Chapter 8

Population or Point-of-Origin Identification

Einar Eg Nielsen
Technical University of Denmark, Silkeborg, Denmark

WHY POPULATION OR POINT-OF-ORIGIN FOR SEAFOOD AUTHENTICITY AND TRACEABILITY

Population or point-of-origin identification represents the intermediate step in a continuum of DNA-based seafood authenticity and traceability applications, ranging from documenting the species to identifying the specific individual present in a product. While DNA methods have gained wide acceptance and application for species identification, population or point-of-origin assignment has received less attention and thus found fewer practical applications. The main reason for this is that these analyses require more background genetic data, more advanced statistical analysis, and a broader insight into the evolution and biology of the focal organisms to interpret the analytical output. However, the obstacles for wider implementation can now be overcome more easily, and point-of-origin identification is beginning to be implemented.

There are many good reasons for engaging in DNA-based population/point-of-origin identification for seafood. First, most seafood resource management is based on a system of spatially defined species-specific "stocks," outlined by international organization, such as FAO, ICES, and NAFO, and supported by fisheries legislation. In order to assure sustainability, the resource is assessed within each area and fishing quotas are given based on stock status. Illegal, unreported, and unregulated (IUU) fishing poses a significant threat to good management through local or regional overfishing and depletion of seafood resources. Thus there is a need for tools that can identify and document that the raw material entering the seafood production chain originates from sustainable fisheries. Secondly, many seafood products can originate from both wild capture fisheries and aquaculture with associated product differences with respect to quality, environmental impact, and animal welfare. Thirdly, seafood from different regions may vary with respect to classical measures of quality as, for example, taste, fat content, and color (Børresen, 1992), but also health aspects such as content of essential fatty acids, heavy

Seafood Authenticity and Traceability. http://dx.doi.org/10.1016/B978-0-12-801592-6.00008-5

metals, and frequency of disease-carrying agents. Collectively, quality measures for seafood products from a specific region, termed "terroir," may be labeled and branded to obtain a higher market price. Finally, there may be personal reasons, including political, why consumers would care about the origin of seafood products. For example, certain consumers would preferentially buy local products to support regional food production or to minimize CO_2 footprints of transporting seafood. To enable consumer choice with respect to these factors, seafood must be accurately identified and labeled.

Based on the considerations discussed earlier, the need for information on the origin of seafood is widely recognized and reflected in international laws. For example, in the European Union catch certificates that state the origin of all traded fish and fish products are required through the European Commission Control and IUU Regulations (EC, 2008, 2009). In addition to legal requirements, voluntary ecolabels, such as the MSC (Marine Stewardship Council, www.msc.org), for seafood products from capture fisheries have emerged, where one of the major pillars regarding certification of fisheries is sustainable fish stocks. A parallel system, ASC (Aquaculture Stewardship Council, www.asc.org) has emerged for aquaculture products, which among other aspects, takes biodiversity, sustainability and both fish and consumer health issues into account. Traditionally, compliance with rules and regulations, as well as voluntary ecolabeling, has to a very large extent relied on a paper trail documenting the origin of the product throughout the seafood supply chain. However, evidence from incidences in other parts of the food sector, for example, through the European BSE scandal has raised the awareness that there is a definite need for independent methods that are not easily fabricated that can validate the paper based traceability scheme.

A number of different methods have been suggested and applied independently or in concert (Higgins et al., 2010) for tracing the origin of seafood. These include morphometrics, meristics, micro/macroparasites, chemical composition, and analysis of fatty acids (Cadrin et al., 2014). Despite some success for inferring the origin of seafood, these methods are often hampered by limited availability of tissue of sufficient quality. This is particularly problematic for the analysis of processed seafood where most of these methods cannot be applied. Furthermore, calibration and standardization among laboratories necessary for forensic purposes (Ogden, 2008; Ogden and Linacre, 2015) is inherently difficult. For example, the establishment of long-term baseline datasets, to which any new specimen or products can be compared, is notoriously difficult. Finally, the statistical power for origin assignment associated with these methods individually is generally low. Thus concerted application of many methods requiring diverse expertise and equipment is beyond the capacity of most seafood laboratories and rarely applicable in a close to real-time framework required for practical seafood identification on a commercial scale, for example, for production line testing.

In contrast to biological- and chemical-based methods for seafood traceability, genetic methods offer many advantages. First, DNA is found in almost all

cells in all organisms and can be retrieved from degraded or processed material, that is, DNA analysis can in principle be conducted at all stages in the production chain from sampling fresh fish onboard vessels to a filet on a dinner plate in a restaurant. Likewise, DNA-based origin assignment relies on a well-established theoretical framework from population and evolutionary genetics, allowing comparisons of new samples to already established genetic databases and at the same time providing a robust statistical framework for evaluating the result, which is essential for forensic purposes. In this context, the calibration of DNA results across laboratories is much simpler than for other methods, allowing global genetic information for various species to be compiled and used across laboratories without the need first to establish a new genetic baseline database. This, in turn, allows swift processing of new samples in a more real-time framework, as only the specific new "case" samples have to be processed before inferences on their geographical origin can be provided. Naturally, there are also limitations to the use of DNA analysis for origin assignment. In contrast to species designation, where mtDNA sequence information is available for all commercially valuable seafood species, databases of genetic information on populations across the species distribution are either incomplete or completely missing for many species. Furthermore, management areas and genetic populations are often not aligned (see Reiss et al., 2009). Management areas may include several genetic populations, or complicating matters further, the same population may be found in different management areas. To a large extent, this reflects that management areas have been defined on a political rather than a biological background. Still, the problems of mismatch between population and management area can be remedied in many cases. This is elaborated further in the following section.

Overall there continues to be a need for population/origin assignment in the seafood industry, and DNA-based methods represent the most promising means for practical use today. There are many very good reasons for using DNA-based methods for population/origin assignment. This chapter provides (1) an introduction to population genetics of marine organisms, (2) a review of the basic principles of population/origin identification, (3) a description of the various methods applied, (4) case studies of population and origin assignment, (5) a summary of caveats and potential pitfalls, and finally (6) a review of the ongoing and expected future developments within the field.

POPULATION GENETICS OF MARINE ORGANISMS

What Is a Genetic Population?

Genetic point-of-origin identification is based on assigning fish back to their genetic or evolutionary population. Under an evolutionary paradigm, a population can be defined as "A group of individuals of the same species living in close enough proximity that any member of the group can potentially mate with any other member" (Waples and Gaggiotti, 2006). That definition is distinct from an ecological paradigm where a population is defined as individuals of

the same species that cooccur in an area and potentially interact. Thus the distinction is related to reproduction and a genetic population should therefore be reproductively isolated to some degree from any other genetic population of the species. The term "to some degree" is, however, quite vague and hardly operational. A more quantitative and practical definition of when groups of individuals are different enough to be considered populations is based on the exchange of effective migrants between populations per generation (N_em, see later). If this number is above 25, populations become very difficult to distinguish using standard genetic tools (Waples and Gaggiotti, 2006). So, when populations are practically genetically indistinguishable they are, in this framework, defined as a single population. This distinction between theoretical delimitations of populations to a more application-based definition is important in applying DNA-based identification of populations to seafood.

Evolutionary Forces and Genetic Population Structure

Following evolutionary theory, individuals within reproductively isolated populations are subject to the same evolutionary forces that determine their genetic composition. These are: mutation, migration, random genetic drift, and selection. In popular terms, mutation is the long-term process that generates the genetic raw material in the form of new genetic variants, "alleles" at any gene locus (position of the DNA sequence in question in the genome). Migration of individuals carries those alleles among populations, and if migrants are successfully interbreeding, spread them through the process of "gene flow". Random genetic drift is the sampling error associated with breeding; that is, if "effectively" few individuals (where N_e is defined as the effective population size) participate in mating, then allele frequencies in the population will change fast and ultimately lead to the loss of allelic variants. Finally, an individual carrying specifically favorable alleles may be at an advantage related to survival and reproduction (fitness), mediated through natural or sexual selection. Thus differential selection pressure among populations can lead to fast changes and large differences in allele frequency between populations. On the relatively short evolutionary timescale often associated with population processes, migration, drift, and selection are the most important processes. The relative impact of the different evolutionary drivers ultimately determines the genetic composition of populations and the genetic differentiation among them. Thus, migration tends to homogenize allele frequencies among populations, while random genetic drift and differential selection acts to differentiate them. As a rule of thumb, small, isolated populations subject to special environmental conditions tend to show the largest genetic differentiation and are therefore most easily distinguished using genetic tools. Genetic differentiation due to population structure is traditionally measured using a fixation index, "F_{ST}" (Wright, 1950; Weir and Cockerham, 1984). The index ranges from zero to one, where zero denotes no differentiation and one represents fixation of different alleles among

populations. As a measure of scale, F_{ST}s among humans on different continents ranges between 0.1 and 0.15 (Jorde and Wooding, 2004).

Types of Population Structure

A prerequisite in order to be able to use genetics to determine the population of origin of seafood is that the marine organism in question display some sort of genetic structuring of populations. In general, there are many evolutionary and ecological models for population structuring of marine organisms, which are beyond the scope of this chapter. However, three crude categories of significant importance for the identification of origin can be recognized (Laikre et al., 2005; Fig. 2.1). First, there is **no genetic differentiation** (panmixia) across the geographical regions of interest, that is, that migration and associated gene flow is sufficient to homogenize populations. This means that genetic tools cannot be used for origin identification as the different regions of the species distribution display non-distinguishable genetic compositions. This may, naturally, be an inherent characteristic of the species in question; however, it may also be an artifact of the sampling strategy and/or the genetic and analytical tools applied (for more details see the following sections). Another type is **continuous genetic change**, that is, allele frequencies shift gradually along a geographical or environmental transect. Accordingly, the genetic compositions at each end of the species distribution are highly genetically differentiated, while intermediate locations display minute and gradual genetic changes. This kind of population structure imposes some problems in relation to determination of origin, as the statistical power associated with referring individuals to specific sites, as opposed to adjacent locations, is expected to be relatively weak. In addition, a significant sampling and genetic typing effort has to be undertaken in order to be able to describe the genetic shape of this "isolation by distance," that is, to establish whether the continuous change is homogenous across the whole distributional area. The final major type is **distinct populations**, where migration among populations is sufficiently small to allow the buildup of distinct genetic differences. This type of population structure not only represents the ideal setting for population-based management and conservation, it also represents the optimal structure for population/origin assignment. As all populations in this scenario are geographically defined and genetically distinct, the population of origin of individuals can be inferred with high probability, dependent on the levels of genetic differentiation among populations. However, as the genetic population represents the reproductive unit, different populations may have distributional areas that significantly differ and overlap outside spawning time. In the latter case, the genetically determined population of origin of an individual fish may provide little information on the geographical origin of the sample. Still the mixture composition, using information from a sufficiently large sample of individuals in concert, may be able to provide insights into the geographical origin. This issue is treated in more detail in a subsequent section.

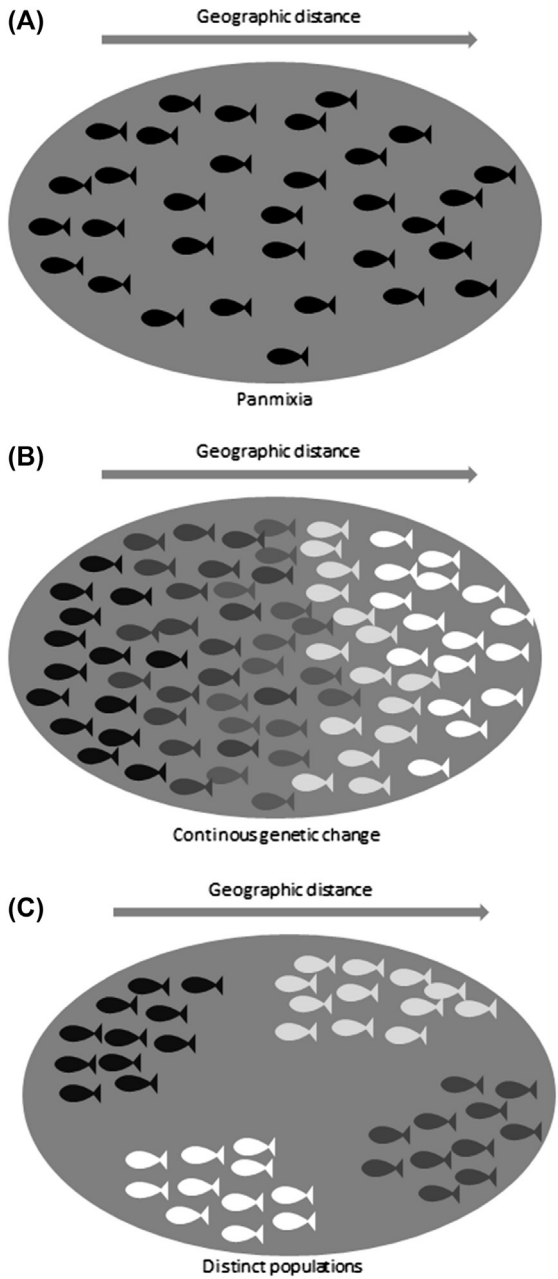

FIGURE 2.1 Three types of population structure for marine organisms (A) no genetic differentiation (B) isolation (C) distinct populations (see text for explanation).

Population Structure of Marine Organisms

The population structure and level of genetic differentiation for important commercial species is of paramount importance for successful origin determination, and subsequently for improved stock management. The level of genetic differentiation among populations (F_{ST}) is typically much lower for marine organisms than for freshwater and anadromous species (Fig. 2.2A, redrawn from Ward et al., 1994). Likewise, it is evident (Fig. 2.2B) that the vast majority of marine fish species display very low levels of genetic differentiation among populations ($F_{ST}<0.03$). The reason for the relatively low levels of genetic differentiation for

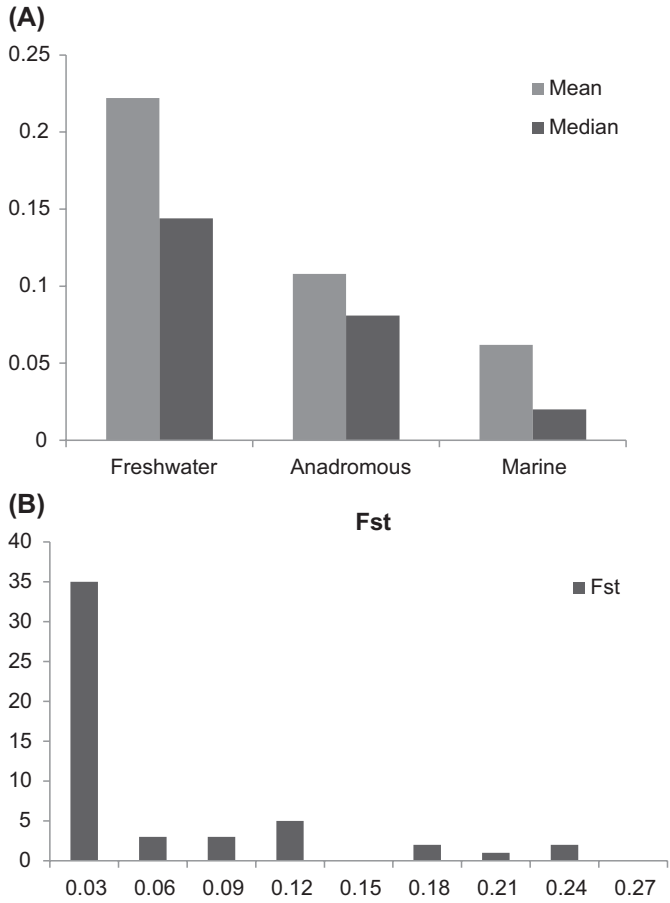

FIGURE 2.2 Levels of genetic differentiation among populations of marine fish. (A) Comparison of genetic differentiation (F_{ST}) among freshwater, anadromous and marine fish. (B) Distribution of F_{ST} values in marine fish. *Redrawn from Ward, R.D., Woodwark, M., Skibinski, D.O.F., 1994. A comparison of genetic diversity levels in marine, freshwater and anadromous fishes. Journal of Fish Biology 44, 213–232.*

"classical marine organisms," including many of our most important commercial species such as clupeoids, gadoids, and scombrids (Nielsen and Kenchington, 2001) relates to a number of inherent characteristics of these species. First, the number of obvious physical barriers in the sea is not so pronounced as in freshwater. For example, while highly mobile marine fish can freely migrate vast distances in the oceans, fish living in a lake are restricted to this particular water body. Likewise, many marine organisms have pelagic eggs and larvae, which can be spread over vast areas by ocean currents before settling. Finally, most marine species have comparatively large (effective) population sizes (Hare et al., 2011) resulting in minute levels of random genetic drift and related low levels of genetic differentiation. However, although it may seem that the oceans are devoid of any physical boundaries, this is not the case. The major oceans are separated by large landmasses, restricting gene flow on a large geographical scale. In addition, factors, such as bathymetry and ocean currents, may serve as barriers to active migration of adult specimens, or act to retain eggs and larvae, so that the juveniles will settle in proximity to the parental population (e.g., see Sinclair and Power, 2015). It has been identified that environmental differences may also restrict migration among populations (Limborg et al., 2009). Thus differences in temperature, salinity, and other environmental factors may define the boundaries between populations. Habitat preference and life history may also restrict gene flow geographically. This phenomenon also sets the scene for the identification of genes subject to differential selection in populations inhabiting different environments. This selection can create vast differences in allele frequencies even in the face of relatively high levels of gene flow, which renders the application of these genes particularly interesting for origin determination. This will be treated in detail in subsequent sections. In conclusion, marine organisms display relatively low levels of genetic differentiation among populations, which in association with the lack of obvious physical boundaries among populations poses a number of challenges for origin determination of seafood products.

PRINCIPLES OF POPULATION ASSIGNMENT

As previously stated, the origin of an individual is equivalent to its genetic population. Thus first, the population of origin has to be determined, and subsequently, this has to be matched with the known or suspected geographical distribution of the population. In order to make this comparison, background knowledge of the populations, their distribution, and their biology must be empirically derived. This may require biological study of an organism and its life history, and/or genotyping of hundreds to thousands of different individuals across a suspected range to identify genetic populations. Where this information is available, population-level assignments can be made. Most often the population of origin of individuals is determined using a method called "individual assignment" (Paetkau et al., 1995; Rannala and Mountain, 1997). In contrast to genetically based species designation (e.g., DNA barcoding), which is

categorical with fixed differences in DNA sequences among species, individual assignment (IA) is probabilistic, exploiting differences in allele frequencies among populations. In essence, the method calculates the probability of observing a given multilocus genotype based on the different allele frequencies in a set of reference populations. It is rarely the case that a single genetic marker is sufficient to provide high statistical power for unambiguous assignment of individuals to population. Instead the method relies on combining allele frequency information from a number of genetic markers, thereby increasing the statistical power for inferring population of origin. If significant genetic differences are found among populations, then in theory, any level of statistical certainty should be attainable, by applying more markers. However, in practice, this is limited by the genotyping error rate, cost, and time constraints.

Population assignment is composed of a number of predefined steps allowing rigorous assessment of the most likely population of origin of individuals and the statistical certainty associated with it. First, a set of baseline genetic data has to be retrieved from potential populations of origin (see Fig. 2.3). Typically,

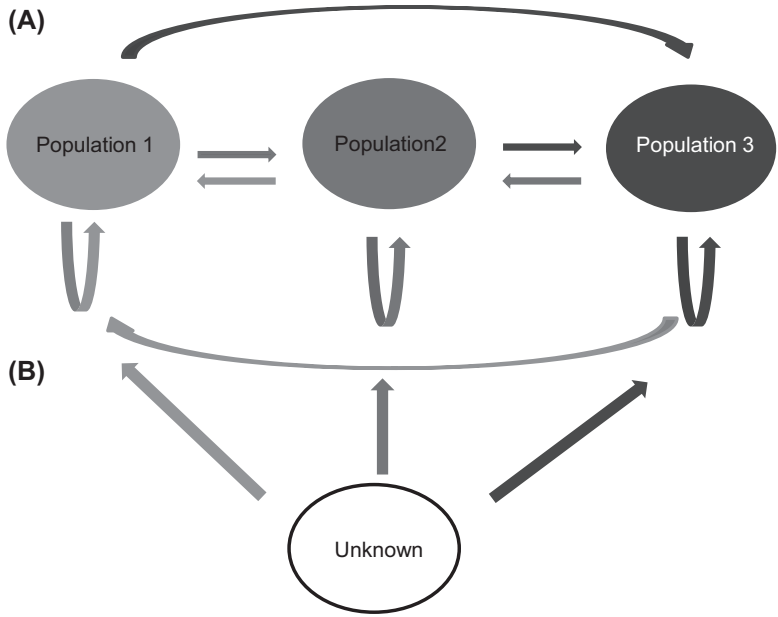

FIGURE 2.3 Principle of individual assignment. (A) "Self-assignment". Likelihoods of observing multilocus genotypes are calculated for all individuals within baseline samples and assigned back to the sample (population) where they have the highest likelihood of occurring. (B) "Assignment of unknown individuals". The likelihood of observing the genotype of an individual of unknown origin is calculated for each of the baseline samples and the individual is assigned to the sample (population) where it has the highest likelihood of occurring (see text for further explanation). *Redrawn from Hansen, M.M., Kenchington, E., Nielsen, E.E., 2001. Assigning individual fish to populations using microsatellite DNA markers: methods and applications. Fish and Fisheries 2, 93–112.*

the markers applied have been single nucleotide polymorphisms (SNPs) or microsatellites (for more information on markers see next section). The second step is to do "self-assignment" to evaluate the statistical power of assignment. For all baseline individuals, the likelihoods of observing their multilocus (across all markers) genotypes in each of the baseline populations are calculated. To avoid biasing allele frequencies, the baseline individual being assigned is commonly excluded in the calculation of allele frequencies in a procedure called "leave one out" (Efron, 1983). The individual is then assigned to the population sample where its genotype has the highest likelihood of occurring based on sample allele frequencies. If the populations are sufficiently genetically distinct, we should expect that most or all baseline individuals are assigned back to their known population of origin. However, the IA procedure entails that all individuals are assigned to a baseline sample no matter how small their likelihoods are. For example, if the population of origin is not included among baseline samples, we may (erroneously) assign the individual to another population included in the baseline. Likewise, individuals may have similar likelihoods in two or more populations, rendering it difficult to state the population of origin with high certainty. So how do we assign a statistical probability to the assignment test? To evaluate the problem of missing baselines Cornuet et al. (1999) devised a method of simulating a large number (>1000) of individuals from each population based on sample allele frequencies to generate a distribution of expected likelihoods for true individuals within the population. The likelihood of each sampled individual is then compared to the simulated distribution of likelihoods and the individual is accepted/rejected from the population if its likelihood is above or below a certain threshold of the distribution (e.g., 0.05 or 0.01). To statistically evaluate, the relative likelihoods of potential alternative origins for a given genotype a number of options are available (Piry et al., 2004). One method is to estimate the relative likelihood scores by dividing the estimated likelihood of observing the genotype in each population by the total likelihood for all populations. Again here a threshold value (e.g., 90% or 95%) can be applied to designate a level above which the assigned population is accepted/rejected as the true population of origin. Alternatively, likelihood ratios between pairs of populations can be calculated, which is often the preferred option for forensic purposes (Ogden, 2008; Ogden and Linacre, 2015). The ideal situation is when the distributions of likelihood ratios from different populations do not overlap and are clearly different from zero (see Fig. 2.4), that is, that the genetic differentiation between populations and the number of markers is sufficient to allow unambiguous assignment. After the assignment power has been assessed, individuals of truly unknown origin can be statistically assigned to the most likely source population among baselines. Again the likelihood of observing each of the "unknown" individual genotypes is calculated for all populations and the individual is assigned to the population where it has the highest likelihood of occurring. Likewise, the statistical evaluation of whether the individual actually could belong to the population where it is assigned and alternative populations

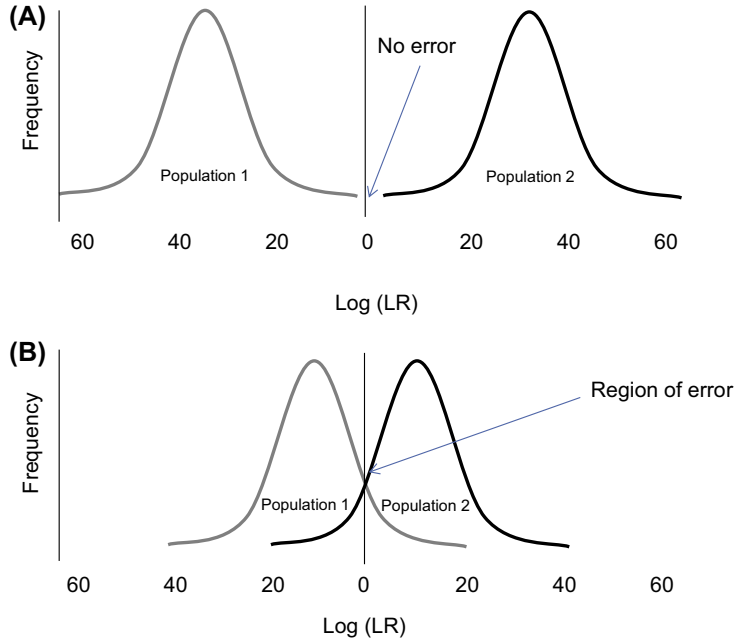

FIGURE 2.4 Distribution of log-likelihood ratios (log LR) for individuals from two different populations where (A) populations are well differentiated and the number of markers are sufficiently high or (B) where genetic differentiation is low and or the number of markers is insufficient for allowing unambiguous assignment. *Redrawn from Ogden, R., Linacre, A., 2015. Wildlife forensic science: a review of genetic geographic origin assignment. Forensic Science International-Genetics 18, 152–159.*

of origin is performed as for the self-assignment of baseline samples. Baseline reference data should be formed from large sample sizes to have appropriate power to discriminate populations with high levels of certainty.

POPULATION OR POINT-OF-ORIGIN IDENTIFICATION IN PRACTICE

Before applying IA methods, there are a number of issues to consider for practical implementation. Of particular concern is the speed and reproducibility of genotyping as well as obtaining sufficiently high statistical power for inferring origin.

Genetic Markers

Mitochondrial DNA is widely used for species identification but only rarely used for origin assignment unless the potential populations of origin are genetically very distinct. MtDNA is (generally) maternally inherited

without recombination, so the whole genome is linked and acts effectively as a single genetic marker. This is not optimal for IA (but see Marko et al., 2011 for an example), where the high statistical power of origin assignment relies on the combination of information across multiple genetic markers. Accordingly, IA typically applies a number of nuclear genetic markers, with microsatellites and single-nucleotide polymorphisms (SNPs) as the preferred options. Since the development of assignment tests 20 years ago, most IA studies have relied on the application of microsatellites. Microsatellites consist of tandemly repeated DNA sequence motifs (2–5 base pairs), commonly found in non-coding regions of the genome and often have a high number of alleles per locus (5–50) (Putman and Carbone, 2014). The high number of alleles provides high information content per locus (see section on assignment power below), making them well suited for studies where a relatively modest number of markers (5–15) can be genotyped. However, there are drawbacks of using microsatellites. As their genotyping typically relies on the relative electrophoretic migration of PCR fragments (alleles) of different sizes, they are prone to genotyping errors and are notoriously difficult to calibrate across laboratories (Ellis et al., 2011). In contrast, SNPs are biallelic markers found in all organisms in both coding and noncoding parts of genomes. Many high-throughput methods are available for SNP genotyping, so a relatively higher number of markers can be genotyped for SNPs compared to microsatellites, compensating for the smaller number of segregating alleles. Another major advantage is that SNP genotyping is platform independent, that is, the genotypes identified in different laboratories can be readily compared. Accordingly, SNPs are currently gaining much wider application for IA. A final note of relevance for all genetic markers is that preferably markers imbedded in short DNA fragments should be applied when using low-quality templates, for example from seafood samples that have been degraded due to processing. As a general rule of thumb segments spanning, more than 200 bps have proven difficult to PCR amplify in low-quality historical samples (e.g., see Nielsen and Hansen, 2008). Thus this lends further support to SNPs as the markers of choice as they allow design of marker segments of smaller size.

Assignment Power

The power of assignment tests is determined by a number of factors: (1) the level of genetic differentiation among the sampled populations, (2) the number and polymorphism of genetic markers applied, and (3) the number of sampled populations and individuals (Cornuet et al., 1999; Hansen et al., 2001; Manel et al., 2005). The level of genetic differentiation among populations is obviously very important for assignment power. Simulation studies (Cornuet et al., 1999; Manel et al., 2002) have shown that with an F_{ST} of 0.1 almost 100% correct assignment can be achieved with a relatively limited number of genetic markers

(30 individuals, 10 loci). In contrast, when F_{ST} is low (i.e., 0.01), as for marine fish populations, many more loci have to be used to achieve high assignment power. In the early 2010s, a simulation program "SPOTG" has been developed (Hoban et al., 2013) for choosing the appropriate number of loci and individuals to achieve a desired level of assignment power. By using user generated input on allele frequency distributions and levels of genetic differentiation, the program reports mean and standard deviation on mis-assignment, incorrect inclusion and incorrect exclusion. This approach is highly recommended to design studies using assignment tests.

High Grading of Genetic Markers

IA has traditionally relied on using a limited number of genetic markers only influenced by "neutral" evolutionary forces (random genetic drift and gene flow). However, the advances in genetic sequencing via "next-generation sequencing" is rapidly changing the number of genetic markers available for IA in a given species (Helyar et al., 2011). It is now possible to high-grade assignment panels by choosing the specific genetic markers that show the largest divergence among populations (Bromaghin, 2008) and thus create "minimum marker panels with maximum power" for IA (Nielsen et al., 2012a and below for examples for marine fish). Many of the high divergence markers, will be gene loci directly, or indirectly influenced by selection (i.e., situated in genomic regions, where the different alleles have an influence on the fitness of the individual or "hitchhiking" through linkage with variants under selection). The high level of differentiation is expected to occur due to the process of differential selection imposed by differences in the environment experienced by different populations (Nielsen et al., 2009). Thus selection increases allele frequency differences among populations at these marker loci compared to neutral loci, and as a consequence, they will provide higher power for IA. This means that fewer markers are needed to obtain similar precision, thereby reducing time and costs associated with IA. However, there are also some potential pitfalls associated with high grading (see Anderson, 2010). Differences in population divergence among markers in the baseline (training) samples may simply be caused by the process of sampling baseline individuals. Thus, some allele frequencies may show large differences among baseline samples just by chance and if those markers are deliberately cherry-picked assignment power can be seriously overestimated. Consequently, another assignment procedure is required to evaluate assignment precision. First, each of the baseline samples is split in two (Fig. 2.5), where half of the individuals are used as a baseline or "training" samples and the other half is treated as if the origin was unknown, termed "holdout" samples (Anderson, 2010). The rationale for using these samples is to provide an unbiased estimate of assignment power. In other words to use a sample of individuals of known origin, which has not been used to estimate

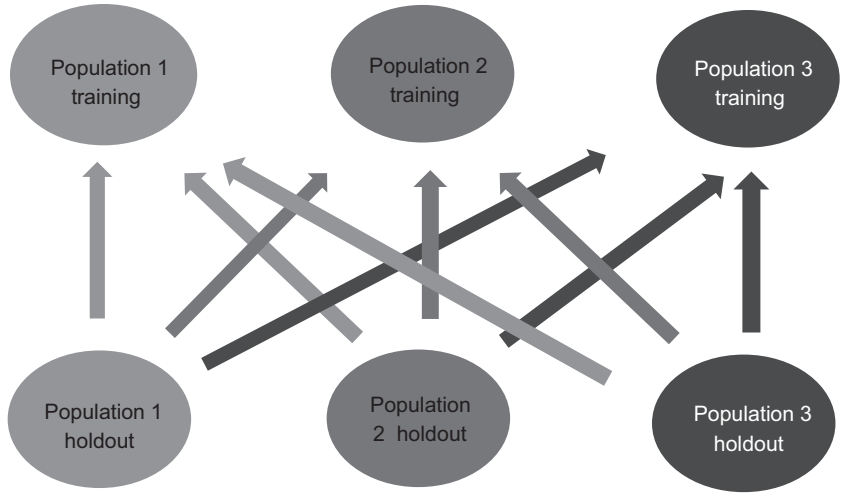

FIGURE 2.5 Procedure for evaluating statistical power when using high grading of loci for IA. Baseline samples are split into "training" and "holdout" samples, where the training sample is used as a baseline for defining population allele frequencies, while the holdout sample is used for evaluation of the statistical power of assignment (see text for more explanation).

allele frequencies within populations, thus resembling a scenario of assigning individuals of truly unknown origin back to baseline (training) samples.

Software for Origin Assignment

A number of statistical analysis methods are available for IA and have been implemented in different software programs. The original frequency-based method developed by Paetkau et al. (1995) has generally been replaced by partly or full probabilistic Bayesian methods (Pritchard, 2000; Piry et al., 2004). The two most commonly applied tools are "GeneClass" (Piry et al., 2004) and "STRUCTURE" (Pritchard, 2000). GeneClass uses a Bayesian approach to estimate baseline allele frequencies, by assuming equal prior probabilities of occurrence of alleles at each locus in each population. This is done to account for potentially missing rare alleles within populations, which has not been detected in individuals sampled for the baseline. The likelihoods of observing a given set of multilocus genotypes in each of the baseline populations and accepting/rejecting that genotype from each of the populations is calculated according to the description in the section on "Principles of population assignment" earlier. The pure Bayesian method implemented in STRUCTURE (Pritchard, 2000), builds on a completely different principle. This method clusters individuals to minimize Hardy–Weinberg and linkage disequilibria within clusters. The rationale behind the model is that it assumes that there is random mating within populations, and therefore, all loci are expected to be in HW and linkage equilibrium. Individuals

are then assigned probabilistically back to a population. Multiple populations may be assigned if their genotypes suggest that they are admixed (i.e., represent hybridization between individuals from different populations). In general, the latter method does not perform particularly well for species with low levels of population structuring (Kalinowski, 2011), thus for the assignment of classical marine organisms on an intermediate to local scale STRUCTURE has limited application. However, a program using discriminant analysis of principal components "DAPC" has been developed (Jombart et al., 2010) to identify the best supported number of groupings assumed to represent populations. The method does not rely on a specific genetic model but generates synthetic variables (principal components), using linear combinations of the original variables (alleles) and seeks the variables that maximize differences between groups (discriminant functions). Based on these, DAPC provides membership probabilities of each individual for the different groups. As for any principal component analysis (PCA), retaining too many principal components may lead to overfitting of the data. However, the program includes a procedure for avoiding this. The three described programs represent the most commonly applied, but fundamentally different, approaches for genetic origin assignment. When performing practical IA, it is advisable to test at least a couple of methods to assess the robustness of the assignment. In particular, for marine organisms with shallow population structure, the different assumptions of the approaches may influence the outcome, in particular for cases when the statistical power of IA is low (e.g., few genetic markers, small baseline samples).

CASE STUDIES

Fishing Competition for Atlantic Salmon in Finland

A classical and one of the first examples of the use of origin identification for "seafood" relates to a fishing competition in the Finnish Lake Saimaa in June 1999 (Primmer et al., 2000). In a local fishing competition, one of the participants presented a 5.5 kg salmon to the judges. This salmon was unexpectedly large compared to normal Lake Saimaa salmon. The judges suspected that the salmon could have been purchased or caught elsewhere and submitted tissue samples for genetic analysis. Based on baseline microsatellite genetic data (7 loci) from Lake Saimaa salmon, the authors used the simulation approach (10,000 individuals) in GeneClass to generate an expected distribution of likelihoods for that population (see section on principles of population assignment). They found that the probability of the large salmon belonging to the Lake Saimaa population was very low ($p < .0001$). In contrast, the likelihood of originating from one of the regions in Finland that supply most fish markets was more than 600 times higher. When confronted with the evidence the angler confessed that he had purchased the fish at a local fish shop and criminal charges were laid. The case demonstrates that even if it is not possible to include all potential populations of

origin for the baseline data, the exclusion approach can still provide statistically robust tests of specific hypothesis about the point of origin of individual fish.

North Sea or Baltic Sea Cod Sold in Sweden

In 2003, journalists from the Swedish TV4 decided to investigate whether Cod illegally caught in the Baltic Sea were sold under false "North Sea" labels of geographical origin by Swedish fishmongers. The journalists had estimated that fish retailers overall could make more than €500.000 extra annually through deliberate mislabeling, illustrating the large potential for illegal economic revenue. In order to investigate the case, the journalists visited a number of fish shops in Sweden and bought two cod filets labeled as North Sea in origin in each shop. If possible they revisited the shops, resulting 42 samples in total. The claimed origin was documented through oral confirmation by the shop assistants on candid camera. The collected cod samples were subsequently tested against baseline data from North Sea and Baltic Sea Cod (Nielsen et al., 2012b) using 10 microsatellite loci. Genetic differentiation between Baltic and North Sea Cod is relatively high with an F_{ST} of 0.045 (Nielsen et al., 2003). Accordingly, assignment power was high with more than 90% of the Baltic Sea baseline samples correctly assigned. Of the 42 Cod filet samples, 20 were assigned to the North Sea, 17 to the Baltic Sea and the remaining five samples were assigned (based on simulations) neither to the North Sea nor Baltic populations though all 42 had been labeled as North Sea Cod. More than half of the filets were apparently mislabeled, and most of these were likely illegally caught Baltic cod. Interestingly, the pattern of mislabeling was not random. While some shops visited in this study always sold correctly labeled products others always sold mislabeled cod. When confronted with the evidence all shop managers admitted that mislabeling was indeed possible, but claimed that the mislabeling occurred from the wholesalers. When the journalists contacted these companies, none of them agreed to give an interview. When published this investigation created an uproar among consumers in Sweden and neighboring countries as the fish wholesale companies allegedly supplying the mislabeled fish were international.

Using SNPs Under Selection for Origin Identification in Classical Marine Fish

Within the European Union, all traded seafood products require catch certificates stating the origin of catch. Likewise, many products are "eco-labeled" from organizations such as the MSC (Marine Stewardship Council), stating that the fishery from where the products originate is sustainable. Accordingly, there is a need for independent methods to identify the population of origin for commercial fish. However, as genetic differentiation in classical marine fish is commonly relatively low, it can be rather difficult to attain sufficient statistical power for unambiguous origin determination through individual assignment.

The EU supported project FishPopTrace set out to develop high-power assignment tools for four important commercial fish species in European waters: Atlantic Cod, Atlantic Herring, European Sole, and European Hake (Nielsen et al., 2012a). They first used next-generation sequencing to identify SNPs distributed across the genomes for all four species. Subsequently, 1536 (Cod), 281 (Herring), 427 (Sole), and 395 (Hake) SNPs were genotyped in individuals across the species' distributions. SNP's with particularly high F_{ST} were identified as being subject to differential selection and used for designing minimum panels with maximum power for IA, by the process known as "high grading" (see previous section). This was done to allow for rapid processing of a high number of samples within a forensic context. Four species specific scenarios were investigated. For cod, there is a need for methods that can discriminate between North Sea, Barents Sea, and Baltic Sea populations, as their health status varies considerably. Using only eight SNP's with the highest levels of genetic differentiation among populations (F_{ST} between 0.07 and 0.51) correctly assigned all fish back to population of origin. Overall 95% of the individuals had likelihoods that were 1500 times higher for the correct population of origin, thus providing robust results suitable for use in legal proceedings. For herring, no method was found that could distinguish North Sea from Northeast Atlantic Herring, which is important to MSC for certifying fisheries. The 32 highest ranking SNPs (F_{ST} between 0.01 and 0.19) could correctly assign origin for 100% of the Northeast Atlantic and 98% of the North Sea Herring. The true population of origin was always more than three times as likely (maximum seven million times more likely), while the median value was 16,800 times more likely. For Sole, landings in Belgian ports are claimed to originate from the Irish Sea/Celtic Sea. However, they may in fact be caught close to the Belgian coast, which is closer to the market, but where fishing is prohibited to allow rebuilding of the local population. An assay of 50 SNPs with the highest F_{ST} values (between 0.005 and 0.054) correctly assigned 93% to area of origin. On average, individuals were more than 60 times more likely in the population of origin, demonstrating the high power of the method even across a very restricted geographical scale. Finally, for Hake, fishing regulations differ between the Mediterranean and the Atlantic with different legal sizes allowed in the two regions. Thus undersized Atlantic Hake are often misreported as being of Mediterranean origin. Thirteen high F_{ST} SNPs (F_{ST} between 0.08 and 0.29) provided 99% correct assignment 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. Overall, this case demonstrates that the combination of next generation sequencing, SNP development and the application of high grading of markers under differential selection is a very powerful method for developing high-powered IA assays (see also Helyar et al., 2011). However, this represents the results of a large-scale research program that requires substantial financial resources

to undertake. Continued development of these assays requires investment in vessel time, research, and analytical costs for determining baseline reference populations for comparisons.

BIOLOGICAL LIMITATIONS AND POTENTIAL PITFALLS OF POINT-OF-ORIGIN IDENTIFICATION

Genetically based individual assignment supports determination of geographical or population of origin of seafood products. The field is developing rapidly in terms of both the type and the number of the genetic markers applied and also with respect to the statistical methods available. In concert, these developments allow origin determination with increasing geographical resolution and precision. However, as mentioned in the introduction to this chapter, there are still a number of factors that could limit the application of point-of-origin identification. These methods are most often applied in natural populations, which implies a need to rely on the far from perfect knowledge of all biological characteristics of the species in question. Lack of population differentiation across the full or main species distribution areas is a major obstacle for genetic origin assignment. This can be an inherent feature of the population, but may also be caused by the choice and number of markers used to attempt description of population differentiation. In addition, many marine species undertake extensive spawning and feeding migrations (e.g., Ruzzante et al., 2006), which may result in extensive mixing of different genetic populations. If these migrations are not documented, erroneous origin determination could take place. Therefore, careful measures must be taken in the design of population-level identification assays to either possess in depth biological knowledge of the species in question, or to be knowledgeable on the potential limitations of applying the method for seafood identification.

In the case of population mixtures, the genetic population signature of individuals may not reveal the geographical origin, while a larger sample consisting of many individuals could provide an overarching sample signature (i.e., proportions of fish from different populations contributing to the mixture), which may expose the origin. This type of analysis is typically conducted using an alternative, but related, approach, a so-called "mixed stock analysis" (Shaklee, 1990), which is optimal for estimating mixture proportions of individuals originating from different populations, rather than the most likely origin of single individuals. However, in order to use such analysis to infer origin, a database of spatiotemporal population mixture signatures for the species in question is required. This may be feasible for specific species and areas, but in general such data are missing. As a conclusion, one should always be cautious when interpreting data on genetic origin assignment. After all, there are few boundaries in the sea and marine organism can swim or potentially get distributed over vast geographical areas. Therefore, it is important to test specific hypothesis about the origin of an individual, in particular for forensic purposes (i.e., prosecutor and defense claims), to see which hypothesis is most strongly supported by the origin assignment and related statistical inferences.

FUTURE PERSPECTIVES FOR POPULATION AND POINT-OF-ORIGIN DETERMINATION

All fields of genetic research are now benefitting from the massive amounts of genomic information generated through "next-generation sequencing" (NGS) technology. This also holds for many seafood species, where full or partial genomes are available for species, such as Atlantic Salmon, Atlantic Cod, and Turbot, (see http://www.ncbi.nlm.nih.gov/genome for a list). Many more are likely to become available within the near future, thus strongly facilitating the development of markers applicable for individual assignment. Sequencing technology is developing extremely quickly, and so in a few years' time, it may be cheaper and faster to apply NGS methods directly for generating genetic data for origin determination. However, at the moment, it is still more practical to develop specific panels of markers with high power for IA for target species and scenarios. Another field where a lot of progress is anticipated is related to the development of portable real-time devices, which at the moment allow for field-based DNA testing in less than 15 min (for an example see the Genie II http://www.optigene.co.uk/instruments/instrument-genie-ii/). The capacity, in terms of numbers of reactions, of these instruments is relatively low, thereby limiting the applicability for origin determination potentially requiring a high number of markers to provide strong statistical inferences for weakly differentiated populations. Still, they may act as a preliminary on-site screening device, where more detailed analysis can subsequently be performed under laboratory conditions. Despite the challenges in applying genetic origin identification to seafood and seafood products outlined here, it is already an applied and generally superior (and sometimes only) method for assigning individuals back to population/geographical origin. It is expected that the genomic revolution will contribute to faster, more cost efficient and precise tools, which can be applied to a wide range of seafood species. This will, however, require that more focus is diverted to provide a better understanding of the biology and genetic population structure for species inhabiting the world's oceans.

REFERENCES

Anderson, E.C., 2010. Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. Molecular Ecology Resources 10, 701–710.

Bromaghin, J.F., 2008. BELS: backward elimination locus selection for studies of mixture composition or individual assignment. Molecular Ecology Resources 8, 568–571.

Børresen, T., 1992. Quality aspects of wild and reared fish. In: Huss, H.H., Jakobsen, M., Liston, J. (Eds.), Quality Insurance in the Fish Industry. Elsevier, Amsterdam, pp. 1–17.

Cadrin, S.X., Kerr, L.A., Mariani, S., 2014. In: Cadrin, S., Kerr, L., Mariani, S. (Eds.), Stock Identification Methods, Stock Identification: An Overview, second ed. Elsevier.

Cornuet, J.M., Piry, S., Luikart, G., Estoup, A., Solignac, M., 1999. Comparison of methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153, 1989–2000.

Council Regulation (EC) No 1224/2009, 20 November 2009. Official Journal of the European Union L 343 (1), 50.

Council Regulation (EC) No 1005/2008, 29 September 2008. Official Journal of the European Union L 286 (1), 32.

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement on cross-validation. Journal of the American Statistical Association 78, 316–320.

Ellis, J.S., Gilbey, J., Armstrong, A., et al., 2011. Microsatellite standardisation and evaluation of genotyping error in a large multi-partner research programme for conservation of Atlantic salmon (*Salmo salar* L.). Genetica 139, 353–367.

Hansen, M.M., Kenchington, E., Nielsen, E.E., 2001. Assigning individual fish to populations using microsatellite DNA markers: methods and applications. Fish and Fisheries 2, 93–112.

Hare, M.P., Nunney, L., Schwartz, M.K., et al., 2011. Understanding and estimating effective population size for practical application in marine species management. Conservation Biology 25, 438–449.

Helyar, S.J., Hemmer-Hansen, J., Bekkevold, D., et al., 2011. Application of SNPs for population genetics of non-model organism: new opportunities and challenges. Molecular Ecology Resources 11, 123–136.

Higgins, R.M., Danilowicz, B.S., Balbuena, J.A., et al., 2010. Multi-disciplinary fingerprints reveal the harvest location of cod *Gadus morhua* in the northeast Atlantic. Marine Ecology Progress Series 404, 197–206.

Hoban, S., Gaggiotti, O., ConGRESS Consortium, Bertorelle, G., 2013. Sample Planning Optimization Tool for conservation and population Genetics (SPOTG): a software for choosing the appropriate number of markers and samples. Methods in Ecology and Evolution 4, 299–303.

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.

Jorde, L.P., Wooding, S.P., 2004. Genetic variation classification and race. Nature Genetics 36, 28–33.

Kalinowski, S.T., 2011. The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. Heredity 106, 625–632.

Laikre, L., Palm, S., Ryman, N., 2005. Genetic population structure of fishes: implications for coastal zone management. Ambio 34, 111–119.

Limborg, M., Pedersen, J.S., Hemmer-Hansen, J., et al., 2009. Genetic population structure of European sprat *Sprattus sprattus*: differentiation across a steep environmental gradient in a small pelagic fish. Marine Ecology Progress Series 379, 213–224.

Manel, S., Berthier, P., Luikart, G., 2002. Detecting wildlife poaching: identifying the origin of individuals with Bayesian assignments tests and multilocus genotypes. Conservation Biology 16, 650–659.

Manel, S., Gaggiotti, O.E., Waples, R.S., 2005. Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology and Evolution 20, 136–142.

Marko, P.B., Nance, H.A., Guynn, K.D., 2011. Genetic detection of mislabeled fish from a certified sustainable fishery. Current Biology 21, 621–622.

Nielsen, E.E., Cariani, A., Mac Aoidh, E., et al., 2012a. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature Communications 3 Article no 851.

Nielsen, E.E., Hansen, M.M., 2008. Waking the dead: the value of population genetic analyses of historical samples. Fish and Fisheries 9, 450–461.

Nielsen, E.E., Hansen, M.M., Ruzzante, D.E., et al., 2003. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea, revealed by individual admixture analysis. Molecular Ecology 12, 1497–1508.

Nielsen, E.E., Hemmer-Hansen, J., Bekkevold, D., 2012b. Development and application of molecular tools to investigate the mislabeling of cod sold in Sweden. In: Hoofar, J. (Ed.), Case Studies in Food Safety and Authenticity: Lessons from Real-life Situations. Woodhead Publishing, Cambridge.

Nielsen, E.E., Hemmer-Hansen, J., Larsen, P.F., Bekkevold, D., 2009. Population genomics of marine fish: identification of adaptive variation in space and time. Molecular Ecology 18, 3128–3150.

Nielsen, E.E., Kenchington, E., 2001. A new approach to prioritizing marine fish and shellfish populations for conservation. Fish and Fisheries 2, 328–343.

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.

Ogden, R., Linacre, A., 2015. Wildlife forensic science: a review of genetic geographic origin assignment. Forensic Science International-Genetics 18, 152–159.

Paetkau, D., Calvert, W., Stirling, I., Strobeck, C., 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology 4, 347–354.

Piry, S., Alapetite, A., Cornuet, J.-M., et al., 2004. GENECLASS2: a software for genetic assignment and first-generation migrant detection. The Journal of Heredity 95, 536–539.

Primmer, C.R., Koskinen, M.T., Piironen, J., 2000. The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proceeding of the Royal Society B-Biological Sciences 267, 1699–1704.

Pritchard, J., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Putman, A.I., Carbone, I., 2014. Challenges in analysis and interpretation of microsatellite data for population genetic studies. Ecology and Evolution 4, 4399–4428.

Rannala, B., Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences of the United States of America 94, 9197–9201.

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

Ruzzante, D.E., Mariani, S., Bekkevold, D., et al., 2006. Biocomplexity in a highly migratory pelagic marine fish, the Atlantic herring. Proceeding of the Royal Society B-Biological Sciences 273, 1459–1464.

Shaklee, J.B., 1990. The electrophoretic analysis of mixed-stock fisheries of Pacific salmon. Progress in Clinical and Biological Research 344, 235–265.

Sinclair, M., Power, M., 2015. The role of "larval retention" in life-cycle closure of Atlantic herring (*Clupea harengus*) populations. Fisheries Research 172, 401–414.

Waples, R.S., Gaggiotti, O., 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology 15, 1419–1439.

Ward, R.D., Woodwark, M., Skibinski, D.O.F., 1994. A comparison of genetic diversity levels in marine, freshwater and anadromous fishes. Journal of Fish Biology 44, 213–232.

Weir, B.S., Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evolution 36, 1358–1370.

Wright, S., 1950. Genetical structure of populations. Nature 166, 247–249.