputation

There are different methods in order to impute missing data. Shapeit and Impute2 are commonly used for genotype phasing and imputation respectively when a reference genome is available. However, the presence of many scaffolds and the lack of a phased reference panel discouraged us from using these software packages. For the imputation of missing values in our study, we used a Support Vector Machines (SVM) classifier. We also applied a method called stratified nested cross validation that initially produces unbiased estimates of the optimum model parameters. Once the model has been tuned, it then estimates the efficiency of the classifier. For efficiency measurement we used the AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic) curve. We measured the AUC of two different types of ROCs: The first, called micro-average ROC curve, is constructed by measuring the sensitivity and specificity of each sample. The second, called macro-average ROC curve, is constructed by averaging the sensitivity and specificity in all samples for each class. Both metrics are required for an unbiased estimate of the efficiency since micro-average is biased towards classes with relatively greater number of samples whereas the second is biased towards samples with lower number of samples. We applied a confidence threshold of 0.9 for both metrics. Namely, we left as missing the genotypes that yielded an AUC metric lower than 0.9 in any of the two metrics. We used the `scikit-learn` python library [34] to implement the above and the complete scripts are available upon request.

Listing 2.8: Imputation

```
# ================ Model Selection ================= #
## PIPELINE ##
from sklearn.pipeline import Pipeline
from sklearn.cross_validation import StratifiedKFold
from sklearn.grid_search import GridSearchCV
from sklearn.metrics import accuracy_score

#=== Stratified  ===#
cv = StratifiedKFold(Known_Labels, n_folds=Min_Folds,random_state=1)
#=== CLASSIFIER ===#
clf = SVC()

param_grid = [{'clf__kernel': ['linear','poly'],
  'clf__C': [0.01, 0.1, 10, 100],
  'clf__probability':[True],
  'clf__decision_function_shape':['ovr'],
  'clf__degree' :[2,3]}]

## Create Pipeline ##
```

```
pipe = Pipeline([('clf', clf)])
gcv = GridSearchCV(estimator=pipe, param_grid=param_grid, scoring='roc_auc', cv=Min_Folds)
#========================================================== #
```

2.8 Linear mixed models

For the association analysis we have extracted only the variants that did not have missing data. Then, we found the significant SNPs, using linear mixed models (lme4 package), taking into account the Family id of each individual [18]. This resulted in a significance value (p-value) for each SNP. We used the GWAStools package for visualization. This packages produces qqplots which demonstrate p-value inflation and manhattan plots which show p-value across the genome.

Listing 2.9: Statistics

```
# ==================== Mixed Models ====================== #
# y = target phenotype (Fat,Weight etc.)
# IDS = Family id
# All Features = Matrix with genotypes. Rows = samples, Columns = SNPs)

library(lme4)
m1 = lmer(Fat~All_Features[,i]+(1|IDS),REML=FALSE)
m0 <- update(m1, .~. - All_Features[,i])
pvals_Fat[i]=(anova(m1,m0)$`Pr(>Chisq)`[2])

library(GWASTools)
qqPlot(pvals_Fat)
manhattanPlot(pvals_Fat,chroms,signif=Threshold,ylim = c(0,4))
#========================================================== #
```

2.9 Feature Selection

Two algorithms were used for feature selection OMP and SES. In the *Orthogonal Matching Pursuit Algorithm* (OMP) the stopping rule was set at two or four units difference in BIC score between the old and the new linear mixed model. In SES we tested different parameters. The maximum conditioning set was set to 2,3,4,5, the threshold equals to 0.01 or 0.05 and "testIndLMM" used as the conditional independence test.

2.9.1 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 [35]. 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: 2.1). The model with the lowest mean sum of errors is selected as best model (Equation: 2.2).

$$ErrOB = \sum_{i=1}^{m} E(y_{i(n_i+1)} - x_{i(n_i+1)}^T \hat{\beta} - z_{i(n_i+1)}^T \hat{b}_i)^2 / m, \tag{2.1}$$

where $y_{i(ni+1)}$, $x_{i(ni+1)}$ and $z_{i(ni+1)}$ 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,

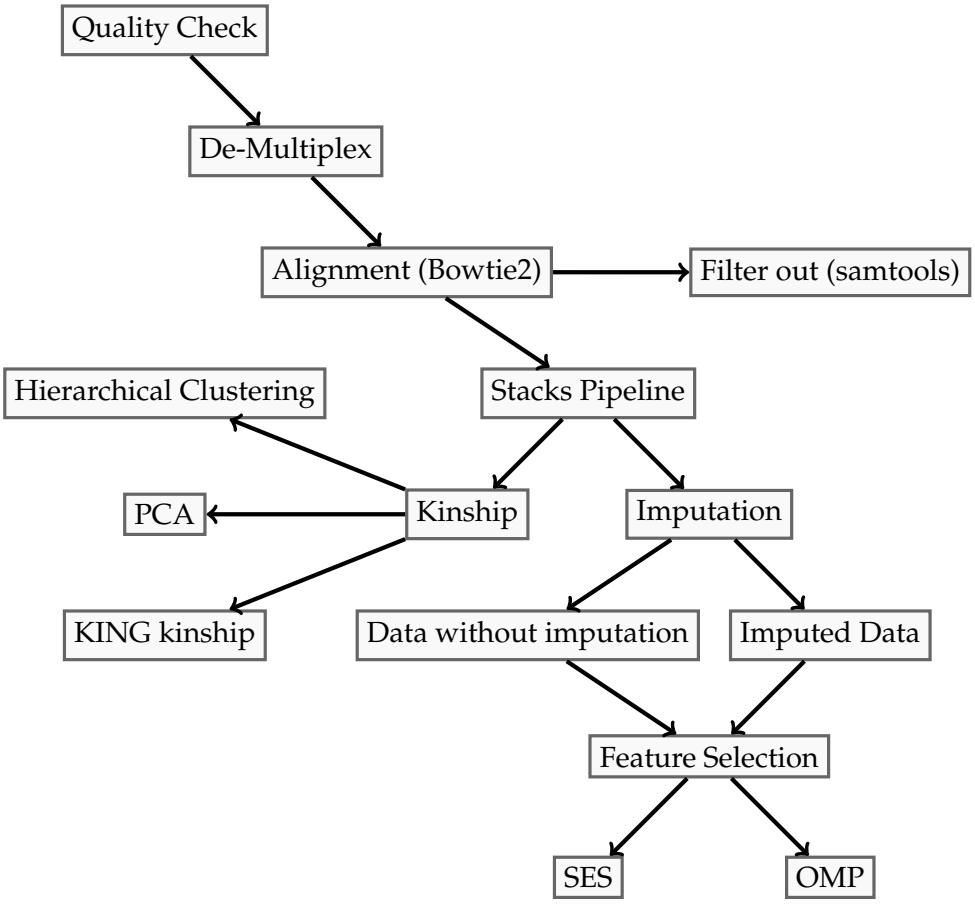
$$LOOCV = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - x_{ij}^T \hat{\beta}^{[i,j]} - z_{ij}^T \hat{b}_i^{[i,j]})^2 / N, \tag{2.2}$$

where $\hat{\beta}^{[i,j]}$ and $\hat{b}_i^{[i,j]}$ are respectively the estimates of β and b_i based on the training data without subject j in cluster i [36].

2.10 Selected SNPs 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 seabream (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.

In a nutshell, the pipeline that was followed is:



3 Results

3.1 Quality Check

The results of the quality control using FASTQC, are illustrated side by side with those from the company, which provided us the data (Figure 3.1). As we can see the results are the same, and the slight difference occurred due to the different number of bins that used.

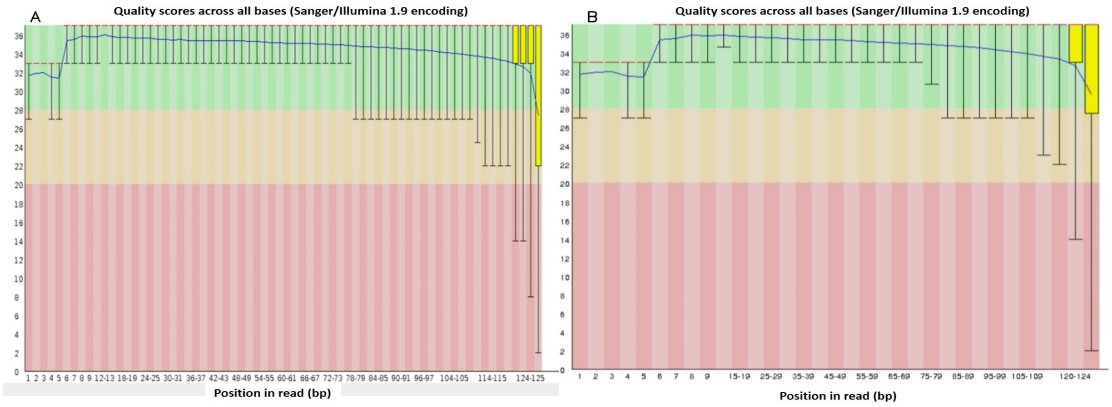


Figure 3.1: **(A)** Quality Check of our analysis. **(B)** Company's quality check.

3.2 Genotyping RAD alleles

Illumina sequencing yielded 559,191,588 raw reads. Following quality control, we filtered out $\sim$ 15.2% due to ambiguous barcodes, $\sim$ 2.9% due to low quality and 1% due to the lack of restriction sites. The rest were successfully assigned to individuals (Suppl. 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 ($\sim$ 8%) and those with more than 3 mismatches ($\sim$ 2.96%), keeping finally 351,781,485 reads for analysis. The ddRAD catalogue built from all parental samples consisted of 15,233 SNPs. 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.

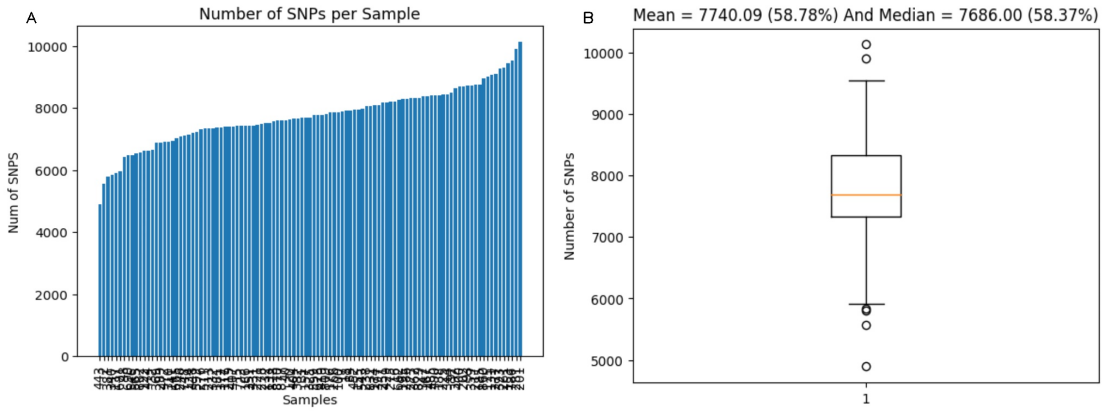


Figure 3.2: **(A)** Number of SNPs distribution in our cohort. **(B)** Boxplot of number of SNPs from our cohort.

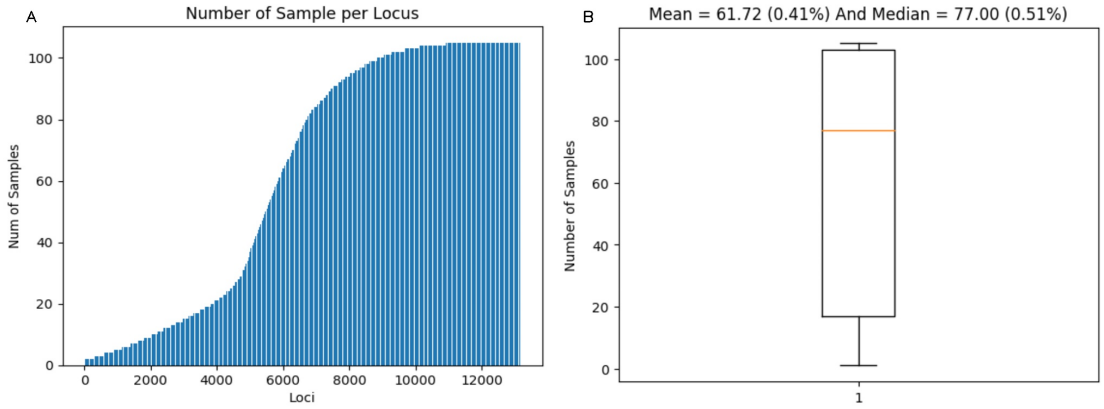


Figure 3.3: **(A)** Number of Samples distribution in Stacks variants. **(B)** Boxplot of number of Samples.

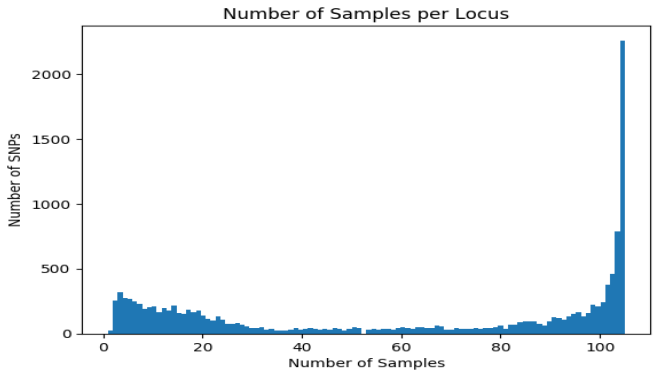


Figure 3.4: The trend of number of genotyped samples per number of SNPs

3.3 Kinship

To verify the family identity of the studied individual, we used three different methods: King kinship, Principal Component Analysis (PCA) and Hierarchical clustering (Figure: 3.5). 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.

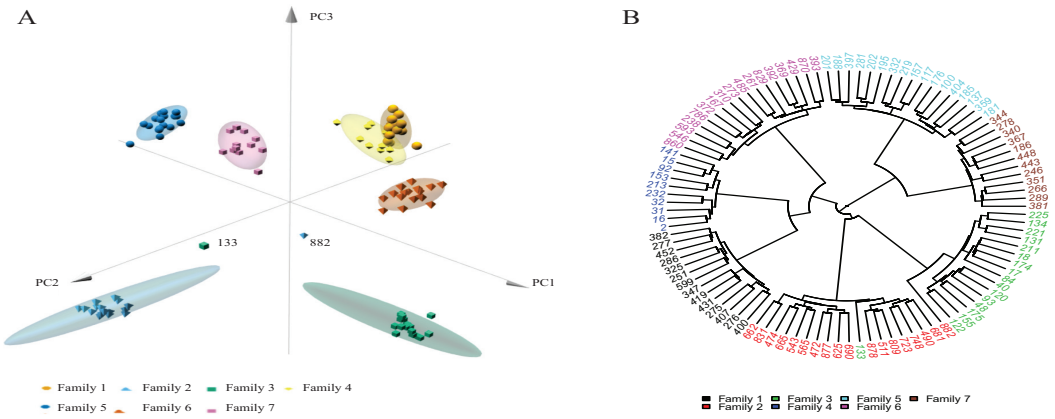


Figure 3.5: (**A**) Principal component analysis of the 105 individual progeny according to the polymorphisms of each individual, for different families.The explained variation is 30% (11% PC1, 9% PC2, 8.9%PC3) (**B**) Cluster dendrogram of the 105 individuals based on the Euclidean Distance of the genotypes. The Sample-ids are colored based on the tagging family id.

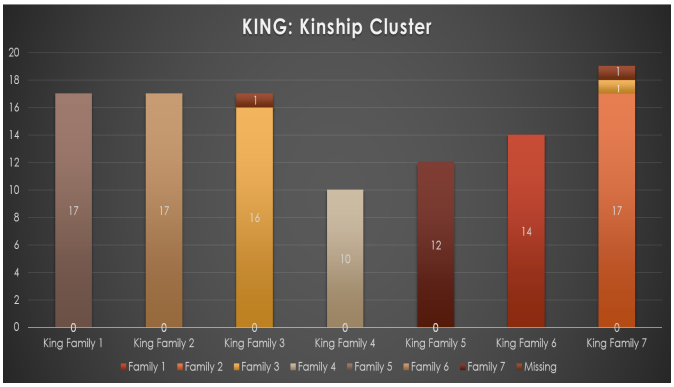


Figure 3.6: KING kinship

3.4 Imputation

Imputation of missing genotypes was implemented with classification algorithms as we mentioned before. For each one of the 3,028 SNPs with at least one missing value, we attempted to impute its missing genotypes. From these, 1,355 SNPs were imputed with a confidence higher than 0.9 in both micro-ROC and macro-ROC AUC scores (see Methods) and were kept for further analysis. Some of the results are illustrated in Figure: 3.7. The rest 1,673 SNPs were filtered out. We refer to the dataset that consists of 1,355 SNPs with imputed missing values and the 2,258 SNPs without any missing value as the "dataset with imputed values".

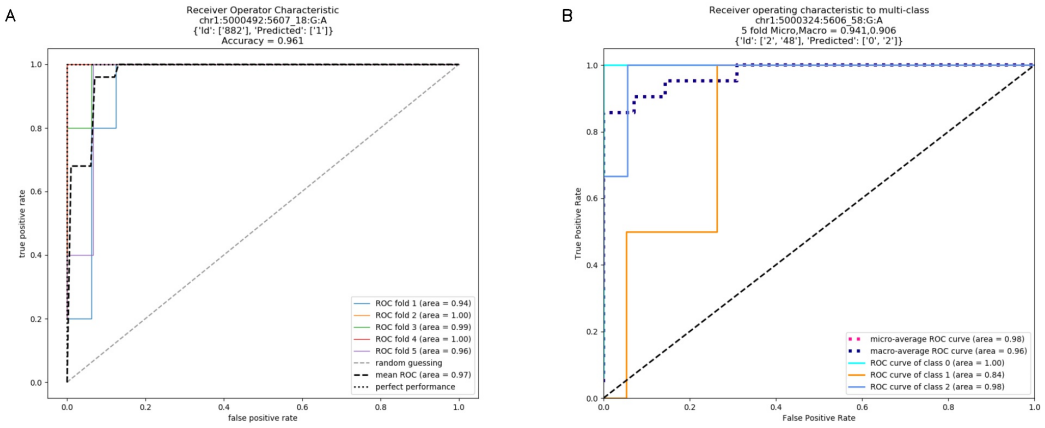


Figure 3.7: Receiver operating characteristic (ROC) curve showing the accuracy of the multi-class or binary-class SVM in predicting specific Variant. The true positive rate (sensitivity) is plotted in function of the false positive rate (100-specificity). The area under the ROC curve is a measure of how well the model distinguishes successful from unsuccessful Variant class. **(A)**ROC curve of binary SVM model predicting specific variant in 5 Folds of cross-validation. **(B)** ROC curve of multiclass SVM model predicting specific variant using one vs rest method.]

3.5 Association Analysis

We proceeded in two different analyses. In the first we used the dataset that did not have missing values (2,258 SNPs). In the second pipeline we used the dataset with imputed values.

3.5.1 Dataset without imputed values

Association analysis through GWAS

The results from the GWAS test among all SNPs and the four phenotypes are shown in Table:3.1. In total, we found five SNPs associated with Weight, four SNPs with Tag Weight, and none for Fat and Length/Width. On Figure 3.8 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: .1. 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: 3.1). 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, 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).

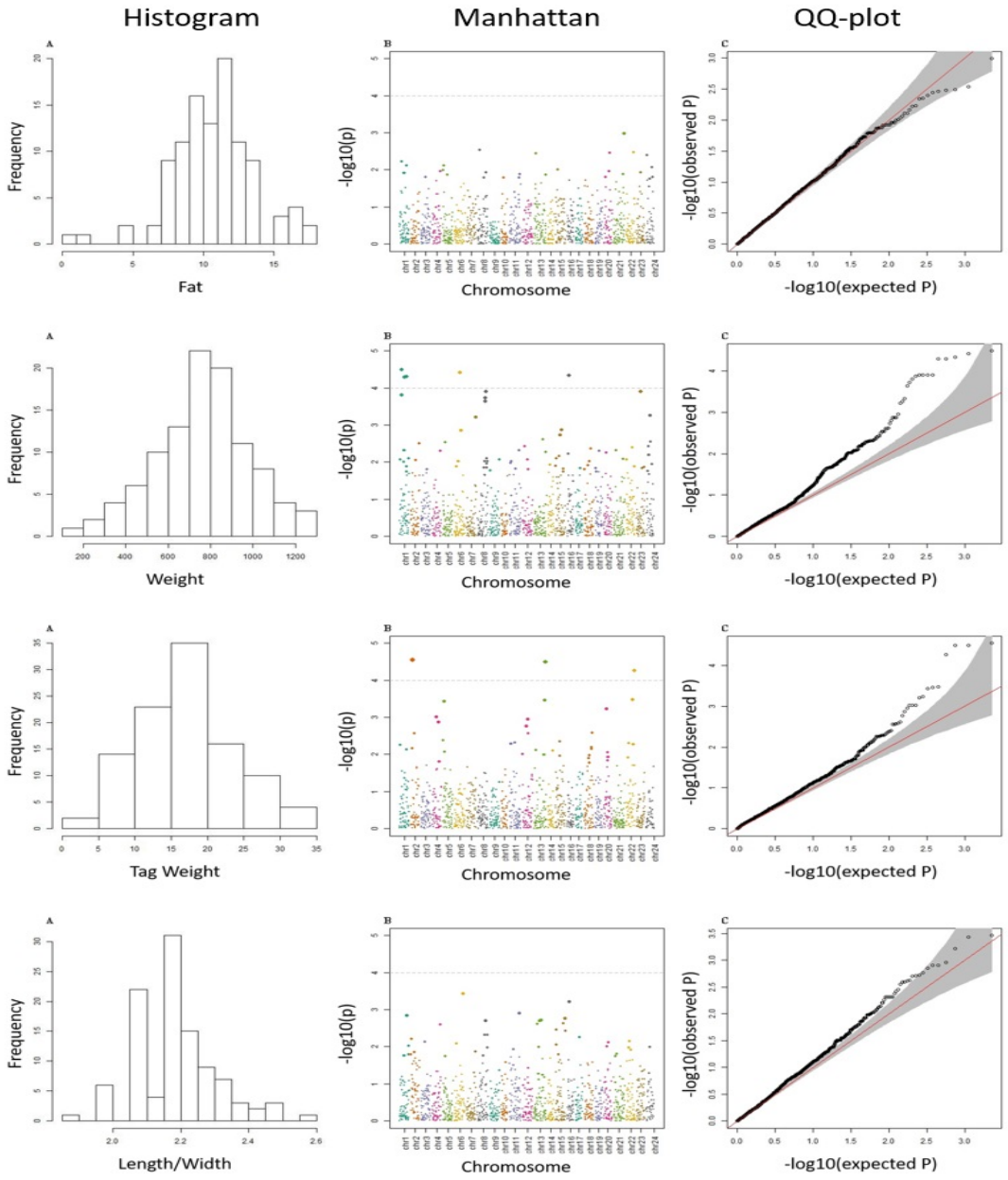


Figure 3.8: (**A**) Distribution of each examined trait in our samples. (**B**) Manhattan plot demonstrating the locations across the chromosomes of the seabream genome (horizontal axis) versus the $-log(p - values)$ of the association between the genetic variants and phenotype (vertical axis). The higher the dots, the stronger the genetic association. The significance threshold was set to 10^{-4}, in order to correct for multiple testing (dashed line). (**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.

Table 3.1: Selected SNPs from GWAs analysis using linear mixed models, with significance threshold equal to 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

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: 3.9]. 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: 3.2-3.5. SES produced different combination of SNPs (signatures) that have the same predictive strength on each one of the examined traits. In Tables 3.2-3.5 we illustrate one of these combinations, while the rest are illustrated in Supplementary Tables .2-.5. Finally, the effects all selected SES SNPs from all traits are presented in Figures: 3.10 - 3.13.

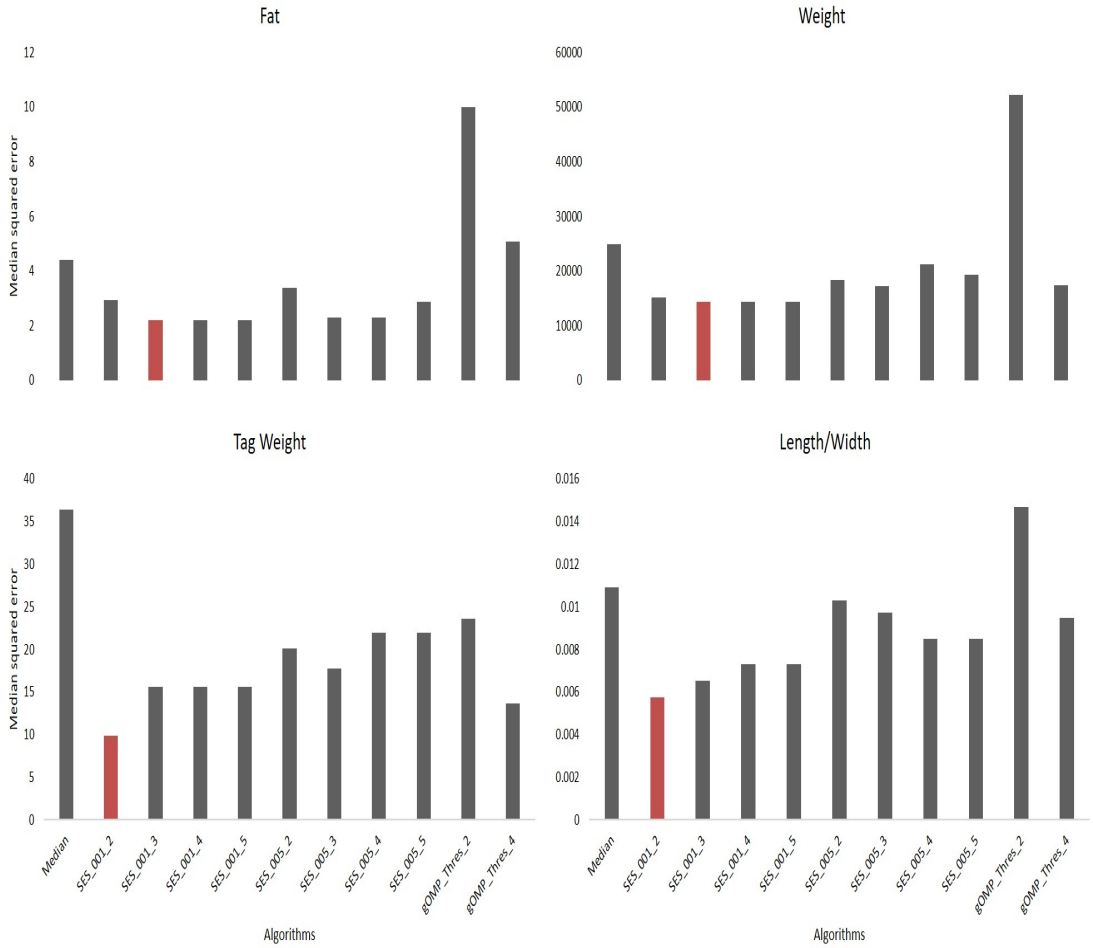


Figure 3.9: Comparison of different algorithms predicting Tag Weight, based on median squared error, after leave one out cross validation. SES algorithm tested for different thresholds (significance threshold equal to 0.01 or 0.05) and for different numbers of SNPs as condition set (k = 2,3,4,5). OMP algorithm tested for different thresholds as stop criterion (Threshold = 2 or 4 units in BIC score).

Listing 3.1: Error Calculation

```
# ====================== Error Calculation ============================== #
Error_fun <- function(fam_num,Train_y,Train_X,Train_id,
Test_y, Test_X, Test_id,ses_thres,ses_k){
  sel <- SES.temporal(target = Train_y, dataset = Train_X,
        group = Train_id, test= "testIndLMM",max_k = ses_k,
        threshold = ses_thres)@selectedVars
  mod1 <- lmer(Train_y ~ Train_X[ , sel ] + (1|Train_id), REML = FALSE )
  b1 <- as.matrix( coef(mod1)$Train_id )
  # Sum of Errors of Different Families
  Error <- sum( ( Test_y - c(1, Test_X[sel]) %*% t(b1) )^2 ) / m
  return(Error)}
# ------------------------------------------------------------------------------
```

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: 3.2). 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: 3.2).

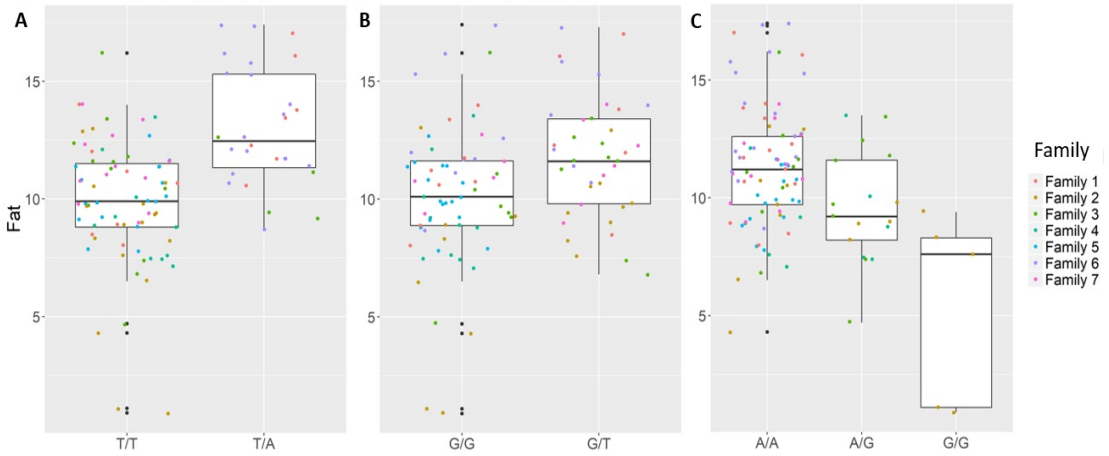


Figure 3.10: The effect of each of the selected SES SNPs associated with fat content. (A-C) Boxplots of selected SNPs. A) chr8:1385781 , B) chr13:1098152 , C) chr21:19924408.

Table 3.2: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Fat						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr13:1098152	rho-related GTP-binding-like	0.007	0.01	-	-	3′ UTR
Chr21:19924408	-	0.006	0.01	-	-	-
Chr8: 1385781	Protection of telomeres 1	0.0024	0.01	-	✓	Intron
Scaffol8147:18634	death-associated kinase 3-like	0.015	0.05	-	✓	Intron
Chr7:2453106	solute carrier family 41 member 1-like isoform X1-2	0.046	0.05	-	-	Intron
Chr4:23265532	NT-3 growth factor receptor isoform X1	0.017	0.05	-	-	Upstream

Selected SNPs for weight at harvest

Four selected variables associated with weight at harvest (800g average weight at harvest), have been retrieved from SES algorithm 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. 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: 3.3).

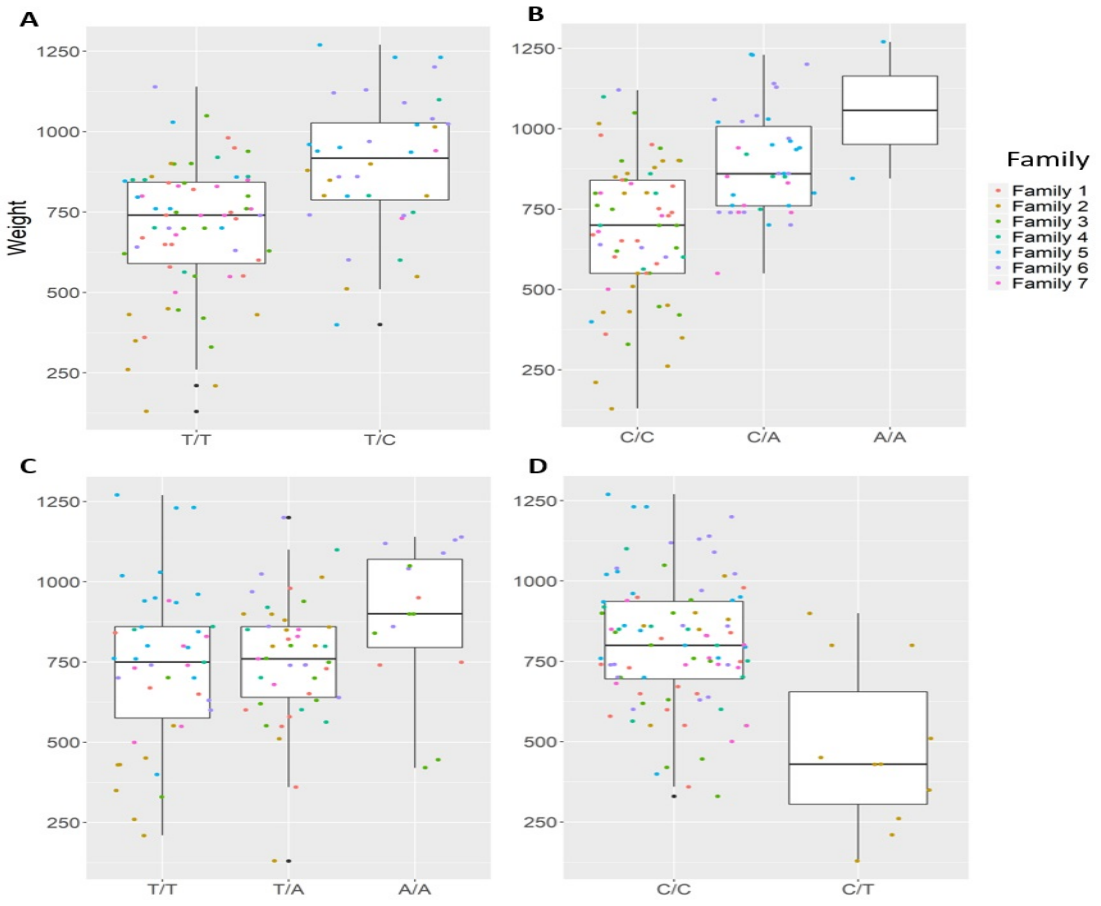


Figure 3.11: The effect of each of the selected SES SNPs associated with weight at harvest. (**A-D**) Boxplots of selected SNPs. A) chr1:16636968 , B) chr6:12617755 , C) chr8:11613979, D) chr16:2232897.

Table 3.3: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

		Weight				
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr1:16636968	ethanolamine-phosphate cytidylyltransferase-like	0.0006	0.01	✓	✓	3′ UTR
Chr6:12617755	myosin-7-like isoform X1	0.0024	0.01	✓	✓	Upstream
Chr:8:11613979	semaphorin-3A	0.0114	0.01	-	✓	Intron
Chr16:2232897	acetylserotonin O-methyltransferase-like	0.0022	0.01	✓	-	3′ UTR
Scaffold29:195838	mitogen-activated kinase-binding 1-like	0.0285	0.05	-	-	Intron
Chr24:8282385	STE20-related kinase adapter beta trafficking kinesin-binding 2 isoform X1	0.0022	0.05	-	✓	Downstream Upstream

Selected SNPs for weight at tagging

Five SNPs were associated with Tag Weight, as retrieved from SES algorithm (Table: 3.4). The first was found at 'RNA-binding 27 isoform X1' gene (chr13:20975921), the second upstream 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: 3.4).

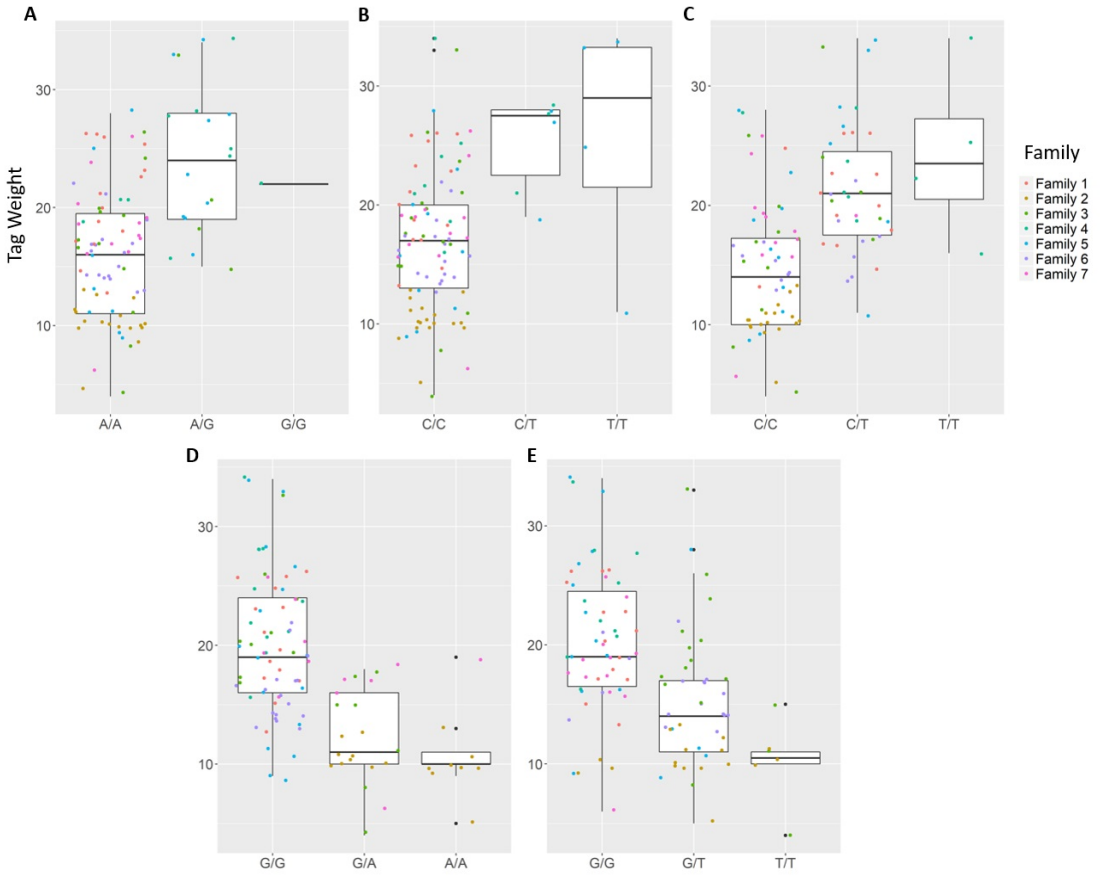


Figure 3.12: The effect of each of the selected SES SNPs associated with tag weight. (A-E) Boxplots of selected SNPs. A) chr2:2623351 , B) chr13:20883924 , C) chr13:20975921, D) chr22:18343985, E) scaffold4139:36071.

Table 3.4: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

		Tag Weight				
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr2:2623351	Tetratricopeptide repeat 36	0.0019	0.01	✓	-	Upstream
Chr13:20883924	DNA repair RAD50	0.0127	0.01	-	✓	Intron
Chr13:20975921	RNA-binding 27 isoform X1	0.0073	0.01	✓	-	Intron
Chr22:18343985	zinc finger BED domain-containing 4-like midasin isoform X2	0.0117	0.01	✓	-	Upstream Downstream
Scaffold4139:36071	predicted uncharacterized protein LOC106518831	0.033	0.01	-	✓	Upstream
Chr15:3260819	follistatin-related 1-like	0.0124	0.05	-	-	Downstream
Chr20:6671436	UBA-like domain-containing 1	0.021	0.05	-	✓	2nd
Chr22:14483563	exostosin-1-like	0.0448	0.05	-	-	Intron
Scaffold14083:12192	-	0.042	0.05	-	-	-

Selected SNPs for length/width phenotype

Finally, five SNPs were associated with Length/Width ratio (at 750 DPH) as retrieved from SES algorithm (Table: 3.5). 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.

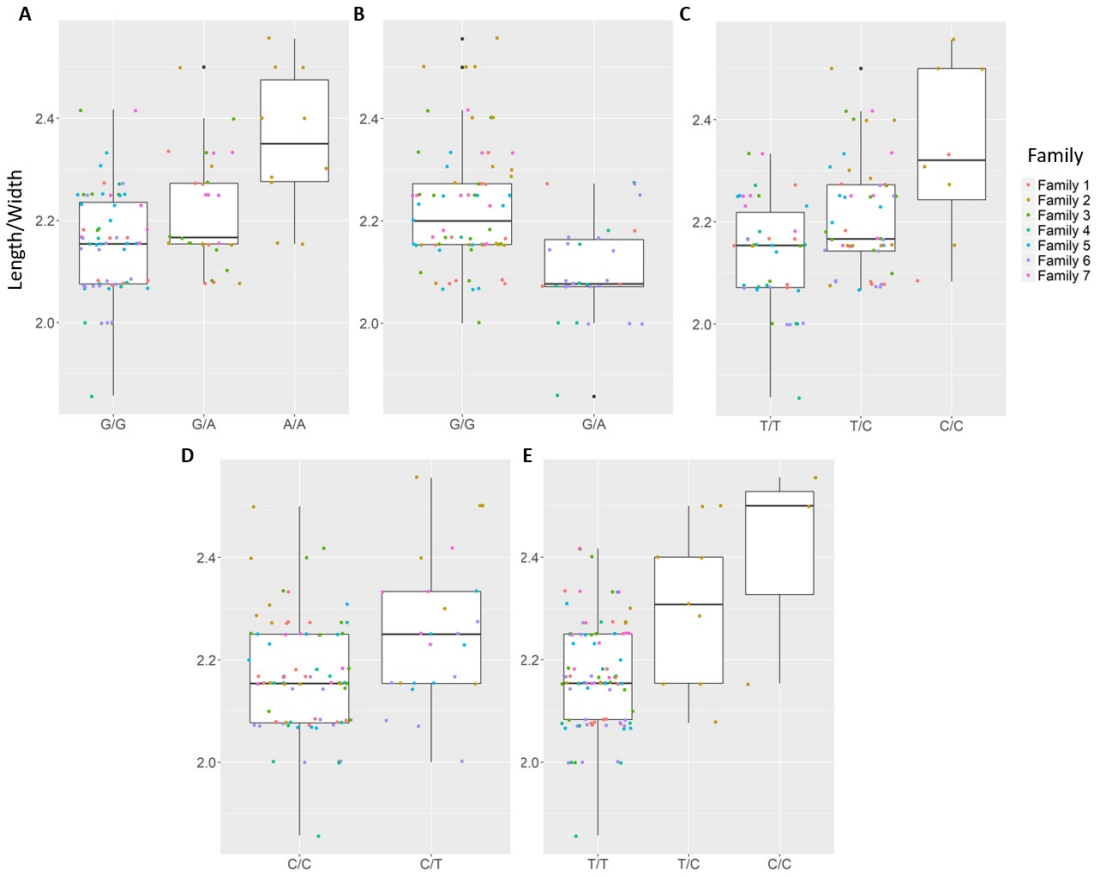


Figure 3.13: The effect of each of the selected SES SNPs associated with length/width. (A-E) Boxplots of selected SNPs. A) chr1:20827142, B) chr3:9671223, C) chr6:23799286, D) chr13:9665394, E) scaffold5661:35982.

Table 3.5: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Length / Width						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr6:23799286	phosphatase 1 regulatory subunit 3D-like	0.0052	0.01	-	✓	3d
Chr1:20827142	Upstream: mucin-5AC-like isoform X1	0.049	0.01	-	✓	Upstream
Chr13:9665394	ATP-dependent RNA helicase DHX33	0.0211	0.01	-	✓	3d
Chr3:9671223	A-kinase anchor 9 isoform X3	0.0144	0.01	-	✓	2nd
Scaffold13177:8369	phosphatase 1 regulatory subunit 3C	0.015	0.01	-	✓	Downstream
Chr8:11613979	semaphorin-3A	0.0193	0.05	-	✓	Intron
Chr22:2545133	neurexin-3b isoform X3	0.049	0.05	-	-	Intron
Scaffold5661:35982	-	0.049	0.05	-	✓	-

Dimitrios Kyriakis

3.5.2 Dataset with imputed values

Association analysis through GWAS

The results from the GWAS test among all SNPs and the four phenotypes are shown in Table: 3.6. In total, we found nine SNPs associated with Weight, five SNPs with Tag Weight, one with length/width. GWAS analysis in fat content did not reveal any associated SNP.
Four more SNPs associated with weight at harvest, have been retrieved from the typical GWAS analysis compare with the previous analysis without imputed data. The first was found in chromosome 1 (chr8:6970078) on 'bile acid receptor' gene, the second (chr8:20827142) in a conserved region upstream of 'interleukin-34' gene. The third (chr8:2887261) in a conserved region upstream of 'interferon regulatory factor 7' gene. The last (Scaffold15653:2041) located downstream of 'adipocyte plasma membrane-associated'. One more SNP found to be associated with weight at tagging, in a conserved region downstream of 'Down syndrome cell adhension molecule isoform X2'. Finally, one SNP found to be associated with Length/Width in a conserved region upstream of 'interferon regulatory factor 7'.

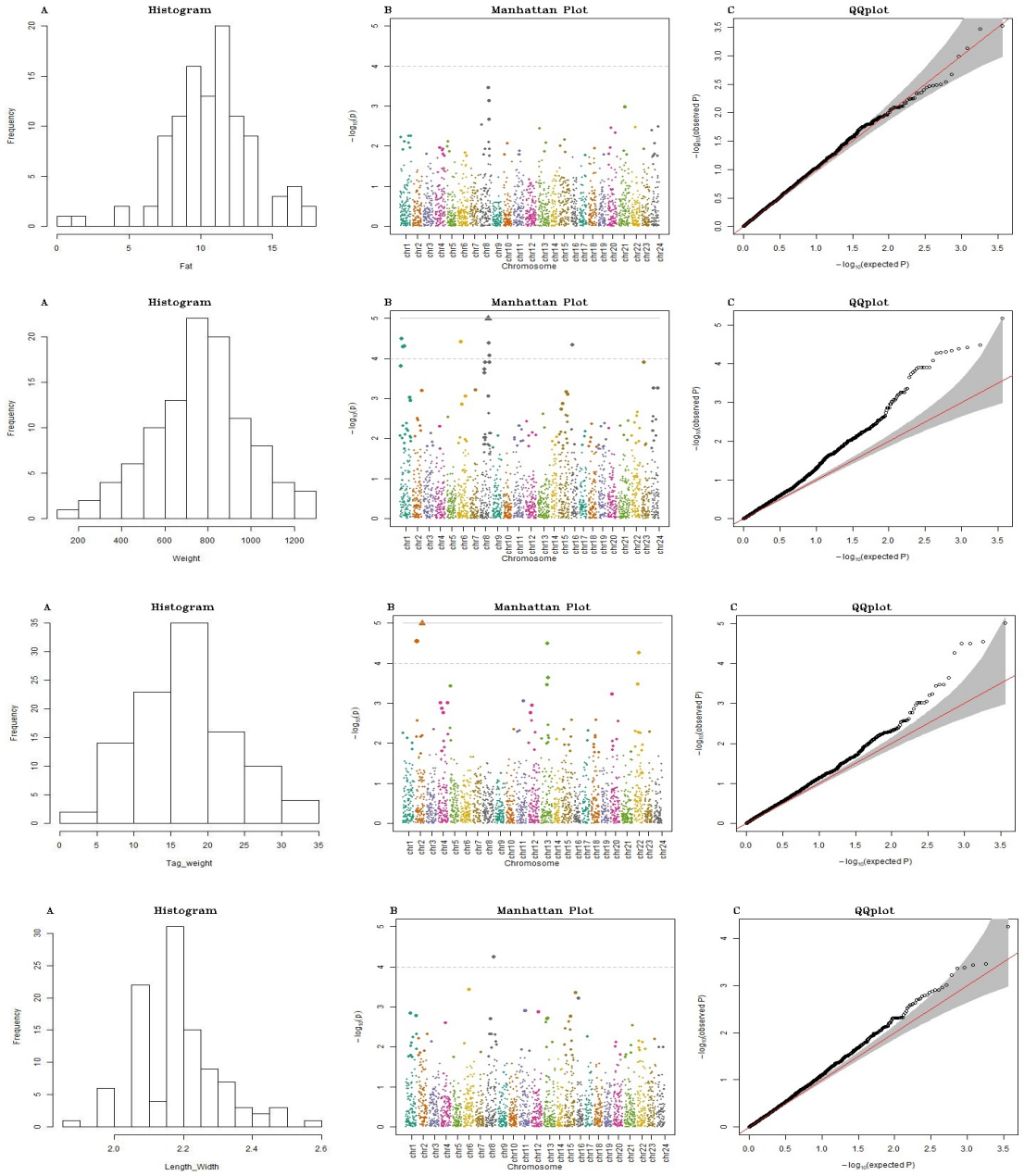


Figure 3.14: (**A**) Distribution of each examined trait in our samples. (**B**) Manhattan plot demonstrating the locations across the chromosomes of the seabream genome (horizontal axis) versus the $-log(p-values)$ of the association between the genetic variants and phenotype (vertical axis). The higher the dots, the stronger the genetic association. The significance threshold was set to 10^{-4}, in order to correct for multiple testing (dashed line). (**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.

Table 3.6: Selected SNPs from GWAS analysis using linear mixed models, with significance threshold equal to to 10^{-4}.

Position	Gene	P-Value	Beta coefficient	Conserved	Position
Weight					
chr1.6970078	lymphoid enhancer-binding factor 1 isoform X1	3.265e-05	174.7210	-	Downstream
Chr1:16636968	ethanolamine-phosphate cytidylyltransferase-like	5.059e-05	189.5556	✓	
Chr1:20827142	mucin-5AC-like isoform X1	4.976e-05	-161.8345	✓	Upstream
Chr16:2232897	acetylserotonin O-methyltransferase-like	4.648e-05	-338.1485	✓	
Chr6:12617755	myosin-7 like	3.838e-05	205.2103	✓	Downstream
Chr8:2887261	interferon regulatory factor 7	6.785e-06	-200.7313	✓	Upstream
Chr8:7819520	bile acid receptor isoform X1	4.105e-05	199.7217	-	
Chr8:9391467	interleukin-34	8.286e-05	180.6165	✓	
Scaffold15653:2041	adipocyte plasma membrane-associated	5.236e-05	-264.4869	-	Downstream
Tag Weight					
Chr13:20975921	RNA-binding 27 isoform X1	3.168e-05	4.747498	-	
Chr13:20975924	RNA-binding 27 isoform X1	3.168e-05	4.747498	-	
Chr2:2623351	Tetratricopeptide repeat 36	2.823e-05	6.182521	-	Upstream
Chr22:18343985	tectonin beta-propeller repeat-containing 2	5.405e-05	-5.139881	-	Upstream
Chr2:8737137	Down syndrome cell adhesion molecule isoform X2	9.843e-06	6.685408	✓	Upstream
Length/Width					
Chr8:2887261	interferon regulatory factor 7	5.664e-05	0.09594116	✓	Upstream

Association analysis through FS

In this part of the analysis, we could not perform the algorithm comparison due to the computational time, so we run SES algorithm with condition set equal to 3 and threshold equal to 0,01.

Selected SNPs for fat content (%)

The selected variables/SNPs associated with Fat content (%) at harvest, retrieved from SES algorithm (threshold 0.01), recovered five SNPs, out of which four were located within or proximal to an annotated gene (Table: 3.7). 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, is located upstream of 'interferon regulatory factor 7' (chr8:2887261). The third one was located in intron of 'serine threonine-kinase WNK1-like isoform X1' (Table 3.7).

Selected SNPs for weight at harvest

SES selected four SNPs to be associated with Weight. A polymorphism found in chromosome 1 on 'ethanolamine phosphate cytidylyltransferase-like' gene. Moreover, a SNP (chr6:12617755) in a conserved region upstream of 'myosin-7-like' gene was found. Also, a polymorphism found in chr8:2887261 located upstream of 'interferon regulatory factor 7'. This polymorphism has been selected as significant also in length/width phenotype. Another one (Scaffold15653:2041) was found downstream of the gene that transcribes 'Adipocyte plasma membrane-associated protein' (APMAP). (Table 3.8).

Selected SNPs for weight at tagging

Five SNPs associated with Tag Weight, were retrieved from SES, from which three was annotated. The first was found at 'RNA-binding 27 isoform X1' gene (chr13:20975921), the second at chromosome 22 upstream of 'zinc finger BED domain-containing 4-like' gene, the third upstream from 'PREDICTED: uncharacterized protein LOC106518831' gene (Scaffold4139:36071) (Table 3.9).

Selected SNPs for length/width phenotype

Finally, there were four SNPs associated with Length/Width ratio. The first SNP (chr6:23799286) is located on the 'phosphatase 1 regulatory subunit 3D-like'. The second SNP (chr8:2887261) located upstream of 'interferon regulatory factor 7'. The third (chr8:11613979) was located in 'semaphorin-3A' gene (Conserved in Asian sea bass, Asian swamp eel) and upstream of 'Piccolo' gene. The last one (chr15:19409307) upstream of an uncharacterized protein (Table 3.10).

Table 3.7: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Fat						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr21:19924408	-	0.007	0.01	-	-	-
Chr8: 1385781	Protection of telomeres 1	0.0013	0.01	-	✓	Intron
Chr8:2887261	interferon regulatory factor 7	0.03	0.01	✓	✓	Upstream
Chr8:4544371	serine threonine- kinase WNK1-like isoform X1	0.05	0.01	-	-	Intron

Table 3.8: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Weight						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr1:16636968	ethanolamine-phosphate cytidylyltransferase-like	0.005	0.01	✓	✓	3′ UTR
Chr6:12617755	myosin-7-like isoform X1	0.001	0.01	✓	✓	Upstream
Chr8:2887261	interferon regulatory factor 7	0.0003	0.01	✓	✓	Upstream
Scaffold15653:2041	adipocyte plasma membrane-associated	0.001	0.01	✓	-	Downstream

Table 3.9: Selected SNPs from SES algorithm with significance tequal to 0.05 (best method based on median squared error score).

Tag Weight						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr13:20975921	RNA-binding 27 isoform X1	0.025	0.01	✓	-	Upstream
Chr22:18343985	zinc finger BED domain-containing 4-like	0.0059	0.01	✓	-	Upstream
Scaffold4139:36071	PREDICTED: uncharacterized protein LOC106518831	0.0099	0.01	-	✓	Downstream
Chr15:1186816	-	0.004	0.01	-	-	-
Chr2:8737137	-	2.10e-5	0.01	-	✓	-

Table 3.10: Selected SNPs from SES algorithm with significance threshold equal to 0.05 (best method based on median squared error score).

Length / Width						
Variables	**Locus**	**P-Value**	**Threshold**	**GWAS**	**Conserved**	**Position**
Chr6:23799286	phosphatase 1 regulatory subunit 3D-like	0.0022	0.01	-	✓	Downstream
Chr8:11613979	semaphorin-3A	0.0138	0.01	-	✓	Intron
Chr8:2887261	interferon regulatory factor 7	0.0004	0.01	✓	✓	Upstream
Chr15:19409307	PREDICTED: uncharacterized protein	0.0027	0.01	-	✓	Upstream

4 Discussion & Analysis

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 seabream. The application of these methods on seven families, each measured on four phenotypes revealed several genetic 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 [37]–[40]. Some of those QTL have been verified in genetically unrelated populations [41]. However, no association study using SNP markers was available for production traits in seabream except this by **Palaiokostas2016** 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. 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 [42]. The difference between these two algorithms is that MMPC does not return multiple solutions. MMPC was shown to achieve excellent false positive rates [43]. Seen from the biological perspective, multiple equivalent signatures may arise from redundant mechanisms, for example, genes performing identical tasks within the cell. For example, [44] 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 which are practically interchangeable in terms of predictive power. SES was tested against LASSO [26] with continuous, binary and survival target variables, resulting in SES outperforming the LASSO algorithm [45] both in predictive performance and computational efficiency. Overall, SES seems to be performing well in smaller datasets, while OMP is known to perform better in larger datasets [46].

Our findings highlight novel SNPs found within or close to coding genes that are significantly associated with our focal traits of interest in seabream. 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, [47]. 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 [47]. 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 [48]. 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 fifteen genes in total. Interestingly, although those two traits represent the same trait at different stages we found no gene associated with both. 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 [49]. Strikingly, follistatin knockout mice have reduced muscle mass at birth underlying the importance of this gene in muscle growth [49]. Apart from Follistatin, the significance association with Myosin, an actin-based motor molecule with ATPase activity essential for muscle contraction, show 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 underling the importance genes regulating the nervous system in length. Also, the same SNP, that 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 [50]. Another associated gene, ethanolamine phosphate cytidylyltransferase, plays a role in lipid metabolism and finally EXT1, a gene regulating important developmental pathways such as hedgehog [51].
In the analysis regarding SES using data with imputation, we identified a SNP located on interferon regulatory factor 7 (IRF7) which seems to be associated with the three out of four phenotypes (Fat,Weight at harvest and Length/Width). IRF7 is a regulator of type I interferon-dependent immune responses. From previous studies on mice, IRF7 seems to play a key role in diet-induced alterations in energy metabolism and insulin sensitivity. IRF7 knockout mice displayed significant decreased weight gain and adiposity on a high fat diet [52].

5 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 seabream 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 of 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 QTL 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.

Literature

[1] X. Geng, D. Zhi, and Z. Liu, "Genome-wide Association Studies of Performance Traits", *Bioinformatics in Aquaculture: Principles and Methods*, pp. 415–433, 2017.

[2] C. S. Tsigenopoulos, B. Louro, D. Chatziplis, J. Lagnel, E. Vogiatzi, D. Loukovitis, R. Franch, E. Sarropoulou, D. M. Power, T. Patarnello, C. C. Mylonas, A. Magoulas, L. Bargelloni, A. Canario, and G. Kotoulas, "Second generation genetic linkage map for the gilthead sea bream Sparus aurata L.", *Marine Genomics*, vol. 18, no. PA, pp. 77–82, 2014.

[3] T. Fernandes, M. Herlin, M. D. L. Belluga, G. Ballón, P. Martinez, M. A. Toro, and J. Fernández, "Estimation of genetic parameters for growth traits in a hatchery population of gilthead sea bream (Sparus aurata L.)", *Aquaculture International*, vol. 25, no. 1, pp. 499–514, 2017.

[4] B. K. Peterson, J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, "Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species", *PLoS ONE*, vol. 7, no. 5, 2012.

[5] P. A. Hohenlohe, S. J. Amish, J. M. Catchen, F. W. Allendorf, and G. Luikart, "Next-generation rad sequencing identifies thousands of snps for assessing hybridization between rainbow and westslope cutthroat trout", *Molecular Ecology Resources*, vol. 11, pp. 117–122, 2011.

[6] N. A. Baird, P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, and E. A. Johnson, "Rapid snp discovery and genetic mapping using sequenced rad markers", *PLOS ONE*, vol. 3, no. 10, pp. 1–7, October 2008.

[7] J. W. Davey, P. A. Hohenlohe, P. D. Etter, *et al.*, "Amphibian molecular ecology and how it has informed conservation", *Molecular Ecology*, vol. 23, no. 4, n/a–n/a, 2015. arXiv: NIHMS150003.

[8] J. M. Catchen, "Stacks: an analysis tool set for population genomics", *Molecular ecology*, vol. 22, no. 11, pp. 3124–3140, 2013. arXiv: NIHMS150003.

[9] S. He, Y. Zhao, M. F. Mette, R. Bothe, E. Ebmeyer, T. F. Sharbel, J. C. Reif, and Y. Jiang, "Prospects and limits of marker imputation in quantitative genetic studies in european elite wheat (triticum aestivum l.)", *BMC Genomics*, vol. 16, no. 1, p. 168, March 2015.

[10] J. Marchini and B. Howie, "Genotype imputation for genome-wide association studies", *Nature Reviews Genetics*, vol. 11, no. 7, pp. 499–511, 2010. arXiv: arXiv:1507.02142v2.

[11] H. Mulder, M. Calus, T. Druet, and C. Schrooten, "Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle", *Journal of Dairy Science*, vol. 95, no. 2, pp. 876–889, 2012.

[12] E. M. Van Leeuwen, A. Kanterakis, P. Deelen, M. V. Kattenberg, P. E. Slagboom, P. I. De Bakker, C. Wijmenga, M. A. Swertz, D. I. Boomsma, C. M. Van Duijn, L. C. Karssen, and J. J. Hottenga, "Population-specific genotype imputations using minimac or IMPUTE2", *Nature Protocols*, vol. 10, no. 9, pp. 1285–1296, 2015.

[13] J. E. Rutkoski, J. Poland, J.-L. Jannink, and M. E. Sorrells, "Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy", *G3: Genes | Genomes | Genetics*, vol. 3, no. 3, pp. 427–439, 2013.

[14] S. Raschka, *Python Machine Learning*, 1. 2014, pp. 1–5. arXiv: arXiv:1011.1669v3.

[15] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", no. March 2001, 2016.

[16] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[17] M. Jordan, J. Kleinberg, and B. Scho, *No Title*.

[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models using lme4", *ArXiv e-prints*, Jun. 2014. arXiv: 1406.5823 [stat.CO].

[19] B. Balliu and N. Zaitlen, "A novel test for detecting SNP to SNP interactions in case-only trio studies", *Genetics*, vol. 202, no. 4, pp. 1289–1297, 2016. arXiv: 1506.08683.

[20] I. Tsamardinos and C. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers", 2003.

[21] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: Astronomical or genomical?", *PLOS Biology*, vol. 13, no. 7, pp. 1–11, Jul. 2015.

[22] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, and I. Tsamardinos, "Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets", vol. 80, no. 7, 2017. arXiv: 1611.03227.

[23] A. Statnikov and C. F. Aliferis, "Analysis and computational dissection of molecular signature multiplicity", *PLOS Computational Biology*, vol. 6, no. 5, pp. 1–9, May 2010.

[24] C. Aliferis, I. Tsamardinos, A. R. Statnikov, and L. Brown, "Causal explorer: A causal probabilistic network learning toolkit for biomedical discovery", pp. 371–376, January 2003.

[25] M. Tsagris, V. Lagani, and I. Tsamardinos, "Feature selection for high-dimensional temporal data", *BMC Bioinformatics*, vol. 19, no. 1, p. 17, 2018.

[26] V. Lagani, G. Athineou, A. Farcomeni, M. Tsagris, and I. Tsamardinos, "Feature Selection with the R Package MXM: Discovering Statistically-Equivalent Feature Subsets", vol. 80, no. 7, 2016. arXiv: 1611.03227.

[27] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise", *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[28] S. A. Miller, D. D. Dykes, and H. F. Polesky, "A simple salting out procedure for extracting DNA from human nucleated cells", *Nucleic Acids Research*, vol. 16, no. 3, p. 1215, 1988.

[29] T. Manousaki, A. Tsakogiannis, J. B. Taggart, C. Palaiokostas, D. Tsaparis, J. Lagnel, D. Chatziplis, A. Magoulas, N. Papandroulakis, C. C. Mylonas, and C. S. Tsigenopoulos, "Exploring a nonmodel teleost genome through rad sequencing—linkage mapping in common pandora, pagellus erythrinus and comparative genomic analysis", *G3: Genes, Genomes, Genetics*, vol. 6, no. 3, pp. 509–519, 2016. eprint: http://www.g3journal.org/content/6/3/509.full.pdf.

[30] S. Andrews, "Fastqc a quality control tool for high throughput sequence data", January 2014.

[31] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2", *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012. arXiv: {\#}14603.

[32] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1. G. P. D. P. Subgroup, "The sequence alignment/map format and samtools", *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[33] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen, "Robust relationship inference in genome-wide association studies", *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, 2010. eprint: /oup/backfile/content_public/journal/bioinformatics/26/22/10.1093_bioinformatics_btq559/1/btq559.pdf.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] R. J. Tibshirani and R. Tibshirani, "A bias correction for the minimum error rate in cross-validation", *Ann. Appl. Stat.*, vol. 3, no. 2, pp. 822–829, Jun. 2009.

[36] Y. Fang, "Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models", *Journal of Data Science*, vol. 9, pp. 15–21, 2011.

[37] D. Loukovitis, E. Sarropoulou, C. S. Tsigenopoulos, C. Batargias, A. Magoulas, A. P. Apostolidis, D. Chatziplis, and G. Kotoulas, "Quantitative Trait Loci involved in sex determination and body growth in the gilthead sea bream (Sparus aurata L.) through targeted genome scan", *PLoS ONE*, vol. 6, no. 1, 2011.

[38] D. Loukovitis, E. Sarropoulou, E. Vogiatzi, C. S. Tsigenopoulos, G. Kotoulas, A. Magoulas, and D. Chatziplis, "Genetic variation in farmed populations of the gilthead sea bream sparus aurata in greece using microsatellite dna markers", *Aquaculture Research*, vol. 43, no. 2, pp. 239–246, 2012. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2109.2011.02821.x.

[39] D. Loukovitis, C. Batargias, E. Sarropoulou, A. P. Apostolidis, G. Kotoulas, A. Magoulas, C. S. Tsigenopoulos, and D. Chatziplis, "Quantitative trait loci affecting morphology traits in gilthead seabream (sparus aurata l.)", *Animal Genetics*, vol. 44, no. 4, pp. 480–483, 2013. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/age.12027.

[40] K. Boulton, C. Massault, R. D. Houston, D. J. de Koning, C. S. Haley, H. Bovenhuis, C. Batargias, A. V. Canario, G. Kotoulas, and C. S. Tsigenopoulos, "QTL affecting morphometric traits and stress response in the gilthead seabream (Sparus aurata)", *Aquaculture*, vol. 319, no. 1-2, pp. 58–66, 2011.

[41] D. Loukovitis, A. Siasiou, I. Mitsopoulos, A. G. Lymberopoulos, V. Laga, and D. Chatziplis, "Genetic diversity of Greek sheep breeds and transhumant populations utilizing microsatellite markers", *Small Ruminant Research*, vol. 136, pp. 238–242, 2016.

[42] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "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, 2003, pp. 673–678.

[43] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local causal and Markov Blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions", *The Journal of Machine Learning Research*, vol. 11, pp. 235–284, 2010.

[44] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?", *Bioinformatics*, vol. 21, no. 2, pp. 171–178, 2005. eprint: /oup/backfile/content_public/journal/bioinformatics/21/2/10.1093/bioinformatics/bth469/2/bth469.pdf.

[45] A. Groll and G. Tutz, "Variable selection for generalized linear mixed models by l1-penalized estimation", *Statistics and Computing*, vol. 24, no. 2, pp. 137–154, March 2014.

[46] M. Tsagris, Z. Papadovasilakis, K. Lakiotaki, and I. Tsamardinos, "Efficient feature selection on gene expression data: Which algorithm to use?", *BioRxiv*, 2018. eprint: https://www.biorxiv.org/content/early/2018/10/03/431734.full.pdf.

[47] R. Sordella, W. Jiang, G. C. Chen, M. Curto, and J. Settleman, "Modulation of Rho GTPase signaling regulates a switch between adipogenesis and myogenesis", *Cell*, vol. 113, no. 2, pp. 147–158, 2003.

[48] M. E. Yalvac, J. Amornvit, L. Chen, K. M. Shontz, S. Lewis, and Z. Sahenk, "AAV1.NT-3 gene therapy increases muscle fiber diameter through activation of mTOR pathway and metabolic remodeling in a CMT mouse model", *Gene Therapy*, pp. 1–10, 2018.

[49] S.-J. Lee and A. C. McPherron, "Regulation of myostatin activity and muscle growth", *Proceedings of the National Academy of Sciences*, vol. 98, no. 16, pp. 9306–9311, 2001.

[50] K. Mukherjee, X. Yang, S. H. Gerber, H.-B. Kwon, A. Ho, P. E. Castillo, X. Liu, and T. C. Sudhof, "Piccolo and bassoon maintain synaptic vesicle clustering without directly participating in vesicle exocytosis", *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6504–6509, 2010.

[51] A. F. Siekmann and M. Brand, "Distinct tissue-specificity of three zebrafish ext1 genes encoding proteoglycan modifying enzymes and their relationship to semitic Sonic Hedgehog signaling", *Developmental Dynamics*, vol. 232, no. 2, pp. 498–505, 2005.

[52] X.-A. Wang, R. Zhang, S. Zhang, S. Deng, D. Jiang, J. Zhong, L. Yang, T. Wang, S. Hong, S. Guo, Z.-G. She, X.-D. Zhang, and H. Li, "Interferon regulatory factor 7 deficiency prevents diet-induced obesity and insulin resistance", *AJP: Endocrinology and Metabolism*, vol. 305, no. 4, E485–E495, 2013.

Supplementary

.1 Figures

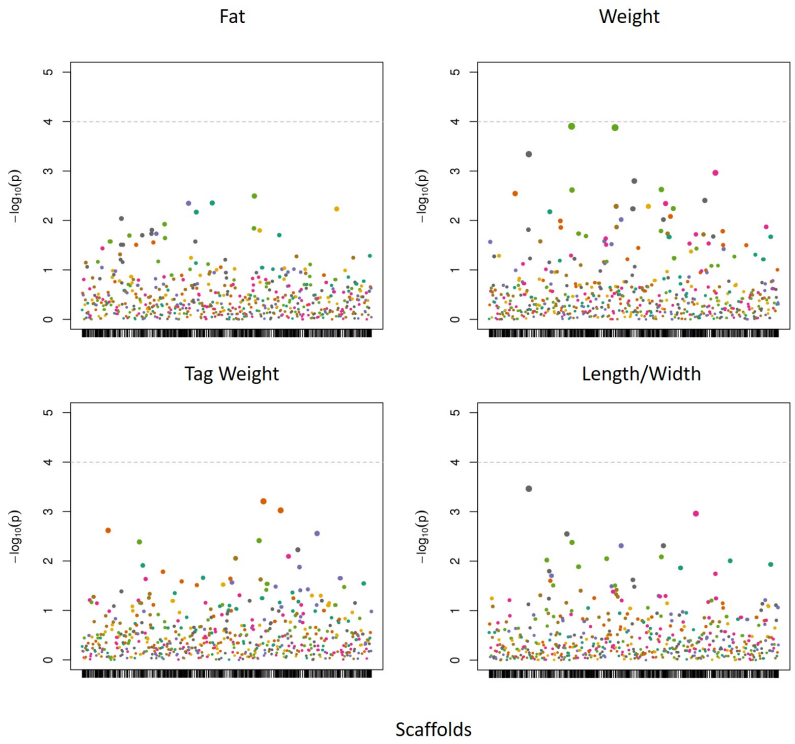


Figure .1: Manhattan plot demonstrating the locations across the chromosomes of the seabream genome (horizontal axis) versus the $-log(p-values)$ of the association between the genetic variants and phenotype (vertical axis). The higher the dots, the stronger the genetic association. The significance threshold was set to 10^{-4}, in order to correct for multiple testing (dashed line).

.2 Tables

Table .1: Number of reads per individual

Individual ID	**Number of reads per individual**
100	6546663
117	2193433
120	1005057
122	1345189
131	609747
133	7155026
134	1269108
137	929301
141	258389
15	781715
153	1355965
155	7195415
157	1886372
16	2372039
167	4138456
17	3981644
174	3579122
175	2620179
176	4682267
18	2819597
181	2189736
185	1583733
186	7738622
188	1702362
195	5168862
2	2605505
201	255634
202	10872537
211	3587667
213	3549243
219	1411377
221	1121468
225	1772750
232	2794623
238	1296888
246	2979025
251	1126376

Table .1: Number of reads per individual

Individual ID	**Number of reads per individual**
261	2538366
266	6096614
273	2467838
275	3823557
276	3326784
277	1100435
278	4455356
281	1395841
286	14708926
289	2700969
31	660023
310	1825366
32	799676
325	670714
332	2053479
340	1232620
344	352136
347	4701092
351	3718157
359	1564283
367	4430484
369	3645438
372	1139189
381	1148960
382	281554
392	1638974
393	3437959
397	10513523
40	3118405
400	1488846
404	8788200
407	2005444
419	1692536
429	2522008
431	2267029
443	1328410
448	172377
452	1124422
472	3674678

Table .1: Number of reads per individual

Individual ID	**Number of reads per individual**
474	961966
48	667114
485	603704
490	3694626
511	3500907
543	1469273
546	2399251
565	542305
593	659475
599	1191062
625	1956917
662	1837502
665	2843772
681	985762
690	358913
723	629737
748	1334033
786	1106786
809	6780868
829	2504467
831	2894008
84	2616821
860	1818399
870	5981632
877	1325468
878	1800465
882	1745418
92	1078240
93	3478466

Table .2: Equivallent signatures retrieved by SES for fat content

Variables	**Signature 1**	**Signature 2**
Chr13:1098152	✓	✓
Chr21:19924408	✓	
Chr8: 1385781	✓	✓
Chr20:19164407		✓

Table .3: Equivallent signatures retrieved by SES for weight at harvest

Variables	**Signature 1**	**Signature 2**	**Signature 3**	**Signature 4**
Chr1:16636968	✓		✓	
Chr6:12617755	✓	✓	✓	✓
chr16:2232897	✓	✓		
Chr:8:11613979	✓	✓	✓	✓
Chr1:6970078		✓		✓
Chr1:20827142			✓	✓

Table .4: Equivallent signatures retrieved by SES for weight at tagging

Variables	**Signature 1**	**Signature 2**	**Signature 3**	**Signature 4**
Chr13:20883924	✓	✓	✓	
Chr2:2623351	✓	✓	✓	✓
Chr22:18343985	✓			
chr13:20975921	✓	✓	✓	✓
scaffold4139:36071	✓	✓		✓
chr13:20975924		✓	✓	✓

Table .5: Equivallent signatures retrieved by SES for length to width ratio

Variables	**Signatures**															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
chr1:20827142	✓								✓							
chr13:9665394	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
chr3:9671223	✓	✓	✓	✓	✓	✓	✓	✓								
chr6:23799286	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
scaffold13177:8369	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Chr11:21805922		✓								✓						
Chr11:21807386			✓								✓					
Chr16:2232897				✓								✓				
Chr4:23265532					✓								✓			
Chr4:23265546						✓								✓		
Chr15:20045450							✓								✓	
Scaffold5661:35982								✓								✓
Chr15:13630715									✓	✓	✓	✓	✓	✓	✓	✓