Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

BIODIVERSITÁ ED EVOLUZIONE

Ciclo XXIV

Settore Concorsuale di afferenza: 05/B1 - ZOOLOGIA E ANTROPOLOGIA

Settore Scientifico disciplinare: BIO/05 - ZOOLOGIA

Novel tools for conservation genetics in marine fish: population structure and evolution of the European hake (*Merluccius merluccius*) inferred by SNP variation and applications to traceability

Presentata da: Dr. Ilaria Milano

Coordinatore Dottorato

Prof. Barbara Mantovani

Relatore

Prof. Barbara Mantovani

Correlatore

Prof. Fausto Tinti

Esame finale anno 2012

CHAPTER 1

THE FISHPOPTRACE PROJECT

1.1 BACKGROUND CONTEXT

1.1.1 THE STATUS OF EUROPEAN FISHERY RESOURCES

The worldwide fishing activity has led to a dramatic reduction of the natural and common fishery resources. According to the Food and Agriculture Organization (FAO), more than half of the stocks of marine fishes are fully exploited, therefore producing catches that are at their maximum sustainable limits, with no opportunity for further expansion, while the proportion of underexploited or moderately exploited stocks declined linearly from 40 percent in the mid-1970s to 20 percent in 2007. Moreover, since the mid-1990s one quarter of the world's fish stocks are either overexploited, depleted or recovering from depletion (FAO 2009). European Community waters are not exempt from this trend: currently, most fish stocks are exploited at levels well beyond their Maximum Sustainable Yield (MSY), the optimal volume of catches that can be taken each year without threatening the future reproductive capacity (Figure 1). European Commission declared that, according to 2009 scientific assessment data, 21.5% of fish stocks are exploited at levels delivering MSY, 35% are over-exploited and 43% are outside safe biological limits, indicating that the fishing pressure exceeds sustainability, i.e. mortality exceeds recruitment and growth. These estimates point out that 78.5% of Community stocks for which scientific advice is available are fished unsustainably (European Commission 2011a). In the Mediterranean Sea the situation is especially alarming, as the proportion of overfished stocks is higher (82%) than in Atlantic (63%), where a slight improvement

has been observed in the last six years, and landings of small juvenile individuals, that have not yet reproduced, are common (European Commission 2011b).

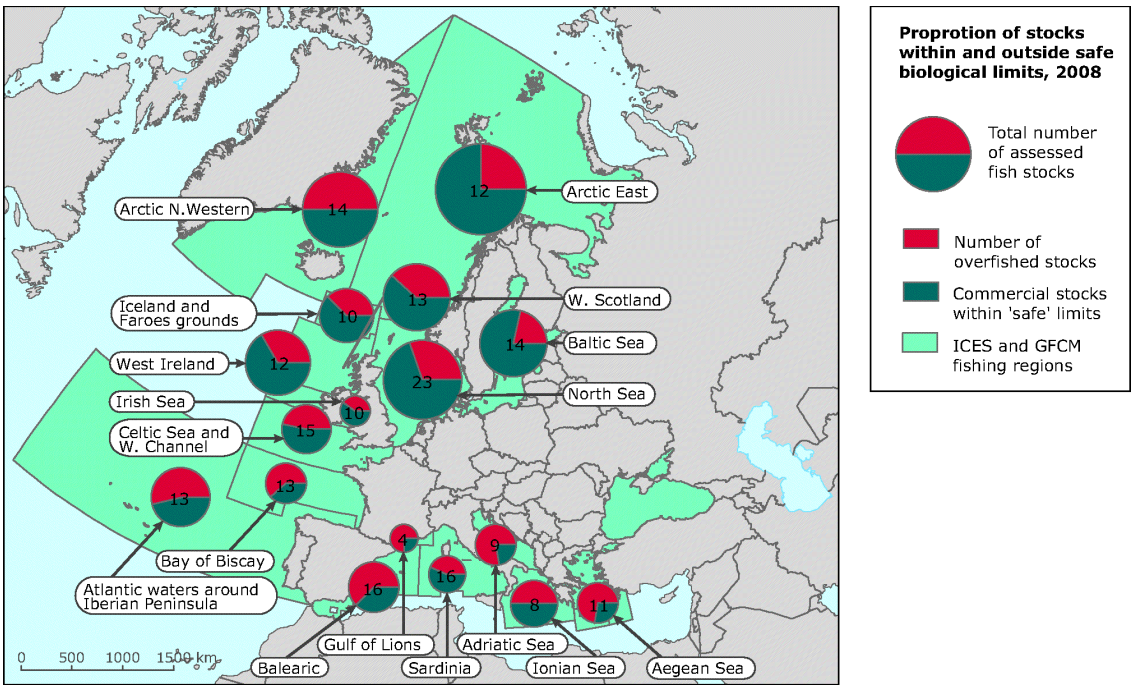


Figure 1. Status of fish stocks in the International Council for the Exploration of the Sea (ICES) and General Fisheries Commission for the Mediterranean (GFCM) fishing regions of Europe. The chart shows the proportion of assessed stocks that are overfished (red) and stocks within safe biological limits (blue). The numbers in the circles indicate the number of stocks assessed within the given region. The size of the circles is proportional to the magnitude of the regional catch. Data refers to the period 2005-2008. Source: European Environment Agency.

The Common Fisheries Policy (CFP) is the principal instrument of the European Union for the management of fisheries and aquaculture. Its underlying rationale is to ensure a sustainable exploitation of living aquatic resources, pursued through extensive regulatory actions, as the setting of Total Allowable Catches (TAC) and quota regulations, the limitation of fishing effort and the adoption of technical measures (minimum mesh sizes for nets, closed areas and seasons, minimum landing sizes). However, despite repeated attempts to overhaul and improve the CFP management scheme, overfishing of European fish stocks still represent a problematical issue, and the economic situation of fishery sector is going through a crisis.

1.1.2 ILLEGAL UNREPORTED UNREGULATED (IUU) FISHING

Efforts to regulate fishing activities aiming at maintaining sustainable levels of harvesting are greatly undermined by Illegal, Unreported and Unregulated (IUU) fishing and fraudulent activities along the supply chain. Common forms of IUU fishing include fishing without permission, catching protected species, breaches of gear restrictions, disregarding catch quotas, illegal trade in undersize fish, and deliberate underreporting or misreporting of catches. The most evident costs of IUU fishing, and probably the most important, are environmental impacts, such as the increased fish mortality, the reduced rates of stock growth and recruitment, the unsustainable harvesting of fish stocks. IUU activities also result in the destruction of marine habitats and can lead to disturbances of whole ecosystems. Besides the ecological impacts, IUU fishing entails economic and social costs, reducing profits in the medium- and long-term and decreasing employment opportunities in fishing and related industries. Worldwide IUU fishing is reported to lead to a loss of many billions of dollars of annual economic benefits (Agnew *et al.* 2009). In addition, the occurrence of unnoticed activities reduces the quality of scientific advice to managers, and the ability of authorities to undertake optimal management of the environment and natural resources.

Illegal activities extend into the supply chain, as shown by numerous fraud cases in Europe and worldwide where fish has been sold under false labels: mislabeling may involve not only the species identity, where lower-quality and less expensive fish is sold as more appreciated and expensive species, but also the geographic origin of fish, probably in relation to different legal and regulatory management frameworks of fishing areas.

1.1.3 SEAFOOD TRACEABILITY

In view of the actual status of most commercially exploited fish stocks, and the extent of illegal fishing, the provision of support to more efficient management strategies on various levels is of paramount importance. In this context, identification of species and tracing of fish back to their

original source constitute a highly valuable tool for monitoring, control and surveillance. In the attempts to circumvent catch limitations, fish may be sold under a false label, reporting a falsified geographic origin or different species name. Origin assignment and species identification are therefore important components of a legal framework underlying the contrast of illegal activities. Moreover, traceability also represent a guarantee for public health and consumer protection, being a useful tool to assure surveillance and certify the origin of fish throughout the entire food supply chain. This is particularly relevant with the increasing interest in consumption of eco-friendly products (Brécard *et al.* 2009) and hence eco-labelling systems, which assure consumers that a product has been produced according to defined environmental standards, such as the sustainability of the resource or the environment impact on the production method. However, while the European Community has launched a wide debate on eco-labelling schemes for fisheries products, the existing applicable EU reference legislation, the Commission Regulations (EC) No 104/2000 and 2065/2001, requires fish products to be labeled with the species name, its approximate catch area and whether it was caught or farmed.

Monitoring the correct application of labeling regulations in fish markets might be particularly hard for fillets, processed or frozen fish, which are not directly identifiable. The application of molecular and chemical techniques have revealed powerful tools in this framework, and have provided evidence of the wide extent of seafood mislabeling in European industry, especially in terms of species substitutions (Filonzi *et al.* 2010; Miller *et al.* 2011; Miller& Mariani 2010). However, this is more likely due to complexity of origin assignment of fish rather than to a real lower number of fraud cases linked to false origin declarations. In fact, in order to assess where fish and fish products come from, first a baseline needs to be created, meaning that for any species of interest in the context of fisheries control and enforcement, genetic or other data has to be available that reveals the existing population structure.

The potential of modern molecular techniques for conservation applications has been emphasized in the Regulation (EC) No 1224/2009, which explicitly refers to "genetic analysis and other fisheries control technologies" as possible future tools to adopt, in order to improve compliance with rules of the CFP in a cost effective way.

1.2 RATIONALE OF THE PROJECT

FishPopTrace (http://fishpoptrace.jrc.ec.europa.eu/) is a collaborative project funded under FP7 (7^{th} Framework Program), launched with the purpose of contributing to more effective enforcement and conservation in European fisheries. In particular, the underlying rationale of the project was aimed at improving the ability to trace fish and fish products through enhanced understanding of the dynamics, temporal stability and distribution of major populations of four key exploited fish species. In the management of fish resources it has been known for many decades that, although it is important to recognize what species an individual belongs to, it is also crucial to identify and monitor the distribution and dynamics of populations. Population diversity represents the focus for a sustainable utilization and conservation of exploited stocks, being the natural unit of evolutionary change. It is only through the careful identification and monitoring of population diversity that it becomes possible to develop strategies to maximize and conserve genetic resources for adaption to natural and human-induced environmental alteration.

FishPopTrace employed three primary criteria in the choice of target fish species, relating to conservation status, traceability issues and representation of marine fish ecology: Atlantic cod (*Gadus morhua*), Atlantic herring (*Clupea harengus*), European hake (*Merluccius merluccius*) and common sole (*Solea solea*). The selected species are all economically important, relatively widespread on a European scale, known to exhibit population structure and fall within European Community priority species for enforcement and/or conservation. For the four species different

levels of population genetic information are available, where cod has been studied genetically for decades, while relatively little is known about the population structure of hake. The multispecies approach encompasses different geographical scales for tracing individuals, representing a range of policy-led traceability scenarios.

Tools for monitoring natural populations and application to fisheries enforcement should meet stringent criteria: they should mirror population identity and stability over an ecological (environmental isolation) and evolutionary (limited interbreeding) scale. Moreover, traceability tools should be functional throughout the food supply chain from capture to a customer's plate and should be amenable to forensic validation. Taking advantage of the rapid progress in life science technologies, FishPopTrace has developed appropriate new tools for traceability studies for the four species throughout the food supply chain, encompassing methods applicable to ultimate processed (DNA based methods) and on-board whole fish (otolith-based methods). Major effort has been focused on genetics and otoliths investigation, selected as primary traceability tools, through the discovery of novel DNA polymorphisms (Martinsohn& Ogden 2009) and the standardization of methodologies for otolith analysis. In addition to the use of state of the art analytical methods, the consortium has also explored the potential of fatty acid analysis, proteomics and gene expression tools.

While starting out as a research project, FishPopTrace set the ambitious goal to move beyond theory, and transfer novel technologies into a forensic framework for fisheries control and enforcement. Such an approach is crucial to ensure that results and information generated by the project do not remain exclusively in the academic realm but are accessible to end-users, particularly control and enforcement authorities and fishery management bodies.

1.3 SINGLE NUCLEOTIDE POLYMRPHISMS (SNPS)

1.3.1 SNP MARKERS

As suggested by the acronym, Single Nucleotide Polymorphisms (SNPs) are single base changes in a DNA sequence (Figure 2). Despite at each position of a sequence stretch any of the four possible nucleotide bases (G, A, T or C) can be present, most of SNPs are biallelic and are therefore by far less variable than microsatellite markers, where often many alleles per locus can be found.

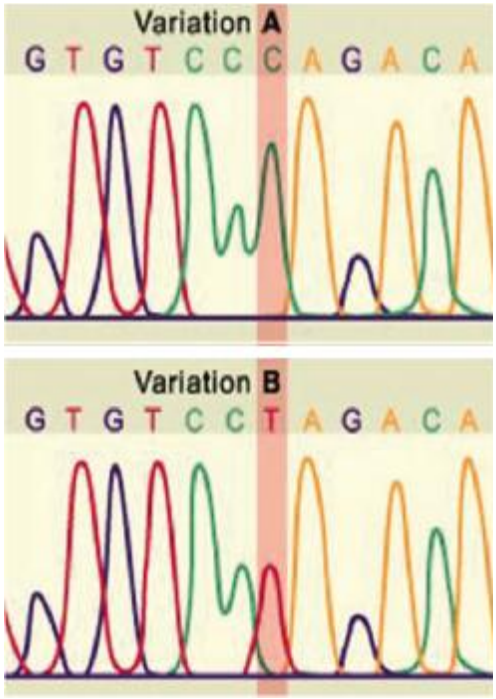


Figure 2. Illustration of a SNP revealed by the alignment of two chromatograms showing the results of DNA sequencing. Source: JRC Reference Report (Martinsohn 2011).

The lack of resolution power for each SNP marker is outweighed by their abundance, being the most common and widespread type of polymorphism in the genome (Morin *et al.* 2004; Wayne& Morin 2004). SNPs are co-dominant markers, evolve following a simple mutation model, the infinite allele model, and are likely less prone to homoplasy than microsatellites. They are rapidly becoming the marker of choice for many applications in population ecology, evolution and conservation genetics, because of the potential for higher genotyping efficiency, data quality and low-scoring error

rates, genome-wide coverage and analytical simplicity. A significant advantage respect to microsatellites lies in the high reproducibility among different laboratories: newly collected data can then be readily compared with reference data with high reliability and ease. In addition, SNPs offer the opportunity of genotyping nuclear loci when using poor quality samples, such as historical, noninvasive or otherwise degraded samples (Morin& McCarthy 2007), not always feasible with microsatellites.

The ongoing improvements in the speed, cost and accuracy of next generation sequencing (NGS) technologies (Glenn 2011; Metzker 2010), together with parallel advances in bioinformatic tools, are rapidly increasing the availability of genomic resources, thus favoring approaches of high-throughput SNP detection, especially in non-model organisms (Helyar *et al.* 2011). In addition to discovery strategies, technological improvements have also facilitated SNP data attainment, allowing to genotype large panels of SNPs in a cost-effective way.

1.3.2 EXPLORING ADAPTIVE VARIATION IN THE WILD

In the last decades fishery genetics studies has produced a wealth of data based on neutral genetic markers, contributing considerably to our understanding of structuring and connectivity among marine populations. However, neutral molecular markers do not necessarily convey information on the extent or importance of adaptive variation, which could shape fish population structure even when putative neutral genetic markers suggest weak genetic differences (Hauser& Carvalho 2008; Larsen *et al.* 2007). Indeed, the marine environment offers a mosaic of environmental dynamic conditions, potentially driving genetic changes at different spatial and temporal scale (Conover *et al.* 2006). Local adaptation represents the missing element to understand the ecological and evolutionary processes that influence biodiversity, as well as a facet of genetic diversity that should be preserved in a framework of conservation, affecting the potential of species to adapt to environmental changes (Crandall *et al.* 2000; Fraser& Bernatchez 2001; Gebremedhin *et al.* 2009).

Accordingly, there is a renewed interest in studying adaptive genetic variation in the wild, and therefore in identifying molecular genetic markers under selection. Population genomics, the analysis of a large number of loci that allow discrimination between locus specific (selection) and genome-wide effects (drift and migration) (Stinchcombe& Hoekstra 2007), have a multitude of potential and applications for elucidating natural selection in marine species (Nielsen *et al.* 2009b). In particular the genome-scan approaches allow the identification of loci that show unusual levels of differentiation respect to the average observed at large numbers of genetic markers spread throughout the genome, possibly representing signatures of selection.

Being abundant and widespread, SNPs represent ideal markers to apply this kind of approach. Genomic resources from which SNPs might be derived include Expressed Sequence Tags (EST), sequences of expressed genes, which have been identified from partial sequencing of a messenger RNA (mRNA) pool that has subsequently been reverse transcribed into cDNA (Bouck& Vision 2007). EST collections sample the gene space of an organism by providing a snapshot of the transcribed mRNA population within a given set of tissues. SNPs can be identified across populations directly from alignments of ESTs obtained from multiple individuals representing different geographical regions, using constantly refined “in silico” detection procedures.

The development of SNPs from coding sequences offers the possibility to screen polymorphisms in and around functional genes, increasing the chance of identifying loci subject to selection (Nielsen *et al.* 2009b). The growing availability of EST resources is rapidly giving access to this approach also to non-model organisms (Seeb *et al.* 2011a). Genome-scan surveys to detect SNPs possibly involved in local adaptation has been successfully applied in cod (Bradbury *et al.* 2011; Nielsen *et al.* 2009a; Poulsen *et al.* 2011), stickleback (Deagle *et al.* 2011), lake whitefish (Renaut *et al.* 2010; Renaut *et al.* 2011) and several salmonid species (Freamo *et al.* 2011; Limborg *et al.* 2011; Seeb *et al.* 2011b).

1.3.3 SNPs STATISTICAL POWER

The proportion of SNPs required to achieve equivalent power than multi-allelic loci such as microsatellites varies according to the application field.

For population structure analysis, statistical power of genetic markers to detect genetic differentiation is primarily related to the total number of independent alleles rather than the number of loci (Kalinowski 2002). For this reason, bi-allelic SNPs have significantly lower individual power for detecting population structure than microsatellites. In a study on human population structure, Liu *et al.* (2005) found that, although on average the informativeness of random microsatellites was 4 to 12 times that of random SNPs, the latter comprised the majority among the most informative markers (showing the greatest allele frequency variation among populations), leading to better inference of population structure (Liu *et al.* 2005). However, estimating the statistical power of SNP markers to detect population structure between two putative populations of whales, Morin et al. (2009) found that 80 or more SNPs may be required to detect demographic independence when F_{ST}< 0.005, and increasing the sample size rather than the number of SNP loci is likely to result in greater gains in statistical power (Morin *et al.* 2009).

A different evaluation concern the power of SNPs in the assignment of individuals to population of origin based on their multilocus genotypes (Manel *et al.* 2005). Individual genetic assignment is an important tool in the management of domesticated and wild genetic resources, and represents the primary investigation tool in a forensic framework (Ogden 2008). The assignment success does not depend on the number of independent alleles, as suggested by studies on chum salmon (Smith& Seeb 2008), Atlantic salmon (Glover *et al.* 2010) and sockeye salmon (Hauser *et al.* 2011) populations, in which SNPs have been demonstrated to outperform microsatellites in assigning individuals to population of origin, considering the lower number of alleles. Moreover, there is increasing evidence that the use of "outlier" SNPs, markers identified as potentially under diversifying selection from genome-scan analysis, may contribute to increase the accuracy of

assignment tests (Helyar *et al.* 2011). In a recent study, Freamo et al. (2011) obtained a better performance in assigning Atlantic salmon individuals to inner or outer Bay of Fundy metapopulations using a subset of 14 non-neutral SNPs, respect to a subset of 67 neutral SNPs (Freamo *et al.* 2011). For applications such as individual assignment, there are many advantages in using a reduced panel of markers that have been identified while maximizing the power available, for example the reduction in costs, time and computational demands. As loci may be have different discriminatory power, it may be desirable to create "minimal panels with maximum power", selected to test specific alternative hypotheses in relation to individual assignment. Although the principle of selecting highly discriminating loci may also be applied to other classes of marker, the continued technological advances in SNP detection genotyping platforms will likely favor this strategy for SNP markers (Glover *et al.* 2010).

CHAPTER 2

TARGET SPECIES: THE EUROPEAN HAKE

2.1 TAXONOMY AND DESCRIPTION

The European hake (Merluccius merluccius, Linnaeus, 1758) is an important species belonging to the Family Merlucciidae (see Figure 3 for a complete taxonomy), which includes 4 Genera and 18 Species: there are 13 recognised species of the genus Merluccius, which makes it the most diverse Genus within the Family. The order Gadiformes, to which the Family Merlucciidae belongs, includes some of the most important commercial fishes in the world: only cods, hakes, and haddocks, gadiform fishes account for approximately 12% of the World's total marine fish catch (FAO 2010).

Taxonomy:

Phylum	Chordata
Class	Actinopterygii
Order	Gadiformes
Suborder	Gadoidei
Family	Merlucciidae
Subfamily	Merlucciinae
Genus	*Merluccius*
Species	*M. merluccius*

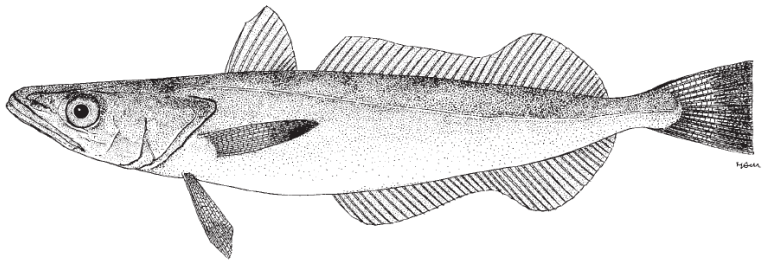


Figure 3. ***Merluccius merluccius***

The European hake is characterized by a long and slender body compared with other hake species. The head is large, being about 25–30% of the body length; the snout and upper jaw are about 30–35% and 48–54% of the head length, respectively. The ocular diameter is around 16–21% and the inter-orbital space 22–28% of the head length. The first dorsal fin contains 1 spine and 7/10 rays, whilst the second dorsal and anal fins have 36–40 rays. The tip of the pectoral fins reaches the

anal opening in young fish, below the 20 cm standard length, but not in adults. The margin of the caudal fin is normally truncated, but becomes forked with growth. The lateral line contains 127–156 small scales. The total number of vertebrae varies from 49 to 54. The body color of the European hake is grey, in general, becoming lighter on the sides and silvery white on the belly.

2.2 GEOGRAPHIC DISTRIBUTION, HABITAT AND ECOLOGY

The European hake is widely distributed throughout the north-east Atlantic, from Norway in the north to the Mauritian coast in the south, where it is quite rare, and throughout the Mediterranean and Black Sea (Figure 4). It is more abundant from the British Isles to the south of Spain (Casey et al. 1994).

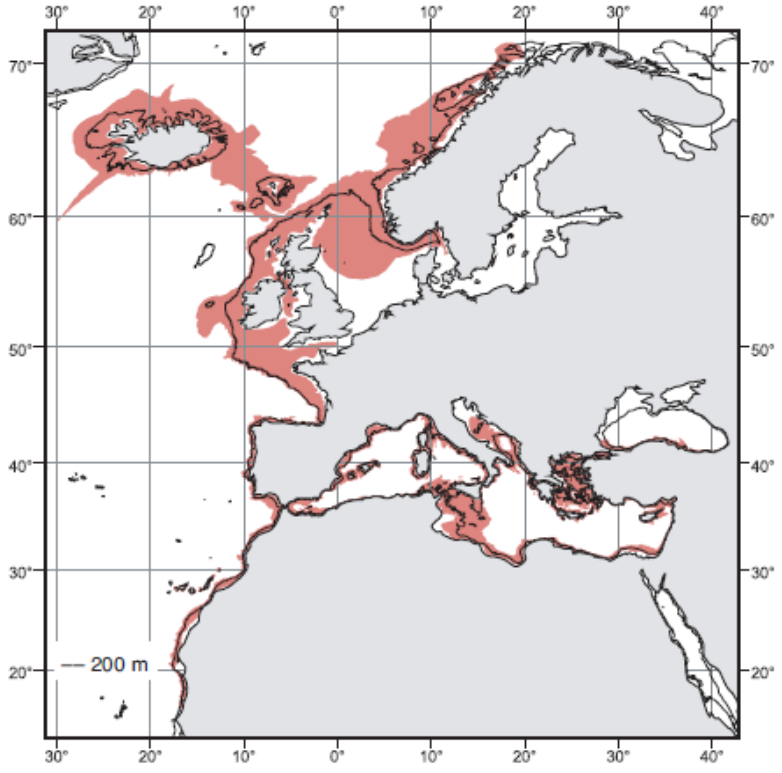


Figure 4. Geographic distribution of European hake. Source: FAO Species Catalogue.

It is a demersal and benthopelagic species, found mainly between 70 and 370 m depth, although it also occurs in inshore waters (30 m) and down to depths of 1000 m (Lloris *et al.* 2005). It lives on

muddy or mud-sand grounds on the continental shelf and slope. Juvenile individuals perform daily vertical migrations, living close to the bottom during the daytime and moving up and down the water column at night, possibly in relation to feeding behavior (Bozzano *et al.* 2005; Carpentieri *et al.* 2005). Bathymetric preferences at different ages has also been observed (Abella *et al.* 2005; Bartolino *et al.* 2008b; Recasens *et al.* 1998): age 0 fish are distributed almost exclusively at depths between 100 and 250 m, where most of nursery grounds are located, intermediate ages (mainly 1-year-old individuals) are concentrated in more shallow waters while large adult individuals are found on the shelf slope. Different bathymetric distribution is likely related to a change in diet during different stages of life cycle (Bartolino *et al.* 2008b). During the first year of life juvenile hakes feed on small crustaceans (mainly euphausiids); then, moving from nursery areas on the shelf-break and upper slope to the middle shelf, the diet change to small pelagic fish prey which form schools on the continental shelf, such as horse mackerel (*Trachurus trachurus*), anchovy (*Engraulis encrasicholus*) and sardines (*Sardina pilchardus*); the diet of larger adult hakes is primarily composed of larger demersal preys, like blue whiting (*Micromesistius poutassou*) (Carpentieri *et al.* 2005; Mahe *et al.* 2007). Cannibalism of juveniles has been also observed in large hakes (>30-36 cm TL), probably influenced by the abundance of juvenile hakes and the overlap between distribution patterns of juveniles and adults.

In Atlantic the European hake grows at higher rates and reaches larger size than in Mediterranean: studies on age and growth rate estimations largely based on otolith microstructure analysis, but also using conventional tagging, have indicated that at the end of the first year individuals reach a size in the range of 15-18 cm in Mediterranean (Arneri& Morales-Nin 2000; Belcari *et al.* 2006; Mellon-Duval *et al.* 2009; Morales-Nin& Aldebert 1997) and around 23-25 cm in Atlantic (Kacher& Amara 2005; Piñeiro *et al.* 2008). The interplay of different factors, including temperature, greater production in the area, and genetics, may contribute to explain the higher growth rate observed in Atlantic.

It is widely recognized that males grow faster than females up to a specific age, probably corresponding to the onset of maturity (Lucio *et al.* 2000; Recasens *et al.* 1998), and that from that age onwards the growth rate largely decrease in males whereas increase in females. However, based on a conventional tagging experiment, Mellon-Duval *et al.* (2009) did not observe any statistical difference in growth rate between sexes for juvenile fish in the Gulf of Lion, while evidence of lower growth rates in males than females has been found in adult fish: this observation is consistent with sexual dimorphism for size-at-first-maturity in the Gulf of Lion, that is 28 cm and 38 cm for males and females, respectively. Nevertheless adult females reach a larger size and grow older than males and consequently the sex ratio is skewed towards females in the largest length classes (Casey *et al.* 1994).

2.3 REPRODUCTION

European hake is a multiple batch-spawner with indeterminate fecundity and asynchronous ovarian organization (El Habouz *et al.* 2011; Murua& Motos 2006; Recasens *et al.* 2008): a continuum of oocyte at all stages of development is present in the ovary of mature female hake during the whole reproductive season; potential annual fecundity is not fixed before the onset of spawning, because previtellogenic oocytes can develop and be recruited into the yolked oocyte stock at any time during the spawning season; eggs are released in several batches over a protracted period that can last days or even months.

Males reach maturity at a smaller size and at an earlier age than females. In Galician and Bay of Biscay waters males mature at around 35 cm (Lm50, total length at which 50% of the fish have reached maturity), whereas females mature between 45 and 50 cm total lengths (Domínguez-Petit *et al.* 2008; Lucio *et al.* 2000; Piñeiro& Saínza 2003), although transitory declines in the size at maturity have been observed, probably in relation to fishing mortality and the age diversity of the stock (Domínguez-Petit *et al.* 2008). In Mediterranean size at which hake reach maturity is smaller and estimates of Lm50 vary according to different geographic areas from around 21 cm to 28 cm for

males and from 30 to 46 for females (Biagi *et al.* 1995; Bouaziz *et al.* 1998; Papaconstantinou& Stergiou 1995; Recasens *et al.* 2008; Recasens *et al.* 1998). A more controversial issue is the assessment of the age at first maturity, which is strictly correlated to the accuracy of the growth rate estimation and the precision and repeatability of methodology used (de Pontual *et al.* 2006; Mellon-Duval *et al.* 2009).

Spawning females are commonly present all year round: however, the levels of egg production varies depending upon the month of the year and according to different geographic areas. In North-East Atlantic, the peak spawning time follows a latitudinal variation from southern to northern waters (Casey *et al.* 1994), showing a northward displacement as the season advance, and is restricted to the first half of the year: in the Bay of Biscay, spawning fraction and batch fecundity were found to be higher between January and May, afterwards the main spawning area is centered on the Celtic Sea and/or extended towards the Porcupine area, where a peak between April and June has been observed (Alvarez *et al.* 2004; Alvarez *et al.* 2001b). In Mediterranean temporal variation of spawning peaks extends throughout the year: a more pronounced spawning peak has been observed during autumn and winter in the Catalan Sea and in the Gulf of Lion (Morales-Nin& Aldebert 1997; Recasens *et al.* 1998), summer and winter in North and Middle Adriatic and Algeria (Arneri& Morales-Nin 2000; Bouaziz *et al.* 1998; El Habouz *et al.* 2011), in the months of February-May and September in North Tyrrhenian (Biagi *et al.* 1995) and between May and September in the Aegean Sea (Papaconstantinou& Stergiou 1995).

The eggs of European hake are pelagic and mainly found in the upper 200 m of the water column over the shelf break, although, depending upon environmental conditions (upwelling, temperature), the depth distribution can be different. In North-East Atlantic eggs are mainly distributed in water at temperatures between 10 and 12.5 °C, the optimum temperature range for the spawning activity of European hake, although they can be found at temperatures up to 15 °C (Alvarez *et al.* 2001a). After the hatching hake small larvae (<8 mm) are still found near the spawning grounds, over the shelf

break, while large larvae (>8 mm) undergo a coastward displacement over the continental shelf, towards the nursery areas (Alvarez *et al.* 2001a). The main nursery areas have been identified over the shelf along the French coast (Le Grand Vasiére), the shelf in the Celtic Sea and off the west coast of Ireland (Casey *et al.* 1994). Transportation of early life stages from the spawning grounds to juvenile recruitment areas is a critical process in the early life history of European hake, its reproductive strategy is probably adapted to match the principal environmental and oceanic events. Larvae have a pelagic existence until they settle on the seabed, around 40-50 days after hatching. In Mediterranean most nursery ground are located between 100 m and 200m (Orsi Relini *et al.* 1989; Orsi Relini *et al.* 2002) They have been identified around throughout the Mediterranean, specifically in the North Adriatic (Pomo/Jabuka Pit) (AdriaMed 2006), around the Gargano peninsula in the South Adriatic Sea (Carlucci *et al.* 2009), in the Tyrrhenian and Ligurian Seas (Abella *et al.* 2005; Colloca *et al.* 2009), where larvae and juveniles of hake are stably retained in areas of relatively high production, in the Strait of Sicily (Abella *et al.* 2008; Fiorentino *et al.* 2003), in the Gulf of Lions (Orsi Relini *et al.* 2002), in well defined areas of the Aegean and Ionian Seas (Papaconstantinou& Stergiou 1995; Tserpes *et al.* 2008). Water temperature plays an important role in egg development and larval growth and survival. Bartolino *et al.* (Bartolino *et al.* 2008a) found that thermal anomalies in summer 2003, characterized by high peaks in water temperature, had a negative effect on the abundance of recruits in autumn in the Tyrrhenian Sea, possibly due to the combination of different mechanisms: increasing of mortality rates of eggs and larvae during higher temperature peaks, water stratification and relative lower phytoplankton production, modification of water circulation systems involved in transport and retention processes of hake larvae in the nursery areas.

2.4 CONSERVATION AND MANAGEMENT

In a management framework, Atlantic and Mediterranean European hake stocks are assessed and regulated separately. This resolution has been established owning to the presence of a geographic barrier, the Strait of Gibraltar, but it is also well supported by a wealth of scientific data: Atlantic and Mediterranean hake shows differences in growth rates and maximum size, and diversity in the vertebrae count has also been observed (Jones 1974); studies based on "biological tags", revealed that significant differences exist between Mediterranean and Atlantic in the relative proportions of the parasitic nematode *Anisakis* identified in hake samples (Mattiucci *et al.* 2004); furthermore, genetic data obtained using allozymes and microsatellite markers (Castillo *et al.* 2005; Castillo *et al.* 2004; Cimmaruta *et al.* 2005; Lundy *et al.* 1999; Roldan *et al.* 1998) and otolith chemistry analysis (Swan *et al.* 2006) confirmed the differentiation between the two macroareas.

2.4.1 ATLANTIC

The Atlantic stock, which represents the main source of production of hake in Europe, is monitored from ICES (International Council for the Exploration of the Sea), that recognizes a further separation between a Northern stock (including ICES Division IIIa, Sub-areas II, IV, VI and VII and Divisions VIIIa, b, d) and the Southern stock (including ICES Divisions VIIIc and IXa) (Figure 5).

This delimitation is based on the presence of a supposed hydrographic barrier, the Cape Breton Canyon, which separates French and Spanish waters in the Bay of Biscay (Casey *et al.* 1994). However, a strong scientific support to the suitability of this management strategy is lacking, as several genetic studies have not corroborated the separation between Atlantic populations northward and southward the stocks boundary (Castillo *et al.* 2005; Lundy *et al.* 1999; Pita *et al.* 2011; Roldan *et al.* 1998). Given the importance of stock delimitation in a proper management, the ICES Working Group on the assessment of southern shelf stocks of hake, monk and megrim agreed on the need for further studies, supporting tagging experiments and genetic studies to attain further

valuable information on hake population dynamics (ICES 2008). The stocks are managed by TAC and quota regulations. In addition, technical measures concerning minimum length at landing (27 cm , except in Division IIIa where a size of 30 cm is recognized), minimum mesh size, effort limitation, seasonal restrictions and closed area (EC No. 850/98) are in place.

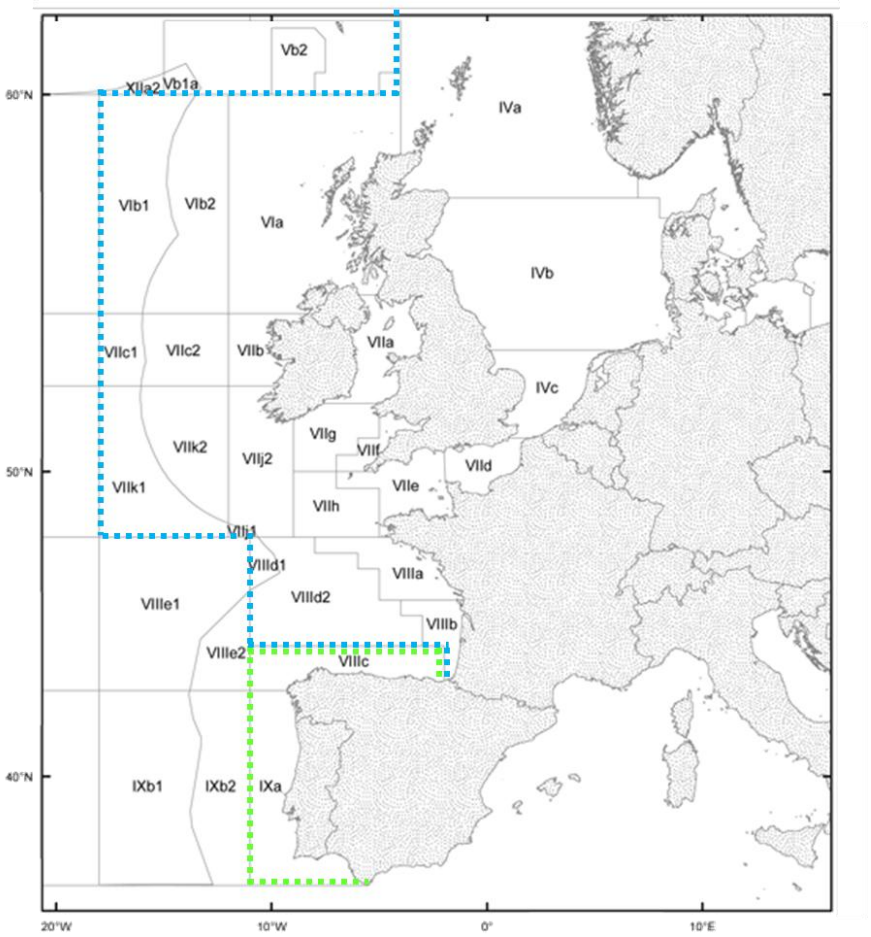


Figure 5. Outline of Northern (blue) and Southern (green) Atlantic stocks of European hake. ICES division areas are indicated in grey.

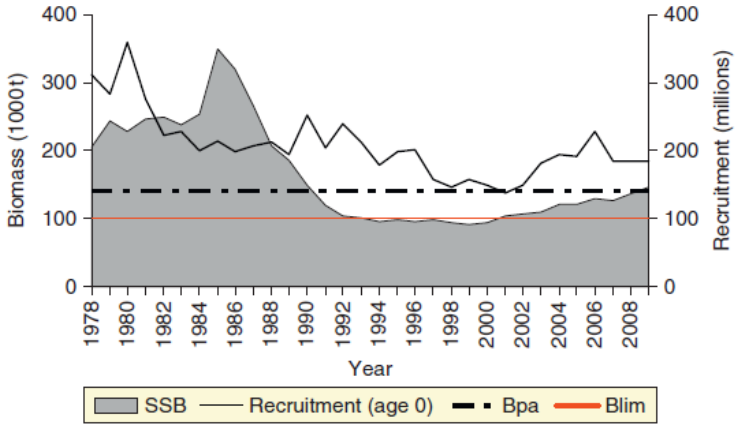
During the last fifty years, global captures in the North East Atlantic area have continuously decreased from 130.000 tonnes in 1960 to 40.000 tonnes in 1997. This decrease has affected both the North and the South sub-stocks.

The Northern stock of European hake supports a major commercial fishery in Atlantic European waters. The stock size underwent an abrupt decline during the late 1990s; the present level appears

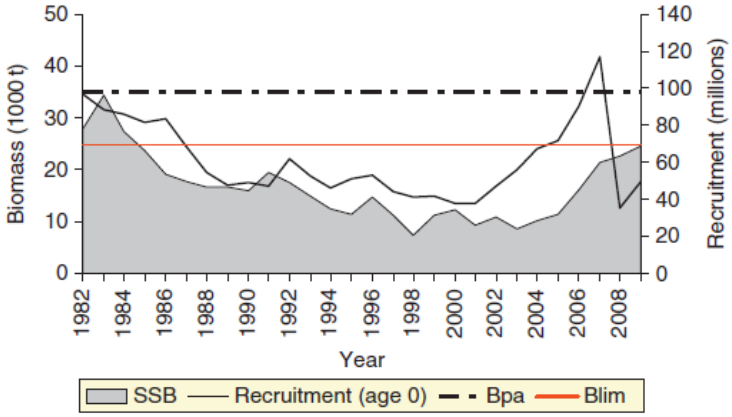
to be only 50% of the level in the 1970s. Based upon the most recent estimates of SSB (Spawning Stock Biomass) and fishing mortality, the ICES considers that the Northern stock has a full reproductive capacity and that it is harvested sustainably. Although the stock currently lies within the safe biological limits, at the beginning of the 1990s the spawning biomass had decreased below the B_{pa} (the biomass below which the stock would be regarded as potentially depleted or over-fished) to around the B_{lim} (limit spawning stock biomass, below which recruitment is impaired or the dynamics of the stock are unknown) until 2001. However, since 2001, the SSB increased and is presently estimated to be just above the B_{pa} (ICES 2008) (Figure 6A). A recovery plan implemented in 2004, under EC Reg. No. 811/2004, has successfully helped the Northern hake stock to recover from almost collapse: during the last few years the SSB has been brought back above the precautionary target level (140.000 tonnes).

The Southern stock supports the commercial fishery on the Atlantic coast of the Iberian Peninsula, and it is especially important for Spanish and Portuguese fishing fleets. The SSB of the stock declined sharply and continuously since 1983 to the lowest observed level of 7.300 tonnes in 1998. The SSB remained relatively stable around 10.000 tonnes from 1999 to 2005 and it increased to around 20.000 tonnes and 25.000 tonnes in 2008 and 2009, respectively. Based on the most recent assessment of this species, the southern hake stock is suffering reduced reproductive capacity and is at risk of being harvested unsustainably (ICES 2008). The stock is considered outside the safe biological limits and the spawning biomass had decreased below the precautionary approach level (B_{pa} = 35.000 tonnes) in 1985, and it has remained at that level ever since (Figure 6B). Moreover, fishing mortality has increased in recent years. Therefore a recovery plan was introduced in December 2005 (see Council Regulation No. 2166/2005), with the objective of bringing the spawning stock biomass above 35.000 tonnes within 10 years. However, a recent report published by the European Commission reveals that this plan has not been effective in reducing fishing mortality and rebuilding the spawning stock biomass to the desired levels. Despite the fishing mortality rate has

remained stable after the implementation of the plan, recent large increases in spawning stock biomass and recruitment have been observed, possibly related to a migration of mixed assemblages of hake from feeding grounds of the Northern stock to spawning grounds in the Bay of Biscay (Pita et al. 2011). This hypothesis would confirm that no biologically based distinctions exist between hake stocks of the North Atlantic.



A.



B.

Figure 6. Spawning stock biomass (SSB) and recruitment (age 0) as well as the biological reference points of the Northern (A) and Southern (B) stocks of European hake. Bpa is the biomass below which the stock is considered potentially depleted or over-fished but, in spite of year-to-year fluctuations, above the Blim. Blim is the limit spawning stock biomass, below which recruitment is impaired or the dynamics of the stock are unknown. (Murua& Michael 2010).

2.4.2 MEDITERRANEAN

European hake represents the most abundant demersal species in Mediterranean. The stock is monitored by the General Fisheries Commission for the Mediterranean (GFCM), that coordinates the collection of scientific and fishery data from different geographical sub-areas (GSA) (Figure 7), in order to evaluate the abundance of the resource, the level of exploitation and the state of the fisheries; supported by the Scientific Advisory Committee (SAC), the Commission consequently recommend the adoption of appropriate measures for a sustainable exploitation.

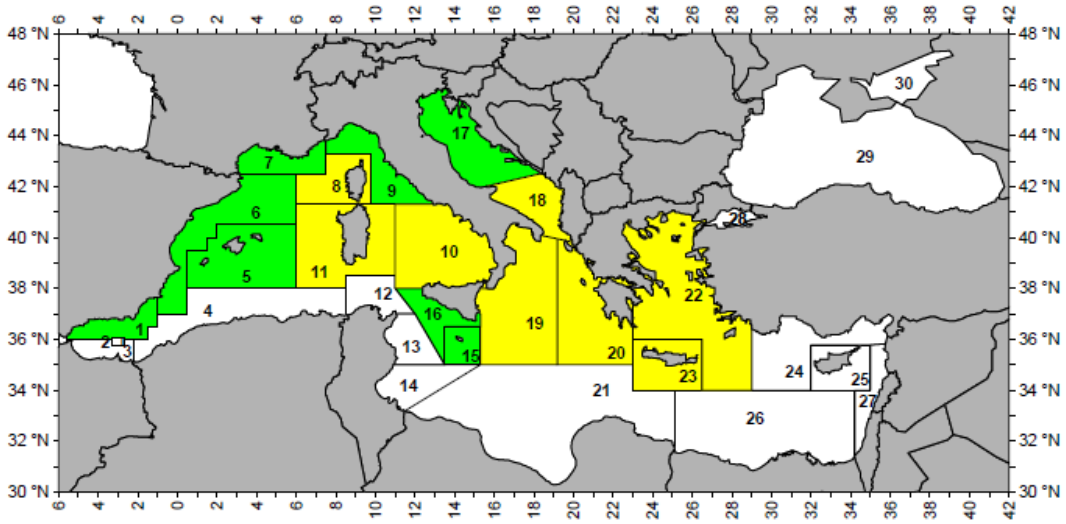


Figure 7. Subdivision of the Mediterranean Sea in GSA. Colors denote different levels of data availability and state of assessments of European hake: green denotes an analytical assessment of exploitation rates and other stock parameters based on commercial catch data available, yellow denotes an assessment based on relative catch rates derived from commercial fisheries or scientific surveys, blank denotes no assessment undertaken as no data available (Cheilari& Rätz 2009).

However, an analytical assessment of exploitation rates and other stock parameters is not available for all Mediterranean GSA, and in some areas (particularly fishing areas along the African coasts and in the Eastern Mediterranean) data are lacking (Cheilari& Rätz 2009).

Since 1994 a number of common conservation regulations for the Mediterranean have imposed restrictive norms in terms of the fishing methods used, in particular the minimum mesh size for bottom trawling nets (40 mm) and minimum size of the fish that can be sold, in the case of hake set at 20 cm (EC 1967/06). In Mediterranean most of the catches occur by means of trawl net, but also

long-line and gillnet. While the bulk of trawl catches consists of immature hakes (<20 cm of total length), artisanal gears mainly affect the adult fraction of population.

The SAC has recently indicated as overexploited the stock status of European hake overall the Mediterranean Sea, particularly in GSA 05 (Balearic islands), GSA 06 (Northern Spain), GSA 07 (Gulf of Lions), GSA 09 (Ligurian and North Tyrrhenian), GSA 18 (Southern Adriatic) (GFCM 2011). According to FAO statistical data, catches have increased in Mediterranean until 1995, and then abruptly declined to less than the half until 2002 (52 000t to 21 000t) (Figure 8).

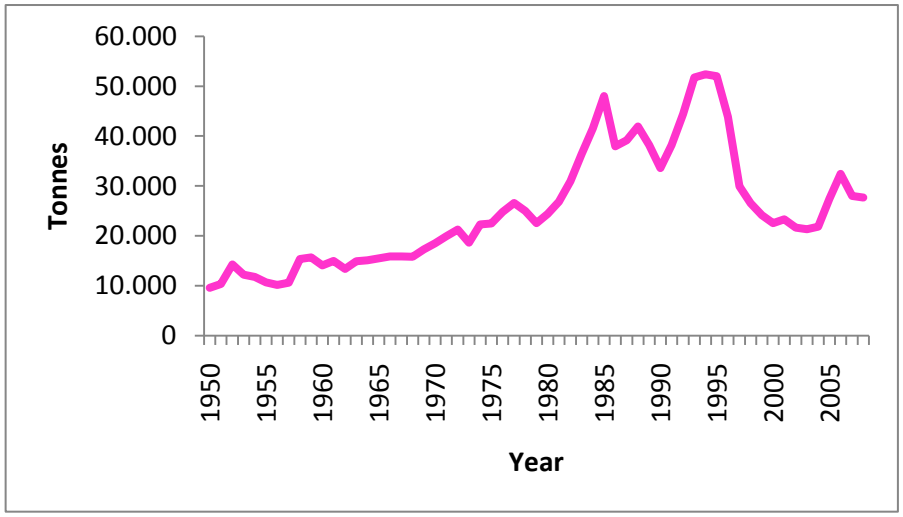


Figure 8. Trend of European hake catches in the Mediterranean Sea, according to FAO data.

CHAPTER 3

STATE OF THE ART AND RESEARCH AIMS

3.1 POPULATION GENETIC STRUCTURE OF EUROPEAN HAKE

In literature the study of European hake genetic structure has been mainly focused on Atlantic populations, particularly on the issue concerning the boundary between Northern and Southern stocks, while only a single extensive survey across the whole Mediterranean Sea has been carried out by Cimmaruta *et. al*. (2005). A limited set of available genetic tools were used, consisting mainly of 6 microsatellite and 21 allozyme markers.

Two major studies have analyzed allozymes genetic variation across the Atlantic and Mediterranean Sea (Cimmaruta *et al.* 2005; Roldan *et al.* 1998). They both found a major subdivision between basins and the evidence of an ongoing gene flow from adjacent Atlantic to proximate Mediterranean populations, across the Gibraltar Strait. However, examining a higher number of Mediterranean samples, Cimmaruta *et al.* (2005) observed a further steep genetic discontinuity across the Almeria-Oran front, and located more precisely the transition zone in the Alboran Sea, rather than in the Gibraltar Strait. Based on six Atlantic and four Mediterranean samples, Roldan *et al.* (1998) found signals of population structure within both regions, despite the genetic pattern within Atlantic did not corroborate the subdivision between Northern and Southern stocks. Interestingly, both studies have detected higher levels of differentiations at one locus, the Glyceraldehyde-3-phosate dehydrogenase (Gapdh), strongly differentiating the two basins (Roldan *et al.* 1998) and showing a longitudinal cline in allele frequencies in Mediterranean (Cimmaruta *et al.* 2005). In addition, Cimmaruta *et al.* (2005) observed a strong correlation between allele frequencies

variation at this locus across 4 Atlantic and 11 Mediterranean populations and the salinity values both at the surface and at 320 m depth, suggesting a possible role for environmental factors in maintaining the genetic differentiation at Gapdh.

Most of population genetic studies focusing on European hake were based on the analysis of putative neutral microsatellites. Lundy *et al.* (1999) carried out the first large-scale population study based on 6 microsatellite loci, analyzing a total of six samples, four from Atlantic and two from Mediterranean. Their results confirmed the main population subdivision between Mediterranean and Atlantic samples, previously revealed by allozymes. In addition, they did not find significant differentiation between two samples from the Mediterranean Sea (Tunisia and Adriatic Sea), while a signal of population structuring was evident within Atlantic (Lundy *et al.* 1999). However, similarly to Roldan *et al.*, they did not notice any genetic difference between samples from southern Bay of Biscay and Celtic Sea, located in different management areas, supporting the inadequacy of the boundary between Atlantic stocks. In a following work they analyzed the temporal stability of genetic variation patterns within the Bay of Biscay (Lundy *et al.* 2000): they found very high levels of inter-annual differentiation, suggesting that in Atlantic population structure may undergone short-term changes, possibly due to large variance in reproductive success.

While the subdivision between Atlantic and Mediterranean was constantly confirmed from subsequent studies (Castillo *et al.* 2005; Castillo *et al.* 2004; Pita *et al.* 2010), more controversial is the existence of sub-structuring within basins. Unlike Lundy *et al.*, Castillo (2004) found significant genetic differences at 5 microsatellites among three Mediterranean samples (Western Mediterranean, Jonian Sea, Aegean Sea), and between locations from Northern and Southern Atlantic stocks, for which no differentiation have been previously found, specifically between southern Bay of Biscay and Celtic Sea samples (Castillo *et al.* 2004). Evidence of small-scale genetic structure was additionally detected within ICES subareas VIIIc, suggesting that populations could be more subdivided than currently considered by ICES for purposes of stock management. However, in a

subsequent study analyzing eight locations around the Iberian Peninsula, the same authors found no differentiation between samples caught in the VIIIa,b,d and in the VIIIc ICES areas, advocating the need to reconsider the boundary between Northern and Southern stocks (Castillo *et al.* 2005). In a recent study based on 5 microsatellites, Pita et al. explored the pattern of connectivity between the two Atlantic European hake stocks analyzing fifty-two samples comprising 1123 mature adults. The analysis revealed large genetic homogeneity, and the occurrence of inter-annual gene flow between the core grounds of the northern stock (Porcupine and Great Sole) and the Northern Iberian grounds. The scenario of high connectivity between stocks have been proposed as the possible explanation of the large recruitment observed in the southern stock after 2003 despite the dramatic low levels of spawning stock biomass. The abundance of recruits overwhelming that of apparent reproductors fits the pattern of high gene flow between stocks found by the authors (Pita *et al.* 2011).

Despite the potential of microsatellites to reveal low but statistically significant genetic differences among European hake populations, Pita *et al.* (2010) proved their shortcoming in assignment tests. Based on the analysis of 27 populations genotyped at 5 microsatellites, they failed to assign individuals, according to their multilocus genotype, to the basin of origin (Atlantic or Mediterranean), indicating that these two geographical stocks can not be reliably identified from each other neither for fishery forensics nor for commercial traceability. (Pita *et al.* 2010).

3.2 RESEARCH AIMS

The research of my PhD was aimed at developing and analyzing novel genetic tools for the European hake (*Merluccius merluccius*) within the framework of FishPopTrace, a European project focused on traceability of fish populations and products, introduced in chapter 1 and 2.

The progress and results of the work can be outlined according to three main steps:

- *Development of novel molecular markers*: a large number of Single Nucleotide Polymorphisms was developed from a massive collection of Expressed Sequence Tags, obtained by high-throughput sequencing;
- *Analysis of population genetic structure*: genome-scan approaches were applied to identify polymorphisms on genes potentially under-selection, and comparative analysis were carried out to disentangle the effects of putative neutral and adaptive evolutionary forces on European hake populations genetic structure;
- *Traceability applications*: a minimum panel of SNP markers showing maximum discriminatory power was selected and applied to a traceability scenario aiming at identifying the basin (and hence the stock) of origin, Atlantic or Mediterranean, of individual fish.

Methodologies and results related to each research topic are illustrated in chapter 4, 5 and 6. While chapter 4 corresponds to the printed copy of the published paper, chapters 5 is given as manuscript in preparation to be submitted and chapter 6 is the copy of the paper currently under revision to the journal Nature Communications.

CHAPTER 4

NOVEL TOOLS FOR CONSERVATION GENOMICS: COMPARING TWO HIGH-THROUGHPUT APPROACHES FOR SNP DISCOVERY IN THE TRANSCRIPTOME OF THE EUROPEAN HAKE

OPEN ACCESS Freely available online

PLoS one

Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake

Ilaria Milano[1]*, Massimiliano Babbucci[2], Frank Panitz[3], Rob Ogden[4], Rasmus O. Nielsen[3], Martin I. Taylor[5], Sarah J. Helyar[5], Gary R. Carvalho[5], Montserrat Espiñeira[6], Miroslava Atanassova[6], Fausto Tinti[1], Gregory E. Maes[7], Tomaso Patarnello[2], FishPopTrace Consortium, Luca Bargelloni[2]

1 Department of Experimental and Evolutionary Biology, University of Bologna, Bologna, Italy, **2** Department of Public Health, Comparative Pathology, and Veterinary Hygiene, University of Padova, Legnaro, Italy, **3** Department of Molecular Biology and Genetics, Faculty of Science and Technology, Aarhus University, Tjele, Denmark, **4** TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Edinburgh, United Kingdom, **5** Molecular Ecology and Fisheries Genetics Laboratory (MEFGL), School of Biological Sciences, Environment Centre Wales, University of Bangor, Bangor, Gwynedd, United Kingdom, **6** ANFACO-CECOPECSA, Vigo, Spain, **7** Laboratory of Animal Diversity and Systematics, Katholieke Universiteit Leuven, Leuven, Belgium

Abstract

The growing accessibility to genomic resources using next-generation sequencing (NGS) technologies has revolutionized the application of molecular genetic tools to ecology and evolutionary studies in non-model organisms. Here we present the case study of the European hake (*Merluccius merluccius*), one of the most important demersal resources of European fisheries. Two sequencing platforms, the Roche 454 FLX (454) and the Illumina Genome Analyzer (GAII), were used for Single Nucleotide Polymorphisms (SNPs) discovery in the hake muscle transcriptome. *De novo* transcriptome assembly into unique contigs, annotation, and *in silico* SNP detection were carried out in parallel for 454 and GAII sequence data. High-throughput genotyping using the Illumina GoldenGate assay was performed for validating 1,536 putative SNPs. Validation results were analysed to compare the performances of 454 and GAII methods and to evaluate the role of several variables (*e.g.* sequencing depth, intron-exon structure, sequence quality and annotation). Despite well-known differences in sequence length and throughput, the two approaches showed similar assay conversion rates (approximately 43%) and percentages of polymorphic loci (67.5% and 63.3% for GAII and 454, respectively). Both NGS platforms therefore demonstrated to be suitable for large scale identification of SNPs in transcribed regions of non-model species, although the lack of a reference genome profoundly affects the genotyping success rate. The overall efficiency, however, can be improved using strict quality and filtering criteria for SNP selection (sequence quality, intron-exon structure, target region score).

Citation: Milano I, Babbucci M, Panitz F, Ogden R, Nielsen RO, et al. (2011) Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake. PLoS ONE 6(11): e28008. doi:10.1371/journal.pone.0028008

Editor: Zhanjiang Liu, Auburn University, United States of America

Received June 29, 2011; **Accepted** October 29, 2011; **Published** November 22, 2011

Funding: The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° KBBE-212399 (FishPopTrace). Financial support for GAII sequencing and SNP discovery was provided by the MerSnip project funded by the Joint Research Council (JRC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ilaria.milano2@unibo.it

Introduction

The European hake (*Merluccius merluccius*, Linnaeus, 1758; Merlucciidae, Actinopterygii) is a widely distributed species inhabiting the North-Eastern Atlantic Ocean and the Mediterranean Sea, whose respective stocks are managed by the International Council for the Exploration of the Seas (ICES) and the General Fisheries Commission for the Mediterranean (GFCM). It represents one of the most important demersal target species of Western European and Mediterranean commercial fisheries [1,2]. According to the FAO report on 2008 fishery statistics [3], European hake ranks first within Mediterranean and among the ten top demersal fish species in North-East Atlantic in terms of catches, and at present stocks are under heavy exploitation and rebuilding action plans [4,5]. Despite its relevance as a fishery resource and growing concerns on fishery stock sustainability, population structure and management units of European hake are only roughly defined and poorly supported by fishery-independent data, such as life-history traits and population genetics. At present, three large stock units of European hake have been recognized in a management framework, two in the North-Eastern Atlantic (*i.e.* the Northern and Southern stocks which extend northward and southward to the Cape Breton submarine canyon in the Bay of Biscay, respectively) and one in the whole Mediterranean, from the Gibraltar Strait to the Levantine Sea. Population genetic structure assessed at six potentially neutral microsatellite loci supported a distinction between Atlantic and Mediterranean stocks [6,7], while the high connectivity and temporal variation among Atlantic populations [6–10] questioned the separation between Northern and Southern Atlantic stocks [11,12]. However, the subtle genetic differentiation among population samples at the neutral loci revealed a relatively weak population structure and

prevented the robust assignment of individual fish to the stock of origin [7]. On the other hand, genetic loci potentially under selection might be more efficient in detecting locally adapted populations and associated potential management units for conservation in marine fish [13,14]. In European hake populations a significant correlation between the allele frequency variation at the glyceraldehyde-3-phosphate dehydrogenase locus and the latitudinal gradient in surface salinity has been previously detected [15]. Identification of other genetic loci potentially under selection might provide promising tools to resolve population structure and traceability of this species at a smaller spatial scale.

Recently, major technological advances have opened innovative opportunities in the application of molecular genetic tools to study evolutionary processes in natural populations of marine species [14,16]. The emergent field of population genomics [17] has shown the potential of exploring genetic variation at a genome-wide scale and offered the opportunity to gain insights into the evolutionary mechanisms of adaptive divergence in the wild [18–20]. The growing interest in developing genomic resources is due mainly to the rapid development of next-generation sequencing (NGS) technologies (see review by Metzker [21]), allowing the production of massive volumes of data at relatively modest and decreasing costs compared to the traditional Sanger sequencing method. The opportunity to obtain extremely large collections of expressed sequence tags (ESTs) at reduced cost, potentially provides an unprecedented trove of genetic markers located in functionally relevant regions of the genome [22–26]. Although EST data mining has mainly been used so far to identify microsatellite loci, the application of single nucleotide polymorphisms (SNPs) is rapidly catching up [27], because they are the most abundant and widespread genetic variants in the eukaryote genome, with great potential in ecological and evolutionary studies [28,29]. Compared to microsatellites, SNPs show lower genotyping error rates, higher data quality and genomic coverage and low probability of homoplasy [28], with the further advantage of easier "portability" of genotypic data across laboratories. A major benefit of utilizing SNP markers associated with transcribed regions of the genome concerns the prospect of identifying outlier genetic markers, showing significantly increased or decreased differentiation among populations compared to neutral expectations [17,30]. Outlier loci are presumed to be either directly linked or in tight linkage disequilibrium with loci subject to natural selection [19,31]. Scaling up the number of available EST-linked SNPs to hundreds of thousands of markers extends the genome-coverage, thereby increasing the probability of identifying loci under selection, and associated insights into population structuring and its determinants [32].

The growing accessibility to high-throughput sequencing technologies and the concurrent development of innovative bioinformatics tools has enabled the application of wide-scale SNP discovery based on transcriptome sequencing even in species for which genomic resources are still limited or absent [25,26,33–37]. Recently, several studies concerning fish species relevant to fishery and aquaculture, such as rainbow trout [38], lake sturgeon [39], lake whitefish [40], sockeye salmon [41], catfish [42] and turbot [43] have successfully used this approach. However, experimental evidence on large scale validation of *in silico*-identified SNPs is still limited and restricted to EST data sets obtained with traditional Sanger sequencing [44,45].

The present study reports the muscle transcriptome characterization and the discovery and validation of a large set of SNP loci for European hake, based on NGS technologies. Existing NGS technologies provide a variety of approaches for transcriptome characterization, although the two most popular platforms are the Roche 454 FLX [46] and the Illumina Genome Analyzer [47,48]. The Roche 454 FLX (hereafter 454) is capable of generating one hundred million nucleotides per run, producing sequences of length approaching 400 base pairs (bp) with an average substitution error rate that is relatively low with respect to other sequencing technologies. The alternative leading sequencing system, the Illumina Genome Analyzer (hereafter GAII) produces shorter reads, currently up to 150 bp, with higher average substitution error rates but much higher throughput and lower costs [48]. As longer reads facilitate *de novo* assembly most published studies on transcriptome sequencing of non-model organism have to date used the Roche 454 pyrosequencing platform; however current advances in bioinformatics have allowed the assembly of shorter reads without reference genome or transcriptome [41,49]. Using the European hake as a case study for non-model species, these two approaches (454 and GAII) were applied in parallel to high-throughput SNP discovery in expressed sequences. Large scale (1,536 SNPs) validation of discovered markers was carried out to properly evaluate and compare their performance. The ultimate goal was to test whether an optimal and cost effective strategy exists to develop a broad set of SNP markers linked to functional loci to be used for improved genetic management of hake fisheries.

Materials and Methods

Samples for cDNA libraries

Muscle tissue samples for transcriptome sequencing were collected from four geographic regions (Figure 1) considered representative for the species range: two locations in the Mediterranean Sea (Aegean Sea and North Tyrrhenian, hereafter respectively AEGS and TYRS), and two in the Atlantic Ocean (North Sea and Iberian Atlantic coast, hereafter NTHS and ATIB, respectively). Animals used in this research were obtained from commercial fishery catches, therefore approval from any ethics committee or institutional review board was not necessary. Muscle tissues were stored in RNA later (Qiagen) at -80°C before RNA extraction. Two distinct non-normalized cDNA libraries were constructed and sequenced using 454 and GAII sequencing systems, following the methods described in detail below. Non-normalized library were used since, according to Hale *et al.* [39], cDNA libraries normalization has little impact on discovery of rare transcripts when NGS platform are used, due to the depth of sequencing coverage.

RNA extraction, cDNA library construction and 454 GS FLX sequencing

Total mRNA was obtained from eight individuals (two for each sampling location, Figure 1) using the RNeasy Lipid Tissue Mini Kit (Qiagen). mRNA was isolated using the Oligotex mRNA Mini Kit (Qiagen) and cDNA was synthesized using the SuperScript Double-stranded cDNA Synthesis Kit (Invitrogen). Due to the low amount of available cDNA compared to that specified in standard protocol for the preparation of a GS FLX sequencing library using Roche multiplex identifiers (MIDs), it was decided to use a customized barcoding protocol, modified from Binladen *et al.* [50], that allows smaller amounts of cDNA to be used for library preparation: a multiplex sequencing library was prepared by labeling each individual sample with two tags by ligation of specific 10-mer barcoding oligonucleotides to allow post-sequencing identification of sequences from the different samples. High-throughput sequencing was performed on a Roche 454 GS FLX (454) sequencer according to the manufacturer's protocol.

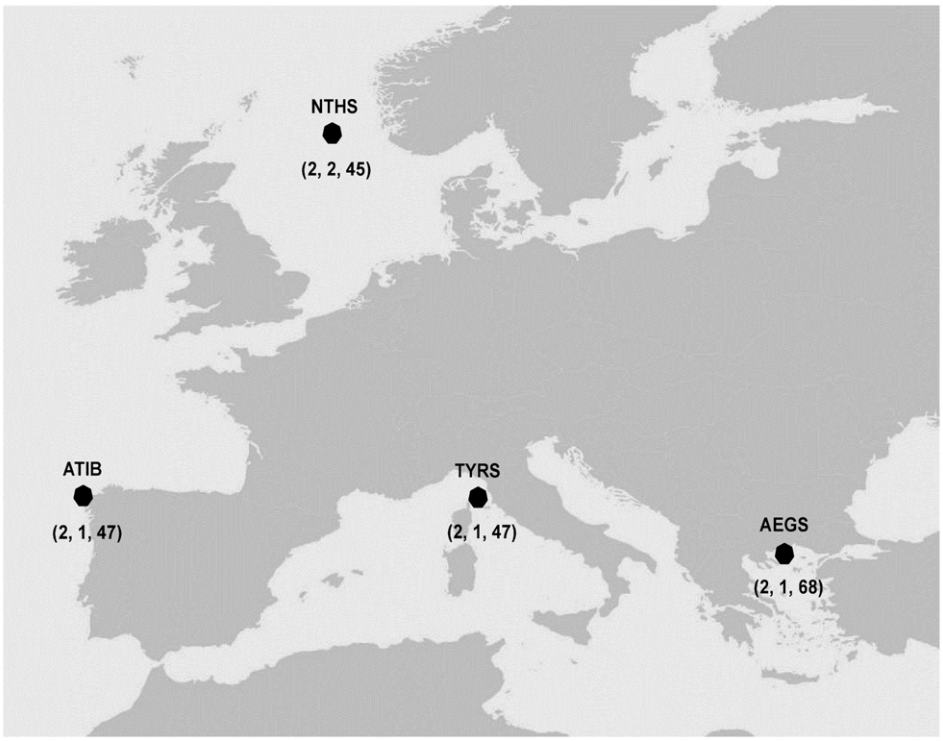


Figure 1. Geographic location of the four sampling sites. In brackets the number of specimens in the discovery panel (454 and GAII) and the number of individuals that have been subsequently genotyped in the validation step. NTHS: North Sea (59°19′N, 1°39′E); ATIB: Iberian Atlantic coast (43°20′N, 8°56′W); TYRS: Tyrrhenian Sea (42°32′N, 10°9′E); AEGS: Aegean Sea (40°19′N, 24°33′E).
doi:10.1371/journal.pone.0028008.g001

RNA extraction, cDNA library construction and Illumina Genome Analyzer II sequencing

Total RNA was purified from muscle tissue from five individual samples, two for NTHS, and one from each of the three other sampling sites (Figure 1) using the mRNeasy isolation kit (Qiagen) following the manufacture's guidelines. A total of 10 μg total RNA from each sample were used for library construction following the protocol described by Marioni *et al.* [51]. Poly-A containing mRNA was purified using poly-T oligonucleotide attached magnetic beads and fragmented using divalent cations under elevated temperature. After copying the RNA fragments into first strand cDNA using reverse transcriptase and random hexamer primers, second strand cDNA synthesis was performed using DNA Polymerase I and RNaseH. The short cDNA fragments were then prepared for sequencing using the Genomic DNA Sequencing Sample Prep Kit (Illumina). Briefly, the cDNA fragments were ''end-repaired'' using T4 DNA polymerase and Klenow DNA polymerase before adding a single A base to the cDNA fragments by using 3′-to-5′ exo-nuclease. Following ligation of Illumina adaptors, fragments of approximately 200 bp in size were gel purified and enriched by PCR amplification for 15 cycles. Library concentrations were measured using a Qubit fluorometer (Invitrogen) and size and purity were assessed using an Agilent 2100 Bioanalyzer (Agilent DNA 1000 Kit). Following dilution in buffer EB (Qiagen) to 10 nM, the libraries were denaturated with 2 M NaOH to a final DNA concentration of 1.0 nM, diluted to 4 pM with pre-chilled Hybridization buffer (Illumina) before loading in individual lanes into a 1.0 mm flowcell together with a single lane of a 2 pM PhiX control library (Illumina). Sequencing (76 cycles) was conducted on an Illumina Genome Analyzer (version II) using the Genomic DNA sequencing primer in combination with clustering and sequencing kits supplied by Illumina. Analysis of the images taken during sequencing was performed using the Genome Analyzer Pipeline Software (version 1.4.0, Illumina) generating the raw fastq files.

454 sequence processing and SNP detection

The sequences were de-multiplexed based on the specific barcoding tags and binned per individual sample. In order to obtain an optimal *de novo* assembly of sequences, repeats and repetitive or low complexity sequences were identified and masked in the reads by RepeatMasker (version open-3.2.7 with RM database version 20090120; Smit AFA, Hubley R & Green P: RepeatMasker at http://repeatmasker.org) using the Zebrafish (*Danio rerio*) repeat library and then cleaned for short sequences using SeqClean (http://compbio.dfci.harvard.edu/tgi/software/; options: -N; -L50). Sequence clustering was performed using CLC GenomicsWorkbench (parameters: match mode = ignore; similarity = 0.99; length fraction = 0.5; insertion cost = 3; deletion cost = 3; mismatch cost = 2) and after parsing the resulting large ace file respective sequences were then assembled 'per contig' using the assembly program CAP3 [52] (options: -o50; -p94) in order to generate the consensus sequences to be used as reference for read mapping in the subsequent SNP detection step. The GigaBayes program, a new implementation and expansion of the PolyBayes SNP detection algorithm by Marth *et al.* [53], was used for SNP discovery (parameters: ploidy = diploid; CRL = 4; CAL = 2; O = 3; D = 0.003), applying a minimum contig depth of four reads covering the polymorphic site with at least two reads for each allele; no insertion or deletion variants (InDels) were considered.

GAII sequence processing and SNP detection

After renaming and trimming of the first base, GAII short reads were assembled *de novo* using Abyss [54,55]. Different settings were tested and the final assembly was run using a k-mer value of 65. Contig trimming and exclusion of sequences shorter than 100 bp was performed by Seqclean (http://compbio.dfci.harvard.edu/tgi/software/; l = 100). MosaikBuild was used to build a database of reads (options: -st illumina, -tp 1, -ts 1) and MosaikAligner was applied (options: -mmp 0.05, -mm 12, unique, -act 35) to map the reads (database) against the Abyss contigs. After running MosaikSort the MosaikAssembler was used to generate the (.gig) input files for the SNP detection by GigaBayes (options: gff, diploid, multiple, CRL 10, CAL 4, D 0.003).

Contig functional annotation and Gene Ontology analyses

In order to characterize 454 and GAII contigs several approaches were explored, using the Basic Local Alignment Search Tool (BLAST) against various protein and nucleotide databases. The Blastn option was used against the NCBI nucleotide database (cut-off e-value of $<1.0\ E^{-5}$), against all annotated transcripts in the draft genomes of *Danio rerio*, *Gasterosteus aculeatus*, *Oryzias latipes*, *Takifugu rubripes*, *Tetraodon nigroviridis*, and *Homo sapiens* available at the Ensembl Genome Browser, and against all unique transcripts for *D. rerio*, *H. sapiens*, *O. latipes*, *T. rubripes*, *Salmo salar*, *Oncorhynchus mykiss* stored in the NCBI UniGene databases. The Blastx option was used (cut-off e-value of $<1.0\ E^{-3}$) to search against the entire UniProtKB/SwissProt and UniProtKB/TrEMBL protein databases as well as against the annotated proteins from the transcriptomes of *D. rerio*, *G. aculeatus*, *O. latipes*, *T. rubripes*, *T. nigroviridis*, and *H. sapiens* available through the Ensembl Genome Browser.

The Gene Ontology (GO) terms (“Cellular Component”, “Biological Process” and “Molecular Function”) were recovered using the Blastx search tool implemented in the software Blast2GO [56] against the NCBI non-redundant protein database.

A pipeline was developed to characterize the SNP mutations at the amino acid level. To obtain the putative reading frame all SNP-containing contigs were compared against six peptide sequence databases (Ensembl genome assembly for *G. aculeatus*, *T. nigroviridis*, *O. latipes*, *T. rubripes*, *D. rerio* and Swissprot database) using the Blastx algorithm (cut-off e-value of $<1.0\ E^{-3}$). The best match was selected and the aligned sequence portions of the query were saved as fasta files and then formatted as a Blast database. A fasta file containing 120 bp SNP-flanking sequences was prepared (two sequences for both alleles of each SNP) and a Blastx analysis was performed against the previously formatted database (cut-off e-value of $<1.0\ E^{-10}$). The aligned sequence portion of the two alleles for each SNP was compared for the presence of a synonymous or not-synonymous mutation.

To evaluate whether SNP-containing contigs were significantly enriched for specific GO terms compared to all annotated hake contigs, the Gossip package [57], which is integrated in the Blast2Go software, was used. Statistical assessment of annotation differences between the two sets of sequences (SNP-containing contigs vs all hake contigs) was carried out using Fisher’s Exact Test with False Discovery Rate (FDR) correction for multiple testing.

Candidate SNP selection

After SNP detection, *in silico* evaluation of candidate SNPs was carried out to select a panel of 1,536 candidate SNPs for high-throughput genotyping validation. Selection criteria were based on the score assigned by the Illumina GoldenGate Assay Design Tool (ADT), the analysis of putative intron-exon boundaries within each contig, and visual inspection of flanking region sequence quality. SNP scores obtained with the ADT take into consideration template GC content, melting temperature, uniqueness, and tendency to form hairpin loops. All SNPs with ADT scores below 0.4 were discarded, while SNPs with an ADT score higher than 0.7 were preferentially selected.

Intron-exon boundary prediction was performed using fish genome and transcriptome sequence resources following two parallel approaches. In the first, SNP-containing contigs were compared against five high-quality draft fish genomes (Ensembl genome assemblies for *G. aculeatus*, *T. nigroviridis*, *O. latipes*, *T. rubripes*, and *D. rerio*) using Blastn (cut-off e-value of $<1.0\ E^{-5}$). After parsing Blast results, the best match was listed including information on alignment length, as well as the start and end of the aligned region. In the case of a positive match, the position of the candidate SNP was evaluated in the framework of the aligned region. If the 60 bp up- and downstream of the SNP position were present in the alignment, the candidate SNP was considered embedded in a single exon, otherwise an intron was assumed to be present in the 121 bp target region for SNP assay design. The same process was repeated against five different databases (see above) and each SNP was assigned a code, either “1” (when in at least one comparison the candidate SNP and its flanking regions were located on a single exon), “0” (an intron was predicted to disrupt the candidate region), or “no” (no significant match against any of the five reference fish genomes). The second approach was designed in order to further increase the likelihood of a positive match and the reliability of intron-exon boundary prediction. SNP-containing contigs were used as a query in a Blast search (Blastn option, cut-off e-value of $<1.0\ E^{-5}$) against the transcriptome of each of the five model species as above. In the case of a positive hit, the matching transcript for each fish model species was downloaded from the Ensembl database, and the nucleotide position in the downloaded sequence corresponding to the candidate SNP in the original hake contig was identified based on the start-end positions of the Blast alignment between the hake contig and the fish model transcript. Then the putative homolog transcript was compared to its own genome sequence using Blast. Based on the inferred SNP position on the fish model transcript, SNPs were assigned a code as above (“1” for SNP candidate region located on a single exon or “0” if SNP region was assumed to be disrupted by an intron). In the case of no matches between the original hake contig and any of the five fish transcriptomes, the SNP was scored “no”. A flowchart for the intron-exon pipeline is depicted in Figure S3.

A final evaluation step of putative SNPs was performed by direct visual inspection of contigs using the assembly viewer software Eagleview [58] and Cluster Viewer (clview; http://compbio.dfci.harvard.edu/tgi/software/), with the aim of ranking candidates within each contig by integrating information on the overall contig assembly quality, depth and length, the quality of flanking regions (number of ambiguous sites), distance and clustering of polymorphic sites. SNPs with highest rank values were selected within each contig.

SNP validation by high-throughput genotyping

A total of 1,536 candidate SNPs were selected to be validated by high-throughput genotyping. Genomic DNA was extracted from fin clip tissues of 207 individuals sampled from the same four locations of origin of specimens used to derive the libraries (AEGS, TYRS, ATIB, NTHS, see Figure 1). A NanoDrop spectrophotomer was used to ascertain that DNA quality and quantity

met the requirements for the genotyping assay. Genotyping was performed using the Illumina GoldenGate Assay platform [59] and the resulting data were visualized and analyzed with the GenomeStudio Data Analysis Software package (1.0.2.20706, Illumina Inc.). Samples with a call rate lower than 0.8 and loci showing poor clustering were excluded. Accepted SNPs were manually re-clustered, to correct errors in allele calling due to inappropriate cluster identification.

Statistical analysis

Different variables were defined to analyze results of SNP discovery and genotyping. The first two are categorical variables that refer to the outcome of individual SNP assays. *SNP_assay_conversion* assumes value 0 (failed) if Illumina SNP assay did not yield a reliable genotype for the examined individuals, either because no clear clustering was observed or due to lack of signal, value 1 (successful) if consistent clustering was obtained, irrespective of the observed genotype(s). *SNP_genotype* has value 0 (monomorphic) if all scored individuals are homozygous for the same allele, value 1 (polymorphic) if two alleles are observed. *SNP_score* reports ADT score for individual SNPs. *I_E_test* and *I_E_species_match* refer to the outcome of the intron-exon boundary analysis pipeline, *I_E_test* = "no" means no significant match could be obtained, value 1 signifies that at least one significant match was reported, with the SNP candidate region putatively contained in a single exonic region, while *I_E_species_match* counts the number of fish model species showing a significant Blast match. *Depth* is a quantitative variable reporting the number of sequence reads supporting individual SNPs, *Individuals* and *Geosites* counts respectively the number of individuals and geographical sites from the original discovery panel that contributed with at least one sequence read to a specific SNP. *MSAF* (minor sequencing allele frequency) represents the frequency of the minor allele detected at the SNP discovery stage. *Rank* is an ordinal variable referring to an order of choice of SNPs within each contig, arbitrarily assigned after a visual inspection (see above). *Q* reports the number of mismatches in the flanking regions of SNP positions based on comparison of sequence reads within a contig. SNP discovery results obtained using 454 sequencing technology were compared with those produced through GAII sequencing using either parametric T-tests for two independent samples (for quantitative, normally distributed variables: *Contig length*, *Depth*) or non-parametric tests, a χ^2 test for categorical variables (*SNP_assay_conversion*, *SNP_genotype*, *I_E_test*) and a Mann-Whitney U test for ordinal variables (*I_E_species_match*, *Geosites*) or quantitative variables not following a normal distribution (*SNP_score*, *MSAF*). All tests were carried out in SPSS ver. 12.0 with Monte Carlo simulation (1,000,000 permutations) to estimate confidence intervals for Mann-Whitney tests.

Binomial logistic regression, implemented in SPSS ver. 12.0, was used to evaluate several predictor variables on two dependent dichotomous variables: *SNP_assay_conversion* and *SNP_genotype*. For both data sets (GAII and 454), positive conversion/clustering of individual SNP assay was analyzed first, assigning single SNPs into two classes (successful-failed). Second, DNA polymorphism was evaluated by filtering out failed SNPs and dividing positive loci respectively into two groups based on observed genotypes (polymorphic-monomorphic, *SNP_genotype*). For 454 data, eight ordinal/quantitative variables (*SNP_score*, *I_E_species_match*, *Depth*, *MSAF*, *Individuals*, *Geosites*, *Q*, *Rank*) and one categorical variable (*I_E_test*) were considered. For GAII data, six ordinal/quantitative variables (*SNP_score*, *I_E_species_match*, *Depth*, *MSAF*, *Individuals*, *Geosites*) and one categorical variable (*I_E_test*) were examined. Predictor variables were either all included in the predicting model (option "enter" in SPSS) or best predictors were selected using a stepwise deletion approach (option "backward"). In the latter approach, the Wald χ^2 statistic was used to estimate the contribution of each predictor.

The Receiver Operating Characteristic (ROC) curve, a widely used method for evaluating the discriminating power of a diagnostic test, was used to assess the significance of specific variables. ROC analysis was implemented in MedCalc ver. 11.5.1.0.

Results and Discussion

Sequencing, *de novo* assembly, and annotation

Approximately 100 Mega base pairs (Mbp) were obtained using 454 sequencing technology, whereas nearly 4,000 Mbp were produced using the GAII sequencer (Table 1). About 6% (30,025) of raw 454 reads, in which no barcoding tag could be reliably recognized, were excluded from further analysis, as individual sample identity of the reads was considered essential to the validation process. The remaining 476,747 reads, which showed the 5′ barcoding tag, were assigned to four groups according to geographic origin (Figure 1, 102,854 for AEGS, 135,494 for TYRS, 97,207 for NTHS, and 141,192 for ATIB). GAII sequencing yielded 8,789,024, 10,327,499, and 10,029,014 reads from single individuals from AEGS, TYRS, and ATIB respectively, while 21,539,868 sequences were obtained from two distinct individuals from NTHS. All 454 and GAII sequence data have been submitted to the EBI Sequence Read Archive (SRA) under the study accession number ERP000950 (http://www.ebi.ac.uk/ena/data/view/ERP000950).

After pre-processing steps (adaptor clipping and read quality filtering), 462,489 454 reads were assembled *de novo* into 5,702 separate contigs of at least 100 bp length (from 4,710 initial clusters). GAII sequence assembly resulted in 9,258 contigs, of which 60% were discarded, having a length shorter than 100 bp. The remaining 3,756 contigs were further processed for SNP discovery. Results on contig length are summarized in Table 1. As expected, mean contig length was significantly higher for 454 than GAII (T-test $p<0.0001$), whereas the opposite was observed for sequence coverage (T-test $p<0.0001$). The observed average length of unique transcripts after assembly of 454 and GAII sequence reads are comparable with values reported in a recent study that applied a simulation approach (based on experimental data) to evaluate the performance of transcriptome sequencing using Sanger, Roche 454, and GAII technologies on either non-normalized or normalized cDNA libraries [47]. For 454 GFLX sequencing on non-normalized libraries (100 Mbp output) average contig length is expected to be 556.6 bp and 217.3 bp for GAII

Table 1. Summary statistics of sequence assembly.

	454	GAII
Total number of sequences	506,772	50,685,405
Average length (min-max)	206 (6–457) bp	74 bp
Sequences suitable for assembly	462,489	50,685,405
Number of contigs	5,702	3,756
Average contig length (min-max)	331 (100–5,103) bp	190 (100–3,063) bp
Annotated contigs (%)	4,221 (74.03)	2,644 (70.39)

doi:10.1371/journal.pone.0028008.t001

sequencing (4,000 Mbp output, non-normalized libraries) [47], similar to what was observed in the present study, whereas the number of unique contigs obtained for hake muscle transcriptome is considerably smaller than expected (respectively 37,853 contigs for 454 and 147,261 for GAII under the same conditions as above). Differences in the experimental setting in the present work (*e.g.* use of a single tissue library and different genome size of the target species) are likely contributing factors to the observed differences.

A total of 4,221 454-contigs (74.02%) showed a significant Blast match against at least one species sequence database among those searched. Annotation by similarity was possible for 2,644 GAII contigs (70.39%), significantly less ($\chi^2 = 14.83$, $p<0.001$) than 454 contigs. A potential explanation for this observation lies in the shorter average length of GAII contigs, which likely reduces the overall probability of obtaining positive Blast hits. In any case, the percentage of annotated contigs is higher for both 454 and GAII data when compared to other studies reporting transcriptome sequencing in teleost fish (40–63% of total annotated contigs, [45,60]). There are several variables that might influence the observed fraction of contigs with a positive Blast match, including phylogenetic distance from species with high quality draft genome sequence, contig length, tissue type(s), developmental stages, and the method used for library construction. The use of a non-normalized cDNA library of adult skeletal muscle may have led to a biased representation of the hake transcriptome, favouring highly expressed genes encoding either housekeeping or structural (muscle contractile fibres) proteins. Housekeeping and structural proteins are known to be expressed at higher levels and to show a higher degree of sequence conservation compared to other proteins [61] (*e.g.* components of the immune system), therefore increasing the chance of finding a positive Blast match. A total of 120 454-contigs (2.1%) and 67 GAII-contigs (1.78%) were identified as mitochondrial sequences. It was possible to associate one or more GO terms to 1,606 454-contigs (28.16%) and 884 GAII contigs (23.53%). Results from level 2 GO assignments within the three categories are summarized in Figures S1 and S2.

In silico SNP detection

A total of 4,034 candidate SNPs were identified *in silico* in 889 454-sequenced contigs (15%), with an average of 0.73 SNPs per 100 bp surveyed in the assembled transcriptome, or 41.1 SNPs per generated Mbp. Approximately 60% of 454-contigs contained one or two SNPs (Figure 2). The total number of SNPs (8,606) found in 2,384 GAII contigs was more than two-fold higher, with a comparable distribution across contigs (Figure 2). A mean of 1.7 SNPs was found every 100 analysed bp, but with a lower output (2.3) per sequenced Mbp. As already observed, sequencing depth at candidate SNP positions was significantly higher (T-test $p<0.0001$) in GAII sequence data (Table 2), which might also explain the near two-fold increase in frequency of SNPs per bp reported above. On the other hand, 3,621 SNPs (89.7%) from 454-generated contigs had at least 60 bp of flanking sequence on either side of the SNP, the minimal requirement for the Illumina GoldenGate genotyping assay, while nearly half of GAII-generated SNPs were discarded due to flanking regions of insufficient size, an observation directly related to the highly significant difference (T-test $p<0.0001$) in the average length of SNP-containing contigs (Table 2). A comparison of 454- and GAII-SNPs revealed that 440 loci were common to both data sets. All suitable SNPs (*i.e.* having sufficient flanking regions) were analysed with the Illumina ADT software, and SNPs with ADT score >0.4 (3,437) were further evaluated. After filtering for the intron-exon boundary prediction results, 2,173 SNPs in the 454 data, presumed to be located in single exons (851 SNPs) or without a Blast match (1,322 SNPs), were selected to be visually inspected. For GAII SNPs, 3,857 loci out of 4,637 with ADT score >0.4 passed the filtering step based on the intron-exon boundary prediction (468 located in single exons and 3,389 with no significant match) (Table 2).

Annotation by similarity of SNP-containing contigs further confirmed that 454-contigs show a higher percentage of putatively annotated sequences (86% vs 71%, $\chi^2 = 72.4$, $p<0.0001$). Blast searches also identified 218 candidate SNPs (positioned in 62 different contigs) that are presumably located on the hake mitochondrial genome. GO term analysis showed a significant enrichment of specific GO terms when comparing the annotations of SNP-containing contigs against all unique transcripts obtained for the hake muscle transcriptome (Figure 3 A, B, Tables S1 and S2 in Supporting Information). Protein synthesis, in particular ribosome assembly/function, and energetic metabolism (*e.g.* carbohydrate metabolism, electron chain transport, ATP synthesis) are significantly over-represented across both 454- and GAII-SNPs, while cytoskeletal components (*e.g.* myosin filaments, microtubules) are enriched in GAII-SNPs. As mentioned before, the use of non-normalized cDNA libraries likely favored a higher representation of abundantly expressed genes. In the skeletal muscle, a high level of protein synthesis is required to fulfill the need for abundant contractile fibres, therefore over-representation of ribosomal/translation components as well as cytoskeletal proteins is expected. Such over-representation likely translates into greater sequence coverage and ultimately in a larger proportion of SNPs being identified in specific functional groups of genes.

454 and GAII SNP validation

A set of 1,536 SNPs (Table 3) was selected mainly on the basis of ADT score and intron/exon analysis for validation on the Illumina Golden Gate platform, using a custom-designed SNP chip. Mitochondrial SNPs (35) were also validated, but were excluded from further analysis as they are not entirely comparable (*i.e.* mitochondrial DNA is a multiple copy, haploid, and intron-less genome). In total, 817 454-SNPs distributed on 516 contigs, 557 GAII-SNPs distributed on 463 contigs, 127 common (454 and GAII) nuclear SNPs, and 35 mitochondrial loci were included in the panel of 1,536 candidate loci to be validated by high-throughput genotyping. The common set of 127 nuclear SNPs was genotyped only once, but all these loci have been independently identified twice and descriptive variables could be estimated in both data sets (454 and GAII). For this reason, they have been considered as independent observations and analysed separately.

Excluding mitochondrial loci, 1,501 unique nuclear SNPs were scored, with nearly identical percentages of successful assay conversion for the two sets of data (409/944, 43.32% (454) and 296/684, 43.27% (GAII), $\chi^2 = 0$, $p = 1$). This leads to the question of how do SNP conversion rates compare to those observed in other studies that reported on high-throughput SNP discovery and validation in non-model species. This remains largely unexplored, and the few available data are not homogeneous. Most studies report *in silico* SNP detection in next-generation sequencing data from non-model species with limited, if any, experimental validation of discovered polymorphisms (*e.g.* [29,33,35,42]). In a recent study on SNP discovery and validation in salmonid species [41], pooled and single tissue cDNA libraries were sequenced with SOLiD technology and short sequence reads aligned to contigs obtained from reference EST databases. This approach yielded a similar number of average SNPs per contig (1.6–4.4) to that observed here (3.6–4.5, Table 2). PCR-based validation was

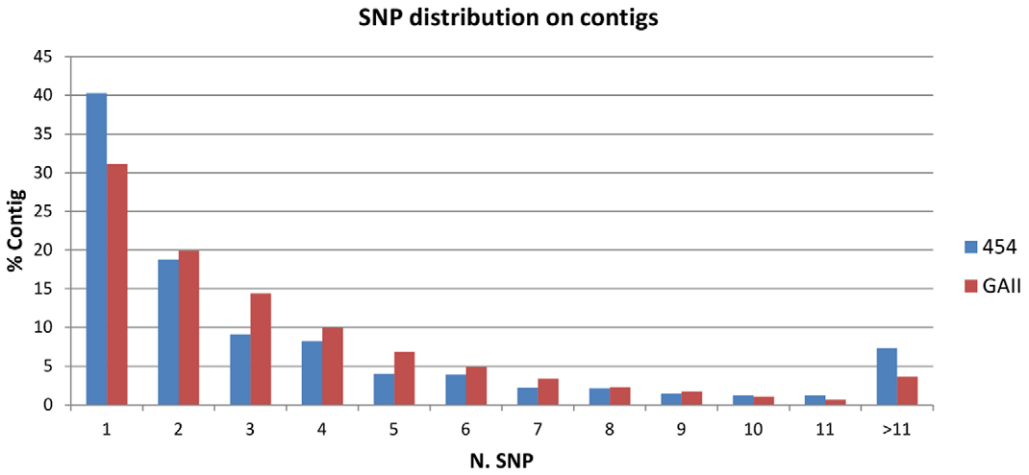


Figure 2. Distribution of SNPs across 454 and GAII contigs. On the x-axis, number of SNPs per contig; on the y-axis, the percentage of contigs showing a specific number of SNPs.
doi:10.1371/journal.pone.0028008.g002

carried out on 96 SNPs, with an assay conversion success rate of 53% (51/96). The conversion rate of validation assays (Illumina Golden Gate) was found to be much higher in two larger studies carried out in non model bird species, the common turkey [62] and the mallard duck [63], for which rates of 88.5% (340/384) and 94.7% (364/384) were observed, respectively. Both studies, however, used Illumina GA technology to sequence reduced representation (genomic) libraries (RRL) of the target species genome. Drastic sequence quality filtering was enforced especially in the turkey study, discarding up to 75% of sequence reads. Finally, assembled turkey contigs were verified through mapping against the high quality draft genome of the closely related chicken species. This step appears crucial as shown by the comparison of conversion rates between genome-mapped turkey SNPs (92.1%, 316/343) and unmapped ones (58%, 24/41) [62]. Likewise, mallard duck genomic contigs were directly aligned against the domestic duck genome, which provided an even closer reference in the SNP discovery process [63]. Additional comparative evidence can be obtained from two studies reporting SNP discovery and high-throughput validation in the channel catfish [44] and the Atlantic cod [45]. In both studies, validation was carried out using Illumina GoldenGate technology on either 384 catfish SNPs or 3,072 cod SNPs. Assay conversion rates were similar (266/384 (69.2%) in catfish, 2,291/3,072 (74.5%) in cod) and intermediate between those observed here and values obtained for turkey and mallard. However, SNP discovery was based on EST libraries produced with traditional Sanger sequencing (with higher read quality), therefore the results are not entirely comparable.

Of the successfully converting assays, a slightly (non-significantly) higher percentage (67.5% vs 63.3%, $\chi^2 = 1.18$, $p = 0.272$) of truly polymorphic sites was detected in GAII-SNPs (Table 3). These values are similar to those described in the channel catfish (156/266, 58.6%) [44] and the Atlantic cod (1,684/2,291, 73.5%) [45], although yet again these two studies are not entirely comparable. For instance, size and composition of the discovery panel for cod and catfish could not be determined. When comparing rates of hake polymorphic SNPs with other NGS-based SNP discovery studies, the percentage obtained from the transcriptome of sockeye salmon (11/51, 21.5%) [41] is lower, but those reported for RRL sequencing in the turkey (324/340, 95.2%) and the mallard (363/364, 99.7%) are significantly higher. Size and composition of the discovery panel appear similar (10 individuals from 5 populations in the sockeye salmon, 6 unrelated individuals in the turkey, 9 individuals from 3 populations in the mallard, 5–8 individuals from 4 locations in the hake), while the choice between transcriptome and genome sequencing as well as

Table 2. Summary statistics of SNP discovery and selection.

	454	GAII
***In silico* candidate SNPs**	4,034	8,606
Contigs with candidate SNPs	889	2,384
Average contig length (min-max)	617.9 (101–5103) bp	212.3 (100–3063) bp
Average depth in SNP position (min-max)	89 (4–3,678)	674 (8–33,079)
SNPs suitable for Illumina assay design	3,621	4,684
Average ADT score	0.76	0.82
SNPs with ADT score >0.4 (%)	3,437 (94.92%)	4,637 (99%)
SNPs with I/E* ''no match'' (%)	1,322 (38.46%)	3,389 (73.09%)
SNPs with I/E* ''single exon'' (%)	851 (24.76%)	468 (10.09%)

*Intron/exon boundary pipeline result.
doi:10.1371/journal.pone.0028008.t002

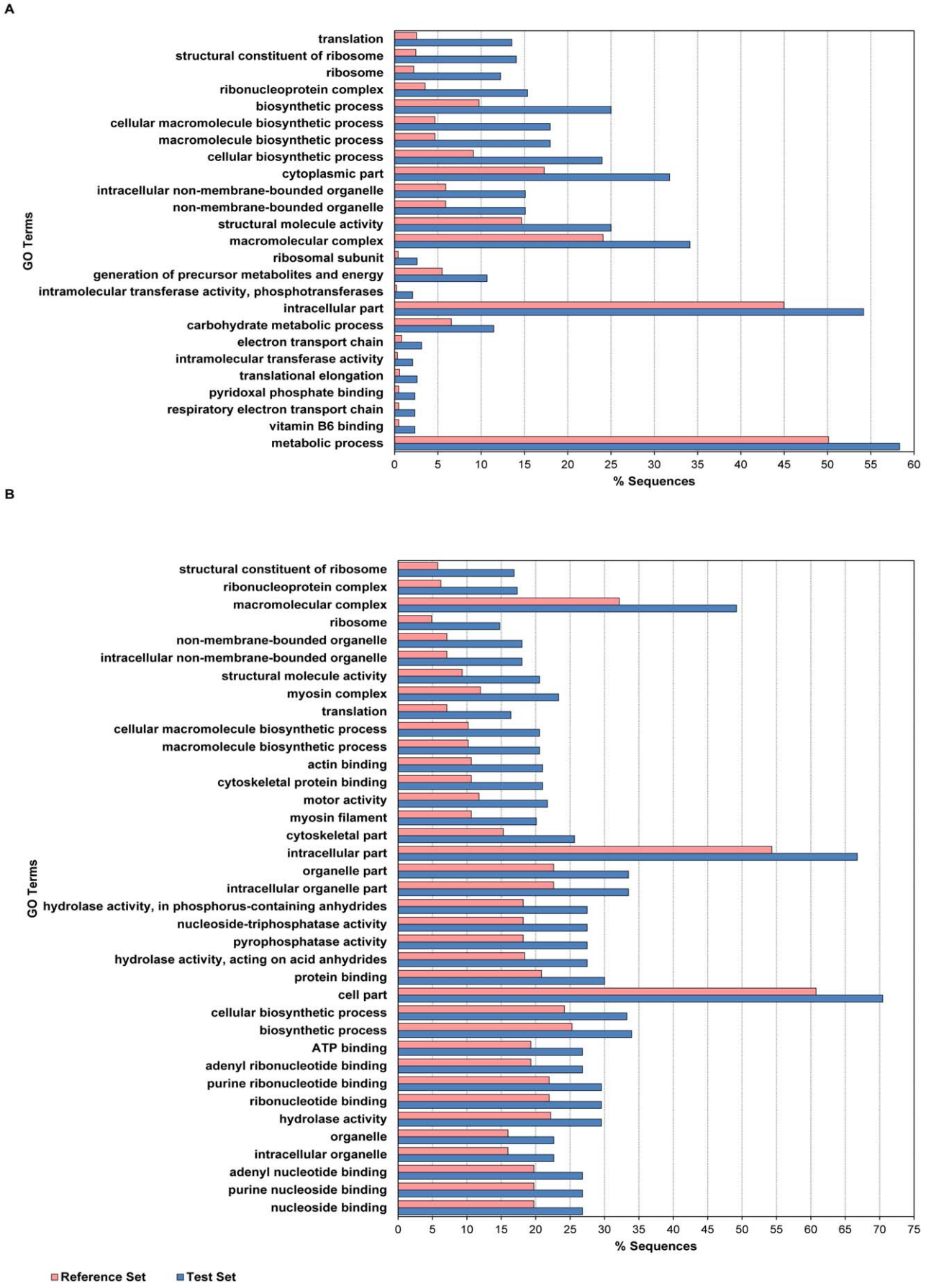

A
translation
structural constituent of ribosome
ribosome
ribonucleoprotein complex
biosynthetic process
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
cellular biosynthetic process
cytoplasmic part
intracellular non-membrane-bounded organelle
non-membrane-bounded organelle
structural molecule activity
macromolecular complex
ribosomal subunit
generation of precursor metabolites and energy
intramolecular transferase activity, phosphotransferases
intracellular part
carbohydrate metabolic process
electron transport chain
intramolecular transferase activity
translational elongation
pyridoxal phosphate binding
respiratory electron transport chain
vitamin B6 binding
metabolic process
GO Terms
0 5 10 15 20 25 30 35 40 45 50 55 60
% Sequences
B
structural constituent of ribosome
ribonucleoprotein complex
macromolecular complex
ribosome
non-membrane-bounded organelle
intracellular non-membrane-bounded organelle
structural molecule activity
myosin complex
translation
cellular macromolecule biosynthetic process
macromolecule biosynthetic process
actin binding
cytoskeletal protein binding
motor activity
myosin filament
cytoskeletal part
intracellular part
organelle part
intracellular organelle part
hydrolase activity, in phosphorus-containing anhydrides
nucleoside-triphosphatase activity
pyrophosphatase activity
hydrolase activity, acting on acid anhydrides
protein binding
cell part
cellular biosynthetic process
biosynthetic process
ATP binding
adenyl ribonucleotide binding
purine ribonucleotide binding
ribonucleotide binding
hydrolase activity
organelle
intracellular organelle
adenyl nucleotide binding
purine nucleoside binding
nucleoside binding
GO Terms
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
% Sequences
Reference Set
Test Set

Figure 3. Enrichment of SNP-containing contigs in GO terms. Differential distribution of GO terms in SNP-containing contigs (test set) compared to all contigs (reference set) in 454 data (A) and GAII data (B).
doi:10.1371/journal.pone.0028008.g003

the availability and quality of reference sequences vary across studies. As already observed for the assay conversion rate, the relevance of the latter factor is emphasized by the significant difference ($\chi^2 = 5.62$, $p = 0.017$) in the percentage of polymorphic "unmapped" SNPs (20/24, 83.3%) and "mapped" ones (304/316, 96.3%) in the turkey data set [62]. Flanking sequences (120 bp) of SNPs validated as polymorphic in this study are available in Table S3.

While similar conversion rates and percentages of polymorphic loci were obtained for both NGS technologies applied to hake SNP discovery, 454-SNPs showed a significantly higher number of Blast matches against at least one fish model coding sequence (*I_E_test* rate 38.4% (454) – 17.9% (GAII), $\chi^2 = 68.05$ $p<0.0001$). Also continuous/ordinal variables were significantly different between the two data sets (data not shown), with average *I_E_species_match* and *MSAF* being higher in 454 SNPs, whereas mean *SNP_score*, *Geosites*, and *Depth* were larger in GAII data. Higher mean number of species matching with 454-SNPs is linked to the significantly larger *I_E test* rate, which in turn likely correlates with the higher percentage of annotated 454-contigs, while greater *Depth* in GAII-SNPs reflects the overall much deeper sequence coverage obtained with GAII technology. Likewise, differential coverage likely explains the more complete representation of sequence variation across all four collection sites in GAII-SNPs. It is not clear why higher average *SNP_score* for GAII-SNPs was observed. It might be an effect of different contig length and/or sequence quality as the parameters that the ADT uses for estimating SNP scores are likely influenced by these factors. The higher *MSAF* estimated for 454-SNPs appears to be the effect of over-estimation of allele frequencies based on sequence data, again putatively related to the lower coverage of SNP sites obtained with 454 technology. When sequence-based and Illumina Golden Gate genotypic results are compared for the same loci in a paired-samples Wilcoxon signed rank test, Minor Allele Frequency (*MAF*) is significantly lower than *MSAF* (Wilcoxon paired rank test $V = 22{,}047.5$, $p<0.0001$) (Figure 4A), whereas it is not different for GAII-SNPs (Wilcoxon paired rank test $V = 9{,}503$, $p = 0.50$) (Figure 4B). Neither method provides accurate values of observed heterozygosity (Ho), although GAII-based data tend to over-estimate Ho (Wilcoxon paired rank test $V = 12{,}830$, $p<0.001$, Figure 5B), whereas 454-based Ho provides a substantial under-estimate of this parameter (Wilcoxon paired rank test $V = 7{,}372$, $p<0.00001$, Figure 5A).

To evaluate the predictive value and thus the potential usefulness of different parameters, all variables were examined together using binomial logistic regression analysis. The success rate for Illumina genotyping assay conversion/clustering show that inclusion of all predictor variables significantly improves model-fitting for 454 ($\chi^2 = 34.944$, degrees of freedom (df) 9, $p<0.001$) and marginally for GAII data ($\chi^2 = 14.641$, df 7, $p<0.05$). The ability to predict the outcome of individual SNP assays is relatively low, with an overall correct classification rate of 60.1% for 454 data and 56.7% for GAII data. Backward stepwise deletion of predictor variables identifies five best predictors (*SNP_score*, *I_E_test*, *MSAF*, *Individuals*, *Q*) for 454 data and three predictors (*SNP_score*, *Individuals*, *Geosites*) for GAII data as summarized in Tables 4 and 5. The number of individuals effectively sequenced for each SNP is positively correlated with the rate of successfully converting assays in both data sets. This is especially significant for GAII data (Table 5), where the maximum number of animals in the discovery panel (5) is lower than in the 454 one (8), suggesting that sequencing coverage *per se* is less relevant than the level of coverage across different individuals. The outcome of the intron-exon boundary pipeline is significant as a predictor only for 454

Table 3. Summary statistics for SNP validation.

	454	GAII
Total number of SNPs tested	966 (829)	707 (570)
Nuclear SNPs tested	944 (817)	684 (557)
SNPs with successful genotype calling*	409 (334)	296 (221)
Polymorphic	259 (195)	200 (136)
SNP with known reading frame	130 (97)	73 (45)
Synonymous	110 (83)	60 (37)
Non-synonymous	20 (14)	13 (8)
Monomorphic	150 (139)	96 (85)
Failed SNPs*	535 (483)	388 (336)

In brackets the number of SNPs after excluding the set of common loci.
*Data referring to nuclear SNPs.
doi:10.1371/journal.pone.0028008.t003

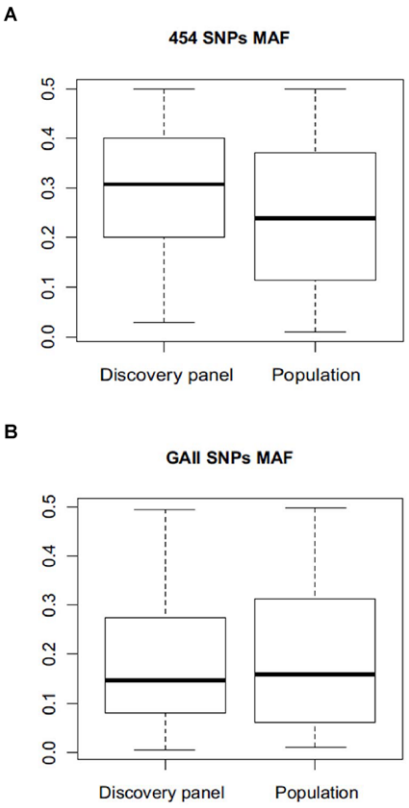


Figure 4. Minor allele frequency distribution. Box plot of minor sequence allele frequency (MSAF) in the discovery panel and Minor allele frequency (MAF) in the validation panel for 454 data (A) and GAII data (B).
doi:10.1371/journal.pone.0028008.g004

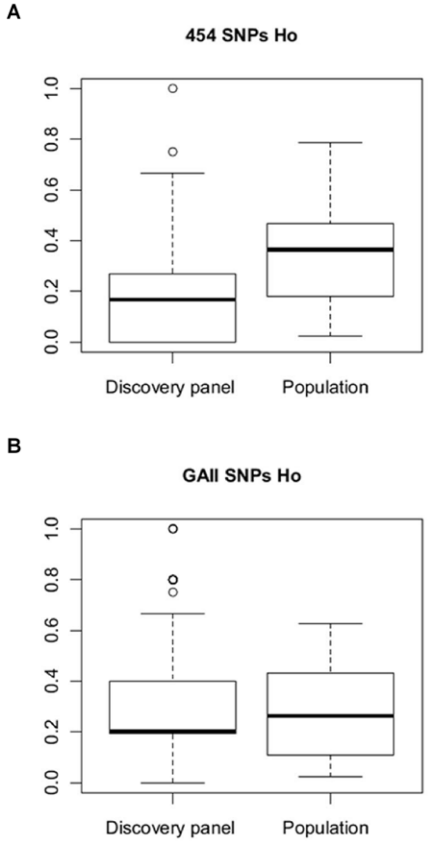


Figure 5. Observed heterozygosity distribution. Box plot of observed heterozygosity (Ho) calculated for the discovery panel and the validation panel of 454 data (A) and GAII data (B).
doi:10.1371/journal.pone.0028008.g005

data, with a higher number of successful assays for SNPs that are verified to be located in a single exon compared to "unknown" SNPs. This is in agreement with what reported by Wang and et al. [44] in the channel catfish, where the presence of an intron was among the causes of assay failure. Comparative genomic analysis to preliminarily exclude SNP candidate regions that are putatively interrupted by one or more introns might therefore be advisable, although it appears not significant for GAII-SNPS, likely in consequence of the lower number of positive Blast matches (Table 2). Less clear is the interpretation for the positive correlation of conversion rate with *MSAF* in 454-SNPs and the negative correlation with number of geographic areas in GAII-SNPs, while the negative correlation of the variable Q, which reports the number of mismatches in the flanking regions of SNP positions based on comparison of sequence reads within a contig, can be easily explained as a result of poor sequence quality on assay performance. This parameter was estimated only on 454-SNPs because the much greater sequencing depth of GAII data did not allow visual estimation of Q. Since this variable appears to convey valuable information, it might be useful to develop an automated pipeline to measure it in the future. Finally, for both data sets, and in particular for 454-SNPs, the ADT score proved to be a significant predictor (Tables 4 and 5).

A specific statistical analysis (Receiver Operating Characteristic (ROC) analysis) was carried out to evaluate the significance of ADT score as predictive variable. The estimated area under the ROC curve is 0.539 ± 0.015, which is significantly different (z statistic = 2.6, $p<0.01$) than expected by chance (0.5), confirming ADT score as a significant predictor of assay conversion. The optimal ADT value, which has the best specificity/sensitivity ratio, is 0.735. It should be noted that, while significant, the overall performance of ADT as diagnostic factor is rather limited.

Binomial logistic regression analysis was also implemented to evaluate the outcome of successful SNP assays, *i.e.* to predict polymorphic and monomorphic loci. Including all predictor variables, model fitting improved significantly for 454 data (χ^2 55.983, df 9, $p<0.0001$) as well as for GAII data (χ^2 29.059, df 8, $p<0.001$), with overall rates of correct classification of 70.9% (454) and 70.6% (GAII). Stepwise deletion of independent variables reduced the list of contributing predictors to *I_E_species_match*, *Rank*, *Individuals*, *Geosites*, and *Q* in the case of 454 data, while *I_E_test*, *I_E_species_match*, *Individuals*, and *Geosites* best predicted *SNP_genotype* for GAII data (Tables 6 and 7). The three predictive variables that are common to the two data sets show opposite correlations with SNP polymorphism, which suggests caution in using them as predictors, while the negative correlation of *I_E_test* outcome with polymorphism in GAII-SNPs is not easily interpreted and is of little use as positive *I_E_test* is relevant for predicting assay conversion, at least in 454-SNPs. More interesting is the negative correlation between the number of mismatches in the flanking regions (*Q*) with the degree of polymorphism, similar to what was observed for assay conversion rate. This evidence is in agreement with the positive predictive value of *Rank*, a subjective measure of overall candidate SNP suitability, which also takes into account contig sequence quality. While *Rank* cannot be translated into an objective, operator-free score, *Q* appears a promising predictor, which might be possible to automatically estimate for a very large number of SNPs and could provide useful information on assay conversion as well as SNP polymorphism. In fact, ROC analysis shows that the area under the ROC curve (0.688 ± 0.0279, Figure 6, see below) is highly significant for *Q* when predicting whether a specific locus is monomorphic (z statistic = 6.737, $p<0.0001$). Optimal trade-off between specificity (56.7%) and sensitivity (72.7%) is obtained with $Q>1$.

Finally, all experimentally validated SNPs were analysed using a bespoke pipeline, which was developed to compare SNP-containing contigs with known protein sequences and to predict protein-coding regions and the corresponding putative reading frame. It was possible to obtain this information reliably for approximately half of the validated SNPs (Table 3) for both data sets, a much higher percentage than that obtained for Atlantic cod (9%) [45]. This is likely attributable to the less 5′ end-biased transcript coverage of NGS technologies compared to Sanger

Table 4. Predictor variables for 454 SNP data (failed/successful), backward stepwise elimination.

	B[a]	Wald[b]	df	P[c]
SNP_score	1.735	14.607	1	**0.000**
I_E_test(1)	−0.361	6.545	1	**0.011**
MSAF	1.415	5.418	1	**0.020**
Individuals	0.075	2.919	1	0.088
Q	−0.007	5.205	1	**0.023**
Constant	−2.339	13.067	1	0.000

[a]Regression coefficient for individual variable,
[b]Wald χ^2 statistic,
[c]associated probability. (*e.g.* average SNP_score for successful SNPs is 0.794, whereas mean SNP_score is 0.759 for "failed" assays).
doi:10.1371/journal.pone.0028008.t004

Table 5. Predictor variables for GAII SNP data (failed/successful), backward stepwise elimination.

	B[a]	Wald[b]	df	P[c]
SNP_score	0.938	2.506	1	0.113
Individuals	1.203	8.795	1	**0.003**
Geosites	−1.206	7.427	1	**0.004**
Constant	−2.196	6.785	1	0.009

[a]Regression coefficient for individual variable,
[b]Wald χ^2 statistic,
[c]associated probability.
doi:10.1371/journal.pone.0028008.t005

ESTs. Of all SNPs located in a coding region, around 20% were putative amino acid replacement substitutions, without significant difference between 454- and GAII-SNPs (Table 3). In this respect, Atlantic cod SNPs were significantly biased toward non-synonymous substitutions (synonymous/non-synonymous 90/51) compared to all hake SNPs (147/28, after excluding common loci) (χ^2 15.06, df 1, $p<0.0001$). Closer examination of non-synonymous SNPs shows that, if amino acids are divided into five classes (non-polar, polar, negatively-charged, positively-charged, and aromatic), the majority of amino acid substitutions are conservative replacements (*i.e.* occurring within the same class), but there are 12 non-conservative mutations, which might cause significant functional changes in the encoded protein.

Conclusions

In the present study, two different NGS methods were applied to high-throughput SNP discovery in the muscle transcriptome of a non-model fish species. Overall, the comparison revealed that despite substantial differences in sequence throughput, average sequencing depth, and sequence read length, similar results were obtained after SNP experimental validation in terms of assay conversion rate and percentage of polymorphic loci. GAII technology yields a larger number of candidate SNPs, but the majority of them are not suitable for SNP genotyping due to short flanking regions. On the other hand, 454-SNPs are less numerous, but are located on longer contigs, which are more easily annotated and screened for putative introns in the candidate region using a comparative genomic approach. Although the platforms we have evaluated have been recently upgraded and superseded by later versions, still our findings should remain valid and relevant. Indeed, during the past few years Illumina and 454 platforms have experienced rapid progress towards the enhancement of reads length and yield in terms of Mb per run produced, mainly resulting in increased throughput; however, raw sequencing error rates have not decreased along with the improvement of instruments and chemistry [64]. For this reason, while using latest platforms could increase transcriptome representation and coverage and improve *de novo* assembly, due to the higher sequencing output, we believe results on SNP discovery and validation, primarily influenced from sequencing error profiles, would be proportionately comparable to our results. In order to evaluate and compare costs of SNP discovery, several aspects should be taken into account. However, considering only the raw cost of reagents per Mb updated to 2011 [64], and referring to the output produced and used in this study for SNP discovery, the Illumina sequencing systems seems to be more appropriate, as similar results in terms of validated SNPs were obtained at less than half the 454 costs (480$ to produce approximately 4,000 Mbp at a reference price of 0.12$/Mbp using the Illumina GAIIx against 1240$ to produce approximately 100 Mbp at a reference price of 12.4$/Mbp using the 454 FLX Titanium; based on [64]). If we consider the respective platforms currently available, the Illumina HiSeq 2000-v3 and the 454 FLX+, the cost-effectiveness of the

Table 6. Predictor variables for 454 SNP data (monomorphic/polymorphic), backward stepwise elimination.

	B[a]	Wald[b]	df	P[c]
I_E_speciesmatch	−0.127	6.369	1	**0.012**
Rank		6.977	2	**0.031**
Rank(1)	0.676	4.232	1	**0.040**
Rank(2)	0.841	4.983	1	**0.026**
Individuals	−0.224	3.489	1	0.062
Geosites	0.802	10.433	1	**0.001**
Q	−0.044	8.907	1	**0.003**
Constant	−2.740	5.915	1	0.015

[a]Regression coefficient for individual variable,
[b]Wald χ^2 statistic,
[c]associated probability.
doi:10.1371/journal.pone.0028008.t006

Table 7. Predictor variables for GAII SNP data (monomorphic/polymorphic), backward stepwise elimination.

	B[a]	Wald[b]	df	P[c]
I_E_test(1)	2.247	8.272	1	**0.004**
I_E_speciesmatch	0.345	2.735	1	0.098
Individuals	1.805	5.162	1	**0.023**
Geosites	−2.188	5.581	1	**0.018**
Constant	−1.577	1.361	1	0.243

[a]Regression coefficient for individual variable,
[b]Wald χ^2 statistic,
[c]associated probability.
doi:10.1371/journal.pone.0028008.t007

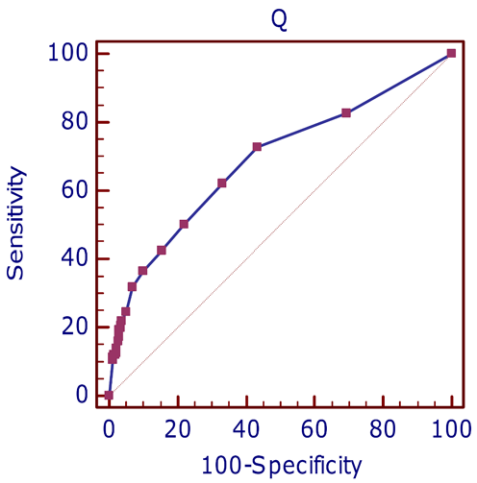


Figure 6. ROC curve of Q score predicting monomorphic/polymorphic SNPs in 454 data.
doi:10.1371/journal.pone.0028008.g006

Illumina technology is maintained, as same results would be achieved with less than a quarter of 454 costs. Additionally, HiSeq technology might allow to increase the number of individuals included in the SNP discovery panel without decreasing coverage depth. As such a variable (Individuals) was found to be positively correlated with successful conversion rate of SNP assays, this represents a further element in favor of Illumina technology. It should be noted, however, that the most expensive steps were the *in silico* analysis of sequence reads (assembly, SNP discovery, quality assessment) and the high-throughput genotyping assays. While the latter showed similar conversion rates, analysis of GAII reads might be more laborious and produce less reliable assemblies [47]. On the other hand, it has been demonstrated that the quality of transcriptome assembly depends on sequence coverage [47], therefore the shift toward HiSeq might allow easier and more reliable assemblies. More generally, the results presented here clearly demonstrate that it is possible to identify and effectively validate many polymorphic SNPs, in transcribed regions of a non model species. The lack of a reference genome, however, dramatically affects the genotyping success rate, although the overall efficiency can be improved using strict quality criteria/filters, especially in the case of 454-SNPs, such as testing for intron-exon boundaries, defining optimal ADT scores, and targeting low *Q* values (number of mismatches in the flanking regions).

The SNP markers developed in the current study represent novel tools for a broad range of future applications in population genetic studies focusing on the European hake. A deeper understanding of ecological and evolutionary dynamics of European hake populations across the entire distribution range provides the necessary means for a proper management and conservation policy, aimed at promoting sustainable fishery and preventing overexploitation and illegal fishing activities.

Supporting Information

Figure S1 Gene Ontology (GO) assignment (2nd level GO terms) of hake 454 contigs. A "Cellular Component". **B** Molecular Function". **C** "Biological Process".
(PDF)

Figure S2 Gene Ontology (GO) assignment (2nd level GO terms) of hake GAII contigs. A "Cellular Component". **B** Molecular Function". **C** "Biological Process".
(PDF)

Figure S3 Flowchart describing the intron-exon pipeline.
(PDF)

Table S1 Results of GO enrichment analysis performed using all 454 contigs as reference set, and 454 SNP-containing contigs as test set. Significantly overrepresented GO terms are listed, together with the respective category (P: "Biological Process"; F: "Molecular Function"; C: "Cellular Component"), FDR (false discovery rate) and Fisher's Exact Test p-value.
(PDF)

Table S2 Results of GO enrichment analysis performed using all GAII contigs as reference set, and GAII SNP-containing contigs as test set. Significantly overrepresented GO terms are listed, together with the respective category (P: "Biological Process"; F: "Molecular Function"; C: "Cellular Component"), FDR (false discovery rate) and Fisher's Exact Test p-value.
(PDF)

Table S3 List of 395 SNPs validated as polymorphic in this study (including 454-, GAII- SNPs and the set of loci detected using both data sets) together with the corresponding flanking sequences of approximately 120 bp (SNP alleles in brackets).
(XLSX)

Acknowledgments

We would like to thank all the members of the FishPopTrace Consortium for their input, and Marco Martini and Michele Drigo for their help with ROC analysis. Sampling was made possible by the generous collaboration of Paolo Sartor from the CBMI (Consorzio per il Centro Interuniversitario di Biologia Marina ed Ecologia Applicata "G. Bacci", Italy), Corrado Piccinetti and Marco Stagioni from the University of Bologna (Italy), Audrey Geffen from the University of Bergen (Norway) and Grigorios Krey from the National Agricultural Research Foundation (Greece). We thank Pernille K. Andersen (Aarhus University, Denmark) for sequencing sample and library management. We are particularly grateful to Jann Martinsohn and Eoin MacAoidh from the European Commission's Joint Research Center, Institute for the Protection and Security of the Citizen (Italy).

Author Contributions

Conceived and designed the experiments: GRC LB TP RO GEM. Performed the experiments: FP RON RO MIT SJH ME MA IM MB. Wrote the paper: IM LB RO GRC FT MB FP. Carried out *in silico* analyses: FP RON RO MIT SJH IM MB ME MA. Analyzed genotype data: IM MB RO LB. Carried out statistical analysis: IM MB RO LB.

References

1. Oliver P, Massutí E, Alheit J, Pitcher TJ (1994) Biology and fisheries of western Mediterranean hake (M. merluccius). In: Noakes DLG, ed. Hake: Biology, fisheries and markets: Springer Netherlands. pp 181–202.
2. Casey J, Pereiro J, Alheit J, Pitcher TJ (1994) European hake (M. merluccius) in the North-east Atlantic. In: Noakes DLG, ed. Hake: Biology, fisheries and markets: Springer Netherlands. pp 125–147.
3. FAO (2010) FAO yearbook. Fishery and Aquaculture Statistics. 2008.
4. FAO (2010) REPORT OF THE 12TH SESSION OF THE SCIENTIFIC ADVISORY COMMITTEE (SAC)-GFCM: XXXIV/2010/Inf.9.
5. Murua H, Michael L (2010) The Biology and Fisheries of European Hake, Merluccius merluccius, in the North-East Atlantic. Advances in Marine Biology: Academic Press. pp 97–154.
6. Lundy CJ, Moran P, Rico C, Milner RS, Hewitt GM (1999) Macrogeographical population differentiation in oceanic environments: a case study of European hake (Merluccius merluccius), a commercially important fish. Molecular Ecology 8: 1889–1898.
7. Pita A, Presa P, Perez M (2010) Gene flow, multilocus assignment and genetic structuring of the European hake (Merluccius merluccius). Thalassas 26: 129–133.
8. Roldan MI, Garcia-Marin J, Utter FM, Pla C (1998) Population genetic structure of European hake, Merluccius merluccius. Heredity 81: 327–334.
9. Castillo AGF, Martinez JL, Garcia-Vazquez E (2004) Fine Spatial Structure of Atlantic Hake (Merluccius merluccius) Stocks Revealed by Variation at Microsatellite Loci. Marine Biotechnology 6: 299–306.
10. Lundy CJ, Rico C, Hewitt GM (2000) Temporal and spatial genetic variation in spawning grounds of European hake (Merluccius merluccius) in the Bay of Biscay. Molecular ecology 9: 2067–2079.
11. Reiss H, Hoarau G, Dickey-Collas M, Wolff WJ (2009) Genetic population structure of marine fish: mismatch between biological and fisheries management units. Fish and Fisheries 10: 361–395.
12. Pita A, Pérez M, Cerviño S, Presa P (2011) What can gene flow and recruitment dynamics tell us about connectivity between European hake stocks in the Eastern North Atlantic? Continental Shelf Research 31: 376–387.
13. Ferguson A (1994) Molecular genetics in fisheries: current and future perspectives. Reviews in Fish Biology and Fisheries 4: 379–383.
14. Hauser L, Carvalho GR (2008) Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. Fish and Fisheries 9: 333–362.

15. Cimmaruta R, Bondanelli P, Nascetti G (2005) Genetic structure and environmental heterogeneity in the European hake (Merluccius merluccius). Molecular ecology 14: 2577–2591.
16. Hauser L, Seeb JE (2008) Advances in molecular technology and their impact on fisheries genetics. Fish and Fisheries 9: 473–486.
17. Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nature reviews Genetics 4: 981–994.
18. Wenne R, Boudry P, Hemmer-Hansen J, Lubieniecki KP, Was A, et al. (2007) What role for genomics in fisheries management and aquaculture? AquatLiving Resour 20: 241–255.
19. Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. Molecular ecology 17: 3583–3584.
20. Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009) Population genomics of marine fishes: identifying adaptive variation in space and time. Molecular ecology 18: 3128–3150.
21. Metzker ML (2010) Sequencing technologies - the next generation. Nature reviews Genetics 11: 31–46.
22. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing. Plant Physiology 144: 32–42.
23. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources 8: 3–17.
24. Wheat C (2010) Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138: 433–451.
25. Helyar SJ, Hemmer-Hansen J, Bekkevold D, Taylor MI, Ogden R, et al. (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources 11: 123–136.
26. Seeb JE, Carvalho G, Hauser L, Naish K, Roberts S, et al. (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources 11: 1–8.
27. Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, et al. (2007) SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. Bioinformatics 23: i387–i391.
28. Morin PA, Luikart G, Wayne RK, the Snp wg (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19: 208–216.
29. Garvin MR, Saitoh K, Gharrett AJ (2010) Application of single nucleotide polymorphisms to non-model species: a technical review. Molecular Ecology Resources 10: 915–934.
30. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Molecular ecology 13: 969–980.
31. Hemmer-Hansen J, Nielsen EE, Meldrup D, Mittelholzer C (2011) Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, Gadus morhua. Molecular Ecology Resources 11: 71–80.
32. Freamo H, O'Reilly P, Berg PR, Lien S, Boulding EG (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. Molecular Ecology Resources 11: 254–267.
33. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. The Plant Journal 51: 910–918.
34. Novaes E, Drost D, Farmerie W, Pappas G, Grattapaglia D, et al. (2008) High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics 9: 312.
35. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular ecology 17: 1636–1647.
36. Galindo J, Grahame JW, Butlin RK (2010) An EST-based genome scan using 454 sequencing in the marine snail Littorina saxatilis. Journal of Evolutionary Biology 23: 2004–2016.
37. Lepoittevin C, Frigerio J-M, Garnier-Géré P, Salin F, Cervera M-T, et al. (2010) In Vitro vs In Silico Detected SNPs for the Development of a Genotyping Array: What Can We Learn from a Non-Model Species? PLoS ONE 5: e11034.
38. Sanchez C, Smith T, Wiedmann R, Vallejo R, Salem M, et al. (2009) Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics 10: 559.
39. Hale M, McCormick C, Jackson J, DeWoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10: 203.
40. Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). Molecular ecology 19: 115–131.
41. Everett MV, Grau ED, Seeb JE (2011) Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. Molecular Ecology Resources 11: 93–108.
42. Liu SK, Zhou ZC, Lu JG, Sun FY, Wang SL, et al. (2011) Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12: 13.
43. Vera M, Alvarez-Dios JA, Milian A, Pardo BG, Bouza C, et al. (2011) Validation of single nucleotide polymorphism (SNP) markers from an immune Expressed Sequence Tag (EST) turbot, Scophthalmus maximus, database. Aquaculture 313: 31–41.
44. Wang S, Sha Z, Sonstegard T, Liu H, Xu P, et al. (2008) Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics 9: 450.
45. Hubert S, Higgins B, Borza T, Bowman S (2010) Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua). BMC Genomics 11: 14.
46. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376–380.
47. Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, et al. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10: 19.
48. Kircher M, Kelso J (2010) High-throughput DNA sequencing - concepts and limitations. Bioessays 32: 524–536.
49. Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Briefings in Bioinformatics 11: 457–472.
50. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, et al. (2007) The Use of Coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by 454 Parallel Sequencing. PLoS ONE 2: e197.
51. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome research 18: 1509–1517.
52. Huang X, Madan A (1999) CAP3: A DNA Sequence Assembly Program. Genome research 9: 868–877.
53. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, et al. (1999) A general approach to single-nucleotide polymorphism discovery. Nature genetics 23: 452–456.
54. Birol İ, Jackman SD, Nielsen CB, Qian JQ, Varhol R, et al. (2009) De novo transcriptome assembly with ABySS. Bioinformatics 25: 2872–2877.
55. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, et al. (2009) ABySS: A parallel assembler for short read sequence data. Genome research 19: 1117–1123.
56. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research 36: 3420–3435.
57. Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, et al. (2005) Biological profiling of gene groups utilizing Gene Ontology. Genome Inform 16: 106–115.
58. Huang W, Marth G (2008) EagleView: A genome assembly viewer for next-generation sequencing technologies. Genome research 18: 1538–1543.
59. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, et al. (2003) Highly Parallel SNP Genotyping. Cold Spring Harbor symposia on quantitative biology 68: 69–78.
60. Ferraresso S, Milan M, Pellizzari C, Vitulo N, Reinhardt R, et al. (2010) Development of an oligo DNA microarray for the European sea bass and its application to expression profiling of jaw deformity. BMC Genomics 11: 17.
61. Zhang LQ, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Molecular Biology and Evolution 21: 236–239.
62. Kerstens HHD, Crooijmans R, Veenendaal A, Dibbits BW, Chin-A-Woeng TFC, et al. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics 10: 11.
63. Kraus RHS, Kerstens HHD, Van Hooft P, Crooijmans R, Van der Poel JJ, et al. (2011) Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). BMC Genomics 12: 11.
64. Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759–769.

Supporting Information

Figure S1

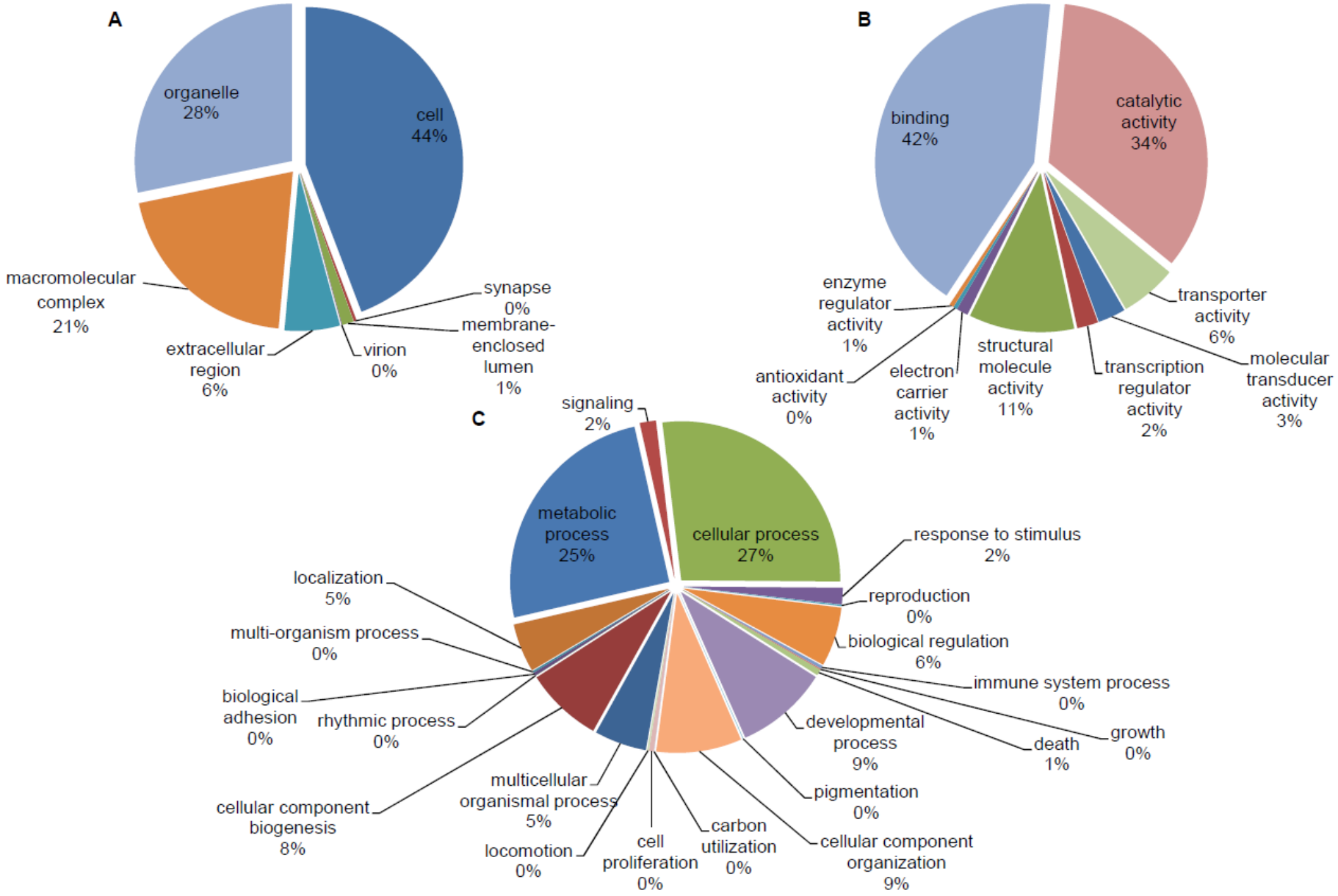


Figure S2

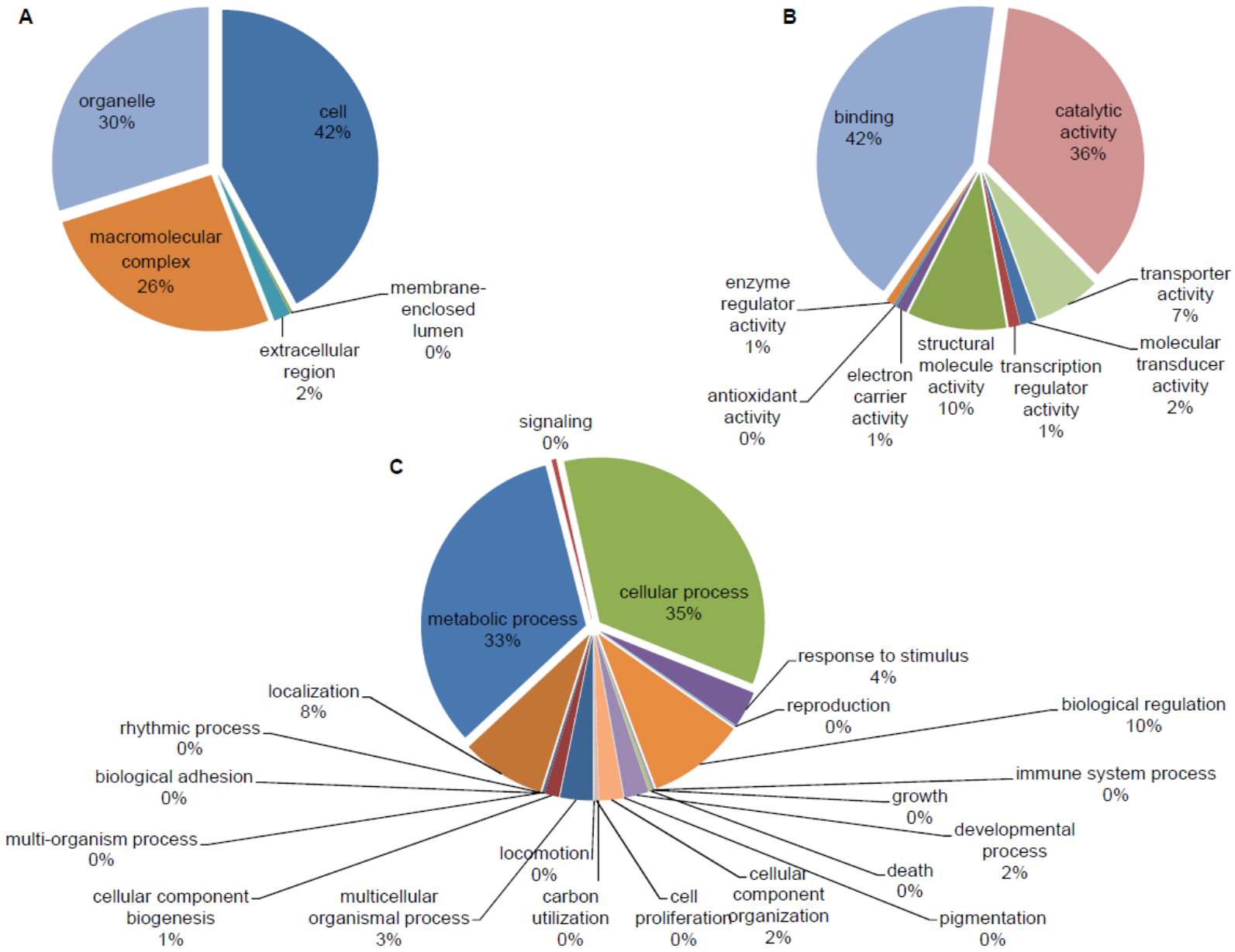

A
organelle 30%
cell 42%
macromolecular complex 26%
membrane-enclosed lumen 0%
extracellular region 2%
B
binding 42%
catalytic activity 36%
transporter activity 7%
molecular transducer activity 2%
transcription regulator activity 1%
structural molecule activity 10%
electron carrier activity 1%
antioxidant activity 0%
enzyme regulator activity 1%
C
signaling 0%
cellular process 35%
metabolic process 33%
response to stimulus 4%
reproduction 0%
biological regulation 10%
immune system process 0%
growth 0%
developmental process 2%
death 0%
pigmentation 0%
cellular component organization 2%
cell proliferation 0%
carbon utilization 0%
locomotion 0%
multicellular organismal process 3%
cellular component biogenesis 1%
multi-organism process 0%
biological adhesion 0%
rhythmic process 0%
localization 8%

Figure S3

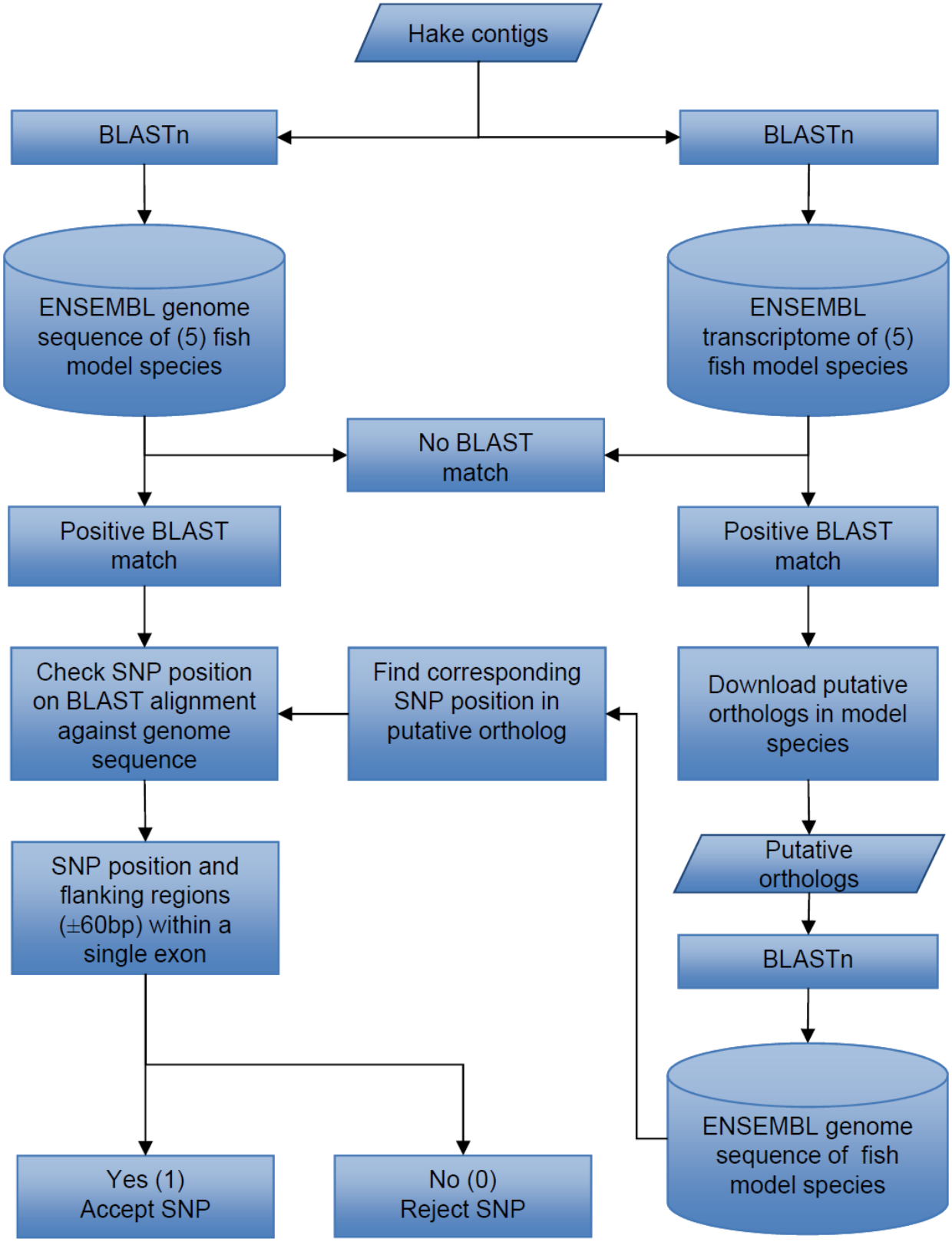

Hake contigs
BLASTn
BLASTn
ENSEMBL genome sequence of (5) fish model species
ENSEMBL transcriptome of (5) fish model species
No BLAST match
Positive BLAST match
Positive BLAST match
Check SNP position on BLAST alignment against genome sequence
Find corresponding SNP position in putative ortholog
Download putative orthologs in model species
SNP position and flanking regions (±60bp) within a single exon
Putative orthologs
BLASTn
ENSEMBL genome sequence of fish model species
Yes (1) Accept SNP
No (0) Reject SNP

Table S1

GO-ID	Term	Category	FDR	P-Value	Over/Under
GO:0006412	translation	P	2.40E-04	0.0	over
GO:0003735	structural constituent of ribosome	F	2.40E-04	0.0	over
GO:0005840	ribosome	C	2.40E-04	0.0	over
GO:0030529	ribonucleoprotein complex	C	2.40E-04	1.32E-08	over
GO:0009058	biosynthetic process	P	2.40E-04	6.80E-08	over
GO:0034645	cellular macromolecule biosynthetic process	P	2.40E-04	9.78E-08	over
GO:0009059	macromolecule biosynthetic process	P	2.40E-04	9.78E-08	over
GO:0044249	cellular biosynthetic process	P	2.40E-04	1.88E-07	over
GO:0044444	cytoplasmic part	C	5.72E-03	2.62E-04	over
GO:0043232	intracellular non-membrane-bounded organelle	C	1.00E-01	4.90E-03	over
GO:0043228	non-membrane-bounded organelle	C	1.00E-01	4.90E-03	over
GO:0005198	structural molecule activity	F	7.90E+00	3.91E-01	over
GO:0032991	macromolecular complex	C	0.00200569	8.51E+00	over
GO:0033279	ribosomal subunit	C	0.0113124	5.11E+01	over
GO:0006091	generation of precursor metabolites and energy	P	0.0113124	5.19E+01	over
GO:0016868	intramolecular transferase activity, phosphotransferases	F	0.02043	8.30E+01	over
GO:0044424	intracellular part	C	0.0231671	9.97E+01	over
GO:0005975	carbohydrate metabolic process	P	0.0332324	0.00171534	over
GO:0022900	electron transport chain	P	0.0332324	0.00178515	over
GO:0016866	intramolecular transferase activity	F	0.0344973	0.00197345	over
GO:0006414	translational elongation	P	0.0350837	0.00206099	over
GO:0030170	pyridoxal phosphate binding	F	0.0422907	0.00290272	over
GO:0022904	respiratory electron transport chain	P	0.0422907	0.00290272	over
GO:0070279	vitamin B6 binding	F	0.0422907	0.00290272	over
GO:0008152	metabolic process	P	0.0422907	0.00292836	over

Table S2

GO-ID	Term	Category	FDR	P-Value	Over/Under
GO:0003735	structural constituent of ribosome	F	5.73E-01	9.95E-03	over
GO:0030529	ribonucleoprotein complex	C	5.73E-01	1.59E-02	over
GO:0032991	macromolecular complex	C	5.73E-01	1.67E-02	over
GO:0005840	ribosome	C	8.87E-01	4.06E-02	over
GO:0043228	non-membrane-bounded organelle	C	8.87E-01	5.62E-02	over
GO:0043232	intracellular non-membrane-bounded organelle	C	8.87E-01	5.62E-02	over
GO:0005198	structural molecule activity	F	2.57E+00	1.71E-01	over
GO:0016459	myosin complex	C	9.29E+00	6.18E-01	over
GO:0006412	translation	P	1.38E+00	1.11E+00	over
GO:0034645	cellular macromolecule biosynthetic process	P	1.38E+00	1.28E+00	over
GO:0009059	macromolecule biosynthetic process	P	1.38E+00	1.28E+00	over
GO:0003779	actin binding	F	1.41E+01	1.55E+00	over
GO:0008092	cytoskeletal protein binding	F	1.41E+01	1.55E+00	over
GO:0003774	motor activity	F	4.85E+01	4.79E-01	over
GO:0032982	myosin filament	C	5.64E+01	6.45E+00	over
GO:0044430	cytoskeletal part	C	8.65E+01	9.14E+00	over
GO:0044424	intracellular part	C	8.98E+01	1.03E+01	over
GO:0044422	organelle part	C	0.00153193	2.07E+01	over
GO:0044446	intracellular organelle part	C	0.00153193	2.07E+01	over
GO:0016818	hydrolase activity, in phosphorus-containing anhydrides	F	0.00512314	6.35E+01	over
GO:0017111	nucleoside-triphosphatase activity	F	0.00512314	6.35E+01	over
GO:0016462	pyrophosphatase activity	F	0.00512314	6.35E+01	over
GO:0016817	hydrolase activity, acting on acid anhydrides	F	0.00644593	8.50E+01	over
GO:0005515	protein binding	F	0.00721798	0.00109514	over
GO:0044464	cell part	C	0.0107931	0.00153039	over
GO:0044249	cellular biosynthetic process	P	0.0139449	0.00177106	over
GO:0009058	biosynthetic process	P	0.0222579	0.00294704	over
GO:0005524	ATP binding	F	0.0401276	0.00505448	over
GO:0032559	adenyl ribonucleotide binding	F	0.0401276	0.00505448	over
GO:0032555	purine ribonucleotide binding	F	0.0434069	0.00597614	over
GO:0032553	ribonucleotide binding	F	0.0434069	0.00597614	over
GO:0016787	hydrolase activity	F	0.047443	0.00745794	over
GO:0043226	organelle	C	0.047443	0.00751754	over
GO:0043229	intracellular organelle	C	0.047443	0.00751754	over
GO:0030554	adenyl nucleotide binding	F	0.047443	0.0080511	over
GO:0001883	purine nucleoside binding	F	0.047443	0.0080511	over
GO:0001882	nucleoside binding	F	0.047443	0.0080511	over

CHAPTER 5

SNP MARKERS ON CANDIDATE GENES FOR LOCAL ADAPTATION REVEAL FINE-SCALE GENETIC STRUCTURE AMONG EUROPEAN HAKE (MERLUCCIUS MERLUCCIUS) POPULATIONS

SNP markers on candidate genes for local adaptation reveal fine-scale genetic structure among European hake (*Merluccius merluccius*) populations

Authors:

Ilaria Milano[1], Massimiliano Babbucci[2], Fausto Tinti[1], Einar E. Nielsen[3], Gary R. Carvalho[4], Tomaso Patarnello[2], FishPopTrace Consortium, Luca Bargelloni[2].

Affiliations:

[1] Department of Experimental and Evolutionary Biology, University of Bologna, via Selmi 3, 40126 Bologna, Italy.

[2] Department of Public Health, Comparative Pathology, and Veterinary Hygiene, University of Padova, viale dell'Università 16, 35020 Legnaro, Italy.

[3] National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK- 11 8600 Silkeborg, Denmark.

[4] Molecular Ecology and Fisheries Genetics Laboratory (MEFGL), School of Biological Sciences, University of Bangor, Environment Centre Wales, Bangor, Gwynedd LL57 2UW, UK.

Abstract

Population structure of marine fish is shaped by the interplay of different evolutionary mechanisms. Despite high dispersal and connectivity and large population size promote wide-scale genetic homogeneity, detected at neutral genetic markers, heterogeneity of environmental conditions might favour divergence among differentially adapted populations. In this study we analysed 381 EST-derived SNPs on 5 Atlantic and 14 Mediterranean samples of European hake (*Merluccius merluccius*), a widely distributed demersal species, in order to find candidate genes for environmental adaptation at different spatial scales and to integrate information on neutral and adaptive evolutionary patterns.

Over a wide spatial range 299 neutral markers confirmed a major genetic break between Atlantic and Mediterranean populations (F_{CT}=0.01559), and the occurrence of a limited but ongoing unidirectional gene flow into Mediterranean, across a transition zone in the Alboran Sea. Higher resolution of population differentiation between basins was observed at 17 outlier SNPs (F_{CT} between 0.275 and 0.705), possibly representing signatures of divergent selection, as suggested by the strong correlation between allele frequencies and environmental data. At a regional scale, basin-specific outlier SNPs resolved finer population structure, as opposed to the weak genetic structuring ascertained at neutral loci, separating North Sea and Portugal from all other Atlantic samples and revealing a clear differentiation among Western, Central and Eastern geographic clusters within Mediterranean. Nevertheless, only a subset of outlier loci showed significant association to environmental parameters variation.

Our results suggest that local adaptation might play a role in shaping population structure of a high gene flow marine fish species as European hake. Although caution should be taken when drawing indirect inferences about adaptive processes in the wild, in this study we found evidence of outlier SNP markers that strongly separate European hake geographic populations at fine scale. Accordingly, we suggest that current stock management policy necessitates to be reassessed. In particular, precautionary management strategies should contemplate the possible presence of locally adapted populations, in order to favour evolutionary resilience to environmental changes.

Introduction

In 'classical' marine fish species high levels of gene flow have been conventionally assumed to constrain the potential for local adaptation, by rapidly homogenizing allelic frequency between populations. However, in the last years this traditional paradigm has been overturned by the increasing evidence that locally self-recruiting populations may occur in marine systems (Cowen& Sponaugle 2009; Ruzzante *et al.* 2006; Skjæraasen *et al.* 2011; Swearer *et al.* 2002), and that weak neutral genetic structures may result from low differentiation via genetic drift due to large population sizes, rather than from actually high migration rates (Hauser& Carvalho 2008; Waples 1998). In such a scenario, adaptive divergence might arise even in the apparent genetic homogeneity revealed by neutral markers (Hemmer-Hansen *et al.* 2007).

The new perception on local adaptation in the marine environment not only brings back the attention on the weight of ecological factors on evolutionary processes, but have also significant implications for conservation purposes, particularly for harvested species, relating to the identification of management units and the preservation of species' adaptive potential (Allendorf *et al.* 2010; Conover *et al.* 2006; Fraser& Bernatchez 2001). The concept of Evolutionary Significant Unit (ESU), defined as the population unit of conservation interest, embraces the principle of reproductive and historical isolation as well as adaptive divergence (Crandall *et al.* 2000). In this perspective, patterns of neutral and adaptive genetic variation should be integrated, in order to maintain diversity between independently evolving units while preserving variability among groups that are uniquely adapted to existing environmental conditions.

Most of recent studies aiming at identifying genetic footprint of selection in marine fish have focused on specific genes having know structural functions or being involved in physiological processes (Andersen *et al.* 2011; Blel *et al.* 2010; Gaggiotti *et al.* 2009; Hemmer-Hansen *et al.* 2007; Larmuseau *et al.* 2009; Larmuseau *et al.* 2010). Targeting candidate genes has become a feasible strategy also for non-model organisms, as genetic information can be achieved from related species using a comparative approach. However, the choice of the right candidate gene indeed represent the major challenge, and may reveal

inadequate in relation to biological and evolutionary differences among species (Zhu& Zhao 2007), making this approach time-consuming and occasionally inefficient (Boutet *et al.* 2008).

Alternatively, genome scan methods based on the screening of high number of genetic markers provide a powerful system to identify outlier loci with higher genetic divergence among natural populations than neutral expectations (Storz 2005), increasing the chance to target candidate gene regions for divergent selection. The application of population genomic approaches to detect potential signatures of local adaptation in non-model organisms have been progressively favored by the growing accessibility of next-generation sequencing technologies (Nielsen *et al.* 2009b). Large panels of molecular markers may be developed in and around functional genes by high-throughput trancriptome sequencing, thus focusing the search in coding regions (Bonin 2008; Stapley *et al.* 2010). Despite the phenotypic effects of any particular gene and the corresponding consequences for fitness are hardly predictable (Naish& Hard 2008), genome scans may be valuable as exploratory searches for candidate genes under selection. Further clues on the adaptive role of outlier loci can be derived by combining information on putative gene function as well as potential association between allele frequencies and environmental variables, using complementary approaches (Vasemägi& Primmer 2005). Currently, large scale genome scans revealing the occurrence of potentially adaptive population divergence in a high gene flow marine species have been limited to cod (*Gadus morhua*) (Moen *et al.* 2008; Nielsen *et al.* 2009a).

The European hake (*Merluccius merluccius*) is a widely distributed marine fish, inhabiting spatially extensive and environmentally heterogeneous regions, spanning from the North Sea in Atlantic to the Levantine Sea in Mediterranean. Atlantic and Mediterranean populations show different demographic and life-history traits, such as growth rate (Belcari *et al.* 2006; Kacher& Amara 2005; Mellon-Duval *et al.* 2009; Piñeiro *et al.* 2008), size at maturity (Biagi *et al.* 1995; Domínguez-Petit *et al.* 2008; Papaconstantinou& Stergiou 1995; Piñeiro& Saínza 2003; Recasens *et al.* 2008; Recasens *et al.* 1998) and spawning season (Alvarez *et al.* 2004; Alvarez *et al.* 2000; Arneri& Morales-Nin 2000; Papaconstantinou& Stergiou 1995; Recasens *et al.* 1998). Moreover, a high variability in spawning peak seasons has been observed across different geographic areas within both basins.

Previous genetic studies based on microsatellites have consistently defined a major subdivision between Mediterranean and Atlantic (Castillo *et al.* 2005; Castillo *et al.* 2004; Lundy *et al.* 1999; Pita *et al.* 2010), while the existence of sub-structuring within basins remains controversial. In Atlantic most of efforts has been focused on the ascertainment of genetic structure between populations from Northern and Southern stocks, as recognized by the International Council for the Exploration of the Sea (ICES). While most of studies did not recognize the effectiveness of the subdivision between the two management areas (Castillo *et al.* 2005; Lundy *et al.* 1999; Pita *et al.* 2011), Castillo et al. (2004) found significant differentiation between southern Bay of Biscay and Celtic Sea locations. Genetic structure in Mediterranean has not been extensively investigated using microsatellites. Slight signals of genetic structure emerged among three locations from Western, Central and Eastern Mediterranean (Castillo *et al.* 2004), despite no differentiation was previously found between Adriatic Sea and Tunisia (Lundy *et al.* 1999). The low number (five or six) and the inadequacy of molecular markers used could limit the potential to effectively resolve population genetic structure, particularly at local scale. Interestingly, studies based on allozyme markers revealed very high levels of differentiation at specific loci both between and within Atlantic and Mediterranean basins (Cimmaruta *et al.* 2005; Roldan *et al.* 1998), leaving an open door on the possible role of selection in shaping European hake genetic structure.

This study present the first evaluation of a panel of SNP markers recently derived from European hake transcriptome sequences (Milano *et al.* 2011), based on an extensive sampling design across Atlantic and Mediterranean. The purpose of our work was to improve the resolving power in detecting spatial structuring on a broad and local scale, using novel and more efficient genetic tools, and provide novel insights in functional regions potentially involved in local adaptation of European hake populations. Different genome scan approaches were applied to detect polymorphisms on candidate genes potentially under divergent selection. A supplementary method was used to verify the possible adaptive role of outlier loci, testing the association between allele frequencies and critical factors in marine environment. Finally, the spatial genetic structure at different spatial scales was examined comparing patterns emerging from putative neutral and non-neutral loci.

Materials and Methods

Population sampling and SNP genotyping

An extensive sampling effort over a broad geographic area has been carried out between 2008 and 2009. A total of 19 samples (n = 1072) were collected from scientific surveys and commercial fishing vessels across the species range, 5 from Atlantic and 14 from Mediterranean locations (Figure 1 and Table 1), spanning steep ecological clines. Abbreviations cited in the text referring to the sampling locations are specified in Table 1.

DNA was extracted from finclip tissues stored in 96% ethanol using the commercial kit Invisorb DNA Universal Clinical HTS 96 Kit/V (Invitek). Quality and concentration of the extracted DNA was checked using a NanoDrop spectrophotomer, in order to optimize the SNP genotyping outcome. A total of 1000 individuals were screened for a panel of 395 EST-derived SNPs (Milano *et al.* 2011), using the Illumina GoldenGate Assay platform (Fan *et al.* 2003). Genotyping data were visualized and examined with the Genome Studio Data Analysis Software package (1.0.2.20706, Illumina Inc.). After excluding individuals with low-quality results (average call rate across loci < 0.8), SNPs were manually re-clustered to obtain highly accurate genotype data. Before proceeding to the actual statistical analysis, a further data quality filter based on the percentage of missing data was applied, removing from each sample individuals with more than 10% missing genotypes overall loci.

Detection and characterization of outlier SNPs

All 381 polymorphic SNPs were screened to detect outlier loci at different spatial scales, analyzing all geographic samples together and Atlantic and Mediterranean samples separately. To provide a more robust statistical support to our analysis we applied two genome-scan approaches, based on independent methods.

The method implemented in the software Arlequin 3.5 (Excoffier& Lischer 2010) uses coalescent simulations to obtain a null distribution of F_{ST} or F_{CT} across loci, depending on the assumed demographic model (finite island or hierarchical island models, respectively), as a function of heterozygosity. Loci showing higher or lower differentiation respect to the simulated confidence intervals are identified as

candidates for divergent or balancing selection (Beaumont& Nichols 1996; Excoffier *et al.* 2009). Excoffier *et al.* proved that the hierarchical island model reduces the excess of false positives that arises when samples belonging to a structured population are analyzed under a finite island model. In order to consider the genetic structure between Atlantic and Mediterranean populations of European hake documented from previous studies, large-scale outlier detection analysis were carried out under the assumption of a hierarchical island model, grouping samples in relation to the basin of origin. The finite island model was used to run analysis on the subset of samples within each basin. We ran 50,000 simulations and assumed 100 demes per group, simulating 10 groups under the hierarchical model assumption.

The Bayesian method implemented in the software Bayescan 2.0 (Foll& Gaggiotti 2008) tests departure from neutrality evaluating the weight of the locus-specific contribution respect to the population-specific effect, shared by all loci, on the observed pattern of genetic diversity: the posterior probability for a locus being under selection is calculated under two alternative models, including or not natural selection. This approach appeared to be robust under complex demographic scenarios (Foll& Gaggiotti 2008), and was found to have lower type I (false positive) and II (false negatives) error rates for divergent selection compared to other outlier detection methods (Narum& Hess 2011). We based our analyses on 20 pilot runs each consisting of 5,000 iterations, followed by 100,000 iterations with a burn-in of 50,000 iterations; the prior odds for the neutral model was set to 10, as suggested for the identification of candidate loci with a few hundreds of markers. Posterior Odds (PO), indicating how more likely the model including selection is compared to the neutral model, were interpreted according to the Jeffreys' scale of evidence for Bayes Factors (Jeffreys 1961). Following this method, a $\log_{10}$PO between 1.5 and 2 denotes a very strong evidence for selection, while values higher than 2 indicate a decisive signal.

Available information on outlier loci characterization, including the gene annotation, the type of mutation and the amino acid change determined at the protein level, were retrieved (Milano *et al.* 2011).

Associations between genetic and environmental variation

We used a Bayesian approach testing for correlation between allele frequency variation observed at outlier SNPs and environmental factors, implemented in the software Bayenv (Coop *et al.* 2010), to further corroborate results obtained with genome-scan analysis. The method proposed by Coop et al. first estimates a null model based on neutral markers describing how allele frequencies covary across populations, and subsequently test whether the correlation observed between allele frequencies at specific markers of interest and an environmental variable is higher than expected under this null model. In this way, the underlying population structure is taken into account, and the probability to find false positive due to shared population history or gene flow is reduced. The software provides the Bayes Factor as a measure in support to the model including a significant correlation. We used putative neutral unlinked SNPs to calculate the covariance matrix, and then for each SNP the Bayes Factor was calculated respect to environmental parameters considered, following the multiple spatial scale approach adopted in outlier analysis. Independent runs were carried out to ensure that results were not sensitive to stochastic errors. Results were evaluated comparing the distribution of Bayes Factor values across putatively neutral and outlier loci and according to the Jeffreys' scale of evidence (Jeffreys 1961).

Environmental data were retrieved from the NODC (National Oceanographic Data Center) database, referring to geographic coordinates as close as possible to our sampling locations for which data were available (Table 2). Annual climatological mean values of temperature and salinity in surface and at 100m depth were acquired. We considered data at two depth levels to account for the different bathymetrical distribution of various life-stages of European hake. NODC statistics on temperature and salinity parameters are based on long-term observations (50 years, from 1955 to 2006), and records vary across a quarter-degree latitude-longitude grid (Antonov *et al.* 2010; Locarnini *et al.* 2010).

Genetic diversity, Hardy-Weinberg equilibrium and Linkage disequilibrium

Allele frequencies, expected (He) and observed (Ho) heterozygosity estimates were calculated using the package GenAlEx 6.41 (Peakall& Smouse 2006). Allelic richness was calculated using the method

implemented in FSTAT v 2.9.3 (Goudet 1995). We estimated departure from Hardy Weinberg Equilibrium (HWE) for each locus in each population using the exact tests implemented in Genepop 4.1.0 (10,000 dememorizations, 100 batches and 5,000 iterations per batch) (Raymond& Rousset 1995). Linkage disequilibrium for each pair of SNPs was tested in each population using the probability test implemented in Genepop 4.1.0 (Raymond& Rousset 1995). Type I error rates for both tests were corrected for multiple comparisons using the sequential Bonferroni procedure (Rice 1989).

Population genetic structure: neutral vs outlier divergence

Population structure analysis were carried out in parallel on putative neutral and outlier loci to compare different patterns of population differentiation.

Arlequin 3.5 was used to perform a locus by locus AMOVA using neutral and global outlier data sets, to assess the genetic structure between Atlantic and Mediterranean populations. Moreover, we used the same software to calculate pairwise Fst between all pairs of samples based on neutral SNPs, using 10,000 permutations to test for significance. The sequential Bonferroni method (Rice 1989) for multiple comparisons was applied to correct the significance level.

Relative genetic distances among samples were also explored by Principal Coordinate Analysis (PCoA) using GenAlEx 6.41 (Peakall& Smouse 2006), based on a matrix of Edwards' genetic distance indexes (Edwards 1971) calculated with Adegenet R package (Jombart 2008).

To infer main genetic clusters a comparative approach was applied, based on the Bayesian clustering method implemented in Structure 2.3.3 (Pritchard *et al.* 2000) and the discriminant analysis of principal components (DAPC) performed with Adegenet R package (Jombart 2008).

Structure uses a Bayesian algorithm to infer the number of distinct K clusters of individuals, based on their multilocus genotypes, assuming Hardy-Weinberg and Linkage Equilibrium. A posterior probability for each inferred K is calculated, allowing to estimate the more likely number of clusters. The algorithm was run assuming the admixture model and correlated allele frequencies among populations, and providing the sampling information as prior, in order to improve accuracy in detecting population structure (Hubisz *et al.* 2009). For each analysis we used 6 iterations per K value, a burnin period length

of 50,000, and 100,000 MCMC repetitions. The Evanno method (Evanno *et al.* 2005) was used to identify the most likely number of clusters accounting for the observed genetic structure. A final long run was performed (10 iterations, 500,000 iterations, 100,000 burn-in) based on the most probable value of K. We combined results from multiple runs at the optimal K using the CLUMPP software v.1.1.2 (Jakobsson& Rosenberg 2007), and obtained the graphical output with Distruct (Rosenberg 2004).

To corroborate the genetic structure inferred from Bayesian clustering, we used the DAPC approach (Jombart *et al.* 2010), a multivariate method that does not rely on specific population genetics models. According to this method, genetic data are first transformed using principal component analysis (PCA) into components explaining most of the genetic variation; these components are used to perform a linear discriminant analysis (DA), that provides descriptive variables of genetic groups minimizing the genetic variance within populations while maximizing among-population variation. The R package Adegenet was used to perform DAPC.

Results

Genotyping

A total of 98 individuals were initially excluded due to scarce genotyping results. Moreover, 52 individuals spread across 15 samples showing more than 10% missing genotypes overall loci were removed, and 850 individuals representing 19 sampling locations (n=19-66) were included in the final dataset (Table1).

Genetic diversity, Hardy-Weinberg equilibrium and Linkage disequilibrium

After evaluating single locus genetic diversity across samples, 14 SNPs showing extremely low levels of polymorphism (minor allele frequency overall populations lower than 0.01) were excluded, and a set of 381 SNPs remained for further analysis. These markers had a mean missing data rate across individuals of 1.3%, and for almost all of them (96.5%) the proportion of missing genotypes did not exceed 5%. While all 381 SNPs were assayed to identify outlier loci, results from Hardy-Weinberg and Linkage Equilibrium tests were considered to fix data sets for population genetic structure analysis. A total of 6510 tests for departure from HWE were performed, excluding monomorphic SNPs within population. Two loci showing significant deviations in more than half of the populations after sequential Bonferroni correction were excluded. A total of 29 pairs of loci in significant linkage disequilibrium after correction for multiple testing were identified in at least 10 out of 19 samples. Loci were sorted into 20 linkage groups, each including two or three markers. From each linkage group, the most informative locus was retained, based on the Shannon-Weaver's index, and SNPs involved in multiple groups were preferentially discarded. Similar levels of observed and expected heterozigosity estimates overall loci (H_o=0.266-0.307 and H_e=0.266-0.296) were found across populations. The percentage of polymorphic SNPs varied between 87.9% (SAD, n=19) and 97.64% (ADB, n=66) (Table 1).

Outlier detection

On a global scale, results from the Arlequin hierarchical method revealed 17 candidate SNPs for divergent selection between Atlantic and Mediterranean at the 1% significance level (F_{CT} between 0.275

and 0.705). Based on the same set of samples, we detected a higher proportion of outlier loci using Bayescan: a total of 58 SNPs, including all markers suggested by Arlequin, showed decisive ($log_{10}PO>2$) evidence for divergent selection (Figure 2 A and B). On the other hand, when screening the panel of 381 SNPs over a local spatial scale, Bayescan identified a subset of Arlequin candidate loci within basins. Both approaches revealed 4 outlier SNPs showing unusually high levels of differentiation among 5 Atlantic samples (F_{ST} between 0.126 and 0.298), even considering more stringent significance thresholds (Figure 2 C and D), and supported 16 candidate SNPs within Mediterranean (F_{ST} between 0.038 and 0.214) with high statistical significance (Figure 2 E and F). The relatively low number of outlier loci found within Atlantic might be biased by the limited number of sampled populations. Foll and Gaggiotti confirmed that the number of population plays an important role for identification of outliers, and observed that 6 populations are usually sufficient to have a good true-positive rate for directional selection (Foll& Gaggiotti 2008). Comparing our findings at different spatial scales, 8 out of 17 substantial candidate SNPs between basins were also identified within Mediterranean, and only one locus was detected independently within both basins.

Based on results of outlier detection, a global neutral data set including 299 putative loci and all 19 samples was obtained. From the complete set of 381 SNPs, we first excluded 69 potential outliers identified in at least one of the tests performed, applying less stringent selection criteria ($p<0.05$ for Arlequin and $log_{10}PO>1.5$ for Bayescan results). Afterward, 2 and 11 loci were further removed, due to Hardy-Weinberg and linkage disequilibrium, respectively.

In order to compare signals occurring from potentially adaptive and neutral markers, three outlier data sets were also constructed, in relation to the spatial scale explored. We adopted stringent criteria for selection of outlier loci ($p<0.01$ for Arlequin and $log_{10}PO>2$ for Bayescan results), to reduce the proportion of false positive. The global outlier data set, including all populations, comprised 17 outlier SNPs supported by both approaches, while basin-specific data sets consisted of 7 and 19 unlinked loci detected by at least one of the two methods within Atlantic and Mediterranean, respectively. We used a more conservative criterion to select candidate loci on a global scale because we expected a higher number of false positive due to the underlying genetic population structure between Atlantic and

Mediterranean. Moreover, unlike results within basin, we observed a substantial difference between the two approaches in the number of candidate loci: the hierarchical model assumed in Arlequin analysis, more appropriate to represent our large-scale scenario, likely provided a less biased estimate.

Candidate loci used for population genetic analysis are listed in Table 3, together with information on gene annotation and the type of mutation: most of annotated outliers were synonymous, while only 4 SNPs resulted in an amino acid change. Non-synonymous mutations were located on genes codifying for nexilin, hemoglobin and AMP deaminase proteins. The titin gene numbered the highest overall number of outlier SNPs (5).

Associations between genetic and environmental variation

After evaluating allele frequency variation across the whole spatial range, all 17 outlier SNPs showed a decisive ($log_{10}BF>2$) or very strong ($log_{10}BF>1.5$) evidence for correlation with temperature and salinity values at surface and 100m depth (Table 3, Figure 3 A and B), except for 3 loci showing $log_{10}BF$ between 1 and 1.5. The strongest locus–environmental variables association characterized the SNP 2186_fpt, a non-synonymous mutation on the hemoglobin gene. High levels of correlation with both salinity and temperature were also observed at loci 3520_fpt (synaptopodyn2-like) and 151_fpt (titin) (Figure 3 A and B). Focusing on a regional scale, in Atlantic 2 SNPs strongly correlated with ecological parameters, one not-annotated and one synonymous mutation (109ms) located on a PDZ/LIM gene, while a less significant association, but higher than neutral average, was found for 2 synonymous SNPs located on the titin and adp/atp translocase genes (927_fpt and 890_fpt) (Figure 3 C and D). Analysis within Mediterranean revealed a complex pattern of correlation between allele frequencies at outlier loci and ecological factors. Highly significant correlations ($log_{10}BF>2$) were observed respect to temperature variation at 100m depth at one not-annotated and one non-synonymous SNPs, the latter (3656_fpt) located on the subunit beta of the hemoglobin gene. However, the same markers showed much lower $log_{10}BF$ values when surface temperature was considered. An analogous situation was observed for three SNPs correlating to a lesser degree with salinity values (Table 3, Figure 3 E and F). This pattern might be due to the heterogeneous climatic and oceanographic properties of the Mediterranean Sea, as

well as to the complex water circulation pattern which determine local environmental conditions and different temperature and salinity depth-profiles.

None of the 299 neutral SNPs significantly correlated with any of the ecological factors tested at different geographic scales (Figure 3).

Genetic population structure at neutral loci

Low overall estimates of genetic differentiation were observed at 299 putative neutral markers, ranging between -0.006 and 0.073, with a mean F_{ST} value of 0.008. We found highly significant ($p<0.005$) differentiation in all pairwise comparisons between Atlantic and Mediterranean samples. F_{ST} values ranged between 0.004 and 0.028, with lower estimates observed between Atlantic populations and samples from the Alboran Sea (ALG and MAL) (Table 4). Shallower genetic differentiation was detected within basins, although we noticed some statistically significant comparisons. In Atlantic, we did not find evidence of significant divergence among WSC, SIR and GAC, while the NTS showed signals of substantial differentiation from SIR, GAC and NPT (F_{ST} between 0.002 and 0.005, $p<0.05$). In the Mediterranean Sea most of significant comparisons involved the ALG population and samples from the Eastern basin (AEG, TKY and CYP) (Table 4). The SAD showed negative F_{ST} values respect to all Mediterranean samples excluding the ALG, likely due to a bias induced from the small sample size (n=19). The PCoA analysis mainly confirmed these results, as nearly half of variance (41%) distinguished the two basins and placed the ALG sample in an intermediate position between them, while the second principal component, accounting for the 13% of variance, differentiated samples within Atlantic and Mediterranean. In accordance to previous studies based on 5 or 6 microsatellite loci (Lundy *et al.* 1999; Pita *et al.* 2010), the AMOVA revealed that most of the total variance (98.24%) was attributable to variation within populations (F_{ST} = 0.01756), 1.56% to variance among groups (F_{CT} = 0.01559) and 0.02% among populations within groups (Table 5). The low but statistically significant differentiation between Atlantic and Mediterranean was corroborated from Bayesian clustering analysis, that suggested K=2 as the most likely scenario: the signal of a steady unidirectional gene flow was evident from the marked west-east introgression of the Atlantic genetic component into Mediterranean (Figure 4A).

Genetic population structure at outlier SNPs

Analysis of outlier data sets increased the resolving power to detect fine-scale population structure, due to the higher levels of genetic differentiation. The panel of 17 global outlier SNPs provided an extremely strong signal of genetic structure between Atlantic and Mediterranean (F_{CT}=0.41883) respect to the average observed at putative neutral loci (Table 5), likely reflecting populations connectivity. At a regional scale, basin-specific outlier data sets unveiled the presence of well differentiated groups within Atlantic and Mediterranean, despite the high levels of gene flow found at neutral loci. In Atlantic, three main genetic clusters were identified from Structure analysis (K=3): the NTS and NPT were genetically divergent from a major cluster including the WSC, SIR and GAC samples (Figure 4B). This subdivision was also confirmed from DAPC analysis (Figure 5A) and PCoA (Figure 6). Bayesian clustering analysis within Mediterranean clearly revealed a spatial genetic heterogeneity, identifying three genetic components (K=3), mainly corresponding to Eastern (AEG, TKY and CYP), Central (STY, NWJ, SAD, NAD) and Western clusters (ALG, MAL, GLF, NTY, SSD, ADB, MAB) (Figure 4C). Genetic groups did not exactly coincide with Mediterranean main sub-areas, as the STY sample, located in the Western Mediterranean, grouped with the Central cluster, while samples from the Sicilian Channel (ADB and MAB), geographically belonging to the Central Mediterranean, were related to the Western cluster (Figure 5B and 6C). Our results showed the ALG sample corresponded to a single and well-defined genetic group, separated from the rest of Western Mediterranean samples, in which the contribution of Central and Eastern clusters was evident to varying degrees. Although most of individuals clustered into their respective population, we observed some samples from Cyprus location having a genetic background similar to the one observed in Central Mediterranean populations (Figure 4C).

Discussion

Despite the spatial genetic structure of European hake has been extensively examined, the efforts of previous studies to elucidate its population dynamics have been limited by the inadequacy and low discriminating power of molecular markers, primarily microsatellites and mitochondrial sequences. In this study we screened a panel of 381 newly discovered SNPs, and identified highly informative loci, following a genome-scan approach, to detect genetic structure of European hake populations at both wide and regional spatial scales. The application of outlier markers in fisheries management has been recently drawn to attention (Ackerman *et al.* 2011; Freamo *et al.* 2011; Russello *et al.* 2011), proving more fruitful respect to conventional genetic tools for designating conservation units. Based on the comparative analysis of putatively neutral and outlier SNPs, we provided a strong support to the efficacy of this approach, and presented significant insights to reconsider the European hake stocks management from a new perspective.

Candidate genes

The outlier analysis revealed 12 annotated candidate genes carrying polymorphisms with unusually higher levels of differentiation at various spatial scales. Most of these genes encoded proteins involved in muscles structural organization, activity and protein synthesis regulation or implicated in energy metabolism.

The percentage of SNPs consistently supported as outliers varied between 1.8% and 4.4%, depending on the spatial scale considered, comparable to proportions found with a similar approach in cod (Bradbury *et al.* 2010; Poulsen *et al.* 2011), several salmonid species (Gomez-Uchida *et al.* 2011; Limborg *et al.* 2011; Seeb *et al.* 2011) and threespine stickleback (Deagle *et al.* 2011). Although stringent criteria were adopted to evaluate outlier detection results, we can not exclude the occurrence of false positive among candidate markers under divergent selection: actually, while global outlier SNPs showed significantly high correlations with environmental variables at a wide spatial scale, within Atlantic and Mediterranean the association was marked only for a few SNPs. However, the contribution of additional environmental factors, not considered in this study, as well as multiple interactions, could also give rise

to negative correlations. It has been also demonstrated that outlier methods are prone to type II error, showing limited sensitive to identify markers when directional selection is weak or patterns of adaptive and neutral genetic variation diverge (Narum& Hess 2011). Nevertheless, in our study putative neutral SNPs did not show significant correlation with temperature or salinity parameters in any of the analysis performed.

Indeed non-synonymous SNPs deserve particular attention, due to potential effect on the protein structure and function. Two out of four non-synonymous sites were located on the hemoglobin gene, already designated as candidate gene for environmental selection in Atlantic cod (Frydenberg *et al.* 1965; Sick 1965), another gadoid species. Interestingly, recent studies identified the molecular characteristics and different oxygen binding properties that differentiate haemoglobin alleles in cod (Andersen *et al.* 2009), despite the evidence of a consistent relationship between genotypes and performance under diverse temperature regimes is still lacking (Behrens *et al.* 2012). Three outlier SNPs, one of which non-synonymous, were located on the nexilin gene, encoding an actin filament-binding protein expressed in heart and skeletal muscle cells, recently isolated and studied in zebrafish in relation to human cardiovascular diseases (Hassel *et al.* 2009). Hemoglobin and nexilin non-synonymous mutations were not identified as outliers among Atlantic populations, possibly because of the limited number of samples. In contrast, a polymorphism on the AMP deaminase gene, codifying for an amino acid change in an enzyme that catalyzes the energy production in muscular tissues, was consistently confirmed within basins and supported only from the Bayesian method between Atlantic and Mediterranean.

The effects of gene mutations on muscular functions alterations has been also observed for the titin gene, the candidate gene showing the highest number of outlier synonymous polymorphisms in our study. Titin is a giant cytoskeletal protein, having important structural functions in developing and mature muscles and functional roles in muscle signaling, interacting with a network of several proteins. Titin mutations have been linked to a variety of sarcomeric diseases, including human cardiomyopathy (Hein& Schaper 2002) and muscular dystrophy in zebrafish (Steffen *et al.* 2007). Moreover, it has been

also demonstrated its role in cytoprotection of zebrafish cardiomyocytes, inferred from the interaction with the heat shock protein Hsp27 during thermal stress conditions (Tucker& Shelden 2009).

The indication of divergent selection at synonymous SNPs should not be disregarded, as different selective mechanisms might be involved. One possible explanations is the hitchhiking effect, in which the footprint of selection on a favorite allele extends to closely linked neutral variations. However, considering that linkage disequilibrium is expected to affect a very small region around adaptive polymorphisms (Turner *et al.* 2010) and to last for a short period of time due to recombination, this hypothesis has been challenged with relation to the substantial proportion of outlier loci usually found in genome-scan analysis (Bierne *et al.* 2011). On the other hand, there is increasing evidence that silent mutations may have functional outcome, having effects on translational efficiency and accuracy, by non-random usage of synonymous codons, (Gingold& Pilpel 2011; Plotkin& Kudla 2011), mRNA stability and splicing events (Chamary *et al.* 2006).

Indeed the genome scan approach is a useful indirect strategy to identify genes or genomic regions potentially involved in divergent selection mechanisms in natural population, particularly for non-model organism. However, in order to demonstrate the direct adaptive significance of observed genetic pattern, the functional relation among factors driving evolution, gene regulation and the effects on phenotype and fitness should be provided. Rather than finding evidence of selective effects, this approach is valuable to identify protein-coding genes to target at when investigating adaptive variation in the wild.

In the case of European hake, the hypothesis that selective processes might have contributed to shape the genetic structure has been raised in previous studies based on allozymes analysis: Cimmaruta *et al.* found significant correlations between the genetic pattern at the Glyceraldehyde-3-phos. dehydrogenase and Glucose phosphate isomerase proteins and salinity variation, suggesting a plausible role for selective processes acting on these genes (Cimmaruta *et al.* 2005). As these enzymes are both involved in energy pathways, likewise many of the candidate genes found in our study, we can infer that metabolic functions play a central role hake physiology and ecology, and could be implicated in environmental adaptation strategies. In addiction, Gonzales *et al.* recently found evidence of

differentiation in protein expression profiles of liver and brain cells among two Atlantic and one Mediterranean locations, noticing that protein with signaling and metabolism-energy functions in brain cells showed higher discriminating power (Gonzalez *et al.* 2010).

Neutral population genetic structure

The analysis of 19 geographic samples using 299 putative neutral SNPs revealed a weak but statistically significant genetic structure over the whole distribution range of European hake, with a major break between Atlantic and Mediterranean populations. The higher levels of subdivision between basins partially represent the footprint of historical isolation events, caused by the lowering of sea level during Pleistocene glaciations. The occurrence of an ongoing but limited unidirectional gene flow from Atlantic into Mediterranean was evident from clustering analysis and PCoA, showing a genetic discontinuity in the Alboran Sea. The widespread availability of samples across the Mediterranean Sea allowed to better delineate the genetic pattern over the Atlantic-Mediterranean transition area near the Strait of Gibraltar. Notably, higher rates of introgression were observed along the Algerian coast respect to opposite side of the Alboran Sea, despite the sampling location on the Iberian continental shelf, Malaga, was closer to the entrance of the Gibraltar Strait than the Algeria sample. The annual variability of local processes affecting water circulation in that area and their role in modulating the inflow of Atlantic waters could explain the genetic pattern observed. In summer, the general circulation model in the Alboran Sea is dominated by the presence of two antyciclonic gyres, determining a northward deviation of the Atlantic inflow through Gibraltar. During winter, the two-gyres systems seems to undergo a temporary instability, and a coastal mode circulation establishes, deflecting southward the incoming waters (Vargas-Yáñez *et al.* 2002). As a consequence, a jet flowing close to the African coast is installed. Winter months correspond to the main annual peak spawning observed along the Iberian and Portuguese Atlantic waters (Alvarez *et al.* 2000; Piñeiro& Saínza 2003): it is therefore likely that the gene flow is mediated by passive drift of eggs and larvae across the Gibraltar Strait, as already hypothesized from Roldan *et al.* (Roldan *et al.* 1998).

Lower levels of differentiations using the same set of putative neutral SNPs were detected within basins, despite significant pairwise F_{ST} estimates were also observed. Population substructure analysis within Atlantic confirmed the substantial genetic connectivity between the Northern and Southern stocks, supporting the hypothesis that Cape Breton Canyon does not represent an effective barrier to gene flow, consistently with previous genetic studies (Lundy *et al.* 1999; Pita *et al.* 2011). No significant pairwise differentiations ($p<0.05$) were found among West of Scotland, Southwest of Ireland and Galician coast samples, and between Southwest of Ireland and Portuguese coast samples; on the other hand, the North Sea population appeared to be relatively more divergent, showing higher and significant F_{ST} values when compared to other samples, with the exception of the West of Scotland.

The genetic pattern in the Mediterranean Sea was much more heterogeneous. Despite the low signal of structuring, significant differences were observed among populations. A west-east longitudinal cline in genetic variation, rather than a clear-cut separation among clusters, was evident from the two-dimensional representation of genetic distances. Most of previous works, based on a few Mediterranean samples, found contradictory results with respect to population structure (Castillo *et al.* 2004; Lundy *et al.* 1999; Roldan *et al.* 1998). However, Cimmaruta *et al.* explored genetic variation across the whole Mediterranean Sea, analyzing 11 samples with allozymes: they found a geographical trend from east to west, compatibly with our results, although rejecting the hypothesis of isolation by distance (Cimmaruta *et al.* 2005).

Different evolutionary mechanisms could contribute to explain the pattern observed within Mediterranean: besides historical processes, the possible incidence of factors that limit dispersal or local events of larval retention could affect the contemporary connectivity among populations. The central role of oceanography in shaping the neutral genetic structure of marine fish is widely documented in the Mediterranean Sea (Galarza *et al.* 2009; Schunter *et al.* 2011). However, also biological features, such as reproduction and recruitment strategies, concur to determine the levels of genetic structuring. In Mediterranean the spawning activity of European hake protracts throughout the year, with characteristic peaks spawning varying according to different geographic areas (Arneri& Morales-Nin 2000; Bouaziz *et al.* 1998; Papaconstantinou& Stergiou 1995; Recasens *et al.* 2008); pelagic eggs are

released in multiple batches, and larvae remain in the water column over one month (Arneri& Morales-Nin 2000), before the settlement on the sea bottom. These peculiar biological features, together with the variability of hydrographic processes in Mediterranean, make the link between spatial genetic differentiation and ocean current trends quite complex to explain. Moreover, other mechanisms might be involved, such as the presence of discrete populations due to adult homing behavior, previously suggested by Lundy *et al.* to account for the geographical structuring of European hake populations in the Bay of Biscay (Lundy *et al.* 2000).

Our inferences about neutral genetic structure could be biased by the incidence of loci under balancing selection, showing lower levels of population differentiation than those expected under neutrality. Due to the limited power of outlier tests in discriminating between loci under balancing selection and loci evolving under neutrality (Helyar *et al.* 2011; Narum& Hess 2011), we were not able to consider this issue. However, our inferences about the evidence of structuring within basins would remain valid, as the bias would result in a slightly underestimation of genetic divergences among populations owning to demographic history.

Outlier divergence interpretation

Compared to supposed neutral markers, outlier loci revealed a sharper signal of genetic structuring at different spatial scales, delineating an abrupt divergence between Atlantic and Mediterranean populations, and distinguishing unambiguous genetic clusters within basins. Outlier analysis revealed the presence of clear-cut discontinuities in the Siculo-Tunisian Strait and south of the Peloponnese in Mediterranean, separating Western, Central and Eastern groups, as well as between the Rockall plateau and the North Sea, and along the Iberian peninsula in Atlantic.

The increased differentiation of outlier loci might actually reflect the effects of directional selection, where genetic clusters correspond to locally adapted populations. The hypothesis that European hake populations might be preferentially adapted to local environments is congruent with our findings on correlations between allele frequencies of some outlier loci and temperature and salinity environmental data, found beyond the neutral genetic structure. Furthermore, Cimmaruta *et al.* have also proposed a

possible role of selective mechanisms in shaping the genetic structure of Mediterranean hake populations (Cimmaruta *et al.* 2005). Besides the direct effect on specific loci and those physically linked to them, environment-driven selection, if sufficiently strong, can also promote reproductive isolation, due to a decrease of immigrants and/or hybrids fitness, facilitating differentiation by genetic drift. Following this assumption, divergent local adaptation might represent an additional mechanism limiting neutral gene flow, generating highly variable levels of genetic differentiation across the genome at neutral loci, due to stochastic processes (Nosil *et al.* 2009). However, the heterogeneity observed across the genome in terms of genetic divergence could give rise to various interpretations. Alternative explanations to plain selection have been proposed to account for the presence of a consistent proportion of outlier loci in genome scan approaches, and the evidence of correlation with ecological variables, the so called genetic–environment association (GEA) (Hedrick *et al.* 1976). In a recent review, Bierne *et al.* advocated the role of endogenous barriers, the occurrence of pre- or post-zygotic genetic barriers apart from environment influence (Bierne *et al.* 2011). Endogenous barriers arise due to incompatibilities between groups of alleles, through selective mechanisms independent from adaption to habitats. The accumulation of incompatibilities does not require environmental changes, and can occur during geographical isolation of populations. Intrinsically incompatible genetic backgrounds can afterwards come into contact by tension zones, independent from exogenous factors and thus geographically unstable, moving according to population density and dispersal rate (Barton& Hewitt 1985). After a secondary contact, raised endogenous barriers may produce a negative effect on hybrids fitness, thus preventing the flow of neutral genes, similarly to the effect determined by environmental selection (Bierne *et al.* 2011). Moreover, tension zones may be "trapped" by natural barriers to dispersal (Barton 1979), thus coinciding with environmental boundaries and mimicking exogenous selection.

Establishing the mechanisms underlying the origin and maintenance of European hake global and regional genetic structuring at outlier SNPs is not straightforward, especially because exogenous and endogenous factors may act concomitantly and one process does not rule out the other (Barton 2001). This is particularly evident in the major large-scale barrier, the Strait of Gibraltar: indeed, it corresponds to a secondary contact zone, as during Quaternary glaciations it has been interested by intermittent

closures, partially o totally limiting connections between Atlantic and Mediterranean populations; however, it also represents a transit between marine environments with different ecological conditions. Therefore the occurrence of both endogenous and exogenous barriers to gene flow, favored during periods of isolation, is plausible. The hybrid zone in the Alboran Sea, particularly evident along Algerian coast, could therefore represent an habitat boundary or a tension zone trapped by environmental barriers.

Management implications

The identification of management units representing meaningful biological entities is one of the primary goal of fishery genetics. In this framework, while the application of neutral genetic markers proved highly valuable in elucidating patterns of genetic connectivity among wild populations, they also might be not appropriate or sensitive enough to detect differences between low levels of gene flow and panmixia, distinguish recently diverged populations, or provide information about adaptive genetic variation, overlooking possible cryptic stocks (Conover *et al.* 2006). European hake stock management have been previously challenged, but available genetic tools have limited the ability to discriminate population structure. Highly informative SNP markers applied in our study, identified on candidate genes potentially under selection, efficiently disentangled population structuring of European hake over its distribution range. On a wide spatial scale, we confirmed the well known separation between Atlantic and Mediterranean populations, finding evidence of an hybrid zone in the Alboran Sea, especially in the southern area. At regional scale our findings unveiled clear and unexpected genetic differentiation patterns, contrasting with the current management system. Assuming that environmental selection is the major driving mechanism underlying the genetic divergence found at outlier loci, results presented in our paper has important implications for conservation purposes. Actually, adaptive diversity denotes the evolutionary potential of populations, affecting their ability to persist in nature and to undergo adaptation in response to environmental changes, stress and diseases (Hoffmann& Willi 2008; Lo Brutto *et al.* 2011; Reusch& Wood 2007), including human-induced environmental alterations and fishing pressure. The idea to incorporate information on ecological traits in identifying conservation units has

been promoted to preserve evolutionary resilience in the wild (Crandall *et al.* 2000; Fraser& Bernatchez 2001; Gebremedhin *et al.* 2009), and the increasing availability of genomic resources, facilitating the search and analysis of adaptive genes, will favor this approach.

Nevertheless, even in case outlier loci signal have raised from genetic incompatibilities due to endogenous factors, the evident sub-structuring within Atlantic and Mediterranean would indicate the occurrence of semipermeable reproductive isolation barriers, and should be considered in a management context.

Following a precautionary approach, our results suggest the need to revise the current management strategies of European hake stocks, particularly within basins. According to our findings the Northern and Southern Atlantic stocks, assessed separately, do not constitute demographically isolated or ecologically differentiated populations, while a genetic discontinuity has been found in the North Sea and along the Portuguese coast. Moreover, the Mediterranean stock, currently managed as a single unit, revealed a more complex structure with a clear subdivision among three main areas, Western, Central and Eastern.

Our results are not definitive, rather open a new prospect on European hake populations dynamics and pose important issues to consider in future studies focusing on its conservation biology. The genome-scan approach we applied in this study allowed to identify candidate genes potentially involved in adaptive processes, showing strong genetic structure at fine spatial scale. In the frame of fishery genetics, the opportunity to explore adaptive variation in the wild represent a step towards a deeper understanding of important biological traits, and the possibility to integrate evolutionary ecology dynamics in conservation and management plans.

References

Ackerman MW, Habicht C, Seeb LW (2011) Single-Nucleotide Polymorphisms (SNPs) under Diversifying Selection Provide Increased Accuracy and Precision in Mixed-Stock Analyses of Sockeye Salmon from the Copper River, Alaska. *Transactions of the American Fisheries Society* **140**, 865-881.

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nature reviews. Genetics* **11**, 697-709.

Alvarez P, Fives J, Motos L, Santos M (2004) Distribution and abundance of European hake Merluccius merluccius (L.), eggs and larvae in the North East Atlantic waters in 1995 and 1998 in relation to hydrographic conditions. *Journal of Plankton Research* **26**, 811-826.

Alvarez P, Motos L, Uriarte A, Egaña J (2000) Spatial and temporal distribution of European hake, Merluccius merluccius (L.), eggs and larvae in relation to hydrographical conditions in the Bay of Biscay. *Fisheries Research* **50**, 111-128.

Andersen O, De Rosa MC, Pirolli D*, et al.* (2011) Polymorphism, selection and tandem duplication of transferrin genes in Atlantic cod (Gadus morhua) - Conserved synteny between fish monolobal and tetrapod bilobal transferrin loci. *BMC Genetics* **12**, 51.

Andersen Ø, Wetten OF, De Rosa MC*, et al.* (2009) Haemoglobin polymorphisms affect the oxygen-binding properties in Atlantic cod populations. *Proceedings of the Royal Society B: Biological Sciences* **276**, 833-841.

Antonov JI, Seidov D, Boyer TP*, et al.* (2010) *World Ocean Atlas 2009, Volume 2: Salinity.* U.S. Government Printing Office, Washington, D.C.

Arneri E, Morales-Nin B (2000) Aspects of the early life history of European hake from the central Adriatic. *Journal of Fish Biology* **56**, 1368-1380.

Barton NH (1979) The dynamics of hybrid zones. *Heredity* **43**, 341-359.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology* **10**, 551-568.

Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics* **16**, 113-148.

Beaumont MA, Nichols RA (1996) Evaluating Loci for Use in the Genetic Analysis of Population Structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **263**, 1619-1626.

Behrens JW, Gräns A, Therkildsen NO, Neuenfeldt S, Axelsson M (2012) Correlations between hemoglobin type and temperature preference of juvenile Atlantic cod Gadus morhua. *Journal of Experimental Marine Biology and Ecology* **413**, 71-77.

Belcari P, Ligas A, Viva C (2006) Age determination and growth of juveniles of the European hake, Merluccius merluccius (L., 1758), in the northern Tyrrhenian Sea (NW Mediterranean). *Fisheries Research* **78**, 211-217.

Biagi F, Cesarini A, Sbrana M, Viva C (1995) Reproductive biology and fecundity of Merluccius merluccius (Linnaeus, 1758) in the Nothern Tyrrhenian Sea. In: *Rapp. Comm. int. Mer Médit.* (ed. CIHEAM), pp. 34-23.

Bierne N, Welch J, Loire E, Bonhomme F, David P (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology* **20**, 2044-2072.

Blel H, Panfili J, Guinand B*, et al.* (2010) Selection footprint at the first intron of the Prl gene in natural populations of the flathead mullet (Mugil cephalus, L. 1758). *Journal of Experimental Marine Biology and Ecology* **387**, 60-67.

Bonin A (2008) Population genomics: a new generation of genome scans to bridge the gap with functional genomics. *Molecular Ecology* **17**, 3583-3584.

Bouaziz A, Bennoui A, Djabali F, Maurin C (1998) Reproduction du merlu *Merluccius merluccius* (linnaeus, 1758) dans la région de Bou-Ismaîl. . In: *Options Mediterraneennes* (ed. CIHEAM), pp. 109-117.

Boutet I, Quéré N, Lecomte F, Agnèse J-F, Guinand B (2008) Putative transcription factor binding sites and polymorphisms in the proximal promoter of the PRL-A gene in percomorphs and European sea bass (Dicentrarchus labrax). *Marine Ecology* **29**, 354-364.

Bradbury IR, Hubert S, Higgins B*, et al.* (2010) Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B: Biological Sciences* **277**, 3725-3734.

Castillo AGF, Alvarez P, Garcia-Vazquez E (2005) Population structure of Merluccius merluccius along the Iberian Peninsula coast. *ICES Journal of Marine Science: Journal du Conseil* **62**, 1699-1704.

Castillo AGF, Martinez JL, Garcia-Vazquez E (2004) Fine Spatial Structure of Atlantic Hake (Merluccius merluccius) Stocks Revealed by Variation at Microsatellite Loci. *Marine Biotechnology* **6**, 299-306.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**, 98-108.

Cimmaruta R, Bondanelli P, Nascetti G (2005) Genetic structure and environmental heterogeneity in the European hake (Merluccius merluccius). *Molecular Ecology* **14**, 2577-2591.

Conover DO, Clarke LM, Munch SB, Wagner GN (2006) Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. *Journal of Fish Biology* **69**, 21-47.

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using Environmental Correlations to Identify Loci Underlying Local Adaptation. *Genetics* **185**, 1411-1423.

Cowen RK, Sponaugle S (2009) Larval Dispersal and Marine Population Connectivity. *Annual Review of Marine Science* **1**, 443-466.

Crandall KA, Bininda-Emonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology. *Trends in Ecology & Evolution* **15**, 290-295.

Deagle BE, Jones FC, Chan YF*, et al.* (2011) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proceedings of the Royal Society B: Biological Sciences*.

Domínguez-Petit R, Korta M, Saborido-Rey F*, et al.* (2008) Changes in size at maturity of European hake Atlantic populations in relation with stock structure and environmental regimes. *Journal of Marine Systems* **71**, 260-278.

Edwards AWF (1971) Distances between Populations on the Basis of Gene Frequencies. *Biometrics* **27**, 873-881.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**, 2611-2620.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity* **103**, 285-298.

Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564-567.

Fan JB, Oliphant A, Shen R*, et al.* (2003) Highly Parallel SNP Genotyping. *Cold Spring Harbor symposia on quantitative biology* **68**, 69-78.

Foll M, Gaggiotti O (2008) A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. *Genetics* **180**, 977-993.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology* **10**, 2741-2752.

Freamo H, O'Reilly P, Berg PR, Lien S, Boulding EG (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources* **11**, 254-267.

Frydenberg OVE, MØLler DAG, NÆVdal G, Sick K (1965) Haemoglobin polymorphism in Norwegian cod populations. *Hereditas* **53**, 257-271.

Gaggiotti OE, Bekkevold D, Jørgensen HBH*, et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution* **63**, 2939-2951.

Galarza JA, Carreras-Carbonell J, Macpherson E*, et al.* (2009) The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. *Proceedings of the National Academy of Sciences* **106**, 1473-1478.

Gebremedhin B, Ficetola GF, Naderi S*, et al.* (2009) Frontiers in identifying conservation units: from neutral markers to adaptive genetic variation. *Animal Conservation* **12**, 107-109.

Gingold H, Pilpel Y (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**.

Gomez-Uchida D, Seeb J, Smith M*, et al.* (2011) Single nucleotide polymorphisms unravel hierarchical divergence and signatures of selection among Alaskan sockeye salmon (<i>Oncorhynchus nerka</i>) populations. *BMC Evolutionary Biology* **11**, 1-17.

Gonzalez EG, Krey G, Espiñeira M*, et al.* (2010) Population Proteomics of the European Hake (Merluccius merluccius). *Journal of Proteome Research* **9**, 6392-6404.

Goudet J (1995) FSTAT (Version 1.2): A Computer Program to Calculate F-Statistics. *Journal of Heredity* **86**, 485-486.

Hassel D, Dahme T, Erdmann J*, et al.* (2009) Nexilin mutations destabilize cardiac Z-disks and lead to dilated cardiomyopathy. *Nat Med* **15**, 1281-1288.

Hauser L, Carvalho GR (2008) Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries* **9**, 333-362.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic Polymorphism in Heterogeneous Environments. *Annual Review of Ecology and Systematics* **7**, 1-32.

Hein S, Schaper J (2002) Weakness of a giant: mutations of the sarcomeric protein titin. *Trends in Molecular Medicine* **8**, 311-313.

Helyar SJ, Hemmer-Hansen J, Bekkevold D*, et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 123-136.

Hemmer-Hansen J, Nielsen EE, Frydenberg J, Loeschcke V (2007) Adaptive divergence in a high gene flow environment: Hsc70 variation in the European flounder (Platichthys flesus L.). *Heredity* **99**, 592-600.

Hoffmann AA, Willi Y (2008) Detecting genetic responses to environmental change. *Nat Rev Genet* **9**, 421-432.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**, 1322-1332.

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**, 1801-1806.

Jeffreys H (1961) *Theory of probability* Clarendon Press, Oxford.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-1405.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* **11**, 94.

Kacher M, Amara R (2005) Distribution and growth of 0-group European hake in the Bay of Biscay and Celtic Sea: a spatial and inter-annual analyses. *Fisheries Research* **71**, 373-378.

Larmuseau MHD, Raeymaekers JAM, Ruddick KG, Van Houdt JKJ, Volckaert FAM (2009) To see in different seas: spatial variation in the rhodopsin gene of the sand goby (Pomatoschistus minutus). *Molecular Ecology* **18**, 4227-4239.

Larmuseau MHD, Vancampenhout KIM, Raeymaekers JAM, Van Houdt JKJ, Volckaert FAM (2010) Differential modes of selection on the rhodopsin gene in coastal Baltic and North Sea populations of the sand goby, Pomatoschistus minutus. *Molecular Ecology* **19**, 2256-2268.

Limborg MT, Blankenship SM, Young SF*, et al.* (2011) Signatures of natural selection among lineages and habitats in Oncorhynchus mykiss. *Ecology and Evolution* **2**, 1-18.

Lo Brutto S, Arculeo M, Stewart Grant W (2011) Climate change and population genetic structure of marine species. *Chemistry and Ecology* **27**, 107-119.

Locarnini RA, Mishonov AV, Antonov JI*, et al.* (2010) *World Ocean Atlas 2009, Volume 1: Temperature.* U.S. Government Printing Office, Washington.

Lundy CJ, Moran P, Rico C, Milner RS, Hewitt GM (1999) Macrogeographical population differentiation in oceanic environments: a case study of European hake (Merluccius merluccius), a commercially important fish. *Molecular Ecology* **8**, 1889-1898.

Lundy CJ, Rico C, Hewitt GM (2000) Temporal and spatial genetic variation in spawning grounds of European hake (Merluccius merluccius) in the Bay of Biscay. *Molecular Ecology* **9**, 2067-2079.

Mellon-Duval C, de Pontual H, Métral L, Quemener L (2009) Growth of European hake (Merluccius merluccius) in the Gulf of Lions based on conventional tagging. *ICES Journal of Marine Science: Journal du Conseil* **67**, 62-70.

Milano I, Babbucci M, Panitz F*, et al.* (2011) Novel Tools for Conservation Genomics: Comparing Two High-Throughput Approaches for SNP Discovery in the Transcriptome of the European Hake. *PLoS ONE* **6**, e28008.

Moen T, Hayes B, Nilsen F*, et al.* (2008) Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genetics* **9**, 18.

Naish KA, Hard JJ (2008) Bridging the gap between the genotype and the phenotype: linking genetic variation, selection and adaptation in fishes. *Fish and Fisheries* **9**, 396-422.

Narum SR, Hess JE (2011) Comparison of FST outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**, 184-194.

Nielsen E, Hemmer-Hansen J, Poulsen N*, et al.* (2009a) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). *BMC Evolutionary Biology* **9**, 276.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009b) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology* **18**, 3128-3150.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology* **18**, 375-402.

Papaconstantinou C, Stergiou KI (1995) Biology and fisheries of eastern Mediterranean hake (*M. merluccius*). In: *Hake: Biology, Fisheries and Markets* (ed. eds AJPTJ), pp. 149-180. London: Chapman & Hall.

Peakall ROD, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**, 288-295.

Piñeiro C, Rey J, de Pontual H, Garcìa A (2008) Growth of Northwest Iberian juvenile hake estimated by combining sagittal and transversal otolith microstructure analyses. *Fisheries Research* **93**, 173-178.

Piñeiro C, Saínza M (2003) Age estimation, growth and maturity of the European hake (Merluccius merluccius (Linnaeus, 1758)) from Iberian Atlantic waters. *ICES Journal of Marine Science: Journal du Conseil* **60**, 1086-1102.

Pita A, Pérez M, Cerviño S, Presa P (2011) What can gene flow and recruitment dynamics tell us about connectivity between European hake stocks in the Eastern North Atlantic? *Continental Shelf Research* **31**, 376-387.

Pita A, Presa P, Perez M (2010) Gene flow, multilocus assignment and genetic structuring of the European hake (Merluccius merluccius). *Thalassas* **26**, 129-133.

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**, 32-42.

Poulsen N, Hemmer-Hansen J, Loeschcke V, Carvalho G, Nielsen E (2011) Microgeographical population structure and adaptation in Atlantic cod Gadus morhua: spatio-temporal insights from gene-associated DNA markers. *Marine Ecology Progress Series* **436**, 231-243.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**, 945-959.

Raymond M, Rousset F (1995) GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism. *Journal of Heredity* **86**, 248-249.

Recasens L, Chiericoni V, Belcari P (2008) Spawning pattern and batch fecundity of the European hake (Merluccius merluccius (Linnaeus, 1758) in the western Mediterranean. *Scientia Marina* **72**, 721-732.

Recasens L, Lombarte A, Morales-Nin B, Tores GJ (1998) Spatiotemporal variation in the population structure of the European hake in the NW Mediterranean. *Journal of Fish Biology* **53**, 387-401.

Reusch TBH, Wood TE (2007) Molecular ecology of global change. *Molecular Ecology* **16**, 3973-3992.

Rice WR (1989) Analyzing tables of statistical tests. *Evolution* **43**, 223-225.

Roldan MI, Garcia-Marin J, Utter FM, Pla C (1998) Population genetic structure of European hake, Merluccius merluccius. *Heredity* **81**, 327-334.

Rosenberg NA (2004) distruct: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**, 137-138.

Russello MA, Kirk SL, Frazer KK, Askey PJ (2011) Detection of outlier loci and their utility for fisheries management. *Evolutionary Applications*, no-no.

Ruzzante DE, Mariani S, Bekkevold D*, et al.* (2006) Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *Proceedings of the Royal Society B: Biological Sciences* **273**, 1459-1464.

Schunter C, Carreras-Carbonell J, Macpherson E*, et al.* (2011) Matching genetics with oceanography: directional gene flow in a Mediterranean fish species. *Molecular Ecology* **20**, 5167-5181.

Seeb LW, Templin WD, Sato S*, et al.* (2011) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. *Molecular Ecology Resources* **11**, 195-217.

Sick K (1965) Haemoglobin polymorphism of cod in the Baltic and the Danish Belt sea. *Hereditas* **54**, 19-48.

Skjæraasen JE, Meager JJ, Karlsen Ø, Hutchings JA, Fernö A (2011) Extreme spawning-site fidelity in Atlantic cod. *ICES Journal of Marine Science: Journal du Conseil*.

Stapley J, Reger J, Feulner PGD*, et al.* (2010) Adaptation genomics: the next generation. *Trends in ecology & evolution (Personal edition)* **25**, 705-712.

Steffen LS, Guyon JR, Vogel ED*, et al.* (2007) The zebrafish runzel muscular dystrophy is linked to the titin gene. *Developmental Biology* **309**, 180-192.

Storz JF (2005) INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology* **14**, 671-688.

Swearer SE, Shima JS, Hellberg ME*, et al.* (2002) Evidence of self-recruitment in demersal marine populations. *Bulletin of Marine Science* **70**, 251-271.

Tucker NR, Shelden EA (2009) Hsp27 associates with the titin filament system in heat-shocked zebrafish cardiomyocytes. *Experimental Cell Research* **315**, 3176-3186.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nat Genet* **42**, 260-263.

Vargas-Yáñez M, Plaza F, García-Lafuente J*, et al.* (2002) About the seasonal variability of the Alboran Sea circulation. *Journal of Marine Systems* **35**, 229-248.

Vasemägi A, Primmer CR (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Molecular Ecology* **14**, 3623-3642.

Waples R (1998) Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *Journal of Heredity* **89**, 438-450.

Zhu MJ, Zhao SH (2007) Candidate gene identification approach: Progress and challenges. *International Journal of Biological Sciences* **3**, 420-427.

Table 1. Sampling information and summary statistics.

Fishery Management Area	Population ID	Sampling Location	Latitude	Longitude	Year	n	H_e	H_o	A_r	% Polymorphic SNPs
Northern Atlantic Stock	NTS	North Sea	57,996	0,219	August 2009	43	0,288	0,307	1,8533	93,18
	WSC	West of Scotland	56,208	-9,212	September 2009	53	0,284	0,285	1,8544	95,28
	SIR	Southwest of Ireland	50,426	-9,975	February 2009	58	0,282	0,282	1,856	96,85
Southern Atlantic Stock	GAC	Galician Coast	43,305	-9,180	March 2009	43	0,286	0,292	1,8552	95,01
	NPT	North Portugal	42,123	-9,394	March 2009	46	0,287	0,290	1,8649	95,01
Mediterranean Stock	ALG	West Algeria	35,605	-2,123	October 2009	41	0,296	0,300	1,8958	97,64
	MAL	Malaga	36,634	-4,369	November 2009	47	0,282	0,280	1,8754	96,85
	GFL	Gulf of Lion	43,350	3,726	May 2009	46	0,276	0,273	1,871	97,11
	NTY	North Tyrrenian	42,515	10,109	May 2009	36	0,279	0,288	1,8637	94,49
	SSD	South Sardinia	38,997	8,510	May 2009	44	0,277	0,287	1,873	97,38
	STY	South Tyrrenian	40,536	14,770	May 2009	48	0,267	0,267	1,8504	95,28
	ADB	Adventure Bank	37,333	12,172	May 2008	66	0,283	0,288	1,8709	97,64
	MAB	Malteste Bank	35,916	14,786	June 2008	46	0,274	0,277	1,8506	95,01
	NAD	North Adriatic	43,928	13,793	May-July 2009	45	0,273	0,279	1,852	95,80
	SAD	South Adriatic	42,260	16,894	July 2009	19	0,266	0,266	1,8421	87,93
	NWJ	Northwestern Jonian Sea	38,193	16,366	November-December 2008	33	0,278	0,278	1,8667	94,23
	AEG	Aegean Sea	40,344	24,515	January 2009	54	0,267	0,274	1,8358	95,28
	TKY	Turkish Coast	36,776	30,945	July 2009	39	0,266	0,270	1,8279	91,34
	CYP	Cyprus	34,520	32,913	January-July 2009	43	0,268	0,266	1,8404	94,75

Table 2. Environmental data related to each population. Annual climatological mean values of temperature and salinity at 0m and 100m depth were retrieved from the NODC database (source data World Ocean Atlas 2009), considering reference geographic coordinates close to our sampling sites. Temperature and salinity climatologies provided from the NODC are based on long-term observations (1955-2006).

		Reference Coordinates NODC		Annual Climatological mean data			
				0 m		100 m	
	Population	Latitude	Longitude	Temperature	Salinity	Temperature	Salinity
Atlantic	NTS	58,125	0,125	9,594	35,057	7,369	35,163
	WSC	56,125	-9,125	11,24	35,274	9,956	35,376
	SIR	50,875	-9,625	12,346	35,19	10,032	35,356
	GAC	43,375	-9,125	15,073	35,564	12,808	35,683
	NPT	42,125	-9,375	15,487	35,648	13,279	35,814
Mediterranean	ALG	35,625	-2,125	18,758	36,681	14,455	37,508
	MAL	36,625	-4,125	17,944	36,656	14,134	37,589
	GFL	43,125	4,625	16,471	37,596	13,275	38,188
	NTY	42,625	10,375	18,404	37,93	13,882	38,114
	SSD	38,875	8,625	18,584	37,745	13,891	38,123
	STY	40,625	14,375	19,157	37,7	14,19	38,12
	ADB	37,375	11,875	19,106	37,41	14,322	37,823
	MAB	35,875	14,875	19,945	37,807	15,404	38,246
	NAD	43,125	14,375	17,559	37,694	13,381	38,593
	SAD	42,375	16,875	18,2	38,154	13,95	38,595
	NWJ	38,125	16,375	19,402	38,16	14,54	38,488
	AEG	40,375	25,375	17,643	36,815	15,192	38,839
	TKY	36,625	30,625	21,244	39,097	16,709	39,033
	CYP	34,625	33,125	21,614	39,077	16,743	38,939

Table 3. List of loci included in outlier data sets and used for population structure analysis. All SNPs were detected as outliers under stringent criteria (log10BF>=2 and p<=0.01, see Figure 1) by at least one of the two genome-scan approaches. On a wide spatial-scale, SNPs supported by both methods were selected, to reduce the proportion of false positive. For each locus information about annotation, amino-acid change and levels of genetic differentiation are specified (F_{CT} values between Atlantic and Mediterranean groups of samples for outliers detected on the whole dataset and F_{ST} values for outliers identified within basin). Relative estimates of correlation between allele frequencies and annual climatological mean values of temperature and salinity at two different depth level, obtained using a Bayesian method that takes the effect of underlying population structure into account, are also reported. Values of $log_{10}BF$ between 1.5 and 2 (in yellow) indicate a "very strong" association, while $log_{10}BF>2$ (in red) designate a "decisive" evidence for correlation.

						Correlation Environmental Variables ($log_{10}BF$)			
						Temperature		**Salinity**	
Samples used	**SNP name**	**Gene**	**SYN-NON SYN**	**Amino acid change**	**F statistics**	**0 m**	**100 m**	**0 m**	**100 m**
All samples	1805_fpt	nexilin (F actin binding protein)	SYN	-	0,275	2,31	1,73	3,01	3,90
	852_fpt	titin	SYN	-	0,380	3,02	2,32	3,41	4,01
	1398_fpt	adipophylin	SYN	-	0,322	3,09	2,14	3,34	3,67
	3474_fpt	nexilin (F actin binding protein)	NON SYN	R/L	0,375	2,09	1,39	2,88	3,30
	3520_fpt	synaptopodin-2 like	-	-	0,650	5,55	4,04	6,68	7,80
	778ms	-	-	-	0,475	2,82	1,67	4,05	4,65
	1580ms	titin	SYN	-	0,340	2,39	1,90	2,86	3,48
	2186_fpt	hemoglobin subunit beta-1	NON SYN	L/M	0,705	8,20	8,01	6,03	8,39
	2184_fpt	nexilin (F actin binding protein)	SYN	-	0,336	3,37	2,32	4,13	4,24
	4539ms	-	-	-	0,317	2,16	1,49	2,88	2,83
	151_fpt	titin	SYN	-	0,504	5,94	5,34	5,99	7,38
	1485_fpt	immunoglobulin-like and fibronectin type iii domain-containing protein	SYN	-	0,297	2,75	2,38	3,11	3,75
	3656_fpt	hemoglobin subunit beta	NON SYN	S/T	0,291	3,92	4,08	4,13	4,27
	1178ms	-	-	-	0,509	4,24	3,33	5,07	5,96
	239_fpt	-	-	-	0,304	2,16	1,23	2,55	2,20
	3039_fpt	titin	SYN	-	0,372	2,91	2,29	3,31	4,04

	2579_fpt	aminoacyl trna synthetase complex-interacting multifunctional protein 1	SYN	-	0,338	3,42	2,88	4,34	4,88
Atlantic	1522ms	-	-	-	0,126	2,52	2,80	2,49	2,49
	927_fpt	titin	SYN	-	0,173	1,14	1,55	0,90	1,05
	109ms	pdz and lim domain protein 7	SYN	-	0,254	2,67	3,04	2,22	2,44
	4472ms	amp deaminase 1	NON SYN	T/P	0,298	-0,34	-0,33	0,04	-0,04
	778ms	-	-	-	0,066	0,36	0,56	0,26	0,26
	890_fpt	adp/atp translocase	SYN	-	0,055	1,275	1,407	1,245	1,143
	239_fpt	-	-	-	0,067	-0,31	-0,30	-0,29	-0,29
Mediterranean	916ms	guanine nucleotide-binding protein subunit beta-2-like	-	-	0,045	-0,59	-0,31	-0,42	0,96
	1034ms	-	-	-	0,056	-0,63	-0,48	-0,61	0,10
	3520_fpt	synaptopodin-2	-	-	0,071	-0,54	-0,56	0,52	1,69
	778ms	-	-	-	0,159	-0,42	-0,39	0,51	1,36
	4472ms	amp deaminase 1	NON SYN	T/P	0,089	-0,65	-0,58	-0,26	-0,39
	2186_fpt	hemoglobin subunit beta-1	NON SYN	L/M	0,214	-0,39	1,02	-0,47	-0,51
	4539ms	-	-	-	0,071	-0,10	-0,42	0,78	0,50
	1330ms	-	-	-	0,064	0,37	0,56	-0,27	-0,38
	151_fpt	titin	SYN	-	0,051	-0,15	0,39	0,18	1,95
	3656_fpt	hemoglobin subunit beta	NON SYN	S/T	0,120	1,09	3,05	0,81	0,90
	1178ms	-	-	-	0,101	-0,54	-0,54	0,32	1,18
	3374_fpt	-	-	-	0,057	0,61	2,02	0,10	1,08
	239_fpt	-	-	-	0,050	-0,02	-0,67	0,67	-0,55
	930_fpt	mid1-interacting protein 1-like	SYN	-	0,038	0,14	-0,41	-0,15	-0,43
	921ms	guanine nucleotide-binding protein subunit beta-2-like	SYN	-	0,089	0,61	1,48	1,15	1,99
	1148ms	guanine nucleotide-binding protein subunit beta-2-like	SYN	-	0,050	0,01	0,64	-0,12	0,36
	3474_fpt	nexilin (F actin binding protein)	NON SYN	R/L	0,079	-0,47	-0,47	0,25	0,44
	179ms	myomesin-1	SYN	-	0,064	-0,69	-0,69	-0,55	-0,54
	1485_fpt	immunoglobulin-like and fibronectin type iii domain-containing protein	SYN	-	0,037	-0,50	-0,45	0,03	0,96

Table 4. Matrix of pairwise F_{ST} values based on 299 neutral SNPs.

* p<0.05; values significant after sequential Bonferroni correction (p<0.0005) in bold.

	Atlantic					Mediterranean												
	NTS	**WSC**	**SIR**	**GAC**	**NPT**	**ALG**	**MAL**	**GFL**	**NTY**	**SSD**	**STY**	**ADB**	**MAB**	**NAD**	**SAD**	**NWJ**	**AEG**	**TKY**	**CYP**
NTS	-																		
WSC	0,0012	-																	
SIR	0,0029*	-0,0003	-																
GAC	0,0026*	-0,0003	-0,0004	-															
NPT	**0,0058***	0,0024*	0,0012	-0,0005	-														
ALG	**0,0101***	**0,0074***	**0,0069***	**0,0047***	**0,0058***	-													
MAL	**0,0134***	**0,0099***	**0,0099***	**0,0094***	**0,0098***	0,0020*	-												
GFL	**0,0177***	**0,0133***	**0,0137***	**0,0118***	**0,0141***	0,0019	0,0005	-											
NTY	**0,0229***	**0,0218***	**0,0210***	**0,0192***	**0,0184***	0,0041*	0,0040*	0,0015	-										
SSD	**0,0195***	**0,0143***	**0,0146***	**0,0131***	**0,0141***	0,0026*	0,0014	0,0009	0,0020*	-									
STY	**0,0224***	**0,0212***	**0,0211***	**0,0182***	**0,0202***	**0,0059***	0,0032*	0,0038*	0,0034*	0,0027*	-								
ADB	**0,0196***	**0,0172***	**0,0171***	**0,0152***	**0,0171***	0,0030*	0,0007	0,0017*	0,0022*	0,0004	0,0021*	-							
MAB	**0,0178***	**0,0160***	**0,0142***	**0,0145***	**0,0145***	0,0039*	0,0003	0,0006	-0,0002	0,0013	0,0013	0,0014*	-						
NAD	**0,0260***	**0,0221***	**0,0212***	**0,0199***	**0,0196***	**0,0048***	0,0026*	0,0023*	0,0018*	0,0006	0,0003	0,0026*	0,0012	-					
SAD	**0,0201***	**0,0165***	**0,0164***	**0,0143***	**0,0161***	0,0017	-0,0020	-0,0026	-0,0017	-0,0029	-0,0037	-0,0018	-0,0043	-0,0065	-				
NWJ	**0,0201***	**0,0187***	**0,0170***	**0,0152***	**0,0168***	0,0029*	-0,0009	0,0012	0,0015	-0,0006	0,0019	0,0014	0,0001	-0,0006	-0,0070	-			
AEG	**0,0243***	**0,0228***	**0,0212***	**0,0213***	**0,0216***	**0,0081***	0,0035*	0,0029*	0,0030*	0,0029*	0,0022*	0,0025*	0,0022*	0,0017*	-0,0026	0,0016	-		
TKY	**0,0285***	**0,0266***	**0,0250***	**0,0238***	**0,0226***	**0,0095***	**0,0062***	0,0044*	**0,0059***	0,0028*	**0,0046***	**0,0038***	**0,0060***	**0,0047***	-0,0004	**0,0055***	0,0020*	-	
CYP	**0,0275***	**0,0251***	**0,0250***	**0,0235***	**0,0257***	**0,0087***	0,0038*	**0,0062***	**0,0051***	0,0026*	0,0017*	0,0031*	**0,0059***	0,0018	-0,0019	0,0020	0,0013	0,0015	-

Table 5. Hierarchical analysis of molecular variance (AMOVA) respect to 299 putative neutral and 17 outlier SNPs. Samples were grouped according to the basin of origin (Atlantic: NTS, WSC, SIR, GAC, NPT; Mediterranean: ALG, MAL, GFL, NTY, SSD, STY, ADB, MAB, NAD, SAD, NWJ, AEG, TKY, CYP).

** P<0.0002

Loci used	Source of variation	d.f.	Variance components	Percentage of variation	Fixation index
299 Neutral	Among groups	1	0.63513	1.56**	F_{CT} = 0.01559
	Among populations within group	17	0.08005	0.20**	F_{SC} = 0.00200
	Within populations	1681	40.02241	98.24**	F_{ST} = 0.01756
	Total	1699	40.73759		
17 Outlier	Among groups	1	1.93472	41.88**	F_{CT} = 0.41883
	Among populations within group	17	0.12502	2.71**	F_{SC} = 0.04657
	Within populations	1681	2.55955	55.41**	F_{ST} = 0.41883
	Total	1699	4.61929		

Figure legends

Figure 1. Geographical representation of sampling locations, listed in Table 1. Dotted lines indicate approximate boundaries among Mediterranean, Northern and Southern Atlantic stocks.

Figure 2. Graphical representation of outlier tests results. Analysis were performed according to two different approaches and at different spatial scales, considering all geographic samples (A, B), and samples within Atlantic (C, D) and Mediterranean (E, F) basins separately. A, C, E: results of outlier test following the method implemented in Bayescan; locus-specific F_{ST} coefficient is plotted against $\log_{10}$PO for the model including selection; values of $\log_{10}$PO>2 indicate a decisive signal for selection. An arbitrary value of 5 was assigned to $\log_{10}$PO when the posterior probability was 1 (corresponding to a PO of infinity). B, D, F: results of Arlequin outlier tests; single locus F_{ST} values are plotted against heterozigosity, with the dotted line representing the 99% confidence threshold estimated under a neutral simulated model; at a broad spatial scale, results were achieved assuming a hierarchical structure between basins, and F_{CT} distribution was considered for outlier detection (B). Filled circles represent SNPs detected as outliers by both approaches under stringent criteria.

Figure 3. Results of correlation tests between allele frequencies and temperature (A, C, E) and salinity (B, D, F) environmental variables. Values of $\log_{10}$BF respect to temperature and salinity at 0m (empty symbols) and 100m (filled symbols) depth are plotted on the y axis for each SNP. Triangles represent 299 neutral loci, while circles indicate outlier loci found at wide spatial scale (A and B), within Atlantic (B and C) and within Mediterranean (E and F), as listed in Table 3.

Figure 4. Clustering analysis results based on: A. 299 neutral SNPs (K=2); B. 7 outlier SNPs within Atlantic (K=3); C. 19 outlier SNPs within Mediterranean (K=3).

Figure 5. Results of DAPC based on outlier loci within Atlantic (A) and Mediterranean (B). The scatterplots (on the left) illustrate differentiation among identified genetic clusters, while barplots (on the right) represent relative proportions of membership to each cluster for each population. Cluster colors correspond to the ones used for Structure results (Figures 4 B and C).

Figure 6. PCoA base on Edwards' genetic distance indexes.

Figure 1

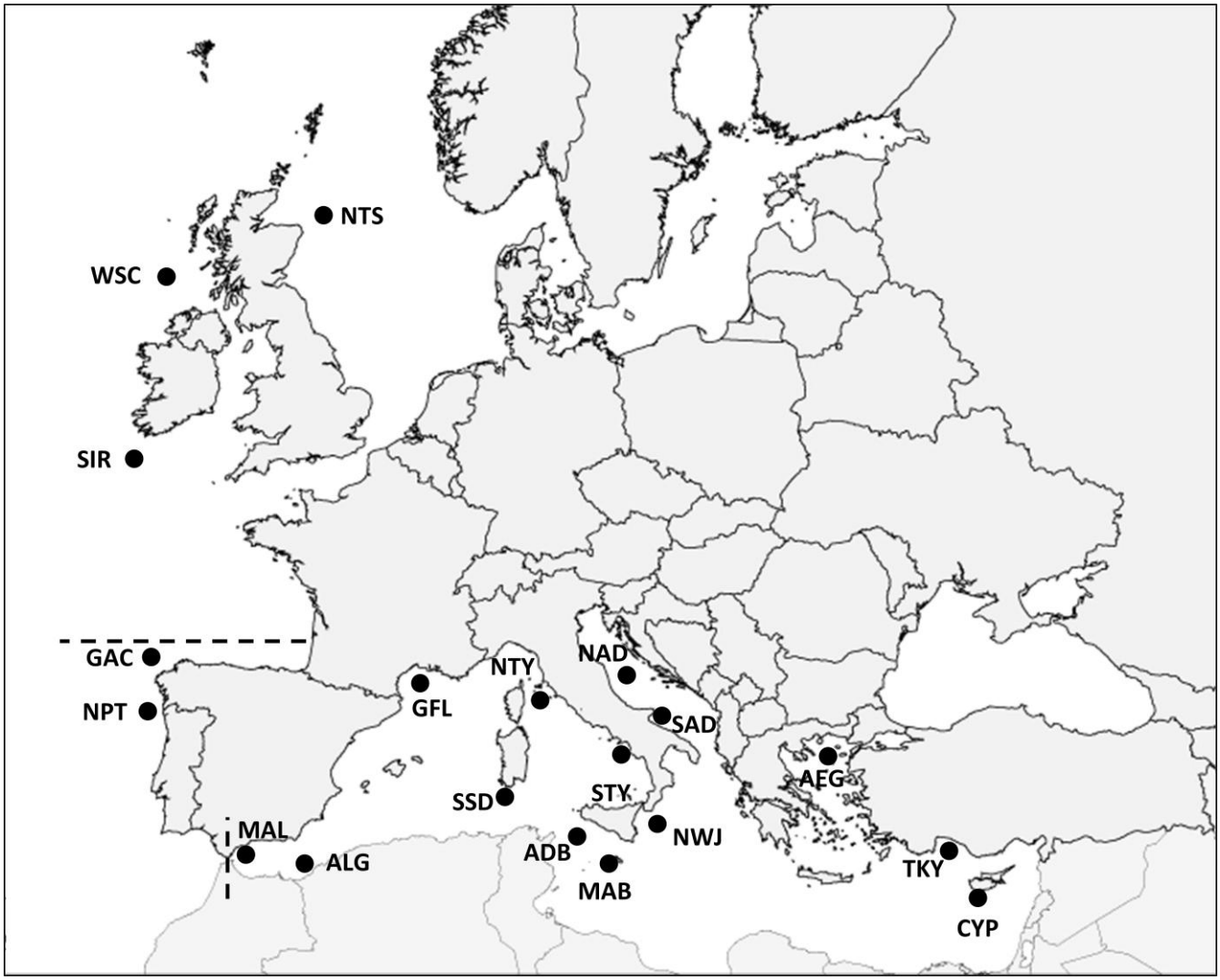
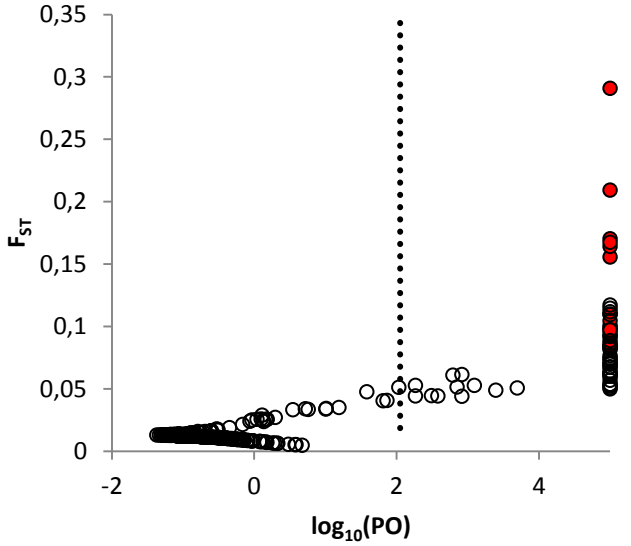
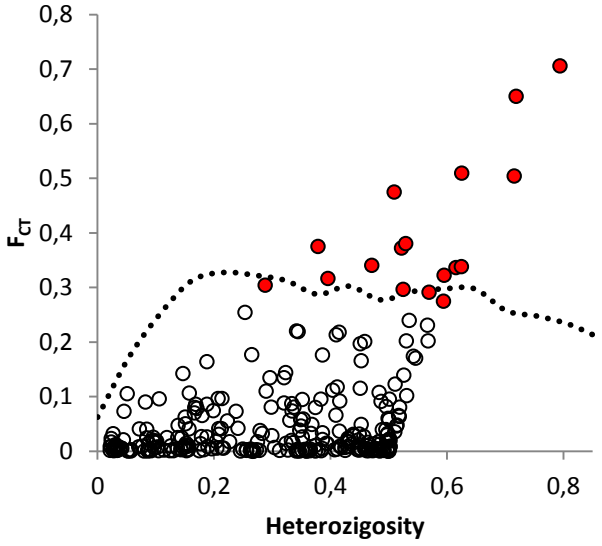

NTS
WSC
SIR
GAC
NPT
GFL
NTY
NAD
SAD
AEG
SSD
STY
NWJ
MAL
ALG
ADB
MAB
TKY
CYP

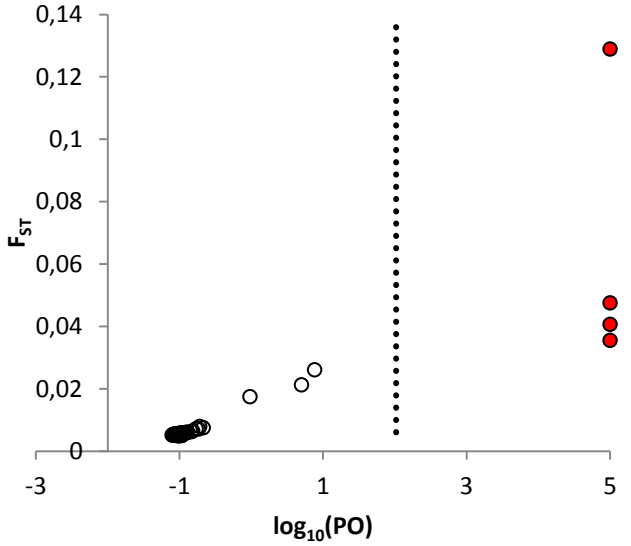
Figure 2



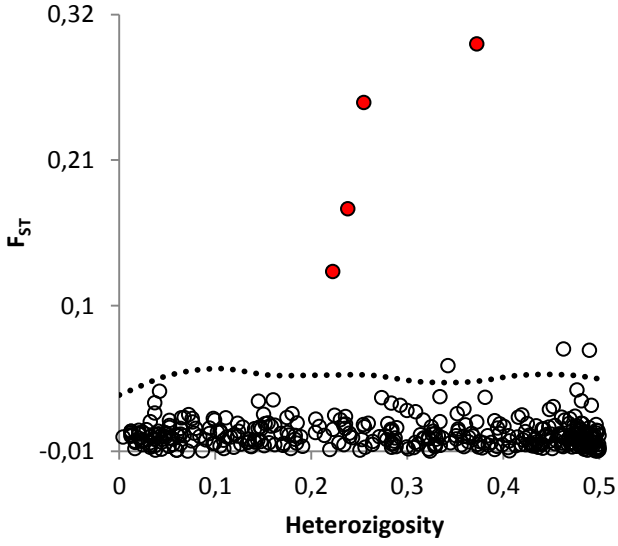
A.



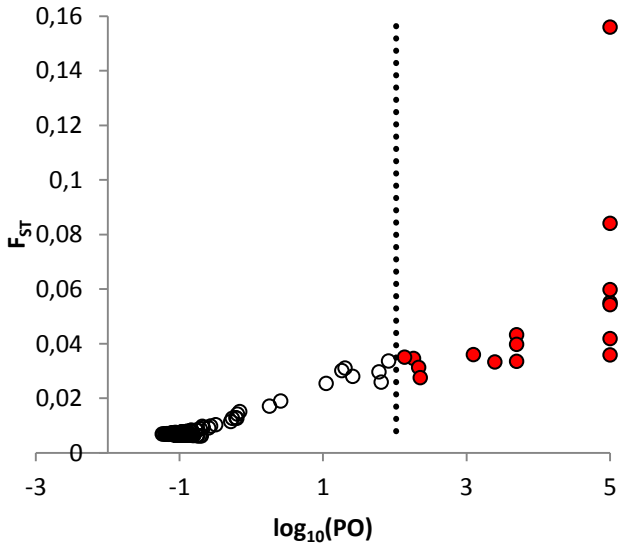
B.



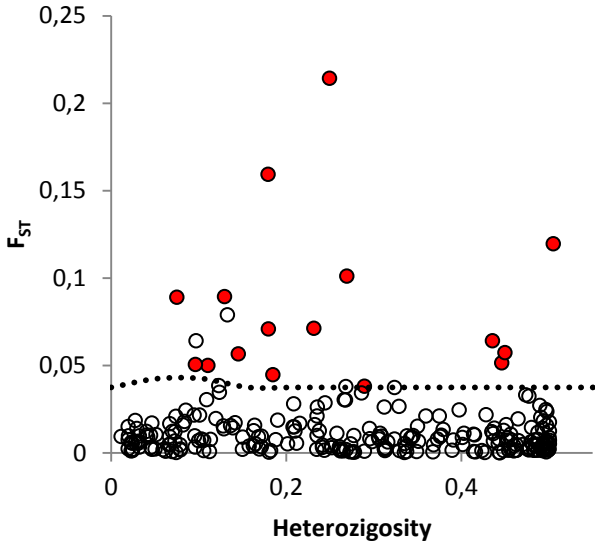
C.



D.



E.



F.

Figure 3

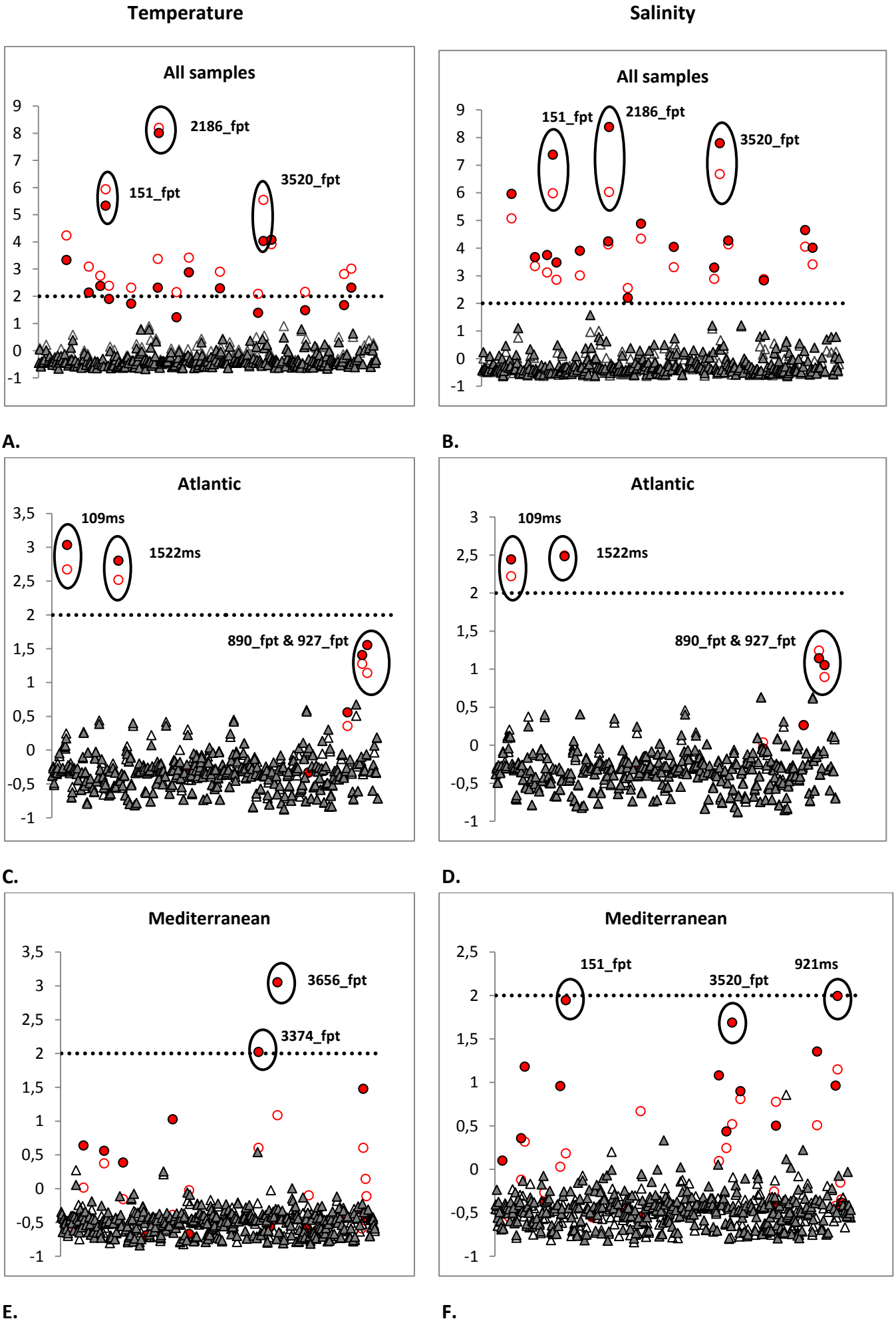
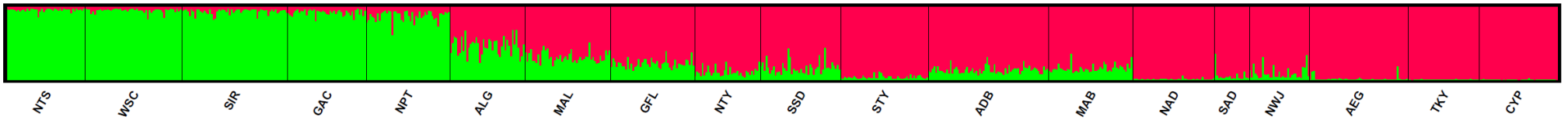
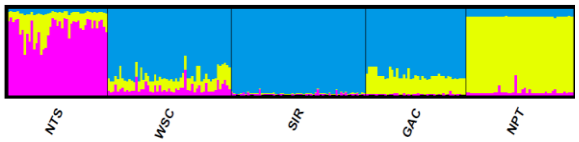

Temperature
Salinity
All samples
2186_fpt
151_fpt
3520_fpt
A.
B.
Atlantic
109ms
1522ms
890_fpt & 927_fpt
C.
D.
Mediterranean
3656_fpt
3374_fpt
151_fpt
921ms
3520_fpt
E.
F.

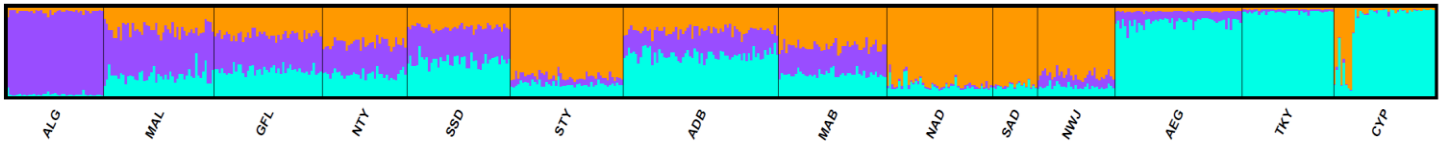
Figure 4


NTS
WSC
SIR
GAC
NPT
ALG
MAL
GFL
NTY
SSD
STY
ADB
MAB
NAD
SAD
NWJ
AEG
TKY
CYP

A.

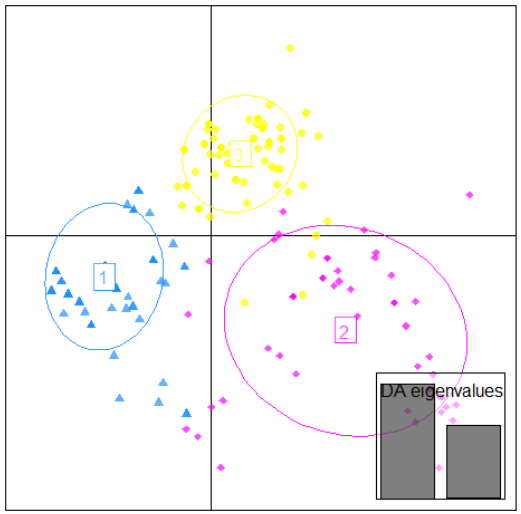

NTS
WSC
SIR
GAC
NPT

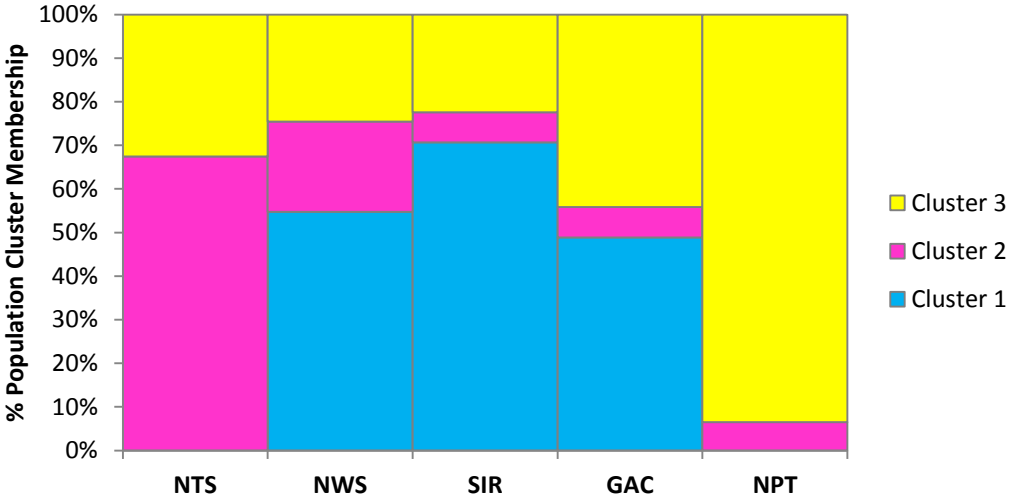
B.


ALG
MAL
GFL
NTY
SSD
STY
ADB
MAB
NAD
SAD
NWJ
AEG
TKY
CYP

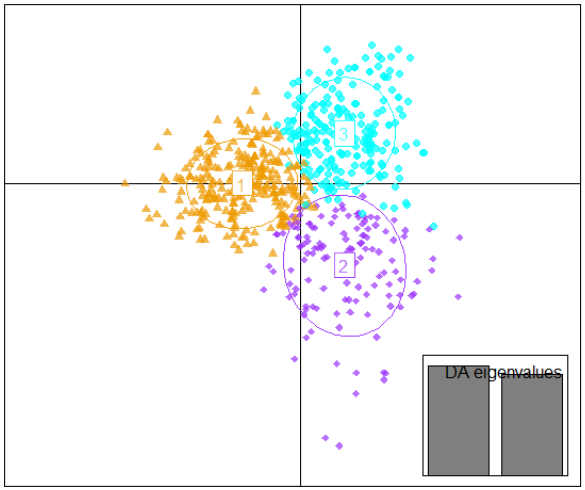
C.

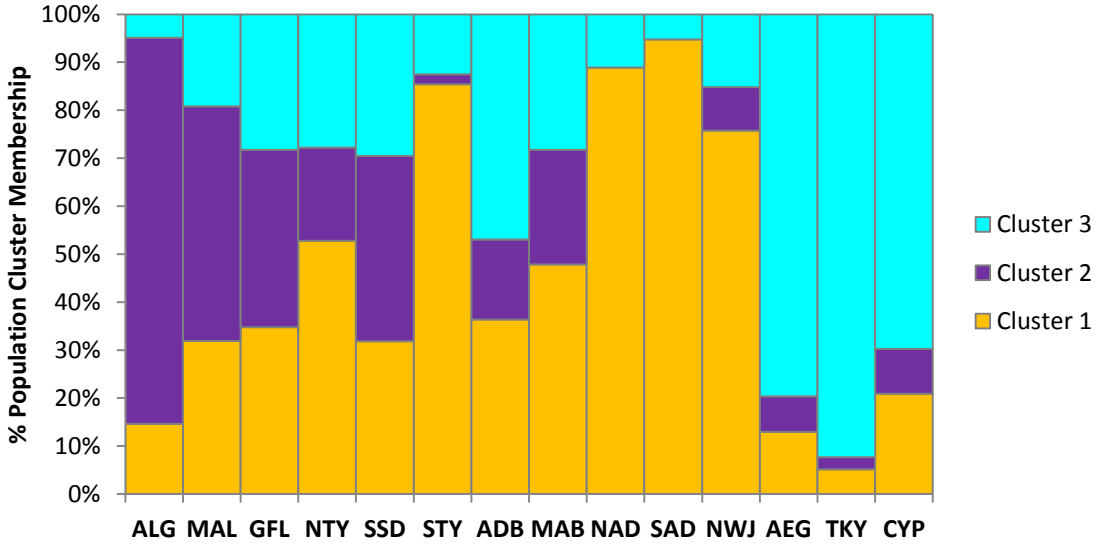
Figure 5


1
2
3
DA eigenvalues


% Population Cluster Membership
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
NTS
NWS
SIR
GAC
NPT
Cluster 3
Cluster 2
Cluster 1

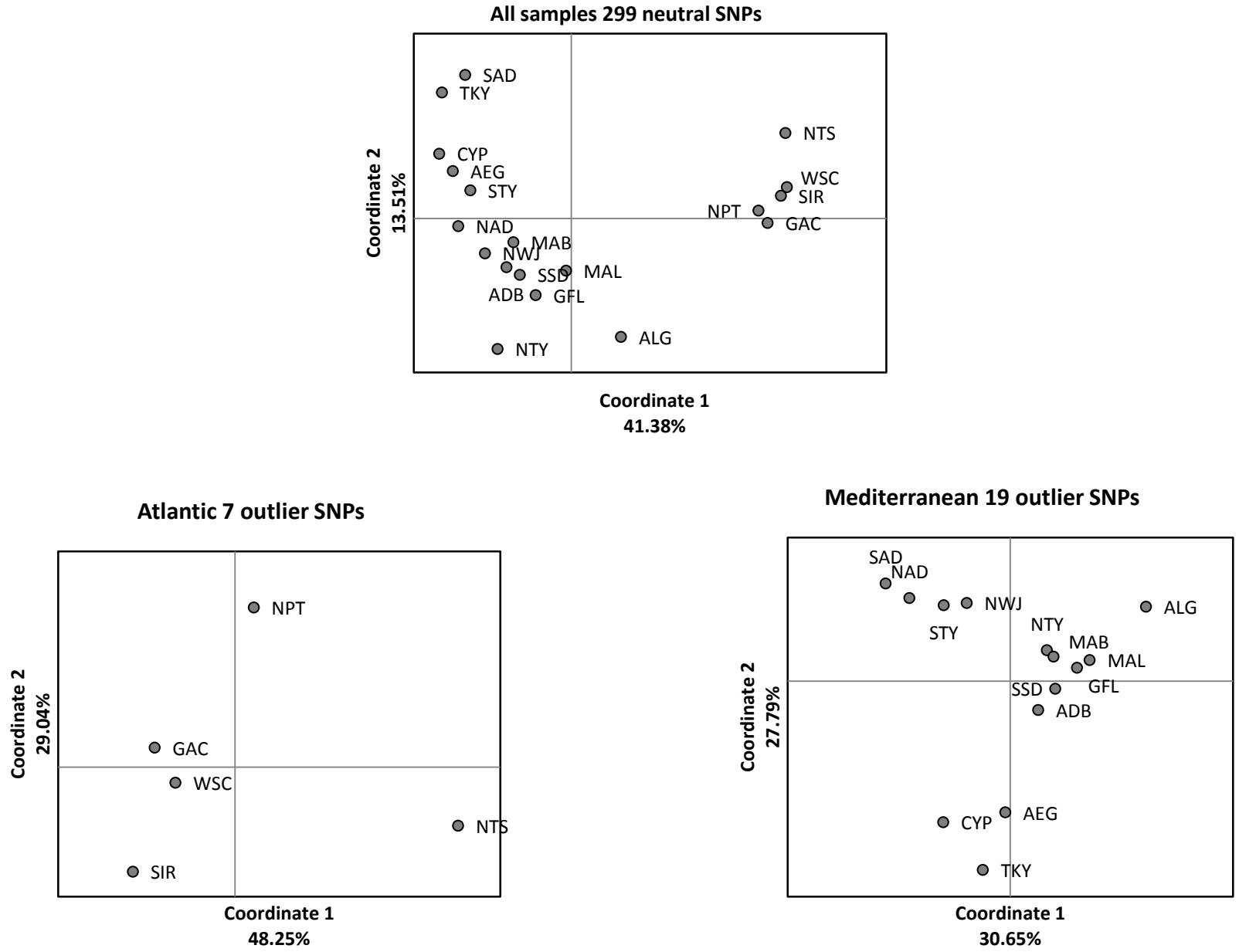
A.


1
2
3
DA eigenvalues


% Population Cluster Membership
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
ALG
MAL
GFL
NTY
SSD
STY
ADB
MAB
NAD
SAD
NWJ
AEG
TKY
CYP
Cluster 3
Cluster 2
Cluster 1

B.

Figure 6


All samples 299 neutral SNPs
SAD
TKY
CYP
AEG
STY
NAD
MAB
NWJ
SSD
MAL
ADB
GFL
NTY
ALG
NTS
WSC
SIR
NPT
GAC
Coordinate 2
13.51%
Coordinate 1
41.38%
Atlantic 7 outlier SNPs
NPT
GAC
WSC
NTS
SIR
Coordinate 2
29.04%
Coordinate 1
48.25%
Mediterranean 19 outlier SNPs
SAD
NAD
NWJ
STY
ALG
NTY
MAB
MAL
SSD
GFL
ADB
CYP
AEG
TKY
Coordinate 2
27.79%
Coordinate 1
30.65%

CHAPTER 6

GENE-ASSOCIATED MARKERS PROVIDE TOOLS FOR TACKLING IUU FISHING AND FALSE ECO-CERTIFICATION

Gene-associated markers provide tools for tackling IUU fishing and false eco-certification

Einar E Nielsen[1], Alessia Cariani[2,3], Eoin Mac Aoidh[4], Gregory E Maes[3], Ilaria Milano[2,5], Rob Ogden[6], Martin Taylor[7], Jakob Hemmer-Hansen[1], Massimiliano Babbucci[5], Luca Bargelloni[5], Dorte Bekkevold[1], Eveline Diopere[3], Leonie Grenfell[6], Sarah Helyar[8], Morten T Limborg[1], Jann T Martinsohn[4], Ross McEwing[6], Frank Panitz[9],Tomaso Patarnello[5], Fausto Tinti[2], Jeroen KJ Van Houdt[3], Filip AM Volckaert[3], Robin S Waples[10], FishPopTrace consortium & Gary R Carvalho[7]

1. Section for Population Ecology and – Genetics, National Institute of Aquatic Resources, Technical University of Denmark, Vejlsøvej 39, DK-8600 Silkeborg, Denmark

2. Department of Experimental Evolutionary Biology, University of Bologna, Bologna, I-40126, Italy

3. Laboratory of Biodiversity and Evolutionary Genomics, Katholieke Universiteit Leuven, Ch. Deberiotstraat, 32, B-3000 Leuven, Belgium

4. Institute for the Protection and Security of the Citizen (IPSC) Joint Research Centre (JRC) European Commission (EC) JRC.G.4 – Maritime Affairs, Via Enrico Fermi 2749 (TP 051) I-21027 Ispra (Va), Italy

5. Department of Public Health, Comparative Pathology, and Veterinary Hygiene, University of Padova, viale dell'Università 16, 35020 Legnaro, Italy

6. TRACE Wildlife Forensics Network, Royal Zoological Society of Scotland, Edinburgh EH12 6TS, UK

7. Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales, Bangor University, Bangor, Gwynedd LL57 2UW, UK

8. Matís, Vínlandsleið 12, 113 Reykjavík, Iceland

9. Department of Molecular Biology and Genetics, Faculty of Science and Technology, Aarhus University Blichers Allé 20, DK-8830 Tjele, Denmark

10. Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Boulevard East, Seattle, WA 98112-2013, USA

Abstract

Illegal, Unreported and Unregulated (IUU) fishing has played a major role in the overexploitation of global fish populations. In response, international regulations have been imposed and many fisheries have been "eco-certified" by consumer organisations, but methods for independent control of catch certificates and eco-labels are urgently needed. Here we show that by using gene-associated Single Nucleotide Polymorphisms (SNPs), individual marine fish can be assigned back to population of origin with unprecedented high levels of precision. By applying high differentiation SNP assays in four commercial marine fish on a pan-European scale we find 93-100% of individuals could be correctly assigned to origin in policy-driven case studies. We show how case-targeted SNP assays can be created and forensically validated using a centrally maintained and publicly available database. Our results demonstrate how application of gene-associated markers will likely revolutionise origin assignment and become highly valuable tools for fighting illegal fishing and mislabelling worldwide.

Commercially important fish populations worldwide have been overexploited and require rebuilding[1,2]. Tight regulations are needed to meet future marine fisheries and conservation objectives. Spatially explicit quotas and closures are common management actions for restoring fish populations and promoting sustainable "ecosystem based fisheries management[3]". IUU fishing poses a major threat to sustainable fisheries, constituting approximately one-fifth of the global catch[4,5]. Accordingly, international rules and laws such as the recent EC Control[6] and IUU Regulation[7]" have been instated, that require catch certificates that state the origin of all fish and fish products traded within the EU. Likewise, consumer awareness regarding vulnerability of fish stocks has been growing and many local fisheries have been awarded or are seeking sustainability certification ("eco-labelling") by organisations such as the Marine Stewardship Council (MSC)[8], despite recent concerns of mislabelling[9]. To enforce fisheries regulations and conservation measures spatially and provide independent control of catch certificates and mislabelling, forensically validated high-throughput identification methods, tracing individual fish to area/population of origin, are urgently needed.

Many tests have been developed to identify the population of origin of fish and fish products[10], that may be applied singly or in combination to address the generally low statistical power of any individual method. However, many such tests have been hampered by limited tissue availability or quality, especially in processed seafood. In addition, inter-laboratory standardisation of operating procedures is inherently difficult and compromises forensic validation[11].

DNA-based tools offer a universal method for assigning fish and fish products to population of origin. DNA is found in all cells in all organisms and can be analysed in any tissue type, from freshly caught fish to a fried fillet. Hitherto, DNA-based population assignment of marine fish has almost exclusively relied on so-called "neutral" genetic markers, compromised by weak population-level genetic signatures[12]. Such patterns coincide with the general lack of obvious physical barriers in the sea and typically large effective population sizes of marine fishes[13].

In contrast to population diversification by neutral genetic drift, evolution by natural selection proceeds more rapidly in large populations[14]. In order to identify genomic regions under divergent selection among natural populations, comparison of genetic differentiation for hundreds or thousands of genetic markers, so-called "genome scans" are increasingly being used[15]. Genome scans identify genetic markers with highly elevated divergence among populations which do not conform to statistical expectations based on a neutral genetic model. Accordingly, these markers are likely to be located within genomic regions with one or more gene loci under selection. Application of markers randomly distributed across the genome has generally provided a relatively low percentage of these high differentiation outliers putatively subject to selection[16]. Thus, targeting gene-associated markers, has been advocated as it increases the probability of identifying footprints of selection[16,17,18].

Many marine fish experience divergent environmental conditions, giving ample opportunity for heritable local adaptation[19,20]. Locally adapted genes will commonly display more divergent allele frequencies among populations than neutral markers and therefore display markedly elevated power for population assignment[12]. A relatively low number of genes with high genetic differentiation have been identified in marine fish to date, and even fewer used for population assignment[12]. Consequently, targeted identification of suites of gene-associated markers likely affected by direct or indirect, "hitch-hiking", selection will vastly enhance our ability to determine the population of origin of individual marine fish and elucidate their temporal and spatial dynamics.

We applied this new population genomics approach to four commercially important marine fish species; Atlantic cod (*Gadus morhua*), Atlantic herring (*Clupea harengus*), sole (*Solea solea*) and European hake (*Merluccius merluccius*), each threatened by overfishing and IUU activities. Policy-led scenarios of illegal fishing and/or mislabelling were selected for all four species. We demonstrated application of SNP tools across different geographical scales and in comparison to

previously published methods. 1) Cod: Northeast Arctic and Eastern Baltic cod populations thrive, while North Sea cod need rebuilding[21]. Strict spatially-based landing regulations are in place. Northeast Arctic and Eastern Baltic cod fisheries are MSC-certified. With the proximity and highly divergent status of these major cod populations there is a large potential for fraud and mislabelling. 2) Herring: No current method can distinguish North Sea from Northeast Atlantic herring (mainly "Norwegian spring" - and "Icelandic summer spawners"). Tracing the geographical origin of herring is important to MSC for certifying fisheries. 3) Sole: Most sole stocks of the Northeast Atlantic Ocean are in the process of rebuilding from high fishing mortalities. It is suspected that a proportion of sole landings in Belgian ports claimed to originate from the Irish Sea/Celtic Sea are in fact caught *en route* between the Irish Sea and the southern North Sea (Thames/Belgian coast), which is closer to market but closed to fishing. 4) Hake: Fishing regulations for hake differ between the Mediterranean and the Atlantic, with legal size limits of 20 and 27 cm respectively. Undersized Atlantic hake are misreported as of Mediterranean origin.

Results

Cod. From 21 geographical samples of Eastern Atlantic cod the genome-scan method identified 132 high differentiation outlier SNPs likely to be influenced by selection out of 1,262 variable and successfully genotyped loci (Fig.1a). For the case scenario of Northeast Arctic, North Sea and Baltic Sea cod a total of 69 out of 1,120 loci showed signs of being affected by divergent selection (Fig.2a) with interpopulation differences (F_{ST}) ranging from 0.10 to 0.51. Simulations identified a minimum assay with maximum power using eight of the highest ranked loci in terms of F_{ST} (between 0.07 and 0.51) which correctly assigned all fish to area of origin, except for one individual identified unambiguously as a North Sea migrant in the Baltic Sea. In a legal framework it is common practice to evaluate the relative likelihoods of observing the evidence under the

prosecution and defence hypothesis (or claims). In this case study the calculated likelihoods of observing a particular genotype was always more than six times higher in the true population of origin than for the second most likely population of origin; for 95% of the cod it was more than 1,500 times higher, while the median value was 600,000 times higher.

Herring: One-step genotyping and validation in herring revealed 281 variable SNPs genotyped in 18 Eastern Atlantic samples (Fig.1b). Overall, 16 SNPs were identified as significant outliers. Between Northeast Atlantic and North Sea herring, nine outlier SNPs were identified (Fig.2b). Simulations revealed that the 32 highest ranking SNPs (F_{ST} between 0.01 and 0.19) could correctly assign 100% of the Northeast Atlantic and 98% of the North Sea herring to their geographic origin (in total 161 out of 163 individuals). The log likelihood ratio between alternative hypotheses of origin, (the prosecutor versus defence claims) revealed that the true population of origin was always more than 3 times more likely (maximum 7 million times more likely) while the median value was 16,800 times more likely. The very few misassigned individuals had low likelihood ratios implicating uninformative genotypes rather than migrant individuals sampled in the other population group.

Sole. For sole, 27 Atlantic and Mediterranean samples were examined. Within the 16 Atlantic samples (Fig.1c), 19 of 427 SNPs appeared to be influenced by selection. For the Thames/Belgian coast versus Irish Sea/Celtic Sea scenario, three outliers were identified (F_{ST} values between 0.037 and 0.054). An *in silico* assay of 50 SNPs showing the highest F_{ST} values (0.005-0.054, Fig.2c) correctly assigned 93% (149 out of 160 individuals) to area of origin. The median log likelihood ratio between alternative populations of origin showed that an "average" individual was more than 60 times more likely in the population of origin even across this very small geographical scale with potentially large population mixing[22].

Hake. Hake collections (19 populations) covered Atlantic and Mediterranean basins (Fig.1d). In total, 72 of 395 SNPs were outliers (Fig.2d). 13 high F_{ST} SNPs (F_{ST} between 0.08 and 0.29) provided 98% (751 of 766 individuals) correct assignment to basin. Fourteen of 15 misassigned individuals originated from western Mediterranean samples (Algerian coast, Malaga) likely to be migrants or the result of admixture with neighbouring Atlantic populations. One individual sampled in the Atlantic was misassigned to the Mediterranean. Excluding likely migrants from the western Mediterranean, 99% of all individuals were assigned unambiguously to basin of origin. Evaluation of the likelihood of alternative hypotheses of origin showed that 95% of all sampled hake were over 500 times more likely to originate from their basin of sampling than to other basins.

Discussion

The policy-led IUU and mislabelling case scenarios demonstrate the large potential for using high differentiation SNPs for assigning individual marine fish to population of origin across a range of geographic scales. For any single assay, the gene-associated SNP framework provides unprecedented levels of assignment power for evaluating hypotheses of fish origin[23,24]. For hake a previous attempt of assigning fish to basin of origin (Atlantic/Mediterranean) using five microsatellites failed due to lack of statistical power[24]. The authors concluded: "…these two geographical stocks cannot be reliably identified from each other neither for fishery forensics nor for commercial traceability". In addition to the elevated power of assignment, these new SNP-based methods are more readily developed, validated and standardized (due to binary nature), in comparison to other markers such as microsatellites that require extensive inter-laboratory calibration[25], thereby providing potentially highly valuable legal evidence. Not only in most cases can we determine the fit of a genotype to a single population of origin, but as likelihoods of alternative explanations are bimodal, unequivocal evaluation of the prosecutor vs. defence claims is

possible. For the few cases where unambiguous assignment of individual fish to area of origin was not possible, statistical inferences from a number of individuals can be combined in order to provide the desired level of certainty. The "minimum markers with maximum power" are transferable across instruments, requiring limited cross-calibration among laboratories; the approach relies on a centrally accessible SNP database maintained by the European Commission Joint Research Centre. Accordingly, upon public release, any potential end-user can create and evaluate *in silico* assays tailored to specific control and enforcement or product certification scenarios. Typically, forensic authenticity testing examines specific alternative hypotheses of claimed geographical origin rather than the potential origin across the full species' distribution. Thus, targeted assays as presented here are faster, cheaper and more flexible than universal all-purpose SNP arrays. It is now possible to process and genotype several hundred fish per day with assays up to 100 SNPs for less than 25$ per individual in almost any reasonably well equipped molecular genetic laboratory. Finally, design flexibility allows the choice among speed, cost and statistical power, e.g. while high individual exclusion power is critical in a court of law; genotyping speed can be more essential in real-time spatially-based fisheries management. Here, rapid genotyping of a few SNP markers in many individuals may provide vital information on the approximate contribution of different populations to a specific marine fishery.

With any method there are potential pitfalls. The gene-associated SNP approach bears the inherent problem of genetic methods that management units are not necessarily equal to biological population units[26] Thus, different management regulations may be imposed for the same genetic population under different jurisdictions (and *vice versa*), leaving genetic methods with reduced discriminatory power. However, these limits to genetic resolution may also reflect ill-defined management areas[26]. Here, other methods such as elemental fingerprinting of otoliths or parasite distribution could prove complementary[10]. Likewise, we focused here on reproductively isolated

populations – i.e. the fundamental population unit. However, there will be some areas and times of year where mixed aggregations of individuals from different spawning populations occur and where assigning single individuals to a specific population may provide little information on geographic origin. Here mixed-stock analysis can be applied[27] potentially providing "mixture signatures" for management areas at certain times. Another special concern is the temporal stability of allele frequencies for genetic markers subject to environmental selection. We expect that most genetic changes will occur over evolutionary time scales; however, if direct or hitch-hiking selection acting on these markers is fast and on-going, allele frequencies in the reference populations could shift. To investigate temporal shifts, analysis of the SNP markers used for the minimum assay developed for cod has been conducted using temporally replicated samples, revealing very small non-significant changes in allele frequencies (p values between 0.11 and 0.92). 97% of contemporary cod samples correctly assigned to historical samples from the same population and the very few "misassigned" individuals are likely to represent migrant individuals from other populations (Fig.3). However, as the functional properties and relationship with environmental changes are unknown for most gene-associated SNPs, validation of the database should be conducted at intervals informed by the biology of the species and local conditions.

The present study examined the application of gene-associated SNPs, which are likely to be affected by adaptive evolution, as high-resolution tools for population traceability to tackle IUU and/or product mislabelling. The issue of SNP-associated gene function in fish has received little attention thus far. However, the many SNPs apparently subject to direct or indirect selection shown here and elsewhere in marine fish[12,20], strongly suggests that populations of marine organisms are genetically adapted to local environmental conditions despite high levels of gene flow. Therefore the examined SNPs are not mere "stamp collections[28]" without biological significance. In many cases they likely represent functional biological diversity in genes influencing survival and

reproduction. Such population diversity, or "biocomplexity[27]", underpins functioning, resilience and productivity of marine ecosystems. The "portfolio effect" of intraspecific biodiversity has been shown to stabilise ecosystem processes and services[29]. It is fortuitous that the adaptive genetic diversity that we aim to conserve underpins the tools that will allow enhanced governance of global fish resources.

Methods

Sampling. Tissue samples (flesh, gills or finclips) of cod (*Gadus morhua*), herring (*Clupea harengus*), sole (*Solea solea*) and hake (*Merluccius merluccius*) were collected on a pan-European scale including additional northwest Atlantic samples for cod (see Supplementary Table 1). Sampling was guided by previous genetic and ecological studies indicating population structuring in respective species. Spawning individuals were collected preferably in order to sample genetic populations. All individuals from an area (population sample) were at all occasions collected on the same cruise. Overall 85% of the baseline samples collected, including temporal replicates, originated from scientific cruises. The remaining 15% were collected by contracted commercial fishermen on designated cruises. The distribution of samples collected on scientific/commercial cruises was relatively uniform among species with scientific collections constituting 83%, 93%, 87%, 77% for cod, herring, sole and hake respectively. All samples included in the database were labelled with information on the approach of sampling (see Supplementary Table 1). As an additional check of the very unlikely event of any sample mislabelling or substitution (i.e. from vessel of sampling to SNP genotype database), patterns of population differentiation among all samples (pairwise F_{ST}) were evaluated after genotyping. This approach was used to identify any population samples that deviated from the general pattern of population structure established by this

or previous studies. However, as expected, no such aberrant population samples were identified. Specific detailed information on individual samples is available at the FishPopTrace database accessible at http//fishpoptrace.jrc.ec.europe.eu/data-access. For cod, samples originated from an extensive tissue bank maintained at the Danish Technical University established from previous studies[30].

SNP discovery and genotyping. SNPs for herring, sole and hake were identified through 454 sequencing (Roche 454 GS FLX sequencer) of the transcriptome. Accordingly, as the trancriptome consists of DNA segments transcribed into RNA molecules encoding at least one gene, the SNPs developed here are all gene-associated. Briefly, RNA was extracted from eight individuals from each species collected from four locations across the species range in order to minimize ascertainment (width) bias due to reduced geographic coverage. SNP discovery was performed by *de novo* sequence clustering and contig assembly, followed by mapping of reads against consensus contig sequences. 1,536 putative SNPs were selected from each species and included on an Illumina Golden Gate array for a one step validation and genotyping approach. Selection was based on information from a) the Illumina Assay Design Tool which assigns scores for each SNP based on the probability of them performing well in the genotyping assay, b) putative intron-exon boundaries within flanking regions of putative SNPs, and c) a visual evaluation of the quality of contig sequences. From these, 281 (herring), 427 (sole) and 395 (hake) SNPs proved variable with reliable genotyping across population samples. For cod we used a second generation Illumina 1,536 Golden Gate array with gene associated loci originating from previous sequencing projects[31,32,33]. Accordingly, a higher number of these (1,258) could be genotyped reliably across cod samples.

Identifying markers likely affected by selection. We used a Bayesian likelihood method implemented in BayeScan 2.01[34] for identifying markers likely to be situated in parts of the genome with one or more genes affected by selection. The method provides Posterior Odds (PO) as the ratio of the posterior probability of a model of selection versus a neutral genetic model for each locus. In addition, the new version of the program allows for setting *prior* odds for the two models. In this case we used the default option that a neutral model was 10 times more likely than a model with selection. Posterior Odds between 32 and 100 (log10(PO) = 1.5 – 2) is considered "very strong evidence" of selection while a Posterior Odds above 100 is viewed as "decisive" and finally, a posterior probability of infinity is assigned a log10(PO) of 5. The power of BAYESCAN for detecting markers affected by selection is significantly reduced for comparisons including few samples[34].

Choice of loci. For each case we chose several loci to create our "minimum assays with maximum power". Accordingly, the overarching aim was to provide assays with high statistical power, but also sufficiently small to be time and cost effective. The rationale behind this approach is that in a court of law the evidence will almost exclusively be weighted in relation to two alternative claimed origins. I.e. were the fish caught illegally in area A as claimed by the prosecutor or legally in area B as claimed by the defence? Choice of loci was based on estimates of pairwise genetic differentiation (F_{ST}) between samples and subsequent ranking of the loci according to the size of estimate from the outputs of BayeScan. This program employs the multinomial Dirichlet model allowing estimation of population specific F_{ST} coefficients. To reduce "high grading bias" in our assignment tests[35] we first estimated pairwise F_{ST} values and ranked our loci based on half of the individuals from each population as recommended by Anderson[35]. In addition we did not expect high grading bias to be particularly prominent when using our concept of application of markers under selection. The very

large differences in F_{ST} values between neutral and loci identified by the statistical model to be affected by selection found in this study is unlikely to be caused by sampling error in contrast to rather minute differences in F_{ST} values commonly found among neutral loci. Tightly linked genetic markers represent redundant information and violate model assumptions used for population assignment which require unlinked loci. Accordingly, we performed linkage analysis among all SNP marker loci intended for use in the minimum assays. In the case of complete or partial linkage between markers, we excluded the locus with the lowest F_{ST} or lowest genotyping success rate. From the final list of SNP loci included in the four described assays (Supplementary Table 2) we re-genotyped a subset of loci using other genotyping platforms to test for consistency, which was generally high.

Assignment procedure. Individuals from the baseline case samples were assigned to the population, or in the case of hake to basin (pooled samples within basin) where their multilocus genotype had the highest likelihood of occurring using the program GeneClass2.0[36]. We employed the Bayesian approach described by Rannala and Mountain[37] to evaluate whether a certain multilocus genotype could occur in (originate from) one or several of the baseline populations using the resampling algorithm described by Paetkau et al.[38]. The method, which simulates 10,000 multilocus genotypes per population from baseline allele frequencies, generates expected distributions of likelihoods within populations to compare with estimated individual likelihoods of real genotypes. To evaluate the relative likelihoods of potential alternative origins for a given genotype we calculated the likelihood ratio of originating from the sampled (home) population divided by the maximum likelihood for any of the other potential alternative populations of origin (L = L_home/L_max_not_home) following Peatkau et al.[38]. This approach is equivalent to a standard statistical evaluation of forensic evidence in relation to opposing claims from prosecutor

and defence in a court of law. I.e. in a potential case of illegal fishing or mislabelling the likelihoods of observing the genotype in question under the prosecutor and defence hypotheses of origin respectively are calculated and evaluated. Values were presented as median and 95% lower percentile values of –log likelihood ratios in order to illustrate the general high discrimination power of our selected SNP *in silico* assays. A few of the case scenario individuals had missing single locus genotypes. To maximize sample sizes they were not excluded for the assignment analyses except for the hake case where basin samples were plentiful. Incomplete genotypes are expected to reduce the assignment power, so the – log likelihood ratio medians and 95% lower percentiles presented here are expected to be upward biased, i.e. more conservative. However, there was no clear indication that misassigned individuals, or individuals assigned with low resolution, were caused by missing genotypic data.

Temporal stability. To evaluate temporal stability for assignment success with the baseline data we used temporal genetic data from cod populations. Short-term temporal stability was assessed through tests for genic differentiation using the program GenePop[39] and a principal component analysis (PCA) conducted with the package ADEGENET v. 1.2-5 for R[40] of individual genotype data from samples collected from Northeast Arctic cod, North Sea cod and Baltic Sea cod at two time points (four to ten years apart). Only the eight loci used in the assignment case for the same populations were used to generate the PCA in order to visualize the stability of population assignment observed for these specific loci.

References

1. FAO, *The State of World Fisheries and Aquaculture 2010* (Fisheries and Aquaculture Department), FAO; Rome (2011).
2. Worm, B. *et al*. Rebuilding global fisheries. *Science* **325**, 578-585 (2009).
3. Gaines, S. D., S.E. Lester, K. Grorud-Colvert, R., Costello, C. & Pollnac, R. Evolving science of marine reserves: New developments and emerging research frontiers *P. Natl. Acad. Sci.USA* **107**, 18251-18255 (2010).
4. Agnew, D. J. *et al.* Estimating the worldwide extent of illegal fishing *PLoS ONE* **4**, e4570 (2009). doi: 10.1371/journal.pone.0004570.
5. Flothmann, S. *et al*. Closing loopholes: getting illegal fishing under control *Science* **328,** 1235-1236 (2010).
6. Council Regulation (EC) No 1224/2009 of 20 November 2009.
7. Council Regulation (EC) No 1005/2008 of 29 September 2008.
8. www.msc.org
9. Marko, P. B., Nance, H. A. & Guynn, K. D. Genetic detection of mislabeled fish from a certified sustainable fishery. *Curr. Biol*. **21**, R6222011.
10. Cadrin, S. X., Friedland, K. D. & Waldman, J. R., eds. *Stock Identification Methods: Application in Fisheries Science* (Academic Press, London, 2005).
11. Ogden, R. Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish. Fish*. **9**, 462-472 (2008).
12. Nielsen, E. E., Hemmer-Hansen, J., Larsen, P. F. & Bekkevold, D. Population genomics of marine fishes: identifying adaptive variation in space and time. *Mol. Ecol*. **18**, 3128-3150 (2009).

13. Waples, R. S. Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species *J. Hered*. **89**, 438-450 (1998).

14. Nei, M. *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York 1987).

15. Strasburg, J. L. *et al.* What can patterns of differentiation across plant genomes tell us about adaptation and speciation? Philos. T. Roy. Soc. B. 367, 364-373 (2012).

16. Shimada, Y., Shikano, T. & Merila, J. A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. *Mol. Biol. Evol.* **28,** 181-193 (2011).

17. Vasemagi, A. & Primmer, C. R. Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. *Mol. Ecol.* **12**, 3623-3642 (2005).

18. Cosart, T. *et al.* Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**, 347, DOI: 10.1186/1471-2164-12-347

19. Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecol. Lett*. **7**, 1225-1241 (2004).

20. Bradbury, I. R. *et al.* Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. *P. R. Soc. B.* **277**, 3725-3734 (2010).

21. http://www.ices.dk/advice/fishstocks.asp

22. Cuveliers, E. L. *et al*. Microchemical variation in juvenile *Solea solea* otoliths as a powerful tool for studying connectivity in the North Sea. *Mar Ecol. Prog Ser*. **401**, 211-220 (2010).

23. Nielsen, E. E., Hansen, M. M., Schmidt, K., Meldrup, D. & Grønkjær, P. Fisheries - Population of origin of Atlantic cod. *Nature* **413**, 272 (2001).

24. Pita, A., Presa, P. & Perez, M. Gene flow, multilocus assignment and genetic structuring of the European hake (*Merluccius Merluccius*). *Thalassas* **26**, 129-133 (2010).

25. Morin, P. A., Luikart, G., & Wayne, R. K. SNPs in ecology, evolution and conservation. *Trends. Ecol. Evol*. **19**, 208-216. (2004).

26. Reiss, H., Hoarau, G., Dickey-Collas, M. & Wolff, W. J. Genetic population structure of marine fish: mismatch between biological and fisheries management units *Fish. Fish*. **10**, 361-395 (2009).

27. Ruzzante, D. E. *et al*. Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. *P. R. Soc. B.* **273**, 1459-1464 (2006).

28. Ferguson, M. M., & Danzman R. G. Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? *Can. J. Fish. Aquat. Sci.* **55**, 1553-1563 (1998).

29. Schindler, D. E. *et al*. Population diversity and the portfolio effect in an exploited species *Nature* **465**, 609-612 (2010).

30. Nielsen, E. E. *et al*. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evol. Biol.* **9**, 276 (2009). doi: 10.1186/1471-2148-9-276.

31. Hubert, S., Higgins, B., Borza, T. & Bowman, S. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*) *BMC Genomics* **11**, 191 (2010). doi: 10.1186/1471-2164-11-191.

32. Hemmer-Hansen, J., Nielsen, E. E. Meldrup, D. & Mittelholzer, C. Identification of single nucleotide polymorphisms in candidate genes for growth and reproduction in a nonmodel organism; the Atlantic cod, *Gadus morhua*. *Mol. Ecol. Resour.* **11**, 71-80 (2011).

33. Moen, T. *et al*. Identification and characterisation of novel SNP markers in Atlantic cod: Evidence for directional selection. *BMC Genet*. **9**, 18 (2008). doi: 10.1186/1471-2156-9-18

34. Foll, M. & Gaggiotti, O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective *Genetics* **180**, 977-993 (2008).

35. Anderson, E. C. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Mol. Ecol. Ressour*. **10**, 701-710 (2010).

36. Piry, S., Alapetite, A., Cornuet, J. M., Baudouin, L. & Estoup, A. GENECLASS2: A software for genetic assignment and first-generation migrant detection *J. Hered.* **95**, 536-539 (2004).

37. Rannala, B. & Mountain, J.L. Detecting immigration by using multilocus genotypes *P. Natl. Acad. Sci. USA* **17**, 9197-9201 (1997).

38. Paetkau, D., Slade, R., Burden, M. & Estoup, A. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power *Mol. Ecol.* **13**, 55-65 (2004).

39. Raymond, M. & Rousset, F. Genepop (version-1.2) – population-genetics software for exact tests and ecumenicism. *J. Hered*. **86**, 248-249 (1995).

40. R development core team 2011.

Acknowledgements We are grateful to the many colleagues who collected fish samples for this project. We thank the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement KBBE 212399 (FishPopTrace) for financial support. .E.E.N was also supported for part of the time by the Greenland Climate Research Center funded by the Danish Agency for Science, Technology and Innovation

Author Contributions G.C. was the Coordinator of the FP7 project, and conceived the project together with E.N, R.O., M.T., D.B., L.B., T.P, and F.V. F.T was in charge of the sample collection and archiving. F.P. headed the bioinformatics analysis for SNP detection assisted by G.M, L.B., S.H., M.L. and J.H. R.O., L.G. and R.M conducted the SNP genotyping and forensic validation. The genetic data analyses were carried out by: J.H-H. and E.N. (cod); D.B., S.H. and M.L. (herring); A.C., G.M. and E.D. (sole); I.M. and M.B. (hake). E.M. and J.M. were responsible for establishing and maintaining the sample and genotype database. R.W. provided overall input to the project as an external expert. E.N. and G.C. drafted the manuscript and coordinated input from all the named contributing authors.

Competing financial interests The authors declare no competing financial interests.

Figure Legends

Figure 1 | Map of sampling localities for the genetic baselines (white circles) and policy-led individual assignment case studies (coloured circles) for the four commercially important marine species. Shown is the percentage of fish assigned to the sample/area of origin and to other samples/areas (arrows). A) Atlantic cod (*Gadus morhua*) case study: Northeast Arctic cod (yellow), North Sea cod (blue), Baltic cod (red). B) Atlantic herring (*Clupea harengus*) case study: Northeast Atlantic herring (yellow), North Sea herring (blue). C) sole (*Solea solea*) case study: Irish Sea/Celtic Sea sole (blue), Thames/Belgian Coast (yellow). D) European hake (*Merluccius merluccius*) case study: Mediterranean hake (blue), Atlantic hake (yellow).

Figure 2 | Identification of outlier loci likely to be subject to selection in the four species using a model based genome scan approach. Each gene locus (grey circle) is represented by the level of genetic differentiation (F_{ST}) and log10 posterior odds (PO) of being under selection. Vertical dashed lines mark the threshold corresponding to a false discovery rate of 5%. Loci included in the minimum assays with maximum power are indicated (red circles). A) Atlantic cod. B) Atlantic herring C) Sole. D) European hake.

Figure 3 | Principal component analysis (PCA) plot based on individual genotypes from Northeast Arctic, North Sea and Baltic cod, illustrating temporal stability of assignment (6, 4 and 10 years between samples respectively) for the designated eight SNP panel.

Figure 1

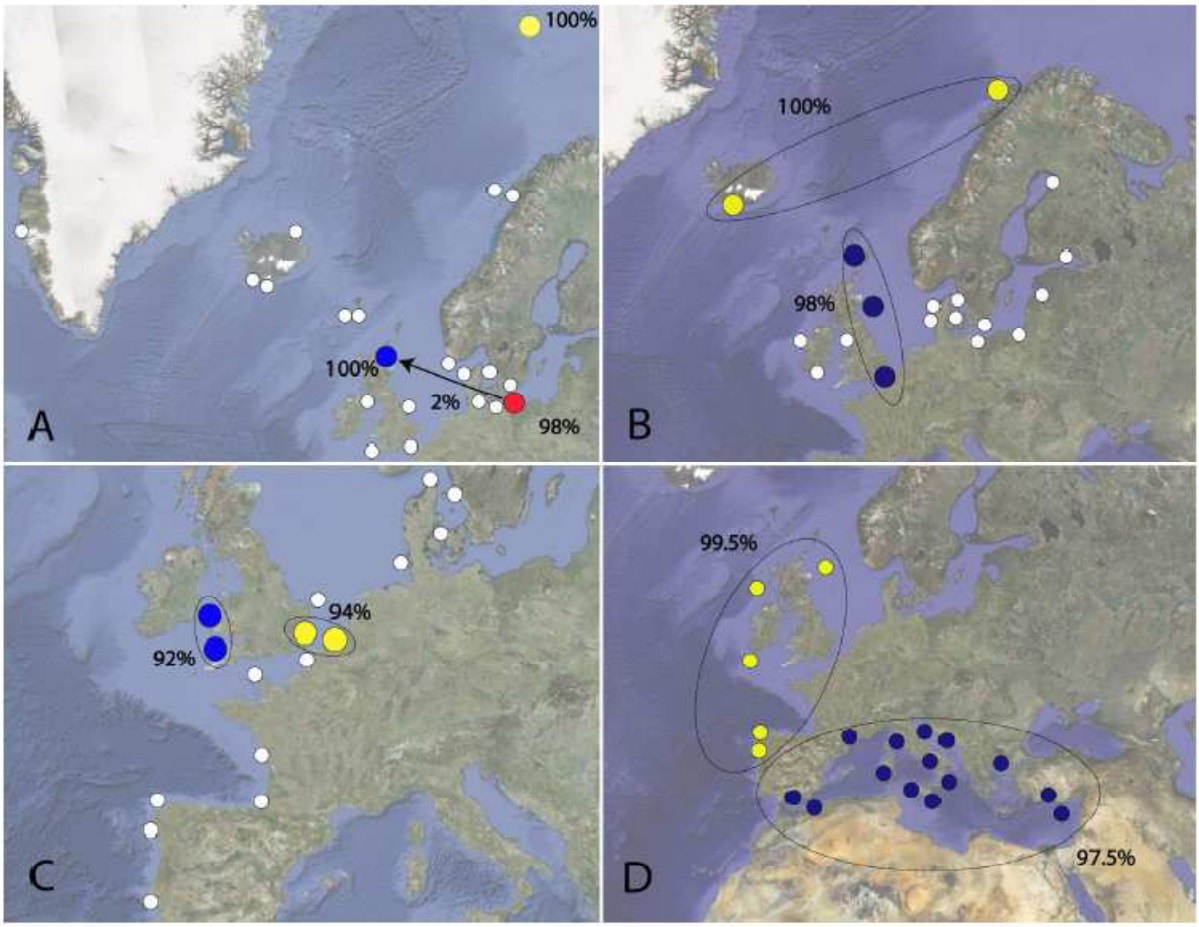


Figure 2

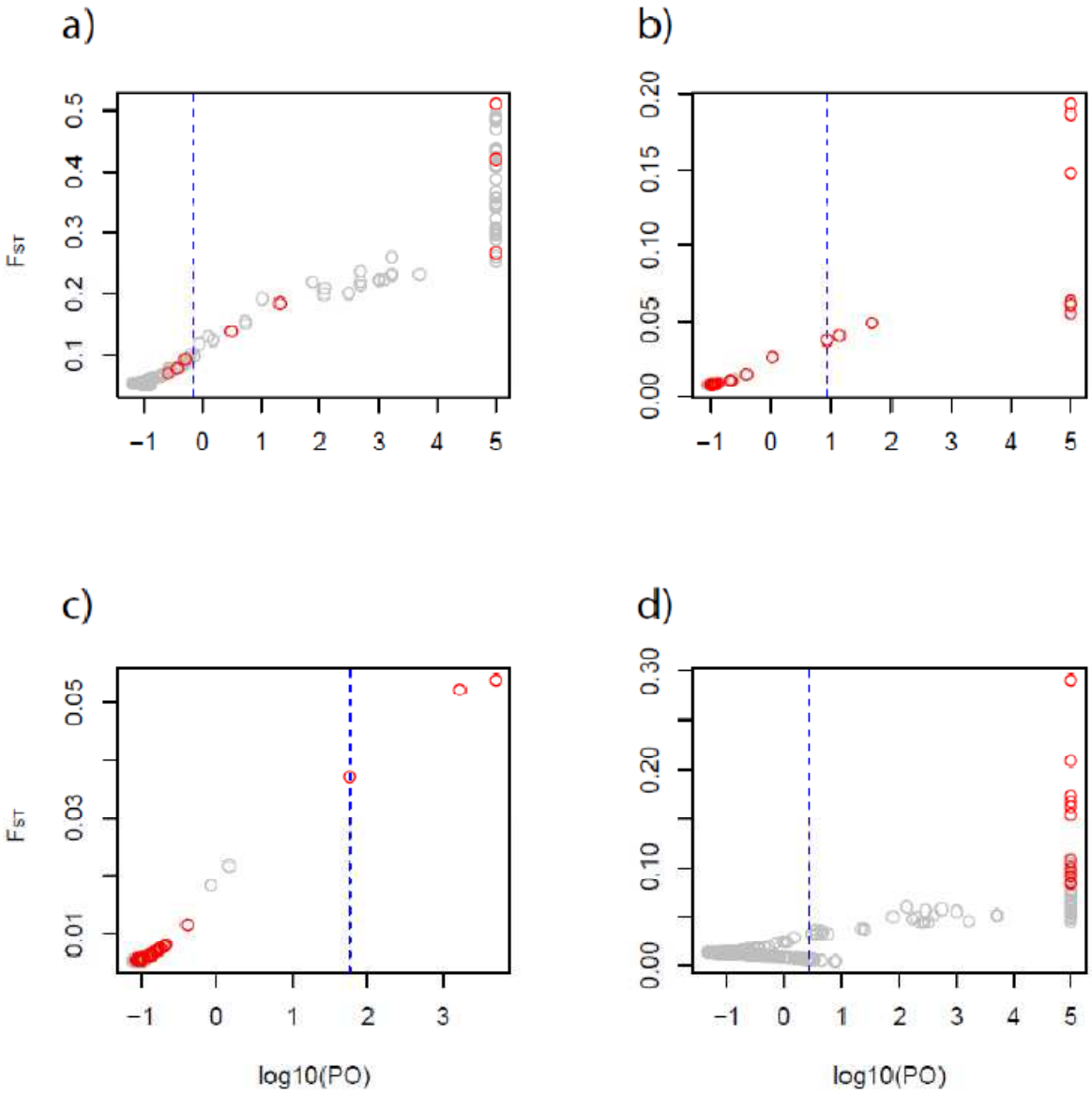


Figure 3

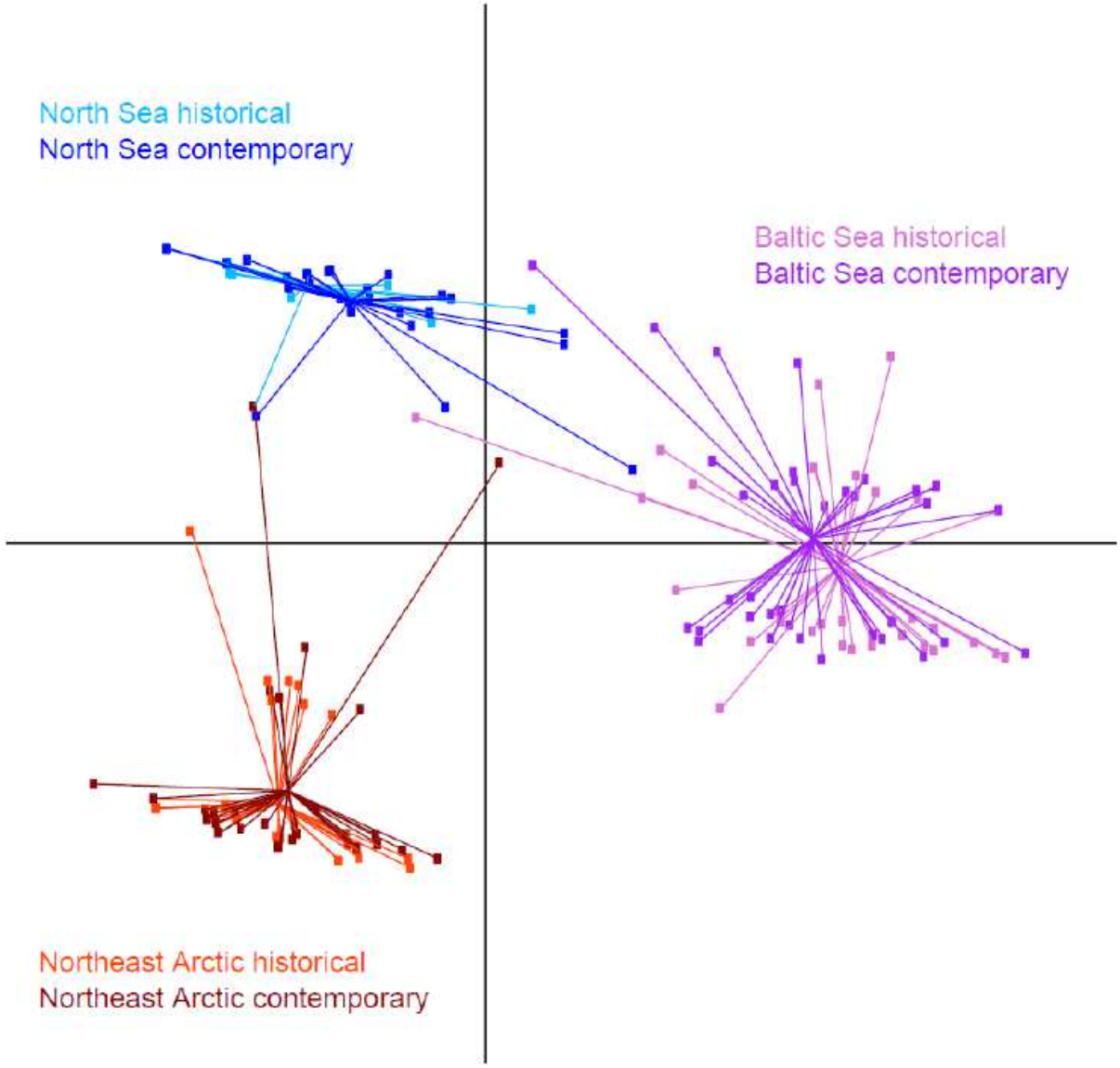

North Sea historical
North Sea contemporary
Baltic Sea historical
Baltic Sea contemporary
Northeast Arctic historical
Northeast Arctic contemporary

Table S1. Summary information on location position and number of individuals for the samples applied to provide a genetic baseline for the four species (B), for identification of markers under selection (S) for the individual assignment case studies (C) and temporal stability (T). Sample type refers to method employed for sampling, i.e. through A) Scientific cruise and B) contracted collection by commercial fishermen.

Species	Sampling location	Sample type	Latitude	Longitude	Number of individuals	Sampling year	Application
Gadus morhua	*27.III.d. Baltic Sea*						
	Arkona Basin	A	54.91	13.55	40	1996	B,S
	Bornhom Basin	A	54.87	15.46	40	1997	B,S,C,T
	Bornholm Basin	A	55.04	15.30	40	2007	T
	27.III.b,c - Transition Area						
	Belt Sea	A	55.11	10.28	40	1996	B,S
	Western Baltic Sea	A	54.56	12.28	35	2007	B,S
	27.III.a. Skagerrak and Kattegat						
	Kattegat	A	57.15	11.35	40	1996	B,S
	27.IV.b. Central North Sea						
	Eastern North Sea	B	56.91	7.83	40	2007	B,S
	Northeastern North Sea	B	57.75	5.50	40	2007	B,S,T
	Southern North Sea	B	54.29	0.02	40	2006	B,S
	27.VII.d. English Channel						
	English Channel	A	50.79	0.48	40	2005	B,S
	27.IV.a. Northern North Sea						
	Northern North Sea	A	58.00	-3.00	39	2003	B,S,T,C
	27.VII.a. Irish Sea						
	Irish Sea	A	54.62	-5.46	39	2006	B,S
	27.VII.f Bristol Channel						
	Celtic Sea	B	50.50	-5.16	39	1998	B,S
	27.Vb2 Faroe Bank						
	Faroe Bank	A	60.95	-8.49	40	2002	B,S
	27.Vb1 Faroe Plateau						
	Faroe Plateau	A	62.53	-6.16	40	2002	B,S
	27.XII.a - Norwegian Sea						
	Lofoten (NEAC)	A	68.35	12.14	39	2003	B,S,T,C
	Lofoten (NEAC)	A	67.33	11.38	39	2009	T
	Lofoten (NCC)	A	68.12	14.44	40	2003	B,S
	27.IIb2 Spitzbergen and Bear Island						
	Barents Sea (NEAC)	A	75.64	16.82	35	2009	B,S
	27.Ib Barents Sea Non-NEAFC Regulatory Area						
	White Sea	A	NA	NA	40	NA	B,S
	27.V.a. Iceland Grounds						
	Iceland south, inshore	A	63.49	-21.05	39	2002	B,S
	Iceland south, offshore	A	63.20	-19.30	39	2002	B,S
	Iceland north, inshore	A	66.30	-53.15	39	2002	B,S
	21.1B						
	Western Greenland	A	66.30	-53.15	39	2005	B,S

	21.4T						
	Canada	A	48.01	-63.55	39	2008	B
Clupea harengus	*27.III.d. Baltic Sea*						
	Bothnian Bay	A	65.05	24.58	33	2009	B,S
	Gulf of Finland	A	60.40	26.70	24	2009	B,S
	Gulf of Riga	A	57.83	22.83	17	2002	B,S
					27	2008	B,S
	Vistula Lagoon	A	54.37	19.67	17	2009	B,S
	Hanö Bay	A	55.57	15.18	24	2002	B,S
	III.c. Western Baltic						
	Greifswalder Bodden (Rügen)	A	54.21	13.62	19	2003	B,S
					36	2009	B,S
	27.III.a. Skagerrak and Kattegat						
	Skagerrak south	A	57.40	11.40	36	2009	B,S
	Kattegat	A	55.73	11.37	23	2003	B,S
	Limfjorden	B	56.60	8.35	33	2009	B,S
	27.IV.b. Central North Sea						
	North Sea	A	56.43	0.20	30	2009	B,S,C
	Ringkøbing Fjord	A	55.97	8.24	33	2009	B,S
	VII.d. English Channel						
	English Channel	A	50.81	1.57	17	1999	B,S
					39	2009	B,S,C
	IV.a. Northern North Sea						
	Shetland	A	60.35	-2.72	34	2009	B,S,C
	VII.a. Irish Sea						
	Douglas Bank	A	54.03	-4.07	36	2009	B,S
	VII.g. Celtic Sea						
	Celtic Sea North	A	51.24	-8.26	39	2008	B,S
	VII.b. West of Ireland						
	West of Ireland	A	53.90	-10.36	28	2004	B,S
	II.a. Norwegian Sea						
	Norwegian Sea	A	65.54	11.26	31	2009	B,S,C
	V.a. Iceland						
	Iceland south	A	63.62	-19.62	34	2009	B,S,C
Solea solea	*27.III.a - Skagerrak and Kattegat*						
	Belt Sea	A	55.65	10.76	40	2007	B
	Kattegat	A	57.16	11.65	40	2007	B
	Skagerrak	A	57.78	9.99	40	2007	B
	27.IV.b - Central North Sea						
	German Bight	A	54.52	7.89	40	2007	B
	27.IV.c - Southern North Sea						
	Norfolk	A	52.92	2.24	28	2008	B
	Belgian coast	A	51.39	3.17	40	2008	B, S, C
		A	51.22	2.83	24	2009	B
	Thames Estuary	A	51.47	1.33	40	2007	B, S, C
	27.VII.d - Eastern English Channel						

	Eastern English Channel	A	50.78	1.48	40	2008	B
	27.VII.e - Western English Channel						
	Western English Channel	A	49.66	-2.13	40	2009	B
	27.VII.a - Irish Sea						
	Bristol Channel	A	52.21	-5.33	40	2008	B, S, C
	27.VII.g - Celtic Sea North						
	Cornwall Coast	A	50.81	-5.01	40	2008	B, S, C
	27.VIII.a - Bay of Biscay - North						
	Pertuis Breton	A	45.92	-1.69	40	2009	B
	27.VIII.c - Bay of Biscay - South						
	Saint Jean de Luz	B	43.55	-1.57	39	2003	B
	Spain	A	43.60	-8.86	40	2009	B
	27.IX.a - Portuguese Waters - East						
	Northern Portuguese Waters	A	42.14	-9.28	40	2009	B
	Lisbon	B	38.35	-9.35	39	2003	B
	37.1.1 - Balearic						
	Castellon	A	40.00	0.12	27	2000	B
	Barcellona	A	41.23	2.33	14	1999	B
	37.1.2 -Gulf of Lion						
	Sète	B	43.02	4.17	40	2003	B
		B	43.17	3.98	39	2009	B
	37.1.3 - Sardinia						
	Viareggio, Northern Tyrrhenian Sea	A	43.30	9.54	40	2009	B
	Salerno Gulf, Southern Tyrrhenian Sea	A	40.53	14.73	30	2000	B
	Sant'Eufemia Gulf, Southern Tyrrhenian Sea	A	38.96	14.48	40	2003	B
	37.2.2 - Ionian						
	South Adriatic Albanian Coast	B	41.28	19.13	14	2000	B
	South Adriatic Italian Coast	A	42.02	15.40	19	2000	B
	37.2.1 - North Adriatic						
	Chioggia Lagoon, North Adriatic	A	44.73	13.27	40	2009	B
	37.3.1 - Aegean						
	Gulf of Kavala, Northern Greece	A	40.85	24.49	40	2009	B
	37.3.2 - Levant						
	Turkish Coast	A	35.95	35.13	40	2002	B
		A	36.75	33.87	27	2009	B
Merluccius merluccius	*27.IV.a - Northern North Sea*						
	North Sea	A	57.94	0.21	44	2009	B, S, C
	27.VI.a - West of Scotland						
	West of Scotland	A	56.21	-9.21	57	2009	B, S, C
	27.VII.j - Southwest of Ireland - East						
	Southwest of Ireland	A	50.43	-9.97	60	2009	B, S, C
	27.VIII.c - Bay of Biscay - South						
	Galician Coast	A	43.30	-9.18	46	2009	B, S, C

	2*7.IX.a - Portuguese Waters - East*						
	North of Portugal	A	42.12	-9.38	55	2009	B, S, C
	37.1.1 - Balearic						
	Algeria	B	35.61	-2.12	46	2009	B, S, C
	Malaga	B	36.63	-4.37	49	2009	B, S, C
	37.1.2 - Gulf of Lions						
	Gulf of Lions	B	43.35	3.73	46	2009	B, S, C
	37.1.3- Sardinia						
	North Tyrrhenian	A	42.51	10.11	44	2009	B, S, C
	South of Sardinia	B	38.99	8.51	46	2009	B, S, C
	South Tyrrhenian	A	40.54	14.77	48	2009	B, S, C
	37.2.2 - Ionian						
	Sicilian Channel - Adventure Bank	A	37.33	12.17	71	2008	B, S, C
	Sicilian Channel - Maltese Bank	A	35.92	14.79	47	2008	B, S, C
	North Western Ionian Sea	A	38.23	16.43	40	2008	B, S, C
	South Adriatic	B	42.26	16.89	22	2009	B, S, C
	37.2.1 - Adriatic						
	North Adriatic	A	43.92	13.79	49	2009	B, S, C
		A			33	2007	
		A			45	2002	
		A			32	1998	
	37.3.1 - Aegean						
	North Aegean Sea	A	40.34	24.52	58	2009	B, S, C
	37.3.2 - Levant						
	Turkish Coast	A	36.78	30.94	39	2009	B, S, C
	Cyprus	A	34.52	32.90	46	2009	B, S, C

Table S2: List of loci used for the case specific assays for the four species. Included is identity number, F_{ST} value, GenBank accession (will be uploaded prior to publication) number and information on loci subject to re-genotyping, the associated technology and the percentage of identical genotypes compared to the standard Illumina Golden Gate genotyping.

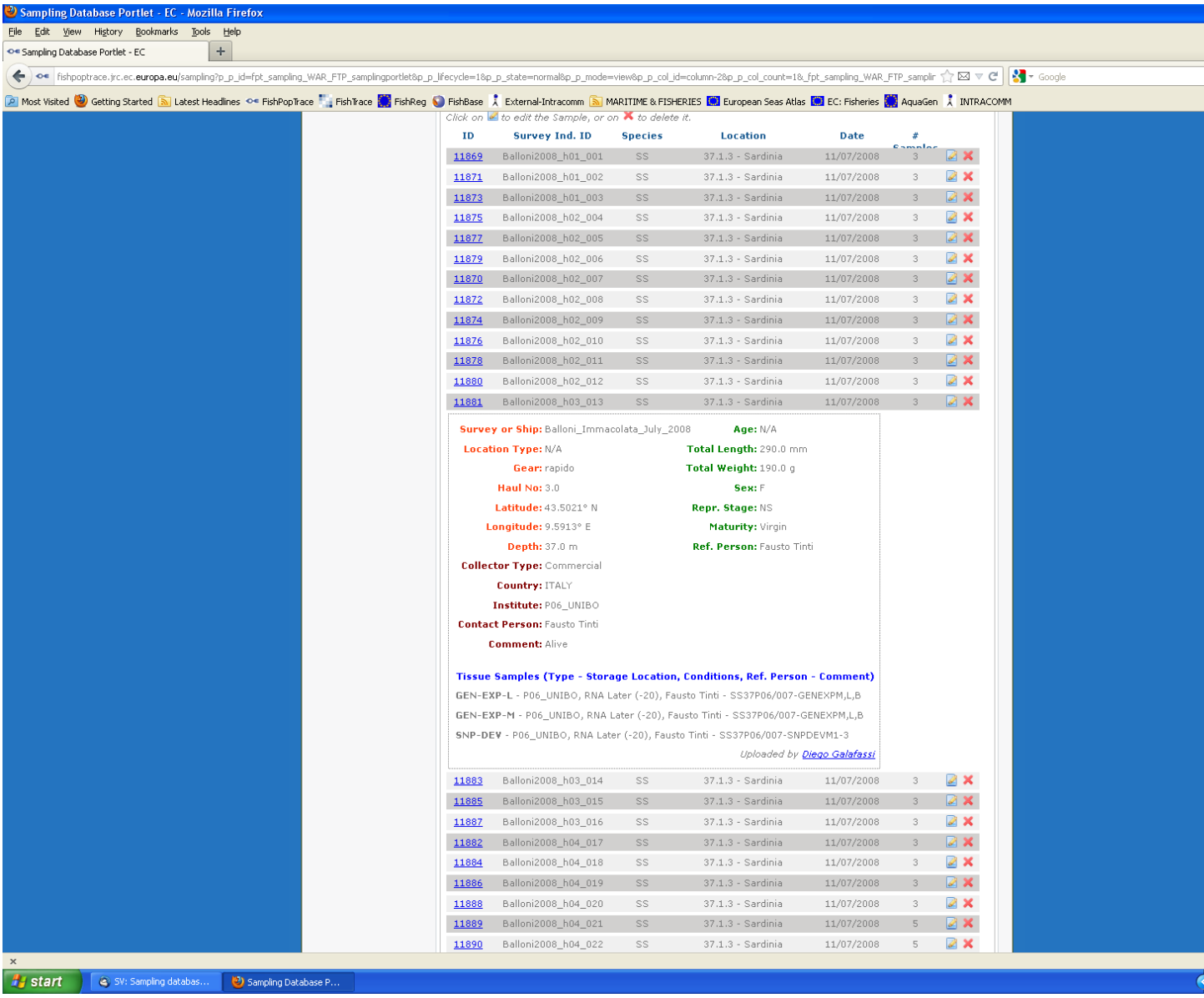
Species	SNP identification number	F_{ST} (BayeScan)	F_{ST} (GenePop, θ)	GenBank accession number	Re-genotyped/ technology	Percentage of identical genotypes
G. morhua	cgpGmo-S1068	0.42	0.75	ss131570937	Kaspar	89
	cgpGmo-S1777	0.09	0.29	rs119055455	-	-
	cgpGmo-S863	0.14	0.43	rs119055265	-	
	cgpGmo-S1066	0.07	0.27	rs119056441	-	-
	Hsp90	0.27	0.74	-	Kaspar	100
	cgpGmo-S1166	0.51	0.95	rs119055013	Kaspar	98
	cgpGmo-S227	0.08	0.23	rs119055651	-	-
	Gm_cyp19a1a_1_10	0.19	0.54	ss252841231	SNaPShot	100
C. harengus	snp.77	0.193	0.536	-	SNaPShot	93

	snp.3279	0.187	0.260	-	SNaPShot	98
	snp.3949	0.148	0.200	-	SNaPShot	99
	snp.675	0.063	0.182	-	SNaPShot	95
	snp.1291	0.061	0.131	-	SNaPShot	93
	snp.288	0.056	0.137	-	-	-
	snp.1360	0.049	0.097	-	SNaPShot	91
	snp.3755	0.040	0.080	-	-	-
	snp.3316	0.037	0.097	-	-	-
	snp.2534	0.026	0.047	-	SNaPShot	100
	snp.737	0.014	0.062	-	SNaPShot	97
	snp.3248	0.010	0.056	-	SNaPShot	100
	snp.3035	0.010	0.055	-	-	-
	snp.799	0.008	0.042	-	SNaPShot	96
	snp.3953	0.008	0.020	-	SNaPShot	97
	snp.840	0.008	0.029	-	-	-
	snp.987	0.008	0.021	-	-	-
	snp.1692	0.008	0.026	-	SNaPShot	98
	snp.2975	0.008	0.025	-	-	-
	snp.3001	0.008	0.022	-	-	-
	snp.500	0.008	0.023	-	SNaPShot	97
	snp.2453	0.008	0.020	-	SNaPShot	97
	snp.1390	0.007	0.018	-	-	-
	snp.2740	0.007	0.014	-	-	-
	snp.659	0.007	0.010	-	-	-
	snp.1874	0.007	0.011	-	-	-
	snp.3543	0.007	0.010	-	-	-
	snp.561	0.007	0.010	-	-	-
	snp.5292	0.007	0.010	-	-	-
	snp.438	0.007	0.010	-	-	-
	snp.2748	0.007	0.010	-	SNaPShot	97
	snp.3499	0.007	0.010	-	SNaPShot	100
S. solea	SNP200	0.052	0.183	-	Kaspar	96
	SNP228	0.054	0.172	-	-	-
	SNP1354	0.037	0.118	-	-	-
	SNP116	0.012	0.097	-	Kaspar	99
	SNP1347	0.008	0.078	-	Kaspar	99
	SNP1515	0.007	0.077	-	Kaspar	95
	SNP923	0.008	0.077	-	-	-
	SNP1466	0.007	0.076	-	Kaspar	99
	SNP499	0.007	0.071	-	-	-
	SNP1478	0.006	0.069	-	-	-
	SNP1052	0.006	0.064	-	Kaspar	98
	SNP1516	0.006	0.060	-	-	-
	SNP933	0.007	0.056	-	-	-
	SNP1003	0.006	0.042	-	-	-
	SNP1094	0.007	0.041	-	-	-
	SNP7	0.008	0.036	-	Kaspar	89
	SNP780	0.006	0.036	-	-	-

	SNP1203	0.006	0.033	-	-	-
	SNP1415	0.006	0.031	-	-	-
	SNP1129	0.006	0.030	-	-	-
	SNP451	0.006	0.030	-	-	-
	SNP932	0.007	0.030	-	Kaspar	96
	SNP1250	0.006	0.029	-	-	-
	SNP1147	0.005	0.028	-	-	-
	SNP1236	0.006	0.026	-	-	-
	SNP1214	0.005	0.026	-	-	-
	SNP962	0.006	0.025	-	-	-
	SNP1512	0.006	0.025	-	-	-
	SNP1168	0.006	0.025	-	-	-
	SNP360	0.006	0.025	-	-	-
	SNP1445	0.006	0.025	-	-	-
	SNP1484	0.006	0.024	-	-	-
	SNP914	0.005	0.023	-	-	-
	SNP1359	0.006	0.023	-	-	-
	SNP1213	0.005	0.022	-	-	-
	SNP1402	0.005	0.021	-	-	-
	SNP235	0.006	0.020	-	-	-
	SNP871	0.005	0.019	-	-	-
	SNP915	0.005	0.019	-	-	-
	SNP778	0.006	0.019	-	Yes/Kaspar	93.75
	SNP1238	0.005	0.018	-	-	-
	SNP769	0.005	0.018	-	-	-
	SNP851	0.005	0.017	-	-	-
	SNP350	0.006	0.017	-	-	-
	SNP577	0.005	0.017	-	-	-
	SNP284	0.006	0.016	-	-	-
	SNP800	0.005	0.016	-	-	-
	SNP410	0.005	0.016	-	-	-
	SNP1114	0.006	0.015	-	-	-
	SNP811	0.005	0.014	-	-	-
M. merluccius	2186_fpt	0.292	0.565	-	Kaspar	96
	3520_fpt	0.208	0.466	-	Kaspar	95
	151_fpt	0.173	0.331	-	Kaspar	95
	1178ms	0.154	0.335	-	Kaspar	97
	778ms	0.167	0.345	-	Kaspar	94
	2579_fpt	0.107	0.193	-	Kaspar	94
	2184_fpt	0.095	0.182	-	Kaspar	93
	3656_fpt	0.161	0.239	-	-	-
	1398_fpt	0.083	0.171	-	Kaspar	92

3474_fpt	0.100	0.212	-	-	-
1805_fpt	0.091	0.157	-	Kaspar	95
1580ms	0.084	0.187	-	-	-
1485_fpt	0.090	0.174	-	-	-

Figure S1: Screen print of a typical sample entry with accompanying information from the FishPopTrace sampling database.



CHAPTER 7

CONCLUSIONS

Based on the findings attained in the framework of my PhD project, it is worth outlining some general concluding remarks.

First of all, using the European hake as a case study, the work presented in my thesis has demonstrated the enormous potential of high-throughput sequencing technologies to facilitate the access to genomic resources of non-model species and accordingly the development of large panels of molecular markers in a cost effective way. A total of 395 SNPs were developed from transcript sequences and successfully validated for the European hake, representing high statistical power tools for a broad range of future applications in population genetics and conservation studies, so far limited by the low number and scarce informativeness of available molecular markers. The rapid improvement in sequencing throughput and the parallel decrease of costs, the availability of refined bioinformatic tools and the ever more efficient strategies explored to develop molecular markers from large-scale sequencing data are going to revolutionize future research on non-model organisms.

Notably, focusing the search for novel genetic tools on the transcriptome, the portion of the genome transcribed into RNA molecules, may provide the opportunity to identify candidate genes for local adaptation in wild populations. This strategy is particularly valuable to explore adaptive variation in marine species, for which common-garden experiments are often time-consuming and not always possible. The screening of European hake transcriptome-derived SNPs allowed the identification of several genes involved in energy pathways and metabolic processes potentially under divergent selection, corroborating the hypothesis advanced in previous studies on the possible role of environmental selection mechanisms in shaping the genetic structure of this species. Indeed

further investigations targeting these genomic regions are needed to consistently support the role played by natural selection. However, the signal of strong genetic differentiation arising from SNPs located on candidate genes left no doubts about the existence of fine-scale structuring of European hake populations that do not match the current stock management, and that conventional neutral markers failed to detect. The increased differentiation of the so called "outlier" loci might be attributable to both exogenous or endogenous barriers. Anyway, following a precautionary approach, the genetic pattern unveiled should be contemplated in a management and conservation context, in order to preserve both demographically independent units and locally adapted populations. Hence, results obtained within my PhD project highlight the importance of integrating patterns of neutral and adaptive genetic variation in a conservation framework.

The application of loci potentially under selection is not limited the identification of evolutionary units and the preservation of species' adaptive potential, but may also interest wildlife forensic science. Due to the increased discrimination power, outlier loci may be applied in a traceability framework moving beyond the species identification, and targeting at the population-level. In the specific case of European hake, the traceability scenario concerned the identification of the putative stock, Atlantic or Mediterranean, managed under different fishing regulations. Once the genetic composition of potential source populations has been outlined, a reduced panel of SNPs having the highest resolution power was selected and successfully applied to correctly assign almost all individuals to the geographic basin of origin. This case study illustrates how molecular analytical technologies have operational potential in real-world contexts, and more specifically, potential to support fisheries control and enforcement and fish and fish product traceability.

REFERENCES

Abella A, Fiorentino F, Mannini A, Orsi Relini L (2008) Exploring relationships between recruitment of European hake (Merluccius merluccius L. 1758) and environmental factors in the Ligurian Sea and the Strait of Sicily (Central Mediterranean). *Journal of Marine Systems* **71**, 279-293.

Abella A, Serena F, Ria M (2005) Distributional response to variations in abundance over spatial and temporal scales for juveniles of European hake (Merluccius merluccius) in the Western Mediterranean Sea. *Fisheries Research* **71**, 295-310.

AdriaMed F (2006) Jabuka/Pomo Pit: a critical area for the Adriatic demersal fisheries resources. Scientific issues and management options. Summary Notes. FAO.

Agnew DJ, Pearce J, Pramod G*, et al.* (2009) Estimating the Worldwide Extent of Illegal Fishing. *PLoS ONE* **4**, e4570.

Alvarez P, Fives J, Motos L, Santos M (2004) Distribution and abundance of European hake Merluccius merluccius (L.), eggs and larvae in the North East Atlantic waters in 1995 and 1998 in relation to hydrographic conditions. *Journal of Plankton Research* **26**, 811-826.

Alvarez P, Motos L, Uriarte A, EgaÃ±a J (2001a) Spatial and temporal distribution of European hake, Merluccius merluccius (L.), eggs and larvae in relation to hydrographical conditions in the Bay of Biscay. *Fisheries Research* **50**, 111-128.

Alvarez P, Motos L, Uriarte A, EgaÃ±a J (2001b) Spatial and temporal distribution of European hake, Merluccius merluccius (L.), eggs and larvae in relation to hydrographical conditions in the Bay of Biscay. *Fisheries Research* **50**, 111-128.

Arneri E, Morales-Nin B (2000) Aspects of the early life history of European hake from the central Adriatic. *Journal of Fish Biology* **56**, 1368-1380.

Bartolino V, Colloca F, Sartor P, Ardizzone G (2008a) Modelling recruitment dynamics of hake, Merluccius merluccius, in the central Mediterranean in relation to key environmental variables. *Fisheries Research* **92**, 277-288.

Bartolino V, Ottavi A, Colloca F, Ardizzone GD, StefÃ¡nsson G (2008b) Bathymetric preferences of juvenile European hake (Merluccius merluccius). *ICES Journal of Marine Science: Journal du Conseil* **65**, 963-969.

Belcari P, Ligas A, Viva C (2006) Age determination and growth of juveniles of the European hake, Merluccius merluccius (L., 1758), in the northern Tyrrhenian Sea (NW Mediterranean). *Fisheries Research* **78**, 211-217.

Biagi F, Cesarini A, Sbrana M, Viva C (1995) Reproductive biology and fecundity of Merluccius merluccius (Linnaeus, 1758) in the Nothern Tyrrhenian Sea. In: *Rapp. Comm. int. Mer Médit.* (ed. CIHEAM), pp. 34-23.

Bouaziz A, Bennoui A, Djabali F, Maurin C (1998) Reproduction du merlu *Merluccius merluccius* (linnaeus, 1758) dans la région de Bou-Ismaîl. . In: *Options Mediterraneennes* (ed. CIHEAM), pp. 109-117.

Bouck AMY, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**, 907-924.

Bozzano A, Sardà F, Rìos J (2005) Vertical distribution and feeding patterns of the juvenile European hake, Merluccius merluccius in the NW Mediterranean. *Fisheries Research* **73**, 29-36.

Bradbury IR, Hubert S, Higgins B*, et al.* (2011) Evaluating SNP ascertainment bias and its impact on population assignment in Atlantic cod, Gadus morhua. *Molecular Ecology Resources* **11**, 218-225.

Brécard D, Hlaimi B, Lucas S, Perraudeau Y, Salladarré F (2009) Determinants of demand for green products: An application to eco-label demand for fish in Europe. *Ecological Economics* **69**, 115-125.

Carlucci R, Giuseppe L, Porzia M*, et al.* (2009) Nursery areas of red mullet (Mullus barbatus), hake (Merluccius merluccius) and deep-water rose shrimp (Parapenaeus longirostris) in the Eastern-Central Mediterranean Sea. *Estuarine, Coastal and Shelf Science* **83**, 529-538.

Carpentieri P, Colloca F, Cardinale M, Belluscio A, Ardizzone GD (2005) Feeding habits of European hake (Merluccius merluccius) in the central Mediterranean Sea. *Fishery Bulletin US* **103**, 411-416.

Casey J, Pereiro J, Alheit J, Pitcher TJ (1994) European hake (M. merluccius) in the North-east Atlantic. In: *Hake: Biology, fisheries and markets* (ed. Noakes DLG), pp. 125-147. Springer Netherlands.

Castillo AGF, Alvarez P, Garcia-Vazquez E (2005) Population structure of Merluccius merluccius along the Iberian Peninsula coast. *ICES Journal of Marine Science: Journal du Conseil* **62**, 1699-1704.

Castillo AGF, Martinez JL, Garcia-Vazquez E (2004) Fine Spatial Structure of Atlantic Hake (Merluccius merluccius) Stocks Revealed by Variation at Microsatellite Loci. *Marine Biotechnology* **6**, 299-306.

Cheilari A, Rätz HJ (2009) Review of possible stock units of European hake, red mullet and deep-water pink shrimp in the Mediterranean Sea by means of trends in survey abundance In: *STECF SG/ECA/RST/MED 09-01*.

Cimmaruta R, Bondanelli P, Nascetti G (2005) Genetic structure and environmental heterogeneity in the European hake (Merluccius merluccius). *Molecular Ecology* **14**, 2577-2591.

Colloca F, Bartolino V, Lasinio G*, et al.* (2009) Identifying fish nurseries using density and persistence measures. *Marine Ecology Progress Series* **381**, 287-296.

Conover DO, Clarke LM, Munch SB, Wagner GN (2006) Spatial and temporal scales of adaptive divergence in marine fishes and the implications for conservation. *Journal of Fish Biology* **69**, 21-47.

Crandall KA, Bininda-Emonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology. *Trends in Ecology & Evolution* **15**, 290-295.

de Pontual Hln, Groison AL, PiÃ±eiro C, Bertignac M (2006) Evidence of underestimation of European hake growth in the Bay of Biscay, and its relationship with bias in the agreed method of age estimation. *ICES Journal of Marine Science: Journal du Conseil* **63**, 1674-1681.

Deagle BE, Jones FC, Chan YF*, et al.* (2011) Population genomics of parallel phenotypic evolution in stickleback across stream-lake ecological transitions. *Proceedings of the Royal Society B: Biological Sciences*.

Domínguez-Petit R, Korta M, Saborido-Rey F*, et al.* (2008) Changes in size at maturity of European hake Atlantic populations in relation with stock structure and environmental regimes. *Journal of Marine Systems* **71**, 260-278.

El Habouz H, Recasens L, Kifani S*, et al.* (2011) Maturity and batch fecundity of the European hake (*Merluccius merluccius*, Linnaeus, 1758) in the eastern central Atlantic

Scientia Marina **75**, 447-454.

European Commission (2011a) Commission Staff Working Paper - Impact Assessment, SEC(2011) 891.

European Commission (2011b) Communication from the Commission concerning a consultation on Fishing Opportunities, COM(2011) 298.

FAO (2009) The state of world fisheries and aquaculture 2008.

FAO (2010) FAO yearbook. Fishery and Aquaculture Statistics. 2008.

Filonzi L, Chiesa S, Vaghi M, Nonnis Marzano F (2010) Molecular barcoding reveals mislabelling of commercial fish products in Italy. *Food Research International* **43**, 1383-1388.

Fiorentino F, Garofalo G, De Santi A*, et al.* (2003) Spatio-temporal distribution of recruits (0 group) of *Merluccius merluccius* and *Phycis blennoides* (Pisces, Gadiformes) in the Strait of Sicily (Central Mediterranean). *Hydrobiologia* **503**, 223-236.

Fraser DJ, Bernatchez L (2001) Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Molecular Ecology* **10**, 2741-2752.

Freamo H, O'Reilly P, Berg PR, Lien S, Boulding EG (2011) Outlier SNPs show more genetic structure between two Bay of Fundy metapopulations of Atlantic salmon than do neutral SNPs. *Molecular Ecology Resources* **11**, 254-267.

Gebremedhin B, Ficetola GF, Naderi S*, et al.* (2009) Frontiers in identifying conservation units: from neutral markers to adaptive genetic variation. *Animal Conservation* **12**, 107-109.

GFCM (2011) GENERAL FISHERIES COMMISSION FOR THE MEDITERRANEAN. Thirty-fifth Session. MANAGEMENT OF MEDITERRANEAN FISHERIES.

Glenn TC (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* **11**, 759-769.

Glover K, Hansen M, Lien S*, et al.* (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genetics* **11**, 2.

Hauser L, Baird M, Hilborn RAY, Seeb LW, Seeb JE (2011) An empirical comparison of SNPs and microsatellites for parentage and kinship assignment in a wild sockeye salmon (Oncorhynchus nerka) population. *Molecular Ecology Resources* **11**, 150-161.

Hauser L, Carvalho GR (2008) Paradigm shifts in marine fisheries genetics: ugly hypotheses slain by beautiful facts. *Fish and Fisheries* **9**, 333-362.

Helyar SJ, Hemmer-Hansen J, Bekkevold D*, et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources* **11**, 123-136.

ICES (2008) Report of the Working Group on the Assessment of Southern Shelf Stocks of Hake, Monk and Megrim (WGHMM).

Jones B (1974) World resources of hakes of the genus Merluccius. In: *Sea fisheries research* (ed. FR H-J), pp. 139-166. Paul Elek Scientific Books, London.

Kacher M, Amara R (2005) Distribution and growth of 0-group European hake in the Bay of Biscay and Celtic Sea: a spatial and inter-annual analyses. *Fisheries Research* **71**, 373-378.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity* **88**, 62-65.

Larsen PF, Nielsen EE, Williams TD*, et al.* (2007) Adaptive differences in gene expression in European flounder (Platichthys flesus). *Molecular Ecology* **16**, 4674-4683.

Limborg MT, Blankenship SM, Young SF*, et al.* (2011) Signatures of natural selection among lineages and habitats in Oncorhynchus mykiss. *Ecology and Evolution* **2**, 1-18.

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**, S26.

Lloris D, Matallanas J, Oliver P (2005) Hakes of the world (Family Merlucciidae). An annotated and illustrated catalogue of hake species known to date. In: *FAO Species Catalogue for Fishery Purposes No. 2*.

Lucio P, Murua H, Santurtun M (2000) Growth and reproduction of hake (Merluccius merluccius) in the Bay of Biscay during the period 1996-1997. *Ozeanografika* **3**, 325-354.

Lundy CJ, Moran P, Rico C, Milner RS, Hewitt GM (1999) Macrogeographical population differentiation in oceanic environments: a case study of European hake (Merluccius merluccius), a commercially important fish. *Molecular Ecology* **8**, 1889-1898.

Lundy CJ, Rico C, Hewitt GM (2000) Temporal and spatial genetic variation in spawning grounds of European hake (Merluccius merluccius) in the Bay of Biscay. *Molecular Ecology* **9**, 2067-2079.

Mahe K, Amara R, Bryckaert T, Kacher M, Brylinski JM (2007) Ontogenetic and spatial variation in the diet of hake (Merluccius merluccius) in the Bay of Biscay and the Celtic Sea. *ICES Journal of Marine Science: Journal du Conseil* **64**, 1210-1219.

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution* **20**, 136-142.

Martinsohn JT (2011) Deterring Illegal Activities in the Fisheries Sector (ed. Union E).

Martinsohn JT, Ogden R (2009) FishPopTraceâ?”Developing SNP-based population genetic assignment methods to investigate illegal fishing. *Forensic Science International: Genetics Supplement Series* **2**, 294-296.

Mattiucci S, Abaunza P, Ramadori L, Nascetti G (2004) Genetic identification of Anisakis larvae in European hake from Atlantic and Mediterranean waters for stock recognition. *Journal of Fish Biology* **65**, 495-510.

Mellon-Duval C, de Pontual H, Métral L, Quemener L (2009) Growth of European hake (Merluccius merluccius) in the Gulf of Lions based on conventional tagging. *ICES Journal of Marine Science: Journal du Conseil* **67**, 62-70.

Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31-46.

Miller D, Jessel A, Mariani S (2011) Seafood mislabelling: comparisons of two western European case studies assist in defining influencing factors, mechanisms and motives. *Fish and Fisheries*, no-no.

Miller DD, Mariani S (2010) Smoke, mirrors, and mislabeled cod: poor transparency in the European seafood industry. *Frontiers in Ecology and the Environment* **8**, 517-521.

Morales-Nin B, Aldebert Y (1997) Growth of juvenile Merluccius merluccius in the Gulf of Lions (NW Mediterranean) based on otolith microstructure and length-frequency analysis. *Fisheries Research* **30**, 77-85.

Morin PA, Luikart G, Wayne RK, the Snp wg (2004) SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**, 208-216.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources* **9**, 66-73.

Morin PA, McCarthy M (2007) Highly accurate SNP genotyping from historical and low-quality samples. *Molecular Ecology Notes* **7**, 937-946.

Murua H, Michael L (2010) The Biology and Fisheries of European Hake, Merluccius merluccius, in the North-East Atlantic. In: *Advances in Marine Biology*, pp. 97-154. Academic Press.

Murua H, Motos L (2006) Reproductive strategy and spawning activity of the European hake Merluccius merluccius (L.) in the Bay of Biscay. *Journal of Fish Biology* **69**, 1288-1303.

Nielsen E, Hemmer-Hansen J, Poulsen N*, et al.* (2009a) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (Gadus morhua). *BMC Evolutionary Biology* **9**, 276.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009b) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology* **18**, 3128-3150.

Ogden R (2008) Fisheries forensics: the use of DNA tools for improving compliance, traceability and enforcement in the fishing industry. *Fish and Fisheries* **9**, 462-472.

Orsi Relini L, Fiorentino F, Zamboni A (1989) Le nurseries del nasello mediterraneo: dove, quando, perché. *Nova Thalassia* **10**, 407-416.

Orsi Relini L, Papaconstantinou C, Jukic-Peladic S*, et al.* (2002) *Distribution of the Mediterranean hake populations (Merluccius merluccius smiridus Rafinesque, 1810) (Ostheichthyes: Gadiformes) based on six years monitoring by trawl-surveys: some implications for management*.

Papaconstantinou C, Stergiou KI (1995) Biology and fisheries of eastern Mediterranean hake (*M. merluccius*). In: *Hake: Biology, Fisheries and Markets* (ed. eds AJPTJ), pp. 149-180. London: Chapman & Hall.

Piñeiro C, Rey J, de Pontual H, Garcìa A (2008) Growth of Northwest Iberian juvenile hake estimated by combining sagittal and transversal otolith microstructure analyses. *Fisheries Research* **93**, 173-178.

Piñeiro C, Saínza M (2003) Age estimation, growth and maturity of the European hake (Merluccius merluccius (Linnaeus, 1758)) from Iberian Atlantic waters. *ICES Journal of Marine Science: Journal du Conseil* **60**, 1086-1102.

Pita A, Pérez M, Cerviño S, Presa P (2011) What can gene flow and recruitment dynamics tell us about connectivity between European hake stocks in the Eastern North Atlantic? *Continental Shelf Research* **31**, 376-387.

Pita A, Presa P, Perez M (2010) Gene flow, multilocus assignment and genetic structuring of the European hake (Merluccius merluccius). *Thalassas* **26**, 129-133.

Poulsen N, Hemmer-Hansen J, Loeschcke V, Carvalho G, Nielsen E (2011) Microgeographical population structure and adaptation in Atlantic cod Gadus morhua: spatio-temporal insights from gene-associated DNA markers. *Marine Ecology Progress Series* **436**, 231-243.

Recasens L, Chiericoni V, Belcari P (2008) Spawning pattern and batch fecundity of the European hake (Merluccius merluccius (Linnaeus, 1758) in the western Mediterranean. *Scientia Marina* **72**, 721-732.

Recasens L, Lombarte A, Morales-Nin B, Tores GJ (1998) Spatiotemporal variation in the population structure of the European hake in the NW Mediterranean. *Journal of Fish Biology* 53, 387-401.

Renaut S, Nolte AW, Bernatchez L (2010) Mining transcriptome sequences towards identifying adaptive single nucleotide polymorphisms in lake whitefish species pairs (Coregonus spp. Salmonidae). Molecular Ecology 19, 115-131.

Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L (2011) SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (Coregonus spp.). Molecular Ecology 20, 545-559.

Roldan MI, Garcia-Marin J, Utter FM, Pla C (1998) Population genetic structure of European hake, Merluccius merluccius. Heredity 81, 327-334.

Seeb JE, Carvalho G, Hauser L, et al. (2011a) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Molecular Ecology Resources 11, 1-8.

Seeb LW, Templin WD, Sato S, et al. (2011b) Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Molecular Ecology Resources 11, 195-217.

Smith CT, Seeb LW (2008) Number of Alleles as a Predictor of the Relative Assignment Accuracy of Short Tandem Repeat (STR) and Single-Nucleotide-Polymorphism (SNP) Baselines for Chum Salmon. Transactions of the American Fisheries Society 137, 751-762.

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100, 158-170.

Swan SC, Geffen AJ, Morales-Nin B, et al. (2006) Otolith chemistry: an aid to stock separation of Helicolenus dactylopterus (bluemouth) and Merluccius merluccius (European hake) in the Northeast Atlantic and Mediterranean. ICES Journal of Marine Science: Journal du Conseil 63, 504-513.

Tserpes G, Politou C-Y, Peristeraki P, Kallianiotis A, Papaconstantinou C (2008) Identification of hake distribution pattern and nursery grounds in the Hellenic seas by means of generalized additive models. Hydrobiologia 612, 125-133.

Wayne RK, Morin PA (2004) Conservation genetics in the new molecular age. Frontiers in Ecology and the Environment 2, 89-97.