Acronym: PerformFISH

Title: Consumer-driven production: Integrating Innovative Approaches for Competitive and Sustainable Performance across the Mediterranean Aquaculture Value Chain

Grant Agreement: 727610

Deliverable 1.1
SNP-chip tool validation

Lead parties for Deliverable: HCMR

Due date of deliverable: 30/4/2019

Actual submission date: 23/4/2020

Version: 3

Dissemination level: PUBLIC

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein. Ref. Ares(2018)647107 - 02/02/2018

Contributors

Tereza Manousaki, Alexandros Tsakogiannis, Katerina Ekonomaki, Vasso Terzoglou & Costas Tsigenopoulos (HCMR),

Serena Ferraresso, Marianna Pauletto & Luca Bargelonni (UNIPD)

Antonio Colli (Selonda), Kostas Tzokas (Andromeda) & Leonidas Papaharisis (Nireus) (FGM)

Massimo Caggiano (PISASrl) (API)

Renata Barić (CROMARIS) (CCE)

Carlos Mazzora (SONRIONANSA) & Javier Ojeda (APROMAR)

Table of contents

1 Introduction

Only 20% of the global aquaculture production is estimated to be based on genetically improved stocks and more than 50% of genetic improvement programs in aquaculture began after 2005. It is therefore, a standing challenge for the aquaculture industry to harness the potential of genetic improvement by developing more selective breeding programs for both established and new/emerging species. Novel scientific approaches can support an increase in aquaculture production by up to 40% aiming to better quality fish that are growing faster and are more efficient at converting feed.

Atlantic salmon is the species where the greatest gains have been achieved, and it is also the only one being farmed from almost 100% improved stock based on family selection. Salmon production in Europe is slightly bigger than those of the European sea bass (BSS) and gilthead sea bream (SBG) taken together; these two species are the most important aquaculture species in the Mediterranean. It has been therefore a need to develop breeding programmes in the two species, probably better both at the same time. For the BSS, there is already a high-quality genome available (Tine et al., 2014). For the SBG, a recently big progress in genetic research has been also made towards a reference genome map which combines genetic information on the inheritance of chromosomes throughout generations with the physical DNA sequences which contain genes that control biological processes (Pauletto et., 2018; Pérez-Sánchez et al., 2019). Having a reference genome map, it was expected to facilitate the development and production of robust genetics research tools that were shown to have a tremendous impact in livestock breeding.

Dense Single Nucleotide Polymorphism (SNP) genotyping arrays provide extensive information on variation across the genome of species of interest. Such information can be used in studies of the genetic architecture of quantitative traits and to improve the accuracy of selection in breeding programmes. In Atlantic salmon (*Salmo salar*) for example, these goals were thought to be hampered by the lack of a high-density SNP genotyping platform; as a result, different arrays were developed and are now publicly available with the objective to be likely used as a platform for high-resolution genetics research into traits of evolutionary and economic importance in salmonids, and in aquaculture breeding programs via genomic selection. Through these arrays, egg producers and in general aquaculture managers are expected to have access to a scientific tool to customise breeding programmes and develop healthier and robust strains of the species in a naturally sustainable way while reducing the use of therapeutic chemicals.

SNP-arrays (or SNP-chips) are powerful tools for genotyping thousands of markers in parallel with an error rate of 0.1% (Saunders et al. 2007) or less (Weller et al. 2010). The development of a SNP-chip sets a baseline for genomic selection, genome-wide association mapping, and QTL mapping in aquaculture species.

This deliverable describes a major task in WP1 “Selective Breeding for Robust Fish” to design, implement and validate a SNP-array that will serve as a basic tool for selective breeding in both species. SNP-chips have already been developed and applied to improve livestock in traditionally domesticated terrestrial animals such as cow (see Nikolazzi et al. 2014 for an overview), chicken (Groenen et al. 2011; Kranis et al. 2013), pig (Ramos et al. 2009) and goat (Tosser-Klopp et al.

2014). Regarding fish, previous experience on SNP-chip development comes among others from the Atlantic salmon (Houston et al. 2014; Yanez et al. 2014), common carp (Xu et al. 2014) and rainbow trout (Palti et al. 2015). Especially for the Atlantic salmon, the chip application has led to GWAS analysis (Correa et al. 2015; Tsai et al. 2015) illustrating the advantages of such a tool.

Till now, the chosen strategies from different studies include either sequencing a part of the genome (e.g. RNASeq, GBS) of many individuals or whole genome re-sequencing of fewer individuals. The former provides higher coverage of the sequenced moiety of the genome, while the latter identifies many more markers evenly distributed across the genome with lower coverage.

In this deliverable, farmed as well as wild populations' samples of both European sea bass (BSS) and gilthead sea bream (SBG) were collected under the responsibility of HCMR and UNIPD, respectively. Samples covered the full range from the Atlantic Ocean to the Eastern Mediterranean; some additional domesticated populations were provided through the MedAID consortium.

The samples were deeply sequenced using the technique PoolSeq (Hivert et al. 2018) which resulted in the discovery of SNPs genome-wide. Following multiple bioinformatic filters, a total of 29,888 and 29,807 SNP markers for BSS and SBG, respectively, have been included in the Med_Fish SNP chip. Both sequencing and bioinformatic analysis effort in means of costs and human resources were jointly undertaken by the PerformFISH and MedAID consortia towards having a common SNP-arrays for both species. Furthermore, a number of SNPs was shared by the French-funded project GèneSEA to allow for comparable results between the two platforms. A first validation of the designed SNP revealed a highly successful tool with 90% high resolution markers for BSS and 85% for SBG.

2 Materials and methods

2.1 Sampling

More than 950 BSS samples (fin-clips) were collected from twenty-eight (28) populations: among them, twelve (12) are farmed broodstock fish and the sixteen (16) are fish captured in the wild; furthermore, through the collaboration of the MedAID consortium, three (3) additional populations from Turkey and Croatia were added to our collection (**Figure 1**). Most fish are now stored in HCMR freezers, preserved in absolute ethanol. From these samples, 24 batches/populations (8 wild and 16 farmed) having a good number of specimens (usually >25) were chosen for the SNP-array design (**Table 1**).

For the SBG samples, UNIPD had the leading role in collecting and storing samples at its premises. Samples from farmed Greek and Croatian companies were sent from HCMR to UNIPD and these were combined to finally have a collection of more than 1,700 fish coming from fifteen (15) farmed and more than forty (40) wild populations, most of them being part of the AQUATRACE (https://aquatrace.eu/index.html) sampling campaign; through the collaboration with the MedAID consortium, three (3) additional populations from Turkey, Spain and Italy were added in our collection (**Figure 1**). From these samples, 27 batches/populations (13 wild and 14

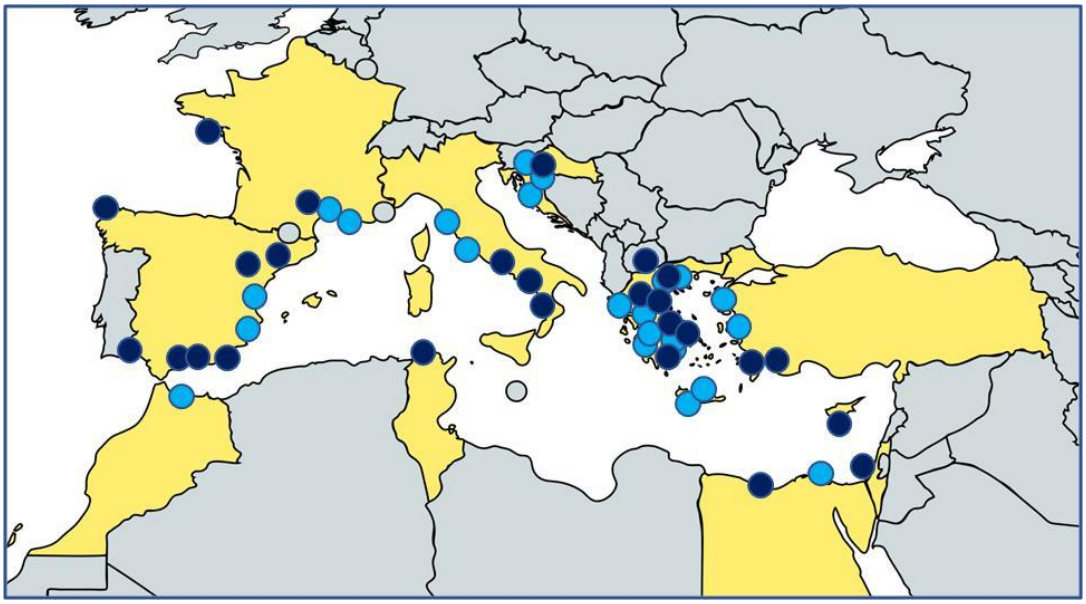

Figure 1. A total of 20 SBG (deep blue circles) and 21 BSS populations (light blue circles) including farmed and wild specimens, were processed following the preparation protocol agreed within the collaboration between PerformFISH and MedAID.

farmed) having a good number of specimens (usually >25) were chosen for the SNP-array design (see also **Table 1**).

2.2 Library Preparations and Next-Generation Sequencing

The next step consisted of performing DNA extractions from the above samples following up to date protocols in order to assure high molecular weight (quality) and quantity (yield) DNA; all these laboratory work has completed, and aliquots were prepared for each fish using fluorometric-based methods for quantification, such as Qubit or Quant-iT PicoGreen. For each sample, dilutions were prepared at 100 ng/µl, and for each population, 25 samples (when available, see **Table 1**) were equimolarly pooled together in order to achieve 100 ng of each sample which should guarantee twice the optimal DNA quantity (2-2.5 µg) necessary for Covaris fragmentation (350 bp insert size). To minimize sampling errors, two technical replicates per pool were created and each one was used for a library preparation (two libraries/pool).

Next Generation Sequencing work has been performed into two sequencing centres; the one selected by HCMR through an open Call for Tenders (NSC, subcontract) and the other in collaboration with MedAID colleagues. A wide range of data sources was used for SNP mining to achieve a coverage of >30X for each individual genome. Further to the initial goal of 25K high-quality SNPs per species, the estimation of the SNPs per species arose to approximately 37K (around 74K total) in collaboration again with the MedAID consortium. HCMR and UNIPD actively participated in the definition and validation of all the experimental protocols for DNA sequencing (including bioinformatics pipelines), which was crucial to obtain a large set of SNPs from each species. In turn, this formed the candidate SNP list to design an optimal SNP-chip.

Both institutes had finally an important role in coordinating task 1.1 efforts and to harmonize them within the collaboration scheme with partners from the MEDAID consortium.

Table 1. Populations of European sea bass and gilthead sea bream sampled

		Atlantic	Strait of Gibraltar	W. Med	France	Adriatic	W.Greece	Aegean	E.Med
European sea bass	**Wild**		1 (Morocco)	1 (Spain)	1	2	1	1	1
	Farmed	1 (Spain)		1 (Spain)		3 (Italy, Croatia)	4	5 (Greece, Turkey)	2 (Cyprus, Egypt)
Gilthead sea bream	**Wild**	2 (France, Spain)	1 (Spain)	3 (Spain, Tunisia)		2	1	4	
	Farmed	1		1	1	2 (Italy, Croatia)	2	5 (Greece, Turkey)	2 (Israel, Egypt)

2.3 Data analysis

2.3.1 SNP chip design: General procedure

SNP calling

A SNP discovery pipeline (see **Figure 2** for an overview) with strict filtering steps has been developed. The raw sequenced reads were first quality controlled using fastp 0.19.10 (Chen et al. 2018). The high quality reads were then mapped against the genomes of SBG (Pauletto et al. 2018) and BSS (Tine et al. 2014) using bwa 0.7.15-r1140 aligner (Li & Durbin 2019). Following the alignment, duplicated reads (those mapped at identical positions) were eliminated using samtools v1.9 (Li et al. 2009). Finally, SNP calling was performed using freebayes v1.2.0-4-gd15209e (https://arxiv.org/abs/1207.3907v2).

SNP filtering

Following the initial SNP discovery a series of filters were applied to ensure the highest possible quality for the SNPs chosen to be included in the SNP array (see **Figure 3** for an overview). The main steps included i. removal of the SNPs near indels or other SNPs to increase the probe effectiveness, ii. Keep only bi-allelic SNPs, removal of SNPs with excessive read depth (<100x), iii. remove those that are either A/T or G/C as using these types of variants would require two probes in the SNPchip instead of one, i.v. genotype quality filter, v. removal of SNPs with minor allele frequency (MAF) below a certain threshold (MAF <0.05) and finally v.i removal of SNPs that are heterozygous in all samples as they most probably represent repetitive regions as well.

SNP selection

Given that the platform accommodates ~30K SNPs per species, a SNP selection strategy was implemented to allow a better representation of high quality SNPs across the two genomes.

First, SNPs with known impact of important traits have been prioritised (for BSS SNPs discovered in Babbucci et al. 2016, Palaiokostas et al. 2018, Faggion et al. 2019 and for SBG those discovered in Kyriakis et al. 2019 and Aslam et al. 2018). In case some of those SNPs were not found in our dataset, the closest discovered SNP was used instead. Second, we prioritised SNPs that are diagnostic of certain populations. Third, we included SNPs developed in the GÈNESEA (R-FEA-4700-16-FA-100-0005) project (Allal et al. 2018 for the BSS) to allow comparable results between the two platforms. Then, we gave priority to SNPs that fall within coding regions and have a high impact on the protein sequences as characterized by SNPeff v 4.3 (Cingolani et al. 2012).

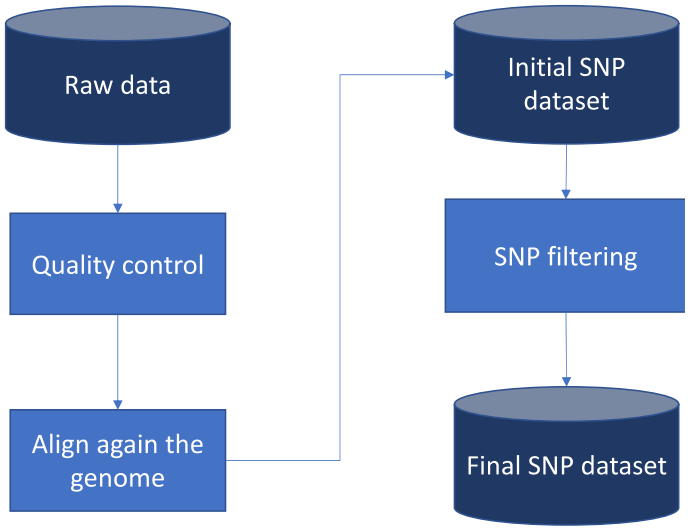

Figure 2. An overview of the bioinformatic pipeline developed.

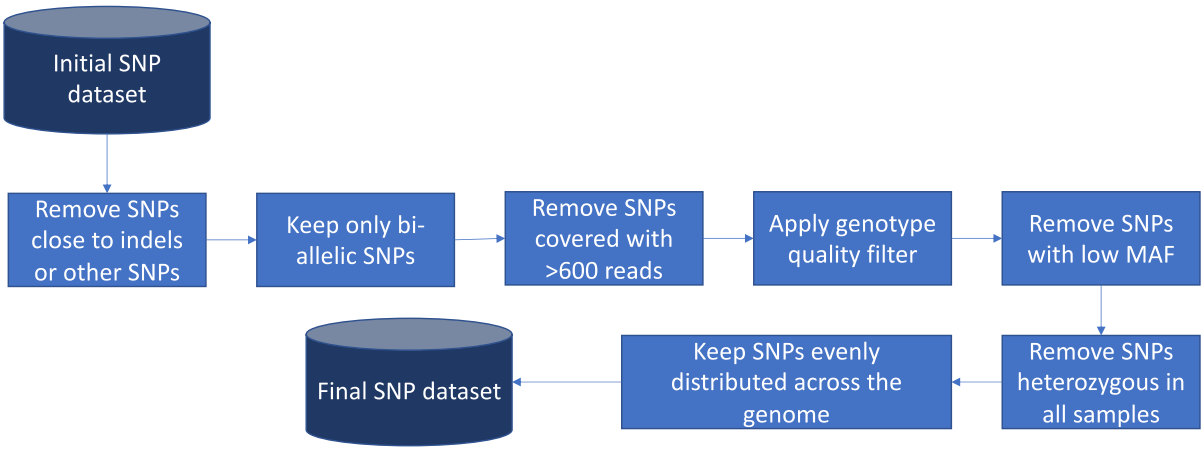

Figure 3. The SNP filtering pipeline followed.

The final layer of selection was based on the discovery that the distribution of SNPs along chromosomes is uneven. Regions with higher SNP density indicate a higher recombination rate and therefore lower linkage disequilibrium. Evidence that nucleotide diversity is negatively correlated to the local recombination rate was shown for the BSS by Tine 2014. Hence, regions with higher SNP density have been sampled more intensely based on nucleotide diversity (π) (Faggion et al. 2019). The two genomes have been divided to non-overlapping windows of 70Kb (for European sea bass) or 85kb (for the gilthead sea bream). In each window local nucleotide diversity was estimated using VCFtools v 0.1.15 (Danecek et al. 2011). Then, we binned the windows based on the π value to five classes (**Table 2**) and selected more SNPs from more polymorphic regions taking into account the MAF as well.

Table 2: Diversity window for both species

			Number of SNPs sampled per allele frequency category		
Type of diversity window	Range	**Number of SNPs sampled per window**	>0.3	0.3-0.2	0.2-0.1
Class 1	π >=0.001	1	1	0	0
Class 2	0.001< π <=0.002	2	2	0	0
Class 3	0.002< π <=0.003	3	2	1	0
Class 4	0.003< π <=0.004	4	2	2	0
Class 5	π >0.004	5	2	2	1

3 Results & Discussion

3.1 SNP discovery

Illumina Sequencing resulted in 2,100 and 1,010 Gb for SBG and BSS out of 20 and 7 Illumina lanes, respectively. This dataset was used for the initial SNP discovery. The unfiltered set of SNPs comprised of 34,754,947 variants for SBG and 17,674,539 variants for BSS (**Figure 4**). The deeper sequencing of SBG led to more variants at the first stage. At this stage, we compared the technical replicates of each pool, and saw that they were highly comparable, ensuring the high-quality application of the pooling process at the library preparation step (**Figure 5**). After applying all quality control filters, we ended up with 1,073,309 and 1,078,567 high-quality SNPs. From this dataset we selected 30K SNPs that were provided to the Thermo Fisher Scientific company based on the priority criteria described in M&Ms. The discovered SNPs were not evenly distributed across the genome. However, following the selection based on SNP density and rate of recombination we managed to get evenly distributed SNPs across the genome (see **Figure 6**).

The final set of SNPs is 29,807 for SBG and 29,888 for BSS. The final SNP chip content is presented in **Table 3**. The list of SNPs included in the array are freely available upon written request to Dr Tereza Manousaki (tereza@hcmr.gr).

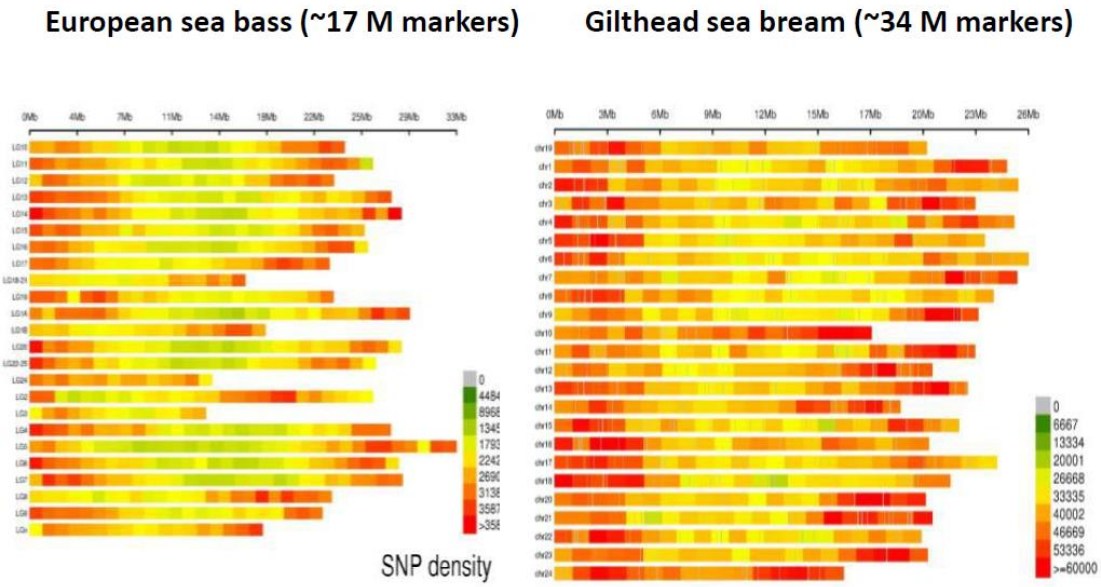

Figure 4. SNP distribution on the two species reference genome maps.

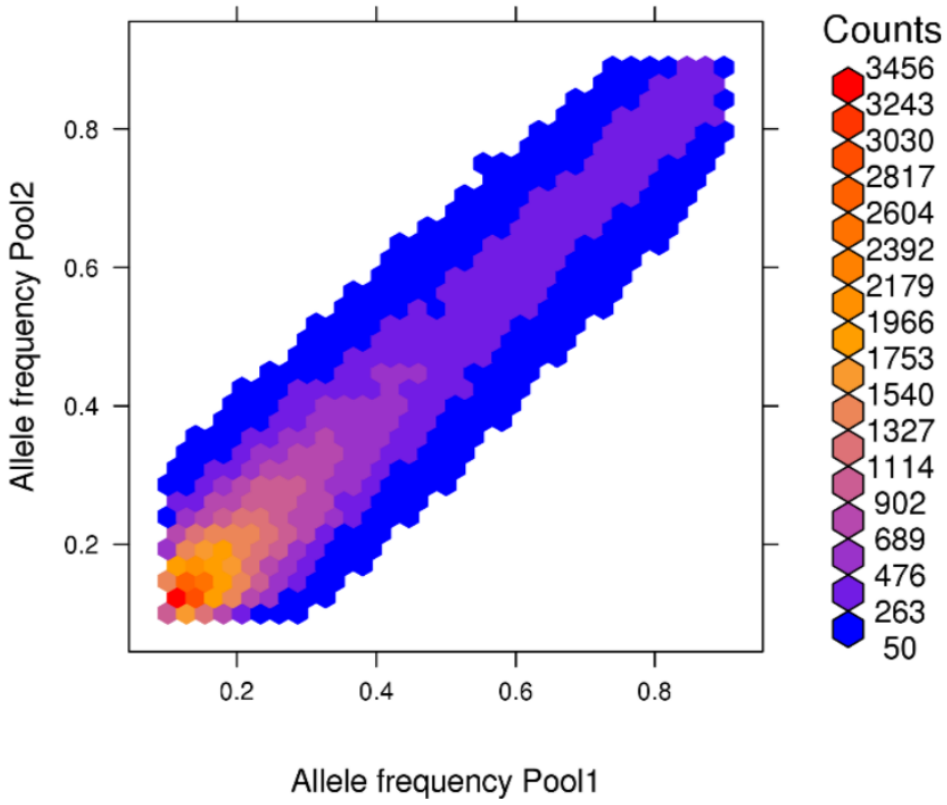

Figure 5. Technical replicates show high consistence. Here, the allele frequencies of each SNP as estimated from the two replicates are shown.

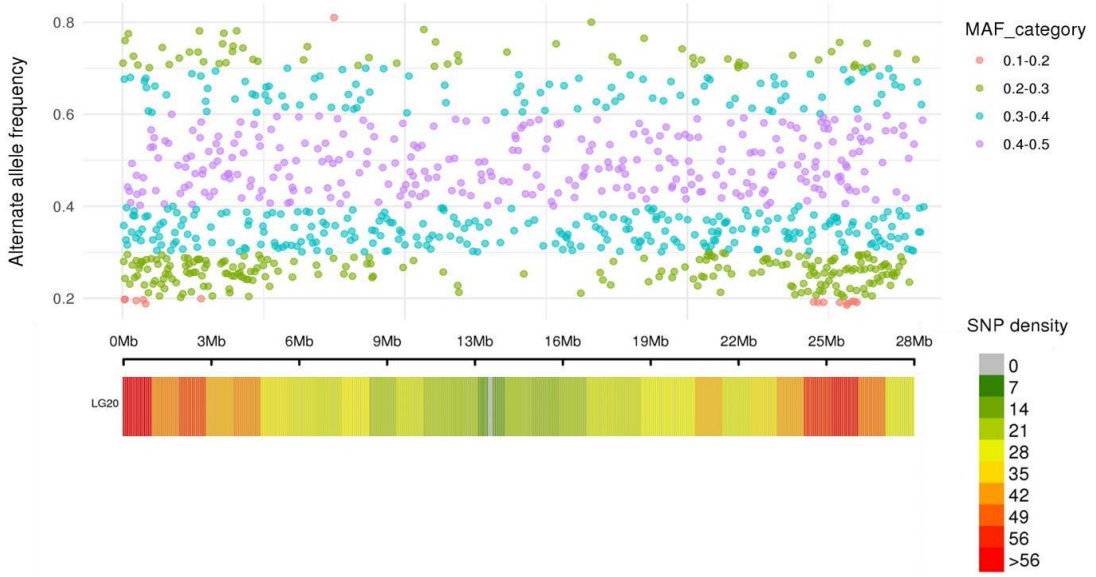

Figure 6. Selected SNP distribution across chromosome LG20 of seabass. Different colors represent different bins of MAF.

Table 3: Final SNP-array content

Priority	Type	BSS	SBG
1	SNPs associated with production traits	22	79
1	Diagnostic	0	24
2	SNPs with a predicted high effect on genes	107	179
2	Shared with the GèneSEA platform	4,560	3,208
3	High quality SNPs selected based on diversity windows	25,199	26,317
Total		29,888	29,807

3.2 SNP Validation

To validate the designed SNP array we have genotyped 79 individuals for SBG and 155 for BSS. The validation results showed an exceptionally functional SNP array with high clustering resolution (PolyHighResolution SNPs) and high call rates for both BSS (89.51 %) and SBG (85.43 %). These preliminary results underline the great SNP array design quality.

This Axiom® MED FISH (Seabass/bream) Genotyping Array is already publicly available through Affymetrix (ThermoFisher Scientific), it has a part number of 551144 and is available in 384 format.

3.3 Perspectives

Our overarching aim was to develop and validate a medium-density SNP genotyping array that will serve as a basic tool for genomic selection, genome-wide association mapping and QTL mapping in both species. The ultimate goal is to provide a functional tool to the whole breeding community of BSS and SBG, being the industrial actors or the scientists.

More particularly and presently, the array will be employed as a genomic tool into the PerformFISH context, to further explore the genetic resistance of a) the BSS to a parasite infection (*Lernanthropus kroyeri*), and b) the SBR larvae to the re-assortant (RGNNV/SJNNV) strain of beta-nodavirus, both considered to be serious emerging threats to Mediterranean aquaculture.

4 Acknowledgements

The contributors would like to warmly thank the MedAID collaborators Carolina Penaloza and Ross D. Houston (The Roslin Institute, University of Edinburgh) and Anna Sonesson (NOFIMA) as well as the GèneSEA collaborators François Allal (IFREMER), Marc Vendeputte (INRA) and Pierrick Haffray (SYSAAF).

5 References

Allal, F., Morvezen, R., Brard-Fudulea, S., et al. 2018. Development and characterization of a 57K Single Nucleotide Polymorphism array for European sea bass, AQUA2018, Montpellier, France, 25-29 August 2018, p.37.

Chen, S., Zhou, Y., Chen, Y., Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics*, 34, 17, i884–i890, https://doi.org/10.1093/bioinformatics/bty560

Cingolani, P., Platts, A., Wang, L., et al. 2012 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, 80-92. https://doi.org/10.4161/fly.19695

Correa, K., Lhorente, J.P., López, M.E. et al. 2015Genome-wide association analysis reveals loci associated with resistance against *Piscirickettsia salmonis* in two Atlantic salmon (*Salmo salar* L.) chromosomes. BMC Genomics, 16, 854. https://doi.org/10.1186/s12864-015-2038-7

Danecek, P., Auton, A., Abecasis, G., et al. 2011 The variant call format and VCFtools. Bioinformatics (Oxford, England) 27, 2156-2158. https://doi.org/10.1093/bioinformatics/btr330

Groenen, M.A., Megens, H., Zare, Y., et al. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics, 12**,** 274. https://doi.org/10.1186/1471-2164-12-274.

Faggion, S., Vandeputte, M., Chatain, B., Gagnaire, P.A., Allal, F. 2019 Population-specific variations of the genetic architecture of sex determination in wild European sea bass *Dicentrarchus labrax* L. Heredity (Edinb) 122, 612-621. https://doi.org/10.1038/s41437-018-0157-z

Houston, R.D., Taggart, J.B., Cézard, T. et al. 2014. Development and validation of a high density SNP genotyping array for Atlantic salmon (*Salmo salar*). BMC Genomics, 15, 90. https://doi.org/10.1186/1471-2164-15-90

Kranis, A., Gheyas, A.A., Boschiero, C., et al. 2013. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics, 14, 59. https://doi.org/10.1186/1471-2164-14-59.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, 25, 14, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 25, 16, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352

Nicolazzi, E.L., Picciolini, M., Strozzi, F., et al. 2014. SNPchiMp: a database to disentangle the SNPchip jungle in bovine livestock. BMC Genomics, 15, 123, https://doi.org/10.1186/1471-2164-15-12.

Palti, Y., Gao, G., Liu, S., Kent, M.P., Lien, S., Miller, M.R., Rexroad, C.E. 3rd, Moen, T. 2015 The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Mol Ecol Resour. 15(3), 662-672. https://doi.org/0.1111/1755-0998.12337

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

Pérez-Sánchez Jaume, Naya-Català Fernando, Soriano Beatriz, Piazzon M. Carla, Hafez Ahmed, Gabaldón Toni, Llorens Carlos, Sitjà-Bobadilla Ariadna, Calduch-Giner Josep A. 2019. Genome Sequencing and Transcriptome Analysis Reveal Recent Species-Specific Gene Duplications in the Plastic Gilthead Sea Bream (*Sparus aurata*), Frontiers in Marine Science, 6, 760, https://doi.org/10.3389/fmars.2019.00760

Ramos, A.M., Crooijmans, R.P.M.A., Affara, N.A., Amaral, A.J., Archibald, A.L., et al. 2009. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLOS ONE, 4(8), e6524. https://doi.org/10.1371/journal.pone.0006524

Saunders, I. W., Brohede, J., Hannan, G. N., 2007. Estimating genotyping error rates from Mendelian errors in SNP array genotypes and their impact on inference, Genomics, 90 (3), 291-296, https://doi.org/10.1016/j.ygeno.2007.05.011.

Schlötterer, C., Tobler, R., Kofler, R. *et al.* 2014 Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat Rev Genet* **15,** 749–763. https://doi.org/10.1038/nrg3803

Tine, M., Kuhl, H., Gagnaire, P. et al. 2014 European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun. 5**,** 5770. https://doi.org/10.1038/ncomms6770

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., et al. 2014 Design and Characterization of a 52K SNP Chip for Goats. PLOS ONE, 9(1), e86227. https://doi.org/10.1371/journal.pone.0086227.

Tsai, H. Y., Hamilton, A., Tinch, A. E., Guy, D. R., Gharbi, K., Stear, M. J., Matika, O., Bishop, S. C., & Houston, R. D. 2015. Genome wide association and genomic prediction for growth traits in juvenile farmed Atlantic salmon using a high density SNP array. BMC Genomics, 16, 969. https://doi.org/10.1186/s12864-015-2117-9

Tsai, H., Hamilton, A., Tinch, A.E. et al. 2016 Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet Sel Evol, 48**,** 47. https://doi.org/10.1186/s12711-016-0226-9

Weller, J.I., Glick, G., Ezra, E., Zeron, Y., Seroussi, E. and Ron, M. 2010. Paternity validation and estimation of genotyping error rate for the BovineSNP50 BeadChip. Animal Genetics, 41, 551-553. doi:10.1111/j.1365-2052.2010.02035.x

Xu, J., Zhao, Z., Zhang, X. et al. 2014. Development and evaluation of the first high-throughput SNP array for common carp (*Cyprinus carpio*). BMC Genomics, 15, 307. https://doi.org/10.1186/1471-2164-15-307

Yanez, J.M., Houston, R.D., Newman, S. 2014a. Genetics and genomics of disease resistance in salmonid species. Frontiers in Genetics, 5, 1–13. https://doi.org/10.3389/fgene.2014.00415

Yanez, J.M., Lhorente, J.P., Bassini, L.N. et al. 2014b Genetic co-variation between resistance against both *Caligus rogercresseyi* and *Piscirickettsia salmonis*, and body weight in Atlantic salmon (*Salmo salar*). Aquaculture, 433, 295–298. https://doi.org/10.1016/j.aquaculture.2014.06.026

List of Tables

List of Figures