Development and testing of a combined species SNP array for the European seabass (Dicentrarchus labrax) and gilthead seabream (Sparus aurata)

C. Peñaloza, T. Manousaki, R. Franch, A. Tsakogiannis, A. K. Sonesson, M. L. Aslam, F. Allal, L. Bargelloni, R. D. Houston, C. S. Tsigenopoulos

Genomics 113 (2021) 2096–2107

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Original Article

Development and testing of a combined species SNP array for the European seabass (*Dicentrarchus labrax*) and gilthead seabream (*Sparus aurata*)

C. Peñaloza [a,1], T. Manousaki [b,1], R. Franch [c], A. Tsakogiannis [b], A.K. Sonesson [d], M.L. Aslam [d], F. Allal [e], L. Bargelloni [c], R.D. Houston [a,*,1], C.S. Tsigenopoulos [b,*,1]

[a] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian EH25 9RG, UK
[b] Hellenic Centre for Marine Research, Thalassocosmos Gournes Pediados, 71500 Irakleio, Crete, Greece
[c] Padova University, Via Ugo Bassi, 58yB, I-35131 Padova, Italy
[d] Nofima, Norwegian Institute of Food, Fisheries and Aquaculture Research, PO Box 210, N-1432 Ås, Norway
[e] MARBEC, University of Montpellier, Ifremer, CNRS, IRD, 34250 Palavas-les-Flots, France

ARTICLE INFO

Keywords:
European seabass
gilthead seabream
SNP array
Aquaculture

ABSTRACT

SNP arrays are powerful tools for high-resolution studies of the genetic basis of complex traits, facilitating both selective breeding and population genomic research. The European seabass (*Dicentrarchus labrax*) and the gilthead seabream (*Sparus aurata*) are the two most important fish species for Mediterranean aquaculture. While selective breeding programmes increasingly underpin stock supply for this industry, genomic selection is not yet widespread. Genomic selection has major potential to expedite genetic gain, particularly for traits practically impossible to measure on selection candidates, such as disease resistance and fillet characteristics. The aim of our study was to design a combined-species 60 K SNP array for European seabass and gilthead seabream, and to test its performance on farmed and wild populations from numerous locations throughout the species range. To achieve this, high coverage Illumina whole-genome sequencing of pooled samples was performed for 24 populations of European seabass and 27 populations of gilthead seabream. This resulted in a database of ~20 million SNPs per species, which were then filtered to identify high-quality variants and create the final set for the development of the 'MedFish' SNP array. The array was then tested by genotyping a subset of the discovery populations, highlighting a high conversion rate to functioning polymorphic assays on the array (92% in seabass; 89% in seabream) and repeatability (99.4–99.7%). The platform interrogates ~30 K markers in each species, includes features such as SNPs previously shown to be associated with performance traits, and is enriched for SNPs predicted to have high functional effects on proteins. The array was demonstrated to be effective at detecting population structure across a wide range of fish populations from diverse geographical origins, and to examine the extent of haplotype sharing among Mediterranean farmed fish populations. In conclusion, the new MedFish array enables efficient and accurate high-throughput genotyping for genome-wide distributed SNPs for each fish species, and will facilitate stock management, population genomics approaches, and acceleration of selective breeding through genomic selection.

1. Introduction

Modern aquaculture selective breeding programmes are embracing the availability of genomic technologies to sustainably increase genetic gain. Genomic tools can also facilitate improvements to methods for forming base populations for breeding programmes by computing well-characterized genetic variability and relationships, which is important for many aquaculture species still in the process of domestication [1,2]. To achieve these goals in target species typically requires the generation of genome-wide genetic marker data (usually SNP markers) across large numbers of individuals. When paired with trait recording on the genotyped individuals, such datasets can be applied to examine the genetic architecture of production traits of interest, including detection of quantitative trait loci (QTL) using genome-wide association studies (GWAS). If the detected QTL are of sufficiently large effect, flanking markers can be utilized to select candidates with favourable alleles at

* Corresponding authors.
E-mail addresses: ross.houston@roslin.ed.ac.uk (R.D. Houston), tsigeno@hcmr.gr (C.S. Tsigenopoulos).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.ygeno.2021.04.038
Received 12 December 2020; Received in revised form 30 March 2021; Accepted 27 April 2021
Available online 30 April 2021

the QTL, also known as Marker-Assisted Selection (MAS). While MAS has been successfully applied for a small number of traits, such as resistance to infectious pancreatic necrosis in Atlantic salmon [3,4], most traits of interest for aquaculture are underpinned by a polygenic architecture [1,5,6]. For such traits, genome-wide SNP markers combined with phenotype data on a reference population can be used to estimate genomic breeding values for selection candidates [7]. Genomic selection is predicted to result in a notably higher selection accuracy and therefore genetic gain in aquaculture breeding programmes, as has also been demonstrated in early studies in several aquaculture species [1,8], including European seabass (*Dicentrarchus labrax*) [9] and gilthead seabream (*Sparus aurata*) [10,11].

The European seabass and the gilthead seabream are the two most important fish species in Mediterranean aquaculture. At the European level, they rank third and fourth, respectively, in terms of value after Atlantic salmon and rainbow trout [12]. Substantial genomic tools have been developed for both species, including the assembly and characterization of high-quality reference genomes [13, 14]. Medium or high-density SNP arrays have been developed for several other important finfish aquaculture species such as rainbow trout (*Oncorhynchus mykiss*) [15], Atlantic salmon (*Salmo salar*) [16,17], catfish (*Ictalurus furcatus and I. punctatus*) [18,19], common carp (*Cyprinus carpio*) [20], Arctic charr (*Salvelinus alpinus*) [21], and Nile tilapia (*Oreochromis niloticus*) [22–24], which have been used for studies into population structure, genetic diversity, signatures of domestication, the genetic architecture of traits of interest, and testing of genomic selection. A 57 K SNP array was also recently developed for European seabass [25] and has been applied to assess the genetic basis of resistance to viral nervous necrosis. However, this array is only available on request from the GeneSea consortium. Therefore, from both an aquaculture and population

Table 1
Summary of the European seabass and gilthead seabream populations sampled for sequencing and SNP discovery.

Species	Origin	Region	Country	Pool ID	N° individuals per pool	N° pools prepared
European seabass	Farmed	Mediterranean	France	Sba_farm_1	12	1
			Spain	Sba_farm_2	25	2
			Spain	Sba_farm_3	25	2
			Italy	Sba_farm_4	25	2
			Croatia	Sba_farm_5	25	2
			Croatia	Sba_farm_6	25	2
			Greece	Sba_farm_7	25	2
			Greece	Sba_farm_8	25	2
			Greece	Sba_farm_9	25	2
			Greece	Sba_farm_10	25	2
			Greece	Sba_farm_11	25	2
			Greece	Sba_farm_12	23	1
			Cyprus	Sba_farm_13	25	2
			Egypt	Sba_farm_14	15	1
	Wild	Mediterranean	France	Sba_wild_1	25	2
			Spain	Sba_wild_2	11	1
			Morocco	Sba_wild_3	25	2
			Italy	Sba_wild_4	25	2
			Croatia	Sba_wild_5	12	1
			Greece	Sba_wild_6	25	2
			Greece	Sba_wild_7	25	2
			Cyprus	Sba_wild_8	15	1
			Turkey	Sba_wild_9	25	2
			Turkey	Sba_wild_10	25	2
					Total N° pools	**42**
Gilthead seabream	Farmed	Mediterranean	France	Sbr_farm_1	25	2
			Spain	Sbr_farm_2	25	2
			Spain	Sbr_farm_3	25	2
			Italy	Sbr_farm_4	25	2
			Croatia	Sbr_farm_5	25	2
			Greece	Sbr_farm_6	14	1
			Greece	Sbr_farm_7	13	1
			Greece	Sbr_farm_8	25	2
			Greece	Sbr_farm_9	25	2
			Greece	Sbr_farm_10	25	2
			Israel	Sbr_farm_11	25	2
			Egypt	Sbr_farm_12	15	1
	Wild	Atlantic	France	Sbr_wild_1	25	2
			Spain	Sbr_wild_2	25	2
			Spain	Sbr_wild_3	25	2
		Mediterranean	Spain	Sbr_wild_4	25	2
			Spain	Sbr_wild_5	25	2
			Tunisia	Sbr_wild_6	25	2
			Italy	Sbr_wild_7	25	2
			Italy	Sbr_wild_8	25	2
			Greece	Sbr_wild_9	25	2
			Greece	Sbr_wild_10	25	2
			Greece	Sbr_wild_11	25	2
			Greece	Sbr_wild_12	25	2
			Greece	Sbr_wild_13	25	2
			Turkey	Sbr_wild_14	25	2
			Turkey	Sbr_wild_15	25	2
					Total N° pools	**51**

genetics perspective, there is a need for a publicly available high-throughput genotyping platform for European seabass and gilthead seabream.

Herein, an extensive and comprehensive SNP database was generated for European seabass and gilthead seabream across Europe by extensive sampling and pooled sequencing of ~25 populations per species from wild and aquaculture sites. From this SNP database, a subset of ~60 K SNPs was chosen based on several filtering criteria to give thorough coverage of each species' genome. The SNP array was created and tested on several of the discovery populations, including highlighting its potential utility for detecting population structure and excess haplotype sharing between farmed populations. This open-access tool will provide new opportunities to the scientific community and industry for genome-scale research and application to improve selective breeding in these two focal European aquaculture species.

2. Materials and methods

2.1. Samples for SNP discovery

A diverse range of farmed and wild populations of European seabass ($n = 24$) and gilthead seabream ($n = 27$) were collected for SNP discovery. A farmed population was defined as that composed of fish originating from the same commercial hatchery or established farm. A total of 538 European seabass individuals were sampled from 14 farmed and 10 wild populations distributed across the Mediterranean Sea, and a total of 642 gilthead seabream individuals were sampled from 12 farmed and 15 wild populations from the Mediterranean and the Atlantic (Table 1). Fin clips were collected from 11 to 30 individuals per population and stored in absolute ethanol until transportation to either the University of Edinburgh (UK), the Hellenic Centre for Marine Research (Greece) or the University of Padova (Italy) for DNA extraction.

2.2. DNA extraction and pooling for sequencing

High quality genomic DNA was isolated from each fin-clip using a salt-based extraction method [26]. The integrity of the DNA extractions was assessed by performing an agarose gel electrophoresis. DNA purity was evaluated by using a NanoDrop ND-1000 (Thermo Fisher Scientific) spectrophotometer. The extracted DNA was quantified in duplicate using the fluorescent-based Qubit® quantitation assay (Thermo Fischer Scientific, cat #Q32850). DNA stocks were diluted to 10–30 ng/ul and then combined at equimolar concentrations into pools of 11–25 individuals per population. The majority of populations had a sample size of 25, and for these populations DNA pools were prepared twice (technical replicates). For the remaining few populations with less individuals (6 and 3 populations in the European seabass and gilthead seabream, respectively), a single population pool was prepared (Table 1).

2.3. Library construction and sequencing

Two sequencing facilities provided the library preparation and sequencing services – the Norwegian Sequencing Centre (NSC) (Oslo, Norway) and Edinburgh Genomics (University of Edinburgh, UK). Both facilities followed the TruSeq® PCR-free library preparation protocol to generate sequencing libraries from the pooled genomic DNA samples. Almost all European seabass population pools were sequenced on a HiSeq 4000 instrument (2 × 150 bp) at NSC, whereas all gilthead seabream pools were sequenced on a HiSeq X Ten platform (2 × 150 bp) at Edinburgh Genomics.

2.4. Bioinformatics analysis for SNP discovery

The sequencing reads of the population pools – 42 and 51 pools for the European seabass and gilthead seabream, respectively (see Table 1) – were processed separately for each species using identical software and parameter values. These reads were filtered using the fastp software v 0.20.0 [27]. Reads with a minimum length of 80 bp for which <20% of their bases showed a BQ ≤ 20 were retained. Cleaned paired-end reads from each population pool were then aligned to either the European seabass [13] or the gilthead seabream [14] genome assemblies using BWA v 0.7.8 [28]. Only primary alignments to the relevant reference genome were retained for further analysis. PCR duplicates were removed from the alignment files using SAMtools v 1.6 [29]. Variants were called separately for each species across all population pools using Freebayes v 1.20 [30] with GNU Parallel [31]. Freebayes was set to call a variant if either (i) a minimum of 3 reads supporting the non-reference allele was observed, or (ii) the allele frequency in the pool was above 0.05, after excluding alignments with a MQ < 20. This SNP calling pipeline led to the discovery of ~17 and 34 million putative polymorphisms in the European seabass and gilthead seabream genomes, respectively.

This initial list of variants was then filtered using vcflib (https://github.com/vcflib/vcflib) to keep bi-allelic SNPs that (i) showed supporting reads on both strands, (ii) a sequence coverage ranging from 17× to 90× for the European seabass and from 25× to 100× for the gilthead seabream, (iii) at least two reads 'balanced' to each side of the variant site, (iv) >90% of the observed alternate and reference alleles supported by properly paired reads, and (v) the ratio of mapping qualities between reference and alternate allele was between 0.9 and 1.1. SNPs were retained only if they had no interfering polymorphic sites within <35 bp upstream and downstream of the variant. The purpose of this filter was to identify markers compatible with array design and eliminate SNPs that could fail the assay due to flanking polymorphisms interfering with probe annealing. The minor allele frequency (MAF) was estimated for all SNPs that were successfully genotyped in more than 18 population pools per species, after averaging the estimated MAF for population pools with technical replicates. To avoid spurious SNPs resulting from sequence differences between paralogues, only SNPs with a MAF between 0.05 and 0.45 were retained for further SNP selection. From this list of candidate markers (~1 million high-quality markers for each fish species), 35 bp probes were extracted downstream and upstream from each SNP. The 71-mer nucleotide sequences were then submitted to Thermo Fisher Scientific for further quality check and in silico probe scoring.

2.5. SNP selection

As a first filtering step, and as recommended by Thermo Fisher Scientific, the remaining SNPs were filtered to avoid *A/T* and *C/G* polymorphisms because they require twice the number of probes for genotyping compared to other types of SNP polymorphisms. The remaining SNPs were divided into selection tiers and were sequentially included in the MedFish platform based on the following hierarchy of importance.

First, SNPs were included as high priority markers based on evidence of their association with relevant production traits. For the European seabass, markers associated with mandibular prognathism [32], resistance to viral nervous necrosis [9], and sex [33] were included. For the gilthead seabream, the set of markers of this type comprised SNPs associated with production traits of high economic importance – i.e., fat content, weight, tag weight and length to width ratio [34] – and resistance to photobacteriosis [11]. Importantly, if the aforementioned SNPs were not identified through our pool-sequencing experiment, they were not included directly on the platform. Instead, the economically relevant marker was substituted by a proxy SNP that was chosen by screening the surrounding region for the closest high quality variant present in our dataset.

A second group of SNPs included in the MedFish SNP array is shared with other platforms that were developed in parallel by the GeneSea consortium [25]. The purpose of including a subset of markers from the

existing platforms was to facilitate backward compatibility and cross-study comparison, especially via the use of genotype imputation.

A third criterion for inclusion of SNPs on the MedFish platform was based on their predicted effect on protein-coding genes. SNPs on genes may affect protein function, for example, by causing truncated proteins. To potentially target variants with a potential functional effect, which may have a direct impact on relevant phenotypes, the list of high confidence variants identified in the European seabass and the gilthead seabream genomes were annotated with SNPEff v 4.3 [35]. For both species, SNPs that were predicted to have a HIGH functional effect on proteins were considered important and included as high priority markers in the array.

Fourthly, from the total number of ~1 million SNPs per fish species that were submitted as 71-mers to Thermo Fisher Scientific for in silico probe evaluation, only those that were categorized as either 'recommended' or 'neutral' became the pool from which array SNPs were selected. From the substantial SNP database generated in this study, markers were selected to achieve good coverage of the reference genomes of the European seabass and gilthead seabream following [33]. In brief, markers were selected along each fish chromosome at a variable density depending on the estimated local nucleotide diversity (π), as in European seabass [13] and other fish species [36] a positive correlation between nucleotide diversity and recombination rate has been observed. For SNPs that were mapped to the "UN" chromosome of the European seabass, the synthetic chromosome was split into contigs that had been previously concatenated by 100 consecutive Ns. The contigs were isolated and the SNPs located within them were remapped with the contigs starting position set to 1. The genomes of both fish species were divided into 70 Kb (for European seabass) or 85 Kb (for gilthead seabream) non-overlapping windows and local nucleotide diversity was estimated with VCFtools v 0.1.15 [37]. Genomic windows were categorized into one of the following classes depending on their estimated π value: $\pi \leq 0.001$ (Class 1), $0.001 < \pi \leq 0.002$ (Class 2), $0.002 < \pi \leq 0.003$ (Class 3), $0.003 < \pi \leq 0.004$ (Class 4) and $\pi > 0.004$ (Class 5). SNPs were chosen to cover all chromosomes of both fish species with a variable SNP density – ranging from 1 to 5 SNPs – depending on the diversity class assigned to each region. For the SNP selection process carried out within each type of diversity class window, two factors were considered as the main inclusion criteria: (i) the MAF for the SNPs in the window and (ii) the physical distance between markers. All discovered markers were divided into three different MAF categories (>0.3, 0.3–0.2 and 0.2–0.1). SNP markers within the MAF >0.3 category were prioritized across all five window classes such that at least 50% of the markers selected for each type of diversity class window came from the most informative SNP category (Table 2). Within each window, SNPs were selected successively from each MAF category by requiring a minimum inter-marker distance of 10,000 bp with any other previously chosen set of markers. To fill the remaining target of ~30 K SNPs per species, the physical distance between pairs of pre-selected SNPs was calculated, and intervals then sorted by length in decreasing order. SNP markers were then included sequentially (one SNP per interval) irrespective of its MAF.

Table 2
Summary of SNP selection approach. A variable number of SNPs was selected along chromosomes according to the local nucleotide diversity (π) estimates for non-overlapping genomic windows.

Genomic window diversity class	Range	Nº of SNPs sampled per window	Nº SNPs sampled per MAF category per window		
			>0.3	0.2–0.3	0.1–0.2
Class 1	$\pi \leq 0.001$	1	1	0	0
Class 2	$0.001 < \pi \leq 0.002$	2	2	0	0
Class 3	$0.002 < \pi \leq 0.003$	3	2	1	0
Class 4	$0.003 < \pi \leq 0.004$	4	2	2	0
Class 5	$\pi > 0.004$	5	2	2	1

A final list of ~70 K SNPs was sent to Thermo Fisher Scientific for the creation of the 60 K SNP array. This 384-format genotyping array was called the MedFish array, reflecting the two European Union funded consortium projects MedAID and PerformFish (see the 'Acknowledgements' section for details).

2.6. Testing of the MedFish array through population genomic analyses

2.6.1. Genotyping

A subset of 502 European seabass and 478 gilthead seabream fin clips from the same populations used for SNP discovery was sent to IdentiGEN (Ireland) for DNA extraction and genotyping with the MedFish SNP array (Table 3). To assess the repeatability and quantify the putative error rate of the platform, a single replicate sample (one per species) was genotyped twelve times across three different arrays. Only SNPs with genotype calls across all replicate samples were considered for evaluation (26,569 SNPs for European seabass and 25,547 SNPs for gilthead seabream). The proportion of SNPs at which the (replicate) individuals shared identical-by-state (IBS) alleles was calculated.

SNP quality control and genotype calling from the intensity files was performed using the Axiom Analysis Suite software v 2.0.035 at default parameter values for diploid species (call rate (CR) > 97; dish QC (DQC) >0.82). Because a significant fraction of the European seabass samples had a CR below the default value of 97 (201 individuals), the threshold was reduced to 93, allowing to recover genotypes for 460 individual samples.

2.6.2. Evaluation of SNP ascertainment bias

SNP markers genotyped with SNP arrays may suffer from a type of bias that is introduced during the array design process at the SNP selection stage. Markers to be included on a platform are typically selected (ascertained) from a larger pool of polymorphisms – discovered in a variable number of individuals from a variable number of populations – based on specific criteria (e.g. MAF threshold, equidistant spacing along the genome, etc.). Particularly when the size and number of SNP discovery populations are limited, this approach can lead to a final panel of markers in which rare SNPs, population-specific SNPs, and more recent SNPs are underrepresented [38]. In turn, this can lead to bias in several downstream population genetic analyses [39], particularly where these analyses are dependent on allele frequencies. To assess whether the SNP selection strategy followed for the development of the MedFish array affects population genetic inferences in European seabass and gilthead seabream, a commonly used summary statistic (F_{ST}) was estimated for pairs of populations (pairwise F_{ST}) based on the pooled whole-genome sequence data before ('non-ascertained' panel) and after SNP selection ('ascertained' panel). The 'non-ascertained' dataset (i.e. initial SNP discovery panel) reflects an unbiased representation of the 'true' genomic variation and allele frequency spectra present in the sampled populations of both fish species, as measured in this study. This dataset comprised ~1.1 million high quality SNPs called across 24 European seabass population pools and ~9 million SNPs genotyped across 27 gilthead seabream population pools. For populations for which two replicate pools were available (see Table 1), only one was selected for evaluation. The SNP QC filters applied to the variants called from the alignment files were the same as described in the 'Bioinformatics analysis for SNP discovery' section, except for the removal of markers with interfering SNPs in the flanking region. To generate the 'ascertained' datasets (one for each fish species), the SNP positions of the array markers were sub-sampled from the former ('non-ascertained') datasets using VCFtools v0.1.15 [37]. The four vcf files – 'non-ascertained' and 'ascertained' SNP panels for both European seabass and gilthead seabream – were imported to the R package poolfstat v1.2.0 [40] for pairwise F_{ST} estimation using the ANOVA method. For each species, the correlation between the pairwise F_{ST} matrices generated from the two

Table 3
Fish samples genotyped using the combined species MedFish SNP array.

Species	Origin	Population ID	Country	N° individuals typed	N° individuals passing QC
European seabass	Farmed	Sba_farm_1	France	12	8
		Sba_farm_2	Spain	25	18
		Sba_farm_3	Spain	25	16
		Sba_farm_4	Italy	24	24
		Sba_farm_6	Croatia	25	16
		Sba_farm_7	Greece	25	18
		Sba_farm_8	Greece	25	16
		Sba_farm_9	Greece	24	16
		Sba_farm_10	Greece	25	17
		Sba_farm_11	Greece	24	11
		Sba_farm_12	Greece	23	6
		Sba_farm_13	Cyprus	23	14
		Sba_farm_14	Egypt	14	12
	Wild	Sba_wild_Mediterranean		208	184
gilthead seabream	Farmed	Sbr_farm_1	France	24	19
		Sbr_farm_2	Spain	18	17
		Sbr_farm_3	Spain	25	19
		Sbr_farm_5	Croatia	19	19
		Sbr_farm_6	Greece	13	12
		Sbr_farm_7	Greece	13	13
		Sbr_farm_8	Greece	21	19
		Sbr_farm_9	Greece	24	22
		Sbr_farm_10	Greece	20	17
		Sbr_farm_11	Israel	13	9
		Sbr_farm_12	Egypt	15	14
	Wild	Sbr_wild_Atlantic		28	27
		Sbr_wild_Mediterranean		245	221

datasets – 'non-ascertained' and 'ascertained' SNP panels – was assessed with a Mantel test. For visualization, a principal coordinate analysis (PCoA) was performed based on the Euclidean distances of the pairwise F_{ST} matrices using the R package LabDSV [41].

2.6.3. Population structure

The combined species ~60 K MedFish array was tested by performing a metric multidimensional scaling (MDS) analysis of a wide range of Mediterranean (and a few Atlantic) European seabass and gilthead seabream farmed and wild populations typed with the platform. These individuals were part of the same set of samples used for the SNP discovery process from Pool-seq data (Table 3). A QC-filtered SNP dataset was created by applying the following filters in PLINK v2.0 [42]. Bi-allelic SNPs were retained for analysis if they had (i) a call rate > 0.95, (ii) MAF > 0.01, (iii) HWE test *p*-value ≥1e-4 (estimated separately for each population) and (iv) no pairs with a squared LD correlation (r^2) > 0.2 occurred within a 100 Kb window. For duplicated or related individuals with a kinship coefficient (KING-rob) > 0.177 (first-degree relatives or closer), only one member of a pair was retained for further analysis. All individuals to be evaluated required to have <10% missing genotypes. The relationship among individuals and populations was visualized using a MDS analysis based on the genome-wide IBS pairwise distances as implemented in PLINK.

2.6.4. Analysis of haplotype sharing among farms

To assess the ability of the SNP array to identify historical connections between farmed populations, a haplotype sharing analysis was performed on the farmed population samples (13 European seabass farms; 11 gilthead seabream farms). A SNP dataset in which all individual and SNP QC filters had been applied (see 'Population structure' section), except the removal of markers based on linkage disequilibrium (measured as r^2) was used for the analysis. Markers that were not located on the assembled chromosomes of the reference genome assemblies (i.e. those located on unplaced scaffolds) were removed from the dataset. Haplotypes were inferred for each individual using the software fastPHASE v1.4.8 [43]. All individuals were phased together in a single analysis, taking into consideration their population labels during the model fitting procedure. For both fish species, the number of random starts of the EM algorithm (T) was set to 20, the number of iterations (C) was set to 35, and the number of haplotype clusters (K) to 8.

The reference genomes of both species were divided into 1 Mb non-overlapping windows using BEDTools v2.25 [44]. SNP-based haplotype variants were defined for each window. The last window of each chromosome was excluded from the analysis. Since the number of haplotypes can be influenced by sample size, the same number of individuals were randomly chosen from each farmed population (6 individuals for the European seabass and 9 for the gilthead seabream). For each individual within a farm, the two haplotypes at any given locus were used to screen the whole dataset for an exact match. All matches with other individuals from a different farm were recorded. The totals were then summed across all individuals that belonged to the same farm, and the proportion of shared haplotypes across farms calculated.

To assess whether a pair of farms had excess haplotype sharing, 1000 permutations were performed. For each permutation, all individuals from a fish species were randomly assigned to an arbitrary farm.

2.7. Ethics statement

The fish fin clip collected in this study were obtained from commercial samples or specific sampling efforts managed and sampled in accordance with the European directive 2010/63/UE on the protection of animals used for scientific purposes.

3. Results

3.1. SNP array development

The pooled DNA sequencing of 24 European seabass and 27 gilthead seabream populations and their replicate pools (see Table 1) produced 8205 and 23,784 million paired-end reads, respectively. The alignment of the post-quality filtered reads of the population pools against each species reference genome resulted in the discovery of ~17 million polymorphisms in the European seabass and ~34 million putative polymorphisms in the gilthead seabream genomes (including both SNPs

and indels). The generated sequence led to an average coverage at SNP variant sites of 36× in the European seabass and 63× in the gilthead seabream (in both cases, averaged across all pools). After applying the QC filters on the variant call set (see 'Bioinformatics analysis for SNP discovery' section), a pool of 1,056,218 and 1,015,264 high confidence SNPs in the European seabass and the gilthead seabream, respectively, remained for SNP selection. The QC filter that removed the largest amount of data was the restriction to retaining SNPs without other polymorphisms in close proximity (within 35 bp on either side). This filter alone removed 88% and 96% of the SNPs discovered in the European seabass and gilthead seabream, respectively.

Following the submission and evaluation of these filtered SNPs by Thermo Fisher Scientific, high-value markers – i.e. those associated with production traits, with an effect on proteins or shared with another array – were tiled on the MedFish array with two sets of probes. The remaining set of SNP from the platform (~24 K in European seabass and ~ 26 K in gilthead seabream) were tiled with a single probe and were obtained from a sampling along chromosomes based on the strategy of selecting SNPs in proportion to the putative local recombination rate (measured as π) of the genomic region. Notably, and in comparison to the European seabass, a particularly high number of polymorphisms were initially discovered along the chromosomes of the gilthead seabream, particularly towards the terminal ends of the chromosome-level scaffolds. The most likely cause was that in this species the higher average sequencing depth of 63× (compared to 36× in the European seabass population pools) enabled the discovery of variants segregating at a lower frequency. Consequently, when the QC filter that removed SNPs with interfering markers in close proximity was applied to the gilthead seabream dataset, a substantial number of markers were filtered out from regions of the genome exhibiting higher levels of genetic polymorphisms. This led to fewer SNPs left to choose from for assay design in regions of the gilthead seabream genome that showed putative higher recombination (e.g. chromosome ends) and for which a higher number of SNPs had to be sampled based on our SNP selection strategy. Therefore, the SNP selection strategy led to a more even sampling of SNPs along the gilthead seabream genome. While in European seabass, the array SNPs followed the expected pattern of the SNP selection strategy, with more markers being assayed towards the terminal ends of the chromosome-level scaffolds (Fig. 1).

The final MedFish SNP array was designed to interrogate 29,888 SNPs in the European seabass genome and 29,807 SNPs in the gilthead seabream genome. Among these markers, 4560 SNPs (15%) in the European seabass and 3208 SNPs (11%) in the gilthead seabream are shared with other SNP arrays that were being developed at the time of this study [25]. A significant fraction of the SNPs on the platform are located in genes (46% seabass; 32% seabream), among which 107 and 179 SNPs, respectively, were predicted in silico to have high functional effects on proteins. For the SNPs included on the array, the physical distance between consecutive markers was similar for both species and averaged 20 Kb in the European seabass and 18 Kb in the gilthead seabream (density plot of inter-marker distances shown in Fig. S1). The largest gaps between markers (200–300 Kb) represented a small fraction of the platform and comprised five regions on chromosomes 1, 4, 9, 13 and 16 of the gilthead seabream, which summed up to ~1.4% of the genome [14]. Detailed examination revealed that these regions lacked suitable markers matching our SNP selection criteria. No large regions in the European seabass genome were devoid of assays, with the highest inter-marker distance being ~120 Kb, which comprised two pairs of markers on chromosomes 5 and 20, and represented <0.05% of the European seabass genome [13].

Two metrics were used to assess the performance of the assays on the array: (i) conversion rate and (ii) platform error rate. Here the conversion rate is defined to be the fraction of probes that yielded strong signals with high-quality clusters discerning different genotypes. The conversion rate of the European seabass fraction of the array was 91.9%, whereas the gilthead seabream assays on the array had a conversion rate of 88.8% (Table 4). In terms of the informativeness of the markers on the platform, for 99.8% of the validated loci in the European seabass the MAF was >5%. In the gilthead seabream, 98.7% of the markers had a MAF > 5%. The process of calculating the platform error rate involved genotyping two samples (one per species) twelve times each. For European seabass, one replicate sample failed to generate a CEL file; consequently, eleven samples remained for evaluation. The repeatability of the assays, after excluding loci with at least one missing value across replicates, was 99.4% for the European seabass and 99.8% for the gilthead seabream. Taken together, these metrics support the high quality and reliability of the genotype data generated by the MedFish SNP array.

To evaluate whether the SNP selection strategy followed for the design of the MedFish array could bias F_{ST} estimates, SNPs common to the array (~30 K for each species) were selected from the whole-genome sequence data of the population pools (consisting of 11–25 individuals each) to mimic array-derived genotype data. Pairwise F_{ST} were calculated from this 'ascertained' dataset and then compared to those estimates obtained for the same population pools but by including all sites discovered through the re-sequencing experiment in the analysis (~1.1 million SNPs in the European seabass; ~9 million SNPs in the gilthead seabream) ('non-ascertained' dataset). Overall, a high correlation between the pairwise F_{ST} matrices calculated from the 'non-ascertained' and 'ascertained' SNP panels was observed, with an $r = 0.97$ for both European seabass and gilthead seabream (Fig. 2A and B). Although for

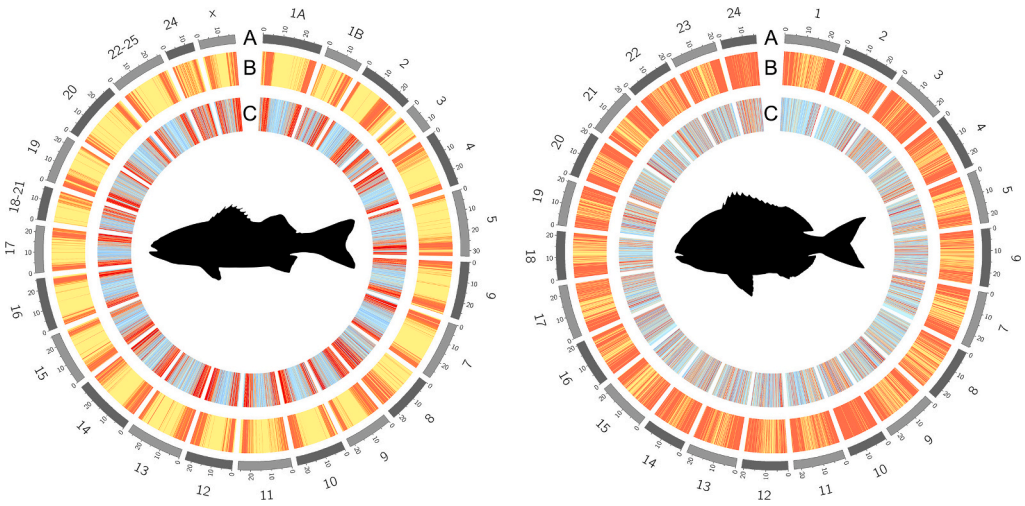

Fig. 1. Schematic representation of the distribution of array markers in the European seabass (left) and gilthead seabream (right) genomes after following a SNP selection strategy based on local nucleotide diversity. (A) Chromosome number. (B) Levels of diversity (π) estimated over 70 Kb and 85 Kb windows in the European seabass and gilthead seabream, respectively. Red bars represent regions with high nucleotide diversity. (C) Genome-wide distribution of markers on the combined-species SNP chip. Light blue bars represent windows for which 1–3 SNPs were selected. Red bars represent windows for which more than four SNPs were selected.

Table 4
Number of SNPs for each species of each Axiom quality class.

Conversion type[a]	N° European seabass (%)	N° gilthead seabream (%)
Polymorphic high resolution	26,466 (88.55%)	26,369 (88.47%)
No minor homozygote	993 (3.32%)	75 (0.25%)
Total high quality polymorphic	**27,459 (91.87%)**	**26,444 (88.72%)**
Monomorphic high resolution	26 (0.09%)	36 (0.12%)
Off-target-variant (OTV)	50 (0.17%)	78 (0.26%)
Call rate below threshold (97%)	889 (2.97%)	1292 (4.33%)
Other	1464 (4.90%)	1957 (6.57%)
Total SNPs on the array	**29,888 (100%)**	**29,807 (100%)**

The categories are based on cluster properties and QC metrics.
[a]The Conversion type follows Thermo Fisher's terminology:
PolyHighResolution = Class with the highest quality probes. SNP is polymorphic and the presence of both the major and minor homozygous clusters is observed.
NoMinorHom = similar to a PolyHighResolution, but no evidence of individuals with minor homozygous genotypes, presumably due to a low genotype frequency.
MonoHighResolution = SNP can reliably be scored as monomorphic.
Off-target variant (OTV) = SNPs where additional (i.e. more than three) clusters are observed, making genotype calling ambiguous.
CallRateBelowThreshold = SNP with the expected number of clusters (usually 3, one for each possible genotype), but where the proportion of samples scored at the SNP falls below a user-defined threshold.
Other = SNPs that do not fall in any of the above categories.

both fish species most of the pairwise F_{ST} values were slightly higher when the estimations were based on the 'ascertained' dataset (Fig. S2), both 'ascertained' and 'non-ascertained' datasets yielded similar population clustering patterns, as observed across the first and second dimension of a PCoA (Fig. 2C and D).

3.2. Population structure

To test the SNP array and to gain a general overview on the population structure within each species, a MDS analysis was performed. The two first dimensions explained 26% and 14% of the total variance for European seabass and gilthead seabream, respectively (Fig. 3).

In European seabass, most of the sampled farmed populations form a loose cluster along D1, which explains 20% of the variance. No geographical cline is observed as farms from the West (France, Spain), centre (Italy, Croatia, Greece) and East (Cyprus, Egypt) of the Mediterranean cluster at least partially in this dimension. On the other hand, three distinctive clusters are recognized for the wild European seabass populations, with a few exceptions corresponding to individuals clustering near farmed populations instead. D2 explains 6% of the total variation and mainly separates (i) a single well-defined wild population cluster, (ii) a large group containing most of the farmed and wild seabass populations, and (iii) a group of individuals that belong to a farm sampled from the centre of the Mediterranean (farm N° 10 sampled from a Greek hatchery) (Fig. 3A).

For the gilthead seabream analysis, the sampled farmed populations appear to form a continuum along D1 rather than discrete units. Although the majority of gilthead seabream wild populations were

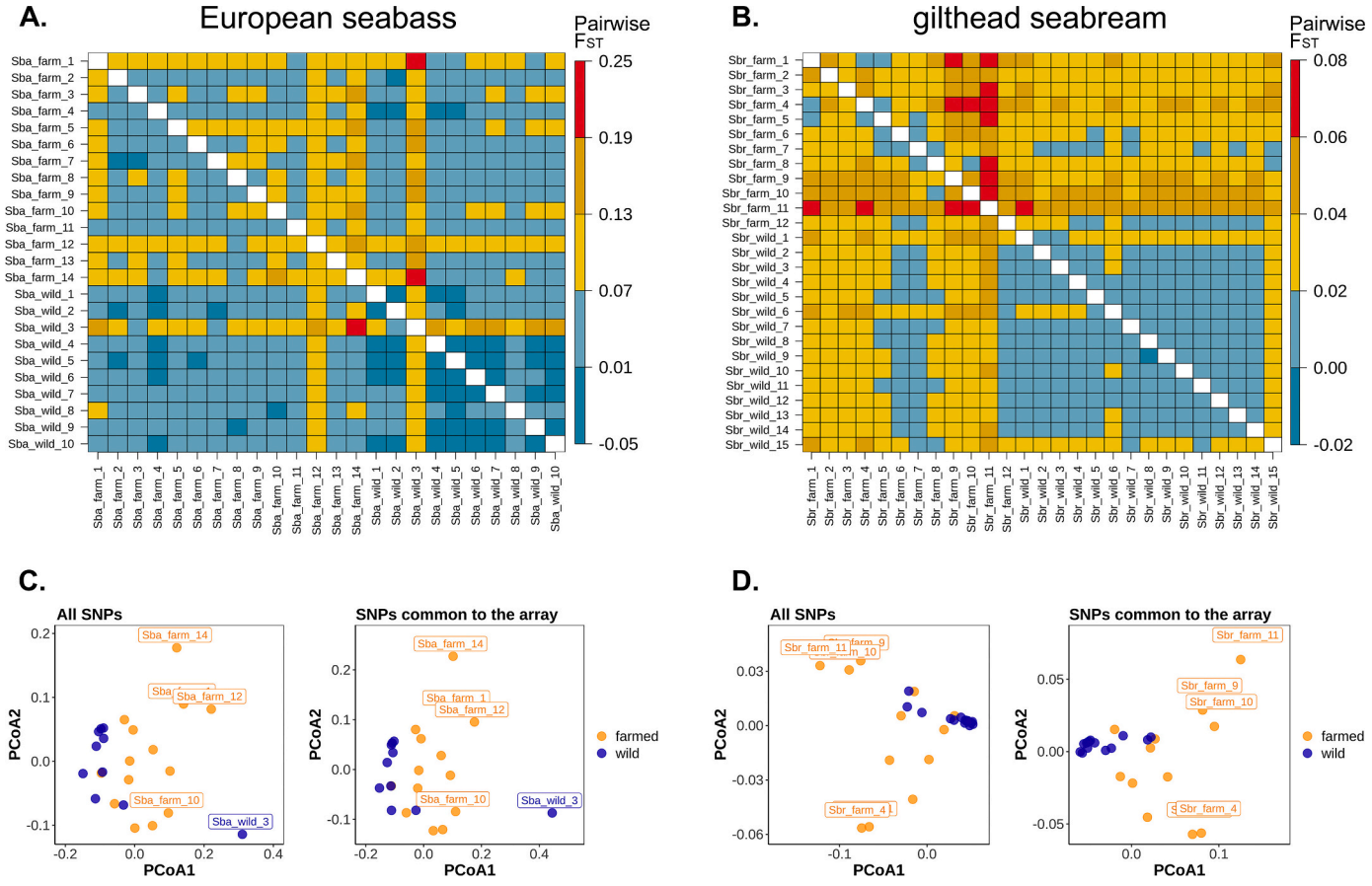

Fig. 2. Evaluation of SNP ascertainment on pairwise F_{ST} estimates in European seabass and gilthead seabream pooled population samples. Pairwise F_{ST} values obtained for (A) European seabass and (B) gilthead seabream populations based on a 'non-ascertained' marker panel (below the diagonal) and an 'ascertained' panel containing the array SNPs (above the diagonal). Principle coordinate analysis of pairwise F_{ST} among (C) European seabass and (D) gilthead seabream populations, and comparison between results obtained from a 'non-ascertained' (left panel) vs 'ascertained' (right panel) set of SNPs.

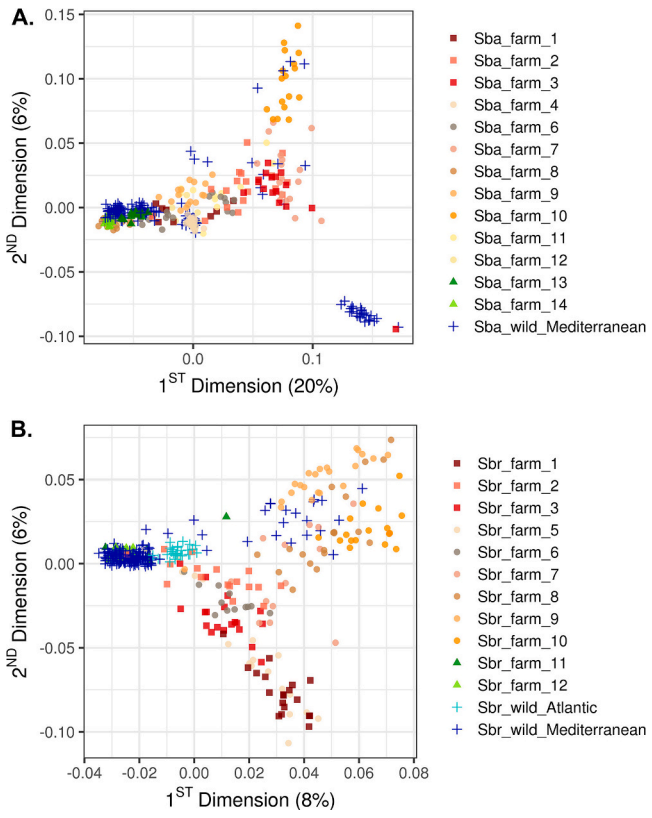

Fig. 3. MDS analysis performed on individuals from farmed and wild (A) European seabass and (B) gilthead seabream populations. The different point symbols separate samples by origin in (i) farms from the West of the Mediterranean (■), (ii) farms from the centre of the Mediterranean (●), and (iii) farms from the East of the Mediterranean (▲), from (iv) wild populations (+).

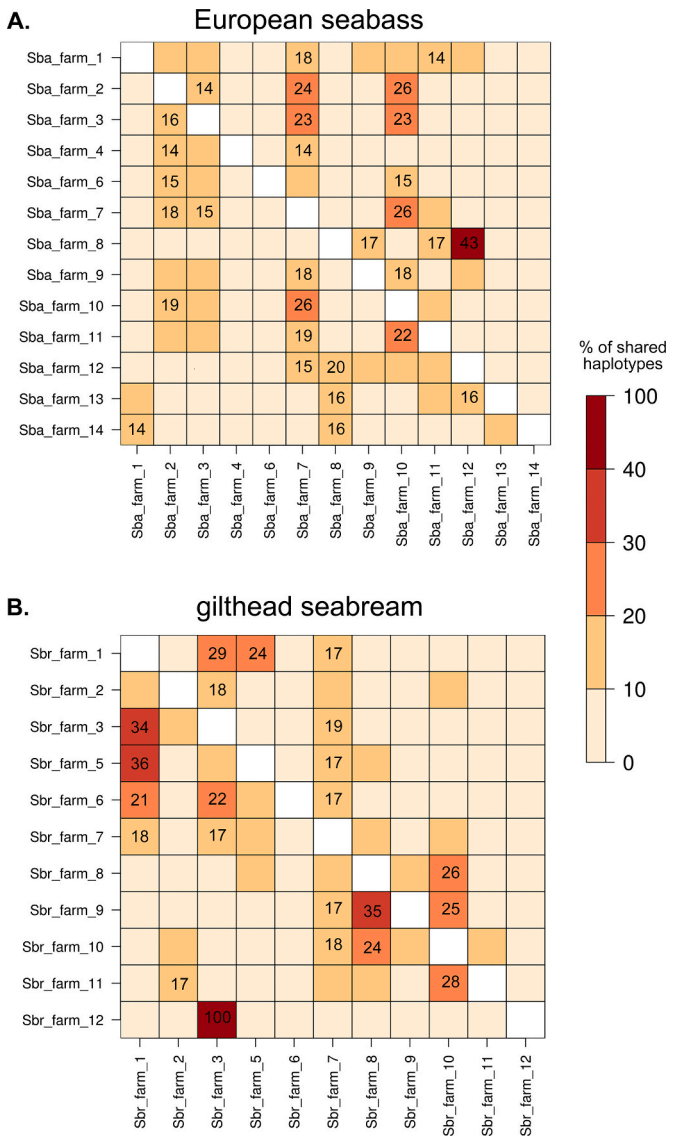

Fig. 4. Heatmaps indicating the percentage of shared haplotypes between pairs of (A) European seabass and (B) gilthead seabream farmed populations. The colorbar on the right indicates the percentage of shared haplotypes. Entries are shown for pairs of farms with a statistically significant excess of shared haplotypes (p-value <0.05).

sampled from the Mediterranean Sea, a few populations from the Atlantic coast of France and Spain were included in the analysis. Individuals sampled from a wide range of wild populations tend to group by origin into either a Mediterranean or Atlantic cluster on one extreme of the D1 axis. D2 accounts for 6% of the variance and distinguishes two groups of overlapping farmed populations that partially coincide with their macro-region of origin. The first group is composed only of farms located in the centre of the Mediterranean (i.e. Greece). A few wild gilthead seabream individuals co-localize with this group of farmed samples. The second group is comprised of a mixture of all three farms sampled from the West of the Mediterranean (i.e. from either France or Spain) and a few populations sampled from farms located in the centre of the Mediterranean, namely farms N⁰ 5 (from Croatia) and N⁰ 6 (from Greece) (Fig. 3B).

3.3. Haplotype sharing analysis among farms

After applying QC filters, a total of 21,822 SNPs in European seabass and 24,765 SNPs in the gilthead seabream remained for the assessment of haplotype sharing between pairs of Mediterranean fish farms.

The pairwise comparison among European seabass farmed population samples revealed that all populations showed an excess of haplotype sharing with at least one other Mediterranean farm (Fig. 4A). The highest percentage of haplotype sharing (43%) was found between two Greek seabass farms (farm N⁰ 8 vs farm N⁰ 12). The reverse relationship between these two farms (i.e. farm N⁰ 12 vs farm N⁰ 8) is also significant but is ranked 9th (20%) in terms of haplotype-sharing between all pairs of farmed populations. This difference in reciprocal comparisons is explained by differences in the total numbers of shared haplotypes identified within each farm (File S1). Haplotypes from individuals of a European seabass farm located in Greece (farm N⁰ 7) were present at significant frequencies in all farms sampled from the West of the Mediterranean (i.e. farms of French or Spanish origin), and most of the seabass farms sampled from the centre of the Mediterranean (i.e. either from Italy, Croatia or Greece).

In common with the results observed for European seabass, all gilthead seabream farms evaluated show an excessive sharing of haplotypes with at least one other Mediterranean farm (Fig. 4B). In gilthead seabream, a clear geographical break (with farm N⁰ 7 forming the boundary) separates the farms from the West and centre of the Mediterranean in two groups of farms between which reduced haplotype sharing is observed. One group includes five farms – farms N⁰ 1, 2, 3, 5, and 6 – from diverse geographical origins (i.e. France, Spain, Croatia and Greece). The second group comprises commercial farms exclusively based in Greece – farms N⁰ 8, 9 and 10. The farm in the boundary of both groups – farm N⁰ 7 – had haplotypes that were present at significant levels in farms belonging to both aforementioned groups. A seabream farm sampled from the East of the Mediterranean (farm N⁰ 12) had the

lowest total number of shared haplotypes among all commercial farms evaluated from both fish species. Most haplotypes identified in farm Nº 12 were unique and specific to the farm, which showed complete absence of shared variants with all but one Mediterranean farm (i.e. farm Nº 3), with which it shared only two haplotypes in total (File S1).

4. Discussion

4.1. Properties of the combined species MedFish SNP array

A publicly available, combined species SNP chip that assays ~30 K SNPs throughout the genome of two prominent Mediterranean fish species - the European seabass and the gilthead seabream – was developed. To evaluate the performance of the MedFish SNP array two metrics were analyzed: conversion rate and platform error rate. The conversion rate is a measure of the number of SNPs successfully assayed by a technology and reflects the quality of both the chosen SNPs and the technology used to score them [45]. Conversion rates were high for the SNP array regardless of the fish species. The assay conversion rate of the European seabass part of the array was 91.9%, while for the gilthead seabream part it was 88.8%. These values are slightly lower than assays designed for terrestrial livestock species (e.g. 92.6% in cattle and 97.5% in pigs), however, they are comparatively higher than similar assays developed for aquatic organisms (e.g. 72.5% in oysters and 86.1% in catfish) [18,46–48], and in the upper range of high performing arrays for fish species [15,25]. As a second metric to assess performance, the platform error rate was calculated based on the genotype concordance of repeated assays on the same individual. By this metric, the MedFish platform shows high genotype accuracy, with a repeatability ranging from 99.4% to 99.7%. These levels of accuracy are comparable to those achieved with Illumina GoldenGate assays in humans (99.7%) or Affymetrix SNP chips in trout (99.4%) and pig (100%) [15,46,49]. Compared to other SNP arrays developed for aquaculture species, the MedFish platform has among the best performance in terms of genotype accuracy and repeatability. To assess potential ascertainment bias arising in chip-derived genotype data, pairwise F_{ST} was estimated for the same populations using a 'non-ascertained' and an 'ascretained' SNP panel for comparison. The patterns of pairwise F_{ST} among populations were highly concordant when results based on the markers on the MedFish platform were compared to the full SNP discovery panel (Fig. 2), which supports the utility of the platform for population genetic studies. However, a slight upward bias of F_{ST} values was observed in the dataset that mimics the array-derived markers and is likely caused by prioritizing common high-frequency markers, at the expense of SNPs with rarer alleles, in the SNP selection process. While the differences between F_{ST} estimates between both datasets ('ascertained' versus 'non-ascertained') were minor (Fig. S2), additional fish populations that were not part of the ascertainment panel should be evaluated to accurately assess the impact of chip-based genotyping data on population genetics estimators for these species [50]. Until recently, high-throughput genotyping analysis was typically only achievable in these fish species by means of reduced-representation sequencing approaches [9,11,34]. Although these genotyping by sequencing techniques can be particularly cost-effective for small numbers of samples and should also display minimal ascertainment bias compared to SNP arrays [51] (however, see [52]), they can suffer from inconsistent marker recovery across experiments and a comparatively lower robustness to low quality input DNA [53]. Hence, the development of this combined species SNP array represents a powerful alternative for high-throughput genotyping in European seabass and gilthead seabream, which may facilitate molecular breeding applications, genetic stock identification and population and evolutionary studies in these emblematic fish species. Moreover, the fact that the two species are represented on the same platform increases the overall volume of arrays that can be purchased, which should reduce the cost of the array per sample due to economy of scale. While it should be noted that the number of arrays required to reach this lower per-sample cost is high (multiples of 384 in the case of the MedFish platform), this improved cost-effectiveness is likely to be important for the uptake of the platform by aquaculture breeding and production companies for the routine application of genomic selection.

As part of the SNP array design, ~25 farmed and wild populations (>500 individuals) per species were sequenced and screened for informative markers. By following a DNA pooling approach, reliable genome-wide allele frequency information was obtained for several fish populations at a fraction of the effort of individual sequencing. Given that the majority of the samples genotyped with the SNP array were also part of the SNP discovery process, metrics such as number and mean MAF of polymorphic markers reflected the performance of the SNP selection strategy. Despite the relatively small DNA pools (11–25 individuals), we were able to reliably identify and select informative markers for inclusion in our SNP array. The number of informative markers (MAF > 0) was high for both fish species. For the European seabass, 23,900 SNPs (99.8%) of the validated markers were polymorphic, whereas for the gilthead seabream, this type of markers comprised 26,017 SNPs (99.3%) of the data. The number of polymorphic markers was remarkably similar in wild and farmed populations of both species (90–99% across populations), demonstrating the efficacy of the SNP selection strategy for recovering highly informative markers in Pool-seq data sequenced at a high to moderate coverage across a wide range of different populations. When evaluating the MAF across European seabass and gilthead seabream populations, the allele frequency profiles were similar within species and did not vary significantly by origin (either wild or farmed) (Fig. S3). The mean MAF across the European seabass (0.33) and gilthead seabream (0.31) populations was higher than that reported when validating SNP arrays in Nile tilapia (0.29) and rainbow trout (0.25) [15,22]. However, the high average MAF observed in this study is most likely influenced by the fact that most of the discovery populations were also used for the validation of the SNP chip. Nonetheless, it should be noted that the discovery population samples cover a large portion of the distribution range in the wild and include the majority of commercial hatcheries for the two species.

A significant obstacle to the uptake of high-throughput genotyping technologies by the industry is the risk that a low fraction of a pre-built platform yields useful information. Indeed ascertainment bias is a common issue for genotyping arrays and can be caused when designing platforms based on a reduced number of individuals [50]. Due to the fact that the MedFish 60 K array was developed based on the screening of genetic data derived from extensive sampling of dozens of Mediterranean fish populations and hundreds of fish of each species, it is tailored to maximize the retrieval of genetic information and provide an increased resolution for the analysis of farmed or wild stocks from this region.

4.2. Population structure and haplotype sharing analysis

To test the MedFish SNP array, the genotyping data obtained from typing a diverse range of wild and farmed European seabass and gilthead seabream fish were used to perform a MDS and haplotype sharing analysis.

Regarding the European seabass populations, the two first dimensions explained 26% of the genotypic variation. Interestingly, the wild Mediterranean populations span a continuum across the range of D1, but has a rather smaller dispersal across the D2 range. However, this continuum in D1 has gaps, and the wild populations seem to be divided into two clearly differentiated clusters located at each end of this major axis. These two clusters are represented by (i) distinct groups of wild individuals showing low levels of genetic differentiation, consistent with a sub-division (albeit weak) of the European seabass Mediterranean lineage [54]; and (ii) a wild population sampled from Morocco, which based on previous findings [54] is expected to cluster with Atlantic rather than Mediterranean populations. All farmed European seabass populations have a more limited distribution across the range of the two

dimensions compared to the wild populations, forming clear clusters although less dense than the wild populations, probably due to their smaller sample sizes (Table 3). Most farm populations fall within the range of the wild populations, with overlap among each other. A few of the wild individuals fall clearly within the range of farm populations, which could indicate the presence of escapees from fish farms in the set of wild samples, reflecting a well-known phenomenon occurring in the Mediterranean [55–57]. Only a single farmed fish population seems to cluster separately from the other farm samples (Fig. 3A; farm Nº 10), suggesting either founder effects, stronger artificial selection, higher number of generations of selection, or any combination thereof. European seabass farms of different geographical origin tend to cluster together in the plot. For instance, farms Nº 2 and 3 (from Spain) group with farm Nº 7 (from Greece). This observation is consistent with the haplotype sharing analysis, as a significant number of 1 Mb SNP-based haplotype variants were jointly present in these farms. A high frequency of shared haplotypes between pairs of populations provides information about their historical relationship, reflecting either a common ancestry and/or gene flow between populations. In the context of aquaculture farming, a high frequency of shared haplotypes between farms might indicate (i) animal transfer between farms or (ii) the recent establishment of these farmed populations from the same wild source (i.e. recent population divergence). Since the MDS analysis revealed that wild populations of European seabass form tight and distinctive clusters, it is likely that pairs of European seabass farms sharing a high frequency of haplotype variants are derived from human-mediated translocations of eggs or juveniles.

For the gilthead seabream populations, MDS results explained much less of the observed genetic variance (only 14% covered by both D1 and D2 summed up), showing a less clear structure than European seabass for most of the populations sampled in this study. In this case, wild populations were sampled from two regions, the Mediterranean and Atlantic. Wild individuals segregate into two closely bound Mediterranean and Atlantic clusters, which is consistent with previous findings indicating a low genetic differentiation between regions [58,59]. Similar to European seabass, a few wild individuals are found scattered throughout farmed populations, likely representing escapees from local fish farms. Farmed gilthead seabream populations seem to be much more differentiated compared to their wild counterparts, with two broader clusters forming a gradient of overlapping farmed populations. The first group is composed only of farms located in Greece. The second cluster groups a mixture of all three farms sampled from the West of the Mediterranean (either France or Spain) and a few farms from Croatia and Greece (Fig. 3B). This pattern may reflect artificial selection and/or different degrees of admixture between farms. The haplotype sharing analysis mirrors this finding and reinforces the idea that most seabream farms from the Mediterranean separate in two clusters, between which a reduced recent contact is observed. However, while the results for both the European seabass and gilthead seabream highlight the potential utility of the SNP array for detecting and studying population structure, more extensive studies are required to further assess these phenomena in the two species.

5. Conclusions

A medium density SNP array suitable for genotyping both the European seabass and the gilthead seabream was developed. The MedFish SNP array is a robust resource with a high assay performance, as demonstrated by its high conversion rate (92% in the European seabass; 89% in the gilthead seabream) and repeatability (99.4% in the European seabass; 99.8% in the gilthead seabream). The platform interrogates ~30 K markers in each fish species and includes features such as SNPs previously shown to be associated with performance traits and enrichment for SNPs predicted to have high functional effects on proteins. The SNP array was highly informative when tested on the majority of the discovery population samples and was further validated by performing a population structure and haplotype sharing analysis across a wide range of fish populations from diverse geographical backgrounds. This recently developed platform will allow the efficient and accurate high-throughput genotyping of ~30 K SNPs across the genomes of each fish species, facilitating population genomic research and the application of genomic selection for the acceleration of genetic improvement in European seabass and gilthead seabream breeding programs.

Data availability

Raw sequence reads from the European seabass and gilthead seabream population pools analyzed for SNP discovery have been deposited in NCBI's Sequence Read Archive (SRA) under accession number PRJEB40423. Details of the allele frequencies of the SNPs on the MedFish array can be found in the Mendeley Data Repository (https://doi.org/10.17632/7w4cb4mdd4.1).

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

This study was made possible by the EU projects MedAID (H2020 grant agreement No 727315) and PerformFish (H2020 grant agreement No 727610). We are grateful to both projects' partners that contributed with samples for SNP discovery as well as the projects' coordinators (A. López-Francos, B. Basurco, D. Furones and K. Moutou) and the H2020 project officer for enhancing a fruitful interaction between research consortia. The authors would also like to thank Christos Palaiokostas and Sara Faggion for providing additional information for trait-associated variants included in the SNP array. CP and RH are supported by funding from the BBSRC Institute Strategic Programme Grants BB/P013759/1 and BB/P013740/1. We also give thanks to the sequencing facilities Edinburgh Genomics (University of Edinburgh, UK) and the Norwegian Sequencing Centre (University of Oslo, Norway). TM would like to thank George Nomikos for valuable support on data analysis. This research was supported through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology, and Aquaculture) of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, LifeWatchGreece RI, and the CMBR (Centre for the study and sustainable exploitation of Marine Biological Resources) RI. We also thank the GeneSea consortium (n° R FEA 4700 16 FA 100 0005) for providing some variants allowing interoperability with existing arrays.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.04.038.

References

[1] R.D. Houston, T.P. Bean, D.J. Macqueen, M.K. Gundappa, Y.H. Jin, T.L. Jenkins, S. L.C. Selly, S.A.M. Martin, J.R. Stevens, E.M. Santos, A. Davie, D. Robledo, Harnessing genomics to fast-track genetic improvement in aquaculture, Nat. Rev. Genet. 21 (2020) 389–409.

[2] J. Fernández, M.Á. Toro, A.K. Sonesson, B. Villanueva, Optimizing the creation of base populations for aquaculture breeding programs using phenotypic and genomic data and its consequences on genetic progress, Front. Genet. 5 (2014).

[3] T. Moen, M. Baranski, A.K. Sonesson, S. Kjøglum, Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (*Salmo salar*): population-level associations between markers and trait, BMC Genomics 10 (2009) 368.

[4] R.D. Houston, C.S. Haley, A. Hamilton, D.R. Guy, A.E. Tinch, J.B. Taggart, B. J. McAndrew, S.C. Bishop, Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic Salmon (*Salmo salar*), Genetics 178 (2008) 1109–1115.

[5] C. Fraslin, E. Quillet, T. Rochat, N. Dechamp, J.-F. Bernardet, B. Collet, D. Lallias, P. Boudinot, Combining multiple approaches and models to dissect the genetic architecture of resistance to infections in fish, Front. Genet. 11 (2020).

[6] S.S. Horn, B. Ruyter, T.H.E. Meuwissen, H. Moghadam, B. Hillestad, A.K. Sonesson, GWAS identifies genetic variants associated with omega-3 fatty acid composition of Atlantic salmon fillets, Aquaculture 514 (2020) 734494.

[7] T.H. Meuwissen, B.J. Hayes, M.E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics 157 (2001) 1819–1829.

[8] K.R. Zenger, M.S. Khatkar, D.B. Jones, N. Khalilisamani, D.R. Jerry, H.W. Raadsma, Genomic selection in aquaculture: application, limitations and opportunities with special reference to marine shrimp and pearl oysters, Front. Genet. 9 (2019).

[9] C. Palaiokostas, S. Cariou, A. Bestin, J.S. Bruant, P. Haffray, T. Morin, J. Cabon, F. Allal, M. Vandeputte, R.D. Houston, Genome-wide association and genomic prediction of resistance to viral nervous necrosis in European sea bass (*Dicentrarchus labrax*) using RAD sequencing, Genet. Sel. Evol. 50 (2018) 30.

[10] C. Palaiokostas, S. Ferraresso, R. Franch, R.D. Houston, L. Bargelloni, Genomic prediction of resistance to pasteurellosis in gilthead sea bream (*Sparus aurata*) using 2b-RAD sequencing, G3: Genes|Genomes|Genetics 6 (2016) 3693–3700.

[11] M.L. Aslam, R. Carraro, A. Bestin, S. Cariou, A.K. Sonesson, J.-S. Bruant, P. Haffray, L. Bargelloni, T.H.E. Meuwissen, Genetics of resistance to photobacteriosis in gilthead sea bream (*Sparus aurata*) using 2b-RAD sequencing, BMC Genet. 19 (2018) 43.

[12] Eurostat, Agriculture, forestry and fishery statistics, in: Eurostat Statistical Books, 2019.

[13] M. Tine, H. Kuhl, P.-A. Gagnaire, B. Louro, E. Desmarais, R.S.T. Martins, J. Hecht, F. Knaust, K. Belkhir, S. Klages, R. Dieterich, K. Stueber, F. Piferrer, B. Guinand, N. Bierne, F.A.M. Volckaert, L. Bargelloni, D.M. Power, F. Bonhomme, A.V. M. Canario, R. Reinhardt, European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation, Nat. Commun. 5 (2014) 5770.

[14] M. Pauletto, T. Manousaki, S. Ferraresso, M. Babbucci, A. Tsakogiannis, B. Louro, N. Vitulo, V.H. Quoc, R. Carraro, D. Bertotto, R. Franch, F. Maroso, M.L. Aslam, A. K. Sonesson, B. Simionati, G. Malacrida, A. Cestaro, S. Caberlotto, E. Sarropoulou, C.C. Mylonas, D.M. Power, T. Patarnello, A.V.M. Canario, C. Tsigenopoulos, L. Bargelloni, Genomic analysis of Sparus aurata reveals the evolutionary dynamics of sex-biased genes in a sequential hermaphrodite fish, Commun. Biol. 1 (2018) 119.

[15] Y. Palti, G. Gao, S. Liu, M.P. Kent, S. Lien, M.R. Miller, C.E. Rexroad 3rd, T. Moen, The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout, Mol. Ecol. Resour. 15 (2015) 662–672.

[16] J.M. Yáñez, S. Naswa, M.E. Lopez, L. Bassini, K. Correa, J. Gilbey, L. Bernatchez, A. Norris, R. Neira, J.P. Lhorente, P.S. Schnable, S. Newman, A. Mileham, N. Deeb, A. Di Genova, A. Maass, Genomewide single nucleotide polymorphism discovery in Atlantic salmon (Salmo salar): validation in wild and farmed American and European populations, Mol. Ecol. Resour. 16 (2016) 1002–1011.

[17] R.D. Houston, J.B. Taggart, T. Cézard, M. Bekaert, N.R. Lowe, A. Downing, R. Talbot, S.C. Bishop, A.L. Archibald, J.E. Bron, D.J. Penman, A. Davassi, F. Brew, A.E. Tinch, K. Gharbi, A. Hamilton, Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*), BMC Genomics 15 (2014) 90.

[18] Q. Zeng, Q. Fu, Y. Li, G. Waldbieser, B. Bosworth, S. Liu, Y. Yang, L. Bao, Z. Yuan, N. Li, Z. Liu, Development of a 690 K SNP array in catfish and its application for genetic mapping and validation of the reference genome sequence, Sci. Rep. 7 (2017) 40347.

[19] S. Liu, L. Sun, Y. Li, F. Sun, Y. Jiang, Y. Zhang, J. Zhang, J. Feng, L. Kaltenboeck, H. Kucuktas, Z. Liu, Development of the catfish 250K SNP array for genome-wide association studies, BMC Res. Notes 7 (2014) 135.

[20] J. Xu, Z. Zhao, X. Zhang, X. Zheng, J. Li, Y. Jiang, Y. Kuang, Y. Zhang, J. Feng, C. Li, J. Yu, Q. Li, Y. Zhu, Y. Liu, P. Xu, X. Sun, Development and evaluation of the first high-throughput SNP array for common carp (*Cyprinus carpio*), BMC Genomics 15 (2014) 307.

[21] C.M. Nugent, J.S. Leong, K.A. Christensen, E.B. Rondeau, M.K. Brachmann, A. A. Easton, C.L. Ouellet-Fagg, M.T.T. Crown, W.S. Davidson, B.F. Koop, R. G. Danzmann, M.M. Ferguson, Design and characterization of an 87k SNP genotyping array for Arctic charr (*Salvelinus alpinus*), PLoS One 14 (2019), e0215008.

[22] C. Peñaloza, D. Robledo, A. Barría, T.Q. Trịnh, M. Mahmuddin, P. Wiener, J.A. H. Benzie, R.D. Houston, Development and validation of an open access SNP array for Nile tilapia (*Oreochromis niloticus*), G3: Genes|Genomes|Genetics 10 (2020) 2777–2785.

[23] R. Joshi, M. Árnyasi, S. Lien, H.M. Gjøen, A.T. Alvarez, M. Kent, Development and validation of 58K SNP-array and high-density linkage map in Nile Tilapia (*O. niloticus*), Front. Genet. 9 (2018).

[24] J.M. Yáñez, G. Yoshida, A. Barria, R. Palma-Véjares, D. Travisany, D. Díaz, G. Cáceres, M.I. Cádiz, M.E. López, J.P. Lhorente, A. Jedlicki, J. Soto, D. Salas, A. Maass, High-throughput single nucleotide polymorphism (SNP) discovery and validation through whole-genome resequencing in Nile tilapia (*Oreochromis niloticus*), Mar. Biotechnol. 22 (2020) 109–117.

[25] R. Griot, F. Allal, F. Phocas, S. Brard-Fudulea, R. Morvezen, A. Bestin, P. Haffray, Y. François, T. Morin, C. Poncet, A. Vergnet, S. Cariou, J. Brunier, J.-S. Bruant, B. Peyrou, P.-A. Gagnaire, M. Vandeputte, Genome-wide association studies for resistance to viral nervous necrosis in three populations of European sea bass (*Dicentrarchus labrax*) using a novel 57k SNP array DlabChip, Aquaculture 530 (2021) 735930.

[26] S.M. Aljanabi, I. Martinez, Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques, Nucleic Acids Res. 25 (1997) 4692–4693.

[27] S. Chen, Y. Zhou, Y. Chen, J. Gu, Fastp: an ultra-fast all-in-one FASTQ preprocessor, Bioinformatics 34 (2018) i884–i890.

[28] H. Li, R. Durbin, Fast and accurate short read alignment with burrows–wheeler transform, Bioinformatics 25 (2009) 1754–1760.

[29] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome project data processing, the sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.

[30] E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, in: arXiv, 2012.

[31] O. Tange, GNU parallel - the command-line power tool, in: Login: The USENIX Magazine, Frederiksberg, Denmark, 2011, p. 5.

[32] M. Babbucci, S. Ferraresso, M. Pauletto, R. Franch, C. Papetti, T. Patarnello, P. Carnier, L. Bargelloni, An integrated genomic approach for the study of mandibular prognathism in the European seabass (*Dicentrarchus labrax*), Sci. Rep. 6 (2016) 38673.

[33] S. Faggion, M. Vandeputte, B. Chatain, P.-A. Gagnaire, F. Allal, Population-specific variations of the genetic architecture of sex determination in wild European sea bass *Dicentrarchus labrax* L, Heredity (Edinb) 122 (2019) 612–621.

[34] D. Kyriakis, A. Kanterakis, T. Manousaki, A. Tsakogiannis, M. Tsagris, I. Tsamardinos, L. Papaharisis, D. Chatziplis, G. Potamias, C.S. Tsigenopoulos, Scanning of genetic variants and genetic mapping of phenotypic traits in gilthead sea bream through ddRAD sequencing, Front. Genet. 10 (2019).

[35] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D. M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3, Fly (Austin) 6 (2012) 80–92.

[36] M. Leitwein, B. Guinand, J. Pouzadoux, E. Desmarais, P. Berrebi, P.-A. Gagnaire, A dense brown trout (*Salmo trutta*) linkage map reveals recent chromosomal rearrangements in the salmo genus and the impact of selection on linked neutral diversity, in: G3 (Bethesda, Md.), 2017, pp. 1365–1376.

[37] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R. E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, G. McVean, R. Durbin, G. Genomes Project Analysis, The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158.

[38] J. Lachance, S.A. Tishkoff, SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it, BioEssays: News Rev. Mol. Cell. Dev. Biol. 35 (2013) 780–786.

[39] D.K. Malomane, C. Reimer, S. Weigend, A. Weigend, A.R. Sharifi, H. Simianer, Efficiency of different strategies to mitigate ascertainment bias when using SNP panels in diversity studies, BMC Genomics 19 (2018) 22.

[40] V. Hivert, R. Leblois, E.J. Petit, M. Gautier, R. Vitalis, Measuring genetic differentiation from pool-seq data, Genetics 210 (2018) 315–330.

[41] D.W. Roberts, Package 'LabDSV', 2012.

[42] C.C. Chang, C.C. Chow, L.C. Tellier, S. Vattikuti, S.M. Purcell, J.J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets, Gigascience 4 (2015) 7.

[43] P. Scheet, M. Stephens, A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, Am. J. Hum. Genet. 78 (2006) 629–644.

[44] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.

[45] P. Hardenbol, F. Yu, J. Belmont, J. MacKenzie, C. Bruckner, T. Brundage, A. Boudreau, S. Chow, J. Eberle, A. Erbilgin, M. Falkowski, R. Fitzgerald, S. Ghose, O. Iartchouk, M. Jain, G. Karlin-Neumann, X. Lu, X. Miao, B. Moore, M. Moorhead, E. Namsaraev, S. Pasternak, E. Prakash, K. Tran, Z. Wang, H.B. Jones, R.W. Davis, T.D. Willis, R.A. Gibbs, Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay, Genome Res. 15 (2005) 269–275.

[46] A.M. Ramos, R.P.M.A. Crooijmans, N.A. Affara, A.J. Amaral, A.L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M.S. Hansen, J. Hedegaard, Z.-L. Hu, H.H. Kerstens, A.S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G.A. Rohrer, M.F. Rothschild, T.P.L. Smith, R.D. Schnabel, C.P. Van Tassell, J.F. Taylor, R.T. Wiedmann, L.B. Schook, M.A.M. Groenen, Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology, PLoS ONE 4 (2009) e6524.

[47] H. Qi, K. Song, C. Li, W. Wang, B. Li, L. Li, G. Zhang, Construction and evaluation of a high-density SNP array for the Pacific oyster (*Crassostrea gigas*), PLoS One 12 (2017), e0174007.

[48] L.K. Matukumalli, C.T. Lawley, R.D. Schnabel, J.F. Taylor, M.F. Allan, M.P. Heaton, J. O'Connell, S.S. Moore, T.P.L. Smith, T.S. Sonstegard, C.P. Van Tassell, Development and characterization of a high density SNP genotyping assay for cattle, PLoS One 4 (2009), e5350.

[49] J.B. Fan, A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. McBride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, M.S. Chee, Highly parallel SNP genotyping, Cold Spring Harb. Symp. Quant. Biol. 68 (2003) 69–78.

[50] A. Albrechtsen, F.C. Nielsen, R. Nielsen, Ascertainment biases in SNP chips affect measures of population divergence, Mol. Biol. Evol. 27 (2010) 2534–2547.

[51] J. Chu, Y. Zhao, S. Beier, A.W. Schulthess, N. Stein, N. Philipp, M.S. Röder, J. C. Reif, Suitability of single-nucleotide polymorphism arrays versus genotyping-by-sequencing for genebank genomics in wheat, Front. Plant Sci. 11 (2020).

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

[53] D. Robledo, C. Palaiokostas, L. Bargelloni, P. Martínez, R. Houston, Applications of genotyping by sequencing in aquaculture breeding and genetics, Rev. Aquac. 10 (2018) 670–682.
[54] E.L. Souche, B. Hellemans, M. Babbucci, E. MacAoidh, B. Guinand, L. Bargelloni, D. A. Chistiakov, T. Patarnello, F. Bonhomme, J.T. Martinsohn, F.A.M. Volckaert, Range-wide population structure of European sea bass *Dicentrarchus labrax*, Biol. J. Linn. Soc. 116 (2015) 86–105.
[55] C. Brown, D. Miltiadou, C. Tsigenopoulos, Prevalence and survival of escaped European seabass *Dicentrarchus labrax* in Cyprus identified using genetic markers, Aquac. Environ. Interact. 7 (2015).
[56] E.-S. Polovina, E. Kourkouni, C.S. Tsigenopoulos, P. Sanchez-Jerez, E. D. Ladoukakis, Genetic structuring in farmed and wild gilthead seabream and European seabass in the Mediterranean Sea: implementations for detection of escapees, Aquat. Living Resour. 33 (2020) 7.
[57] T. Šegvić-Bubić, L. Grubišić, Ž. Trumbić, R. Stanić, J. Ljubković, J. Maršić-Lučić, I. Katavić, Genetic characterization of wild and farmed European seabass in the Adriatic Sea: assessment of farmed escapees using a Bayesian approach, ICES J. Mar. Sci. 74 (2016) 369–378.
[58] F. Maroso, K. Gkagkavouzis, S. De Innocentiis, J. Hillen, F. do Prado, N. Karaiskou, J.B. Taggart, A. Carr, E. Nielsen, A. Triantafyllidis, L. Bargelloni, C. the Aquatrace, Genome-wide analysis clarifies the population genetic structure of wild gilthead sea bream (*Sparus aurata*), PLoS ONE 16 (2021) e0236230.
[59] J.A. Alarcón, A. Magoulas, T. Georgakopoulos, E. Zouros, M.C. Alvarez, Genetic comparison of wild and cultivated European populations of the gilthead sea bream (*Sparus aurata*), Aquaculture 230 (2004) 65–80.