UNIVERSITÀ DEGLI STUDI DI PADOVA

Università degli Studi di Padova

Dipartimento di Biomedicina Comparata e Alimentazione

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE VETERINARIE

INDIRIZZO COMUNE

CICLO XXVIII

DEVELOPMENT OF TOOLS FOR TRACEBILITY AND FOR ASSESSING THE GENETIC IMPACT OF AQUACULTURE

Direttore della Scuola: Prof. Gianfranco Gabai

Coordinatore d'indirizzo: Prof. Giuseppe Radaelli

Supervisore: Prof. Bargelloni Luca

Dottorando: Francesco Maroso

INDEX

Sommario 1

Abstract 3

Flowchart: rationale of the thesis 5

General introduction

Overexploitation of marine fish resources 7

Solutions to overexploitation of marine fish 8

Risks for wild populations related with aquaculture 9

Loss of genetic diversity within populations 11

Loss of inter-population genetic diversity 12

Traceability 13

Loss of fitness 13

Monitoring 14

Aquatrace 15

Gilthead seabream 15

Restriction site Associated DNA (RAD) sequencing 16

Phd 17

Performance and precision of double digestion RAD (ddRAD) genotyping in multiplexed datasets of marine fish species

Abstract 20

Introduction 21

Material And Methods 24

Results 29

Discussion 34

Conclusion 39

Supplementary Material 40

RAD SNP markers as a tool for conservation of dolphinfish *Coryphaena hippurus* in the Mediterranean Sea: identification of subtle genetic structure and assessment of populations sex-ratios

Abstract 43

Introduction 44

Material And Methods 46

Results 49

Discussion 52

Conclusion 57

Supplementary Material 58

ddRAD SNPs markers reveal subtle genetic structure of Gilthead Sea Bream (*Sparus aurata*) in European wild populations and high divergence between farm broodstocks: implication for aquaculture and natural stock management

Abstract 72

Introduction 73

Material And Methods 77

Results 83

Discussion 91

Conclusion 96

Supplementary Material 97

SNPs identification and validation in *Thunnus albacares* and *Scomberomorus brasiliensis* by double digest RADseq using a 454 pyrosequencing platform

Abstract 109

Introduction 110

Material And Methods 111

Results and Discussion 114

Conclusion 118

Supplementary Material 119

General conclusion 133

Bibliography 135

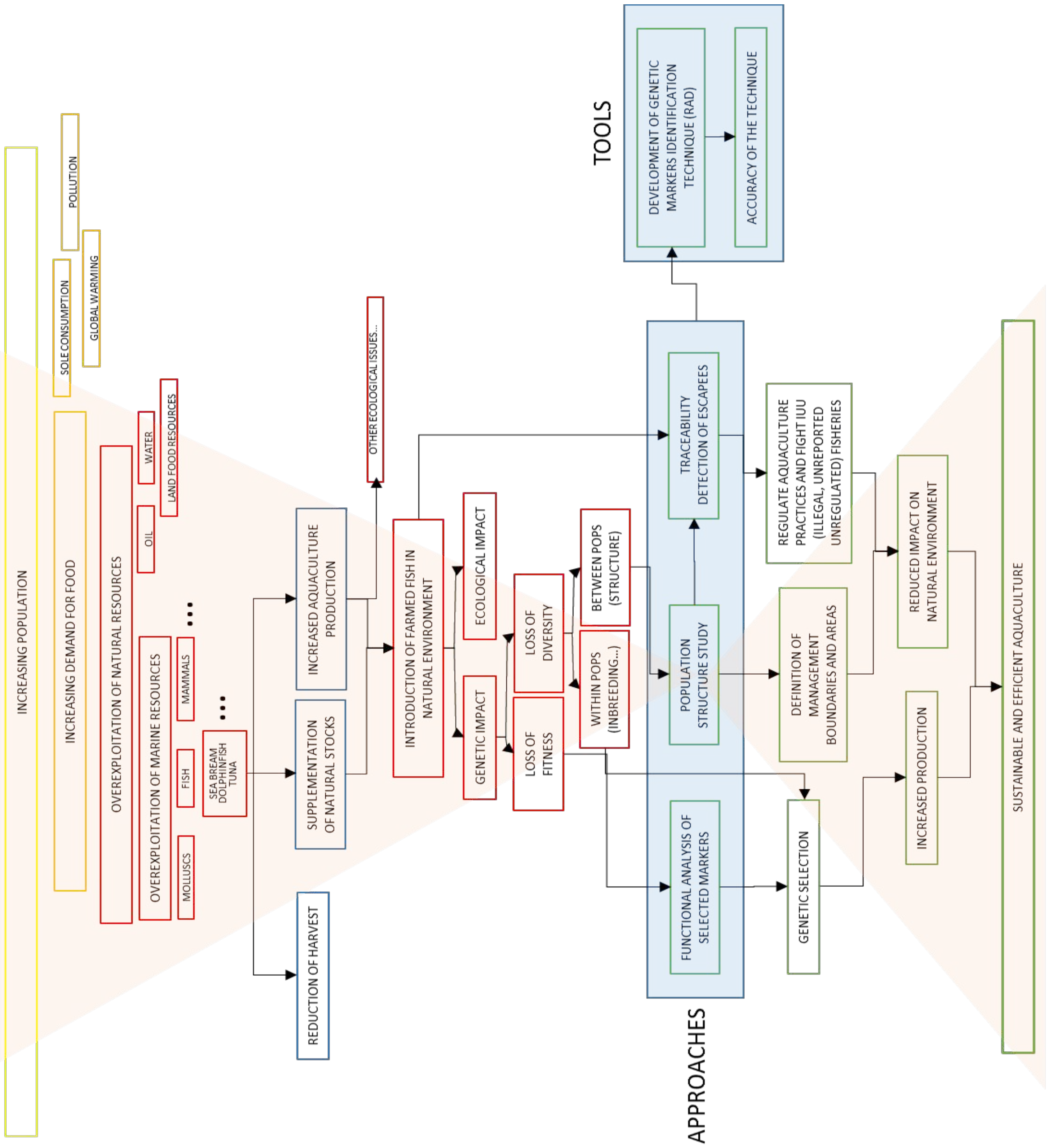
SOMMARIO

Il sovrasfruttamento delle risorse naturali di cibo, tra cui quelle marine, ha messo in serio pericolo la sopravvivenza di molte specie e la loro disponibilità per il consumo da parte dell’uomo. La riduzione del prelievo, il ripopolamento degli stock naturali con pesci allevati e l’allevamento stesso come fonte alternativa di prodotto sono mezzi comunemente utilizzati per risolvere il duplice problema di garantire cibo di qualità e preservare l’ambiente naturale. Tuttavia, le ultime due misure presentano potenziali effetti collaterali, tra i quali un impatto sulla diversità genetica delle popolazioni naturali soggette a ripopolamenti o fughe dagli allevamenti. Per comprendere i rischi legati a questi due eventi, è fondamentale studiare le caratteristiche genetiche delle popolazioni selvatiche e dei riproduttori usati in allevamento. Allo stesso tempo, le tecniche di analisi sviluppate possono essere sfruttate per la tracciabilità del prodotto allevato e selvatico, aspetto che sta guadagnando sempre maggiore importanza tra i consumatori. L’analisi basata su tecniche di caratterizzazione genetica di tipo RAD ha permesso lo studio di più di 1000 campioni di orata con 1240 marcatori SNPs. I risultati suggeriscono una suddivisione dei campioni naturali in quattro gruppi geneticamente distinti: Atlantico, Mediterraneo Ovest, Ionio e Egeo. L’analisi dei broodstocks dei maggiori allevamenti europei ha rivelato una differenziazione genetica tra i gruppi più elevata di quella osservata tra i selvatici, probabilmente dovuta all’uso di un ridotto numero di riproduttori e alla deriva genetica; è stata rilevata anche una minore variabilità genetica all’interno dei gruppi allevati, talvolta al di sotto dei limiti considerati sicuri per evitare l’inbreeding; infine, alcuni riproduttori portano tratti genetici che potrebbero rendere la prole non adatta all’ambiente naturale che incontrerebbe in caso di fuga o rilascio. Il confronto delle caratteristiche genetiche dei gruppi allevati e selvatici ha permesso di discutere il potenziale impatto dell'acquacoltura sulla fitness e sul potenziale adattativo delle popolazioni selvatiche. Le tecniche messe a punto e i risultati ottenuti sono di grande importanza per lo sviluppo del settore dell'acquacoltura di orata e per la corretta gestione e salvaguardia delle popolazioni naturali delle zone coinvolte nella produzione.

ABSTRACT

Overexploitation of natural food resources, among which marine resources, put in serious risk the survival of many species and its availability as human food. Reduction of harvest, restocking with farmed fish and farming as alternative source of product, are commonly used to grant high quality food in a sustainable way. Nevertheless, the last two solutions have potential side effects, among which the genetic impact on natural populations that are involved in restocking actions or escapees from fattening cages and farms. Study the genetic structure of the species in the wild and farming environment is a key aspect to understand the real risks related to aquaculture. At the same time, genetic tools developed in the process can be used to trace wild and farmed origin of fish product, which is an aspect that is gaining great interest among consumers. In the study presented in the thesis, genetic analysis based on RAD genotyping allowed the study of more than 1000 wild and farmed samples with 1216 SNP. The results obtained suggest a subdivision of natural samples in four genetically distinct groups: Atlantic, West Mediterranean, Ionian Sea and Aegean Sea. The analysis carried out on many European broodstocks revealed a higher genetic differentiation compared to wild groups, probably due to founder effects and genetic drift; broodstocks are characterized by lower genetic variability, that in some cases fell below the minimum threshold to avoid inbreeding; finally, some of the broodstocks showed genetic traits that could make offspring unfit to the natural environment they would find in case of restocking of escapees. Comparing wild and farmed groups stimulated a discussion on the potential impact of aquaculture on natural populations, considering the reduction in fitness and the loss of inter/intra groups genetic variability, that cause a loss in long-term adaptation potential. The analytical techniques used and the results obtained are important for the development of gilthead sea bream aquaculture in Europe and for the correct management and protection of natural populations from the areas involved in production.

FLOWCHART: RATIONALE OF THE THESIS


INCREASING POPULATION
SOLE CONSUMPTION
POLLUTION
GLOBAL WARMING
INCREASING DEMAND FOR FOOD
OVEREXPLOITATION OF NATURAL RESOURCES
WATER
OIL
LAND FOOD RESOURCES
OVEREXPLOITATION OF MARINE RESOURCES
MAMMALS
FISH
MOLLUSCS
SEA BREAM
DOLPHINFISH
TUNA
INCREASED AQUACULTURE PRODUCTION
SUPPLEMENTATION OF NATURAL STOCKS
REDUCTION OF HARVEST
OTHER ECOLOGICAL ISSUES...
INTRODUCTION OF FARMED FISH IN NATURAL ENVIRONMENT
ECOLOGICAL IMPACT
GENETIC IMPACT
LOSS OF DIVERSITY
LOSS OF FITNESS
BETWEEN POPS (STRUCTURE)
WITHIN POPS (INBREEDING...)
TOOLS
DEVELOPMENT OF GENETIC MARKERS IDENTIFICATION TECHNIQUE (RAD)
ACCURACY OF THE TECHNIQUE
TRACEABILITY DETECTION OF ESCAPEES
POPULATION STRUCTURE STUDY
FUNCTIONAL ANALYSIS OF SELECTED MARKERS
APPROACHES
REGULATE AQUACULTURE PRACTICES AND FIGHT IUU (ILLEGAL, UNREPORTED UNREGULATED) FISHERIES
DEFINITION OF MANAGEMENT BOUNDARIES AND AREAS
GENETIC SELECTION
REDUCED IMPACT ON NATURAL ENVIRONMENT
INCREASED PRODUCTION
SUSTAINABLE AND EFFICIENT AQUACULTURE

Rationale of the work carried out during the PhD. From top to center, the problem is introduced starting from a wider point of view and then focusing on the particular issue tackled. In the center of the graph (blue background), the approaches used to study the problem and the development of the techniques. Form center to the bottom, how results found within the work can be implemented in a wider perspective to enhance the role of aquaculture in providing food resources for the growing population and, at the same time, avoid overexploitation of the marine resources

GENERAL INTRODUCTION

Overexploitation of marine fish resources

Recently, human population has grown over 7 billion people [1] and it is increasing at a rate of more than 1% every year. One of the most evident consequences of human population increase of the last century is the overexploitation of the resources humans use. From oil to food, the rate of exploitation is unsustainable by the rate of renewal. Applied to ecology, the concept of ‘overexploitation’ has been addressed among the activities that threaten global biodiversity more [2]. In the marine environment context, overexploitation involves a wide range of living organism: humans use over 400 species as food resources. In addition, the trophic level of the exploited species is in general higher than for land organism, and this fact makes exploitation even worse in ecological terms. It is interesting to notice that hunter-gatherers’ attitude toward marine organism is comparable to that abandoned over 10’000 ago in the land and, also due to this, management of marine resources are far less developed than for land animals [3]. Anyhow, exploitation technology has been developing continuously and the combination of the two things has led to the actual situation where many marine species are threatened to extinction. According to the United Nations Food and Agriculture Organization (FAO) over 25% of all the world's fish stocks are either overexploited or depleted and 52% are fully exploited [4]. Thus a total of almost 80% of the world's fisheries are fully to overexploited, depleted, or in a state of collapse. Although, these estimates are considered rather conservative. Recently, a study showed that 29% of fish and seafood species have collapsed (i.e. their catch has declined by 90%) and are projected to collapse within by 2048, unless immediate action is taken [5]. Worldwide, about 90% of the stocks of large predatory fish stocks are already collapsed. While the most visible and known problems affect the open ocean environment, and mostly large species (e.g. tunas, dolphins, turtles…) affect public opinion on the problem, also coastal and intertidal areas show high level of decline and population crushing worldwide [6], due to overfishing or, indirectly, to other human related activities (pollution, exploitation of the ground and water...).

As happened with land animals, also for marine resources the concept of overexploitation gained public consciousness much later than the problem itself became real. As a consequence, the correct understanding of the seriousness of human impact is not always achieved (for more detail see 'shifting baseline syndrome' described by Pauly(1995) [7]. Anyhow, after consciousness raising about the health status of our oceans and seas, the first management laws started appearing around 200 years ago and were applied to Norwegian fisheries [8].

Solutions to overexploitation of marine fish

Generally speaking, three approaches are used to cope with marine population declining due to overexploitation: the first and most intuitive one is the reduction of the fishing efforts. Though, this approach cannot cope with the problem of increasing demand for fish product for human consumption, can create social issues in communities where fishing is an important economical activity and is therefore feasible only in rare cases or for small, defined areas. The aim of sustainable fisheries management is therefore to balance wildlife conservation and harvesting.

Intuitively, a wise application of such an approach cannot overlook a deep knowledge of the biology, demography and genetics of the addressed species. For example, genetic data based on molecular analysis can provide (in a non-invasive way) useful information about the good environmental status of a population as well as about the effects of the past harvesting pressure and the risk of harvesting at particular levels.

The second possible solution tackles the problem of the increasing demand for fish product by developing farmed production of fish in inland and sea-based facilities. This is exactly what happened, starting from 10'000 years ago, for land livestock with animal farming, and is nowadays the most important source of animal products worldwide, without which it would be impossible to provide people with sufficient animal food. While being still far from the levels reached by land livestock farming, fish production and associated technologies have been increasing rapidly.

Side effects of an unregulated increase in farm production can be bad, but the knowledge of the involved dynamics carried by our experience with land animals can help preventing and dealing with them. To briefly summarize, the best known drawbacks are indirect effects on the environment where farming takes place (e.g. land exploitation and destruction, use and release of drugs, pollution caused by animal manure) and, less known but equally important, the effect on the genetic composition of the harvested species caused by farming practices.

The third approach considered here is applied in case of serious depletion of a natural population, or when natural population’s restoring potential is not enough to guarantee survival of the population. In these cases, restocking can be a solution, that is releasing animals in the wild in order to enhance the biomass of a species in an area. While in some cases individuals from other wild populations are released, more often animals produced in farms are used and the practice of generating individuals for restocking is called 'supportive breeding'. In this latter case, the potential bad effect of ‘genetic pollution’ can be highly relevant, especially if the restocking practices are not undertaken with particular care for important parameters such as the genetic characteristics of the breeders and the genetic variability of the released stock.

Risks for wild populations related to aquaculture

One consequence of farming and restocking is the introduction of fish of farm origin in the wild, either released on purpose or accidentally by escape events. Ecological aspects (competition for food and reproduction, alteration of food chain, alteration of environment) [9] and genetic aspects are involved.

To set up a study on the potential genetic impact of aquaculture, previous knowledge of the species is required. Firstly, the genetics of the wild populations needs to be studied and understood. Information provided by this type of study are useful also as a base for management policy of any species that is harvested in nature. Secondly, it is important to gain genetic information for the major broodstocks and information about the farming/selection practices going on in at least the major farms that work in the distribution area of the species. The latter point includes both information at the hatchery level (to know where and how juveniles are produced) and at the fattening farms (that are the main source of escapees). Altogether, this information is fundamental to understand the potential effect of aquaculture on the genetic makeup of the species and be able to provide tools for a sustainable development of farming.

Escape events play an important role for the presence of farmed fish in the wild [10–12]. Severity of these events can range from relatively constant leakage of small numbers of individuals to large catastrophic events involving thousands of individuals, gametes or larvae. Farm type affects risks of introgression: the probability of escapes is highest from ponds and net pen cages in near-shore or open-ocean sites. Since genetic effects spread with reproduction, whether released/escaped individuals mate is fundamental. Therefore, the life stages involved also affect the likely genetic consequences: a large event involving thousands of escaped juveniles would not be expected to lead to the same level of introgression as the same number of escaped adults, as many juveniles would be expected to die before maturity. A variety of factors (e.g. source of broodstock, selection practices and proximity to spawning grounds) affects escaped individuals' reproductive success in the wild.

Minimizing opportunities to escape ca be achieved by using land base systems, improving cages resistance to environment and better placing cages. To reduce opportunities for reproduction the use of sterile specimens is advised, which has also the additional advantage of increasing growth rates in many species [13]; moreover, the use of highly domesticated fish (i.e. adapted to the peculiar farm environment) can reduce their chance to survive the natural environment and thus reproduce.

When these precautions are not taken, genetic introgression can happen and need to be evaluated.

Taking example from the well-studied case of salmon breeding, three points deserve particular focus from a genetic perspective: i) the loss of genetic diversity within populations; ii) the loss of population structure and iii) the loss of fitness for wild populations [14].

Genetic variation in a species provides the raw material for its evolution and survival. All else being equal, populations with low levels of genetic variability have less capacity to respond to stressful conditions or environmental changes. Overall genetic variability can be split in several components (i.e. within individuals, among individuals in a population, among populations...) that, summed up, provide a view of the genetic "health" of a species and its flexibility to face a changing environment. The first two points listed above are related to the potential loss of genetic variability of the species in the wild.

Loss of genetic diversity within populations

Intuitively, the census size of a population is the main parameters to estimate when it comes to evaluating the 'health status' of a species subject to exploitation. Anyhow, in a genetic and evolutionary perspective, the amount of genetic variation that can be passed to future generation is much more important [15], as offspring might be challenged by future environmental changes. Therefore, the actual number of breeding individuals should be taken into account. Nevertheless, additional variables such as unequal sex ratios, skewed distribution of reproductive success or high relatedness between breeders should be considered to develop more reliable indicators, as all these factors can reduce the actual genetic variability. To cope with the issues described above, Effective Population Size (N_e) is used to characterize wild population and to evaluate the genetic variability of a group, which reflects its 'health status' in terms of potential for adaptation. This is therefore a fundamental parameter for conservation and management [16]. Low N_e means fast loss of genetic variability because of drift in wild populations. At the same time, large N_e plays an important role in facilitating the action of natural selection, because genetic changes promoted by selection are overwhelmed by those arising from genetic drift especially in small populations.

Also when studying farms' broodstocks, N_e can be an important tool as it is correlated to 'inbreeding'. Therefore, its calculation can help farmers in understanding how inbred their broodstocks are, and avoid inbreeding depression. In the perspective of conservation and wild stocks management, low N_e values of reared broodstocks can represent a serious risk in case of production of offspring for restocking purposes or in case of escapees, as it is expected that the N_e of a system that combines wild and farmed individuals will be a combination of the (normally high) N_e of the wild population ($N_e W$) and the (often low) N_e of the captive population ($N_e C$).

Maintaining high genetic variability within breeders is advised to reduce the impact of released/escaped animals on the evolutionary potential of impacted wild population [17]. On the other side, when restocking seriously depleted wild populations with small population sizes, genetically variable (i.e. characterized by high N_e) batches can help increasing the genetic diversity of the wild stock [18].

While the importance of an accurate estimation of N_e is clear, its calculation is not straightforward. Several methods have been developed to estimate it: i) temporal methods [19–21], based on two or more samples of the same population collected at different times; ii) single sample methods, performed using linkage disequilibrium [22], or heterozygote excess in the offspring [23]. Linkage disequilibrium method was shown to be biased when sample size is lower than estimated N_e, but a correction can be applied [24], which is normally implemented in recent Ne estimator software.

Loss of inter-population genetic diversity

Another important source of genetic variability is provided by genetic diversity among populations. This is to some extent reflected in the species specific population structure, which is an aspect commonly studied in population genetics. Together with intra population variability, it gives a species the potential to cope with environmental changes in a long term perspective, maintaining overall productivity high under a wider ranges of conditions. In the marine environment, mainly due to lack of barriers to dispersal, population structure is expected to be weaker than what is normally found in land animals [25]. Though, strong structure is more likely to develop for coastal species, which are not expected to have long range migratory habits, and whose dispersal potential should be limited only to eggs and larvae, which might be passively transported by water currents.

While more and more information is being accumulated, it is getting clear that a wide range of population structuring levels characterize marine fish, therefore one has to be prudent to assume no population differentiation exists.

A commonly used strategy to minimize loss of genetic variability is using fish of local origin as breeders. While offspring produced in this way should not affect the natural structure of wild populations, it is still important to maintain high genetic variability within the broodstock, to avoid the aforementioned problems related with with low N_e. Another issue of using local breeders is that they can be sub-optimal in terms of growth rates and resistance to pathologies compared to non-local breeders, which is why, in real cases, broodstocks are often composed by animals of different origin, which further complicate the situation.

Traceability

The existence of genetic differentiation between populations in the species distribution range suggests that it could be possible to detect the origin of a sample based on its genetic profile. Traceability is one of the most interesting tool that exploit the genetic information provided by the analysis of wild or farmed animals.

The identification of the geographic origin of wild samples can be used in the fight of unregulated, unreported and illegal fishing (UUI) or for labeling purposes, especially in a market in which the consumers are increasingly interested in the origin and the supply chain of the products they buy.

In addition, in a context where farming is gaining importance every day, the ability of tracking individuals back to the origin farm can serve at least two important needs: fish can be assigned to origin farm in case of problem related to food safety; in case of release/escapes of farmed fish, individuals can be tracked back to origin farm once caught in the wild. This last tool is fundamental for keeping aquaculture activities under control and to assess the impact of fish farming and restocking on the wild populations. It is also important to notice that the analysis presented above are barely invasive, as DNA can be extracted from a small portion of tissue, whose excision doesn’t have consequences on the health of the fish.

Loss of fitness

In a farm, animals are selected (either directly by farmers or indirectly by the peculiar condition of captivity) for characteristic that are remarkably different from the optimal in natural environment. Therefore, offspring’s fitness to natural environment is likely to be altered as a consequence of two mechanisms: domestication [26] and inbreeding depression [27]. Domestication happens when farmers select their breeders for traits that enhance production. The traits most commonly targeted are growth, morphology and disease resistance [28]. The extent at which these traits are selected reduces fitness to the natural environment. When fish from farms cross with wild specimens, the resulting offspring might therefore be less fit. In the context of supportive breeding, the actual effectiveness of the strategy in the long term can therefore be compromised. Similarly, inbreeding depression is a result of farm practices. In the selection process, individuals that share favorable traits (and are therefore kept as breeders) are often closely related. As a consequence, the selection of these individuals as novel breeders can reduce the overall fitness of offspring in the long term.

It is expected that phenotypic traits that affect fitness are linked with genetic markers, but they are expected to be difficult to find, as they are probably scattered in the genome and are low in number if compared to the "unselected" markers. Nevertheless, newly developed genotyping techniques parse the genome at much higher resolution, and the chance to detect loci linked to phenotypic traits under selection consequently increases. Often, these markers show a peculiar behavior, which is not expected if they evolved under natural selection (e.g. odd differences in allele frequencies between groups, correlation between frequency and environmental variable...). These characteristic are exploited by methods that are used to detect these markers, in jargon called "outliers".

Monitoring

Monitoring should be an integral part of both production and management programs. It has the threefold function of allowing escapees detection and understand their effect on natural populations, evaluate effectiveness of measures to reduce risk and, if well designed, it can reduce the need of unnecessary or expensive sampling efforts [29].

With regard to marine stock enhancement, given the continental or global scale at which it takes place, monitoring can be very difficult without the coordination between the different parts interested. Anyhow, the same tools developed for studying populations and risks associated with aquaculture can be used to implement monitoring practices. If a project is carried out at large scale (i.e. covering most of the species distribution area) and is well coordinated, results will be maximized so that no effort is wasted.

Aquatrace

It is exactly in the framework previously described that Aquatrace set its roots. Aquatrace is a project funded by European Union in the context of Framework Program 7 and involves 22 academic and private entities. Its first aim is "the development of tools for tracing and evaluating the genetic impact of fish from aquaculture". In other words, this project takes advantage of cutting edge genetic and genomic analytical approaches to support aquaculture activity and management, as well as the protection of our marine and freshwater environments. The rationale behind Aquatrace is to develop reliable and cost-effective molecular tools for the identification of the genetic origin of both wild and farmed fish (genetic traceability), as well as for the detection of interbreeding between farmed and wild stocks. This work is carried out on three marine fishes of economic significance and with growing aquaculture activities, the European sea bass, the gilthead sea bream and the turbot. The project is willing to give its contribution to the common challenge of Europe to develop sustainable aquaculture and, at the same time, preserve the environment from the potentially adverse effects of uncontrolled development of aquaculture, that mainly spread through escapees or releases.

Gilthead sea bream

Within the Aquatrace, our group was responsible for the analysis of the gilthead sea bream (*Sparus aurata*). Sea bream is an important demersal commercial species, highly appreciated as food fish for its flesh. It is a coastal species and is characterized by protandrous hermaphroditism, with males reaching maturity at the second year of life, and changing sex generally in the second spawning season.

It is a subtropical fish distributed from 62°N-15°N, 17°W-43°E in the Mediterranean Sea, the Black Sea and the Eastern Atlantic Ocean, from the British Isles to Cape Verde [4]. In North-East Atlantic waters, the species is still considered rare, as colder waters limit its distribution to the English Channel and the Celtic Sea; capture records have recently increased in England and Ireland [30].

Wild populations have not been well characterized yet from a population genetic point of view, and for many geographic areas inconsistencies and lack of information don’t allow a clear understanding of population structuring. This unclear picture requires further studies to determine the genetic and phenotypic structure of the gilthead sea bream over its whole geographical range in order to develop strategies for the conservation of wild populations and for the genetic-based management of farmed stocks. Population genetic results published so far rely only on allozymes, microsatellites and mitochondrial DNA markers [31–35]. For this reason, the development of species specific SNP markers may be very informative for understanding the genetic pattern of the species in its distribution area.

Together with European sea bass, it’s the main marine aquaculture species in the Mediterranean region, with a global production that reached almost 170,000 tonnes in 2012 [36]. Both sea cages and land based facilities are used [37]. Although breeding programs are already in progress for the most important phenotypic traits, marker assisted selection is at its very beginning, but is highly promising for production efficiency. Important genetic information has been independently collected for several European farm broodstock, but so far not much is known about origin of broodstock, exchange of breeders, eggs or juveniles and it is therefore difficult to drew a general interpretation about the potential consequences of farming on wild populations. In general, strong founder effects and loss of genetic diversity are recorded for broodstock, leading to high characterization of each strain, which would make distinction of wild from farmed individuals easier [38,39]. However, no universal domestication markers are available yet.

Restriction site Associated DNA (RAD) sequencing

Genetic markers are features in the genomes that differentiate one taxonomic entity from another (either an individual, a sub-population, a population or even a species). Several types of markers exist and approaches to identify them vary accordingly. The earlier genotyping approaches included a “discovery” step, in which genomes were scan to identify informative regions; a selection step, in which the most informative markers were selected and filtered; and finally the development of a tool to characterize selected markers in a fast and cheap way (i.e. SNP-chip, array…).

Recently, novel approaches, generally referred to as Genotyping By Sequencing (GBS), have been developed that allow SNP discovery and genotyping steps to be performed simultaneously, substantially reducing analysis time and efforts. This was possible mainly because costs of sequencing technologies dropped substantially in the last years [40]. Following this trend, many protocols have been developed, among which Restriction site Associated DNA (RAD) genotyping is obtaining a rising interests for many reason. The amount of information contained in a species genome is much more than what is needed to answer relatively simple questions about evolution, life history, demographic history and phenotypic traits. Therefore, the possibility to analyze a reduced portion of the genome is appealing for reducing sequencing cost and analysis time. In addition to this, RAD sequencing ensures that the same portions of the genome are analyzed across specimens, as only fragments nearby restriction enzymes recognition sites are sequenced. Finally, many RAD techniques also allow the selection of subsets of the sequences cut, via specific adapters (e.g. 2bRAD) or fragment size selection (e.g. ddRAD). This feature makes RAD techniques very flexible and adaptable to many taxa and scientific purposes, as the amount of information obtained and cost can be decided a priori. In addition to this, the increasing throughput and better accuracy of newly developed sequencing machines means that more individuals can be analyzed simultaneously and lower coverage is needed to achieve reliable results. Variations of the original RAD technique have been developed (e.g. 2bRAD [41], ddRAD [42,43]), providing a variety of approaches, whose pros and cons have to be evaluated considering the species analyzed and the aim of the study [44,45].

PhD

The present work is a collection of the results obtained in the context of my PhD at the Veterinary School of the University of Padua (Italy). During four years, my main focus was the EU funded project Aquatrace, but in order to develop the skills needed to accomplish my tasks, I also collaborated in other projects, that brought to the publication of the scientific papers attached in the thesis.

In the first year of my PhD, I followed the coordination of sample collection for the three marine species (sea bass, sea bream and turbot), looking for possible tissue samples sources (e.g. project partners, fishermen and farms owners). DNA extraction was performed in the first year too. After sample collection and extraction was completed, in the second year my main effort focused on ddRAD library preparation. Since the ddRAD protocol selected for the analysis was developed by one of the partners (Dr. John Taggart, University of Stirling, UK), a period of one week was spent at Taggart’s laboratory in order to learn the technology that would be later transferred to UNIPD group, where is still used also for other projects. After library sequencing, I moved to the bioinformatic analysis of the outcome data, that took most of the third year. In this stage of the PhD, I worked on the technical paper presented here. With the aim of increasing my knowledge of population genetics tools and approaches, I also worked on another project focused on studying the population structure of the marine fish *Coryphaena hippurus*. The results of the work are reported in the published manuscript “RAD SNP markers as a tool for conservation of dolphinfish *Coryphaena hippurus* in the Mediterranean Sea: Identification of subtle genetic structure and assessment of populations sex-ratios” by Maroso et al. (Marine Genomics, 2016).

Finally, results obtained from the analysis of sea bream samples collected within the Aquatrace project were reported in a manuscript including the analysis of the wild populations and broodstocks and an evaluation of the potential risk posed by aquacuture to natural populations, providing useful tools and approaches for management of the species and tools that could be used by farms for monitoring their breeders.

During the second year of my PhD I spent six months at the research group led by prof. Paulino Martìnez at University of Santiago de Compostela (campus de Lugo), where I followed the development of a linkage map of turbot, based on the markers discovered in the Aquatrace, and in general I had the possibility to continue my research in a stimulating environment. There, I collaborated with prof. Martinez’s partners from University of San Paulo (Brazil) in writing a paper on the development of SNP in two species of tuna from the south America’s coasts, using a combination of ddRAD and 454 pyrosequencing.

With the aim of evaluating the RAD technique used in the project Aquatrace, we worked on a comparative analysis of the results obtained with ddRAD technique in three marine teleost species (i.e. Mediterranean Sea bass, gilthead sea bream and turbot). Different bioinformatic approaches were tested and results presented under different point of view, in order to provide other researchers with a comprehensive evaluation of the technique, including pros and cons of its use.

Performance and precision of double digestion RAD (ddRAD) genotyping in multiplexed datasets of marine fish species

Maroso, F.[a], Hillen, J.E.J.[b], Pardo, B.G.[c], Gkagkavouzis, K.[d], Coscia, I.[b,e], Hermida, M.[c], Franch, R.[a], Hellemans, B.[b], Van Houdt, J.[f], Simionati, B.[g], Taggart, J.B.[h], Nielsen, E.E [i], Maes, G.[b, f, l], Volckaert, F.A.M.[b], Martinez, P[c], Bargelloni, L. [a], AquaTrace Consortium, Ogden, R.[m]

[a] Department of Compared Biomedicine and Food Science, University of Padova, 35020 Legnaro, ITALY

[b] Laboratory of Biodiversity and Evolutionary Genomics, University of Leuven, Ch. de Bériotstraat 32 box 2439, B-3000 Leuven, Belgium

[c] Departmento de Zoología, Genética y Antropología Física, Universidade de Santiago de Compostela,27002, Lugo, Spain

[d] Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

[e] Current address: School of Environmental and Life Science, Rm 332, Peel building, University of Salford, Salford, M5 4WT, UK

[f] Department of Human Genetics, University of Leuven, O&N I Herestraat 49 - box 602, B-3000 Leuven, Belgium

[g] BMR Genomics, Via Redipuglia 21a, Padova, Italy

[h] Division of Environmental and Evolutionary Biology, School of Biology and Biochemistry, The Queen's University of Belfast, Belfast BT7 INN, Northern Ireland, U.K.

[i] National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, 8600 Silkeborg, Denmark

[l] Centre for Sustainable Tropical Fisheries and Aquaculture, Comparative Genomics Centre, College of Marine and Environmental Sciences, Faculty of

Science and Engineering, James Cook University, Townsville, 4811 QLD, Australia
[m] Royal Zoological Society of Scotland, WildGenes Laboratory, Edinburgh EH12 6TS, UK

Abstract

The development of Genotyping-By-Sequencing (GBS) technologies enables cost-effective analysis of relatively large numbers of Single Nucleotide Polymorphisms (SNPs), especially in ‘non-model’ species. Nevertheless, as such technologies enter a mature phase, biases and errors inherent to GBS are becoming evident. Here, the performance of an increasingly popular GBS approach, double digest Restriction enzyme Associated DNA (ddRAD) sequencing, was assessed in population level SNP screening studies. Three large sets of sequence data were generated from three marine teleost species (>2.5x10^{12} bases in total), using the same standardized protocol. A common bioinformatics pipeline was established, based on the widely used STACKS software, with and without the use of a reference genome. We performed analyses throughout the production and analysis of ddRAD data in order to explore (i) the amount of information lost due to heterogeneity in the number of raw reads across samples; (ii) the discrepancy between expected and observed tag length and tag coverage; (iii) the difference in performance of reference based vs. *de novo* approaches; and (iv) the sources of potential genotyping errors of the library preparation/bioinformatics protocol, based on the comparison of technical replicates. Our results showed that DNA integrity and time from sample collection affect the output in terms of percentage and absolute number of high quality sequence reads. Likewise, using a reference genome and *a posteriori* genotype correction improved genotyping precision. Individual read coverage revealed to be a key variable for reproducibility, but also variance in sequencing depth between loci in the same individual was identified and found to correlate to tag length. The results and insights presented here will help to select and improve approaches to the analysis of large datasets based on RAD-like methodologies.

Introduction

The options for studying the genomic constitution of individuals and populations are increasing rapidly thanks to the development of powerful and accurate sequencing technologies that provide higher throughput at decreasing costs [46]. Meanwhile, efficient reduced representation methods have been proposed to provide high sequence coverage for selected genomic regions, collectively named as Genotyping-By-Sequencing (GBS) technologies [45]. One of these GBS methods, Restriction-site Associated DNA sequencing (RAD-seq) [47] has become particularly popular as it allows the cost-effective analysis of thousands of markers for tens/hundreds of individuals in a single sequencing lane. The original RAD protocol has also been modified to optimize throughput and ease of use, generating several alternative RAD-like methods (*e.g.* Peterson et al. 2012[42]; Wang et al. 2012[41]; and the review by Andrews et al. 2016[48]).

As GBS technologies enter a more mature phase, biases and errors inherent to such methods are becoming apparent [49] and comparative analysis of the most popular RAD-like protocols have addressed some of these subjects [50]. Two recent studies [51,52] focused specifically on genotyping issues relating to double digest Restriction enzyme Associated DNA (ddRAD) [42]. ddRAD is one of the most recently developed RAD variants, known for its relative flexibility and ease of use. In addition to the sources of error that also affect other methodologies, the authors recorded ddRAD-specific issues such as the recovery of restriction fragments shorter than expected, amplification bias toward GC-rich fragments, non-specific cutting by restriction enzymes, newly formed restriction enzyme sites and drop of fragment number due to loss of restriction sites.

Beyond laboratory-based assessments of variation in ddRAD performance, there is a need to better understand the risk of errors associated with the production and use of ddRAD data, which is becoming increasingly relied upon for population genetic inference. Unawareness of the presence of biased markers can indeed lead to artificial excess of homozygotes [53], false departure from Hardy–Weinberg equilibrium [54], overestimation of inbreeding [55] and unreliable inferences about population structure that have the potential to distort research conclusions. As a consequence, natural resource management and policy can be seriously affected. In this study, we seek to expand the experimental evaluation of ddRAD by focusing on the performance of common bioinformatics approaches as applied to multiple, comparable, large ddRAD datasets of marine fish species. A technical evaluation focused on marine fish data is interesting due to some biological characteristics of this taxon, such as high relatively high SNP frequency and high heterozygosity, that can further affect genotyping accuracy.

The species analyzed in this study are the European sea bass (*Dicentrarchus labrax*), the gilthead sea bream (*Sparus aurata*) and the turbot (*Scophtalmus maximus*).

Available genomic resources are increasing for three species studied. Sea bass [56] and turbot [57] genomes have already been published and a draft sea bream genome will soon be published (L. Bargelloni, personal communication) and made available for this work. The three differ in the quality of their assembly, as indicated by the contig length (i.e. their respective N50 values, which is defined as the length N for which 50% of all bases in the sequences are in a sequence of length L < N). However, they share similar genome size and can thus be used to implement comparative and functional genomics analysis (Table 1). The use of species with different levels of genome sequence development permits assessing effects of the reference genome quality on approaches that use genomes to improve the performance of clustering methods for RAD data (e.g. reference based analysis in STACKS).

Table 1 Details of the genome resources used for European sea bass, gilthead sea bream and turbot.

Species	Length (Mbp)	N° of contigs	Average contig length	N50 (kbp)	Reference
European sea bass	668.3	37,783	17,687	62	Tine et al., 2014
Gilthead sea bream	770.3	259,783	2,965	13.35	Bargelloni et al., unpublished
Turbot	544.2	16,463	33,058	31.2	Figueras et al. 2016

In this study, we set out to examine how variation in multispecific ddRAD sequence datasets and the application and quality of available reference genome sequences affect the consistency and accuracy of resulting data, generated through commonly used analytical approaches. The laboratory and bioinformatic pipeline used to generate the ddRAD datasets followed standard published methods (see below) and has been summarized in a flowchart (Figure 1). The performance of the ddRAD pipeline was evaluated at different stages in order to investigate the causes and effects of variation in individual sample coverage, RAD-tag sequence length and application and quality of reference genomes on the eventual accuracy and error rates of individual genotyping. We specifically addressed the following questions:

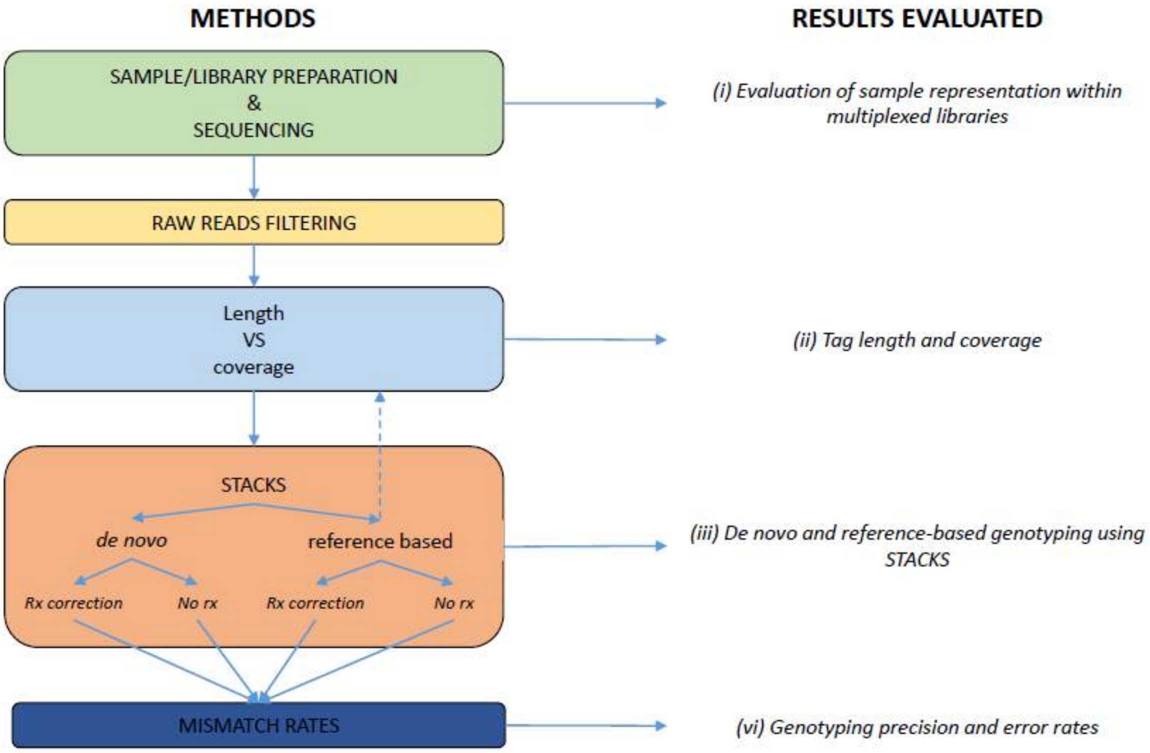
(i) *Evaluation of sample representation within multiplexed libraries*. What is the typical variation in terms of number of raw reads per sample when multiple individuals (144 in our case) are multiplexed in a single sequencing lane?

(ii) *Tag length and coverage*. Is there any difference between the expected and observed length of analyzed tags? Does any relationship exist between tag length and depth of coverage?

(iii) *De novo and reference-based genotyping using STACKS.* What is the effect of different clustering approaches (e.g. *de novo* vs reference-based, *a posteriori* genotyping correction) on the number of markers identified?

(iv) *Genotyping precision and error rates*. What are the effects of the variables described above on the number of mismatches between technical replicates?

Figure 1 Flowchart of the analysis pipeline followed in this study, indicating the results evaluated in order to understand the performances of ddRAD sequencing technique


METHODS
RESULTS EVALUATED
SAMPLE/LIBRARY PREPARATION & SEQUENCING
(i) Evaluation of sample representation within multiplexed libraries
RAW READS FILTERING
Length VS coverage
(ii) Tag length and coverage
STACKS
de novo
reference based
Rx correction
No rx
Rx correction
No rx
(iii) De novo and reference-based genotyping using STACKS
MISMATCH RATES
(vi) Genotyping precision and error rates

Based on these insights we suggest approaches which can help to mitigate the identified risks of error in ddRAD analysis.

Material and Methods

Samples and library preparation

Specimens of European sea bass, gilthead sea bream and turbot were collected in the context of the European Union’s FP7 funded project ‘AQUATRACE’ (KBBE 311920). The entire sample set included more than 5,581 specimens (2128 European sea bass, 2156 gilthead sea bream and 1297 turbot) from the species’ distribution range, some of which were collected specifically for the project (years 2013-2014, from now on referred to as “fresh” samples), while others had been collected earlier (“archived” samples). For fresh samples, fin clips were preserved separately in 95% ethanol at 4°C until genomic DNA (gDNA) extraction. Samples were extracted either with Invisorb® DNA tissue HTS 96 kit (Stratec biomedical) or with a standard NaCl isopropanol precipitation protocol [43]. Extracted DNA samples were then classified as “high”, “mid” or “low” quality according to the level of degradation assessed with agarose gel electrophoresis (see Supplementary material).

The same ddRAD protocol, with minor modifications, was used for the three species. The library preparation followed the original guidelines of Peterson et al. (2012) [42], with some modifications that facilitate the screening of large number of individuals (see Supplementary Material for details), and was carried out in three different laboratories within the AquaTrace consortium, each focusing on a single species: the sea bass at the Laboratory of Biodiversity and Evolutionary Genomics, University of Leuven, sea bream at the Department of Compared Biomedicine and Food Science, University of Padova and turbot at the Departmento de Zoología,Genética y Antropología Física, Universidade de Santiago de Compostela. To promote a common standardized approach, staff from the three laboratories completed a hands-on training course in library preparation at the Institute of Aquaculture, Stirling, where the modified ddRAD protocol originated. Multiple ddRAD libraries were prepared for each species (sea bream, n=14; sea bass n=14; turbot, n=9). Each library comprised 144 samples, and in all the libraries the same three or four control samples for each species were included, to enable cross-library comparisons and mismatch rates between replicates to be assessed. In particular, four sea bream specimens (SAC3, SAC4, SAC5 and SAC6 from Sardinia, Italy); three sea bass specimens (DLTY40, from the Central Mediterranean Sea; DLM44, from the Atlantic and DLFF1, from a European broodstock); and four turbot specimens (SMFF1, SMFF2 and SMFF3 from a Spanish broodstock; SMNS32 from North Sea’s wild population) were used.

Sequence data analysis – standard pipeline

The following approach to sequence data analysis was used for all datasets as the basis for subsequent comparative analysis. Raw data were filtered to retain only high quality reads, using STACKS’ [58,59] *process_radtags* program, which allows simultaneous quality filtering and sample demultiplexing. After barcode removal (5-7 bases), the sequences were 3’ end-trimmed to a standard 90 nucleotides length. Each read was then analyzed to assess sequence quality using default parameters. Briefly, a 3-base sliding window (STACKS’ option –w) was used to parse each read and where the average phred score of three consecutive bases was lower than 20 (STACKS' option –s) the entire read was discarded.

STACKS was also used for clustering reads and for SNP discovery, following standard *de novo* and reference based pipelines, well described in the program website (http://catchenlab.life.illinois.edu/stacks/). In our case parameter –m (minimum number of reads to call a stacks) was set to four and –M (maximum number of mismatches between reads to be considered as part of the same cluster) was set to five. For the *de novo* approach, reads from primer P1 were concatenated with the reverse complement sequence of reads from primer P2, obtaining 180 bp *pseudo-contigs*. This approach was used to create longer sequence tags which reduces the risk of over-merging (i.e. clustering together tags coming from different genomic regions) by keeping the information about relative proximity of Read 1 and Read 2. As an added benefit, this approach allowed to be fully aware of linkage issues. Since reference based approach require reads to be mapped against a reference, we used the software package BOWTIE [60], considering read pairing in the alignment process. We kept only read pairs that matched a single genomic position.

When building the RAD-tag catalog a maximum number of five mismatches between tags was set. For the reference based approach, clustering was based on mapping position. Consensus genotypes were called by *sstacks* (with minor allele frequency (MAF) <0.01 homozygote is called; with MAF between 0.01 and 0.1 the genotype is considered 'unknown'; with MAF>0.1 a heterozygote is called). *Rxstacks,* STACKS' component that corrects genotypes on the basis of population information, was also implemented for comparison. Finally, we used the algorithm implemented in STACKS' *populations* step to retain only individual loci represented with at least 10 reads per individual sample and genotyped in at least 80% of the samples analyzed. This is an important step when the genotypes of multiple individuals need to be compared, as only shared loci provide useful information for genetic analysis.

Analysis of the pipeline

Here, we describe the methods used to assess the pipeline based on the four issues described in the Introduction (Figure 1).

(i) Evaluation of sample representation within multiplexed libraries

Considering the number of samples multiplexed and the average output of the sequencing platform/chemistry (180 M reads), approximately 1.3 M reads per sample are theoretically expected. However, even if initial DNA quantification is accurate and input DNA is equal among samples, subsequent library preparation steps may alter individual representation within the library resulting in variability in inter-sample sequencing effort. To investigate sample read homogeneity in libraries with up to 144 pooled individuals, we first established a threshold number of reads per sample against which to filter individual sample data. A threshold of 150 k reads was chosen as a minimum to accept an individual sample for downstream data processing, based on an expected number of 7,000 stacks per sample (estimated from *in-silico* analysis) and an average coverage of 20x. This threshold was used in the analysis of the sequencing output for all available ddRAD data including more than 5,000 samples.

To identify the factors correlated with fewer reads, we tested the correlation between number of reads (above or below the threshold) and variables such as "DNA quality", whether a sample was "fresh" or "archived", "individual sample collector" (i.e. the project partner that collected the sample), and "index barcode" (different length/sequence barcodes could perform differently in the amplification or sequencing by synthesis steps), testing the effect of each variable under a Generalized Linear Model (GLM), as implemented in R 3.2.3 library function Rcmdr [61,62]. Chi-squared tests were applied to check association between tested variables.

For the analysis described further on, only replicate samples with sufficient read numbers were used.

(ii) Tag length and coverage

To understand whether the length of the RAD-tags observed corresponded to the expected length and to investigate association between tag length and coverage, we extracted fragment length and DNA sequences of ddRAD-tags from BOWTIE alignment results. Data on coverage depth was extracted for each single locus of each sample, separately. To allow comparison between samples with different average coverage, standardized coverage depth was obtained by dividing locus specific values by the average coverage across all loci for each sample. Similarly, when comparing the distribution of the number of tags with different lengths, 10 bp bins were used and the relative number of tags was calculated dividing the number of tags of a certain length bin by the average number of tags across all the bins. A Wilcoxon signed-rank test, as implemented in R 3.2.3 library Rcmdr, was used to test for differences between distributions from the three datasets.

(iii) De novo and reference-based genotyping using STACKS

In order to understand how the alignment to a reference genome influences SNP genotyping, we obtained individual genotypes using both *de novo* and reference-based analysis in STACKS. Since we expected *de novo* approach to detect also tags that are not contained in the reference genome, we wanted to evaluate the amount of *de novo* tags that could be found in the genome. In order to do this, RAD-tags resulting from *de novo* analysis (180 bp long) were subsequently split in two (in order to reconstitute the original 90 bp tags) and mapped against the reference genome using BOWTIE, with the same parameters used while aligning reads for reference based analysis. Under both *de novo* and reference-based analysis, results were compared with and without the final step in *rxstacks*. Statistical differences between approaches were tested with chi-squared tests.

(iv) Genotyping precision and error rates

To investigate the level of reproducibility across different bioinformatic approaches we examined the level of consistency among scored SNP genotypes within the sets of 11 to 14 replicated samples for each species. The most frequent genotypes were considered as the “correct” ones, and mismatches were counted for each locus in each sample to estimate genotyping error.

When comparing results from different approaches, statistical significance was tested using either on-line applications (e.g. Kruskall-Wallis: http://vassarstats.net) or the Rcmdr library for R 3.2.3. A first global analysis was carried out to assess the effect of several parameters (“coverage”, “genome reference” mapping, “*rxstacks* correction”, percentage of high-quality reads) on mismatch rate across the entire dataset. Individual mismatch rates were classified either as a binary outcome (0 for values lower than the overall median mismatch rate, 1 for those equal or greater), or grouped into quartiles for a finer evaluation of the effects of different explanatory factors. In both cases, either a Generalized Linear Model (used with binary outcome) or Ordinal Linear Regression (used with samples grouped into quartiles) were used to detect the most influential variables. The same statistical approach was then implemented, within each dataset, across single specimens, to look more into detail at individual-specific features that could affect genotyping quality and to avoid dataset-specific biases and errors. This additional analysis was possible thanks to the large number of replicates available for each species (three to four specimens replicated nine to 15 times) and the standardization of library preparation technique and bioinformatics protocols. Lastly, mismatch rates were analyzed across loci within single individuals, to check for association with locus-specific coverage.

Results

The first part of the study addressed the loss of analytical power in terms of number of samples filtered due to unequal representation of individuals within libraries; it was based on a data set of more than 5,581 samples, in which the replicate individuals were included.

(i) Evaluation of sample representation within multiplexed libraries

After quality filter was applied, an average of 74.5%±10.8% reads remained available for further analysis. After filtering, the average number of high quality reads was similar across species, 687,426±447,701 in European sea bass, 614,099±406,018 in gilthead sea bream and 610,703±707,152 in turbot. As indicated by high values of standard deviation (in particular for turbot), variation among individuals within species was very high. In fact, 129 samples were represented by less than 1,000 reads and three samples had more than 5,000,000 reads in the three species. Using the threshold of 150,000 raw reads, 6.8% of sea bass samples, 8.1% of sea bream samples and 16.0% of turbot samples were discarded. Regression analysis indicated that better quality DNA resulted in higher number of high quality reads (t= -11.4 p<0.001); similarly, “fresh” samples had a higher amount of high quality reads than “archived samples” (t= -3.1 p<0.005). “DNA quality” of individual samples was neither significantly associated with species (X^2=4.6 p>0.25), nor with fresh/archived condition (X^2=3.1 p>0.25).

After filtering and quality checking, the final number of replicated samples available for downstream analysis was 111: 43 sea bream samples (11 replicates for SAC3, 11 for SAC4, 10 for SAC5 and 11 for SAC6) genotyped across 11 independent libraries, 34 sea bass samples (5 replicates for DLTY_40, 14 for DLT_1 and 15 for DLM_44) genotyped across 15 libraries and 34 turbot samples (9 replicates for SMFF1, 8 for SMFF2, 9 for SMFF3 and 8 for SMNS32) genotyped across 9 libraries.

(ii) Tag length and coverage

On average across species, 78.4% of the reads were successfully mapped on the reference genomes and mapping rates ranged from 71.3% uniquely mapped reads in sea bream to 85.4% in sea bass.

Average fragment length across datasets was 288.9±110.5 bp. Most of the tags (79.5%) were 100-380 bp. In addition, substantial fractions (21.1% sea bream, 24.5% sea bass, 15.9% turbot) of analyzed RAD-tags were shorter than 190 bp (the minimum size expected according to the library construction protocol) (Figure 2).

Figure 2: Graph of fragment length vs number of fragments in European sea bass (square), gilthead sea bream (diamond) and turbot (triangle). The graph is based on the reference-based analysis, as only for this it was possible to obtain information about fragments’ length. Dash vertical line indicates the limit under which pair-end tags present overlapping between Read 1 and Read 2.

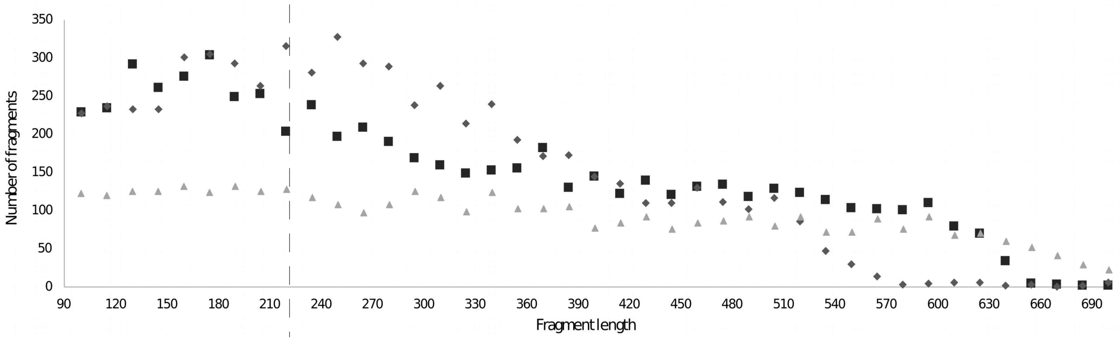
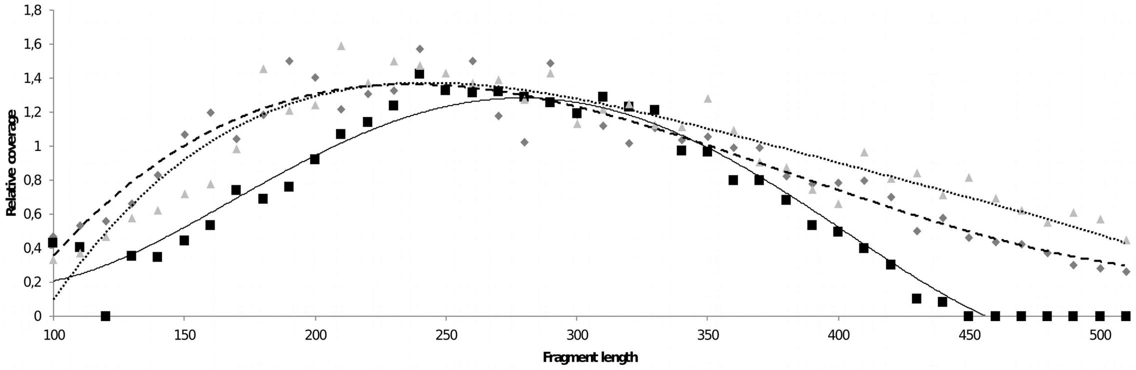


Figure 3: Graph of fragment length vs coverage depth in European sea bass (squares), gilthead sea bream (diamonds) and turbot (triangles). The graph is based on the reference-based analysis, as only here it was possible to obtain information about fragments' length. Coverage is expressed as relative to specific average coverage, in order to account for difference between species in average coverage depth. Trend lines were calculated as polynomial, third order for sea bass (solid line, R2=0.70), sea bream (dash, R2=0.93), turbot (point, R2=0.89)



Significant (p<0.01) positive linear correlations between length and coverage were also found for fragments in the range from 100 to 250 bp (Spearman rho=0.903 in sea bream, 0.957 in sea bass and 0.918 in turbot). Fragments longer than 250 bp showed significant (p<0.01) negative linear correlation between length and coverage (Spearman rho=-0.969 in sea bream; -0.968 in sea bass, -0.952 in turbot). No significant correlation between GC content of fragments and coverage depth was observed.

(iii) De novo and reference-based genotyping using STACKS

The number of independent RAD-tags identified varied depending on the approach. In all cases the number of tags found by the reference genome-based approach was much lower than that found with the *de novo* approach (up to 5.5 times, in turbot dataset) (Table 2).

Table 2 Summary of the STACKS' analyses on European sea bass, gilthead sea bream and turbot using de novo and reference based approaches. Application of the correction sub-program rxstacks is indicated under column 'Correction'. SNP frequency is calculated as the number of base pairs analyzed (180 bp x number of tags for the de novo approach; 90 bp x number of tags for the reference based approach) and the SNPs detected. 'Tags 80%' indicates the number of tags after filtering for those shared by at least 80% of individuals analyzed.

Species	Type of analysis	Correction	Tags	SNPs	SNP freq	Tags 80%	Average coverage
European sea bass	*de novo*	No correction	19,672	16,342	216.7	3,246	111.0 ± 65.9
		rxstacks	19,595	15,612	225.9	1,347	101.51 ± 59.6
	reference based	No correction	13,458	3,013	402.0	4,913	156.8 ± 94.3
		rxstacks	13,379	3,007	400.4	1,764	153.9 ± 92.9
Gilthead sea bream	*de novo*	No correction	25,322	39,842	114.4	3,913	151.5 ± 72.0
		rxstacks	24,257	31,790	137.3	2,353	89.3 ± 48.3
	reference based	No correction	13,659	5,161	238.2	7,091	247.7 ± 126.4
		rxstacks	12,293	4,388	252.1	5,796	109.9 ± 52.6
Turbot	*de novo*	No correction	58,171	26,635	393.1	1,674	272.1 ± 226.8
		rxstacks	56,320	21,582	469.7	1,631	157.3 ± 150.2
	reference based	No correction	8,887	2,530	316.1	4,175	700.9 ± 544.6
		rxstacks	5,595	1,440	346.7	4,106	255.4 ± 230.3

However, when a filter was applied to retain only tags shared by at least 80% of samples analyzed, higher proportion was retained for reference-based analysis (on average 44.9%±19.7%) than *de novo* analysis (on average 9.1%±6.0%). This made that in most cases the final number of retained tags was higher using the reference-based approach. Similarly, a higher number of SNPs was observed in the reference-based approach after filtering. The application of the genotype correction implemented in *rxstacks* reduced the number of tags by different extents: a minimum of 63% of total tags were retained in the turbot reference-based analysis and a maximum of 99.6% in the sea bass *de novo* analysis. The proportion of SNPs retained was comparable, ranging from 56.9% to 99.8% in turbot (reference-based) and sea bass (*de-novo*), respectively. Mapping tags from *de novo* analysis against the reference genomes produced 11,121 matches for sea bass (28.3% of *de novo* RAD tags); 11,650 for sea bream (23.0% of *de novo* RAD tags) and 7,889 for turbot (6.8% of *de novo* RAD tags). This figures are in agreement with the relative length of the genomes utilized (Table 1), while the lower than expected difference between sea bass and sea bream results can be explained by the lower quality of the bream assembly, as indicated by the N50 value.

(iv) Genotyping precision and error rates

Our analysis suggested that "*rxstacks* correction" and "coverage" significantly affected the level of accuracy in the comparison of different approaches, regardless the species. In particular, lower mismatch rate were recorded when *rxstacks* was implemented and when coverage depth per sample was higher. However, variation in mismatch rates were found between different species datasets (Table 3); they were apparently linked with differences in species-specific coverage, which varied significantly both for *de novo* RAD-tags (Kruskall-Wallis test, H=15.27 p<0.001) and reference-based ones (Kruskall-Wallis test, H=30.74 p<0.0001).

Table 3 Summary of mismatch analysis on European sea bass, gilthead sea bream and turbot using de novo and reference based approaches. Values are given as average or median percentage of genotypes that differ from the consensus (most frequently recorded) genotype over the total number of genotypes analyzed (number of individuals analyzed x number of SNPs). Application of correction subroutine rxstacks is indicated under column 'Correction'.

Species	Type of analysis	Correction	Average % of mismatches	Median % of mismatches
Sea bass	*de novo*	No correction	2.9	0.9
		rxstacks	2.9	0.9
	reference based	No correction	1.9	0.5
		rxstacks	1.7	0.4
Sea bream	*de novo*	No correction	0.7	0.3
		rxstacks	1.3	0.3
	reference based	No correction	0.2	0.2
		rxstacks	0.1	0.1
Turbot	*de novo*	No correction	0.5	0.2
		rxstacks	0.6	0.1
	reference based	No correction	0.4	0.2
		rxstacks	0.3	0.1

To overcome biases linked to species-specific differences, more specific tests were carried out within single datasets. In fact, additional factors were found to be significantly affecting mismatch rate. In addition to "*rxstacks* correction", also "library", "reference-mapping" and "sample" (only in the turbot database) showed significant correlations. "Coverage" showed a significant correlation in two out of three datasets (sea bream ($p<0.05$) and turbot ($p<0.001$)). At the individual level (i.e. across loci) no significant correlation between mismatch and coverage across loci was found.

Discussion

The aim of the present work was to quantify the actual amount of genetic information that can be obtained with ddRAD approach, net of information loss due to reasons presented in the introduction; and to evaluate the performance of different bioinformatics approaches on the number of markers detected and the precision of the genotype calling. The use of large datasets of marine fish species and the application of the same approaches as those used in real case studies make our results informative on the practical application of this technique.

(i) Evaluation of sample representation within multiplexed libraries

The first step in which genotyping information is lost is quality filtering, which is fundamental in order to get reliable results with NGS analysis [63,64]. Up to 20% reads can be lost here. In STACKS, quality filter is based on the average phred quality of a portion of the analyzed sequence, that can cause the entire read to be discarded if average quality is below a certain threshold. Parameters can be set in order to have more reads passing the filter, but this increases the risk of including error-containing reads in the subsequent analysis. Similarly, trimming a certain number of bases at the very end of each read (usually characterized by lower quality) can help rescue more sequences. On the other hand, this procedure causes additional loss of genetic information. In our specific case, the total amount of information lost ranged up to 28%, considering sequence trimming and quality filter. To reduce this loss, the best approach would be to implement the base call quality within the marker specific significance statistics, or trim only bases affected by low quality instead of the entire read.

One of the main advantages of RAD techniques is the possibility of multiplexing many individuals in the same sequencing run thanks to individual sample barcoding. However, as the number of multiplexed individual samples increases, the chance to have poorly represented samples increases as well [42,47], causing lower coverage and in the worst case, too few reliably genotyped or false homozygote excess for a number of individuals. In particular, the combination of samples at different quality/concentration, rather than the quality of single samples is the influencing variable and even using the same starting DNA result might vary in relation with the other samples genotyped in the same library. The threshold at 150,000 raw reads used here is much lower than the expected average number of reads per individual (1.3 millions) and may not be appropriate for other species. In fact, it should be set taking into consideration the number of expected tag and the desired average coverage depth. However, "losing" a certain amount of samples (up to 16% in our case) needs to be considered when planning a ddRAD sequencing project, even when significant effort was given to equalize DNA input under library preparation.

Not surprisingly, DNA quality was a good predictor of poorly performing samples. Gel-based quality analysis essentially reflects the level of DNA degradation, that can be caused by many factors that act before or after extraction. In our specific case, pre-extraction factors are probably the most relevant, as extraction and post-extraction protocols were the same for all the samples. Ethanol has been recognized as a good media for long term tissue storage [65,66], and it is easily available and not hazardous. Nevertheless, Seutin, White, and Boag (1991)[67] reported that ethanol conservation can decrease DNA yield and cause significant degradation to the extracted DNA, that can be reduced by keeping samples refrigerated as soon as possible after sampling. DNA from long-term stored specimens might have some additional features reducing the efficiency in library preparation. Therefore, when selecting the DNA samples to be pooled as part of the same library, it is advisable to avoid mixing samples of heterogeneous DNA quality as well as mixing "fresh" with "archived" specimens. When this is not possible (e.g. for those projects that use only one or few sequencing pools), an upward correction for the starting amount of DNA of poor quality samples and DNA from “archived” samples might be considered. However, further analysis is necessary to better understand how this procedure should be applied.

(ii) Tag length and coverage

Accuracy and consistency in size selection is not easily achievable, but fragment size distribution was not significantly different across species in our study. From this point of view, the period of training of the personnel proved to be effective in order to have consistent results. Nevertheless, loci shorter than 190 bp were retained in our analysis, which was unexpected considering that size selection step was implemented. Indeed, low accuracy has been documented in particular for manual *vs* automated gel band extraction [68]. A similar result was found by DaCosta e Sorenson (2014)[51], who recovered loci down to a length of 10 bp. In our case, the number of fragments below the 100 bp length threshold was extremely low. This was probably achieved by the purification steps performed at the very end of the library preparation protocol, which eliminated most fragments shorter than 200 bp, that translates into RAD tags longer than 75 bp, after removing adapters. It is important to notice that, considering the 100 pair end sequencing protocol used, all the analyzed fragments shorter than 190 bp are affected by read1-read2 overlapping of the final regions, potentially causing SNP duplication, redundant data and a waste of sequencing effort that further lower the actual power of ddRAD technique. Improvement in size selection step is fundamental to optimize the performance of the ddRAD technique.

Davey et al. (2013)[44], using data from a *Caenorhabditis elegans* RAD library, found a strong positive correlation between fragment length and coverage depth. In other published ddRAD studies, such as DaCosta e Sorenson (2014)[51], the relationship between coverage and length was similar to our work. Tags with different lengths show variable coverage within individual samples. This means that additional care should be taken when multiplex size is calculated, in order to achieve a desired minimum depth of coverage across loci. According to our results, loci in the shortest and longest length range will be underrepresented if coverage was calculated just by dividing the number of individual reads by the number of expected loci. Upward correction in the number of reads per individual should be applied to obtain minimum coverage also for loci in short and long fragments.

(iii) De novo and reference-based genotyping using STACKS

The possibility to use RAD techniques in species without genomic resources (i.e. *de novo* approach) has been highlighted as one of the method's biggest advantages [69,70]. However, we showed that using a reference genome improves RAD genotyping performance, i.e. better precision and higher number of shared markers. With reference based approach, only reads correctly mapped against the genome are used. Hence, the quality of reference-based analysis is also dependent on the quality of the assembly used. In particular, N50 seemed to better predict mapping percentage compared to average contig length. Turbot shows the longest average contig length, but ranked second in terms of positive mapping matches, in agreement with N50 ranking (Table 2). J. Catchen et al. (2013)[59] showed that in three-spined stickleback *de novo* approach yielded a higher number of tags (42,300) than the reference based one (37,600), mostly due to loss of loci that could not be mapped against the reference genome (>4,700). Likewise, in our analysis, using the genome as a reference returned a lower number of tags compared to the *de novo* approach (Table 3). In any case, the number of *de novo*-based tags that mapped correctly to the reference genome was in good agreement with the number of tags identified by the reference based analysis. The larger number of *de novo* ddRAD tags might then be explained in part by the incomplete mapping of reads against the reference genome as in the case of three-spined stickleback. A second possibility is that a fraction of tags, which STACKS identified as separate "loci" in the *de novo* analysis, is likely represented by divergent alleles of the same locus. However, STACKS controls for such phenomenon through the –M parameter and, in the present study, a less conservative value (–M=5) than the default one (–M=2) was set for all species. More likely, *de-novo* approach might include some "spurious" loci at individual level. In support of this hypothesis, a filter that exclude loci shared by less than 80% of individuals, filter out most of *de-novo* loci. The origin of these tags is difficult to find but some sources can be the presence of exogenous DNA from viral or bacterial contaminants or sequencing errors introduced with amplification in library preparation and sequencing steps. While we cannot exclude that these sequences can provide useful information or could be used as dominant markers [71](Fu et al, 2013), we recognize that they need to be studied more in detail to understand their origin and whether they can have bad effects on certain downstream applications (i.e. those requiring the use of markers shared by a percentage of individuals). Without deeper knowledge of the origin of these sequences, it is therefore advisable to use the above mentioned filters to reduce source of bias in the final filtered datasets. In general, even if in the form of a draft, a reference genome should allow more efficient SNP detection.

(iv) Genotyping precision and error rates

Genotyping reproducibility across technical replicates is one of the most important test to evaluate genotyping methods. A first analysis on over 100 replicates over the three species datasets, showed that "coverage" represented a significant explanatory variable for differences in mismatch rates. In fact, sea bass' technical replicates, which were characterized by a significantly lower coverage, also showed lower precision than the two other species. The effect of reduced coverage also appears to be affecting samples characterized by a high DNA quality.

(Davey et al. (2013)[44] suggested at least 30x average coverage depth for reference genome-based analysis and at least 60x coverage depth for *de novo* analysis in order to have reliable results. In the present study, the average coverage for all the three species was higher than that suggested, but also the variability across loci was high (36x-386x in sea bass, 31x-2840x in sea bream and 69x-2731x in turbot), which might influence the outcome in term of mismatch rates. However, we couldn't find any significant correlation between mismatch rate and coverage per locus when analyzing results within individual samples.

The same analysis showed that the SNPs in the reference-based tags are more consistently genotyped than *de novo* ones in both turbot and sea bream. The positive effect of using a reference genome on genotyping reproducibility is an additional one to the advantage of avoiding inflation of tag number described above. More reproducible genotypes are also obtained when *a posteriori* genotype correction was implemented. In our opinion, even if both approaches (reference-based analysis and *a posteriori* correction) come at a price as the total number of tags/SNPs analyzed gets reduced, they should be used to obtain more reliable data.

Conclusions

Application of new genotyping techniques is rapidly increasing as they potentially allow more accurate, easier and less expensive population genetic analysis of any species. However, several issues might affect the quality of the results. In the present study, it was demonstrated that some factors, i.e. DNA fragmentation and archived-fresh samples, affect the throughput in terms of percentage and absolute number of high quality sequence reads in ddRAD datasets. Similarly, actual fragment length and coverage can differ from expectations, leading to redundant loci and loci with too low coverage. Although RAD has been proven to be applicable on non-model species, the use of a preliminary draft genome sequence increase genotyping performance enabling to obtain higher numbers of loci shared between multiplexed individuals. Finally, we showed the critical importance of introducing replicate individuals among the samples in order to assess the performance of the approach used. Our results are useful for setting up genotyping project and for considering the features that can affect genotyping throughput and precision.

Acknowledgements

The research leading to these results has received funding from the European Community’s Seventh Framework Programme under agreement no. KBBE-311920 (AQUATRACE). Samples used in this study were collected by the AQUATRACE partners: Laboratory of Biodiversity and Evolutionary Genomics (LBEG), University of Leuven (European sea bass); Departmento de Genética, Universidad de Santiago de Compostela (turbot); Sabina De Innocentiis, ISPRA (gilthead sea bream). JH benefited from a PhD scholarship from Flanders Innovation & Entrepreneurship (IWT)

SUPPLEMENTARY MATERIAL

Detailed library preparation protocol

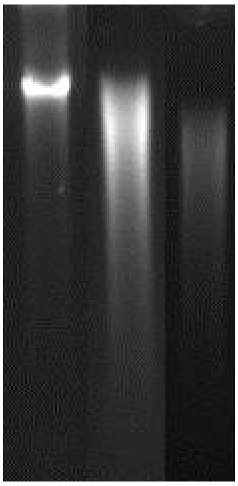
Each group used biochemical consumables from the same manufacturers and were supplied with custom barcoded ddRAD adapters mixes, sourced from the same original stocks prepared at the Institute of Aquaculture, Stirling.

The original protocol of Peterson et al. (2012)[42] involved processing each sample separately (i.e. restriction digestion, adapter ligation, fragment size selection, PCR amplification and purification, quantitation) prior to pooling into a single library for sequencing. A modified protocol (described in detail elsewhere[72,73]), which was more convenient for screening large numbers of individuals, was used for this project. The methodology allowed for pooling of samples after the adapter ligation step, which greatly reduced the number of manipulations required, ensured consistent size selection within libraries and reduced construction time to two to three working days. Library preparation began with basic qualitative and quantitative assessment of extracted DNA samples. DNA quality was evaluated by gel electrophoresis (0.8% agarose 0.5x TAE) and concentration was accurately measured by fluorimetry with each sample being finally diluted to 7 ng/μL in 5 mM Tris pH 8.5. For a library (144 samples), individual DNA samples (21 ng) were first simultaneously digested with *Sbf*I (recognition site CCTGCA'GG) and *Sph*I (recognition site GCATG'C) restriction enzymes. An adapter mix comprising individual-specific barcoded combinations of P1 (*Sbf*I-compatible) and P2 (*Sph*I-compatible) adapters (compatible with Illumina sequencing chemistry) were then added / ligated. Adapters were designed such that adapter–genomic DNA ligations did not reconstitute RE sites, residual RE activity limiting concatemerization of genomic fragments. Each adapter included an inline five- or seven-base barcode, allowing for post-sequencing identification of individuals (P1-P2 combinatorial barcoding). The ligation reactions were terminated by heat inactivation and all 144 samples combined in a single pool. Following column purification of the pooled sample, DNA fragments in the range of 320 bp to 590 bp were size selected by agarose gel electrophoresis, followed by gel-based column purification. The eluted size-selected DNA template was then PCR amplified (14 cycles, 400 uL volume), column purified down to a 50 uL volume and then subjected to a further clean-up using an equal volume of AMPure magnetic beads (Perkin-Elmer, UK) (used in sea bream and turbot), to maximize removal of small fragments (less than ca. 200 bp). The final library was eluted in c.20 µL10 mM Tris pH 8.5.

Libraries were sequenced on Illumina HiSeq 2500 sequencers with pair-end (PE) 100 base option to allow sequencing of both barcodes at the Genomics Core of the University of Leuven, Belgium (sea bass and sea bream) and BMR S.r.l, Padova, Italy (turbot).

DNA quality from agarose gel electrophoresis

Example of "high" (a), "mid" (b) and "low" (c) quality DNA taken from agarose gel of DNA samples used in the study



Population genetic approaches are applied to study dolphinfish population from the Mediterranean Sea. The study highlighted the presence of loci linked to sex determination and led to the hypothesis of presence of sexual chromosome in the species.

RAD SNP markers as a tool for conservation of dolphinfish *Coryphaena hippurus* in the Mediterranean Sea: identification of subtle genetic structure and assessment of populations sex-ratios

Francesco Maroso[a,1], Rafaella Franch[a,1], Giulia Dalla Rovere[a], Marco Arculeo[b] and Luca Bargelloni[a]

[a] Department of Compared Biomedicine and Food Science, University of Padova, viale dell’Università, 16, 35020 Legnaro, ITALY
[b] Dipartimento STEBICEF, Università di Palermo, via Archirafi, 18, 90123 Palermo, ITALY

Abstract

Dolphinfish is an important fish species for both commercial and sport fishing, but so far limited information is available on genetic variability and pattern of differentiation of dolphinfish populations in the Mediterranean basin. Recently developed techniques allow genome-wide identification of genetic markers for better understanding of population structure in species with limited genome information. Using restriction-site associated DNA analysis we successfully genotyped 140 individuals of dolphinfish from eight locations in the Mediterranean Sea at 3,324 SNP loci. We identified 311 sex-related loci that were used to assess sex-ratio in dolphinfish populations. In addition, we identified a weak signature of genetic differentiation of the population closer to Gibraltar Strait in comparison to other Mediterranean populations, which might be related to introgression of individuals from Atlantic. No further genetic differentiation could be detected in the other populations sampled, as expected considering the known highly mobility of the species. The results obtained improve our knowledge of the species and can help managing dolphinfish stock in the future.

Keywords: 2bRAD; genetic differentiation; outliers; sex determination markers

Introduction

Dolphinfish, *Coryphaena hippurus* [74], is an important target species for artisanal, recreational, and commercial fisheries. It is considered a mid-trophic level pelagic fish [75], with high potential for dispersal. Dolphinfish are found in tropical and subtropical waters, including the Mediterranean Sea. It is considered a fast growing species with an estimated maximum longevity of four years in Caribbean Sea [76,77], even if in Mediterranean no individuals exceeding two years have been found [78]. It is available for fishing during the summer season [79], with sport fishing targeting larger individuals, and commercial fishing aiming at the juvenile stage (25-60 cm fork length (FL) which corresponds to individuals aged between 2 and 8 months [80.] While the bad effects of fishing for juveniles are known (e.g. reduction of future yield and recruitment for the species), there is no specific minimum size regulation for this species (though some exceptions exist, as for example Sardinia, with 60 cm minimum size, www.sardegnaambiente.it/documenti/19_4_20080215151247.pdf).

Specifically, for the dolphinfish this can be a major issue, considering the different behavior of fish of different age. It is indeed known that under fish aggregating devices (FADs), used by commercial and sometimes also by recreational fishermen mostly female and young males are found, while adult males prefer open waters as they move between female dominated rafts [81]. Thus, fishing around FADs could lead to alteration in the sex ratio at particular life stages. Sexual dimorphism is present but morphological differences arise only when sexual maturity is reached, usually from May to October of the first year, at 60 cm FL [78], in both sexes. Sexual dimorphism is evident in large individuals as males develop a typical bulging squared-off forehead, which is not present in females [79]. For younger specimens, sex can only be determined by histological analysis of gonads, which is a time consuming and often not easy task.

Restriction enzyme Associated DNA (RAD) refers to a family of genotyping techniques that use the cutting activity of restriction enzyme and selection of resulting fragments to obtain a reduced representation of a specimen’s DNA that will subsequently be sequenced. Indeed, for the aims of many genetic approaches (e.g. population genetic studies) the information provided by only a small portion of the entire genome is sufficient, and requires less genotyping effort to achieve enough depth of coverage for reliable analysis. These library preparation techniques, combined with high read throughput (up to 1,5 terabases for the latest Illumina technologies machines), of Next Generation Sequencing (NGS) machines, allow the simultaneous analysis of multiple individuals at the same time (which is typical for population genetic studies) at reduced cost. One of the approaches derived from the original RAD is the 2bRAD [41], which exploit the cutting activity of type IIB restriction enzymes to cut specific site in the genome and retrieve uniform length fragments (centered at the enzyme’s recognition sequence), shared by all the individuals analyzed. Among the advantages of this particular RAD technique, the relatively simple laboratory approach (i.e. no need for shearing, no agarose gel size selection…) and its flexibility are the two most frequently addressed. In particular, the availability of different combinations of enzymes-adaptors can be used to trim the number of markers analyzed according to the needs of any specific study or to the species addressed. From a bioinformatic point of view, the homogeneous length of the fragments and the presence of the restriction enzyme recognition site in the center of the sequences are advantages for clustering steps, especially in species lacking reference genome resources. This approach has already been proved effective in fish population genetic analysis (e.g. in tuna by Pecoraro et al. (2016)[82]), and allowed the identification of previously undiscovered population genetic structure.

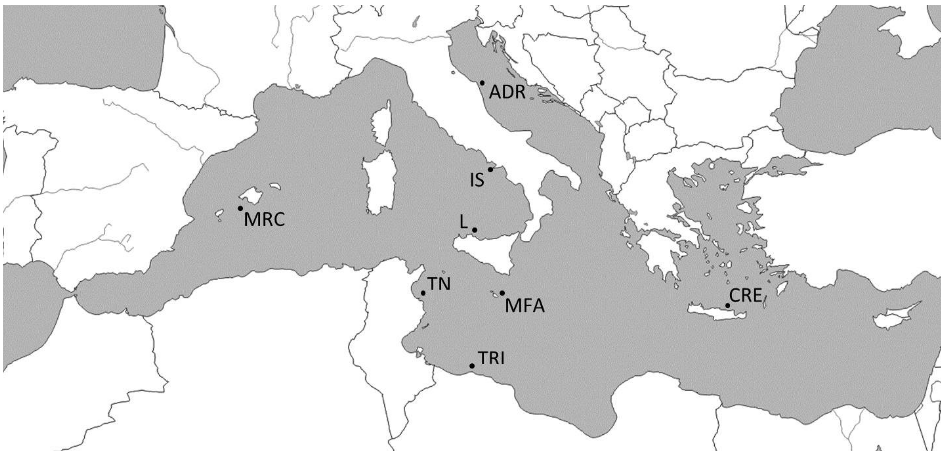
In the present study, for the first time, a large set of SNP markers was identified with 2bRAD and used to study the population genetic variability of dolphinfish, providing a robust tool for determining sex and showing preliminary evidence for subtle genetic divergence within the Mediterranean basin despite the large potential for dispersal of the species. This work was carried out in the framework of the Ritmare project (http://www.ritmare.it/en/), the Italian flagship research project on marine biology for the period 2012-2015.

Material and Methods

Sampling design, libraries preparation and sequencing

Fin clips from 169 juvenile dolphinfish (FL range 36-64 cm) from eight different landing localities across the Mediterranean Sea (Figure 1) were collected.

Figure 1 Sampling locations. Mediterranean sites surveyed in the present study (MRC=Spain, IS= Ischia, L= Porticello, TN= Tunisia, MFA= Malta, TRI= Libya, ADR= Ancona, CRE= Crete)



Genomic DNA (gDNA) was extracted using either commercial kits (Invisorb® Spin Tissue Mini Kit (Invitek, STRATEC Biomedical, 242 Germany) and Real Genomics Tissue DNA Extraction kit (RBC Bioscience, Taiwan)) or the SSTNE buffer, a modified TNE buffer added of spermidine and spermine [83].

Genomic libraries were constructed following the 2bRAD protocol with minor modifications. In brief, gDNA (300 ng) was digested with 2 U of the enzyme CspCI (New England Biolabs, NEB, Ipswich, Massachusetts, USA) for 1 h at 37°C. The digested DNA was ligated in a 25 µL total volume reaction consisting of 0.4 µM for each of the two library-specific adaptors, 0.2 mM ATP (New England Biolabs) and 1 U T4 DNA ligase (SibEnzyme Ltd., Academ town, Siberia). To reduce marker density, one adaptor with fully degenerate 3’ overhangs NN and one with reduced 3’degeneracy NG were chosen. Sample-specific barcodes were designed with Barcode Generator (http://comailab.genomecenter.ucdavis.edu/258

index.php/Barcode_generator) and introduced by PCR with platform-specific barcode-bearing primers (P6-BC). In order to minimize PCR amplification bias [52], 2b-RAD tags were amplified splitting in three wells a 60 µL mixture consisting of 12.5 µL of ligated DNA, 0.5 µM each primer (P4 and P6-BC, Eurofins Genomics S.r.l, Italy), 0.2 µM each primer (P5 and P7, Eurofins Genomics), 0.3 mM dNTP (New England Biolabs), 1X Phusion HF buffer and 1 U Taq Phusion high-fidelity DNA polymerase (ThermoFisher Scientific). Cycling conditions were: 98°C for 4 min; 98 °C for 5 s, 60° C for 20 s, 72° C for 5 s for 14 for 5 cycles, 72°C for 5 min. The reduced number of amplification cycles (n=14) is crucial to decrease the amount of PCR amplification errors and the ratio of GC rich fragments. PCR products were purified with the SPRIselect purification kit (Beckman Coulter, Pasadena, California, USA), to exclude any low-molecular weight DNA remaining after PCR amplification. The concentration of purified individual libraries was quantified using Qubit® ds DNA BR Assay Kit (Invitrogen–ThermoFisher Scientific) and Mx3000P qPCR instrument. The quality of a subset of purified libraries was checked on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA), before sending for sequencing. Samples were equally pooled into three libraries and sequenced on an Illumina HiSeq2500 platform with 50 bp single-end read module at the Genomix4Life S.r.l. facilities (Baronissi, Salerno, Italy). To assess the robustness of the method, technical replicates (TRs) for 14 specimens were prepared by constructing three independent libraries for each replicated individual.

SNP discovery and filtering

Standard demultiplexing and quality filtering of raw data were performed by the sequencing service provider following Illumina protocols. Subsequently, a custom-made script (available upon request) was used to retain only reads with the CspCI recognition site and trim them to 32 base pairs (bp) long fragments (centering on the recognition site).

Stacks' pipeline 'denovo_map.pl' (version 1.35) was used to cluster obtained reads and identify SNPs across samples [58,59]. Demultiplexed reads were first clustered on a single-sample basis (subroutine *ustacks*), with minimum coverage (parameter –m) of 5x and maximum number of three mismatches between reads (parameter –M). *cstacks* was then used to merge tags from single individuals and define a catalog of tags with maximum number of three mismatches (parameter –n). Consensus genotypes for all the samples analyzed were defined with subroutine *sstacks*.

Additional filters were applied to Stacks' output: i) the first and last two nucleotide positions of each tag were trimmed as prone to artifacts produced by Stacks, which in the used version could not handle indels in the clustering process; ii) tags with more than four SNPs were excluded, and, when more than one SNP was present, only the one with the highest minor allele frequency (MAF) was considered; iii) loci with overall MAF<2.5% were discarded; iv) tags shared by less than 80% of all individuals or with coverage < 10x were eliminated; v) samples genotyped at less than 80% of the loci were excluded; vi) pairwise genetic relatedness between samples were calculated with Coancestry [84], and pairs of samples showing coancestry values > 0.95 (i.e. “genetic clones”) were considered being erroneously collected twice and one was discarded. Pruning of potentially duplicated samples is highly recommended as they can affect subsequent analysis such as estimates of genetic variability and clustering [85].

Genetic analysis

Genetic analysis was carried out with GenAlEx 6.501 [86] to estimate expected (He) and observed (Ho) heterozygosity, private alleles and to test for deviations from Hardy-Weinberg (HW) equilibrium. Pairwise Linkage Disequilibrium (LD), using r^2 estimator, was evaluated using Plink [87]. Overall and pairwise Fst values were calculated with Genepop 4.3 [88,89] and Arlequin 3.5.2.1 [90], respectively. Related p-values were calculated with 50,000 permutations. Sequential Bonferroni correction was applied to multiple tests. Outlier markers analysis was performed to identify signal of differentiation between populations, using two different approaches: i) FDIST approach implemented in Lositan [91], with “Neutral mean Fst” and “Force Fst” options and strict Confidence Interval (CI) (99%) and False Discovery Rate (FDR) (0.01); and ii) Bayescan [92–94], with default parameters and Prior odd 100, to avoid false positive [95].

The existence of population structure within the Mediterranean basin was performed using Structure’s [96] clustering analysis. The software was run with the entire dataset and with a reduced data set including only outlier loci. Cluster numbers from k=1 to k=10 were evaluated. Three replicates for each k value were run, with 100,000 burn-in and 100,000 replicates per run. In order to find the most likely k value, results were analyzed following Evanno et al. (2005)[97] methodology as implemented in the website http://taylor0.biology.ucla.edu/structureHarvester/ [98]. A first clustering analysis revealed a strong differentiation in two clusters (see Supplementary

Material, A), which was found to be related to sex (see section 3, Results). This feature was used to discriminate males and females across the sample set. The accuracy of sex discrimination was tested using 35 samples previously sexed by histological analysis of gonads. Clusters were then separated and Bayescan was used to detect loci related to sex differentiation. In addition, presence/absence of RAD tags in different sexes was tested in order to identify sex-specific (“sex-private”) markers, defined as RAD tags present in at least 80% of the individuals of one sex and no individual of the other sex. The analysis was carried out among those tags found in at least six individuals, in order to reduce background noise in the data derived from tags with too much missing data.

Consensus sequences of tags (32 bp long) were compared with GenBank non redundant nucleotide database (nr/nt) using BASIC LOCAL ALIGNMENT TOOL (BLAST) available at NCBI website (https://blast.ncbi.nlm.nih.gov/Blast.cgi) to annotate markers found, with particular focus on tags carrying outliers SNPs or SNPs correlated to sex and considering as “good” matches those with e values below 1e-04.

Results

A total of 438 million reads was obtained for 197 samples (including 28 TRs). Sequencing procedure failed for four samples, which were eliminated. For the remaining 165 individuals, the number of reads per individual ranged from 647,000 to 6,135,241 (average 2,593,523). Stacks identified 61,754 unique tags with 17,495 SNPs, distributed in 10,532 tags. The number of SNPs per tag ranged from 1 to 10, with on average 1.66 SNPs per tag. After filtering, the dataset consisted of 3,324 SNPs (MAF range 0.025-0.500) located on distinct tags (see Supplementary Material, D for additional information about tags’ sequences and SNP variants). Eleven individuals were discarded because of the low number of loci genotyped (<80%). Additional 14 individuals were found to be potential sampling duplicates and thus eliminated.

Analysis of TRs confirmed the good level of precision achievable with the utilized protocol. Across 14 within-replicate comparisons, the average percentage of mismatches was 0.6% (minimum 0.0%; maximum 1.9%).

Ho and He did not differ across samples (Ho range: 0.253-0.259, He range: 0.250-0.258). Out of 21,864 tests for deviation from HW equilibrium, no locus presented a significant deviation, after sequential Bonferroni correction. No private alleles were found in the analyzed populations.

Sampling location	ID	N°	S.R.	Ho	He	F_{is}
Palma de Majorca. Majorca (Spain)	MRC	10	1:0.7	0.253	0.250	-0.014
Ischia. Italy	IS	21	1:0.6	0.256	0.258	0.009
Porticello. Sicily (Italy)	L	19	1:1.4	0.254	0.253	0.005
Port de Teboulba. Tunisia	TN	19	1:18.0	0.252	0.255	0.014
La Valletta. Malta	MFA	20	1:5.3	0.259	0.258	0.002
Tripoli. Libya	TRI	19	1:3.0	0.255	0.257	0.011
Ancona. Italy	ADR	18	1:0.8	0.255	0.256	0.005
Heraklion. Crete (Greece)	CRE	14	1:0.8	0.254	0.254	0.002

Table 1 Summary statistics for eight populations analyzed. N°= number of samples analyzed; S.R.=sex-ratio (M:F); Fis=Fixation index.

Three hundred and eleven SNPs (8.8% of the total number of filtered markers analyzed) were potentially associated with specimen sex. Validation of “genetic” sex identification with gonads histological analysis in 35 individuals showed complete agreement between the results of the two approaches. The genotypes at 65 of these loci were homozygous in all female while some heterozygous were found among males and 185 followed an even more differentiated pattern: female individuals showed all homozygous genotypes, while males had only heterozygous genotypes. The opposite situation was found more rarely, as only 28 loci showed complete homozygosity in males and some heterozygous genotypes in females. As expected, pairwise linkage disequilibrium between these loci was always high (>0.7). The presence of “sex-private” tags was also tested and 386 sequences (almost 1% of the tags analyzed) were found to be present only in males, while only four (0.01%) were found only in females.

When sex-related genetic markers were analyzed in all 140 samples tested, 54 were identified as males and 86 as females, suggesting an overall sex-ratio of 1:1.59 (male:female). At the level of sampling locations, sex-ratio showed broad variations (Table 1), ranging from 1:0.6 (Ischia) to 1:18 (Tunisia), with highly significant heterogeneity across sites (Chi-square test with seven d.f.: X_2=25.73 p<0.001).

Population analysis was carried out on 3,013 loci, after removing loci correlated to sex. Lositan detected eight outlier loci (OL) for divergent selection at stringent CI and FDR. Fifty-seven loci showed signs of balancing selection. No loci were identified by Bayescan as possible outliers.

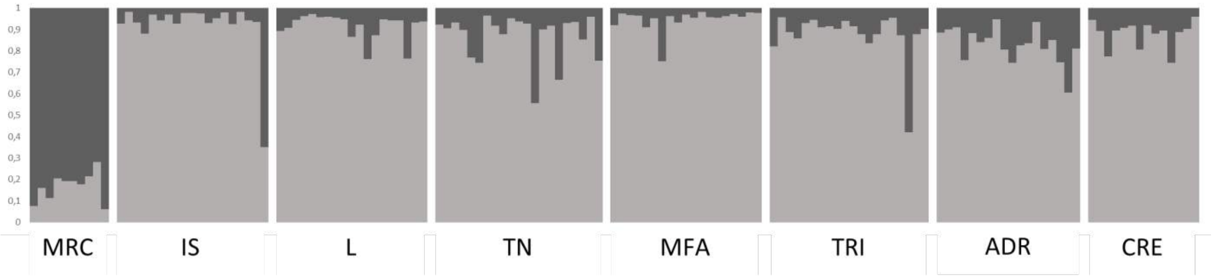
Overall Fst value was 0.0013 (p-value<0.01), using the entire SNP dataset, but no significant pairwise Fst values were detected, and Structure analysis suggested the most probable value for k=1. However, when using eight OL, pairwise Fst between MRC and L and between MRC and MFA were high and significant also after sequential Bonferroni correction (Table 2). Overall Fst for OL dataset was 0.0729 (p-value<0.001).

Pop	MRC	IS	L	TN	MFA	TRI	ADR	CRE
MRC	*	0.0019	0.0029	0.0030	0.0003	0.0002	0.0005	0.0004
IS	0.1267	*	0.0003	0.0035	0.0001	0.0008	0.0006	0.0007
L	**0.2938**	0.0090	*	0.0008	0.0001	0.0000	0.0000	0.0000
TN	0.1429	0.0000	0.0000	*	0.0037	0.0030	0.0035	0.0029
MFA	**0.4168**	0.1150	0.0125	0.0957	*	0.0019	0.0018	0.0001
TRI	0.0806	0.0000	0.0753	0.0000	0.2046	*	0.0000	0.0000
ADR	0.2436	0.0000	0.0000	0.0000	0.0322	0.0447	*	0.0018
CRE	0.0313	0.0020	0.1113	0.0146	0.2451	0.0000	0.0771	*

Table 2 Pairwise Fst table. Values based on the entire SNPs dataset are indicated above the diagonal. Values based on eight "potential outlier" dataset are indicated under the diagonal. Underlined values indicate p<0.05. Bold values indicate statistical significance also after sequential Bonferroni correction.

STRUCTURE HARVESTER analysis suggested that the most likely value of the analysis performed with 8 OL loci was k=7 (see Supplementary Material, B). Nevertheless, the plot obtained for this value of k, was characterized by the presence of uninformative clusters, that didn't differentiate the populations. For this reason, the second most likely value was selected as the most informative (i.e. k=2). The presence of uninformative cluster has already been reported by other authors [99,100], and it is suggested to remove these clusters from the analysis. Structure analysis obtained with OL at k=2 suggested the presence of differentiation between the most Western Mediterranean group (MRC) and the other Mediterranean samples, which formed a more homogeneous group (Figure 2).

Figure 2 Structure's plot. Analysis performed using eight outlier loci and sampling location 'a priori', k=2 (MRC=Spain, IS= Ischia, L= Porticello, TN= Tunisia, MFA= Malta, TRI= Libya, ADR= Ancona, CRE= Crete)

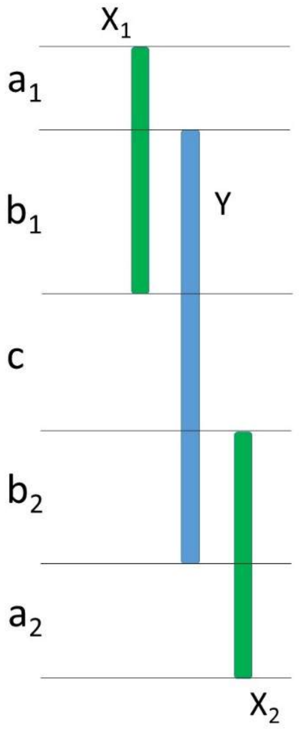


Out of 319 test for local alignment with BLAST (eight outliers and 311 sexual differentiation loci), only five (all of them sex differentiation loci) gave significant result (see Supplementary Material, D). Among the remaining 3,013 "neutral" loci, six loci had significant matches.

Discussion

Two relevant results were achieved in the present study: (i) the identification of a large set of genetic markers tightly associated with sex and (ii) the identification of unexpected, albeit weak genetic differentiation of the Western Mediterranean population in comparison to the other ones. Over 300 SNPs were found to be associated with sex in dolphinfish and most of them were characterized by a typical genotyping pattern with one sex showing only homozygous genotypes and the other sex showing heterozygous genotypes, suggesting the existence of sexual chromosome. In marine fish species, the presence of cytologically differentiated sex chromosome is rare (around 10% according to Devlin and Nagahama (2002)[101]), and where present they are quite diverse, including single as well as multiple chromosome systems (XX/XY, ZZ/ZM, XX/XY_1Y_2, $X_1X_1X_2X_2/X_1X_2Y$). These systems usually arise from either translocation or centric/tandem fusion between ancestral sex chromosomes and autosomes. The closest species for which karyotype is known for both sexes is *Coryphaena equiselis,* while for *C. hippurus* only female has known karyotype (Soares et al., 2014). *C. equiselis* showed a $X_1X_1X_2X_2/X_1X_2Y$ karyotype, which is also the most widely reported for fish species with differentiated sex chromosome (e.g. Ueno and Takai (2008)[102]). The apparently high number of sex associated markers (8.8% of the total number of SNPs) found in this study is expected in the case of $X_1X_1X_2X_2/X_1X_2Y$ system, that involves genetic dimorphism at two chromosomes out of 24. Moreover, it is known that in this karyotype pattern, remaining X_1 and X_2 chromosome in the heterogametic sex (in our case the male) can maintain some level of recombination with Y chromosome [102]. This situation generates five separated areas of markers in the sex chromosomes, according to their position and the possibility of recombination between X_1-Y and X_2-Y (Figure 3). This can also explain the homozygous/heterozygous pattern found in this study. Some markers (28 out of 311) showed a pattern in which males (the heterogametic sex) was characterized by all homozygous loci, while heterozygosity was present in females. These SNPs probably belong to regions a_1 and a_2 in Figure 3, and follow the “classical” pattern of heterozygosity found in sex chromosomes. Nevertheless, a much greater number of SNPs (250) were homozygous in females and heterozygous in males. According to our hypothesis, these markers belongs to region c of Figure 3. In fact, if Y chromosome derived from the centric fusion of two acrocentric chromosomes, then males’ heterozygous loci derived from the persistence of the same homologous regions in the X_1/ X_2 chromosomes and Y. When mutations happen in the Y part close to enzyme’s recognition site, these cause heterozygosity in 2b RAD tags of heterogametic sex. In accordance with this hypothesis, the existence of almost 400 tags detected only in one sex (i.e. male) was expected and already documented in another species with the same chromosomic sex determination system [103]. These markers come from Y chromosome and arose from mutations that created new enzyme cutting sites, not present in the original acrocentric chromosome. Further studies focusing on the chromosomic sexual dimorphism in *Coryphaena hippurus* are required to better understand the link between chromosome pattern and the genotypes we obtained. For the moment, from our results (i.e. higher number of markers homozygous in females and heterozygous in males) we can assume that the regions of chromosomes X_1 and X_2 not recombining with Y chromosome (a_1 and a_2 in Figure 3) are relatively small when compared to the region of Y that doesn’t recombine (region c). Regions b_1 and b_2, in which recombination with Y persists, are known as pseudoautosomial regions (PAR)[104]. In our case they can vary in their size and are hard to measure with the tools we used, as markers in these regions are expected to behave as autosomes. Unfortunately, annotation of these sex related markers showed low success rates, as only five markers had significant match with nr/nt database. Anyhow, this was expected, considering that 2bRAD sequences are only 32 bp long and thus the ability to annotate them is low [50]. The availability of a reference genome would give the possibility to extract flanking regions of these markers, thus increasing the chance of finding significant matches with already studied sequences. The significant match of one of the tags with a microtubule-associate protein ('cytoskeleton associated protein 5 (ckap5)' encoded by the CKAP5 gene) could indicate an intersexual differential expression of this protein involved in spindle formation, that is expected given the hypothesized different chromosomal asset of the sexes.

Figure 3 Supposed recombination pattern between sexual chromosome X_1, X_2 and Y. Different letters indicate different regions of recombination: in a_1 and a_2 no recombination happens and genotypes are expected to be all homozygous in males while heterozigosity can be present in females; in b_1 and b_2 recombination happens and markers should behave as autosomal markers; c represent Y region that doesn't recombine. In c "male-private" markers can be found and markers that are heterozygous in males and homozygous in females, originated from mutations in Y chromosomes around the enzyme's recognition site.



From a practical point of view, the identification of sex related markers represents an important starting point to develop a quick and inexpensive diagnostic tool for genetic sex identification especially because, in dolphinfish, most of the catches focuses on sexually immature individual. Robust estimates of sex-ratios in large samples are an important tool for management and conservation of biological resources. Moreover, non-invasive genetic sex identification helps in determine sex in tagging studies, providing essential information to assess sex-biased dispersal (e.g. Galindo et al. (2011)[105]). The average sex-ratio obtained in the present study (1:1.6 M:F) is close to what reported in previous works carried out on dolphinfish populations of the western central Atlantic (1:2-1:3) [81,106]. Although caution should be exerted because of the limited sample sizes, it is interesting to notice that sex-ratios significantly differ between sampling areas. Such differences might be due to (i) sampling strategy, which aim at different average fish sizes, (ii) sampling sites, e.g. around FADs or in the open sea and also (iii) sampling season [78,81]. Anyhow, such large differences in sex-ratio have not been reported before and further studies specifically focused on this feature can be of great interest to understand the behavior of the species at different life stages and the potential sex-ratio-biased harvesting of different fishing techniques.

The second important finding is that, using OL, preliminary evidence of genetic differentiation between different localities was found. Considering the high mobility of dolphinfish and according to previous works on the species, it was not expected to find significant population structure within a single basin, such as the Mediterranean Sea. In fact, dolphinfish are known to move over large areas, as studies carried out especially on the American Atlantic coasts suggest [107,108]. The use of a large set of robust, highly informative SNP loci is likely the key to such a finding. The approach based on 2bRAD proved already to reveal hidden population structure in another large pelagic highly mobile species, the yellowfin tuna [82]. In the case of dolphinfish, the availability of over 3,000 SNPs allowed the identification of a few OL, which provided greater sensitivity compared to all remaining markers. It was already reported in the European hake that only the use of OL could detect relevant population structure within the Mediterranean basin [109]. While it might be possible that such markers represent random deviations from the mean in a large set of observations, this seems unlikely considering that outlier analysis was carried out across all eight population samples, but just one site (MRC) seems genetically divergent based on allele frequencies (see also Supplementary Material, C). Biological interpretation of OL remains under debate. While most authors consider them as “loci under local, divergent selection”, others have proposed the concept of genomic incompatibilities as a consequence of secondary contact after historical separation to explain the evidence of outlier loci [110]. Thus, at least two hypotheses might explain the weak genetic differentiation of the population sample in the Western Mediterranean. In the first scenario, either local environmental conditions or past separation of dolphinfish between sub-basins within the Mediterranean should be invoked. In fact, pairwise Fst showed an unexplained heterogeneity between the westernmost sample (MRC) and only two sampling locations (L and MFA) with no apparent link between them. In the second scenario, considering that MRC samples are the closest ones to Gibraltar trait, admixture of individuals from the Atlantic and the Mediterranean basins might occur. In general, Atlantic-Mediterranean exchange is expected for long distance migratory species such as dolphinfish even if, as reported by the “Dolphinfish Research Program” (www.dolphintagging.com), no fish tagged in the Atlantic has been recovered in the Mediterranean Sea. Moreover, a case was reported of a specimens tagged in the Island of Majorca and recovered in the Sicilian strait (D.L. Hammond, *pers. comm.*). For highly motile species such as dolphinfish, understanding the effect of local environmental or ecological factors driving genetic differentiation is not easy. In addition, little is known about the behaviour of this species during the spawning season (e.g. homing behaviour, seasonal behaviour), thus we cannot exclude that fish caught in the West Mediterranean are Atlantic that migrate into Mediterranean for different reasons. Nevertheless, taking into consideration the most influencing factors for marine fish genetic differentiation (i.e. salinity and temperature) it is likely that fish coming from the Atlantic are better adapted to an environment such as that of Western Mediterranean, which is characterized by lower salinity and surface temperature, and prefer to swim in these areas. Even if BLAST analysis didn’t suggest any link between OL RAD tags and known functional sequences, this is probably due to the length of 2b RAD tags, as for the sexual differentiation markers. Genetic differentiations related with environmental factors has been already describe for a species with similar high motility, the bluefin tuna, in the Mediterranean basin, despite the long distances covered by specimen of this species, as recorded by electronic tagging [111,112]. In these studies, salinity and temperature were in fact identified as parameters shaping the genetic population structure more than geography. For *Coryphaena,* additional analyses including eastern Atlantic specimen are required to understand, if it exists, the wider scale genetic structure and the level of migration and gene exchange/flow between the Atlantic and the Mediterranean population. These findings would suggest a preliminary hypothesis for future management strategies which should take into account or include a greater knowledge of the biology and ecology of this important and valuable resource.

Conclusions

2bRAD-based analysis of Mediterranean dolphinfish confirmed the great potential of genotyping-by-sequencing methods applied to fishery genetics, with the identification of markers for genetic sex determination and preliminary evidence for a possible hidden population genetic structure. This information will help management and conservation of this species.

Data accessibility

Raw sequences were deposited in SRA (Accession Number: SRP074965).

Acknowledgements

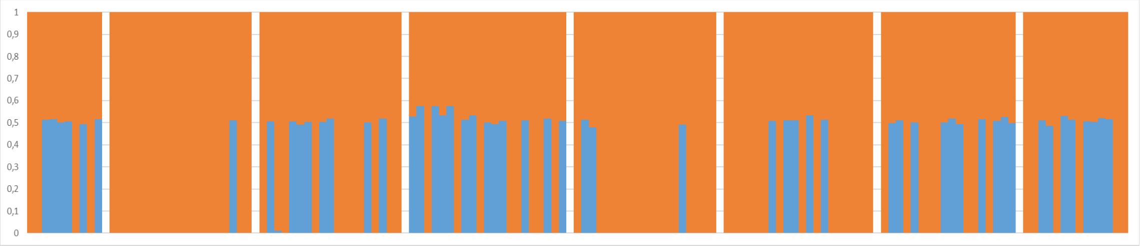
This work was supported by the project RITMARE (www.ritmare.it) and CoNISMa. Samples were collected with the help and coordination of FAO Regional Projects ADRIAMED, MEDSUDMED and COPEMED II which are here acknowledged. The authors thanks J. A. Camiñas, E. Massuti, F. Grati, G. Scarcella, P. Psomadakis, E. Arneri, L. Ceriola, A. Pešić, E. Ricevuto, C. S. Tsigenopoulos, A. Besbes, F. Fiorentino, M. Gatt and S. Saber for the help they provided.

SUPPLEMENTARY MATERIAL

A) Structure clustering highlighting genetic sexual dimorphism

As reported in the main text, the first analysis performed with Structure, using the entire SNPs datasets (3,324 loci), highlighted a clear differentiation of the sample in two clusters (Figure 1). Clusters were related to sex differentiation, and this feature was tested with samples of known sex (from histological analysis).

Figure 1 Structure plot (k=2) for 144 dolphinfish samples, using 3,324 2bRAD markers. Samples from different geographical areas are separated by blank lines. The clusters identified by Structure are related to sex. Specifically, "full orange" cluster identifies females and "blue-orange" cluster identifies males.



B) STUCTURE HARVESTER results

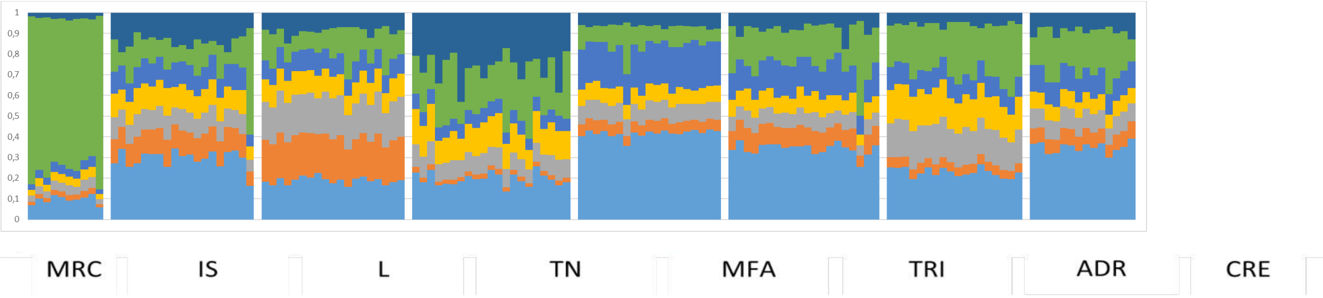
STRUCTURE HARVESTER was used to parse results from Structure run with eight outlier loci in order to find the number of clusters that best fitted the data.

k	*reps*	*Mean LnP (k)*	*StDev LnP (k)*	*Ln' (k)*	*Ln'' (k)*	*Delta k*
1	*3*	*-1175.9667*	*0.2082*	*NA*	*NA*	*NA*
2	*3*	*-1144.0667*	*4.5938*	*31.900*	*49.333*	*10.739*
3	*3*	*-1161.5000*	*7.7544*	*-17.433*	*41.200*	*5.313*
4	*3*	*-1220.1333*	*57.3941*	*-58.633*	*72.133*	*1.257*
5	*3*	*-1206.6333*	*40.5789*	*13.500*	*9.266*	*0.228*
6	*3*	*-1183.8667*	*15.4212*	*22.766*	*47.333*	*3.069*
7	*3*	*-1208.4333*	*3.2716*	*-24.566*	*44.000*	*13.449*
8	*3*	*-1189.0000*	*16.6640*	*19.433*	*10.433*	*0.626*
9	*3*	*-1180.0000*	*23.8447*	*9.000*	*9.102*	*0.456*
10	*3*	*-1289.6582*	*18.3258*	*10.500*	*NA*	*NA*

Table 1 Results from STRUCTURE HARVESTER's analysis of the results of Structure's analysis with eight outlier loci for 140 individuals.

In the plot for k=7 (Figure 1), 'green' and 'light blue' are the most informative clusters for groups differentiation, reducing the plot to a total of two informative clusters.

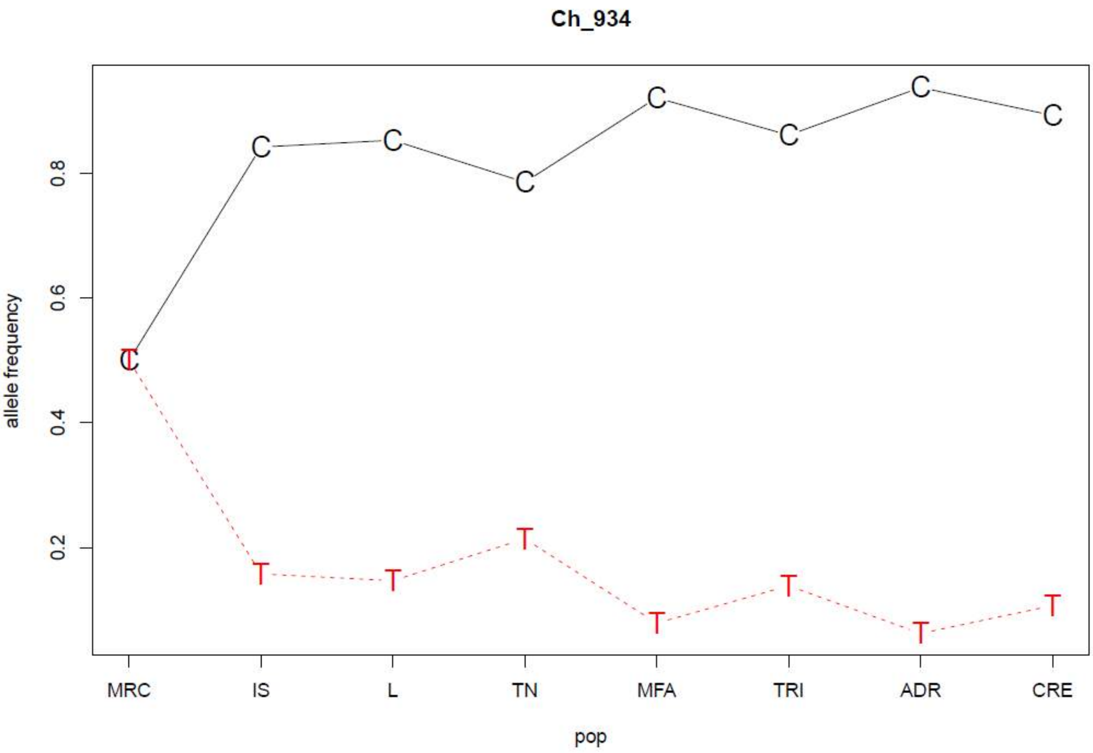
Figure 2 Structure plot for k=7; 140 dolphinfish samples divided in seven populations (MRC=Spain, IS= Ischia, L= Porticello, TN= Tunisia, MFA= Malta, TRI= Libya, ADR= Ancona, CRE= Crete).

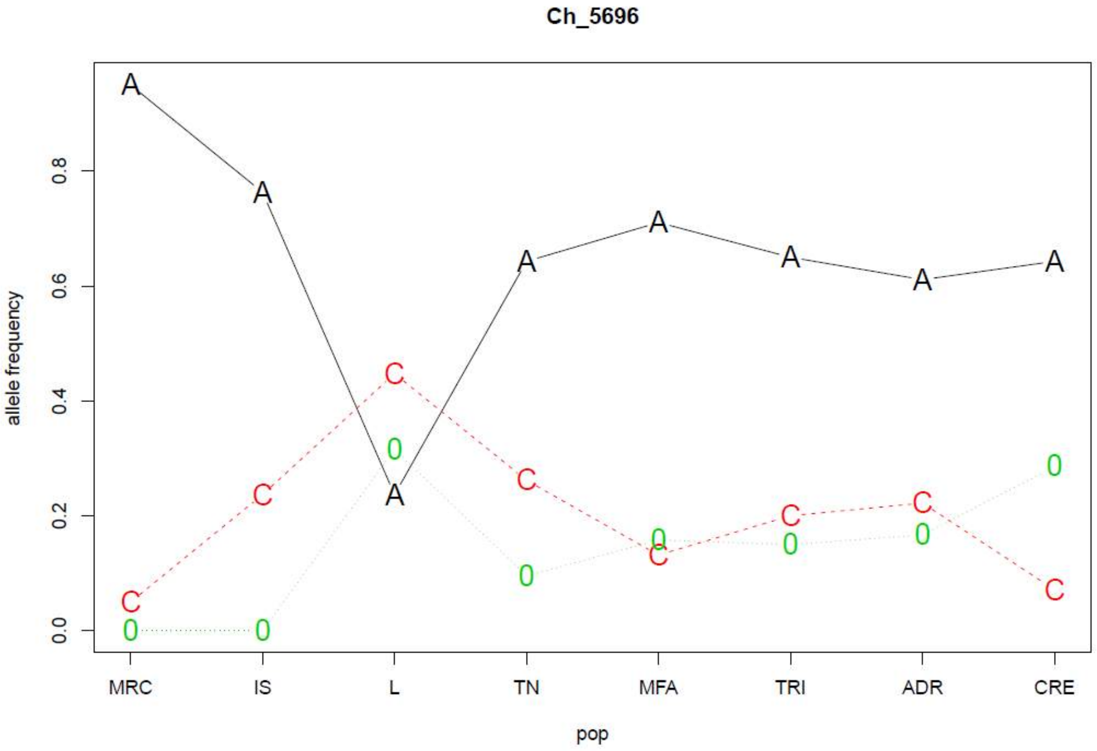


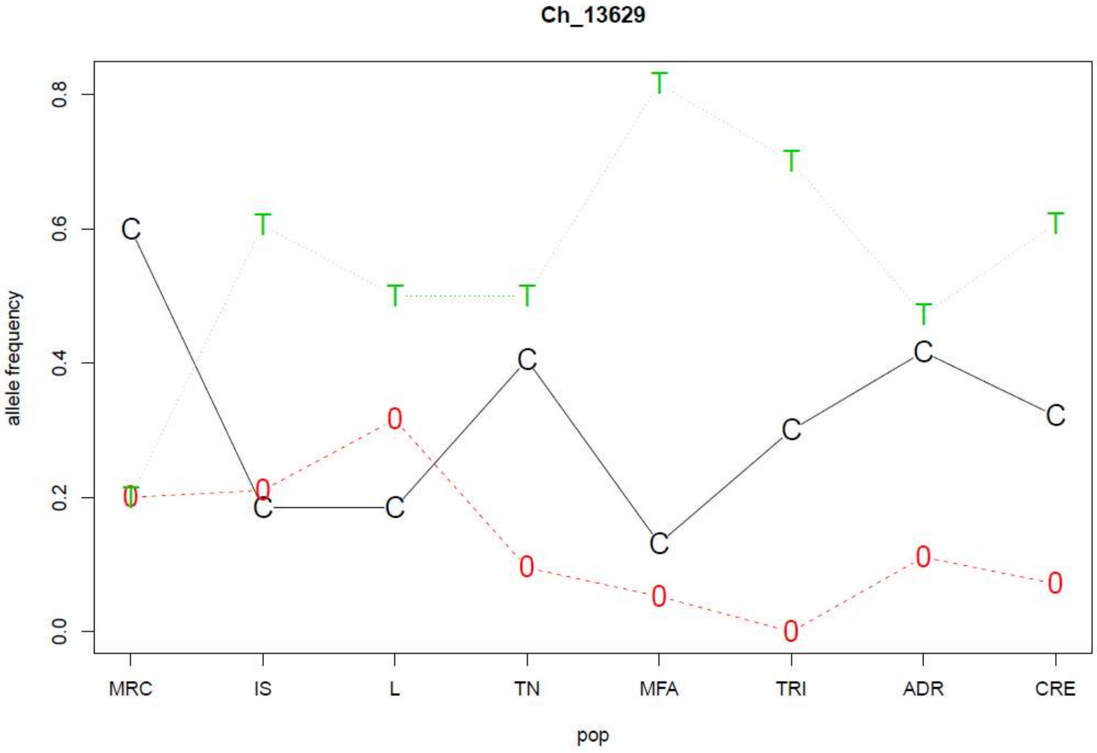
C) Allele frequencies at outlier loci

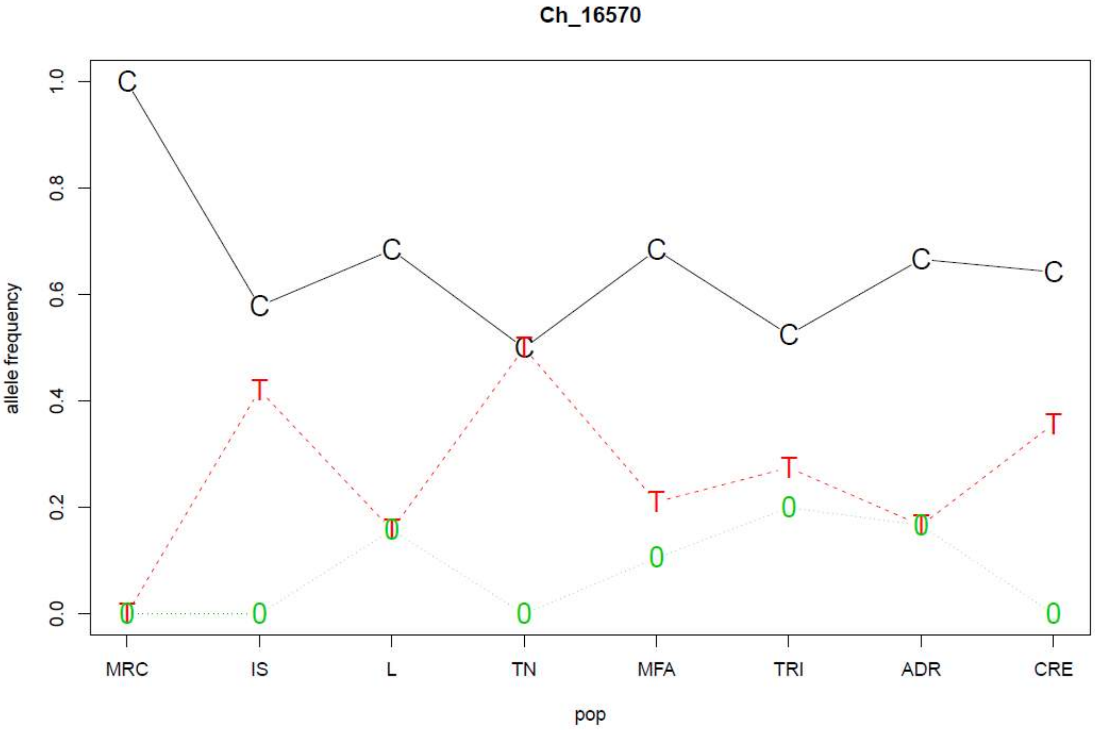
As stated in the main text, outlier analysis was carried out using eight separated populations. Nevertheless, clustering analysis using these loci suggested the presence of shallow genetic differentiation only between the Western Mediterranean sample and the other groups (that appear homogeneous). This feature is better understandable looking at the outliers' allele frequencies at each population (Figure 1). Many loci present different allele frequencies when comparing Spain population and the remaining populations, while the other groups have more homogeneous frequencies.

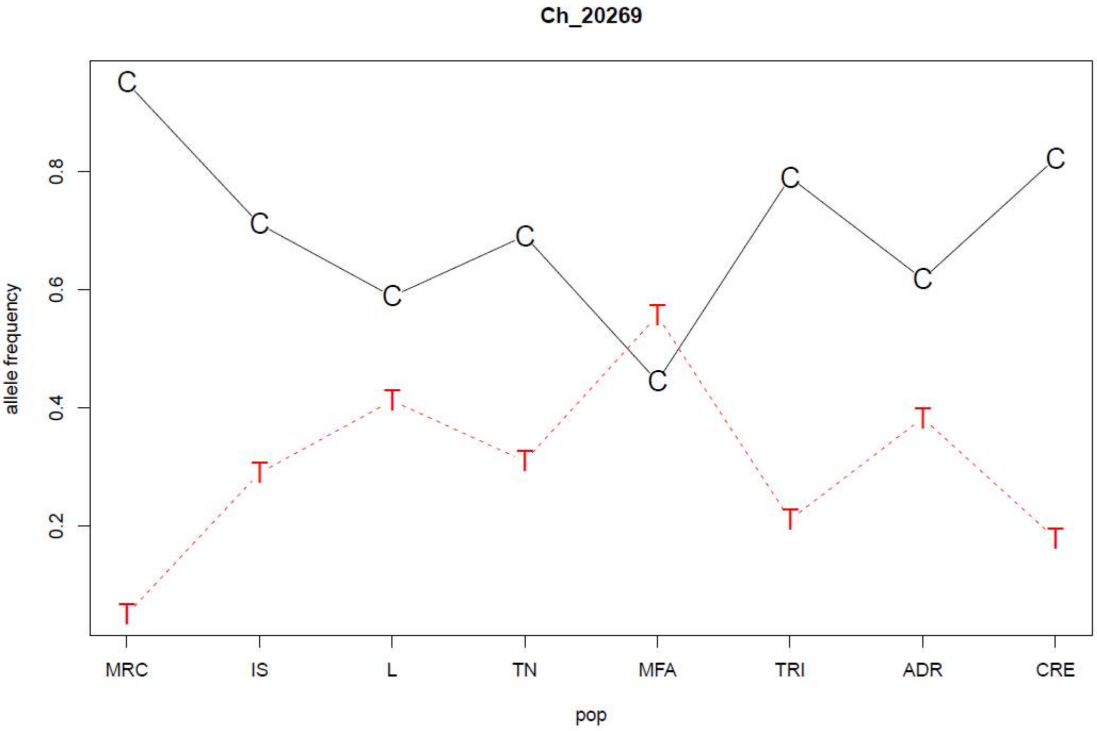
Figure 3 Allele frequencies at eight outlier loci (Ch_934, Ch_5696, Ch_13629, Ch_16570, Ch_20269, Ch_21050, Ch_21915, Ch_29182) for each population (MRC=Spain, IS= Ischia, L= Porticello, TN= Tunisia, MFA= Malta, TRI= Libya, ADR= Ancona, CRE= Crete). A, T, C, G represent different alleles, 0 indicates missing data.

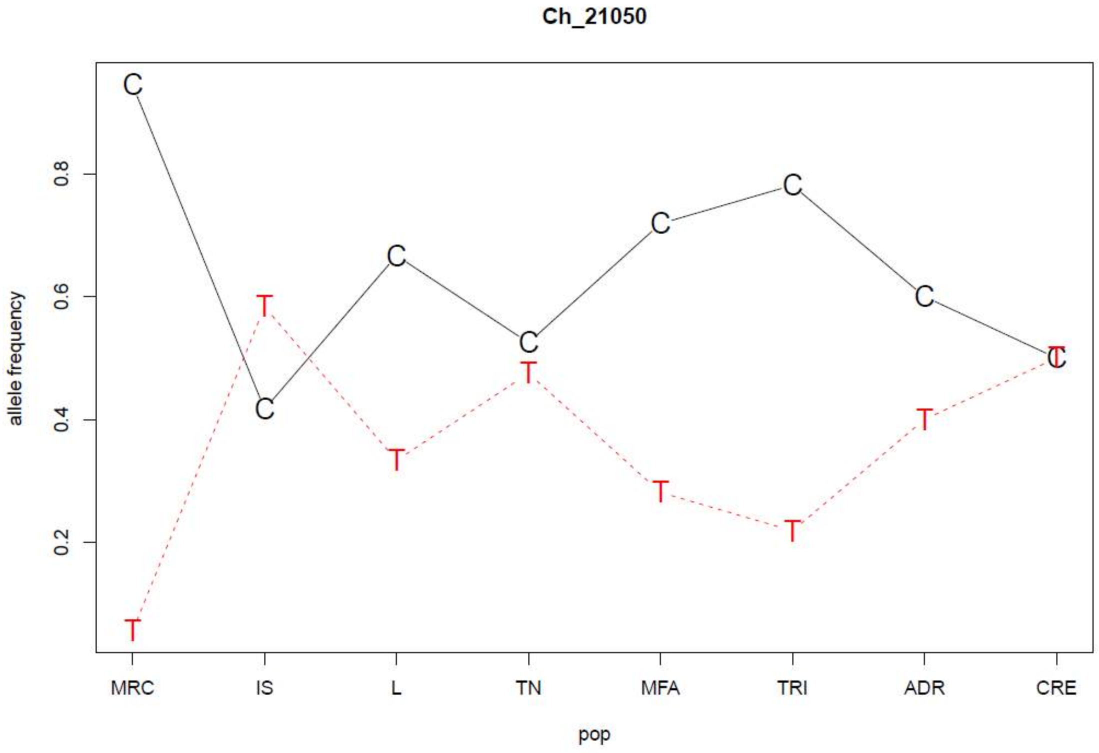


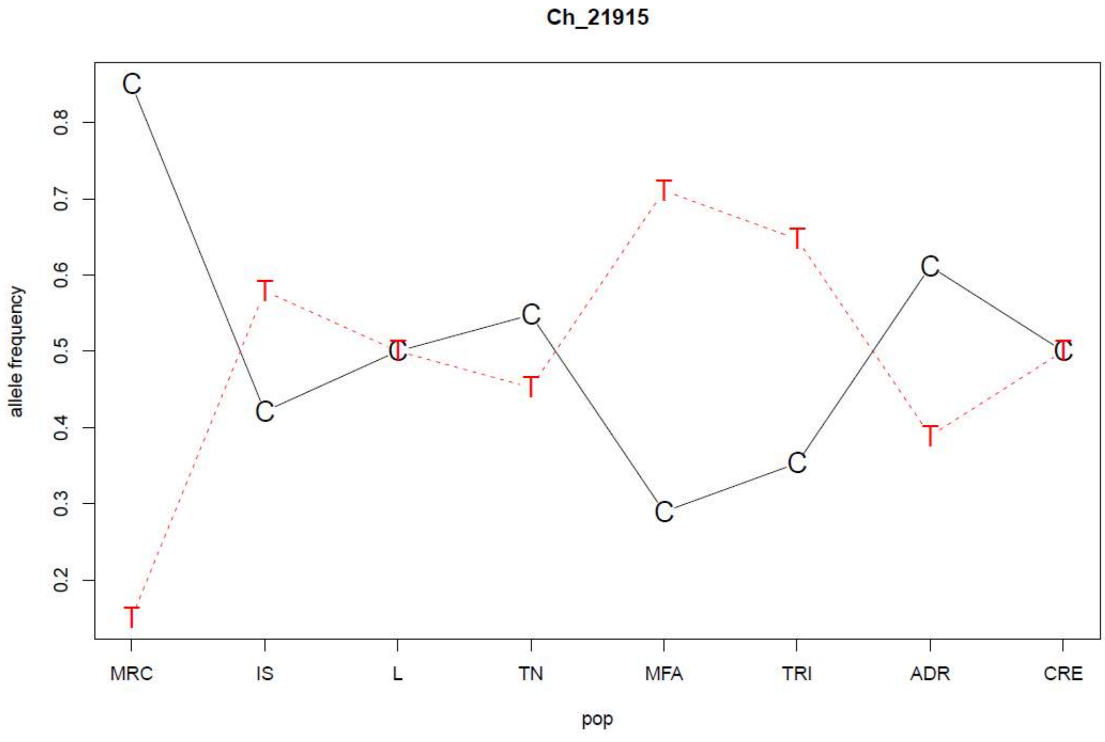

Ch_5696
allele frequency
0.0
0.2
0.4
0.6
0.8
MRC
IS
L
TN
MFA
TRI
ADR
CRE
pop

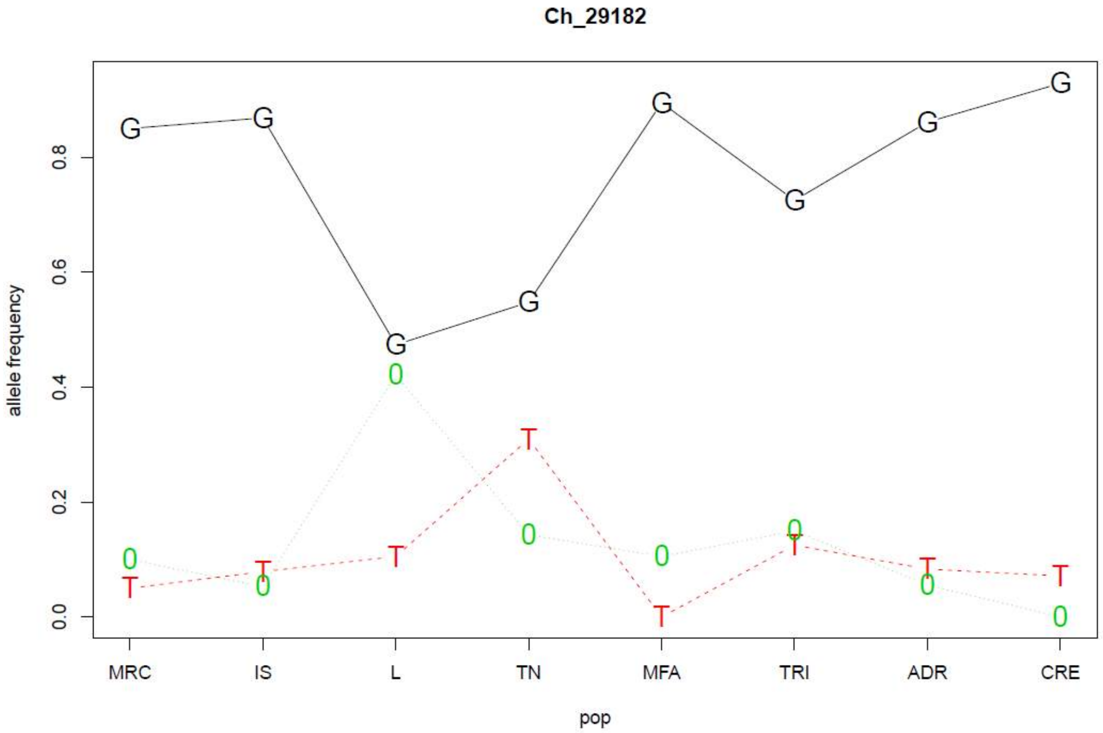

Ch_13629
allele frequency
0.0
0.2
0.4
0.6
0.8
MRC
IS
L
TN
MFA
TRI
ADR
CRE
pop


Ch_16570
allele frequency
0.0
0.2
0.4
0.6
0.8
1.0
MRC
IS
L
TN
MFA
TRI
ADR
CRE
pop
C
T
0


Ch_20269
allele frequency
0.2
0.4
0.6
0.8
MRC
IS
L
TN
MFA
TRI
ADR
CRE
pop
C
T


Ch_21050
allele frequency
0.2
0.4
0.6
0.8
pop
MRC
IS
L
TN
MFA
TRI
ADR
CRE
C
T


Ch_21915
allele frequency
0.2
0.3
0.4
0.5
0.6
0.7
0.8
pop
MRC
IS
L
TN
MFA
TRI
ADR
CRE
C
T


Ch_29182
allele frequency
0.0
0.2
0.4
0.6
0.8
MRC
IS
L
TN
MFA
TRI
ADR
CRE
pop
G
T
0

D) Table of RAD sequences, SNP positions and variants, BLAST alignments

We report the list of 311 2bRAD SNPs correlated to sex markers, eight outlier SNPs identified by Lositan and tags for which a match was found in BLAST analysis. Position of SNPs within the sequence and the variants are included.

Marker name	SNP pos	Alt alleles	Sexual marker/Outliers	BLAST alignment	e-value
13629_26	26	G/T	Outlier		
16570_6	6	C/T	Outlier		
20269_16	16	C/T	Outlier		
21050_25	25	A/G	Outlier		
21915_9	9	C/T	Outlier		
29182_14	14	G/T	Outlier		
5696_2	2	A/C	Outlier		
934_15	15	A/G	Outlier		
10049_7	7	A/G	Sexual marker		
10316_13	13	C/T	Sexual marker		
10510_22	22	A/C	Sexual marker		
10593_26	26	A/G	Sexual marker		
10653_22	22	G/T	Sexual marker		
10853_17	17	C/G	Sexual marker		
11087_6	6	G/T	Sexual marker		
11089_28	28	A/G	Sexual marker		
11093_7	7	C/T	Sexual marker		
11229_26	26	C/T	Sexual marker		
11283_14	14	C/T	Sexual marker		
11411_6	6	C/T	Sexual marker		
11497_7	7	G/T	Sexual marker		
11582_5	5	A/G	Sexual marker		
11679_5	5	G/T	Sexual marker		
11926_28	28	C/T	Sexual marker		
11928_3	3	G/T	Sexual marker		
11936_9	9	A/T	Sexual marker		
1218_5	5	A/G	Sexual marker		
12222_5	5	A/G	Sexual marker		
12253_16	16	C/T	Sexual marker		
12314_23	23	C/T	Sexual marker		
12345_29	29	A/T	Sexual marker		
12387_6	6	A/G	Sexual marker		
12440_3	3	G/T	Sexual marker		
12453_8	8	G/T	Sexual marker		
12538_6	6	A/G	Sexual marker		
12659_24	24	A/G	Sexual marker		

1290_28	28	C/G	Sexual marker		
1296_7	7	A/G	Sexual marker		
13022_15	15	C/T	Sexual marker		
13145_8	8	C/T	Sexual marker		
13188_2	2	A/G	Sexual marker		
13340_23	23	C/T	Sexual marker		
13381_26	26	A/G	Sexual marker		
13631_5	5	C/T	Sexual marker		
1364_15	15	A/T	Sexual marker		
13739_14	14	A/G	Sexual marker		
13874_4	4	C/T	Sexual marker		
1398_25	25	A/C	Sexual marker		
14021_26	26	A/G	Sexual marker		
14058_17	17	A/G	Sexual marker		
14154_14	14	A/T	Sexual marker		
14275_26	26	C/T	Sexual marker		
14315_15	15	A/C	Sexual marker		
14334_25	25	A/T	Sexual marker		
14387_8	8	A/G	Sexual marker		
14436_9	9	A/T	Sexual marker		
14448_2	2	A/G	Sexual marker		
14583_17	17	G/T	Sexual marker		
14635_13	13	C/T	Sexual marker	TPA_asm: Oryzias latipes strain Hd-rR, complete genome assembly, chromosome 6	9,25E-05
14869_4	4	C/T	Sexual marker		
14899_24	24	A/G	Sexual marker		
1493_3	3	A/G	Sexual marker		
14951_28	28	A/C	Sexual marker		
14969_25	25	A/G	Sexual marker		
14986_2	2	C/T	Sexual marker		
15023_16	16	A/G	Sexual marker		
15079_14	14	C/T	Sexual marker		
15353_29	29	C/T	Sexual marker		
1540_2	2	A/C	Sexual marker		
15563_8	8	C/T	Sexual marker		
15574_25	25	A/G	Sexual marker		
15635_24	24	A/G	Sexual marker		
15693_5	5	C/T	Sexual marker		
15836_4	4	C/T	Sexual marker		
15899_4	4	A/G	Sexual marker		
15944_26	26	A/G	Sexual marker		
15948_26	26	C/T	Sexual marker		
15969_22	22	G/T	Sexual marker		
16061_14	14	C/T	Sexual marker		
16084_16	16	A/C	Sexual marker		

16139_9	9	A/T	Sexual marker
1623_7	7	A/G	Sexual marker
16303_2	2	A/C	Sexual marker
16517_3	3	A/C	Sexual marker
16809_7	7	C/T	Sexual marker
16993_9	9	A/G	Sexual marker
17019_3	3	C/T	Sexual marker
17053_2	2	C/T	Sexual marker
17065_27	27	A/T	Sexual marker
17104_5	5	G/T	Sexual marker
17111_5	5	C/T	Sexual marker
17215_23	23	A/G	Sexual marker
17289_28	28	A/G	Sexual marker
17359_4	4	G/T	Sexual marker
17496_25	25	A/G	Sexual marker
1763_7	7	G/T	Sexual marker
17633_14	14	A/G	Sexual marker
17650_15	15	A/G	Sexual marker
17684_29	29	A/T	Sexual marker
17705_6	6	C/T	Sexual marker
17815_13	13	A/G	Sexual marker
17849_4	4	C/T	Sexual marker
17994_17	17	C/T	Sexual marker
18015_8	8	C/G	Sexual marker
18075_3	3	A/G	Sexual marker
18108_5	5	G/T	Sexual marker
18163_3	3	C/T	Sexual marker
18276_2	2	A/T	Sexual marker
18524_9	9	A/T	Sexual marker
187_7	7	A/T	Sexual marker
18821_5	5	A/T	Sexual marker
19069_14	14	A/T	Sexual marker
19072_5	5	G/T	Sexual marker
19084_7	7	G/T	Sexual marker
19153_15	15	C/T	Sexual marker
19170_17	17	A/G	Sexual marker
19336_13	13	A/G	Sexual marker
1944_16	16	A/C	Sexual marker
19449_26	26	A/T	Sexual marker
19683_3	3	A/G	Sexual marker
19709_8	8	A/C	Sexual marker
19925_17	17	C/G	Sexual marker
19958_2	2	A/G	Sexual marker
20092_27	27	A/G	Sexual marker

20112_7	7	A/T	Sexual marker		
20113_17	17	G/T	Sexual marker		
20306_8	8	C/G	Sexual marker		
20342_3	3	A/G	Sexual marker		
20361_14	14	A/G	Sexual marker		
20514_9	9	A/G	Sexual marker		
20522_28	28	C/T	Sexual marker		
20581_6	6	A/G	Sexual marker		
20583_14	14	A/G	Sexual marker		
2059_6	6	A/T	Sexual marker		
20646_15	15	A/T	Sexual marker		
20973_7	7	A/G	Sexual marker		
20996_2	2	C/T	Sexual marker		
20999_17	17	A/T	Sexual marker		
21004_16	16	A/G	Sexual marker		
21010_25	25	C/T	Sexual marker		
21025_14	14	A/G	Sexual marker	Cyprinus carpio clone 286704 microsatellite sequence	9,25E-05
21027_9	9	A/T	Sexual marker		
21197_26	26	A/G	Sexual marker		
21239_2	2	A/G	Sexual marker		
21310_24	24	A/G	Sexual marker		
2143_13	13	A/T	Sexual marker		
21435_9	9	C/T	Sexual marker		
21454_28	28	G/T	Sexual marker		
21581_24	24	A/G	Sexual marker		
22048_8	8	A/T	Sexual marker		
22085_28	28	A/G	Sexual marker		
22139_2	2	C/T	Sexual marker		
2217_15	15	A/G	Sexual marker		
22305_28	28	C/T	Sexual marker		
22362_4	4	A/G	Sexual marker		
22463_8	8	C/T	Sexual marker		
22485_26	26	A/G	Sexual marker		
22526_5	5	A/G	Sexual marker		
22970_5	5	C/G	Sexual marker		
22980_5	5	A/G	Sexual marker		
23138_17	17	C/T	Sexual marker		
23157_15	15	A/G	Sexual marker		
23173_16	16	C/T	Sexual marker		
2331_17	17	A/T	Sexual marker		
23313_4	4	C/T	Sexual marker		
23429_2	2	C/T	Sexual marker		
23441_24	24	C/T	Sexual marker		
23579_9	9	A/C	Sexual marker		

23653_29	29	C/T	Sexual marker		
23659_27	27	C/T	Sexual marker		
23771_4	4	A/T	Sexual marker		
2379_5	5	C/G	Sexual marker		
2387_28	28	C/T	Sexual marker		
23998_27	27	A/C	Sexual marker		
24036_4	4	A/T	Sexual marker		
24251_24	24	A/G	Sexual marker		
24339_3	3	A/T	Sexual marker		
2451_4	4	A/G	Sexual marker	PREDICTED: Poecilia formosa pre-B-cell leukemia homeobox 4 (pbx4), transcript variant X4, mRNA	2,00E-05
24586_8	8	C/T	Sexual marker		
24653_14	14	A/G	Sexual marker		
24657_13	13	A/G	Sexual marker		
24741_4	4	C/T	Sexual marker		
2482_16	16	C/T	Sexual marker		
24887_17	17	C/T	Sexual marker		
24915_6	6	C/T	Sexual marker		
25055_26	26	A/C	Sexual marker		
25080_2	2	C/T	Sexual marker		
25093_15	15	C/T	Sexual marker		
25153_2	2	A/G	Sexual marker		
25518_4	4	A/G	Sexual marker		
25551_15	15	C/T	Sexual marker		
2580_3	3	A/T	Sexual marker		
25848_3	3	A/G	Sexual marker		
25998_7	7	C/G	Sexual marker		
26038_6	6	G/T	Sexual marker		
26047_22	22	C/T	Sexual marker		
26104_23	23	A/G	Sexual marker		
26163_3	3	C/G	Sexual marker		
26211_5	5	C/T	Sexual marker		
26229_13	13	C/T	Sexual marker		
2630_26	26	C/T	Sexual marker		
26323_24	24	A/C	Sexual marker		
26341_6	6	A/C	Sexual marker		
264_28	28	A/G	Sexual marker		
2655_23	23	C/T	Sexual marker		
26737_2	2	A/G	Sexual marker		
26969_29	29	A/G	Sexual marker		
27000_4	4	C/T	Sexual marker		
27039_24	24	C/G	Sexual marker		
27079_8	8	A/G	Sexual marker		
27106_5	5	C/T	Sexual marker		
2720_27	27	C/G	Sexual marker		

27251_17	17	C/T	Sexual marker		
27327_2	2	C/T	Sexual marker		
2746_27	27	A/G	Sexual marker		
27631_7	7	A/G	Sexual marker		
27702_26	26	A/C	Sexual marker		
2771_7	7	C/T	Sexual marker		
27724_2	2	C/T	Sexual marker		
27755_2	2	A/G	Sexual marker		
27767_4	4	C/T	Sexual marker		
28013_17	17	A/G	Sexual marker		
28220_8	8	A/G	Sexual marker	PREDICTED: Esox lucius cytoskeleton associated protein 5 (ckap5), transcript variant X4, mRNA	9,25E-05
28258_2	2	A/G	Sexual marker		
28291_28	28	C/G	Sexual marker		
28349_5	5	C/G	Sexual marker		
28623_5	5	A/G	Sexual marker		
28754_2	2	A/C	Sexual marker		
28795_22	22	C/T	Sexual marker		
29134_5	5	C/T	Sexual marker		
29194_28	28	A/G	Sexual marker		
29403_23	23	A/G	Sexual marker		
29441_2	2	C/T	Sexual marker		
29477_13	13	C/T	Sexual marker		
29543_13	13	A/C	Sexual marker		
29574_8	8	C/T	Sexual marker		
29780_14	14	C/T	Sexual marker		
29814_9	9	C/G	Sexual marker		
29919_15	15	A/G	Sexual marker		
30079_16	16	A/G	Sexual marker		
30093_13	13	C/T	Sexual marker		
30138_2	2	C/T	Sexual marker		
30632_2	2	A/G	Sexual marker		
3094_23	23	A/G	Sexual marker		
3108_14	14	C/T	Sexual marker		
3152_3	3	A/G	Sexual marker		
3186_5	5	C/T	Sexual marker		
3224_2	2	C/T	Sexual marker		
3234_15	15	C/T	Sexual marker		
3254_14	14	C/T	Sexual marker		
3270_5	5	A/C	Sexual marker		
34800_2	2	C/G	Sexual marker		
3633_23	23	C/T	Sexual marker		
3677_6	6	A/G	Sexual marker		
3860_28	28	A/G	Sexual marker		
3883_7	7	A/T	Sexual marker		

3975_22	22	C/T	Sexual marker
4040_26	26	C/T	Sexual marker
4073_24	24	A/G	Sexual marker
4340_27	27	G/T	Sexual marker
4359_22	22	C/G	Sexual marker
4523_17	17	A/G	Sexual marker
4629_23	23	C/T	Sexual marker
4679_4	4	C/T	Sexual marker
4766_26	26	A/C	Sexual marker
4825_29	29	A/G	Sexual marker
5070_5	5	C/G	Sexual marker
5152_17	17	A/G	Sexual marker
5180_7	7	A/G	Sexual marker
5261_2	2	A/G	Sexual marker
5263_4	4	A/G	Sexual marker
5315_28	28	G/T	Sexual marker
5506_14	14	G/T	Sexual marker
5514_13	13	C/G	Sexual marker
5576_15	15	A/C	Sexual marker
5585_3	3	A/G	Sexual marker
5597_27	27	A/C	Sexual marker
5611_6	6	C/T	Sexual marker
5817_8	8	A/G	Sexual marker
592_26	26	C/G	Sexual marker
6158_17	17	A/G	Sexual marker
6300_22	22	A/C	Sexual marker
6448_5	5	A/G	Sexual marker
6472_2	2	A/G	Sexual marker
65_5	5	C/T	Sexual marker
6594_25	25	A/C	Sexual marker
6622_2	2	A/G	Sexual marker
6704_7	7	G/T	Sexual marker
6718_28	28	G/T	Sexual marker
6819_9	9	A/G	Sexual marker
6977_5	5	A/G	Sexual marker
7050_8	8	G/T	Sexual marker
7053_5	5	A/G	Sexual marker
7110_8	8	A/G	Sexual marker
7275_3	3	C/G	Sexual marker
7429_15	15	C/G	Sexual marker
7437_2	2	A/G	Sexual marker
7563_14	14	G/T	Sexual marker
7703_17	17	A/G	Sexual marker
8084_4	4	C/T	Sexual marker

8101_24	24	C/G	Sexual marker		
8145_17	17	A/G	Sexual marker		
8241_14	14	A/C	Sexual marker		
8256_16	16	A/G	Sexual marker		
8526_25	25	C/T	Sexual marker		
8548_29	29	A/G	Sexual marker		
8724_26	26	G/T	Sexual marker		
8997_25	25	G/T	Sexual marker		
913_8	8	C/T	Sexual marker		
9180_24	24	A/G	Sexual marker	Variabilichromis moorii voucher Matthew D. McGee:4237 ultra conserved element locus uce-981 genomic sequence	9,25E-05
9262_7	7	C/T	Sexual marker		
9653_13	13	A/C	Sexual marker		
9668_9	9	A/G	Sexual marker		
9775_26	26	A/G	Sexual marker		
982_29	29	A/G	Sexual marker		
9857_14	14	C/T	Sexual marker		
9860_9	9	A/T	Sexual marker		
9874_23	23	C/T	Sexual marker		
9905_29	29	A/T	Sexual marker		
15022_2	2	G/T		PREDICTED: Larimicthys crocea cortactin (cttn), transcript variant X4, mRNA	7,00E-05
15550_28	28	G/T		PREDICTED: Notothenia coriiceps rho guanine nucleotide exchange factor 10-like (LOC104954371), mRNA	7,00E-05
169_29	29	C/T		PREDICTED:Sinocyclochelius rhinocerous vescicular glutamate transporter 2.1 (LOC107740422), mRNA	6,00E-06
23606_26	26	C/G		PREDICTED: Stegastes partitus tenascin-like (LOC103356464), mRNA	2,00E-05
4556_28	28	C/T		PREDICTED: Nothobranchius furzeri ATP-binding cassette sub-family A member 1-like (LOC107391047), mRNA	7,00E-05
7456_4	4	C/G		PREDICTED: Maylandia zebra extracellular calcium-sensing receptor-like (LOC101474764), mRNA	2,00E-05

The development of measure to monitor and reduce the genetic impact of aquaculture to natural environment requires a deep knowledge of the species at wild level (i.e. genetic structure, diversity within populations) and at farm level (i.e. number of farms operating, approaches to selection and the genetics of the broodstocks). In the context of the Aquatrace, information about the state of the aquaculture of the three target species was collected through a specific survey. To complete the background knowledge, we performed a study of the genetics of the gilthead sea bream, focused on understanding the genetic arrangement of the wild populations and that of the major broodstocks. The results stimulated a discussion on the potential impact of sea bream aquaculture in the European seas.

ddRAD SNPs markers reveal subtle genetic structure of Gilthead Sea Bream (*Sparus aurata*) in European wild populations and high divergence between farm broodstocks: implication for aquaculture and natural stock management

Maroso, F. [a 1], Gkagkavouzis, K. [b 1], De Innocentiis, S. [c], Tryantafyllidis, A. [b], AquaTrace Consortium, Bargelloni, L. [a]

[a] Department of Compared Biomedicine and Food Science, University of Padova, 35020 Legnaro, ITALY
[b] Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
[c] Italian National Institute for Environmental Protection and Research, Marine Molecular Biology Lab (BMM), Rome, Italy
[1] These authors contributed equally to the paper

Keywords**:** *gilthead seabream, aquaculture, population structure, introgression, stock management*

Abstract

Farming marine fish in sea cage systems increases the probability of massive escapes and can lead to gene introgression across farmed and wild individuals. The gilthead sea bream (*Sparus aurata*) is one of the most important fish species farmed in the Mediterranean and trade of sea bream eggs and breeders as well as selective breeding further increase the aforementioned risks. A detailed assessment on the genetic composition of both wild and farmed populations is fundamental for a proper risk management. More than 1,000 sea bream, from 16 wild populations and eigth major European hatcheries were analyzed at 1240 high quality SNP markers based on double digest Restriction-site Associated DNA (ddRAD). Weak population structure was detected in the wild, suggesting shallow genetic differentiation between Atlantic and Mediterranean basins and, within the Mediterranean, along the West-East transect. Broodstocks displayed a stronger differentiation, likely due to genetic drift associated with reduced population size and selective breeding/domestication, which also affected genetic variability (lower in farmed populations). Strong genetic divergence was detected between wild and farm samples, though broodstocks with less generations of captivity appeared genetically similar to their wild counterparts. Allele frequencies at loci potentially under selection were divergent in farm and wild samples. Our results suggested that risks for wild population might exist in case of escapees or restocking, in terms of loss of genetic variability and fitness. Preliminary signs of presence of farmed individuals into the wild were also observed. Overall, the approach we developed (based on genetic markers) and the results obtained could be used to improve sea bream farming and to implement sustainable aquaculture practices through effective risk management.

Introduction

Sea bream is a demersal species living in warm coastal euryhaline waters of Mediterranean Sea and North-East Atlantic Ocean. It is highly appreciated for the quality of its flesh and it is an important target for commercial as well as for recreational fishing. Nevertheless, harvest from natural population contributes nowadays only to a very small percentage of total sea bream production. In fact, while capture fisheries have provided almost constant production since the 60’s (around 8000 tons per year), aquaculture production has increased constantly from the early nineties and reached almost 160000 tons of production [36]. Sea bream can be farmed in various ways: in coastal ponds and lagoons (extensive and semi-intensive methods) or in land-based installations and in sea cages (intensive farming systems).

These methods are very different, especially regarding fish farming density and food supply.

Following the increase diffusion of farming, concerns arose regarding the potential effect of these practices for the natural populations. For example, introduction of farmed fish can happen intentionally through restocking or unintentionally through sea cage escape events. The extent of introduction can go from few individuals that might be lost from cages during routine operations or leaking, to very important escapes of hundreds of individuals in case a cage is damaged or breaks. Since marine cage culture is rapidly evolving, new risks to natural populations emerged from escape events [113,114]. In most cases offspring are interested, as breeders are usually kept in breeding tanks located inland. Hence, the way offsprings are produced should be taken into account when monitoring their impact. From a broader point of view, effects of introduction can be either ecological or genetic, but often natural populations are affected in both ways. The severity of genetic consequences depends on many factors related to both the origin of the released/escaped fish and the genetic characteristic of local population that is affected. Waples (2012) [14] effectively described the possible effects dividing them into three aspects described below and indicated genetic parameters that could be used to study and face these issues. Although this report is focused on salmonids, that differ from other marine fish for ecological reasons and for the commercial/breeding practices used, still the “salmon example” provides a useful tool to use as a guide for management plans of less studied species (i.e. sea bream).

Consequences on genetic arrangement include modification to (i) population genetic structure, (ii) variability and (iii) adaptation to environment, all of which play fundamental role in the future survival of the species in the wild. This risk is enhanced by the fact that, along with increase in mass production, also exchange of fingerlings and breeders, as well as selective breeding practices, are becoming more common.

It is therefore important to better define and understand i) the population structure of the species and ii) the broodstocks genetic structure (i.e. origin, variability, breeding practices…) of the main breeding companies operating in the areas where natural populations occur. This information will be used to assess the potential effect of introgression through a comparative analysis of wild populations and broodstocks addressing variability, genetic pattern of diversity and effects at traits under natural selection.

The current knowledge of the genetics of gilthead sea bream is scarce and fragmented for both wild and reared populations. Previous studies of the natural genetic structuring along sea bream distribution area didn’t provide a consistent scenario, and while some surveys report absence of genetic differentiation between basins [115], other reported subtle genetic structures or even evident population subdivision even at small geographical scale [30,35,116]. Remarkably, these works are mainly based on markers (e.g. microsatellites, mitochondrial DNA) that are normally outperformed by SNPs for defining population structure. For what regards aquaculture, there is little or none public available information on the origin of broodstocks, nor on the exchange of breeders, eggs or juveniles. In addition, unlike salmon, origin farm traceability tools are missing for this species, despite the increasing interest of the consumers for the geographical origin of the food they eat [117]. New genetic analysis techniques and approaches offer nowadays the opportunity to study the genetic impact at an accuracy level never achieved before. For example, Next Generation Sequencing (NGS) and Restriction site Associated DNA (RAD) have already been proved efficient to detect hidden population structures and to better define already known genetic subdivision and differentiation in fish [82,118–120]. A higher resolution in terms of marker density also increases the chance of finding genomic regions affected by natural selection [121], that could therefore be used in a breeding context to select best breeders or, in a conservation context, to assess the potential effect of introducing in the wild animals adapted/selected to farm environment. For farm industry, it could also be possible to trace products, record and quantify exchange of fry and broodstock among countries. Tools for geographical traceability of wild captured fish can be developed, which is one of the most important and, at the same time, most difficult task for management and for fighting illegal, unreported and unregulated (IUU) fishing. Another important advantage of RAD genomic libraries preparation/bioinformatic procedure is the possibility of studying species that lack well developed genetic reference, using *de-novo* approaches that exploit the high throughput of NGS to extract reliable data.

The development of traceability tools is actually one of the main goals of EU FP7 funded ‘Aquatrace’ project, that also aims at understanding the genetic impact of fish from aquaculture in the wild environment. Thanks to the efforts of more than 22 partners, it was possible to collect samples from almost the entire distribution area of the species and have temporal replicates for some areas. Similarly, farmed samples were collected from farm directly involved in the project and in several countries thanks to partners from different nationalities, all well connected with breeding and farming reality of their origin country.

In this paper we present the result of a broad scale population genetic study of gilthead sea bream (*Sparus aurata*), including both wild and farmed samples. More than one thousands SNPs were screened using one of the most recently developed genotyping technique, namely double digest RAD (ddRAD). We analyzed the genetic structuring of wild population and the genetic arrangement of broodstocks of many farms; we thus compared and discussed the results obtained in the light of the potential effects of escapes or intentional release of farmed sea bream in the wild. The results represent a step toward a deeper knowledge of the genetics of the species and the discussion on the potential effect of aquaculture can feed the debate on the management practices needed to protect wild sea bream populations, improving at the same time the growth of its aquaculture production in Europe.

Material and Methods

A total of 601 wild individuals from 16 different locations, covering great part of the distribution area of the species, were sampled for this work (Fig 1).

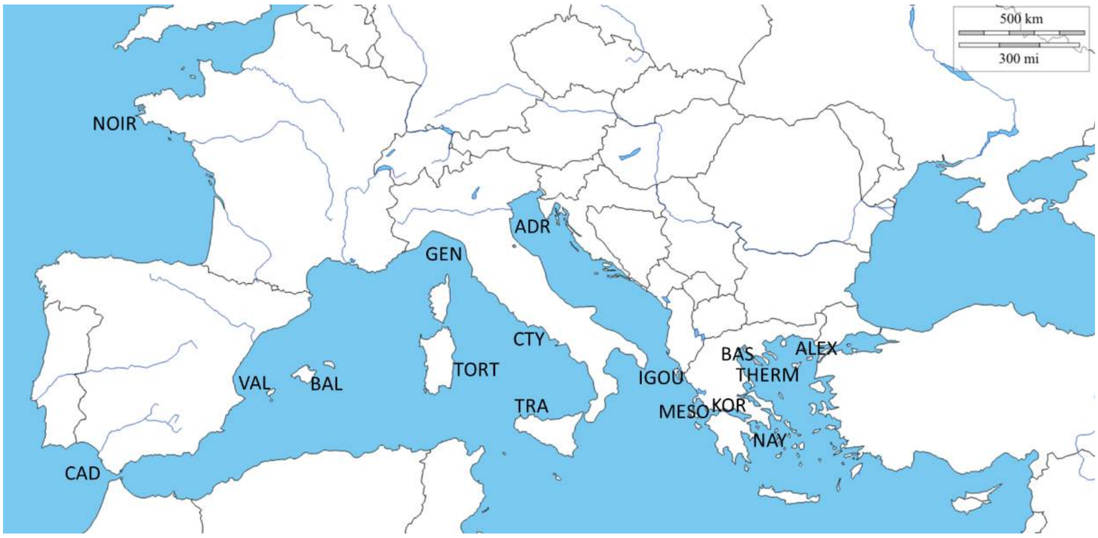


Fig 1 Population maps for wild samples, indicating the geographic positions of 16 wild samples from Atlantic and Mediterranean Sea.

Specimens were either collected specifically in the context of Aquatrace project or had already been collected and were provided by the partners of Aquatrace. Additionally, eight different aquaculture broodstocks were sampled in Greece, Israel, France, Italy, Malta and Spain specifically for the AT project, adding a total of 559 individuals to the sample dataset (Tab 1).Only origin country of sampled broodstocks are reported here, while farms' names and detailed locations were kept reserved for privacy reason in agreement with project partners. Information about the sampled broodstocks were recorded, such as the number of generation of selection (in case of ongoing selective breeding programs) and the presumptive geographical origin of breeders (where available). Samples consisted in either fin clips or muscle tissue, preserved in 95% ethanol as soon as possible after sampling. Genomic DNA was extracted using a commercial kit (Invisorb® Spin Tissue Mini Kit (Invitek, STRATEC Biomedical, 242 Germany) or the SSTNE buffer, a modified TNE buffer added of spermidine and spermine [43], that allowed a more efficient (thought more time consuming) extraction for samples that failed with commercial kits.

Multiple ddRAD libraries were prepared, each including 144 samples and splitting samples from the same population in different libraries, in order to avoid confounding library-specific biases. Library preparation protocol followed the original one of Peterson et al. (2012)[42], with some modifications that facilitate the screening of large number of individuals (see Supplementary Material S1). Libraries were sequenced on Illumina HiSeq 2500 sequencers with pair-end (PE) 100 base option to allow sequencing of both barcodes at the Genomics Core of the University of Leuven, Belgium.

Tab 1 List of wild and farmed samples used in the present paper. Level of selection is indicated as number of selected generations. Number of samples indicate the number of individuals analyzed after filtering for those samples genotyped at least at 80% of the markers), while in brackets the number of individuals before filtering.

Type	ID	Location	Long	Lat	N samples	Year	Ho	He	% polym. SNPs
W	NOIR	Noirmoutier (FR)	-2,169	46,988	22	2003	0,129	0,146	69,4%
	CAD	Cadiz (SP)	-5,953	36,263	21	2001	0,138	0,147	65,9%
	VAL	Valencia (SP)	-0,281	38,289	44	2009-2014	0,128	0,146	80,8%
	BAL	Balearic Is. (SP)	2,680	39,403	36 (37)	2013	0,137	0,151	78,7%
	GEN	Genova (IT)	8,900	44,359	33	2005	0,129	0,148	74,9%
	CTY	Central Tirrenean (IT)	12,624	41,405	52 (54)	2013	0,136	0,151	84,1%
	TORT	Tortolì (IT)	9,756	39,924	29 (30)	2002	0,136	0,150	76,6%
	TRA	Trapani (IT)	12,449	38,005	22	2007	0,137	0,150	71,2%
	ADR	North Adriatic (IT)	12,409	45,321	40	2014	0,115	0,136	71,5%
	IGOU	Igoumenitsa (GR)	20,162	39,485	53	2006	0,134	0,150	82,4%
	MESO	Mesologgi (GR)	21,314	38,303	49	2005	0,133	0,150	82,2%
	NAY	Nayplio (GR)	22,757	38,045	32 (36)	2013	0,135	0,151	76,2%
	KOR	Korinthiakos (GR)	22,944	37,270	32 (33)	2005	0,135	0,151	75,8%
	BAS	Basova Kavalas (GR)	24,495	40,846	29 (34)	2013	0,132	0,146	72,4%
	THERM	Thermaikos gulf (GR)	22,846	40,262	45 (46)	2013	0,132	0,149	80,8%
	ALEX	Alexandroupolis (GR)	25,916	40,777	46 (47)	2013	0,132	0,149	80,0%
	Farm ID		Level of selection						
F	FARM 1	GREECE	3		78	2014	0,139	0,150	68,5%
	FARM 2	GREECE	1		88	2014	0,131	0,148	73,3%
	FARM 3	GREECE	1		56	2014	0,142	0,150	76,8%
	FARM 4	ISRAEL	6		174	2014	0,130	0,147	75,4%
	FARM 5	FRANCE	4		66	2014	0,125	0,143	67,8%
	FARM 6	MALTA	2		39	2014	0,133	0,151	78,9%
	FARM 7	ITALY	1		34	2014	0,133	0,144	66,9%
	FARM 8	SPAIN	1		24	2014	0,135	0,148	66,6%

Raw reads were checked for quality using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, reads missing valid restriction site were discarded, and barcodes were searched (allowing up to 1 error) for demultiplexing. Barcodes were trimmed and remaining sequences cut to 90 bp length. Four bases at the end of each read were cut in order to increase the number of reads passing the filter and to obtain higher coverage at the end of the genotyping process. Reads with one or more uncalled bases were filtered out, as well as reads with 11 or more consecutive bases with average quality score less than 20 (1% error rate). If a sample was sequenced on more than one lane, reads were combined into a single file before processing. Stacks 1.3 [58,59] was used to cluster reads into consensus tags and call high quality SNPs. Typical *de-novo* pipeline was run (refer to Stacks' website for details about how the pipeline works). Main clustering parameters used were -m (minimum depth of coverage to call a stack) set to 4; -M and –n (maximum number of differences between stacks to be considered as the same tag in *ustacks* and *cstacks*, respectively) set to 7; SNP calling model was set to 'bounded'. Correction module *rxstacks* was run after the analysis to correct genotypes based on population-wide information. As including all samples in the catalogue would be prohibitively slow with the version of Stacks used, 500 samples were selected for this step, including those with higher number of reads from each population, in order to have all of them represented. Examination of SNP distribution along the length of the read revealed a dramatic increase towards the end of loci. These polymorphisms were not true SNPs but were due to indels, that caused a shift in the alignment resulting in all/many nucleotides after the indel being erroneously identified as SNPs. Frequently these indels appear to occur within SSRs. For this reason, SNPs at the very end of the reads (last two bases) were discarded and tags with more than five SNPs were filtered out. Additionally, an *ad-hoc* program was used to identify and remove false SNPs that arose from indels. Finally, SNPs were filtered out when they were found in less than 80% of the analyzed samples and when Minor Allele Frequency (MAF) was lower than 0.5%. Similarly, samples were filtered in order to retain only those genotyped at more than 80% of the remaining markers.

Four samples were replicated 12 to 13 times in different libraries in order to assess genotyping precision of the library preparation and sequencing techniques. For each group of replicates and for each marker, the most frequent genotypes were considered as the correct one and the number of mismatches were recorded through the entire filtered SNP dataset.

GenAlEx 6.501[86] was used to calculate expected (He) and observed (Ho) heterozygosity, percentage of polymorphic markers, population specific allele frequencies and to detect private alleles in both natural populations and broodstocks. Deviations from Hardy-Weinberg equilibrium was also tested for each locus and for each population.

Fst matrices were calculated with Arlequin 3.5.2.1 [90] and significance were calculated with 50,000 permutations. Sequential Bonferroni correction were applied when multiple tests were performed simultaneously.

Linkage Disequilibrium between loci arises when the frequency of association of alleles at different loci is higher than expected if the loci were independent. Physical association, proximity in the genome and co-selection of traits located far apart in the genome can cause loci to be in LD. From an analytical point of view, statistically linked markers can seriously affect downstream analysis that assume loci are independent. While recent versions of many programs (i.e. Structure, from version 2.0) cope with this issue, some other (e.g. software for Effective Population Size estimation) are not able to deal with them yet. We searched loci pairs for unusually high LD using r^2 estimator implemented in Plink [87], parsing all loci pairs in the wild populations.

Two different approaches were used to summarize and visualize genetic relationship between groups: the model-based clustering method implemented in Structure [96] and Discriminant Analysis of Principal Components (DAPC) as implemented in *adegenet* R package [122–124].

Structure 2.3.4 was run through Parallel Structure [125] to allow faster and more efficient parallel running using different *k* values and replicates of each *k* value. We run separately all the wild samples, using 'a-priori' information; all the farmed samples and wild and farmed together to be able to highlight similarities between farmed broodstock and wild counterparts. All the analysis run with k ranging from one to ten, each repeated three times to allow evaluation of likelihood of different population structures. Burn in was set to 50’000 and number of iterations to 100’000. Results from different runs were collated and most likely k values were detected using the Evanno’s method implemented in STRUCTURE HARVESTER [98].

An alternative way to search for optimal number of cluster and visualize samples subdivision is Discriminant Analysis of Principal Components (DAPC). The approach puts together the advantages of reducing the number of explanatory variables provided by the traditional Principal Components Analysis (PCA), with the power of Discriminant analysis (DA), that separates between- and within-groups variability extracting variables that influence most the first, while overlooking the latter. We used its implementation in *adegenet* to understand the genetic relationships between our wild populations. To avoid the effect of retaining too many principal components (PC), which would discriminate better the sampled individuals, whilst performing poorly with newly sampled ones, repeated cross-validation was used to select the best number of SNPs and to obtain a trade-off between stability and power of discrimination. Finally, the markers with major influence in group division were detected for any discriminant axis, providing lists of SNPs potentially under differential selection in different populations.

The possibility to calculate Effective Population Size (N_e) from a single sample (i.e. without temporal replicates) is appealing for conservation biologist for its potential to provide useful information about the health status of a population [126]. Additionally, this parameter can be used in broodstocks analysis to assess the level of diversity of a particular group and therefore help in estimating the potential effect on natural genetic variation in case of escapees or intentional release. Nevertheless, at the moment, estimating this parameter is a challenging statistical problem faced by conservationists and bioinformaticians. In this work, we estimated contemporary Ne of our natural populations and broodstocks using a single sample method based on Linkage Disequilibrium between loci that arise when populations with low Ne are sampled. The algorithm is implemented in NeEstimator 2.01 [127]. Ne estimations were calculated from polymorphic SNPs with minor allele frequency (MAF) >1%.

Pairwise genetic relatedness between individuals were calculated with Coancestry 1.0 [84]. This analysis is commonly used to assess the level of inbreeding in broodstocks, which is an important parameter to avoid inbreeding depression. In our case, the same method was used to detect unexpected high relatedness between individuals from wild (which is not expected considering that wild populations are usually of large sizes) or between wild specimens and breeders. High relatedness might indicate sibling individuals or parent-offspring pairs, and both can be considered as signs of the presence of animals of farmed origin. In the first case siblings might have escaped from sea cages (where usually individuals from the same few families are kept); in the second case offspring from farms are directly detected and linked to the parent breeders.

One of the most interesting and useful advantages of genome wide genotyping is the increased chance of finding loci potentially under natural selection. These markers can be used to highlight finer-scale genetic structure in wild populations and to link genetic and phenotypic traits selected in a particular environment. Several approaches that aim at detecting outliers have been proposed, but today the two most used are probably the Bayesian approach implemented in Bayescan [92–94] and the Fdist approach, that aims at detecting unusually high or low locus specific Fst values [128]. We run both on our wild dataset to search for loci showing higher than expected differences in allele frequencies. Bayescan 2.1 was run with default parameters. Fdist algorithm was run through the graphical interface implemented in Lositan [91], performing 50,000 simulations, with '"Neutral" mean Fst' and 'Force mean Fst' options checked, a confidence interval of 0.99 and False Discovery Rate (FDR) of 0.01. Bayescan and Lositan were also used to detect markers showing unusual differences in allele frequencies between broodstocks, which could be used as tools to detect the origin of fish from farms.

To have a more robust and more representative outlier list for wild populations, we also searched for loci whose allele frequencies showed unusual correlation with environmental variables (i.e. loci potentially involved in local adaptation) using the Bayesian approach implemented in Bayenv [129]. Results are given in terms of Bayes Factor (BF), and normally BFs higher than two indicate highly probable outliers. As suggested by the authors, we double checked also correlation values between allele frequencies and environmental variables, as sometimes high BF can be obtained also with low correlation, but this results should be taken with care. Environmental data was extracted from SeaDataNet portal (http://www.seadatanet.org/) and included temperature and salinity values at surface and at 20 m, as these represent functional proxy indicators of more complex environmental variation. We referred to the geographic coordinates as close as possible to actual sampling locations for which data were available.

Finally, we searched for loci most influencing wild populations genetic differentiation looking at loci contribution to DAPC axes provided by *adegenet.* Different groups were separated along different explanatory axes, and we could then differentiate loci according to their effect on different axes.

After collecting information from four different approaches to outlier discovery, “outlier panel” was defined selecting those loci detected as potential outliers by at least two of the methods used. A combination of different methods is indeed advised to obtain more information from the data [130]. Differently, we defined the neutral dataset excluding all those loci found by at least one of the approaches used.

A comparative analysis between farmed and wild individuals, based on outlier loci, was used to understand the potential impact escapees or released animals can have on the overall fitness of wild populations. We tested if there were any significant pairwise difference between wild and farmed populations at outlier loci with Fst test implemented in Arlequin. Additionally, we compared allele frequencies in wild and farmed groups for each outlier locus to allow for pairwise comparison for each wild population-broodstock pair.

Finding potential links between genetic data and phenotypes is interesting to better understand and validate results from genetic data and to highlight potential effects on fitness of different populations. With this aim, we use BLAST to try and annotate RAD tag sequences containing outlier markers.

Results

Initial number of SNPs before filtering was 11662. After filtering out low quality markers, a total of 1246 SNPs (10.6%) were retained (suppl. Material table S2). Seven wild individuals were excluded due to low quality genotypes, so that a total of 1144 individuals, 585 wild and 559 farmed individuals, were kept for subsequent analysis. After filtering, the level of missing data per sample ranged from 0.8% to 19.8%, with an average of 5.1%.

22298 tests for departure from H-W equilibrium were carried out. After sequential Bonferroni correction, only two loci showed significant deviation from H-W equilibrium (both for excess of heterozygous) in more than half of the natural populations and were excluded.

A total of 767560 tests for LD were carried out and only four loci pairs showed r^2 values higher than 0.7 and for each pair, the locus with lower missing data was retained. Remaining 1240 SNPs were used for subsequent analysis.

Out of 63240 tests, the level of mismatch within replicates at 1240 filtered loci ranged from 3.4% to 5.8 %, with an average of 4.0%.

Natural pop analysis: genetic structure and outlier detection

The overall level of differentiation between wild populations was low (Fst=0.45%). Pairwise Fst values calculated on 1240 SNPs ranged from 0 to 1.9% (CAD-THERM comparison, see Suplementary Material S3) and were significant mostly in the comparisons between Atlantic samples and Mediterranean samples. Within Mediterranean basin, values ranged from 0 to 0.7% and tend to increase and be more significant in the comparisons between samples from Western Mediterranean and the Eastern part of the basin (Ionian and Aegean basins).

Fst values suggested that wild populations are differentiated into two main groups: Atlantic and Mediterranean. A further (but weaker) subdivision within the Mediterranean basin was found by clustering analysis, that separated West Mediterranean basin (WMED, including VAL, BAL, GEN, CTY, TORT and TRA) from Ionian (ION, including IGOU, MESO, KOR and ADRIATIC) and Aegean basin (AEG, including NAY, BAS, THERM, ALEX).

Structure suggested the presence of a weak subdivision within the Mediterranean. In the analysis with the full 1240 SNPs dataset, the most likely *k* values according to Evanno’s method was *k*=2. Anyhow, a “spurious” cluster was identified that separated five Cadiz samples from the others (see Supplementary Material S4). Such a problem can arise when closely related individuals are present among a group of non-related individuals and it is known to affect Structure clustering analysis [85]. In fact, pairwise relatedness values for these samples were much higher than average in the same group, suggesting that those animals could be either siblings or parent-offspring pairs (i.e. relatedness around 0.5). After removing this cluster, the three remaining were used to describe population structure. Individuals are generally admixed but samples from west Mediterranean (VAL, BAL, GEN, TORT, CTY, TRA) appeared differentiated from samples from Ionian Sea (IGOU, MESO, KOR) and those from Aegean (BAS, THERM, ALEX, NAY), that formed two separated groups (Fig 2).

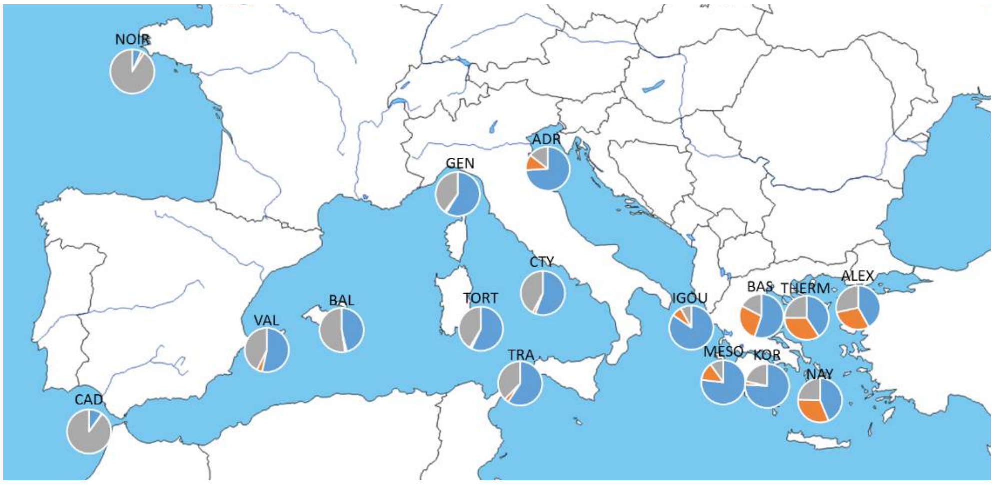


Fig 2 Pie chart representation of populations' average admixture patterns suggested by Structure. The analysis was carried out simulating four clusters (*k*=4). Different admixture patterns can be identified in ATL, WMED, ION and AEG populations (see Supplementary Material S11 for Structure's typical plot chart).

ADR samples looked more similar to Ionian Sea samples. When grouped in this way, these four clusters showed low (from 0.3% to 1.4%) but highly significant ($p<0.001$) Fst values for all pairwise comparisons.

DAPC analysis was based on 150 PCs, after cross validation analysis. Scatterplots confirmed the pattern of differentiation for wild samples already detected by Structure, that is a stronger separation between Atlantic and Mediterranean samples along first (i.e. the most discriminant) axis and a weaker differentiation within Mediterranean along the second and the third axis (see Supplementary Material Figure S5).

Private alleles were searched at two different levels: analyzing all the populations separately, 13 loci showed private alleles across 16 wild populations, with a frequency ranging from 1% to 9.5%. Using four groups identified by Structure and DAPC analysis (ATL, WMED, ION and AEG) 22 private loci were identified, with frequency ranging from 0.2% to 4.7% (see Supplementary Material Table S2)

A total of 60 SNPs were identified as potential outliers by at least one of the approaches used in the analysis with 16 wild populations of sea bream (see Supplementary Material Table S2).

Bayescan detected in total 12 loci with log10(PO)>2; Lositan detected 13 potential diverging outliers (plus four outliers for balancing selection, that were not considered in subsequent analysis); the outlier approach based on contribution of loci to DAPC Discriminant Factors (DF) detected 42 loci whose contribution to either first (16), second (12) or third (14) DF were higher than 0.7%. Based on correlation between allele frequencies and environmental factors, we detected 12 loci with Bayes factor higher than 2 and correlation higher than 0.5. Overlapping and differences between different approaches are highlighted in the Venn graph (Figure 3). Only one marker (8727_39) was identified as outlier by all the approaches used.

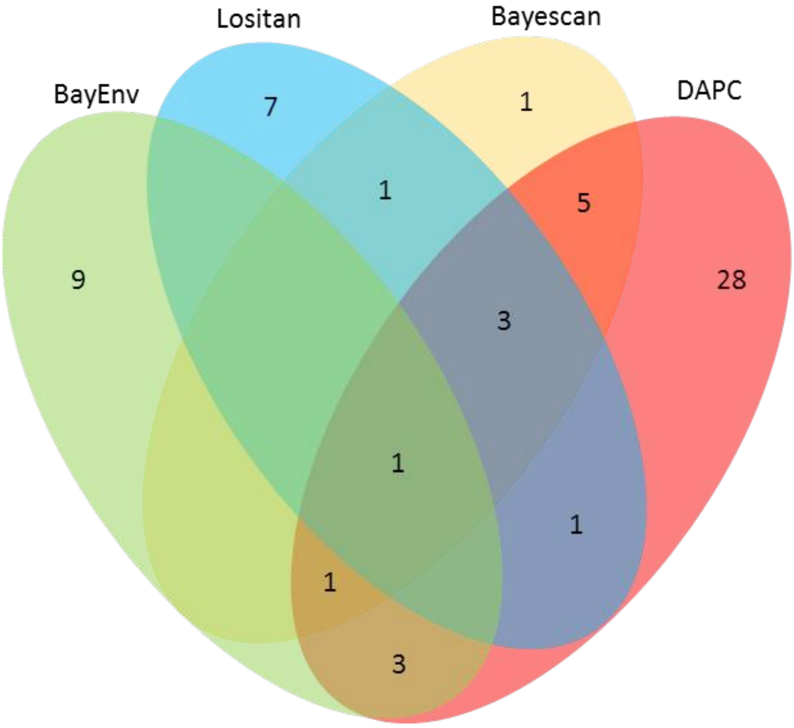


Fig 3 Venn diagram showing the results of different outlier detection methods

Different stringencies were used to define an outlier panel and a neutral panel (see Supplementary Material Table S2). A total of 15 loci were selected to create the ‘outlier dataset’, that was subsequently used for analysis focused at understanding the functional divergence between populations.

The pattern of differentiation at locus-specific level was studied plotting allele frequencies of OL loci at different populations arranged in west-to-east order (see Supplementary Material S6). The frequency patterns of many loci showed abrupt change in mean allele frequencies between groups identified with genetic analysis presented above (e.g. locus 8727_39 or 13776_28), or gradient of change moving from more western to more eastern populations (e.g. 2689_62 and 10524_58), supporting the subdivision previously hypothesized.

The most remarkable match found with BLAST analysis of 15 OL was locus 8727_39, that showed high similarity with sea bream’s Carnitine Palmitoyltransferase 1B mRNA (98% identity, E-value 6E-47), a mitochondrial enzyme responsible for the formation of acyl carnitines and therefore involved in energetic metabolism.

Sixty loci were excluded from the entire marker dataset in order to define a neutral dataset, leaving 1180 SNPs as part of the neutral panel.

Genetic structure at these “non-selected” loci was much weaker than with the entire dataset, especially within the Mediterranean. This is visible both in Structure plot (Supplementary Material S7) and from Fst values (Supplementary Material S3). Differentiation between Atlantic and Mediterranean is still present and Fst are positive and significant at most of the pairwise comparisons.

Ne values for wild groups (NeW) at neutral and not linked markers aranged from 58.6 for Cadiz population to ‘Infinite’ (Table 2). Notably, the very low value of Cadiz sample is influenced by the presence of highly related individuals (see above) that biased the estimation downward. Indeed, removing these individuals and running again the analysis increased ten-fold Ne estimate (561.7).

Table 2 Effective Population Size values for 16 wild and 8 farmed population analyzed.
* CAD value is biased by the presence of highly related individuals. In brackets value calculated for this population after removing “potential siblings”

	Polymorphic	Lowest allele freq.		Polymorphic	Lowest allele freq.		Polymorphic	Lowest allele freq.
Population	loci	0.01	Population	loci	0.01	Population	loci	0.01
NOIR	794	Infinite	ADR	825	Infinite	FARM 1	804	63.1
Lower 95% bound		2536,2	Lower 95% bound		Infinite	Lower 95% bound		61.8
Upper 95% bound		Infinite	Upper 95% bound		Infinite	Upper 95% bound		64.5
CAD	751	58.6 (561.7)*	IGOU	943	11911.9	FARM 2	864	61.6
Lower 95% bound		54,8	Lower 95% bound		2518.3	Lower 95% bound		60.5
Upper 95% bound		62,9	Upper 95% bound		Infinite	Upper 95% bound		62.7
VAL	928	6213.5	MESO	944	Infinite	FARM 3	905	39.6
Lower 95% bound		1945.7	Lower 95% bound		11711.3	Lower 95% bound		38.8
Upper 95% bound		Infinite	Upper 95% bound		Infinite	Upper 95% bound		40.4
BAL	901	Infinite	NAY	875	1643.5	FARM 4	884	31.7
Lower 95% bound		61495.2	Lower 95% bound		863.5	Lower 95% bound		31.4
Upper 95% bound		Infinite	Upper 95% bound		15722.6	Upper 95% bound		32
GEN	858	1856	KOR	866	Infinite	FARM 5	797	114.5
Lower 95% bound		905	Lower 95% bound		Infinite	Lower 95% bound		110.1
Upper 95% bound		Infinite	Upper 95% bound		Infinite	Upper 95% bound		119.2
CTY	966	1833.4	BAS	830	Infinite	FARM 6	933	217.1
Lower 95% bound		1169.8	Lower 95% bound		Infinite	Lower 95% bound		199.3
Upper 95% bound		4191.9	Upper 95% bound		Infinite	Upper 95% bound		238.2
TORT	878	Infinite	THERM	928	Infinite	FARM 7	791	31.4
Lower 95% bound		Infinite	Lower 95% bound		31874.8	Lower 95% bound		30.7
Upper 95% bound		Infinite	Upper 95% bound		Infinite	Upper 95% bound		32.2
TRA	811	1712.1	ALEX	928	Infinite	FARM 8	785	75.9
Lower 95% bound		683.6	Lower 95% bound		5033.7	Lower 95% bound		71
Upper 95% bound		Infinite	Upper 95% bound		Infinite	Upper 95% bound		81.6

Farmed pops differentiation

Among the most striking features of broodstocks when compared to wild stocks is i) higher genetic differentiation between broodstocks and between broodstock and wild counterpart and ii) much lower genetic variability within broodstocks when compared to variability within wild populations.

Fst values and DAPC analysis suggest that broodstock populations are much more genetically divergent between each other. Fst values calculated on 1240 loci ranged from 1.2% to 5.7%, and all comparisons were highly significant also after sequential Bonferroni correction. *adegenet*'s DAPC based on 100 PC displayed a general pattern of high differentiation for farmed groups with FARM 2 and FARM 3 appearing genetically similar, as well as FARM 6, FARM 7 and FARM 8 that clustered together. (see Supplementary Material Figure S8). Also Structure analysis showed a best fitting number of clusters (*k*) equal to five, suggesting the same pattern of differentiation as DAPC. When focusing on the farmed groups separately, 64 loci showed presence of private alleles with frequencies reaching 14.5% in FARM 4 for locus 2379_6 (see Supplementary Material Table S2). In addition, 18 OL markers were found by Bayescan and Lositan in the analysis focused on farmed samples, of which two shared by the two approaches (see Supplementary Material Table S1). Broodstocks' Ne values (NeC) at neutral and not linked markers are much smaller than those recorded for natural populations (Table 2). Upper and lower 95% bond are close to the estimated values, which increase the confidence on this results.

Comparison between wild populations and broodstocks

While some broodstocks look similar to the wild counterparts, FARM 1, FARM 2, FARM 3 and FARM 4 are genetically divergent from the natural populations. DAPC scatterplot, based on 250 PCs, offered a clear visualization of the genetic structure of the groups (Figure 3).

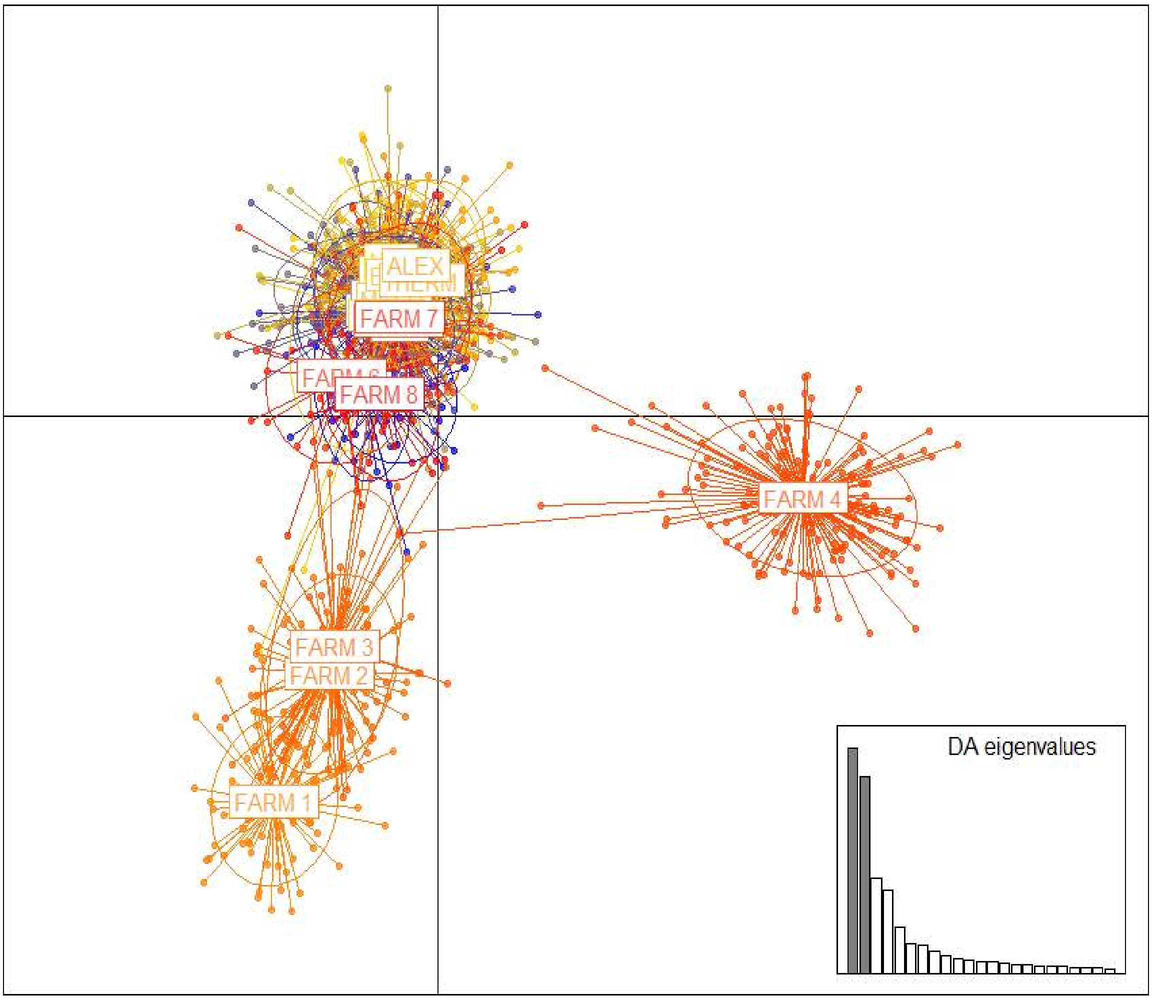


Figure 3 Scatter plot representation of DAPC analysis carried out for wild and farmed populations.

Similarly, Structure analysis including both farmed and wild groups indicated the same pattern of differentiation (Supplementary Material Figure S9). Unexpectedly, farms with higher number of generations of selection are more divergent from the wild counterpart only in some cases. Indeed, FARM 5 (4 generations of selection) looked more similar to wild cluster, while FARM 2 and FARM 3 (1 generation of selection) looked more divergent from the natural groups and similar to FARM 1. Using both wild and farmed samples in single analysis provided further signs of presence of escaped/released individuals among supposedly wild specimens. Two individuals in KORIN population showed admixture pattern similar to those of FARM 1 in Structure analysis.

Considering only the 15 OL, Fst diversity between farmed and wild populations ranged from 0 to 13.3% and tends to increase in pairwise comparisons between farms and ATL samples and between farms and ION-AEG samples.

At the level of single loci, comparison between farmed and wild populations highlighted that in some cases allele frequencies are different. Locus 13518_71 showed the most discordant pattern, with only one broodstock showing higher frequency for allele 1, as in all the wild, and all the other broodstocks showing inverted frequency than the wild populations. Locus 13129_86 showed a more variable pattern for both wild and farmed groups, with prevalence of allele 1 in some populations and allele 2 in other. Locus 8727_39 behaved similarly (see Supplementary Material S10).

Discussion

Understanding the genetic structure of wild population and major broodstocks is the first step toward the development of proper management of gilthead seabream. The possibility to analyze more than one thousand samples based 1216 SNPs allowed an accurate analysis of the genetic arrangement of natural populations and broodstocks of this important commercial species. Results collected stimulated a discussion about the potential effect of escapees/intentional release of fish with farm origin into the wild. In this paper we often use the “salmon example” for our discussion and conclusions.

For the first time to our knowledge, a population genetic study of this species was carried out with a high number of polymorphic markers and covering great part of the distribution area of the species. Previously, a similar broad range analysis was performed with allozymes and microsatellites by Alarcon et al (2004)[115]. In that case, authors concluded that structuring pattern could not be associated with geographic nor oceanographic known factors. In our work, more sensitive approaches were used to uncover hidden genetic structures. Clustering analysis suggests that genetic structure of wild populations is characterized by a weak subdivision into four main ”sub-basins”, following a geographic pattern: Atlantic, west Mediterranean, Ionian Sea and Aegean. Fst levels are lower than what usually found in fish (Fst=0.062; Ward, 2006[25]), in agreement with previous studies on the same species [115,116], but most of the pairwise comparisons between groups from different “sub-basins” are significant. A weak differentiation between Atlantic and Mediterranean basins persists also after removing loci that are potentially under environmental selection. This was already found using other typically neutral markers (i.e. microsatellites in Garcia-Celdran, 2016 [116]). In the same study, significant differentiation was also found between Atlantic samples from north and south coasts of Spain, that was not the case in our analysis (NOIR vs CAD). Analysis at neutral loci suggested that the differentiation between Atlantic and Mediterranean might have historical or demographical causes as major drivers, affecting neutral loci and persisting after OL removal. It is curious to see that CAD samples is more strongly differentiated than NOIR samples when compared to Mediterranean groups, despite its closer location. A similar result was found by Alarcon et al.[115], using allozymes and microsatellites. In their study the most differentiated population was in fact from the Atlantic south coast of Spain.

On the other side, differentiation within the Mediterranean might have arisen from adaptation to different environments, whose signature eventually disappeared when selected loci are removed from the database. Genetic structuring in the Mediterranean was previously found also by Ben-Slimen (2004)[33] analyzing samples from Tunisian coast using protein loci. These findings suggest that gene flow through Strait of Sicily is probably reduced, due to the depth of the strait and being sea bream a coastal species that usually doesn’t swim deeper than 150 m [131]. Anyhow, further studies specifically focused on small scale populations are encouraged to provide a more detailed view of the situation. Seabream’s undefined structure at basin level contrasts with higher differentiation found in other species with similar biological and ecological traits [115]. It is reasonable to think that both ecological and biological factors might be involved, such as bottleneck or expansions. In addition, steps are being done to understand the behavior of sea bream in the wild [132], that could be used in the future to explain the genetic structure of the species.

Analysis focused on broodstock provided interesting information on the genetics of eight of the biggest hatcheries in the European area. Despite different selection practices all the farms analyzed showed a much lower level of diversity than the wild counterpart and were characterized by higher between-groups differentiation than wild groups. This feature was already found by other authors, and affects reared seabream even after only one generation of selection [116]. The most likely causes have to be searched in the genetic drift acting in broodstocks, founder effects and by the practice of reintroducing offspring to increase the number of breeders available [115,133]. Because of high fecundity of sea bream and differential mating success, loss of variability is a serious issue when control of mating is not implemented [134]. In addition, as a hermaphrodite species, farmers can rely on the same individual first as male and then as female (or the opposite) and effective population size can be further reduced. In the last decade, selection practices have been implemented in several sea bream farms and nowadays gilthead sea bream is one of the species with higher number of selective breeding programs [28]. While more attention is being taken by farmers to avoid inbreeding depression, that is recognized as a serious threat to broodstock fitness, its signature might be still visible in those broodstock that started more effective selection program without completely changing breeders.

If broodstock are not properly managed and grown offspring are reintroduced as breeders, variability is expected to decrease when number of generation of selection increases [135]. From our analysis the correlation between Ne and level of selection was very weak, and some of the broodstocks for which no or low level of selection were declared showed lower variability than broodstock with established (> 3 generations) selection practices. This is an important point for breeders and is an indication that, if properly managed, broodstock can maintain acceptable level of variability and avoid inbreeding. Nowadays, often breeders from different basins are put together, in order to create fitter individuals (i.e. hybrid vigor). This raises issues about the possibility to detect the exact geographic origin of a breeder from genetic data. Nevertheless, with the data in our hand we don’t expect to be able to achieve high confidence in telling the exact geographical origin of a broodstock. In fact, differentiation between wild stocks is extremely low when compared to differentiation between natural and farmed groups, as can be seen from DAPC scatterplot comparing wild and farmed groups. On the other side, accurate traceability to the level of origin farm is more feasible. Indeed, high Fst values as those recorded between different broodstocks allow easier assignment based on genetic data. In addition, many private alleles and the presence of outlier loci could facilitate this task. Nevertheless, as some farms look weakly differentiated, high level of confidence cannot be achieved in some cases. Efficient tools for tracing origin of fingerlings and juveniles used in fattening farms are fundamental for the conservation and management of the species, as well as for implementing regulations aimed at preserving the natural stocks by the consequences of escapees and improperly conducted restocking. In this case, frequent and not monitored exchanges of breeders between farms pose a challenge to traceability. An approach based on origin farm traceability is already used in salmon aquaculture to monitor/prevent escapes, and its benefit has been recognized [136,137].

In addition to provide useful information of the genetic background of the species, these results can be used to understand the potential effects of aquaculture on the genetics of wild populations.

The first aspect of interest is the "Loss of diversity within wild populations". Due to differences in diversity values of wild population (NeW) and broodstocks (NeC), in case the two groups are mixed (as happens with escape and release events) the resulting genetic variability of wild counterpart (NeT) would decrease. More specifically, a formula has been proposed to calculate the extent of this variation in Ne [14]. The presence in the formula of a variable that takes in consideration the relative number of escapees suggests that escape of relatively low number of specimens shouldn't have deleterious effects on natural population. This is particularly true if one considers the dispersal capacity of sea bream [12]: after a short period from an escape/release event, individuals with farmed origin have probably dispersed over a wide area and therefore their effect on genetic variability of local populations should be less relevant. Therefore, while we acknowledge that caution should be taken when using these formula, we think that it can be useful when it is required to set a threshold for the amount of escapees that can constitute a real risk to natural stocks.

Another source of genetic variability that can be potentially eroded by aquaculture practices is the diversity among populations, which shapes the genetic structure of the species. Given the weak differentiation among wild populations, potential risk related to escapees might be reduced. Nevertheless, significant genetic differentiation was detected between the

Atlantic and Mediterranean basins. Despite being weak, this differentiation might be important for a species to respond to future environmental changes. Some gene flow is expected between Atlantic and Mediterranean basins and genetic contribution of one basin could, in the future, support the species in case of environmental changes affecting the other basin.

A third important point to be considered is the potential effect on fitness of wild population: as reported, we found signs of directional selection acting on some of the loci we analyzed. OL provided clues about the possible effects of farmed animals released/escaped into the wild. In particular, significant high Fst values and differences in allele frequencies at specific loci between some farmed lines and wild groups indicate that the spread of farm-originated individuals in the natural environment might change the overall fitness of the affected populations. Though, it is important to keep in mind that the data analyzed here are variation in allele frequencies that are statistically correlated with environmental variables. Whether they are actually linked to genes involved in local adaptation is to evaluate through more focused approaches. For the moment, we found an evidence that at least one of these loci is linked to putatively important phenotypic traits. Indeed, locus 8727 seems to be related to a protein involved in energetic metabolism, and might therefore be a sign of selection to different temperature or salinity.

Considering that restocking and sea cage fattening have been implemented for years nowadays, it is expected to find individuals of farmed origin already among wild populations. In fact, we found two signs of presence of farmed individuals among wild specimens. First, highly genetically related (i.e. full-sib) individuals were found in CAD sample, and they might come from a restocking action undertaken few years before the sampling in that area [138]. If confirmed, this result would be a good example of the possibility to recognize escaped/released individuals in the wild without knowing their origin farm. A limit to this method come from the fact that one needs to sample at least a pair of individuals coming from the same escape/release event, as it cannot be used as a single-individual based test. Alternatively, samples from local farms could be used to check if siblings of wild-caught specimens are present among the farmed individuals, as used in a recent paper by Glover et al. (2016) [136]. A second sign of presence of farmed individuals among wild specimens involves a couple of individuals from KORIN populations, that show admixture pattern very similar to those of farmed individuals from FARM 1. In this case, direct evidence of the origin of the fish was obtained, thanks to the fact that most breeders from the putative origin farm were sequenced.

In the perspective of a more efficient and accurate traceability of sea breams from farms, at least most of the breeders of major hatcheries should be genotyped, so that a genetic database of breeders from different farms can be created and used in case of escapees from unknown source. While farm assignment (or exclusion) is already a reality in salmon [137], this is not the case for sea bream yet.

Conclusions

An efficient and sustainable aquaculture is fundamental to guarantee fish food for the living populations without compromising the possibility for future generations to feed on the same resources. To pursue this aim, a deep knowledge of the biology and genetics of fish species is fundamental, as well as the genetic characterization of the aquaculture counterpart. In this paper, we exploit the potential of one of the most advanced genotyping technique available nowadays to improve our knowledge of the wild population structure of the gilthead sea bream and the genetic characteristic of some important European broodstocks. The results obtained allowed us to discuss the possible genetic effects of aquaculture on wild populations, in case of escapees or intentional release of farmed breams. For the future, we envisage that the development of aquaculture will be coupled with the development of accurate and reliable tools for estimating its effect on the natural environment. We are confident that the results and the discussion provided by this paper will be helpful for resources management and regulations but will also further stimulate the application of molecular approaches to farming practices, in order to increase aquaculture production in a sustainable way.

SUPPLEMENTARY MATERIAL

S1 Detailed library preparation protocol

The original protocol of Peterson et al. (2012)[42] involved processing each sample separately (i.e. restriction digestion, adapter ligation, fragment size selection, PCR amplification and purification, quantitation) prior to pooling into a single library for sequencing. A modified protocol (described in detail elsewhere;[72,73]), which was more convenient for screening large numbers of individuals, was used for this project. The methodology allowed for pooling of samples after the adapter ligation step, which greatly reduced the number of manipulations required, ensured consistent size selection within libraries and reduced construction time to two to three working days. Library preparation began with basic qualitative and quantitative assessment of extracted DNA samples. DNA quality was evaluated by gel electrophoresis (0.8% agarose 0.5x TAE) and concentration was accurately measured by fluorimetry with each sample being finally diluted to 7 ng/µL in 5 mM Tris pH 8.5. For a library (144 samples), individual DNA samples (21 ng) were first simultaneously digested with SbfI (recognition site CCTGCA'GG) and SphI (recognition site GCATG'C) restriction enzymes. An adapter mix comprising individual-specific barcoded combinations of P1 (SbfI-compatible) and P2 (SphI-compatible) adapters (compatible with Illumina sequencing chemistry) were then added / ligated. Adapters were designed such that adapter–genomic DNA ligations did not reconstitute RE sites, residual RE activity limiting concatemerization of genomic fragments. Each adapter included an inline five- or seven-base barcode, allowing for post-sequencing identification of individuals (P1-P2 combinatorial barcoding). The ligation reactions were terminated by heat inactivation and all 144 samples combined in a single pool. Following column purification of the pooled sample, DNA fragments in the range of 320 bp to 590 bp were size selected by agarose gel electrophoresis, followed by gel-based column purification. The eluted size-selected DNA template was then PCR amplified (14 cycles, 400 uL volume), column purified down to a 50 uL volume and then subjected to a further clean-up using an equal volume of AMPure magnetic beads (Perkin-Elmer, UK) (used in sea bream and turbot), to maximize removal of small fragments (less than ca. 200 bp). The final library was eluted in c.20 µL10 mM Tris pH 8.5.

S2 Table of SNPs

SNP	Outlier Wild	Outlier Farmed	Private	Private All. Freq.	Blast Match	e-value
10078_42		L				
1028_11		D				
10524_58	D,BS					
10527_9	D					
10601_55		L,BS	FARM7	0.076		
10734_15	D,BS					
11061_20			MESO	0.010		
11177_32			ALEX	0.011		
11292_67			FARM1	0.084		
11434_75			**FARM4**	0.129		
11530_57			FARM4	0.012		
11535_27	BE					
11697_75			FARM1	0.022		
11783_62	D	D				
11808_17	D	L,BS				
11829_46	D	D				
11921_18	D					
11978_7			FARM4	0.009		
1225_14			FARM2	0.006		
12380_10			FARM7	0.015		
12382_21		L				
12386_22	D					
12443_28		D				
12479_70			FARM1	0.006		
12494_20	L					
12615_64	L,BS		FARM2	0.011		
12743_53	D					
12969_12			BAL	0.014		
13024_5		D				
13053_40	D					
13124_73	L					
13129_86	D,BS					
132_61	BS					
13310_71	D,BS,BE					
13398_81	D					
13518_71	D,BS					
13574_31		L,BS				
13664_39	BE					
13674_61	D					
13732_16			FARM7	0.015		

13734_5			FARM4	0.040
13741_35			FARM6	0.019
13776_28	L,D,BS			
13810_61			FARM1	0.013
13928_33	D			
1417_27			FARM6	0.014
1452_7			CTY	0.010
1485_41			FARM8	0.042
1522_21			**FARM4**	0.139
15345_61			ADR	0.013
17347_79			FARM6	0.032
1819_36			FARM6	0.028
1881_71		L,BS		
1922_16				
1927_48		L,BS	FARM4	0.066
1994_73	D	D		
2023_56			FARM4	0.031
2182_62			FARM5	0.015
2344_51			FARM5	0.092
2347_81			FARM5	0.131
2359_76	BE			
2379_6			FARM4	0.145
2623_8	BE			
2689_62	L,D,BS			
270_23	L		FARM8	0.021
2824_79			BAS	0.017
2830_23			FARM4	0.003
2879_85		D		
2893_59		D		
3013_72		D		
3039_9	L		FARM4	0.029
309_46			FARM4	0.095
3185_18		D		
3230_60	D			
3233_44			FARM4	0.011
3274_31		D		
3299_40	D			
3441_67	D,BS			
3519_78			FARM8	0.021
3550_26		D		
3716_88			FARM4	0.019
4135_65	BE			
4410_30		D		
4439_49	D	D		
4448_77			FARM7	0.044
4455_78		D		
4466_70		D		
4504_57			FARM4	0.011

4539_26			FARM6	0.105
4636_26			FARM7	0.017
466_34			FARM4	0.039
4676_56	BE			
4715_35			FARM7	0.059
5068_16	L		CAD	0.095
5074_30			FARM7	0.015
5199_27	BE			
526_42			IGOU	0.019
5321_48		L,BS	FARM4	0.064
5440_33	D			
5470_34			FARM2	0.006
5517_59			FARM6	0.013
5636_7			FARM5	0.028
567_85			FARM1	0.084
571_51			FARM2	0.056
5717_84			FARM6	0.013
5836_11		L,BS	FARM1	0.077
5928_12	BE			
60_69	D			
6025_36			TORT	0.017
6441_64		D		
6632_78	D			
6755_85			FARM2	0.006
6857_36			FARM6	0.091
7017_85			FARM4	0.032
7045_86			FARM8	0.021
7148_37	D			
7170_38			ADR	0.014
7206_5	D			
7216_76	D			
7262_48	D			
7339_17			FARM8	0.021
7352_81	BE			
7400_19			FARM4	0.013
7416_10		D		
7501_60		D		
7594_36			VAL	0.011
7610_82	D			
7641_63	D			
7684_6			FARM4	0.053
7951_79		D		
8136_42	L			
8150_73	D			
825_13	D			
8278_44	D			
8301_44	L			
8327_14		D		

8466_15		D				
8657_46			FARM6	0.026		
867_76			VAL	0.011		
8727_39	L,D,BS,BE	L			sea bream's Carnitine Palmitoyltransferase 1B mRNA	6.00E-047
8813_23	D					
8835_47	D	D				
89_46			FARM5	0.038		
8913_85	L,D					
9006_10			FARM6	0.026		
9012_31			FARM4	0.015		
9025_25			FARM7	0.029		
9150_19		L,BS	FARM3	0.077		
9474_38			FARM1	0.013		
9633_68	D,BE	D				
9641_33			FARM4	0.011		
9677_31	L,D,BS					
9869_38	D					
9871_62			NOIR	0.023		

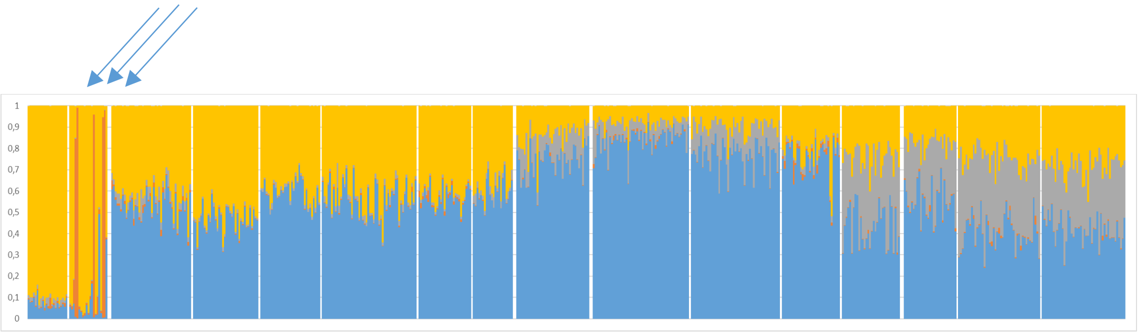
S3 Fst table for wild populations (full and neutral SNP datasets)

Fst values for the neutral dataset (under the diagonal) and for the full dataset (above the diagonal). Underlined values indicate significance <0.01, bold values indicate significance after sequential Bonferroni correction (p<0.05)

	NOIR	CAD	VAL	BAL	GEN	CTY	TORT	TRA	ADR	IGOU	MESO	NAY	KOR	BAS	THERM	ALEX
ALEX	**0,0099**	**0,0142**	0,0015	0,0020	0,0000	0,0017	0,0029	0,0031	0,0000	**0,0042**	0,0038	**0,0053**	0,0021	0,0008	0,0028	*
THERM	**0,0138**	**0,0185**	0,0028	**0,0046**	0,0024	**0,0036**	**0,0064**	0,0029	0,0009	**0,0046**	0,0046	**0,0068**	0,0000	0,0035	*	0,0027
BAS	**0,0128**	**0,0174**	0,0011	0,0034	0,0000	0,0034	0,0040	0,0061	0,0026	0,0040	0,0035	0,0038	0,0017	*	0,0031	0,0014
KOR	**0,0115**	**0,0154**	0,0017	0,0036	0,0001	0,0035	0,0050	0,0025	0,0000	0,0032	0,0037	**0,0064**	*	0,0011	0,0000	0,0015
NAY	**0,0131**	**0,0151**	0,0025	0,0045	0,0000	**0,0053**	0,0048	0,0034	0,0000	0,0012	0,0031	*	0,0046	0,0021	**0,0049**	0,0034
MESO	**0,0136**	**0,0173**	0,0011	**0,0060**	0,0009	**0,0046**	0,0041	0,0029	0,0000	0,0011	*	0,0028	0,0026	0,0023	0,0034	0,0027
IGOU	**0,0151**	**0,0174**	0,0028	**0,0041**	0,0000	0,0030	0,0026	0,0028	0,0000	*	0,0009	0,0012	0,0018	0,0032	0,0026	0,0023
ADR	**0,0085**	**0,0136**	0,0000	0,0000	0,0000	0,0003	0,0011	0,0000	*	0,0000	0,0000	0,0000	0,0000	0,0009	0,0000	0,0000
TRA	0,0062	0,0099	0,0006	0,0003	0,0000	0,0000	0,0006	*	0,0000	0,0013	0,0023	0,0028	0,0014	0,0050	0,0018	0,0026
TORT	**0,0076**	**0,0113**	0,0000	0,0000	0,0000	0,0000	*	0,0006	0,0000	0,0000	0,0021	0,0033	0,0020	0,0027	0,0040	0,0026
CTY	**0,0054**	**0,0095**	0,0000	0,0000	0,0000	*	0,0000	0,0000	0,0000	0,0000	0,0024	0,0031	0,0010	0,0026	0,0015	0,0017
GEN	**0,0068**	**0,0126**	0,0000	0,0000	*	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0000	0,0000	0,0000	0,0009	0,0000
BAL	0,0036	**0,0097**	0,0000	*	0,0000	0,0000	0,0000	0,0002	0,0000	0,0014	0,0036	0,0025	0,0013	0,0028	0,0030	0,0014
VAL	0,0053	**0,0102**	*	0,0000	0,0000	0,0000	0,0000	0,0011	0,0000	0,0018	0,0000	0,0018	0,0004	0,0012	0,0012	0,0016
CAD	0,0055	*	**0,0095**	0,0079	**0,0108**	**0,0081**	**0,0096**	0,0094	**0,0096**	**0,0133**	**0,0132**	**0,0113**	**0,0117**	**0,0144**	**0,0135**	**0,0115**
NOIR	*	0,0066	0,0026	0,0006	0,0030	0,0031	0,0049	0,0035	0,0013	**0,0071**	**0,0063**	0,0062	0,0045	0,0062	0,0054	0,0050

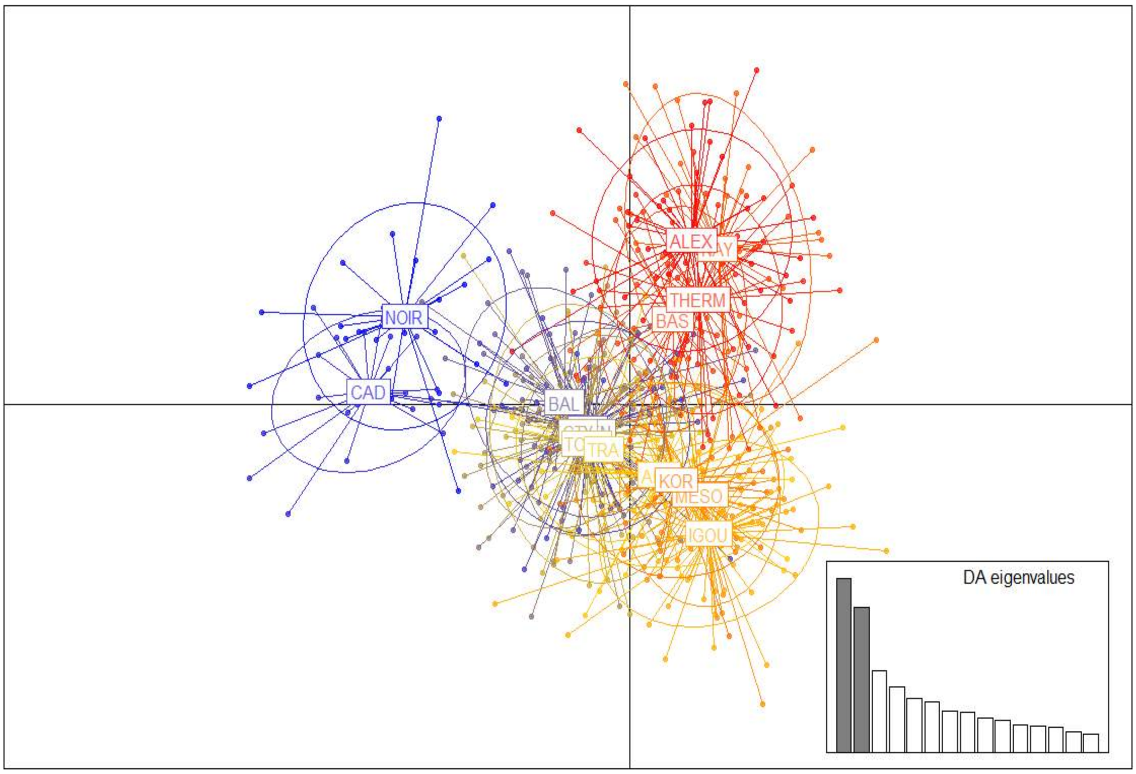
S4 Structure plot with CAD “outlier” samples

In first structure analysis, five CAD samples were separately clustered from all the other individuals. The issue was related to the fact that these samples were closely related and therefore Structure's algorithm grouped them in a separated cluster.



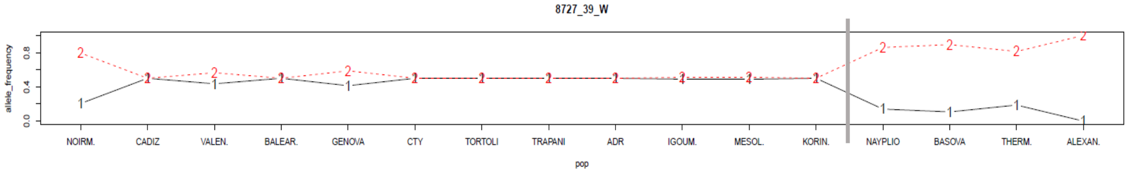
S5 DAPC scatter plot for wild populations

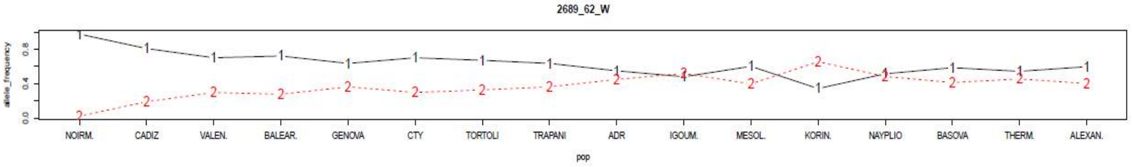
DAPC analysis highlighting the differentiation between Atlantic and Mediterranean on the first axes and within Mediterranean between W Med, Ionian and Aegean on the second axes

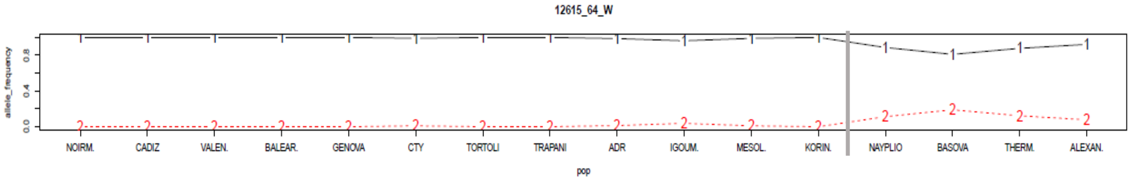


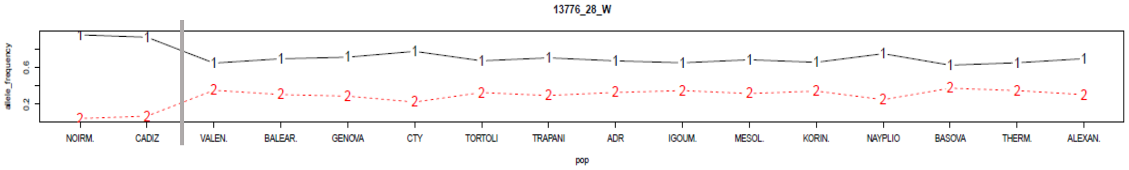
S6 Allele frequencies for single loci in the wild populations

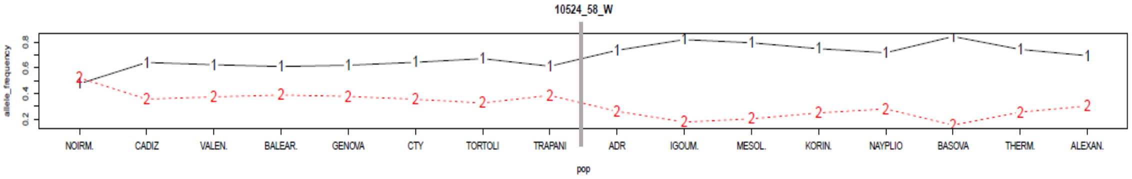
Examples of allele frequencies at eight potential outlier loci. Vertical lines indicate the “transition area” where allele frequency present stronger shifts. E.g. locus 2689 presents a gradient instead of a clear shift moving from one basin to another, with first allele being more frequent in the Atlantic and western Mediterranean and equally represented as second allele Aegean basin. Locus 13129_86 shows an even more drastic pattern, with most frequent allele being allele 1 in populations from ATL and WMED, and second allele being more frequent in almost all populations of ION and AEG basins.

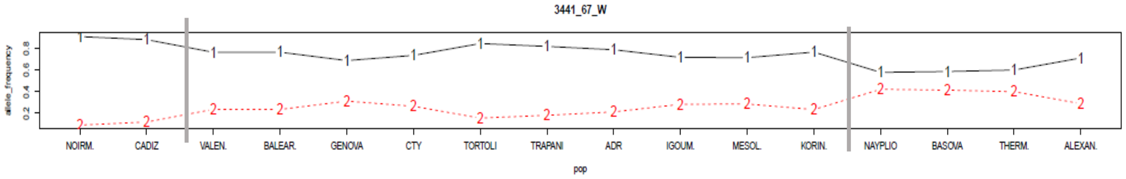

8727_39_W

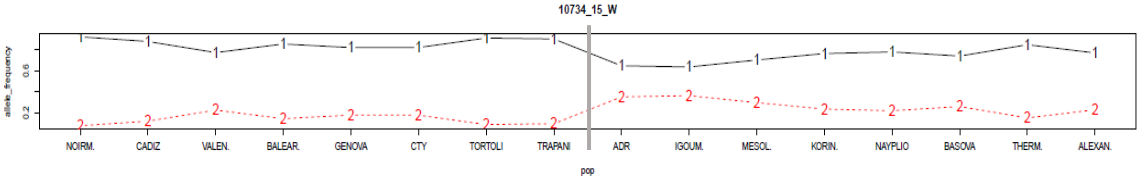

2689_62_W

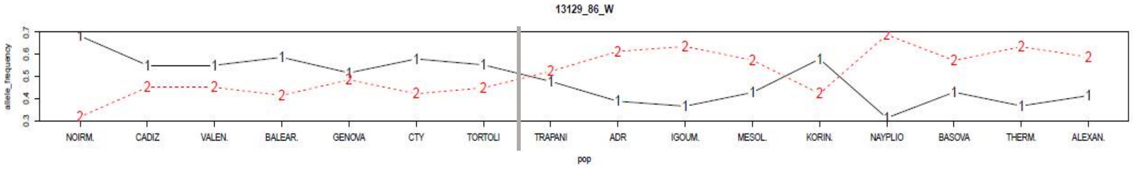

12615_64_W


13776_28_W


10524_58_W

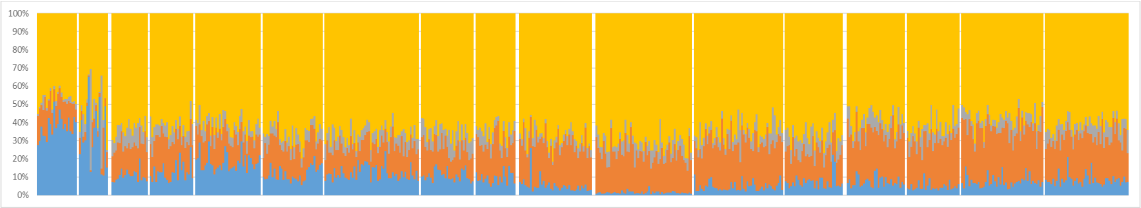

3441_67_W



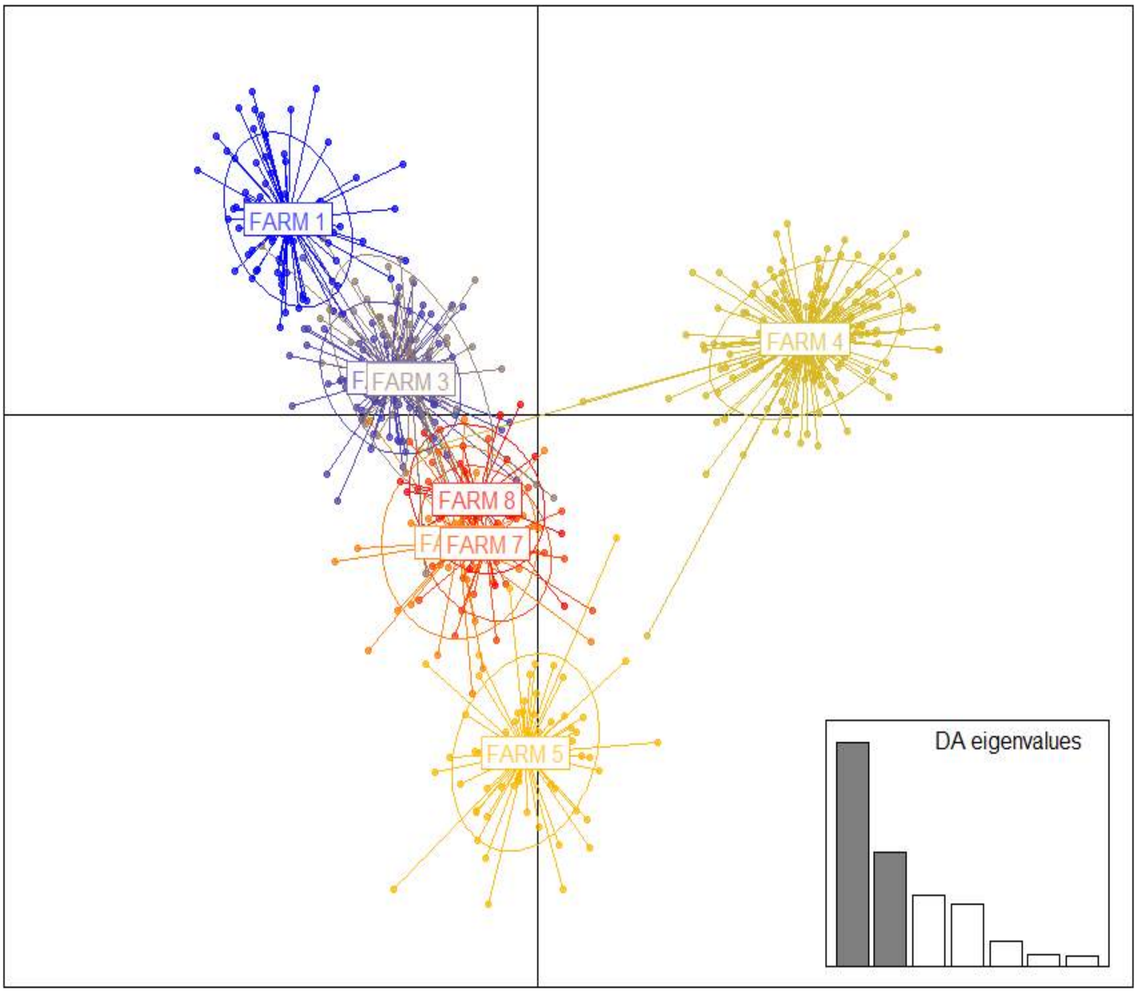


S7 Structure plot at putatively neutral loci for wild populations

Structure analysis at 1180 putatively neutral loci for all the wild populations, k=4. A fifth cluster that divided five CAD samples was removed from the analysis.

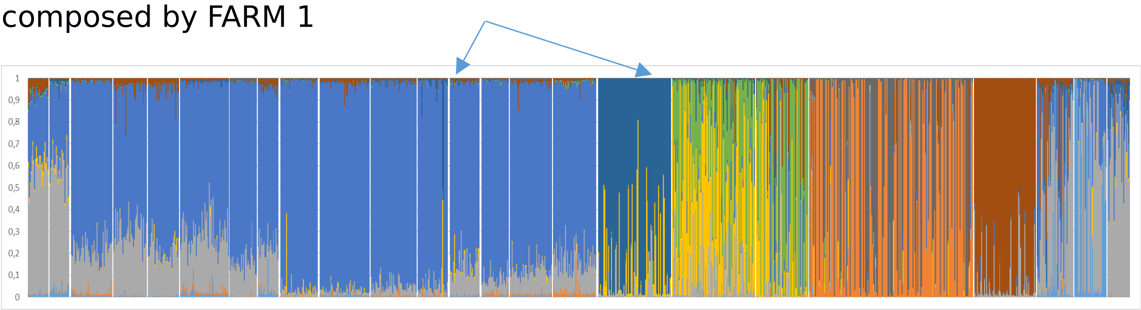


S8 DAPC farmed



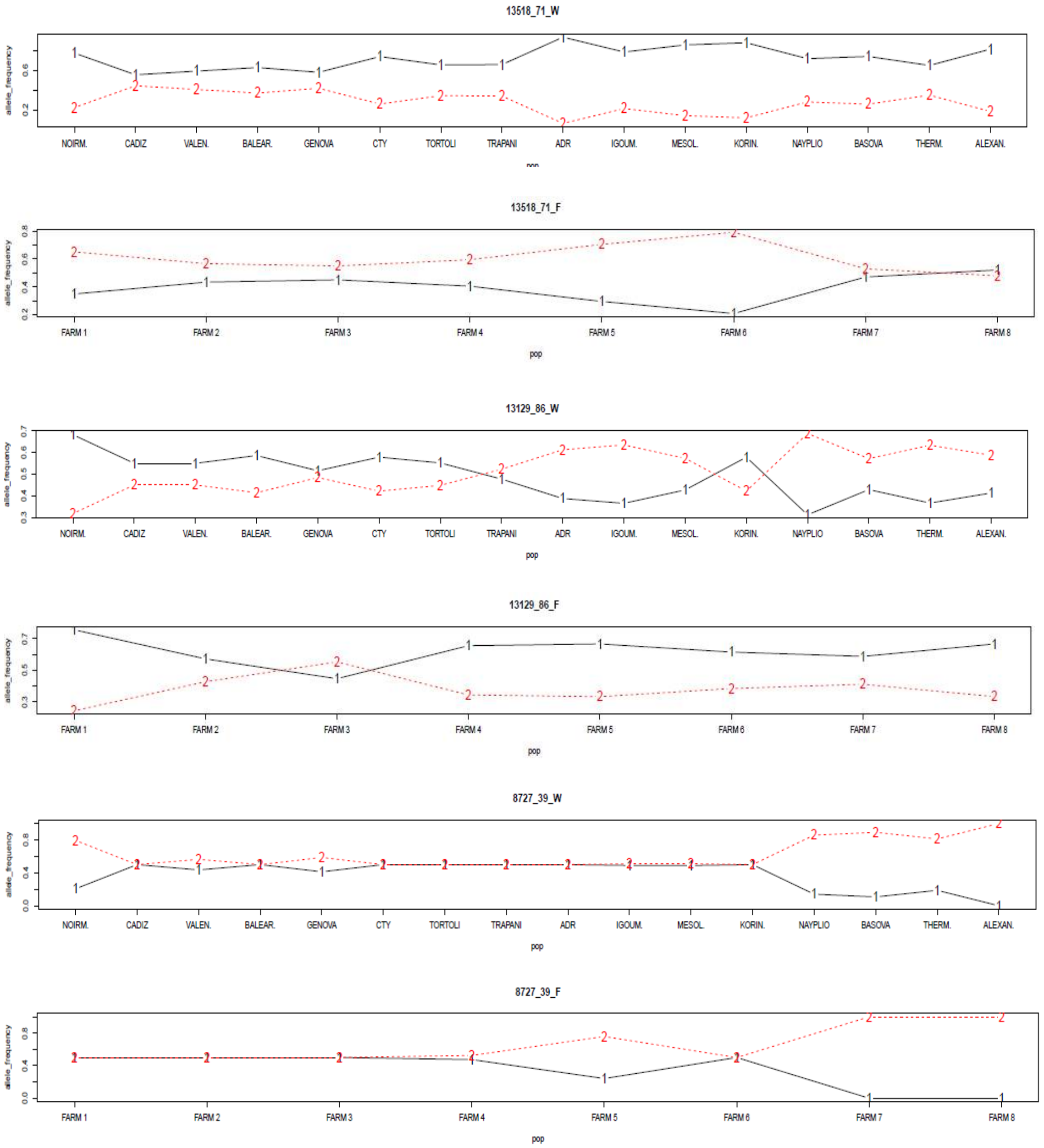
S9 Structure Wild and farmed

Structure plot (k=9) for wild and farmed samples analyzed together. Arrow indicates two “outlier” samples from KOR that were assigned to the cluster composed by FARM 1



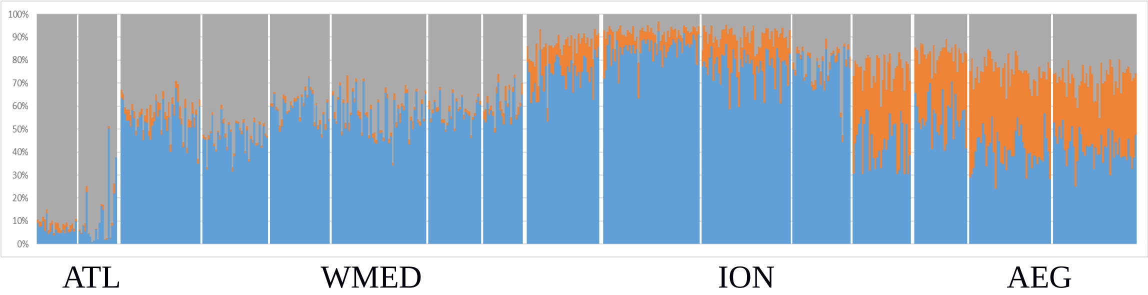
S10 OL allele frequency comparison between wild and farmed populations

Examples of comparisons of allele frequencies at outlier loci between farm groups and wild populations. According to the source of escapees/release and the amount of released individuals, allele frequencies at natural populations might drift and overall fitness might be compromised



S11 Structure plot for wild populations analysis

Structure plots for wild populations (ATL-WMED-IONIAN-AEGEAN) with k=4, after removing “outlier samples” cluster, that separated five individuals from CAD population, characterized by high pairwise relatedness.



During the period spent in Spain, working in the group led by prof. Paulino Martínez, I collaborated with the research group of Claudio Oliveira (University of San Paulo, Brazil) at the identification and validation on SNPs in *Thunnus albacares* and *Scomberomorus brasiliensis*. The approach used was a combination of ddRAD and 454 pyrosequencing, which revealed useful for marker development in species without genomic resources.

SNPs identification and validation in *Thunnus albacares* and *Scomberomorus brasiliensis* by double digest RADseq using a 454 pyrosequencing platform

Siccha-Ramirez R[1]*, Paes V[1], Pardo BG[2], Fernandez C [2], Maroso F [2,3], Martinez P[2] and Oliveira C[1]

[1] Departamento de Morfologia, Instituto de Biociências, UNESP, Botucatu, São Paulo, Brazil

[2] Departamento de Xenética, Facultade de Veterinaria, Universidade de Santiago de Compostela, Campus de Lugo, Lugo, 27002, Spain

[3] Dipartimento di Biomedicina Comparata e Alimentazione (BCA), Università degli Studi di Padova, Italy

Abstract

The world of genomics is advancing rapidly and new methodologies are being developed which increase data and decrease costs and working time. Here, a combination of ddRAD method with 454 pyrosequencing was developed in order to identify and validate single nucleotide polymorphisms (SNPs) in *Thunnus albacares* and *Scomberomorus brasiliensis*. For SNPs identification DNA from 11 and 21 individuals of *T. albacares* and *S. brasiliensis,* respectively, was individually digested using two restriction enzymes (*Sbf*I and *Sph*I) and fragments from 300 to 600 bp were selected. Barcode sequences (5-7 bp) for combinatorial barcoding were included in the adapters for each restriction site (P1 and P2) ligated to digested DNA. Samples were pooled and size-selected from agarose gels, subsequently amplified by PCR, and finally sequenced on 454 GS Junior (Roche Diagnostic). A total of 180,779 reads were produced with an average length of ~ 287 bp and 26x coverage. A set of SNPs was *in silico* selected for *T. albacares* (60) and *S. brasiliensis* (79) and two SNP multiplex reactions were

designed for each species and tested on a panel of 42 and 23 individuals, respectively, in the MassARRAY platform (Sequenom, San Diego, CA, USA). Finally, 36 and 42 SNPs were polymorphic in *T. albacares* and *S. brasiliensis*, respectively. Our results demonstrate the possibility of combining ddRAD with the longer reads from 454 pyrosequencing to obtain genomic information for marker development in species without genomic resources. This methodology demonstrated to be useful for identification and validation of SNPs in the species studied and could be easily applied for many other non-model organisms.

Introduction

Next generation sequencing (NGS) has revolutionized the field of genetics [139,140] allowing investigation on species considered non-models, with a genomic unprecedented coverage. One application is the search, validation and large-scale genotyping of single-nucleotide polymorphisms (SNPs), using different methodologies (Snapshot, Sequenom, Veracode, Goldengate, etc.) [141]. SNPs are stable and usually bi-allelic polymorphisms [142], found in coding and non-coding regions of the genome [143], uniformly distributed at high density [144], thus, being more common than other markers in the genome [145]. In fish, a SNP is found every ~100 bp [145,146]. These properties make these markers ideal for comparative studies, evolutionary genomics [144], fine mapping of genes associated with important features, conservation genetics, enabling to estimate the evolutionary history of populations and genetic differentiation between populations [147,148], and hybridization and impact of biological invasions [149].

The Reduced Representation Library technique (RRL) provides a high yield for the efficient discovery of SNPs [150] having a great advantage because it reduces by large magnitudes the analyses of complex genomes [144]. This method analyzes an identical small portion of the genome in all individuals or populations analyzed, represented by a particular set of fragments randomly distributed without the need to sequence their entire genome [151]. A derivation of this technique, the RADseq (Restriction-site Associated DNA sequencing), has gained popularity in non-model organisms because it allows obtaining useful genomic information at low cost [47,152,153]. To increase the breadth of RADseq applications, the double-digest RADseq (ddRADseq) method was developed by eliminating random shearing and explicitly using size selection to recover a tunable number of regions according to the goals of the study [42]. The ddRADseq tags not only possess the advantages of RAD tags, such as allowing high-throughput, multiplexed sequencing and being amenable to genotyping, but they also provide improved efficiency and robustness compared to RADseq [154].

The yellowfin tuna (*Thunnus albacares*) is a migratory species found in tropical and subtropical waters all over the world with high commercial interest. This species is currently overfished and appears on the Red List as near threatened species (IUCN, 2014). A variety of studies have been done trying to assess the population structure of *T. albacares* using different approaches, including genetic approaches [155–161], but its genetic structure is yet not clear. On the other hand, the serra Spanish mackerel (*Scomberomorus brasiliensis*) is a neritic species [162] distributed from Belize to south Brazil [163] of high commercial interest in Trinidad and Tobago, Venezuela [164] and in Brazil, especially in the state of Maranhão [165], however, no research has been conducted in this species to evaluate genetic resources and structure essential for their management. Considering the importance of these species in an ecological context and its global importance as an economic resource, in addition to the urgent need for conservation measures, this work aimed at the identification, validation and genetic diversity evaluation of SNPs, testing a new combination and adaptation of techniques as ddRADseq with 454 pyrosequencing in *T. albacares* and *S. brasiliensis*.

Material and Methods

Biological material and DNA extraction

Twenty-one *S. brasiliensis* and 11 *T. albacares* individuals were used for library construction, sequencing and SNP discovery and validation. Additionally, 23 and 42 samples of *S. brasiliensis* and *T. albacares*, respectively, coming from a single wild population each were used to evaluate genetic parameters of validated SNPs. Samples of *S. brasiliensis* and *T. albacares* were collected in the North and South Brazil from Rio Grande do Norte to Santa Catarina. Representative specimens and all tissues were deposited in the fish collection of Laboratório de Biologia and

Genética de Peixes of the Universidade Estadual Paulista UNESP (Botucatu, São Paulo, Brazil). Genomic DNA samples were obtained from ethanol-preserved tissues, lysed in 300 µl of SSTNE extraction buffer [166] plus SDS (0.1%) and 5 µl of proteinase K (20 mg/ml) for 3 h at 55°C. After 20 min at 70°C, RNAse treatment was performed adding 7.5 µl of RNAse (10 mg/ml) and incubated 1 h at 37°C. Total DNA was purified after protein precipitation (5M NaCl) with freezing absolute ethanol (1 ml). DNA quality (high molecular weight > 20 kb) was first evaluated on agarose gels and the DNA quantity was measured using the NanoDrop ® ND-1000 spectrophotometer (NanoDrop® Technologies Inc) and PicoGreen kit (Molecular Probes) according to the kit instructions. Finally, DNA concentration was accurately measured on a fluorometry based device (Life Technology Qubit fluorometer).

Library construction and sequencing

A reduced portion of the genome of the two species was sequenced using a modified ddRAD protocol [42]. DNA from the 32 samples (21 *S. brasiliensis* and 11 *T. albacares*) was analyzed all together in a single sequencing run. Briefly, for each sample, the same amount of DNA (78 ng) was individually digested using *Sbf*I and *Sph*I restriction enzymes (RE). Adapters for each RE site were subsequently ligated to digested DNA fragments including: i) complementary cohesive ends for REs; ii) barcodes to identify individuals (Supplementary Table S1); and iii) a couple of primers for an intermediate PCR amplification. Labeled samples were then pooled and run in agarose gels 1.1% for fragment selection (300 to 600 bp), followed by extraction using Qiagen MinElute Gel Extraction kit. After selection, the target DNA fragments were amplified by polymerase chain reaction (PCR): initial denaturation and enzyme activation at 98°C x 30s; 14 cycles at 98°C x 10s (denaturation step) 65°C x 30s (annealing) and 72°C x 30s (extension); final extension at 72°C x 5min. The PCR product was purified using the Qiagen MinElute PCR clean up kit followed by a magnetic bead clean-up / size selection using an equal volume of Beckman Coulter AMPure XP beads. This protocol ensured that only those fragments including SbfI and SphI target sites were amplified and further sequenced.

454 sequencing, assembly contigs and SNP identification and

selection

The final library was sequenced in a single shotgun run on a 454 GS Junior sequencer (Roche Diagnostic) available at the Sequencing and Functional Genomics Platform of the University of Santiago de Compostela (USC, Campus Lugo, Spain), starting in section 3.2 Fragment End Repair of the Rapid Library Preparation Method Manual. Sequencing reads were filtered using default parameters, classified per individual according to barcodes and assembled with Newbler software (specifically designed for 454 GS-series data). Alignments were then parsed with Tablet [167] in order to identify SNPs in the assembled sequences. Only contigs containing a sequencing depth >4 were retained for subsequent analysis to reduce SNPs attributable to sequencing errors.

SNPs were selected according to the presence of enough flanking regions for primer design (±100bp) and the absence of other DNA polymorphism (SNPs and indels) in those regions that could interfere with primer annealing and genotyping. Additionally, only those SNPs with at least three sequences of the least common allele were selected.

SNP genotyping

Identified SNPs were validated and genotyped with the MassARRAY platform (Sequenom, San Diego, CA, USA) at the USC node of the Spanish National Centre of Genotyping (CeGen ISCIII) following the protocols and recommendations provided by the manufacturer. Briefly, the technique consists of an initial locus-specific PCR, followed by single-base extension using mass-modified dideoxynucleotide terminators of an oligonucleotide primer that anneals immediately upstream of the polymorphic site (SNP) of interest [168,169]. The distinct mass of the extended primer identifies the SNP allele. MALDI-TOF mass spectrometry analysis in an Autoflex spectrometer was used for allele scoring. Two SNPs multiplexes were designed *in silico* using Assay Design 3.1 program (Sequenom, San Diego, CA), which maximizes the number of SNP per multiplex and minimizes the number of multiplex, and tested on a panel of 42 and 23 individuals from a single wild population in *T. albacares* and *S. brasiliensis,* respectively. Feasible SNPs (markers with proper and reliable genotypes) and “failed assays” (majority of genotypes not scored or difficult to cluster according to genotype) were classified by manual inspection.

Gene diversity and annotation

Genepop on the web tools [88,89] were used to estimate genetic diversity parameters (He: expected heterozygosity; Ho: observed heterozygosity; MAF: minimum allele frequency), to test for deviations from H-W equilibrium and their sense (Fis), and to check for linkage disequilibrium. Complete enumeration approach was used to calculate p-values for H-W equilibrium test [170] and Weir and Cockerham test [171] was used for F*is*. Linkage (genotypic) disequilibrium was analyzed for each pair of loci using the log likelihood ratio statistic (G-test). The p-value threshold was set after Bonferroni correction when multiple tests were performed.

BLASTn was used to look for hits in whole genome shotgun contigs (wgs) databases of six different fish species, selected among those with best characterized genomes: Pacific bluefin tuna (*Thunnus orientalis*), fugu (*Takifugu rubripes*), stickleback (*Gasterosteus aculeatus*), zebrafish (*Danio rerio*), medaka (*Oryzias latipes*) and tetraodon (*Tetraodon nigroviridis*). Threshold significance was set at 10e-5. In addition to BLASTn analysis, SNP containing sequences in both species were blasted against non redundant (nr) protein database from NCBI, using BLASTx. As the ddRAD protocol is a random representative genomic reduction approach, most tags are expected not to match protein coding regions.

Results and discussion

SNP discovery and genotyping

A single Roche 454 GS-Junior run was performed and a total of 180,779 reads passed the quality filter (73.3% of a total of 246,663 reads obtained in the GS-Junior run). Average read length was 287.0 bp and the average read quality Phred was 30.2. High quality reads were separately assembled per species and a total of 1,715 contigs were detected for *T. albacares*, with an average length of 374.9 bp and average coverage depth of 25.5 reads. For *S. brasiliensis* the number of identified contigs was 2,274, with an average length of 374.6 bp and an average coverage of 26.2 reads per contig (Table 1).

Table 1. Characteristics of 454 GS Junior run and genotyping SNPs

Sequencing results	**Roche 454 GS-Junior stats**
Number of HQ reads	180,779
Total megabases (Mb)	51,879,629
Average length of reads	287.0
N° individuals sequenced	32

Assembly results	*Thunnus albacares*	*Scomberomorus brasiliensis*
N° individuals sequenced	11	21
Number of aligned reads	42,875	59,124
Total n° of contigs	1,715	2,274
Average contig length	375.3	375.3
Average coverage per contig	25.3	26.2

Genotyping	*Thunnus albacares*	*Scomberomorus brasiliensis*
N° individuals	42	23
N° markers	50	55
N° variable markers	36	42

A total of 60 SNPs for *T. albacares* and 79 for *S. brasiliensis* were initially selected according to the criteria outlined above for subsequent validation steps. Contigs containing selected SNPs averaged 290.1 bp for *T. albacares* and 298.7 bp for *S. brasiliensis*. Following manual inspection 50 feasible SNPs (83.3% of the 60 SNPs selected) for *T. albacares* and 55 (69.6% of the 79 SNPs selected) for *S. brasiliensis* were finally chosen for validation. Markers were combined in two multiplex reactions for each species including 30 and 20 SNPs each for *T. albacares* and 30 and 25 for *S. Brasiliensis,* respectively. Primer sequences, SNP position, expected variants and annotation for the 50 and 55 tested SNPs in *T. albacares* and *S. brasiliensis*, respectively, are shown in Supplementary Table S1.

The consensus sequences of these 105 SNP-containing contigs were compared using NCBI BLASTn with public databases, showing, as expected, a high similarity between *T. albacares* and *S. brasiliensis* sequences with the available *T. orientalis* genome. Best significant BLASTn (Supplementary Table S2) matches were always against this species. Average E-values showed a higher similarity between *T. albacares* and *T. orientalis* (8 e-63) as expected, being much lower in the case of *S. brasiliensis* (3.71 e-28 see Supplementary Table S2 for details). Out of 105 sequences blasted against NCBI's nr protein database 12 for *T. albacares* and 14 for *S. brasiliensis* showed a significant hit (E-value < 10^{-5}; Supplementary Table S2) but without a consistent annotation (all predicted, hypothetical or unnamed proteins).

Using 42 individuals of *T. albacares* and 23 of *S. brasiliensis*, 36 and 42 markers were variable of the 50 and 55 feasible SNPs chosen, respectively (72.0% and 76.4% of total markers analyzed, respectively). Sequences with these SNPs and the used primers were deposited in GenBank. The ddRAD technique combined with 454 pyrosequencing was successful for SNP identification and primer design, mainly due to the size of reads (~ 300 bp).

Gene diversity and annotation

Number of transitions and transversions were calculated for SNPs in both species. Out of 36 variable markers in *T. albacares* dataset, 30 (83.3%) were transitions and 6 (16.7%) transversions. In *S. brasiliensis*, out of 42 variable markers, 16 (38.1%) were transitions and the remaining 26 (61.9%) transversions (Fig.1).

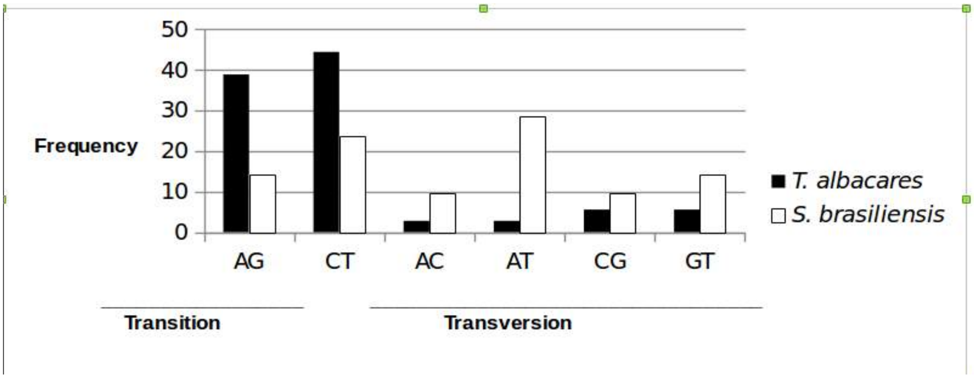


Figure. 1 Frequency of transitions and transversions in *T. albacares* and *S. brasiliensis.*

Accordingly, transition/transversion (ts/tv) ratio largely differed between both species (5.000 vs 0.615 for *T. albacares* and *S. brasiliensis*, respectively). These values are somewhat different to those found in other fish species. Ts/tv ratios ranging between 1.354 (Vera et al. 2013) and 1.885

(Vera et al. 2011) have been reported in turbot (*Scophthalmus maximus*); 1.310 in common carp (*Cyprinus carpio*) (Zhu et al. 2012) and 1.375 in gilthead sea bream (*Sparus aurata*) (Cenadelli et al. 2007). A higher number of transitions *vs* transversions is a common observation, despite higher number of transversion events can happen. We cannot discard some bias produced during the process of SNP selection. So, in *T. albacares* the mean MAF for transition SNPs (0.155) was lower than mean MAF for transversion SNPs (0.231), although differences were not statistically significant (Mann-Whitney test; P=0.170). On the contrary, MAF for transition SNPs and transversion SNPs in *S. brasiliensis* were very similar (0.217 *vs* 0.244). Anyhow, the extreme values observed in *T. albacares* and *S. brasiliensis* are uncommon and further data will be required to confirm this trend. The minimum allele frequency (MAF) for *T. albacares* SNPs ranged from 0.011 (Talb0149, Talb0153, Talb0417, Talb2746) to 0.452 (Talb0891), with an average value of 0.168. For *S. brasiliensis* MAF ranged from 0.021 (Sbra0059, Sbra0256, Sbra1146) to 0.5 (Sbra1095), averaging 0.234. He ranged from 0.024 to 0.495 (average 0.242) for *T. albacares* and from 0.043 to 0.500 (average 0.312) for *S. brasiliensis*. Hardy-Weinberg equilibrium tests detected only two significant deviations due to heterozygote deficit after Bonferroni correction ($p < 0.0007$) both in *S. brasiliensis* SNPs (loci Sbra0660 and Sbra1038). The high Fis value observed (0.894 and 0.803, respectively) suggests the presence of null alleles. The average heterozygosity in *T. albacares* (0.242) and *S. brasiliensis* (0.312) are higher than those found by Ward et al. (1994) after the analysis of 57 marine species, reporting an average of 0.059, but closed to that found by Vera et al. (2013) analyzing 130 SNPs in *Scophthalmus maximus* (average value of 0.344), and by Albaina et al. (2012) analyzing 41 and 15 SNPs in *Thunnus alalunga* and *T. thynnus* reporting mean values of 0.278 and 0.272, respectively.

No pair of loci showed significant deviation from linkage equilibrium after Bonferroni correction in both species (630 and 859 tests, respectively), although four possible linked pairs of loci ($p < 0.01$) were identified in *T. albacares* (Talb1155 and Talb2058; Talb0337 and Talb0417; Talb1258 and Talb1549; Talb488 and Talb568), and five in *S. brasiliensis* (Sbra1397 and Sbra1706; Sbra0484 and Sbra2880; Sbra0447 and Sbra2061; Sbra787 and

Sbra1706; Sbra0455 and Sbra0933).

Conclusions

Here we described a combination of technology of laboratory using ddRAD approach and 454 pyrosequencing to identify *in silico* SNPs markers which were validated using high-throughput genotyping Sequenom MassARRAY platform. This method enabled highly repeatable and tunable recovery of hundreds to thousands of sampled regions from *T. albacares* and *S. brasiliensis* genomes. Our results demonstrated the utility of this new approach to obtain a rapid and cost-effective discovery of SNPs useful for population genetics in *T. albacares* and *S. brasiliensis,* easily used for many other non-model organisms.

Acknowledgements

We are grateful to the following people, who helped during the sampling process: Leonardo Machado (UNESP), Jonathan Ready (UFPA), Ricardo Britzke (UNESP) and all fishery owners who kindly help us to get the samples in their boats. We thank Susana Sánchez, Lucía Insua and María López for technical support. We also thank CeGen for the genotyping of our samples. Finally, a special thanks to funding source FAPESP (grant 2011/00881-1 and grant 2014/2013/07560-1).

SUPPLEMENTARY MATERIAL

Supplementary Table 1. SNPs Information

Thunnus albacares

SNPs	Annotation	Variants	MAF	He	P (HW)	Fis (W&C)
Talb0043	PREDICTED: protein HEG homolog 1 [*Larimichthys crocea*]	C/T	C=0,059	0.112	0.118	0.372
Talb0074						
Talb0149		A/G	A=0,011	0.024		
Talb0153		A/T	T=0,011	0.024		
Talb0213						
Talb0259		A/C	C=0,428	0.490	0.527	0.137
Talb0337		C/T	C=0,369	0.466	0.335	-0.164
Talb0401		A/G	A=0,130	0.228	0.528	0.071
Talb0417	hypothetical protein *[Rhinecanthus aculeatus]*	A/G	A=0,059	0.112	1.000	-0.051
Talb0435		C/T	T=0,011	0.024		
Talb0463	unnamed protein product [*Tetraodon nigroviridis*]					
Talb0488		A/G	A=0,333	0.444	1.000	0.048
Talb0507		C/T	C=0,047	0.091	1.000	-0.038
Talb0516						
Talb0546		C/T	C=0,047	0.091	1.000	-0.038
Talb0568		C/T	T=0,095	0.172	0.307	0.183
Talb0685		C/G	G=0,071	0.133	1.000	-0.065
Talb0717						
Talb0807	PREDICTED: anoctamin-7-like isoform X1 [*Pundamilia nyererei*]	A/G	G=0,083	0.153	1.000	-0.079
Talb0822	PREDICTED: opioid-binding protein/cell adhesion molecule-like [*Takifugu rubripes*]					
Talb0826	PREDICTED: sodium/glucose cotransporter 4-like [*Larimichthys crocea*]					
Talb0827		A/G	G=0,035	0.069	1.000	-0.025
Talb0851						
Talb0889						
Talb0891		A/G	G=0,452	0.495	0.764	0.051
Talb0895						
Talb0930						
Talb0952		C/T	T=0,341	0.450	0.035	0.360
Talb1040	PREDICTED: ankyrin repeat domain-Talbtaining protein 6 [*Larimichthys crocea*]	C/T	T=0,142	0.245	0.573	-0.155
Talb1083	PREDICTED: uncharacterized protein C15orf52 homolog isoform X2 [*Maylandia zebra*]	C/T	T=0,154	0.262	1.000	0.011
Talb1093		A/G	A=0,083	0.153	1.000	-0.079
Talb1154		C/T	T=0,35	0.455	0.731	-0.086
Talb1155		A/G	A=0,071	0.133	1.000	-0.065

Talb1211						
Talb1258		C/T	T=0,440	0.493	0.753	0.094
Talb1308		A/G	A=0,273	0.398	0.696	-0.126
Talb1383		A/G	G=0,154	0.262	0.232	0.193
Talb1549	PREDICTED: sentrin-specific protease 7 [*Larimichthys crocea*]	C/T	C=0,142	0.245	1.000	0.040
Talb1580		A/G	G=0,166	0.278	0.299	0.155
Talb1582		A/G	A=0,238	0.363	1.000	-0.038
Talb1911	PREDICTED: trophoblast glycoprotein-like [*Stegastes partitus]*	C/T	T=0,107	0.191	1.000	-0.108
Talb1933						
Talb1952		C/G	C=0,273	0.398	1.000	-0.006
Talb2008		A/G	G=0,142	0.245	0.573	-0.155
Talb2058	PREDICTED: uncharacterized protein LOC101166008 [*Oryzias latipes]*	G/T	G=0,166	0.278	1.000	-0.017
Talb2064						
Talb2335		G/T	G=0,439	0.493	0.530	0.121
Talb2481		C/T	T=0,023	0.046	1.000	-0.012
Talb2545	cytochrome P450 3A69 [*Micropterus salmoides*]	C/T	T=0,095	0.172	1.000	-0.093
Talb2746		C/T	C=0,011	0.024		

Scomberomorus brasiliensis

SNPs	**Annotation**	**Variants**	**MAF**	**He**	**P (HW)**	**Fis (W&C)**
Sbra0059		C/T	C=0,021	0.043		
Sbra0087	PREDICTED: anoctamin-7-like isoform X2 [Pundamilia nyererei]	A/G	G=0,478	0.499	1.000	-0.023
Sbra0099		G/T	T=0,347	0.454	0.662	-0.128
Sbra0112						
Sbra0123						
Sbra0126		A/C	C=0,065	0.122	1.000	-0.048
Sbra0130		A/C	A=0,108	0.194	1.000	-0.100
Sbra0142		G/T	G=0,086	0.159	1.000	-0.073
Sbra0145		C/G	C=0,282	0.405	0.626	-0.158
Sbra0197	unnamed protein product [Tetraodon nigroviridis]					
Sbra0200	Uncharacterized protein [Dicentrarchus labrax]					
Sbra0256	PREDICTED: zinc finger and BTB domain-Sbrataining protein 8A [Stegastes partitus]	C/T	T=0,021	0.043		
Sbra0265						
Sbra0286		A/T	T=0,222	0.346	1.000	0.064
Sbra0292						
Sbra0325		C/T	C=0,181	0.298	1.000	-0.200
Sbra0374	PREDICTED: activated CDC42 kinase 1-like isoform X2 [Haplochromis burtoni]					
Sbra0438	PREDICTED: zinc finger protein 513-like isoform X5 [Maylandia zebra]	A/G	G=0,239	0.364	0.279	-0.294
Sbra0447		C/T	C=0,454	0.496	1.000	-0.077
Sbra0455		C/T	T=0,434	0.491	0.674	0.137
Sbra0458		G/T	T=0,021	0.043		
Sbra0463						

Sbra0484		A/T	A=0,239	0.364	1.000	-0.053
Sbra0502		A/C	A=0,113	0.201	0.224	0.344
Sbra0660		G/T	G=0,309	0.427	0.000	0.894
Sbra0693		A/T	T=0,173	0.287	1.000	-0.189
Sbra0700		A/C	C=0,043	0.083	1.000	-0.023
Sbra0713	PREDICTED: far upstream element-binding protein 1-like isoform X2 [Xiphophorus maculatus]	A/T	T=0,136	0.236	0.324	0.250
Sbra0787		G/T	T=0,159	0.268	0.057	0.508
Sbra0801		A/T	A=0,087	0.159	0.002	1.000
Sbra0932	PREDICTED: ryanodine receptor 1-like [Poecilia reticulata]	C/T	C=0,413	0.485	0.670	-0.144
Sbra0933		A/T	T=0,413	0.485	0.010	0.567
Sbra0981		A/G	G=0,043	0.083	1.000	-0.023
Sbra1035		A/T	T=0,431	0.491	0.419	0.189
Sbra1038		A/T	A=0,304	0.423	0.000	0.803
Sbra1095		C/G	C=0,5	0.500	0.402	0.239
Sbra1146		A/T	A=0,021	0.043		
Sbra1196	PREDICTED: NADPH oxidase organizer 1-like [Larimichthys crocea]					
Sbra1304						
Sbra1361		C/T	T=0,326	0.440	1.000	-0.066
Sbra1397		A/G	A=0,130	0.227	0.310	0.254
Sbra1532		A/T	A=0,260	0.386	1.000	-0.106
Sbra1614		G/T	G=0,434	0.491	1.000	-0.039
Sbra1626	PREDICTED: ubiquitin-associated protein 2-like isoform X1 [Stegastes partitus]					
Sbra1675	PREDICTED: DENN domain-Sbrataining protein 5A isoform X2 [Larimichthys crocea]	C/T	T=0,086	0.159	1.000	-0.073
Sbra1706	PREDICTED: myomegalin-like isoform X6 [Cynoglossus semilaevis]	C/G	G=0,434	0.491	0.002	0.659
Sbra1957	PREDICTED: phosphatidylinositol 4-phosphate 5-kinase type-1 gamma-like [Larimichthys crocea]	A/G	G=0,130	0.227	0.310	0.254
Sbra1985		C/G	G=0,434	0.491	0.205	0.313
Sbra2061		A/T	A=0,333	0.444	0.624	0.167
Sbra2083		C/T	T=0,043	0.083	1.000	-0.023
Sbra2272		A/T	T=0,391	0.476	0.034	0.470
Sbra2735						
Sbra2880		C/T	T=0,304	0.423	0.619	-0.211
Sbra2947	PREDICTED: paired box protein Pax-3-like isoform X2 [Notothenia coriiceps]					

Supplementary table 2. Thunnus albacares and Scomberomorus brasiliensis results for blastn. Five best hits showed

<table>
<thead>
<tr><th>SNPs</th><th>Subject ids</th><th>Species</th><th>% identity</th><th>Alignment length</th><th>E-value</th></tr>
</thead>
<tbody>
<tr><td colspan="6"><b><i>Thunnus albacares</i></b></td></tr>
<tr><td><b>Talb0043</b></td><td>gi|515322231|dbj|BADN01039616.1|</td><td><i>Thunnus orientalis</i></td><td>99.34</td><td>301</td><td>1E-149</td></tr>
<tr><td><b>Talb0074</b></td><td>gi|515229138|dbj|BADN01096052.1|</td><td><i>Thunnus orientalis</i></td><td>94.90</td><td>294</td><td>5E-129</td></tr>
<tr><td></td><td>gi|347787886|emb|CAAB02005086.1|</td><td><i>Takifugu rubripes</i></td><td>84.93</td><td>219</td><td>1E-61</td></tr>
<tr><td></td><td>gi|500898618|gb|AOOT01061080.1|</td><td><i>Takifugu flavidus</i></td><td>83.11</td><td>219</td><td>3E-56</td></tr>
<tr><td></td><td>gi|86296913|gb|AANH01005775.1|</td><td><i>Gasterosteus aculeatus</i></td><td>83.81</td><td>210</td><td>3E-56</td></tr>
<tr><td></td><td>gi|145778951|dbj|BAAF04060440.1|</td><td><i>Oryzias latipes</i></td><td>85.71</td><td>182</td><td>2E-52</td></tr>
<tr><td><b>Talb0149</b></td><td>gi|515372793|dbj|BADN01010141.1|</td><td><i>Thunnus orientalis</i></td><td>99.67</td><td>300</td><td>1E-150</td></tr>
<tr><td><b>Talb0153</b></td><td>gi|515285576|dbj|BADN01061533.1|</td><td><i>Thunnus orientalis</i></td><td>98.33</td><td>300</td><td>9E-145</td></tr>
<tr><td></td><td>gi|86297219|gb|AANH01005469.1|</td><td><i>Gasterosteus aculeatus</i></td><td>71.96</td><td>214</td><td>5E-21</td></tr>
<tr><td><b>Talb0213</b></td><td>gi|515262844|dbj|BADN01074885.1|</td><td><i>Thunnus orientalis</i></td><td>92.86</td><td>182</td><td>9E-69</td></tr>
<tr><td><b>Talb0259</b></td><td>gi|515272685|dbj|BADN01068374.1|</td><td><i>Thunnus orientalis</i></td><td>98.13</td><td>267</td><td>7E-127</td></tr>
<tr><td><b>Talb0337</b></td><td>gi|515265284|dbj|BADN01073347.1|</td><td><i>Thunnus orientalis</i></td><td>98.25</td><td>286</td><td>4E-136</td></tr>
<tr><td><b>Talb0401</b></td><td>gi|515319814|dbj|BADN01041435.1|</td><td><i>Thunnus orientalis</i></td><td>98.87</td><td>266</td><td>4E-130</td></tr>
<tr><td></td><td>gi|86294982|gb|AANH01007706.1|</td><td><i>Gasterosteus aculeatus</i></td><td>84.85</td><td>99</td><td>1E-22</td></tr>
<tr><td><b>Talb0417</b></td><td>gi|515374391|dbj|BADN01009230.1|</td><td><i>Thunnus orientalis</i></td><td>99.67</td><td>300</td><td>1E-150</td></tr>
<tr><td></td><td>gi|86292181|gb|AANH01010507.1|</td><td><i>Gasterosteus aculeatus</i></td><td>91.00</td><td>300</td><td>2E-115</td></tr>
<tr><td></td><td>gi|144277214|dbj|BAAE01051408.1|</td><td><i>Oryzias latipes</i></td><td>89.67</td><td>300</td><td>5E-110</td></tr>
<tr><td></td><td>gi|145749074|dbj|BAAF04089395.1|</td><td><i>Oryzias latipes</i></td><td>89.00</td><td>300</td><td>3E-107</td></tr>
<tr><td></td><td>gi|347784048|emb|CAAB02008924.1|</td><td><i>Takifugu rubripes</i></td><td>88.33</td><td>300</td><td>1E-104</td></tr>
<tr><td><b>Talb0435</b></td><td>gi|515336119|dbj|BADN01031196.1|</td><td><i>Thunnus orientalis</i></td><td>97.93</td><td>242</td><td>8E-114</td></tr>
<tr><td></td><td>gi|515322449|dbj|BADN01039455.1|</td><td><i>Thunnus orientalis</i></td><td>84.52</td><td>155</td><td>3E-37</td></tr>
<tr><td></td><td>gi|515374408|dbj|BADN01009219.1|</td><td><i>Thunnus orientalis</i></td><td>80.43</td><td>184</td><td>3E-37</td></tr>
<tr><td></td><td>gi|515301475|dbj|BADN01052850.1|</td><td><i>Thunnus orientalis</i></td><td>78.11</td><td>201</td><td>1E-36</td></tr>
<tr><td></td><td>gi|515302323|dbj|BADN01052309.1|</td><td><i>Thunnus orientalis</i></td><td>79.57</td><td>186</td><td>4E-35</td></tr>
<tr><td><b>Talb0463</b></td><td>gi|515295033|dbj|BADN01056989.1|</td><td><i>Thunnus orientalis</i></td><td>95.32</td><td>278</td><td>2E-122</td></tr>
<tr><td></td><td>gi|47214121|emb|CAAE01014641.1|</td><td><i>Tetraodon nigroviridis</i></td><td>81.82</td><td>275</td><td>1E-67</td></tr>
</tbody>
</table>

	gi|500970834|gb| AOOT01010581.1|	*Takifugu flavidus*	83.57	213	6E-58
	gi|347791743|emb| CAAB02001229.1|	*Takifugu rubripes*	83.10	213	3E-56
	gi|144240657|dbj| BAAE01007965.1|	*Oryzias latipes*	84.66	189	2E-52
Talb0488	gi|515356145|dbj| BADN01019576.1|	*Thunnus orientalis*	98.67	300	6E-147
Talb0507	gi|515318313|dbj| BADN01042425.1|	*Thunnus orientalis*	95.38	173	5E-72
	gi|515318313|dbj| BADN01042425.1|	*Thunnus orientalis*	98.11	106	8E-44
	gi|515325380|dbj| BADN01037364.1|	*Thunnus orientalis*	94.39	107	8E-38
	gi|515252999|dbj| BADN01082483.1|	*Thunnus orientalis*	93.46	107	3E-36
	gi|515281925|dbj| BADN01063082.1|	*Thunnus orientalis*	93.46	107	3E-36
Talb0516	gi|515328168|dbj| BADN01035698.1|	*Thunnus orientalis*	98.90	272	8E-133
Talb0546	gi|515370723|dbj| BADN01011470.1|	*Thunnus orientalis*	99.32	292	9E-145
Talb0568	gi|515335671|dbj| BADN01031462.1|	*Thunnus orientalis*	99.31	291	9E-145
Talb0685	gi|515224176|dbj| BADN01099401.1|	*Thunnus orientalis*	98.68	228	2E-109
Talb0717	gi|515260549|dbj| BADN01076265.1|	*Thunnus orientalis*	98.67	301	2E-146
	gi|515270409|dbj| BADN01069663.1|	*Thunnus orientalis*	74.86	183	2E-25
	gi|515324324|dbj| BADN01038173.1|	*Thunnus orientalis*	84.26	108	1E-23
	gi|515328219|dbj| BADN01035669.1|	*Thunnus orientalis*	73.10	197	4E-23
	gi|515287475|dbj| BADN01060510.1|	*Thunnus orientalis*	78.79	132	1E-22
Talb0807	gi|515227109|dbj| BADN01097501.1|	*Thunnus orientalis*	84.08	314	3E-93
	gi|86291212|gb| AANH01011476.1|	*Gasterosteus aculeatus*	84.44	135	2E-33
Talb0822	gi|515250264|dbj| BADN01083826.1|	*Thunnus orientalis*	99.64	277	3E-138
	gi|86302182|gb| AANH01000518.1|	*Gasterosteus aculeatus*	84.15	183	1E-48
	gi|347787760|emb| CAAB02005212.1|	*Takifugu rubripes*	87.12	132	3E-37
	gi|500980279|gb| AOOT01005022.1|	*Takifugu flavidus*	87.12	132	3E-37
	gi|145732098|dbj| BAAF04106369.1|	*Oryzias latipes*	85.00	140	4E-35
Talb0826	gi|515381068|dbj| BADN01005155.1|	*Thunnus orientalis*	98.33	300	7E-146
Talb0827	gi|515369341|dbj| BADN01012321.1|	*Thunnus orientalis*	99.00	300	5E-148
Talb0851	gi|515229589|dbj| BADN01095636.1|	*Thunnus orientalis*	99.62	266	3E-132
	gi|515188049|dbj| BADN01123372.1|	*Thunnus orientalis*	88.29	299	2E-101
	gi|515176813|dbj| BADN01130391.1|	*Thunnus orientalis*	87.67	300	3E-100

	gi|515212905|dbj|BADN01106887.1|	*Thunnus orientalis*	87.63	299	3E-100
	gi|515191948|dbj|BADN01121078.1|	*Thunnus orientalis*	87.63	299	1E-99
Talb0889	gi|515325515|dbj|BADN01037269.1|	*Thunnus orientalis*	98.79	247	4E-118
	gi|515251071|dbj|BADN01083425.1|	*Thunnus orientalis*	87.79	172	5E-53
	gi|515369106|dbj|BADN01012405.1|	*Thunnus orientalis*	76.99	226	4E-41
Talb0891	gi|515209531|dbj|BADN01109204.1|	*Thunnus orientalis*	98.29	293	5E-142
Talb0895	gi|515260647|dbj|BADN01076199.1|	*Thunnus orientalis*	99.00	300	5E-148
	gi|515360150|dbj|BADN01017516.1|	*Thunnus orientalis*	87.44	207	2E-65
	gi|515355542|dbj|BADN01019906.1|	*Thunnus orientalis*	83.49	218	2E-57
	gi|515371553|dbj|BADN01010970.1|	*Thunnus orientalis*	82.25	231	2E-57
	gi|515372463|dbj|BADN01010376.1|	*Thunnus orientalis*	84.39	205	2E-57
Talb0930	gi|515355232|dbj|BADN01020121.1|	*Thunnus orientalis*	99.26	270	8E-133
	gi|515290065|dbj|BADN01060069.1|	*Thunnus orientalis*	76.31	287	2E-44
	gi|515383127|dbj|BADN01004014.1|	*Thunnus orientalis*	76.07	234	7E-39
	gi|515233830|dbj|BADN01092858.1|	*Thunnus orientalis*	85.61	132	4E-35
	gi|515259859|dbj|BADN01076704.1|	*Thunnus orientalis*	84.25	146	5E-34
Talb0952	gi|515337580|dbj|BADN01030232.1|	*Thunnus orientalis*	98.29	293	2E-141
Talb1040	gi|515233275|dbj|BADN01093250.1|	*Thunnus orientalis*	98.68	302	2E-146
	gi|86294233|gb|AANH01008455.1|	*Gasterosteus aculeatus*	87.94	257	1E-86
	gi|144295549|dbj|BAAE01073073.1|	*Oryzias latipes*	79.01	262	8E-57
	gi|347771988|emb|CAAB02020984.1|	*Takifugu rubripes*	79.47	190	8E-38
	gi|347773494|emb|CAAB02019478.1|	*Takifugu rubripes*	79.47	190	8E-38
Talb1083	gi|515277127|dbj|BADN01065696.1|	*Thunnus orientalis*	99.17	242	1E-117
	gi|86298629|gb|AANH01004059.1|	*Gasterosteus aculeatus*	86.99	123	5E-34
	gi|347792035|emb|CAAB02000937.1|	*Takifugu rubripes*	80.15	136	2E-25
	gi|500975981|gb|AOOT01007783.1|	*Takifugu flavidus*	80.15	136	2E-25
	gi|145825269|dbj|BAAF04014125.1|	*Oryzias latipes*	78.32	143	3E-24
Talb1093	gi|515369672|dbj|BADN01012153.1|	*Thunnus orientalis*	98.02	303	9E-145
	gi|86294972|gb|AANH01007716.1|	*Gasterosteus aculeatus*	79.10	177	4E-35
Talb1154	gi|515316918|dbj|BADN01043245.1|	*Thunnus orientalis*	97.83	277	1E-131

<table>
<tr><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td><b>Talb1155</b></td><td>gi|515294433|dbj|BADN01057371.1|</td><td><i>Thunnus orientalis</i></td><td>99.00</td><td>300</td><td>5E-148</td></tr>
<tr><td></td><td>gi|515224360|dbj|BADN01099300.1|</td><td><i>Thunnus orientalis</i></td><td>85.25</td><td>122</td><td>1E-29</td></tr>
<tr><td></td><td>gi|515228967|dbj|BADN01096198.1|</td><td><i>Thunnus orientalis</i></td><td>88.24</td><td>102</td><td>5E-28</td></tr>
<tr><td></td><td>gi|515356441|dbj|BADN01019402.1|</td><td><i>Thunnus orientalis</i></td><td>82.93</td><td>123</td><td>7E-26</td></tr>
<tr><td></td><td>gi|515339993|dbj|BADN01028723.1|</td><td><i>Thunnus orientalis</i></td><td>83.76</td><td>117</td><td>2E-25</td></tr>
<tr><td><b>Talb1211</b></td><td>gi|515273344|dbj|BADN01067979.1|</td><td><i>Thunnus orientalis</i></td><td>98.65</td><td>296</td><td>9E-145</td></tr>
<tr><td><b>Talb1258</b></td><td>gi|515279313|dbj|BADN01064614.1|</td><td><i>Thunnus orientalis</i></td><td>88.16</td><td>228</td><td>6E-77</td></tr>
<tr><td><b>Talb1308</b></td><td>gi|515326303|dbj|BADN01036641.1|</td><td><i>Thunnus orientalis</i></td><td>98.66</td><td>298</td><td>3E-145</td></tr>
<tr><td></td><td>gi|515371735|dbj|BADN01010863.1|</td><td><i>Thunnus orientalis</i></td><td>92.26</td><td>297</td><td>4E-117</td></tr>
<tr><td></td><td>gi|515351645|dbj|BADN01022397.1|</td><td><i>Thunnus orientalis</i></td><td>84.07</td><td>295</td><td>9E-88</td></tr>
<tr><td></td><td>gi|347792126|emb|CAAB02000846.1|</td><td><i>Takifugu rubripes</i></td><td>79.42</td><td>243</td><td>3E-50</td></tr>
<tr><td></td><td>gi|500957774|gb|AOOT01021028.1|</td><td><i>Takifugu flavidus</i></td><td>79.42</td><td>243</td><td>3E-50</td></tr>
<tr><td><b>Talb1383</b></td><td>gi|515212583|dbj|BADN01107107.1|</td><td><i>Thunnus orientalis</i></td><td>98.20</td><td>222</td><td>1E-104</td></tr>
<tr><td></td><td>gi|515374580|dbj|BADN01009104.1|</td><td><i>Thunnus orientalis</i></td><td>84.13</td><td>189</td><td>3E-50</td></tr>
<tr><td></td><td>gi|515261504|dbj|BADN01075708.1|</td><td><i>Thunnus orientalis</i></td><td>81.37</td><td>204</td><td>1E-47</td></tr>
<tr><td></td><td>gi|515232946|dbj|BADN01093439.1|</td><td><i>Thunnus orientalis</i></td><td>80.95</td><td>189</td><td>4E-42</td></tr>
<tr><td></td><td>gi|515282024|dbj|BADN01063026.1|</td><td><i>Thunnus orientalis</i></td><td>75.78</td><td>223</td><td>3E-37</td></tr>
<tr><td><b>Talb1549</b></td><td>gi|515241644|dbj|BADN01088514.1|</td><td><i>Thunnus orientalis</i></td><td>96.73</td><td>306</td><td>6E-141</td></tr>
<tr><td><b>Talb1580</b></td><td>gi|515263687|dbj|BADN01074396.1|</td><td><i>Thunnus orientalis</i></td><td>98.47</td><td>262</td><td>9E-126</td></tr>
<tr><td></td><td>gi|86299967|gb|AANH01002721.1|</td><td><i>Gasterosteus aculeatus</i></td><td>93.83</td><td>81</td><td>7E-26</td></tr>
<tr><td></td><td>gi|145822080|dbj|BAAF04017314.1|</td><td><i>Oryzias latipes</i></td><td>91.46</td><td>82</td><td>1E-23</td></tr>
<tr><td></td><td>gi|144303692|dbj|BAAE01064930.1|</td><td><i>Oryzias latipes</i></td><td>91.46</td><td>82</td><td>1E-23</td></tr>
<tr><td></td><td>gi|347791187|emb|CAAB02001785.1|</td><td><i>Takifugu rubripes</i></td><td>92.21</td><td>77</td><td>1E-22</td></tr>
<tr><td><b>Talb1582</b></td><td>gi|515387962|dbj|BADN01001409.1|</td><td><i>Thunnus orientalis</i></td><td>98.23</td><td>282</td><td>2E-133</td></tr>
<tr><td><b>Talb1911</b></td><td>gi|515389512|dbj|BADN01000529.1|</td><td><i>Thunnus orientalis</i></td><td>98.33</td><td>300</td><td>3E-145</td></tr>
<tr><td></td><td>gi|86300261|gb|AANH01002427.1|</td><td><i>Gasterosteus aculeatus</i></td><td>82.37</td><td>295</td><td>4E-79</td></tr>
<tr><td></td><td>gi|347777495|emb|CAAB02015477.1|</td><td><i>Takifugu rubripes</i></td><td>79.33</td><td>300</td><td>3E-68</td></tr>
<tr><td></td><td>gi|145726451|dbj|BAAF04112016.1|</td><td><i>Oryzias latipes</i></td><td>76.74</td><td>301</td><td>1E-55</td></tr>
<tr><td></td><td>gi|144396827|dbj|BAAE01171872.1|</td><td><i>Oryzias latipes</i></td><td>76.74</td><td>301</td><td>1E-55</td></tr>
<tr><td><b>Talb1933</b></td><td>gi|515328979|dbj|BADN01035253.1|</td><td><i>Thunnus orientalis</i></td><td>90.53</td><td>190</td><td>1E-67</td></tr>
</table>

	gi|515314516|dbj| BADN01044767.1|	*Thunnus orientalis*	89.34	197	4E-67
	gi|515338381|dbj| BADN01029584.1|	*Thunnus orientalis*	89.89	188	2E-65
	gi|515227250|dbj| BADN01097412.1|	*Thunnus orientalis*	88.32	197	6E-65
	gi|515340897|dbj| BADN01028215.1|	*Thunnus orientalis*	88.32	197	6E-65
Talb1952	gi|515368960|dbj| BADN01012487.1|	*Thunnus orientalis*	98.55	138	4E-61
	gi|515368960|dbj| BADN01012487.1|	*Thunnus orientalis*	97.78	135	3E-56
	gi|515285598|dbj| BADN01061520.1|	*Thunnus orientalis*	69.05	336	7E-39
	gi|515388858|dbj| BADN01000924.1|	*Thunnus orientalis*	88.15	135	7E-39
	gi|515350063|dbj| BADN01023102.1|	*Thunnus orientalis*	86.67	135	3E-36
Talb2008	gi|515270198|dbj| BADN01069778.1|	*Thunnus orientalis*	96.04	278	5E-123
Talb2058	gi|515314351|dbj| BADN01044864.1|	*Thunnus orientalis*	98.48	263	2E-127
	gi|515363616|dbj| BADN01015594.1|	*Thunnus orientalis*	83.97	262	1E-73
	gi|144300082|dbj| BAAE01068540.1|	*Oryzias latipes*	74.05	262	5E-40
	gi|144239829|dbj| BAAE01008793.1|	*Oryzias latipes*	74.05	262	5E-40
	gi|144239828|dbj| BAAE01008794.1|	*Oryzias latipes*	73.66	262	2E-38
Talb2064	gi|515268300|dbj| BADN01070998.1|	*Thunnus orientalis*	98.35	243	2E-115
Talb2335	gi|515344409|dbj| BADN01026112.1|	*Thunnus orientalis*	97.89	285	2E-134
Talb2481	gi|515238413|dbj| BADN01090213.1|	*Thunnus orientalis*	98.21	280	2E-134
	gi|515274429|dbj| BADN01067323.1|	*Thunnus orientalis*	82.75	313	3E-88
	gi|515330753|dbj| BADN01034285.1|	*Thunnus orientalis*	82.14	308	8E-82
	gi|515272947|dbj| BADN01068214.1|	*Thunnus orientalis*	83.16	285	1E-79
	gi|515372151|dbj| BADN01010594.1|	*Thunnus orientalis*	81.61	299	2E-78
Talb2545	gi|515263884|dbj| BADN01074266.1|	*Thunnus orientalis*	99.33	300	1E-149
Talb2746	gi|515317961|dbj| BADN01042628.1|	*Thunnus orientalis*	98.97	292	1E-143
	gi|515317938|dbj| BADN01042637.1|	*Thunnus orientalis*	77.57	321	8E-63
	gi|515272344|dbj| BADN01068580.1|	*Thunnus orientalis*	74.28	311	7E-51
	gi|515254950|dbj| BADN01081307.1|	*Thunnus orientalis*	76.88	160	7E-26
	gi|515317949|dbj| BADN01042634.1|	*Thunnus orientalis*	82.05	117	1E-23
Scomberomorus brasiliensis					
Sbra0050	gi|515299986|dbj| BADN01053815.1	*Thunnus orientalis*	89.36	282	3E-100

<table>
<tr><td><b>Sbra0059</b></td><td>gi|515363776|dbj|BADN01015500.1</td><td><i>Thunnus orientalis</i></td><td>88.06</td><td>310</td><td>9E-107</td></tr>
<tr><td></td><td>gi|86294149|gb|AANH01008539.1</td><td><i>Gasterosteus aculeatus</i></td><td>70.49</td><td>349</td><td>2E-38</td></tr>
<tr><td></td><td>gi|47228171|emb|CAAE01014992.1</td><td><i>Tetraodon nigroviridis</i></td><td>95.83</td><td>96</td><td>1E-34</td></tr>
<tr><td></td><td>gi|347788540|emb|CAAB02004432.1</td><td><i>Takifugu rubripes</i></td><td>93.55</td><td>93</td><td>3E-30</td></tr>
<tr><td></td><td>gi|500934807|gb|AOOT01036385.1</td><td><i>Takifugu flavidus</i></td><td>93.55</td><td>93</td><td>3E-30</td></tr>
<tr><td><b>Sbra0087</b></td><td>gi|515227109|dbj|BADN01097501.1</td><td><i>Thunnus orientalis</i></td><td>83.81</td><td>315</td><td>5E-91</td></tr>
<tr><td></td><td>gi|86291212|gb|AANH01011476.1</td><td><i>Gasterosteus aculeatus</i></td><td>83.82</td><td>136</td><td>7E-32</td></tr>
<tr><td><b>Sbra0099</b></td><td>gi|515312988|dbj|BADN01045683.1</td><td><i>Thunnus orientalis</i></td><td>87.96</td><td>299</td><td>1E-99</td></tr>
<tr><td><b>Sbra0112</b></td><td>gi|515294585|dbj|BADN01057280.1</td><td><i>Thunnus orientalis</i></td><td>85.81</td><td>310</td><td>2E-96</td></tr>
<tr><td><b>Sbra0123</b></td><td>gi|515334059|dbj|BADN01032377.1</td><td><i>Thunnus orientalis</i></td><td>84.54</td><td>304</td><td>2E-90</td></tr>
<tr><td></td><td>gi|515248859|dbj|BADN01084536.1</td><td><i>Thunnus orientalis</i></td><td>81.71</td><td>328</td><td>3E-87</td></tr>
<tr><td></td><td>gi|515294290|dbj|BADN01057452.1</td><td><i>Thunnus orientalis</i></td><td>81.40</td><td>328</td><td>3E-87</td></tr>
<tr><td></td><td>gi|515383989|dbj|BADN01003491.1</td><td><i>Thunnus orientalis</i></td><td>81.50</td><td>319</td><td>3E-87</td></tr>
<tr><td></td><td>gi|515205391|dbj|BADN01111944.1</td><td><i>Thunnus orientalis</i></td><td>81.27</td><td>331</td><td>1E-86</td></tr>
<tr><td><b>Sbra0126</b></td><td>gi|515289916|dbj|BADN01060168.1</td><td><i>Thunnus orientalis</i></td><td>88.97</td><td>290</td><td>4E-98</td></tr>
<tr><td><b>Sbra0130</b></td><td>gi|515285576|dbj|BADN01061533.1</td><td><i>Thunnus orientalis</i></td><td>92.86</td><td>294</td><td>1E-117</td></tr>
<tr><td><b>Sbra0142</b></td><td>gi|515348011|dbj|BADN01024151.1</td><td><i>Thunnus orientalis</i></td><td>87.33</td><td>300</td><td>3E-99</td></tr>
<tr><td></td><td>gi|86297915|gb|AANH01004773.1</td><td><i>Gasterosteus aculeatus</i></td><td>86.29</td><td>124</td><td>6E-33</td></tr>
<tr><td></td><td>gi|86297915|gb|AANH01004773.1</td><td><i>Gasterosteus aculeatus</i></td><td>83.58</td><td>134</td><td>9E-31</td></tr>
<tr><td></td><td>gi|47228376|emb|CAAE01014764.1</td><td><i>Tetraodon nigroviridis</i></td><td>71.03</td><td>252</td><td>4E-29</td></tr>
<tr><td></td><td>gi|347789642|emb|CAAB02003330.1</td><td><i>Takifugu rubripes</i></td><td>84.38</td><td>96</td><td>5E-21</td></tr>
<tr><td><b>Sbra0145</b></td><td>gi|515253081|dbj|BADN01082439.1</td><td><i>Thunnus orientalis</i></td><td>90.41</td><td>292</td><td>2E-108</td></tr>
<tr><td><b>Sbra0197</b></td><td>gi|515264243|dbj|BADN01074058.1</td><td><i>Thunnus orientalis</i></td><td>96.49</td><td>285</td><td>1E-130</td></tr>
<tr><td></td><td>gi|86293133|gb|AANH01009555.1</td><td><i>Gasterosteus aculeatus</i></td><td>82.62</td><td>282</td><td>7E-76</td></tr>
<tr><td></td><td>gi|500962888|gb|AOOT01017702.1</td><td><i>Takifugu flavidus</i></td><td>75.68</td><td>259</td><td>2E-44</td></tr>
<tr><td></td><td>gi|347784301|emb|CAAB02008671.1</td><td><i>Takifugu rubripes</i></td><td>75.68</td><td>259</td><td>8E-44</td></tr>
<tr><td></td><td>gi|144262136|dbj|BAAE01026486.1</td><td><i>Oryzias latipes</i></td><td>75.46</td><td>273</td><td>1E-42</td></tr>
<tr><td><b>Sbra0200</b></td><td>gi|515382484|dbj|BADN01004351.1</td><td><i>Thunnus orientalis</i></td><td>79.39</td><td>330</td><td>2E-84</td></tr>
<tr><td></td><td>gi|515379735|dbj|BADN01005955.1</td><td><i>Thunnus orientalis</i></td><td>92.13</td><td>216</td><td>2E-82</td></tr>
<tr><td></td><td>gi|515343862|dbj|BADN01026441.1</td><td><i>Thunnus orientalis</i></td><td>90.13</td><td>223</td><td>3E-81</td></tr>
</table>

<table>
<tr><th></th><th></th><th></th><th></th><th></th><th></th></tr>
<tr><td></td><td>gi|515327079|dbj|BADN01036109.1</td><td><em>Thunnus orientalis</em></td><td>90.18</td><td>224</td><td>1E-80</td></tr>
<tr><td></td><td>gi|515326981|dbj|BADN01036163.1</td><td><em>Thunnus orientalis</em></td><td>89.78</td><td>225</td><td>3E-80</td></tr>
<tr><td><strong>Sbra0256</strong></td><td>gi|515327628|dbj|BADN01035977.1</td><td><em>Thunnus orientalis</em></td><td>99.00</td><td>300</td><td>5E-148</td></tr>
<tr><td></td><td>gi|347792070|emb|CAAB02000902.1</td><td><em>Takifugu rubripes</em></td><td>87.89</td><td>256</td><td>4E-86</td></tr>
<tr><td></td><td>gi|500928793|gb|AOOT01039262.1</td><td><em>Takifugu flavidus</em></td><td>87.89</td><td>256</td><td>4E-86</td></tr>
<tr><td></td><td>gi|47219502|emb|CAAE01015009.1</td><td><em>Tetraodon nigroviridis</em></td><td>85.07</td><td>268</td><td>3E-80</td></tr>
<tr><td></td><td>gi|86298680|gb|AANH01004008.1</td><td><em>Gasterosteus aculeatus</em></td><td>79.81</td><td>312</td><td>4E-73</td></tr>
<tr><td><strong>Sbra0265</strong></td><td>gi|515304550|dbj|BADN01050863.1</td><td><em>Thunnus orientalis</em></td><td>84.10</td><td>195</td><td>9E-50</td></tr>
<tr><td></td><td>gi|515354125|dbj|BADN01020894.1</td><td><em>Thunnus orientalis</em></td><td>75.49</td><td>204</td><td>5E-34</td></tr>
<tr><td></td><td>gi|515372155|dbj|BADN01010591.1</td><td><em>Thunnus orientalis</em></td><td>75.12</td><td>201</td><td>2E-32</td></tr>
<tr><td></td><td>gi|515306374|dbj|BADN01049731.1</td><td><em>Thunnus orientalis</em></td><td>74.40</td><td>207</td><td>9E-31</td></tr>
<tr><td></td><td>gi|515341827|dbj|BADN01027665.1</td><td><em>Thunnus orientalis</em></td><td>74.63</td><td>201</td><td>1E-29</td></tr>
<tr><td><strong>Sbra0286</strong></td><td>gi|515336401|dbj|BADN01031032.1</td><td><em>Thunnus orientalis</em></td><td>89.14</td><td>304</td><td>2E-108</td></tr>
<tr><td><strong>Sbra0292</strong></td><td>gi|515361942|dbj|BADN01016496.1</td><td><em>Thunnus orientalis</em></td><td>98.66</td><td>299</td><td>7E-146</td></tr>
<tr><td><strong>Sbra0325</strong></td><td>gi|515290967|dbj|BADN01059655.1</td><td><em>Thunnus orientalis</em></td><td>78.31</td><td>189</td><td>4E-35</td></tr>
<tr><td></td><td>gi|515363351|dbj|BADN01015734.1</td><td><em>Thunnus orientalis</em></td><td>78.16</td><td>174</td><td>1E-34</td></tr>
<tr><td></td><td>gi|515367094|dbj|BADN01013607.1</td><td><em>Thunnus orientalis</em></td><td>77.97</td><td>177</td><td>1E-34</td></tr>
<tr><td></td><td>gi|515314293|dbj|BADN01044900.1</td><td><em>Thunnus orientalis</em></td><td>78.57</td><td>168</td><td>5E-34</td></tr>
<tr><td></td><td>gi|515233122|dbj|BADN01093336.1</td><td><em>Thunnus orientalis</em></td><td>77.90</td><td>181</td><td>6E-33</td></tr>
<tr><td></td><td>gi|515312184|dbj|BADN01046252.1</td><td><em>Thunnus orientalis</em></td><td>79.82</td><td>114</td><td>6E-20</td></tr>
<tr><td><strong>Sbra0374</strong></td><td>gi|515318867|dbj|BADN01042090.1</td><td><em>Thunnus orientalis</em></td><td>91.29</td><td>264</td><td>1E-98</td></tr>
<tr><td></td><td>gi|86297270|gb|AANH01005418.1</td><td><em>Gasterosteus aculeatus</em></td><td>89.53</td><td>172</td><td>2E-58</td></tr>
<tr><td></td><td>gi|347783047|emb|CAAB02009925.1</td><td><em>Takifugu rubripes</em></td><td>87.06</td><td>170</td><td>6E-52</td></tr>
<tr><td></td><td>gi|500894733|gb|AOOT01064962.1</td><td><em>Takifugu flavidus</em></td><td>87.06</td><td>170</td><td>6E-52</td></tr>
<tr><td></td><td>gi|47227898|emb|CAAE01014542.1</td><td><em>Tetraodon nigroviridis</em></td><td>86.55</td><td>171</td><td>2E-51</td></tr>
<tr><td><strong>Sbra0438</strong></td><td>gi|515252402|dbj|BADN01082787.1</td><td><em>Thunnus orientalis</em></td><td>94.63</td><td>298</td><td>5E-129</td></tr>
<tr><td></td><td>gi|86295846|gb|AANH01006842.1</td><td><em>Gasterosteus aculeatus</em></td><td>81.88</td><td>298</td><td>2E-77</td></tr>
<tr><td></td><td>gi|347788573|emb|CAAB02004399.1</td><td><em>Takifugu rubripes</em></td><td>81.14</td><td>297</td><td>5E-72</td></tr>
<tr><td></td><td>gi|500969705|gb|AOOT01011105.1</td><td><em>Takifugu flavidus</em></td><td>81.98</td><td>283</td><td>6E-71</td></tr>
<tr><td></td><td>gi|47224214|emb|CAAE01015003.1</td><td><em>Tetraodon nigroviridis</em></td><td>79.86</td><td>278</td><td>8E-63</td></tr>
</table>

Sbra0447	gi|515325784|dbj| BADN01037064.1	*Thunnus orientalis*	92.33	300	8E-120
	gi|86301570|gb| AANH01001130.1	*Gasterosteus aculeatus*	80.00	130	3E-23
	gi|347783156|emb| CAAB02009816.1	*Takifugu rubripes*	88.51	87	1E-21
	gi|500965491|gb| AOOT01015116.1	*Takifugu flavidus*	87.36	87	2E-20
Sbra0455	gi|515370723|dbj| BADN01011470.1	*Thunnus orientalis*	89.04	301	2E-107
	gi|515245812|dbj| BADN01086174.1	*Thunnus orientalis*	81.68	131	8E-25
	gi|515199904|dbj| BADN01115926.1	*Thunnus orientalis*	85.19	108	1E-23
	gi|515292618|dbj| BADN01058572.1	*Thunnus orientalis*	77.33	150	1E-23
	gi|515319827|dbj| BADN01041424.1	*Thunnus orientalis*	81.45	124	1E-23
Sbra0458	gi|515354108|dbj| BADN01020906.1	*Thunnus orientalis*	93.36	301	1E-124
	gi|86297605|gb| AANH01005083.1	*Gasterosteus aculeatus*	72.80	261	5E-34
Sbra0463	gi|515295033|dbj| BADN01056989.1	*Thunnus orientalis*	94.98	279	7E-121
	gi|47214121|emb| CAAE01014641.1	*Tetraodon nigroviridis*	81.52	276	4E-66
	gi|500970834|gb| AOOT01010581.1	*Takifugu flavidus*	83.18	214	9E-56
	gi|347791743|emb| CAAB02001229.1	*Takifugu rubripes*	82.71	214	1E-54
	gi|144240657|dbj| BAAE01007965.1	*Oryzias latipes*	84.21	190	7E-51
Sbra0484	gi|515234814|dbj| BADN01092220.1	*Thunnus orientalis*	83.91	317	7E-95
	gi|86295010|gb| AANH01007678.1	*Gasterosteus aculeatus*	76.01	296	1E-53
Sbra0502	gi|515377060|dbj| BADN01007716.1	*Thunnus orientalis*	73.53	204	3E-30
	gi|515254585|dbj| BADN01081522.1	*Thunnus orientalis*	85.71	119	4E-29
	gi|515361462|dbj| BADN01016769.1	*Thunnus orientalis*	82.31	130	4E-29
	gi|515270294|dbj| BADN01069733.1	*Thunnus orientalis*	79.70	133	2E-25
	gi|515308232|dbj| BADN01048740.1	*Thunnus orientalis*	80.00	130	2E-25
Sbra0660	gi|515325339|dbj| BADN01037392.1	*Thunnus orientalis*	78.99	138	2E-26
Sbra0693	gi|515368097|dbj| BADN01013044.1	*Thunnus orientalis*	89.00	291	3E-99
	gi|515205391|dbj| BADN01111944.1	*Thunnus orientalis*	87.33	300	1E-97
	gi|515323293|dbj| BADN01038832.1	*Thunnus orientalis*	86.75	302	6E-96
	gi|515294634|dbj| BADN01057248.1	*Thunnus orientalis*	87.63	291	2E-95
	gi|515294290|dbj| BADN01057452.1	*Thunnus orientalis*	86.67	300	3E-94
Sbra0700	gi|515330200|dbj| BADN01034581.1	*Thunnus orientalis*	93.21	265	2E-108

Sbra0713	gi|515215954|dbj|BADN01104931.1	*Thunnus orientalis*	93.43	289	3E-118
	gi|86297275|gb|AANH01005413.1	*Gasterosteus aculeatus*	86.34	183	4E-54
	gi|47187288|emb|CAAE01021646.1	*Tetraodon nigroviridis*	82.63	167	5E-40
	gi|347788310|emb|CAAB02004662.1	*Takifugu rubripes*	82.04	167	6E-39
	gi|500955977|gb|AOOT01021985.1	*Takifugu flavidus*	81.44	167	3E-37
Sbra0787	gi|515239114|dbj|BADN01089826.1	*Thunnus orientalis*	80.86	303	6E-71
	gi|47222409|emb|CAAE01015120.1	*Tetraodon nigroviridis*	89.33	150	7E-45
	gi|86295984|gb|AANH01006704.1	*Gasterosteus aculeatus*	89.33	150	7E-45
	gi|500916540|gb|AOOT01047462.1	*Takifugu flavidus*	88.51	148	3E-42
	gi|347783374|emb|CAAB02009598.1	*Takifugu rubripes*	87.84	148	4E-41
Sbra0801	gi|515286291|dbj|BADN01061105.1	*Thunnus orientalis*	83.22	286	1E-80
	gi|515221310|dbj|BADN01101127.1	*Thunnus orientalis*	86.08	194	2E-57
	gi|515379668|dbj|BADN01005998.1	*Thunnus orientalis*	87.36	182	3E-56
	gi|515384366|dbj|BADN01003282.1	*Thunnus orientalis*	86.96	184	9E-56
	gi|515223688|dbj|BADN01099714.1	*Thunnus orientalis*	86.49	185	3E-55
Sbra0932	gi|515320066|dbj|BADN01041240.1	*Thunnus orientalis*	94.31	299	5E-129
	gi|86302013|gb|AANH01000687.1	*Gasterosteus aculeatus*	85.71	210	8E-63
	gi|347792372|emb|CAAB02000600.1	*Takifugu rubripes*	86.70	188	6E-58
	gi|500909073|gb|AOOT01052940.1	*Takifugu flavidus*	86.70	188	6E-58
	gi|145808719|dbj|BAAF04030672.1	*Oryzias latipes*	84.50	200	3E-55
Sbra0933	gi|515340160|dbj|BADN01028629.1	*Thunnus orientalis*	88.21	195	7E-64
Sbra0981	gi|515293715|dbj|BADN01057796.1	*Thunnus orientalis*	82.08	307	2E-76
	gi|86299576|gb|AANH01003112.1	*Gasterosteus aculeatus*	84.21	95	6E-20
Sbra1035	gi|515279179|dbj|BADN01064697.1	*Thunnus orientalis*	82.26	248	1E-60
	gi|515351417|dbj|BADN01022572.1	*Thunnus orientalis*	80.71	254	8E-57
	gi|515300363|dbj|BADN01053599.1	*Thunnus orientalis*	78.26	276	1E-54
	gi|515336158|dbj|BADN01031179.1	*Thunnus orientalis*	80.25	243	4E-54
	gi|515298885|dbj|BADN01054481.1	*Thunnus orientalis*	79.45	253	1E-53
Sbra1038	gi|515321953|dbj|BADN01039821.1	*Thunnus orientalis*	87.74	106	4E-28
Sbra1095	gi|515285505|dbj|BADN01061574.1	*Thunnus orientalis*	80.49	287	2E-69

Sbra1146	gi|515336324|dbj|BADN01031083.1	*Thunnus orientalis*	84.15	328	3E-99
Sbra1196	gi|515320769|dbj|BADN01040632.1	*Thunnus orientalis*	89.90	297	1E-105
Sbra1304	gi|515322312|dbj|BADN01039558.1	*Thunnus orientalis*	83.82	204	5E-53
	gi|515368700|dbj|BADN01012640.1	*Thunnus orientalis*	82.14	196	5E-46
	gi|515308477|dbj|BADN01048604.1	*Thunnus orientalis*	79.80	203	1E-42
	gi|515370677|dbj|BADN01011505.1	*Thunnus orientalis*	80.30	203	1E-42
	gi|515312528|dbj|BADN01045991.1	*Thunnus orientalis*	80.00	200	1E-40
Sbra1361	gi|515320769|dbj|BADN01040632.1	*Thunnus orientalis*	89.52	248	1E-85
Sbra1397	gi|515318162|dbj|BADN01042512.1	*Thunnus orientalis*	90.24	123	6E-39
	gi|515209778|dbj|BADN01109002.1	*Thunnus orientalis*	89.52	124	3E-37
	gi|515325028|dbj|BADN01037637.1	*Thunnus orientalis*	89.92	119	3E-37
	gi|515295297|dbj|BADN01056821.1	*Thunnus orientalis*	88.43	121	1E-35
	gi|515370397|dbj|BADN01011677.1	*Thunnus orientalis*	88.62	123	1E-35
Sbra1532	gi|515372779|dbj|BADN01010149.1	*Thunnus orientalis*	93.52	247	4E-98
	gi|86298006|gb|AANH01004682.1	*Gasterosteus aculeatus*	85.71	231	4E-66
	gi|347789903|emb|CAAB02003069.1	*Takifugu rubripes*	81.02	137	2E-25
	gi|144405276|dbj|BAAE01163423.1	*Oryzias latipes*	91.67	84	8E-25
	gi|145781486|dbj|BAAF04057905.1	*Oryzias latipes*	90.48	84	1E-23
Sbra1614	gi|515262913|dbj|BADN01074846.1	*Thunnus orientalis*	88.89	306	9E-107
	gi|347777507|emb|CAAB02015465.1	*Takifugu rubripes*	89.74	117	1E-35
	gi|500932711|gb|AOOT01037466.1	*Takifugu flavidus*	88.89	117	1E-34
	gi|47225224|emb|CAAE01015008.1	*Tetraodon nigroviridis*	85.09	114	5E-27
Sbra1626	gi|515316561|dbj|BADN01043469.1	*Thunnus orientalis*	86.60	321	9E-107
	gi|86300421|gb|AANH01002267.1	*Gasterosteus aculeatus*	90.91	121	5E-40
	gi|86300421|gb|AANH01002267.1	*Gasterosteus aculeatus*	88.18	110	4E-29
	gi|347790278|emb|CAAB02002694.1	*Takifugu rubripes*	71.52	323	2E-33
	gi|500824872|gb|AOOT01088351.1	*Takifugu flavidus*	71.52	323	2E-33
Sbra1675	gi|145833099|dbj|BAAF04006295.1	*Oryzias latipes*	88.16	245	7E-83
	gi|86302332|gb|AANH01000368.1	*Gasterosteus aculeatus*	87.60	250	7E-83
	gi|500956212|gb|AOOT01021857.1	*Takifugu flavidus*	87.25	251	8E-82

<table>
<tr><td></td><td>gi|47230322|emb|CAAE01014581.1</td><td><i>Tetraodon nigroviridis</i></td><td>86.85</td><td>251</td><td>1E-80</td></tr>
<tr><td></td><td>gi|347791685|emb|CAAB02001287.1</td><td><i>Takifugu rubripes</i></td><td>86.85</td><td>251</td><td>1E-80</td></tr>
<tr><td><b>Sbra1706</b></td><td>gi|515364789|dbj|BADN01014892.1</td><td><i>Thunnus orientalis</i></td><td>93.17</td><td>293</td><td>8E-120</td></tr>
<tr><td></td><td>gi|86297200|gb|AANH01005488.1</td><td><i>Gasterosteus aculeatus</i></td><td>93.64</td><td>110</td><td>6E-39</td></tr>
<tr><td></td><td>gi|145809177|dbj|BAAF04030214.1</td><td><i>Oryzias latipes</i></td><td>92.73</td><td>110</td><td>8E-38</td></tr>
<tr><td></td><td>gi|347784554|emb|CAAB02008418.1</td><td><i>Takifugu rubripes</i></td><td>90.65</td><td>107</td><td>2E-33</td></tr>
<tr><td></td><td>gi|500949407|gb|AOOT01026151.1</td><td><i>Takifugu flavidus</i></td><td>90.65</td><td>107</td><td>2E-33</td></tr>
<tr><td><b>Sbra1957</b></td><td>gi|515301779|dbj|BADN01052651.1</td><td><i>Thunnus orientalis</i></td><td>88.89</td><td>288</td><td>1E-97</td></tr>
<tr><td></td><td>gi|86297218|gb|AANH01005470.1</td><td><i>Gasterosteus aculeatus</i></td><td>76.51</td><td>281</td><td>2E-46</td></tr>
<tr><td></td><td>gi|347784203|emb|CAAB02008769.1</td><td><i>Takifugu rubripes</i></td><td>70.92</td><td>282</td><td>4E-28</td></tr>
<tr><td></td><td>gi|145809068|dbj|BAAF04030323.1</td><td><i>Oryzias latipes</i></td><td>71.53</td><td>274</td><td>4E-28</td></tr>
<tr><td></td><td>gi|500935077|gb|AOOT01036238.1</td><td><i>Takifugu flavidus</i></td><td>71.26</td><td>261</td><td>7E-26</td></tr>
<tr><td><b>Sbra1985</b></td><td>gi|515382811|dbj|BADN01004184.1</td><td><i>Thunnus orientalis</i></td><td>79.58</td><td>240</td><td>5E-59</td></tr>
<tr><td><b>Sbra2061</b></td><td>gi|515375547|dbj|BADN01008566.1</td><td><i>Thunnus orientalis</i></td><td>85.91</td><td>298</td><td>9E-94</td></tr>
<tr><td></td><td>gi|515342334|dbj|BADN01027367.1</td><td><i>Thunnus orientalis</i></td><td>77.91</td><td>172</td><td>3E-30</td></tr>
<tr><td></td><td>gi|515258476|dbj|BADN01077931.1</td><td><i>Thunnus orientalis</i></td><td>78.47</td><td>144</td><td>2E-26</td></tr>
<tr><td></td><td>gi|515320937|dbj|BADN01040486.1</td><td><i>Thunnus orientalis</i></td><td>78.87</td><td>142</td><td>7E-26</td></tr>
<tr><td></td><td>gi|515320937|dbj|BADN01040486.1</td><td><i>Thunnus orientalis</i></td><td>76.19</td><td>147</td><td>1E-22</td></tr>
<tr><td><b>Sbra2083</b></td><td>gi|515297224|dbj|BADN01055556.1</td><td><i>Thunnus orientalis</i></td><td>87.60</td><td>242</td><td>1E-79</td></tr>
<tr><td><b>Sbra2272</b></td><td>gi|515212324|dbj|BADN01107264.1</td><td><i>Thunnus orientalis</i></td><td>75.00</td><td>224</td><td>5E-34</td></tr>
<tr><td><b>Sbra2735</b></td><td>gi|515256898|dbj|BADN01079457.1</td><td><i>Thunnus orientalis</i></td><td>87.41</td><td>286</td><td>3E-94</td></tr>
<tr><td><b>Sbra2880</b></td><td>gi|515369373|dbj|BADN01012303.1</td><td><i>Thunnus orientalis</i></td><td>95.93</td><td>246</td><td>7E-108</td></tr>
<tr><td><b>Sbra2947</b></td><td>gi|515322651|dbj|BADN01039306.1</td><td><i>Thunnus orientalis</i></td><td>96.90</td><td>258</td><td>3E-118</td></tr>
<tr><td></td><td>gi|47224785|emb|CAAE01014974.1</td><td><i>Tetraodon nigroviridis</i></td><td>90.16</td><td>254</td><td>3E-93</td></tr>
<tr><td></td><td>gi|347786836|emb|CAAB02006136.1</td><td><i>Takifugu rubripes</i></td><td>89.11</td><td>248</td><td>4E-86</td></tr>
<tr><td></td><td>gi|145760046|dbj|BAAF04078933.1</td><td><i>Oryzias latipes</i></td><td>89.08</td><td>238</td><td>7E-83</td></tr>
<tr><td></td><td>gi|144299777|dbj|BAAE01068845.1</td><td><i>Oryzias latipes</i></td><td>88.66</td><td>238</td><td>1E-80</td></tr>
</table>

GENERAL CONCLUSIONS

Restriction site Associated DNA (RAD) genotyping proved to be a powerful tool for detecting and analyzing polymorphisms in fish species and it can be used in the context of population genetics study and to characterize broodstocks.

In the works presented here, thanks to the higher analytical resolution achievable with this method, we were able to discover a subtle genetic structure in dolphinfish and characterize sex related markers never found before in the species.

Moreover, through the analysis of more than 1000 sea bream samples we were able to provide for the first time a wide geographical scale population genetics analysis based on more than one thousand SNP, we characterized some of the major European broodstocks and collected useful information for assessing the potential impact of sea bream aquaculture in the wild populations, which is a fundamental step toward the development of sustainable aquaculture of the species.

Nevertheless, as reported in the first paper presented here, we acknowledge that care should be taken when developing and using RAD technique (and in particular ddRAD), since biases can arise by sub-optimal library preparation technique and bioinformatic approaches used. In general, particular attention should be put in mixing DNA samples of different qualities, in fragment size selection steps (when applied) and in the selection of number of samples analyzed simultaneously, that should be set taking in consideration biases in samples representation. A reference genome, even if not of high quality, ensured detection of higher number of shared markers and also more reliable results.

Altogether, the works collected make up an important source of information whose ultimate usefulness is to support the management of marine fish resources, including the development of aquaculture and the preservation of marine biodiversity. Finally, we acknowledge that the results presented might still not be sufficient to draw ultimate conclusions on the best management approaches to be used, thus we encourage additional studies with the purpose of increasing the genetic information available to stakeholders and improve effectiveness of conservation policy in the future.

BIBLIOGRAPHY

1. Bureau, P. R. *World population data sheet*. (2007).
2. Wilcove, D. S., Rothstein, D., Dubow, J., Phillips, A. & Losos, E. Quantifying threats to imperiled species in the United States. *BioScience* **48,** 607–615 (1998).
3. Atalah, javier. Over exploitation. (2010).
4. Pauly, D. & Froese, R. Comments on FAO’s State of Fisheries and Aquaculture, or ‘SOFIA 2010’. *Mar. Policy* **36,** 746–752 (2012).
5. Worm, B. *et al.* Impacts of biodiversity loss on ocean ecosystem services. *science* **314,** 787–790 (2006).
6. Thompson, R. C., Crowe, T. P. & Hawkins, S. J. Rocky intertidal communities: past environmental changes, present status and predictions for the next 25 years. *Environ. Conserv.* **29,** 168–191 (2002).
7. Pauly, D. Anecdotes and the shifting baseline syndrome of fisheries. *Trends Ecol. Evol.* **10,** 430 (1995).
8. Christy, F. T. *Territorial use rights in marine fisheries: definitions and conditions*. (Food & Agriculture Org., 1982).
9. Papoutsoglou, S. E. Impact of aquaculture on the aquatic environment in relation to applied production systems. *EAS Spec. Publ.* (1992).
10. Naylor, R. *et al.* Fugitive salmon: assessing the risks of escaped fish from net-pen aquaculture. *Bioscience* **55,** 427–437 (2005).
11. Arechavala-Lopez, P., Uglem, I., Fernandez-Jover, D., Bayle-Sempere, J. T. & Sanchez-Jerez, P. Immediate post-escape behaviour of farmed seabass (Dicentrarchus labrax L.) in the Mediterranean Sea. *J. Appl. Ichthyol.* **27,** 1375–1378 (2011).
12. Arechavala-Lopez, P., Uglem, I., Fernandez-Jover, D., Bayle-Sempere, J. T. & Sanchez-Jerez, P. Post-escape dispersion of farmed seabream (Sparus aurata L.) and recaptures by local fisheries in the Western Mediterranean Sea. *Fish. Res.* **121,** 126–135 (2012).
13. Taranger, G. L. *et al.* Control of puberty in farmed fish. *Gen. Comp. Endocrinol.* **165,** 483–515 (2010).
14. Waples, R. S., Hindar, K. & Hard, J. J. Genetic risks associated with marine aquaculture. (2012).

15. Reed, D. H. & Frankham, R. Correlation between fitness and genetic diversity. *Conserv. Biol.* **17,** 230–237 (2003).

16. Lande, R. & Barrowclough, G. F. Effective population size, genetic variation, and their use in population management. *Viable Popul. Conserv.* 87–123 (1987).

17. Ryman, N. & Laikre, L. Effects of supportive breeding on the genetically effective population size. *Conserv. Biol.* 325–329 (1991).

18. Lorenzen, K., Beveridge, M. & Mangel, M. Cultured fish: integrative biology and management of domestication and interactions with wild fish. *Biol. Rev.* **87,** 639–660 (2012).

19. Nei, M. & Tajima, F. Genetic drift and estimation of effective population size. *Genetics* **98,** 625–640 (1981).

20. Waples, R. S. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121,** 379–391 (1989).

21. Wang, J. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genet. Res.* **78,** 243–257 (2001).

22. Hill, W. G. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* **38,** 209–216 (1981).

23. Pudovkin, A. I., Zaykin, D. V. & Hedgecock, D. On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics* **144,** 383–387 (1996).

24. Waples, R. S. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci*. *Conserv. Genet.* **7,** 167–184 (2006).

25. Ward, R. D. The importance of identifying spatial population structure in restocking and stock enhancement programmes. *Fish. Res.* **80,** 9–18 (2006).

26. Araki, H., Cooper, B. & Blouin, M. S. Genetic effects of captive breeding cause a rapid, cumulative fitness decline in the wild. *Science* **318,** 100–103 (2007).

27. Charlesworth, D. & Charlesworth, B. Inbreeding depression and its evolutionary consequences. *Annu. Rev. Ecol. Syst.* 237–268 (1987).

28. Chavanne, H. *et al.* A comprehensive survey on selective breeding programs and seed market in the European aquaculture fish industry.

Aquac. Int. 1–21

29. Kapuscinski, A. R. *Environmental risk assessment of genetically modified organisms*. **3,** (CABI, 2007).
30. Coscia, I., Vogiatzi, E., Kotoulas, G., Tsigenopoulos, C. S. & Mariani, S. Exploring neutral and adaptive processes in expanding populations of gilthead sea bream, Sparus aurata L., in the North-East Atlantic. *Heredity* **108,** 537–546 (2012).
31. Palma, J. *et al.* Developmental stability and genetic heterozygosity in wild and cultured stocks of gilthead sea bream (Sparus aurata). *J. Mar. Biol. Assoc. UK* **81,** 283–288 (2001).
32. Rossi, A., Perrone, E. & Sola, L. Genetic structure of gilthead seabream, Sparus aurata, in the Central Mediterranean Sea. *Open Life Sci.* **1,** 636–647 (2006).
33. Ben Slimen, H. *et al.* Genetic differentiation between populations of gilthead seabream (Sparusaurata) along the Tunisian coast. *Cybium* **28,** 45–50 (2004).
34. Karaiskou, N., Triantafyllidis, A., Katsares, V., Abatzopoulos, T. J. & Triantaphyllidis, C. Microsatellite variability of wild and farmed populations of Sparus aurata. *J. Fish Biol.* **74,** 1816–1825 (2009).
35. Šegvić-Bubić, T. *et al.* Population genetic structure of reared and wild gilthead sea bream (Sparus aurata) in the Adriatic Sea inferred with microsatellite loci. *Aquaculture* **318,** 309–315 (2011).
36. FAO. FishStat plus. Capture production 1950-2012; Aquaculture production 1950-2012. (2014).
37. Pavlidis, M. & Mylonas, C. *Sparidae: Biology and aquaculture of gilthead sea bream and other species*. (John Wiley & Sons, 2011).
38. Loukovitis, D. *et al.* Genetic variation in farmed populations of the gilthead sea bream Sparus aurata in Greece using microsatellite DNA markers. *Aquac. Res.* **43,** 239–246 (2012).
39. Bilgen, G., Akhan, S., Arabaci, M. & Oguz, I. Genetic diversity of gilthead seabream (Sparus aurata) broodstocks as determined by RAPD-PCR. (2007).
40. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30,** 418–426 (2014).

41. Wang, S., Meyer, E., McKay, J. K. & Matz, M. V. 2b-RAD: a simple and flexible method for genome-wide genotyping. *Nat. Methods* **9,** 808–810 (2012).

42. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **7,** e37135 (2012).

43. Cruz, V. P. *et al.* Identification and validation of single nucleotide polymorphisms as tools to detect hybridization and population structure in freshwater stingrays. *Mol. Ecol. Resour.* (2016).

44. Davey, J. W. *et al.* Special features of RAD Sequencing data: implications for genotyping. *Mol. Ecol.* **22,** 3151–3164 (2013).

45. Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R. & Hohenlohe, P. A. Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* **22,** 2841–2847 (2013).

46. Liu, L. *et al.* Comparison of next-generation sequencing systems. *BioMed Res. Int.* **2012,** (2012).

47. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One* **3,** e3376 (2008).

48. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17,** 81–92 (2016).

49. Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22,** 3179–3190 (2013).

50. Puritz, J. B. *et al.* Demystifying the RAD fad. *Mol. Ecol.* **23,** 5937–5942 (2014).

51. DaCosta, J. M. & Sorenson, M. D. Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol. (2014).

52. Mastretta-Yanes, A. *et al.* Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15,** 28–41 (2015).

53. Taberlet, P. *et al.* Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* **24,** 3189–3194 (1996).

54. Xu, J., Turner, A., Little, J., Bleecker, E. R. & Meyers, D. A. Positive results

in association studies are associated with departure from Hardy-Weinberg equilibrium: hint for genotyping error? *Hum. Genet.* **111,** 573–574 (2002).

55. Gomes, I. *et al.* Hardy–Weinberg quality control. *Ann. Hum. Genet.* **63,** 535–538 (1999).
56. Tine, M. *et al.* European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. *Nat. Commun.* **5,** (2014).
57. Figueras, A. *et al.* Whole genome sequencing of turbot (Scophthalmus maximus; Pleuronectiformes): a fish adapted to demersal life. *DNA Res.* dsw007 (2016).
58. Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W. & Postlethwait, J. H. Stacks: building and genotyping loci de novo from short-read sequences. *G3 Genes Genomes Genet.* **1,** 171–182 (2011).
59. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22,** 3124–3140 (2013).
60. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10,** R25 (2009).
61. Team, R. C. R: A language and environment for statistical computing. (2013).
62. Fox, J. Getting started with the R commander: a basic-statistics graphical user interface to R. *J. Stat. Softw.* **14,** 1–42 (2005).
63. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* **12,** R112 (2011).
64. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10,** 57–59 (2013).
65. Gillespie, J. W. *et al.* Evaluation of non-formalin tissue fixation for molecular profiling studies. *Am. J. Pathol.* **160,** 449–457 (2002).
66. Dawson, M. N., Raskoff, K. A. & Jacobs, D. K. Field preservation of marine invertebrate tissue for DNA analyses. *Mol. Mar. Biol. Biotechnol.* **7,** 145–152 (1998).
67. Seutin, G., White, B. N. & Boag, P. T. Preservation of avian blood and

tissue samples for DNA analyses. *Can. J. Zool.* **69,** 82–90 (1991).

68. Puritz, J. B. Fishing for Selection, but Only Catching Bias: Examining Library Effects in Double-Digest RAD Data in a Non-Model Marine Species. in *Plant and Animal Genome XXIII Conference* (Plant and Animal Genome, 2015).
69. Willing, E.-M., Hoffmann, M., Klein, J. D., Weigel, D. & Dreyer, C. Paired-end RAD-seq for de novo assembly and marker design without available reference. *Bioinformatics* **27,** 2187–2193 (2011).
70. Pegadaraju, V., Nipper, R., Hulke, B., Qi, L. & Schultz, Q. De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics* **14,** 556 (2013).
71. Fu, X. *et al.* RADtyping: an integrated package for accurate de novo codominant and dominant RAD genotyping in mapping populations. *PloS One* **8,** e79960 (2013).
72. Palaiokostas, C. *et al.* A novel sex-determining QTL in Nile tilapia (Oreochromis niloticus). *BMC Genomics* **16,** 171 (2015).
73. Manousaki, T. *et al.* Exploring a Nonmodel Teleost Genome Through RAD Sequencing—Linkage Mapping in Common Pandora, Pagellus erythrinus and Comparative Genomic Analysis. *G3 Genes Genomes Genet.* **6,** 509–519 (2016).
74. Linnaeus, C. Systema naturae, vol. 1. *Syst. Naturae Vol 1* (1758).
75. Merten, W., Appeldoorn, R. & Hammond, D. Movement dynamics of dolphinfish (Coryphaena hippurus) in the northeastern Caribbean Sea: Evidence of seasonal re-entry into domestic and international fisheries throughout the western central Atlantic. *Fish. Res.* **175,** 24–34 (2016).
76. Beardsley Jr, G. L. Age, growth, and reproduction of the dolphin, Coryphaena hippurus, in the Straits of Florida. *Copeia* 441–451 (1967).
77. Rose, C. D. & Hassler, W. W. Food habits and sex ratios of dolphin Coryphaena hippurus captured in the western Atlantic Ocean off Hatteras, North Carolina. *Trans. Am. Fish. Soc.* **103,** 94–100 (1974).
78. Gatt, M., Dimech, M. & Schembri, P. J. Age, Growth and Reproduction of Coryphaena hippurus (Linnaeus, 1758) in Maltese Waters, Central Mediterranean. *Mediterr. Mar. Sci.* **16,** 334–345 (2015).
79. Massutí, E. & Morales-Nin, B. Reproductive biology of dolphin-fish

(Coryphaena hippurus L.) off the island of Majorca (western Mediterranean). *Fish. Res.* **30,** 57–65 (1997).

80. Benseddik, A. B. *et al.* Détermination de l’âge et de la croissance de la coryphène, Coryphaena hippurus, des côtes tunisiennes par l’analyse des microstructures des otolithes. *Cybium* **35,** 173–180 (2011).
81. Oxenford, H. A. Biology of the dolphinfish (Coryphaena hippurus) in the western central Atlantic: a review. *Sci. Mar.* **63,** 303–315 (1999).
82. Pecoraro, C. *et al.* Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin tuna (Thunnus albacares). *Mar. Genomics* (2016).
83. Pardo, B. G. *et al.* Phylogenetic analysis of flatfish (Order Pleuronectiformes) based on mitochondrial 16s rDNA sequences. *Sci. Mar.* 531–543 (2005).
84. Wang, J. COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* **11,** 141–145 (2011).
85. Rodriguez-Ramilo, S. T., Toro, M. A., Wang, J. & Fernandez, J. Improving the inference of population genetic structure in the presence of related individuals. *Genet. Res.* **96,** e003 (2014).
86. Peakall, R. O. D. & Smouse, P. E. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* **6,** 288–295 (2006).
87. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
88. Raymond, M. & Rousset, F. GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *J. Hered.* **86,** 248–249 (1995).
89. Rousset, F. genepop’007: a complete re-implementation of the genepop software for Windows and Linux. *Mol. Ecol. Resour.* **8,** 103–106 (2008).
90. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10,** 564–567 (2010).
91. Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A. & Luikart, G. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier

method. *BMC Bioinformatics* **9,** 323 (2008).

92. Fischer, M. C., Foll, M., Excoffier, L. & Heckel, G. Enhanced AFLP genome scans detect local adaptation in high-altitude populations of a small rodent (Microtus arvalis). *Mol. Ecol.* **20,** 1450–1462 (2011).
93. Foll, M., Fischer, M. C., Heckel, G. & Excoffier, L. Estimating population structure from AFLP amplification intensity. *Mol. Ecol.* **19,** 4638–4647 (2010).
94. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180,** 977–993 (2008).
95. Lotterhos, K. E. & Whitlock, M. C. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* **23,** 2178–2192 (2014).
96. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
97. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).
98. Earl, D. A. & vanHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4,** 359–361 (2012).
99. Guillot, G., Estoup, A., Mortier, F. & Cosson, J. F. A spatial statistical model for landscape genetics. *Genetics* **170,** 1261–1280 (2005).
100. Puechmaille, S. J. The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Mol. Ecol. Resour.* (2016).
101. Devlin, R. H. & Nagahama, Y. Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture* **208,** 191–364 (2002).
102. Ueno, K. & Takai, A. Multiple sex chromosome system of X1X1X2X2/X1X2Y type in lutjanid fish, Lutjanus quinquelineatus (Perciformes). *Genetica* **132,** 35–41 (2008).
103. Fowler, B. L. & Buonaccorsi, V. P. Genomic characterization of sex-

identification markers in Sebastes carnatus and S. chrysomelas rockfishes. *Mol. Ecol.* (2016).

104. Johnson, N. A. & Lachance, J. The genetics of sex chromosomes: evolution and implications for hybrid incompatibility. *Ann. N. Y. Acad. Sci.* **1256,** E1–E22 (2012).

105. Galindo, H. M., Loher, T. & Hauser, L. Genetic sex identification and the potential evolution of sex determination in Pacific halibut (Hippoglossus stenolepis). *Mar. Biotechnol.* **13,** 1027–1037 (2011).

106. Alejo-Plata, C., Gómez, J. & Serrano-Guzmán, S. Variabilidad en la abundancia relativa, estructura por tallas y proporción por sexos del dorado Coryphaena hippurus (Pisces: Coryphaenidae) en Golfo de Tehuantepec, Mexico. *Rev. Biol. Trop.* **62,** 611–626 (2014).

107. Díaz-Jaimes, P. *et al.* Global phylogeography of the dolphinfish (Coryphaena hippurus): the influence of large effective population size and recent dispersal on the divergence of a marine pelagic cosmopolitan species. *Mol. Phylogenet. Evol.* **57,** 1209–1218 (2010).

108. Merten, W., Appeldoorn, R. & Hammond, D. Movements of dolphinfish (Coryphaena hippurus) along the US east coast as determined through mark and recapture data. *Fish. Res.* **151,** 114–121 (2014).

109. Milano, I. *et al.* Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius). *Mol. Ecol.* **23,** 118–135 (2014).

110. Gagnaire, P.-A. *et al.* Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. *Evol. Appl.* **8,** 769–786 (2015).

111. Cermeño, P. *et al.* Electronic tagging of Atlantic bluefin tuna (Thunnus Thynnus, L.) reveals habitat use and behaviors in the Mediterranean Sea. *PloS One* **10,** e0116638 (2015).

112. Riccioni, G. *et al.* Genetic structure of bluefin tuna in the Mediterranean Sea correlates with environmental variables. *PloS One* **8,** e80105 (2013).

113. Triantafyllidis, A. Aquaculture escapes: new DNA based monitoring analysis and application on sea bass and sea bream. in *CIESM Workshop Monographs* **32,** 67–71 (2007).

114. Arechavala-Lopez, P. *et al.* Differentiating the wild or farmed origin of

Mediterranean fish: a review of tools for sea bream and sea bass. *Rev. Aquac.* **5,** 137–157 (2013).

115. Alarcón, J. A., Magoulas, A., Georgakopoulos, T., Zouros, E. & Alvarez, M. C. Genetic comparison of wild and cultivated European populations of the gilthead sea bream (Sparus aurata). *Aquaculture* **230,** 65–80 (2004).
116. García-Celdrán, M. *et al.* Genetic assessment of three gilthead sea bream (Sparus aurata L.) populations along the Spanish coast and of three broodstocks managements. *Aquac. Int.* 1–12 (2016).
117. Dalvit, C., De Marchi, M. & Cassandro, M. Genetic traceability of livestock products: A review. *Meat Sci.* **77,** 437–449 (2007).
118. Hohenlohe, P. A., Catchen, J. & Cresko, W. A. in *Data Production and Analysis in Population Genomics* 235–260 (Springer, 2012).
119. Keller, I. *et al.* Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol. Ecol.* **22,** 2848–2863 (2013).
120. Wagner, C. E. *et al.* Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol. Ecol.* **22,** 787–798 (2013).
121. Hohenlohe, P. A., Amish, S. J., Catchen, J. M., Allendorf, F. W. & Luikart, G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Mol. Ecol. Resour.* **11,** 117–122 (2011).
122. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24,** 1403–1405 (2008).
123. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27,** 3070–3071 (2011).
124. Jombart, T. *A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 1.4-0*. (2013).
125. Besnier, F. & Glover, K. A. ParallelStructure: a R package to distribute parallel runs of the population genetics program STRUCTURE on multi-core computers. *PLoS One* **8,** e70651 (2013).
126. Hare, M. P. *et al.* Understanding and estimating effective population size for practical application in marine species management. *Conserv. Biol.* **25,** 438–449 (2011).

127. Peel, D., Ovenden, J. R. & Peel, S. L. NeEstimator: software for estimating effective population size, Version 1.3. *Qld. Gov. Dep. Prim. Ind. Fish.* (2004).

128. Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B Biol. Sci.* **263,** 1619–1626 (1996).

129. Coop, G., Witonsky, D., Di Rienzo, A. & Pritchard, J. K. Using environmental correlations to identify loci underlying local adaptation. *Genetics* **185,** 1411–1423 (2010).

130. Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M. & Holderegger, R. A practical guide to environmental association analysis in landscape genomics. *Mol. Ecol.* **24,** 4348–4370 (2015).

131. Muus, B. J. & Nielsen, J. G. *Sea fish*. (Scandinavian Fishing Year Book, 1999).

132. Arabaci, M. *et al.* A review on population characteristics of gilthead seabream (Sparus aurata). *J. Anim. Vet. Adv.* **9,** 976–981 (2010).

133. Brown, R. C., Tsalavouta, M., Terzoglou, V., Magoulas, A. & McAndrew, B. J. Additional microsatellites for Sparus aurata and cross-species amplification within the Sparidae family. *Mol. Ecol. Notes* **5,** 605–607 (2005).

134. Perez-Enriquez, R., Takagi, M. & Taniguchi, N. Genetic variability and pedigree tracing of a hatchery-reared stock of red sea bream (Pagrus major) used for stock enhancement, based on microsatellite DNA markers. *Aquaculture* **173,** 413–423 (1999).

135. Frost, L. A., Evans, B. S. & Jerry, D. R. Loss of genetic diversity due to hatchery culture practices in barramundi (Lates calcarifer). *Aquaculture* **261,** 1056–1064 (2006).

136. Glover, K. A., Skilbrei, O. T. & Skaala, Ø. Genetic assignment identifies farm of origin for Atlantic salmon Salmo salar escapees in a Norwegian fjord. *ICES J. Mar. Sci. J. Cons.* **65,** 912–920 (2008).

137. Glover, K. A. Forensic identification of fish farm escapees: the Norwegian experience. *Aquac. Environ. Interact.* **1,** 1–10 (2010).

138. Sanchez-Lamadrid, A. Stock enhancement of gilthead sea bream (Sparus aurata, L.): assessment of season, fish size and place of release in SW Spanish coast. *Aquaculture* **210,** 187–202 (2002).

139. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24,** 133–141 (2008).

140. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11,** 31–46 (2010).

141. Helyar, S. J. *et al.* Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Mol. Ecol. Resour.* **11,** 123–136 (2011).

142. Mullikin, J. C. *et al.* An SNP map of human chromosome 22. *Nature* **407,** 516–520 (2000).

143. Bruneaux, M. *et al.* Molecular evolutionary and population genomic analysis of the nine-spined stickleback using a modified restriction-site-associated DNA tag approach. *Mol. Ecol.* **22,** 565–582 (2013).

144. Du, Y. *et al.* Comprehensive evaluation of SNP identification with the Restriction Enzyme-based Reduced Representation Library (RRL) method. *BMC Genomics* **13,** 1 (2012).

145. Vera, M. *et al.* Development and validation of single nucleotide polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (Scophthalmus maximus) using high-throughput genotyping. *Int. J. Mol. Sci.* **14,** 5694–5711 (2013).

146. Pardo, B. G. *et al.* Expressed sequence tags (ESTs) from immune tissues of turbot (Scophthalmus maximus) challenged with pathogens. *BMC Vet. Res.* **4,** 1 (2008).

147. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156,** 439–447 (2000).

148. Glaubitz, J. C., Rhodes, O. E. & DeWoody, J. A. Prospects for inferring pairwise relationships with single nucleotide polymorphisms. *Mol. Ecol.* **12,** 1039–1047 (2003).

149. BERS, N. E. V. *et al.* Genome-wide SNP detection in the great tit Parus major using high throughput sequencing. *Mol. Ecol.* **19,** 89–99 (2010).

150. Sánchez, C. C. *et al.* Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *Bmc Genomics* **10,** 559 (2009).

151. Davey, J. W. *et al.* Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12,**

499–510 (2011).

152. Gompert, Z. *et al.* Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Mol. Ecol.* **19,** 2455–2473 (2010).

153. Vandepitte, K. *et al.* Conservation genetics of an endemic from the Mediterranean Basin: high genetic differentiation but no genetic diversity loss from the last populations of the Sicilian Grape Hyacinth Leopoldia gussonei. *Conserv. Genet.* **14,** 963–972 (2013).

154. Zhou, R., Cheng, H. & Tiersch, T. R. Differential genome duplication and fish diversity. *Rev. Fish Biol. Fish.* **11,** 331–337 (2001).

155. Ward, R. D., Elliott, N. G., Grewe, P. M. & Smolenski, A. J. Allozyme and mitochondrial DNA variation in yellowfin tuna (Thunnus albacares) from the Pacific Ocean. *Mar. Biol.* **118,** 531–539 (1994).

156. Alvarado Bremer, J. R., Naseri, I. & Ely, B. Orthodox and unorthodox phylogenetic relationships among tunas revealed by the nucleotide sequence analysis of the mitochondrial DNA control region. *J. Fish Biol.* **50,** 540–554 (1997).

157. Appleyard, S. A., Renwick, J. M. & Mather, P. B. Individual heterozygosity levels and relative growth performance in Oreochromis niloticus (L.) cultured under Fijian conditions. *Aquac. Res.* **32,** 287–296 (2001).

158. Dammannagoda, S. T., Hurwood, D. A. & Mather, P. B. Evidence for fine geographical scale heterogeneity in gene frequencies in yellowfin tuna (Thunnus albacares) from the north Indian Ocean around Sri Lanka. *Fish. Res.* **90,** 147–157 (2008).

159. Wu, G. C.-C., Chiang, H.-C., Chen, K.-S., Hsu, C.-C. & Yang, H.-Y. Population structure of albacore (Thunnus alalunga) in the Northwestern Pacific Ocean inferred from mitochondrial DNA. *Fish. Res.* **95,** 125–131 (2009).

160. Aires-da-Silva, A. & Maunder, M. N. Status of bigeye tuna in the eastern Pacific Ocean in 2010 and outlook for the future. *IATTC Stock Assess. Rep.* **13,** (2012).

161. Kunal, S. P., Kumar, G., Menezes, M. R. & Meena, R. M. Mitochondrial DNA analysis reveals three stocks of yellowfin tuna Thunnus albacares (Bonnaterre, 1788) in Indian waters. *Conserv. Genet.* **14,** 205–213 (2013).

162. Fonteles Filho, A. A. Recursos pesqueiros: biologia e dinâmica populacional. (1989).
163. CarvalhoÁFilho, A. Peixes: costa brasileira. (1999).
164. Hodgkinson-Clarke, F. M. *The Carite (Scomberomorus Brasiliensis) Fishery of South Trinidad: A Comparison of Catch Rates, Catch Composition, Use and Operation of the Monofilament and Multifilament Gillnets*. (Centre for Resource Management and Environmental Studies, University of the West Indies, Cave Hill Campus, 1990).
165. Fonteles-Filho, A. A. Sinopse de informacoes sobre a cavala, Scomberomorus cavalla (Cuvier) ea serra, Scomberomorus brasiliensis Collette, Russo and Zaval-Camin (Pisces: Scombridae), no estado do Ceara, Brasil. *Arq. Defic. Mar Fortaleza* **27,** 21–48 (1988).
166. Blanquer, A. Phylogéographie intraspécifique d'un poisson marin, le flet Platichthys flesus L.(Heterosomata): polymorphisme des marqueurs nucléaires et mitochondriaux. (Montpellier 2, 1990).
167. Milne, I. *et al.* Tablet—next generation sequence assembly visualization. *Bioinformatics* **26,** 401–402 (2010).
168. Buetow, K. H. *et al.* High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl. Acad. Sci.* **98,** 581–584 (2001).
169. Oeth, P., del Mistro, G., Marnellos, G., Shi, T. & van den Boom, D. in *Single Nucleotide Polymorphisms* 307–343 (Springer, 2009).
170. Louis, E. J. & Dempster, E. R. An exact test for Hardy-Weinberg and multiple alleles. *Biometrics* 805–811 (1987).
171. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38,** 1358–1370 (1984).
172. Vera, M. *et al.* Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, Scophthalmus maximus, database. *Aquaculture* **313,** 31–41 (2011).
173. Zhu, C., Cheng, L., Tong, J. & Yu, X. Development and characterization of new single nucleotide polymorphism markers from expressed sequence tags in common carp (Cyprinus carpio). *Int. J. Mol. Sci.* **13,** 7343–7353 (2012).

174. Cenadelli, S. *et al.* Identification of nuclear SNPs in gilthead seabream. *J. Fish Biol.* **70,** 399–405 (2007).

175. Albaina, A. *et al.* Single nucleotide polymorphism discovery in albacore and Atlantic bluefin tuna provides insights into worldwide population structure. *Anim. Genet.* **44,** 678–692 (2013).