

Research



**Cite this article:** Farrell ED *et al.* 2022

A baseline for the genetic stock identification of Atlantic herring, *Clupea harengus*, in ICES Divisions 6.a, 7.b–c. *R. Soc. Open Sci.* **9**: 220453. <https://doi.org/10.1098/rsos.220453>

Received: 12 April 2022

Accepted: 19 August 2022

**Subject Category:**

Ecology, Conservation, and Global Change Biology

**Subject Areas:**

ecology/genetics/ecosystems

**Keywords:**

fisheries, management, genetic assignment, stock identification, Northwest herring, West of Scotland herring

**Author for correspondence:**

Edward D. Farrell

e-mail: [edward.d.farrell@gmail.com](mailto:edward.d.farrell@gmail.com)

Electronic supplementary material is available online at <https://doi.org/10.6084/m9.figshare.c.6168324>.

# A baseline for the genetic stock identification of Atlantic herring, *Clupea harengus*, in ICES Divisions 6.a, 7.b–c

Edward D. Farrell<sup>1,7</sup>, Leif Andersson<sup>2,3,4</sup>, Dorte Bekkevold<sup>5</sup>, Neil Campbell<sup>6</sup>, Jens Carlsson<sup>7</sup>, Maurice W. Clarke<sup>8</sup>, Afra Egan<sup>8</sup>, Arild Folkvord<sup>9</sup>, Michaël Gras<sup>8,10</sup>, Susan Mærsk Lusseau<sup>6,11</sup>, Steven Mackinson<sup>12</sup>, Cormac Nolan<sup>8</sup>, Steven O'Connell<sup>6</sup>, Michael O'Malley<sup>8</sup>, Martin Pastoors<sup>13</sup>, Mats E. Pettersson<sup>2</sup> and Emma White<sup>8</sup>

<sup>1</sup>EDF Scientific Limited, Rathaha, Ladysbridge, Cork, Ireland

<sup>2</sup>Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

<sup>3</sup>Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, Texas, USA

<sup>4</sup>Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

<sup>5</sup>National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark

<sup>6</sup>Marine Scotland Science, 375 Victoria Road, Aberdeen AB11 9DB, Scotland

<sup>7</sup>Area 52 Research Group, School of Biology and Environmental Science/Earth Institute, Science Centre West, University College Dublin, Dublin, Ireland

<sup>8</sup>Marine Institute, Rinville, Oranmore, Co. Galway, Ireland

<sup>9</sup>Department of Biological Sciences, University of Bergen, Bergen, Norway

<sup>10</sup>European Commission, Joint Research Centre (JRC), Ispra, Italy

<sup>11</sup>National Institute of Aquatic Resources, Willemoesvej 2, Hovedbygning, 067, 9850 Hirtshals, Denmark

<sup>12</sup>Scottish Pelagic Fishermen's Association, Heritage House, 135-139 Shore Street, Fraserburgh, Aberdeenshire, Scotland

<sup>13</sup>Pelagic Freezer-trawler Association, Louis Braillelaan 80, 2719 EK Zoetermeer, The Netherlands

EDF, 0000-0002-0070-9154; LA, 0000-0002-4085-6968; DB, 0000-0002-5297-032X; JC, 0000-0002-9262-5627; MWC, 0000-0002-6036-9040; AF, 0000-0002-4763-0590; MEP, 0000-0002-7372-9076

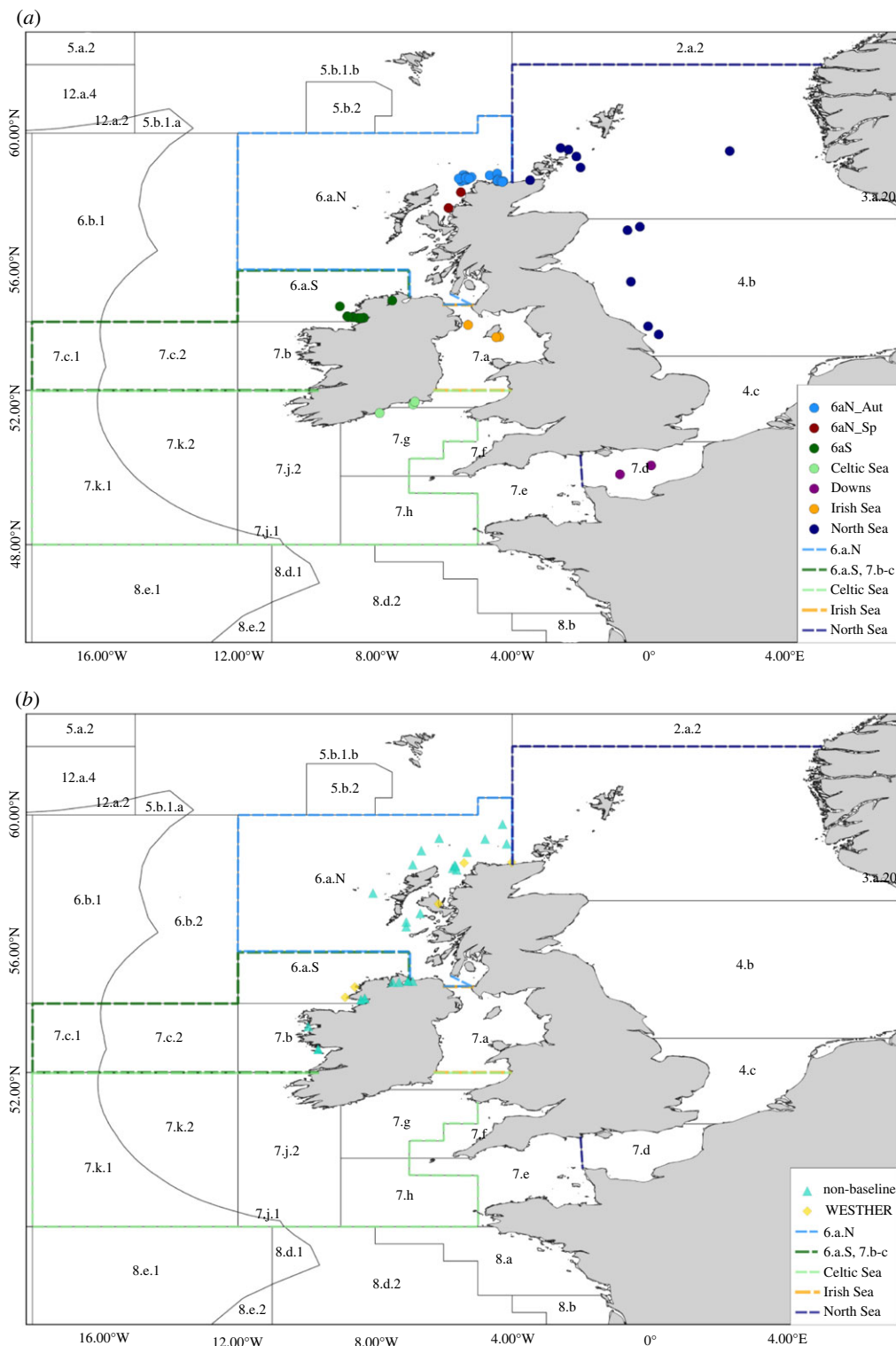
Atlantic herring in International Council for Exploration of the Sea (ICES) Divisions 6.a, 7.b–c comprises at least three populations, distinguished by temporal and spatial differences in spawning, which have until recently been managed as two stocks defined by geographical delineators. Outside of spawning the populations form mixed aggregations, which are the subject of acoustic surveys. The inability to distinguish the populations has prevented the development of separate survey indices and separate stock assessments. A panel of 45 single-nucleotide polymorphisms, derived from whole-genome sequencing, were used to genotype 3480 baseline spawning samples (2014–2021). A temporally stable baseline comprising 2316 herring from populations known to inhabit Division 6.a was used to develop a genetic assignment method, with a self-assignment accuracy greater than 90%. The long-term temporal stability of the assignment model was validated by assigning archive (2003–2004) baseline samples (270 individuals) with a high level of accuracy. Assignment of non-baseline samples (1514 individuals) from Divisions 6.a, 7.b–c indicated previously unrecognized levels of mixing of populations outside of the spawning season. The genetic markers and assignment models presented constitute a ‘toolbox’ that can be used for the assignment of herring caught in mixed survey and commercial catches in Division 6.a into their population of origin with a high level of accuracy.

## 1. Introduction

Fish stock identification has been an important prerequisite for fisheries stock assessment throughout its history [1]. However, the central fundamental weakness that remains in many existing stock assessments is the inaccurate recognition, definition and delineation of ‘stocks’ for data collection and aggregation. Traditionally, exploited stocks have been defined, assessed and managed according to geographical and political features or regions. Such is the case in the northeast Atlantic (Food and Agriculture Organization of the United Nations (FAO) Major Fishing Area 27) where the European Union defines the term ‘stock’ as ‘a marine biological resource that occurs in a given management area’ and delineates and names stocks using International Council for Exploration of the Sea (ICES) statistical areas [2]. As more information becomes available, it is evident that the temporal and spatial distributions of most fisheries resources are not aligned to these artificial divisions [3] and that biological populations are more dynamic and complex [4,5].

While delineation by predefined area may be convenient for management and regulation purposes, accurately assessing the status, biomass and sustainable exploitation rates of mixed ‘stocks’ is inherently difficult if not impossible, as they do not correspond to biological units. Fisheries-dependent and independent data may be confounded, which may mask changes in the abundance of individual populations and lead to biased estimates of population abundance and overexploitation of smaller populations [6]. It is thus critical to identify the underlying population structure of fisheries resources in order to identify the appropriate level at which to aggregate or segregate data for defining assessment and management units. It is also important to be able to assign individuals in mixed survey and commercial catches to the population or assessment unit to which they belong [6,7] in order to ensure the validity of data for inclusion in stock-specific assessments. An ideal method of stock identification should be reproducible among laboratories and enable monitoring of the spatial and temporal integrity of a stock.

There is a long history of research into the characterization of Atlantic herring (*Clupea harengus* Linnaeus, 1758) populations using a wide variety of different techniques, including life-history characteristics, morphometric and meristic characters of whole bodies and otoliths, parasite analyses, physical tagging and genetic approaches (see [8–10]). While many of the approaches have purported to offer reliable methods of discrimination between different populations, the reality is that confusion surrounding the population structure in herring across its distribution has persisted. This has prevented the identification of populations and hampered the delineation of stocks in many cases, for instance in the waters around Ireland and Britain where ICES currently assesses five herring stocks. The North Sea autumn spawning stock (ICES Subarea 4, Divisions 3.a and 7.d.) is the most abundant and well-studied [11] and is considered to be a complex of four spawning components (the autumn spawning Shetland/Orkney, Buchan, Banks components and the winter spawning *Downs* component), which are largely managed as one unit [12,13]. The definition of the western herring stocks has changed considerably over the last five decades (see [8,14]) and the main stocks are currently recognized as *6aN\_Aut* (6.a.N); *6aS* (6.a.S, 7.b and 7.c); *Irish Sea* (7.a north of 52°30' N) and *Celtic Sea* (7.a south of 52°30' N, 7.g, 7.h, 7.j



**Figure 1.** (a) The distribution of the baseline herring samples collected and analysed as the *full baseline dataset* in the current study. (b) The distribution of WESTHER and non-baseline herring samples in the current study. In both figures, the current stock boundaries are indicated according to the legend. The ICES Divisions are indicated by the numbers and letters within the plots, and the latitude and longitude are indicated outside the frame.

and 7.k) (figure 1; electronic supplementary material, table S1; [15]). The *6aN\_Aut* herring spawn in autumn (September/October) off Cape Wrath on the north coast of Scotland, the *6aS* herring spawn in winter (November–February) primarily off the coast of Donegal in the northwest of Ireland, Irish Sea herring spawn in autumn (September/October) mainly on the Douglas Bank east of the Isle of Man in

the Irish Sea, and Celtic Sea herring spawn in winter (November–February) off the south coast of Ireland. Several groups of spring spawning (February–May) herring are also known to occur in the Minch (*6aN\_Sp*), Clyde and Milford Haven, though these are not currently assessed and are believed to be small populations (see review in [8]). Other autumn/winter spawning herring groups are also found in the western English Channel and Bristol Channel (ICES Divisions 7.e and 7.f, respectively), though no assessment is made of these groups and there are no management measures in place.

The stock divisions for herring assessments and management around Ireland and Britain are largely based on the recognition of temporal and spatial differences in spawning season and grounds and are believed to broadly align with biological population structure [14]. Though some geographical and political boundaries are still in place, the mixing across these boundaries is unclear. This is evident in ICES Division 6.a, where the *6aN\_Aut* stock is separated from the North Sea autumn spawning stock by the 4° west line of longitude, despite there being no biological evidence that these represent different populations (see review in [8]). Within Division 6.a, the herring are subdivided into two stocks (figure 1) by the 56° north line of latitude and 7° west line of longitude [16]. Herring caught or surveyed to the north or east of this boundary (excluding the Clyde area) are included as part of the *6aN\_Aut* stock regardless of their population of origin or their spawning time. This includes herring caught in Lough Foyle in NW Ireland, whose waters are bisected by the 7° west line. Most of Lough Foyle is west of this line; however, the mouth of Lough Foyle is east of this line and hence the herring in the whole Lough are considered to be part of the *6aN\_Aut* stock despite having no biological affinity to this population i.e. they spawn in a different area at a different time and are from a different population. Herring caught to the south and west of the 56° and 7° lines are considered to be part of the *6aS* stock in combination with herring in Divisions 7.b and 7.c.

Adult herring from different populations, both within Division 6.a (*6aN\_Aut*, *6aS*, *6aN\_Sp*) and possibly from adjacent populations (*Irish Sea* and *Celtic Sea*) are believed to form mixed aggregations on common feeding grounds in Division 6.a during summer [9]. It is during this time that they are surveyed by the annual Malin Shelf Herring Acoustic Survey (MSHAS), part of the internationally coordinated Herring Acoustic Survey (HERAS), which is the primary tuning index used in the stock assessments of Division 6.a herring. The inability to assign herring catches from the MSHAS into their population of origin prevents the development of separate indices of abundance for the populations in Division 6.a; therefore, ICES has conducted a combined assessment of these populations since 2015 [14], which provides combined management advice. Combined management of separate stocks can only be precautionary if the two stocks are of similar size and are homogeneously distributed together in commercial catches. If these conditions are not met, uncertainty of the status of each of the individual stocks increases, as does the risk that one stock may sustain higher fishing mortality than the other.

Genetic assignment methods, which compare genetic data from individuals with genetic profiles of reference samples from potential source populations to determine the population of origin [17], offer the potential to resolve these issues. However, the incorporation of genetic assignment methods into regular fisheries data collection, assessment and management has been slow [4,18,19], as many existing genetic studies have been hampered by high cost, inadequate sampling coverage, low numbers of suitable molecular markers and low power to detect genetic structure. The advent of high-throughput sequencing technologies fundamentally changed the way in which genetic sequence data are generated (see [20,21]). It is now possible to generate large genomic datasets for non-model species, which facilitate the identification of genetic loci with high discriminatory power for resolving specific population differentiation questions [22,23]. There has also been a shift toward the analysis of sequence variation of functional, adaptive significance rather than just neutral DNA sequence variation [24]. This approach focuses on identifying adaptive markers that are under diversifying selection and may reflect distinctive features of local populations [23]. Small panels of high-graded markers may be selected to develop efficient and cost-effective genetic assignment tools for informing marine fisheries assessment and management [25].

Assignment methods that attempt to solve classification problems rely on computing a discriminant function based on samples from potential source populations and then classify unknown individuals to the group with the highest discriminant score [17]. Genetic assignment methods have traditionally relied on using the genotypic frequency distribution under the assumption of Hardy–Weinberg equilibrium (HWE) and linkage equilibrium in each source population as their discriminant function [17]. These genetic assignment methods can be broadly divided into Bayesian [26], frequency [27] and distance [28] based methods. The underlying assumptions of the methods are quite similar although the distance-based methods may be less sensitive to violations of population genetic expectations such as HWE and linkage equilibrium [28]. These methods are commonly implemented in the software

GeneClass2 [29]. In the absence of baseline data to guide classification, Bayesian clustering methods may be used to delineate clusters of individuals based on their multi-locus genotypes and assign individuals to their individual clusters [17]. However, these Bayesian clustering analyses such as that implemented in the software *Structure* [30] are also constrained by the underlying assumptions of HWE and linkage equilibrium. Multivariate analysis has several advantages over other classical approaches used in population genetics and genetic assignment, the foremost of which is that they do not require the assumptions of HWE or linkage equilibrium [31]. Multivariate approaches are particularly suited to solving classification problems when used in the form of supervised machine-learning (SML) approaches. SML is concerned with predicting the value of a response label/category on the basis of the input variables/features [32]. When empirical data are available, SML trains an algorithm based on a training set of the labelled data, which can then be used to predict the category of unknown data. Support vector machines (SVM) are a set of SML methods that can be used for classification problems. The objective of SVM algorithms is to find a hyperplane in an N-dimensional space (N—the number of features) that distinctly classifies the data point (see [33]). SVM models can also be used to classify nonlinear data through the use of nonlinear kernels [33] and can be optimized by adjusting parameters, including cost and gamma, which control the stringency of the boundary and the influence of single training datapoints, respectively. The R package *assignPOP* [34] has recently made the use of SVM models for assignment more accessible and also allows for the integration of genetic and non-genetic data within the same model, which is an advantage in many stock identification studies which also collect morphometric data.

Recent studies using whole-genome sequencing (WGS) approaches have comprehensively clarified the genetic population structure of Atlantic herring across its distribution and have illustrated that herring populations show a strong population structure inferred to be associated with a high level of local ecological adaptation [22,35,38]. Genetic markers associated with loci under selection have also been proven to provide a significantly better resolution to distinguish population structure than neutral genetic markers [35]. From the *ca* 10 million single-nucleotide polymorphisms (SNPs) identified by Han *et al.* [35] a small subset of *ca* 800 SNPs, associated with ecological adaptation to different geographical areas and spawning conditions, were shown to be able to discriminate all the sampled populations of herring (53 population samples,  $n = 35\text{--}110$  individuals pooled per sample) from across the species distribution. The populations around Ireland and Britain are the southernmost ecomorphs of herring in the Northeast Atlantic and are genetically distinct from the other adjacent Northeast Atlantic herring populations, including Norwegian spring spawning herring to the north and the Baltic herring to the east, from which they can be discriminated with a small subset of 12 independent loci [35]. The populations sampled around Ireland and Britain could be further subdivided into four main groups: the spring spawning herring from the Minch (*6aN\_Sp*) and the Clyde; the *6aN\_Aut* and North Sea autumn spawning herring, which were indistinguishable from each other, thus supporting the conclusion these stocks are most likely a single population; the *6aS* herring, which were differentiated from all populations but more closely related to the southern group consisting of Celtic Sea, Irish Sea and Downs herring, which had the lowest level of genetic differentiation between them. Identification of the primary genome level differences between the herring populations around Ireland and Britain offers the potential to develop a genetic assignment method for discriminating and resolving the outstanding issues of separating mixed survey and commercial catches in ICES Divisions 6.a, 7.b–c.

The aims of the current study were therefore to

- (i) validate a small ‘toolbox’ of genetic markers, from those identified by Han *et al.* [35], that could be used to develop genetic baselines for the individual populations in Divisions 6.a, 7.b–c;
- (ii) develop a temporally stable genetic baseline dataset by collecting and analysing multiple years of spawning baseline samples from each population; and
- (iii) develop, test and validate an assignment model for the genetic assignment of individuals of unknown origin collected in Divisions 6.a, 7.b–c back to their population of origin.

## 2. Materials and methods

### 2.1. Sampling and DNA isolation

Samples of herring were collected from the catches of fisheries surveys and commercial fishing operations, between 2014 and 2021 in the core ICES Divisions 6.a, 7.b–c area and on the adjacent populations where

possible. Each fish was measured for total length (to the 0.5 cm below), total body weight to the nearest 1 g and assessed for sex and maturity. Samples processed by Marine Scotland Science were maturity staged using the 9-point scale, those processed by the Irish Marine Institute were maturity staged using the 8-point scale, and samples processed by the Wageningen University and Research on behalf of the Dutch Pelagic Freezer Trawler Association were maturity staged using the ICES 6-point scale [36]. All maturity stages were converted to the ICES 6-point scale according to Mackinson *et al.* [37]. A 0.5 cm<sup>3</sup> piece of tissue was excised from the dorsal musculature of each specimen, taking care to avoid skin and scales, and stored in absolute ethanol at 4°C. Archive fin clips were also available from the spawning baseline samples collected during the WESTHER project 2003–2004 (FP5-LIFE QUALITY Q5RS-2002–01056; [9]). Total genomic DNA (gDNA) was extracted from *ca* 10 mg of tissue or fin clip from each fish using 300 µl of 10% Chelex suspension and 5 µl of Proteinase K (20 mg µl<sup>-1</sup>). Extracted DNA was stored in 96-well PCR plates at –20°C until ready for genotyping.

## 2.2. Genetic marker identification

The SNPs used in the current study (electronic supplementary material, table S2) were identified during the GENSINC project (GENetic adaptations underlying population Structure IN herring; Research Council of Norway project 254774) and were derived from the analyses of WGS of pooled samples from herring populations across the species distribution, which was undertaken to study the biological significance of the genetic variants underlying ecological adaptation in the Atlantic herring [22,35,38,39]. The subset of SNPs was selected following testing of candidate SNPs with the highest delta allele frequency values from the major genomic regions of divergence in the contrasts between populations around Ireland and Britain. The 45 SNPs selected were distributed across eight chromosomes and comprised 14 loci (electronic supplementary material, table S2). Linked SNPs were retained in the panel in order to add a level of redundancy and ensure that key genomic regions (loci) were well represented even in the instance of missing genotype data from an individual.

## 2.3. Genotyping

The majority of samples were genotyped using a genotyping by sequencing approach [40] described in detail and validated in Farrell *et al.* [41]. In short, locus-specific forward and reverse primers were designed for SNP loci with the Primer3 application [42] in Geneious® 7.0 [43] with optimal primer length set at 20 bp and product size range at 120–180 bp. Primers were designed to bind in conserved flanking regions to minimize the possibility of null alleles and were cross-referenced with existing genome sequence data to identify primers that annealed to multiple regions, which if detected were excluded. The forward and reverse locus-specific primers were adapted, to facilitate combinatorial barcoding of amplicons, by adding either an M13-R (5'-GGAAACAGCTATGACCAT-3') or CAG (5'-CAGTCGGGCGTCATCA-3') universal tail to the 5' end and were divided into multiplex panels in MultiPLX 2.1 [44]. A set of ninety-six 11 bp combinatorial barcodes were used to identify individuals within pooled sequencing runs. An M13-R universal tail was added to the 3' end of 48 of the barcodes and a CAG universal tail to the 3' end of the remaining 48 barcodes, yielding 2304 possible combinations. The tagged primers and tagged barcodes were tested for the formation of secondary structures (hairpins, primer dimers and hetero dimers) with the IDT OligoAnalyzer Tool 3.1 (<http://eu.idtdna.com/calc/analyzer>).

Amplification and barcoding reactions were carried out using a two-step PCR as described in Farrell *et al.* [8,41]. In short, the first PCR involved the amplification of the target SNPs and the second PCR involved the incorporation of the combinatorial barcodes for individual identification. Following PCR amplification, each plate of amplicons was pooled and then standardized for concentration and combined into a single sample before sending for library preparation and amplicon sequencing by a third-party sequencing service provider. Six different herring amplicon sequencing runs were conducted over the course of the current project using both the Illumina MiSeq and HiSeq platforms. The raw data from these runs were treated following the same protocols in order to derive the final individual genotypes. Raw FASTQ sequence data were downloaded from Illumina BaseSpace, and initial quality control was performed using FastQC [45]. Reads were sorted and grouped using a modified python script [40] based on the Levenshtein distance metric. The raw sequence data were processed by identifying sequence reads containing the forward and reverse combinatorial barcodes and the locus-specific primers. Reads were sorted hierarchically and grouped into five separate FASTA files as reads with: no barcode, one barcode, two barcodes and no primers, two barcodes and

two non-matching primers, two barcodes and two matching primers. Only reads containing two barcodes and two matching primers were included in further analyses. These reads were grouped by locus and individual before removing the barcode from the sequences.

SNP genotyping was automated by using a modified Perl script from the genotyping-in-thousands by sequencing (GT-seq) approach [46], which counts amplicon-specific sequences for each allele, and uses allele ratios to determine the genotypes. The Perl scripts were modified in the current project to use the output of the custom python scripts as the input. The default settings of the GT-Seq Perl script designated allele ratios greater than 10.0 to be called as homozygous for allele 1, ratios less than 0.1 to be called as homozygous for allele 2 and ratios between 0.2 and 5.0 to be called as heterozygous [46]. These ratios were optimized for the data and markers in the current study by analysing each marker separately and plotting the genotyping calls from which new ratios were calculated for each marker. The average designated allele ratios in the current study were greater than 5.0 to be called as homozygous for allele 1, ratios less than 0.2 to be called as homozygous for allele 2, ratios between 0.3 and 3.33 to be called as heterozygous and ratios between 3.34–4.9 and 0.201–0.29 were called as NA (NA = no genotype call made). Individuals with less than 10 reads at a particular locus were also designated as NA. Only individuals with greater than 89% genotyping success (i.e. 40/45 genotypes) were retained in the dataset.

Genotyping of the majority of samples collected from quarter three 2019 to 2021 was undertaken by a commercial provider; IdentiGEN, Dublin, Ireland, using their proprietary *IdentiSNP* genotyping assay chemistry, which uses target-specific primers and universal hydrolysis probes. Following an endpoint PCR reaction, different genotypes were detected using a fluorescence reader. Concordance between the two genotyping methods was confirmed by genotyping a subset of samples from each of the target populations ( $n = 24$  per population) and confirming that the same genotypes were called with each method (data not shown).

## 2.4. Baseline dataset analyses

It should be noted that the aim of the current study was not to undertake an exhaustive population genetics and demographic study of the herring populations around Ireland and Britain but was to develop a genetic-based method to separate the herring caught in putatively mixed survey and commercial catches in ICES Divisions 6.a, 7.b–c into their population of origin. The analytical approaches followed were tailored to this specific task. The limited number of genetic markers used in the current study were high graded to maximize the power of discrimination between the core Division 6.a populations and in some instances comprised multiple SNPs from a small number of loci. Therefore, the dataset may not be suitable for conventional population genetic analyses and as such some of the analyses presented (e.g. estimation of fixation indices) were for exploratory purposes only.

Deviations from HWE, linkage disequilibrium (LD) and excess and deficiency of heterozygotes in the *full baseline dataset* (see Results) were tested with Genepop 4.2 using default settings [47]. Microsatellite analyser (MSA) 4.05 [48] was used, under default settings, to assess multi-locus pairwise  $F_{ST}$  with 1000 bootstrap replications and 10 000 permutations. In all cases with multiple tests, significance levels were adjusted using the sequential Bonferroni technique [49]. In order to visualize the pairwise  $F_{ST}$  results and to explore the relationships between the different samples, principal coordinate analysis (PCoA) using the covariance standardized method was conducted in GenAlEx 6.51b2 [50].

Discriminant analysis of principal components (DAPC), from the R package *adegenet* [51,52], is a multivariate approach that transforms multi-locus genotype data using principal component analysis (PCA) to derive a set of uncorrelated variables, which serve as input for discriminant analysis (DA). The DA aims to maximize among-group variation and minimize within-group variation. DAPC does not make assumptions of underlying population genetic processes (e.g. neutrality, linkage equilibrium, HWE); therefore, it was appropriate to use this approach with the data in the current study. In the first instance, DAPC was run using the 64 baseline samples as the input groups and retaining all PCs and discriminant functions. The DAPC was run again after the temporal samples were combined to form seven groups (*6aS*, *Celtic Sea*, *Irish Sea*, *6aN\_Aut*, *6aN\_Sp*, *North Sea*, *Downs*) which represented the putative populations in the study area. Following further analyses (see Results), a reduced *6a baseline dataset* consisting only of samples from groups that are confirmed as being present in Division 6.a, i.e. *6aS*, *6aN\_Aut* and *6aN\_Sp*, was analysed using DAPC. In this instance, DAPC was conducted as before with prior definition of group membership (*Approach 1*) and also following a second approach (*Approach 2*) using the *find.clusters* function to infer genetic clusters. This function transforms

the data using PCA, then runs the  $K$ -means algorithm (function *kmeans* from the *stats* package) with increasing values of  $K$  and computes Bayesian information criterion to assess the best supported model.

## 2.5. Assignment model development

The R package *assignPOP* [34], which performs population assignment using a machine-learning framework, was used to develop the assignment model. *assignPOP* uses Monte Carlo cross-validation (*assign.MC*) to divide the baseline data into a training dataset and test dataset. The assignment model is developed with the training dataset and subsequently tested with the independent test dataset, which avoids introducing ‘high-grading bias’ (see [53]). As the Monte Carlo procedure samples random individuals each time, it does not guarantee that every individual is sampled. Therefore, *assignPOP* can perform an additional method of  $K$ -fold cross-validation (*assign.kfold*), which involves randomly dividing the individuals from each population into  $K$  groups and then using one group from each population as test individuals and the remaining  $K - 1$  groups as the training individuals. Assignment tests are performed until every group and hence individual is tested, resulting in  $K$  tests. *assignPOP* has a number of classification model options including the SVM model from the R package *e1071* [54]. Based on the results of the aforementioned baseline dataset analyses, it was decided to develop the assignment model in *assignPOP* trained on the *6a baseline dataset* as the constituent populations had the highest level of discrimination between them with the current marker panel. As per the DAPC analyses, the assignment model was developed using two different approaches with, in this case, each approach conducted at two hierarchical levels.

*Approach 1* used the *6a baseline dataset* with the predefined *6aS*, *6aN\_Aut* and *6aN\_Sp* population groups. *Approach 2* was initially independent of the assumptions of prior populations and instead used the output of the  $K$ -means clustering analyses of the *6a baseline dataset* to identify different baseline assignment clusters. In the cases where multiple clusters represented a single assumed population of origin, these clusters were combined. The assignment approaches were conducted at two hierarchical levels based on the power to discriminate the different groups in the DAPC analyses. In *Level 1*, the winter and spring spawning herring (*6aS* and *6aN\_Sp*) were combined and tested against the autumn spawning herring (*6aN\_Aut*). In *Level 2*, the combined winter and spring spawning group was split, and the individual groups tested against each other.

In order to avoid over-fitting the model and to objectively determine the optimum number of PCs to be used in both assignment approaches, DAPC cross-validation was conducted with the *xvalDapc* function in *adegenet*. Exploratory analyses were conducted in *assignPOP* to determine the optimum model and kernel for the assignment model, and the *tune*, *tune.control* and *best.svm* functions in R package *e1071* [54] were used to perform a grid search for the optimum values for cost and gamma. These parameters were used for testing the rate of self-assignment using both Monte Carlo and  $K$ -fold cross-validation to estimate membership probability. In order to avoid unbalanced sample sizes among the baseline groups, the number of individuals in the training sets was specified and was limited by the number of individuals in the smallest group. *Level 1* assignments were tested with 200, 400, 600 and 800 individuals in the training set, while the *Approach 1-Level 2* assignment was tested with 50 and 75 individuals and *Approach 2-Level 2* assignment with 100, 150 and 200 individuals. Both Monte Carlo and  $K$ -fold cross-validation were performed using 25%, 50%, 75% and 100% of the highest  $F_{ST}$  loci (*loci.sample='fst'*), and all tests were conducted with 100 iterations.

An important consideration when developing the assignment model was to determine how many genetic markers were required for accurate assignment using either of the approaches and at either of the levels. This enabled the threshold for missing data of unknown samples to be set with a robust basis without compromising the integrity of the assignments. In order to do this, the Monte Carlo cross-validation analyses were run again with random sampling of loci (*loci.sample='random'*) rather than highest  $F_{ST}$  loci and were run with 20–100% of loci in 10% intervals. All other parameters were the same as the previous runs.

## 2.6. Assignment model validation with archive samples

As an additional validation of the baseline assignment models, the WESTHER baseline samples collected in 2003/2004 in Division 6.a were used as known–unknown samples and assigned to the contemporary baseline in order to test the long-term temporal stability of the assignment models. The WESTHER samples were processed and genotyped following the same method as the other samples in the study. The assignments were conducted using the *assign.X* function in *assignPOP* [34] using the two



hierarchical approaches and with the same model parameters as described above. For each approach, the *Level 1* assignment was conducted with all the individuals in the sample and the *Level 2* assignment on a subset of individuals that required further assignment. A successful assignment probability threshold was set at 0.67, which indicated a situation where one assignment outcome was twice as likely as the alternative outcome. This was deemed an acceptable level of confidence given the high level of self-assignment accuracy of the baseline datasets. Individuals with an assignment probability below the assignment threshold were considered to be unassigned and denoted as *NA*. The final assignment call of each individual was based on a combination of the *Level 1* and *Level 2* assignments. Final sample assignments were plotted using the *draw.pie* function of the R package *mapplots*. In order to test the potential effect of increasing the assignment threshold and also to compare the relative assignment rates between the assignment approaches, the proportion of individuals falling below thresholds of 0.67, 0.7, 0.8 and 0.9 were also calculated.

## 2.7. Exploratory analyses with contemporary non-baseline samples

Additional samples, which were not considered to be baseline samples (i.e. they were not collected on known spawning grounds or were not in the correct maturity stage), were also collected during the study. These non-baseline samples were used to further test the assignment model and also to provide an exploratory analysis of potential mixing of populations within Divisions 6.a, 7.b–c. The samples were assigned as per the WESTHER samples, however. In this instance, the samples were divided by quarter for plotting using *mapplots* in order to minimize overlap between the samples and allow clearer interpretation of the assignments. As with the archive samples, the effect of the range of assignment thresholds was also tested with these samples.

# 3. Results

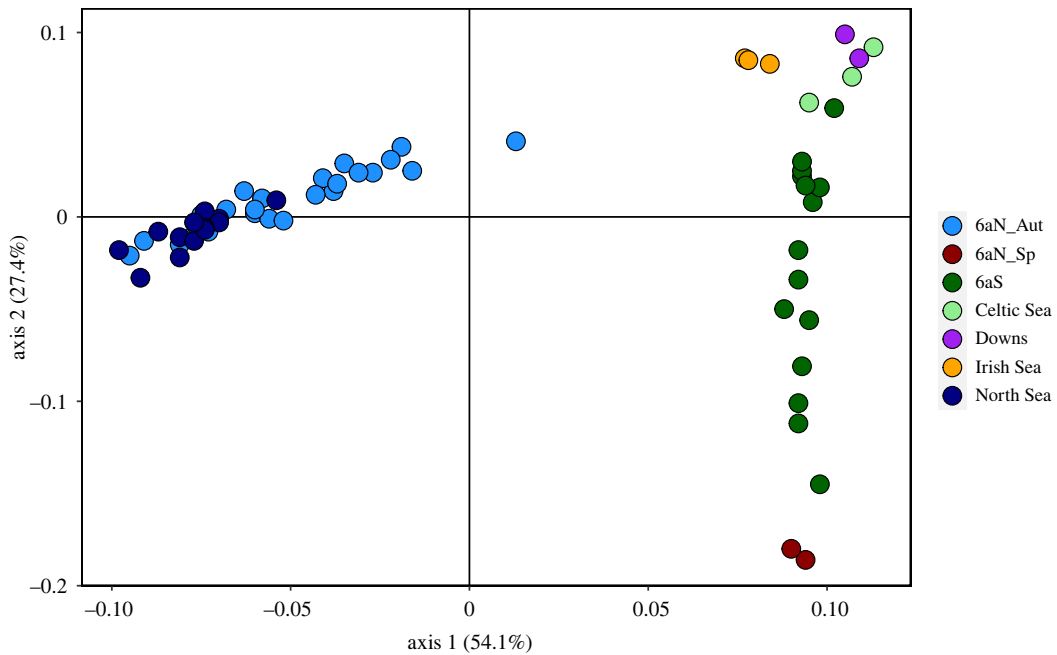
## 3.1. Sampling and genotyping success rate

Due to the opportunistic nature of the sampling, the samples contained a significant mix of length classes and maturity stages (electronic supplementary material, tables S3 and S4). In total, 92 contemporary samples were collected (figure 1; electronic supplementary material, table S3), comprising 6591 individual herring of which 5638 individuals (86%) passed the genotyping threshold of 89% (40/45 SNPs genotyped). All 45 SNPs were successfully genotyped in more than 92% of retained individuals. For the purposes of developing robust baselines for genetic assignment, it is critical to avoid including individuals with uncertain origin. Therefore, only samples with a significant number of maturity stage 3 (spawning) individuals [36] caught in close proximity to known spawning grounds at recognized spawning times were selected to be baseline samples. In order to further limit the potential for misclassification as a baseline spawning sample, only individuals classified as maturity stage 3 were included in the *full baseline dataset*. The resulting contemporary *full baseline dataset* contained 64 samples, comprising a total of 3480 herring (figure 1).

The 28 remaining samples (1514 individuals) were retained in a separate *non-baseline dataset*, to be used to test the assignment model and provide an exploratory analysis of potential mixing of populations within Divisions 6.a., 7.b–c. In addition, five archive baseline samples comprising 340 individuals collected in 2003 and 2004 were also available from the WESTHER project, of which 270 surpassed the genotyping quality control threshold. These samples were retained in an *archive dataset* for the purposes of having independent baseline samples for the validation of the assignment model and for testing the long-term temporal stability of the assignment model.

## 3.2. Baseline dataset analyses

There were no significant patterns of deviation from HWE, heterozygote deficiency or heterozygote excess at the locus level (45 SNPs). At the population level, significant deviations from HWE were observed in samples *6aN\_Sp\_18b* (10/45 SNPs), *6aS\_17d* (11/45), *6aS\_17e* (12/45) and *6aS\_19c* (6/45). Samples *6aS\_17d* and *6aS\_17e* also displayed indications of a significant heterozygote deficiency in 8 and 13 loci, respectively, which was probably the result of some of the *6aS* samples containing a mixture of early and later spawning components (see pairwise  $F_{ST}$  results). Samples *6aN\_Sp\_18b*, *6aS\_19c* and *DWN\_18* displayed indications of significant heterozygote excess at 10, 7 and 7 loci,



**Figure 2.** PCoA of the multi-locus pairwise  $F_{ST}$  analyses of the baseline samples.

respectively. The significant indications of LD were in keeping with the loci already identified (electronic supplementary material, table S2) from Han *et al.* [35]. All markers and all samples were retained in the *full baseline dataset* for further analyses.

The analyses of multi-locus pairwise  $F_{ST}$  (electronic supplementary material, table S5) indicated proportionately higher  $F_{ST}$ s and significant differentiation between samples collected from the different putative populations except between the *6aN\_Aut* and *North Sea* populations, which displayed little if any significant genetic differentiation among or between the temporal samples. There was no significant differentiation between the temporal samples from the *Irish Sea* or *Downs* population, and there was little if any significant genetic differentiation between the temporal samples from the *Celtic Sea* and *6aN\_Sp*. There were some indications of differentiation between the temporal samples from *6aS*, with some of the samples collected in January and February (*6aS\_17a*, *6aS\_17b*, *6aS\_18a*) showing a low level of differentiation from the other samples. The PCoA of the pairwise  $F_{ST}$  results enabled a clearer interpretation and illustrated the clustering of samples within and between putative populations (figure 2). The temporal samples from each population clustered together though some intrapopulation diversity was evident particularly among the *6aS* and *6aN\_Aut* samples. The *6aN\_Sp* samples were distinct from the *6aN\_Aut* samples and were closely aligned with the late *6aS* samples, i.e. those collected in quarter 1.

The DAPC results supported the previous indications of temporal stability within each of the putative population areas with samples from the same putative populations clustering together (electronic supplementary material, figure S1). Therefore, the temporal samples were combined to form seven groups (*6aS*, *Celtic Sea*, *Irish Sea*, *6aN\_Aut*, *6aN\_Sp*, *North Sea* and *Downs*), which represented the putative populations in the study area. The DAPC and MSA analyses were run again on the pooled samples. Pairwise multi-locus  $F_{ST}$  analyses of the pooled *full baseline dataset* indicated significant differentiation between all baseline population groups (table 1). The lowest level of differentiation was between the *6aN\_Aut* and *North Sea* groups. The level of differentiation between these groups ( $F_{ST}=0.016$ ) was lower than the average differentiation (0.032) between all of the samples within the *6aN\_Aut* pool (electronic supplementary material, table S5). The highest level of differentiation was between the *6aN\_Sp* group and the other groups. There was also a very low level of differentiation between the *Downs* group and the *Celtic Sea* and *Irish Sea* groups, while the *Downs* group had a high level of differentiation from the *North Sea* group.

The DAPC results indicated the same pattern of structure as the  $F_{ST}$  analyses (figure 3a) and also as those observed in Han *et al.* [35] based on whole-genome analyses. The highest level of discrimination observed in the DAPC analyses was along the primary axis (74%) and concerned the *6aS* and the

**Table 1.** Pairwise multi-locus  $F_{ST}$  (above the diagonal) for the pooled *full baseline dataset* and associated  $p$ -values (below the diagonal) with the temporal replicates condensed.

	6aS	Celtic Sea	Irish Sea	6aN_Aut	6aN_Sp	North Sea	Downs
6aS		0.09	0.13	0.20	0.35	0.24	0.12
Celtic Sea	0.0001		0.08	0.23	0.64	0.32	0.01
Irish Sea	0.0001	0.0001		0.20	0.67	0.29	0.08
6aN_Aut	0.0001	0.0001	0.0001		0.57	0.02	0.24
6aN_Sp	0.0001	0.0001	0.0001	0.0001		0.60	0.68
North Sea	0.0001	0.0001	0.0001	0.0001	0.0001		0.33
Downs	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	

*6aN\_Aut* groups, though some outliers were evident. The *6aS* and *6aN\_Sp* groups were discriminated primarily on the secondary axis (18%). These groups partially overlapped, indicating a lower potential to accurately discriminate between them. The *6aS*, *Celtic Sea* and *Irish Sea* groups overlapped, indicating that the current marker panel cannot be used to distinguish these groups with a high level of accuracy. Therefore, the *Celtic Sea* and *Irish Sea* groups were removed from the baseline data and excluded from further analyses. There is no evidence of significant numbers of herring from these groups being present in Division 6.a [8]. DAPC also indicated an overlap and an inability to distinguish between the *6aN\_Aut* and *North Sea* groups. There is currently no evidence to support the assertion that the *North Sea* autumn spawning herring comprise a different population to the *6aN\_Aut* herring [8]; however, this distinction was not the focus of the current study; as such, the *North Sea* samples were removed from further analyses. The *Downs* group was confirmed to be distinct from the *North Sea* group though it could not be reliably discriminated from the *Celtic Sea* and *Irish Sea* groups with the current panel of markers and, as such, the *Downs* group was removed from further analyses. The resulting reduced *6a baseline dataset* consisted only of the 43 samples from populations that are confirmed as being present in Division 6.a i.e. *6aS*, *6aN\_Aut* and *6aN\_Sp* (electronic supplementary material, table S3; figure 3b).

Clustering analyses of the *6a baseline dataset* indicated that six clusters were the optimum number to provide the most accurate division of the samples based on their assumed population of origin (electronic supplementary material, table S6). DAPC of the *clustered 6a baseline dataset* indicated clear division between the clusters with minimal overlap (figure 3c), suggesting that an SVM model-based assignment using this approach would have a high accuracy. The majority of the *6aN\_Aut* individuals were represented by the combined *Clusters\_1+3+5*, and the majority of the *6aS* individuals by the combined *Clusters\_4+6* (table 2). The majority of *6aN\_Sp* individuals were in *Cluster\_2*; however, this cluster also contained a significant proportion of *6aS* individuals. These individuals were primarily from the samples of late-spawning herring collected in Division 6.a.S in January and February. In terms of cluster composition, *Clusters\_1+3+5* comprised 98% *6aN\_Aut* samples and as such were considered, for the purposes of the assignment model, a proxy for that population group. *Clusters\_4+6* comprised 89% *6aS* and 10% *6aN\_Aut*. There is some evidence that the *6aN\_Aut* individuals in these clusters may be misidentified *6aS* herring or strayers from *6aS* (see [8]); therefore, *Clusters\_4+6* were considered to represent *6aS* for the purposes of assignment. *Cluster\_2* comprised 54% *6aS* and 44% *6aN\_Sp* and was considered, for the purposes of the assignment, to represent a mix of *6aS* and *6aN\_Sp* herring. The resulting *clustered 6a baseline dataset* comprised *Clusters\_1+3+5*, *Clusters\_4+6*, *Cluster\_2*. In order to simplify the nomenclature and align it with the *Approach 1* assignment, from this point on these clusters will be referred to as *6aN\_Aut*<sup>A2</sup>, *6aS*<sup>A2L2</sup> and *6aS/6aN\_Sp*<sup>A2L2</sup>, respectively.

### 3.3. Assignment model development

The optimum numbers of PCs for the *Approach 1-Level 1* dataset and *Approach 1-Level 2* dataset, determined as the values with the lowest root-mean-squared error following DAPC cross-validation, were 40 and 35, respectively. The optimum number of PCs for the *Approach 2-Level 1* dataset and *Approach 2-Level 2* dataset were 30 and 5, respectively. There was, however, little difference between the number of PCs retained in all cases, suggesting that the assignment is not sensitive to this

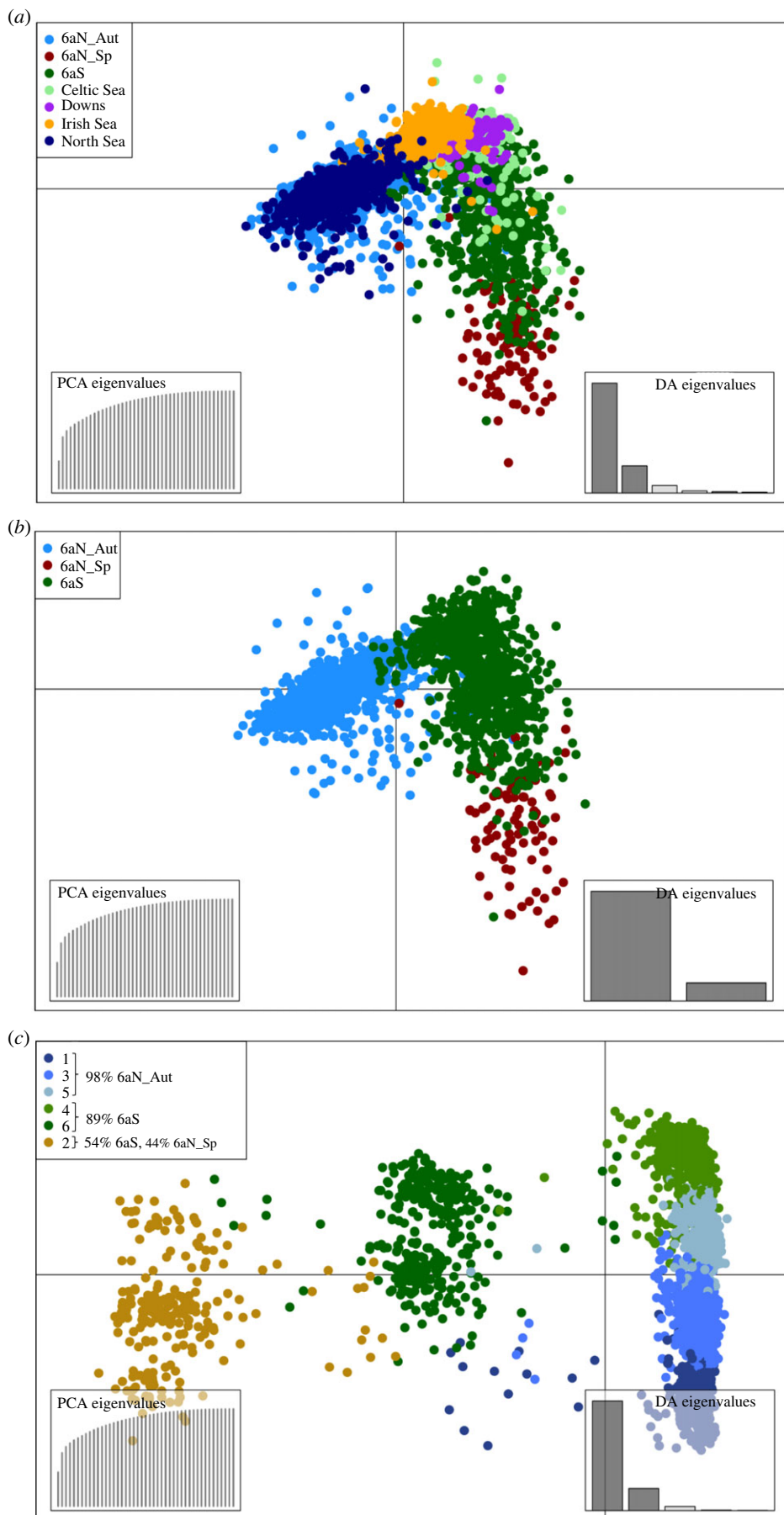


Figure 3. DAPC of (a) the pooled full baseline dataset, (b) the 6a baseline dataset and (c) the clustered 6a baseline dataset.

**Table 2.** Clustering analyses, using the *find.clusters* function in *adeigenet*, of the *6a* baseline dataset. The percentage of each population group split by cluster and the percentage of each cluster split by population are shown.

	1 + 3 + 5	2	4 + 6
<b>Populations by cluster</b>			
6aS	3.5	13.8	82.7
6aN_Aut	93.6	0.4	6.0
6aN_Sp	2.0	95.1	2.9
<b>Clusters by population</b>			
6aS	2.3	53.6	89.4
6aN_Aut	97.5	2.3	10.5
6aN_Sp	0.2	44.1	0.4

parameter. The optimum model and kernel for the assignment model were the SVM model and the radial basis function kernel. Grid search indicated the optimum values for cost and gamma in *Approach 1-Level 1* and *Approach 2-Level 1* were 1 and 0.33, respectively, and in *Approach 1-Level 2* and *Approach 2-Level 2* were 1 and 0.5, respectively.

There was little difference between the self-assignment accuracy of *Approach 1-Level 1* and *Approach 2-Level 1* (electronic supplementary material, figures S2 and S3; table 3). Both approaches resulted in self-assignment rates greater than 90% and neither approach was observed to be particularly sensitive to the number of individuals in the training data. Similarly, neither approach was observed to be particularly sensitive to the proportion of highest  $F_{ST}$  loci used in the analyses. The main difference between the two approaches at *Level 1* was the higher probabilities of assignment and lower error observed in the *K*-fold analyses in *Approach 2* (electronic supplementary material, figure S3; table 3). Conversely there were large differences between the two approaches in the *Level 2* assignments, where *Approach 1-Level 2* did not confidently assign *6aN\_Sp* samples to their baseline, whereas *Approach 2-Level 2* achieved near-perfect self-assignment to both the  $6aS^{A2L2}$  and  $6aS/6aN\_Sp^{A2L2}$ .

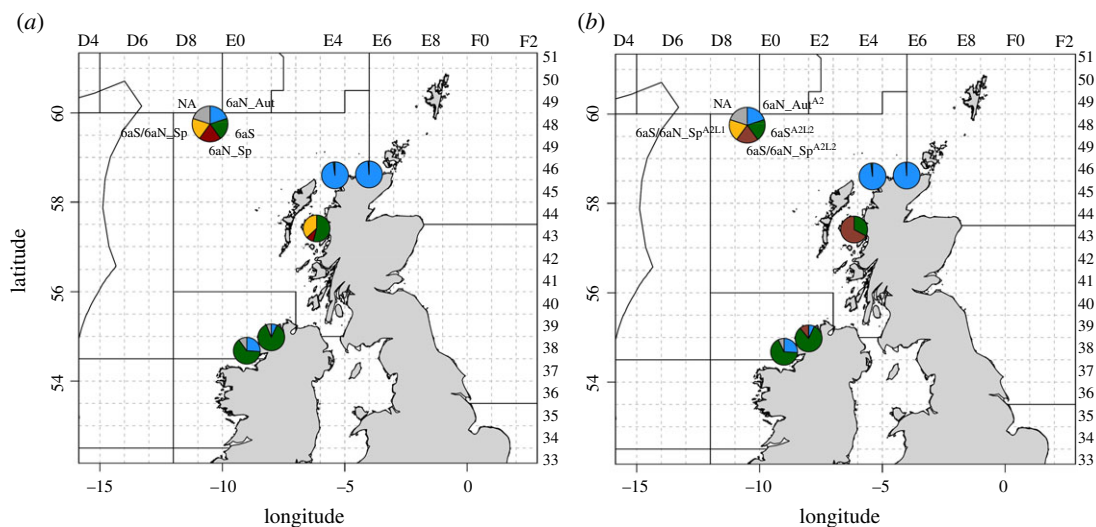
The *Approach 1-Level 1* assignment was more sensitive to the number of loci than *Approach 1-Level 2* (electronic supplementary material, figure S4 and table S7). This was particularly notable for the *6aS/6aN\_Sp* group in *Level 1*, where there was a significant drop in assignment accuracy and an increase in the number of outliers below 50% of loci. This indicated that at least 23 of the 45 loci were required for accurate assignment at this level. Ideally over 60% (27 loci) should be genotyped at this level to ensure assignment accuracy. The *Approach 1-Level 2* assignment was not very sensitive to the number of loci, and there was little difference in the accuracy of assignment down to 20% of the loci. The *Approach 2-Level 1* assignments had a similar pattern of sensitivity to the number of loci as the *Approach 1* assignments indicating that a minimum of 60% of loci should be genotyped at this level to ensure assignment accuracy. The *Approach 2-Level 2* assignments had a higher level of accuracy than the *Approach 1-Level 2* assignments and again were less sensitive to the number of markers than the *Level 1* assignments.

### 3.4. Assignment model validation with archive samples

The final assignment call of each individual in the archive WESTHER samples was based on a combination of the *Level 1* and *Level 2* assignments according to the assignment decision table (table 4). The assignments of the WESTHER *6aN\_Aut* samples from 2003 (*6aN\_03*) and 2004 (*6aN\_04*) displayed near-perfect assignment to the *6aN\_Aut* groups, with only one individual in each of the years being misassigned to the *6aS* groups (figure 4; electronic supplementary material, table S8). The assignment of the 2003 *6aS* WESTHER sample (*6aS\_03*) was more uncertain, with a quarter of individuals misassigned to the *6aN\_Aut* groups (electronic supplementary material, table S8). The 2004 *6aS* sample (*6aS\_04*) had a higher level of correct assignment (greater than 80%). The assignment of the 2004 *6aN\_Sp* (*6aN\_Sp\_04*) sample indicated perfect assignment, at *Level 1* in both approaches, to the *6aS/6aN\_Sp* groups. As expected, the *Approach 1-Level 2* assignment displayed a high rate of misassignment (53%) to the *6aS* group and a high rate of below-threshold individuals (37%), which could not be confidently split below the level of *6aS/6aN\_Sp*. The *Approach 2-Level 2* assignment provided a

**Table 3.** Monte Carlo and  $K$ -fold cross-validation results from the *assignPOP* analyses of the two assignment approaches.

approach	level	method	proportion loci	training individuals	$K$	origin	baseline samples	assignment group membership probability $\pm$ s.d.	
1	1	MC	1	800	–	6aS/6aN_Sp	956	6aS/6aN_Sp 0.94 $\pm$ 0.02	6aN_Aut 0.06 $\pm$ 0.02
						6aN_Aut	1360	0.08 $\pm$ 0.01	0.92 $\pm$ 0.01
1	1	$K$ -fold	1	–	10	6aS/6aN_Sp	956	0.93 $\pm$ 0.03	0.07 $\pm$ 0.03
						6aN_Aut	1360	0.06 $\pm$ 0.03	0.94 $\pm$ 0.03
1	2	MC	1	75	–	6aS	854	6aS 0.86 $\pm$ 0.01	6aN_Sp 0.14 $\pm$ 0.01
						6aN_Sp	102	0.06 $\pm$ 0.04	0.94 $\pm$ 0.04
1	2	$K$ -fold	1	–	10	6aS	854	0.96 $\pm$ 0.02	0.04 $\pm$ 0.02
						6aN_Sp	102	0.53 $\pm$ 0.11	0.47 $\pm$ 0.11
2	1	MC	1	800	–	6aS/6aN_Sp <sup>A2L1</sup> (Clusters_2 + 4+6)	1011	6aS/6aN_Sp <sup>A2L1</sup> 0.96 $\pm$ 0.01	6aN_Aut <sup>A2</sup> 0.04 $\pm$ 0.01
						6aN_Aut <sup>A2</sup> (Clusters_1 + 3+5)	1305	0.04 $\pm$ 0.01	0.96 $\pm$ 0.01
2	1	$K$ -fold	1	–	10	6aS/6aN_Sp <sup>A2L1</sup> (Clusters_2 + 4+6)	1011	0.95 $\pm$ 0.02	0.05 $\pm$ 0.02
						6aN_Aut <sup>A2</sup> (Clusters_1 + 3+5)	1305	0.03 $\pm$ 0.01	0.97 $\pm$ 0.01
2	2	MC	1	200	–	6aS <sup>A2L2</sup> (Clusters_4 + 6)	791	6aS <sup>A2L2</sup> 0.99 $\pm$ 0.00	6aS/6aN_Sp <sup>A2L2</sup> 0.01 $\pm$ 0.00
						6aS/6aN_Sp <sup>A2L2</sup> (Cluster_2)	220	0.00 $\pm$ 0.00	1.00 $\pm$ 0.00
2	2	$K$ -fold	1	–	10	6aS <sup>A2L2</sup> (Clusters_4 + 6)	791	1.00 $\pm$ 0.00	0.00 $\pm$ 0.00
						6aS/6aN_Sp <sup>A2L2</sup> (Cluster_2)	220	0.01 $\pm$ 0.02	0.99 $\pm$ 0.02



**Figure 4.** Assignment output of the archive samples from the WESTHER project following (a) Approach 1 and (b) Approach 2.

**Table 4.** Assignment decision table, indicating the assignment steps in relation to the assignment threshold probability ( $P$ ).

approach	level	assigned group	$P$	action	final assignment
1	1	6aN_Aut	$\geq 0.67$	assigned	6aN_Aut
1	1	6aS/6aN_Sp	$\geq 0.67$	move to level 2	–
1	1	6aN_Aut	$< 0.67$	not assigned	NA
1	1	6aS/6aN_Sp	$< 0.67$	not assigned	NA
1	2	6aS	$\geq 0.67$	assigned	6aS
1	2	6aN_Sp	$\geq 0.67$	assigned	6aN_Sp
1	2	6aS	$< 0.67$	not assigned	6aS/6aN_Sp
1	2	6aN_Sp	$< 0.67$	not assigned	6aS/6aN_Sp
2	1	6aN_Aut <sup>A2</sup> (Clusters_1 + 3+5)	$\geq 0.67$	assigned	6aN_Aut <sup>A2</sup>
2	1	6aS/6aN_Sp <sup>A2L1</sup> (Clusters_2 + 4+6)	$\geq 0.67$	move to level 2	–
2	1	6aN_Aut <sup>A2</sup> (Clusters_1 + 3+5)	$< 0.67$	not assigned	NA
2	1	6aS/6aN_Sp <sup>A2L1</sup> (Clusters_2 + 4+6)	$< 0.67$	not assigned	NA
2	2	6aS <sup>A2L2</sup> (Clusters_4 + 6)	$\geq 0.67$	assigned	6aS <sup>A2L2</sup>
2	2	6aS/6aN_Sp <sup>A2L2</sup> (Cluster_2)	$\geq 0.67$	assigned	6aS/6aN_Sp <sup>A2L2</sup>
2	2	6aS <sup>A2L2</sup> (Clusters_4 + 6)	$< 0.67$	not assigned	6aS/6aN_Sp <sup>A2L1</sup>
2	2	6aS/6aN_Sp <sup>A2L2</sup> (Cluster_2)	$< 0.67$	not assigned	6aS/6aN_Sp <sup>A2L1</sup>

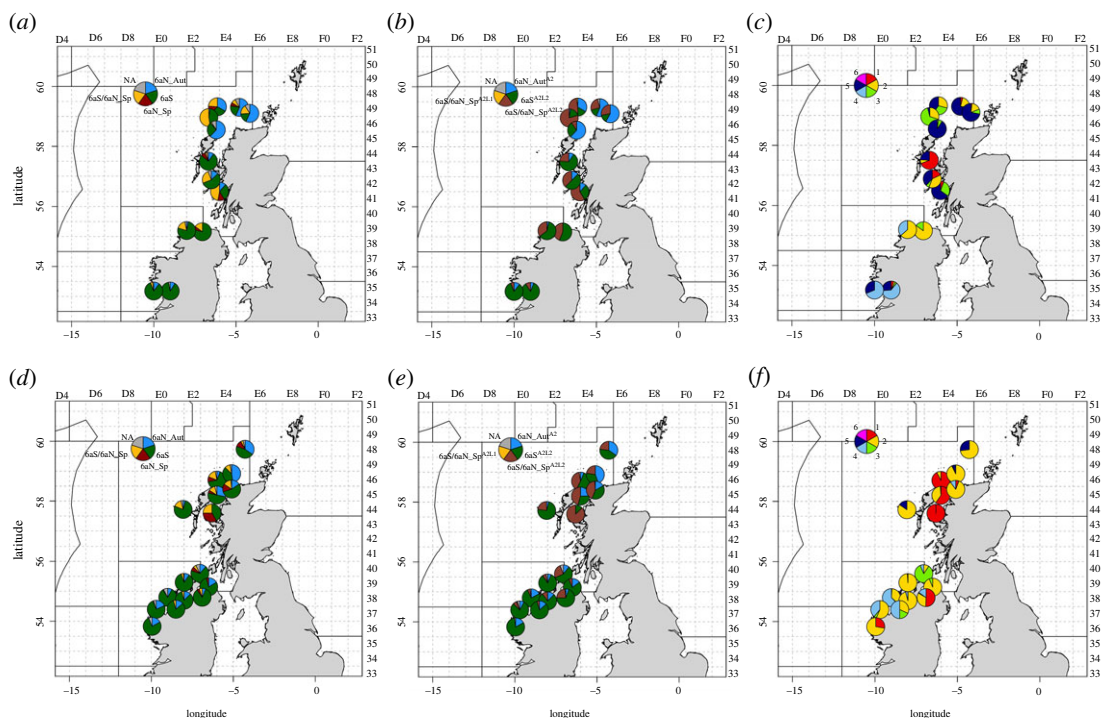
more robust assignment with 67% of individuals assigned to 6aS/6aN\_Sp<sup>A2L2</sup>. It should be noted that the 6aN\_Sp\_04 WESTHER sample did not fulfil the criteria of baseline samples as defined in the current study, as 27% of the individuals were maturity stage 2 individuals (electronic supplementary material, table S3).

The average proportion of unassigned individuals in the WESTHER samples increased with increasing assignment threshold for *Approach 1-Level 1* and 2 but only for *Level 1* in *Approach 2* (table 5). Analysis of the individual rather than average values (electronic supplementary material, table S9) showed differences between the individual samples. All of the 6aN\_Aut samples' individual assignments had a probability greater than 0.9, indicating they were at least nine times more likely that the alternative assignment (electronic supplementary material, table S9). The assignment probabilities for the 2003 6aS individuals were more variable at *Level 1* for *Approach 1* and 2, highlighting a degree of uncertainty around the assignment of some of the individuals; however, the

**Table 5.** The average proportion of non-baseline and WESTHER samples falling below a range of assignment thresholds for the *Approach 1-Level 1* and *2* and *Approach 2-Level 1* and *2* assignments. The individual sample proportions are in electronic supplementary material, table S9.

type	no. samples	no. individuals	approach	$\leq 0.67$		$\leq 0.7$		$\leq 0.8$		$\leq 0.9$	
				Level 1	Level 2	Level 1	Level 2	Level 1	Level 2	Level 1	Level 2
non-baseline	28	1514	1	0.03	0.16	0.04	0.19	0.06	0.25	0.11	0.28
			2	0.02	0.00	0.02	0.00	0.04	0.01	0.08	0.01
WESTHER	5	270	1	0.03	0.07	0.03	0.08	0.05	0.12	0.09	0.13
			2	0.01	0.00	0.02	0.00	0.03	0.00	0.06	0.00





**Figure 5.** The assignment outputs and maturity stages of the contemporary non-baseline samples are divided by quarter; (a–c) = quarters 1 and 2; (d–f) = quarters 3 and 4. Note the exact catch positions have been adjusted to minimize the overlap of the pie charts. The latitude and longitude are indicated outside the frame on the left and bottom. The ICES statistical rectangles are indicated outside the frame on the top and right. (a) Non-baseline Q1 & Q2—Approach 1, (b) non-baseline Q1 & Q2—Approach 2, (c) non-baseline Q1 & Q2—maturity, (d) non-baseline Q3 & Q4—Approach 1, (e) non-baseline Q3 & Q4—Approach 2 and (f) non-baseline Q3 & Q4—maturity.

Level 2 assignments all had a probability greater than 0.9 (electronic supplementary material, table S9). The 2004 *6aS* individuals also showed some uncertainty at Level 1 and at Approach 1–Level 2. The Approach 2–Level 2 assignments all had a probability greater than 0.9. The individuals of the *6aN\_Sp\_04* sample had the highest proportion of unassigned individuals at the higher thresholds for Approach 1–Level 2. The Approach 2–Level 2 assignments all had a probability greater than 0.9.

### 3.5. Exploratory analyses with contemporary non-baseline samples

The quarter 1 non-baseline samples from Division 6.a.N came primarily from the Scottish West Coast International Bottom Trawl Survey and comprised a number of small samples of herring of mixed maturity stages (figure 5; electronic supplementary material, tables S3 and S4). The most northerly samples (*6aN\_Sp\_19a* and *6aN\_Sp\_19b*) were dominated by resting (Stage 5) individuals, and the assignments indicated a significant proportion of *6aN\_Aut* individuals, though assignments to both the *6aS* and *6aS/6aN\_Sp* groups were also evident (electronic supplementary material, table S8). One haul to the north of the Hebrides (*6aN\_Sp\_19d*) comprised primarily maturing (Stage 2) and spawning (Stage 3) herring that assigned to the *6aS* and *6aS/6aN\_Sp* groups. In the south Minch area, there was a significant proportion of immature juvenile (Stage 1) individuals in one sample (*6aN\_Sp\_19h*), which assigned primarily to *6aS*. The other two hauls in this area (*6aN\_Sp\_19f* and *6aN\_Sp\_19g*) also had the majority of assignments to the *6aS/6aN\_Sp* groups.

The quarter 1 and 2 samples from Irish coastal waters were collected from by-catch of commercial vessels and comprised four samples, one sample from Lough Foyle (*6aS\_19b*), one from Lough Swilly (*6aS\_19a*) and two from Galway Bay (*6aS\_18b* and *6aS\_18c*) (figure 5; electronic supplementary material, table S3). The Lough Foyle sample contained mainly Stage 2 individuals with a small proportion of Stage 3 individuals and the Lough Swilly sample, which was caught 2 days earlier, contained predominately Stage 2 individuals and some spent (Stage 4) fish. The assignment of both samples was quite similar with the majority of individuals assigned to *6aS* and most of the remaining samples to *6aS/6aN\_Sp*. The Galway Bay samples were caught later in quarter 1 and 2 and had similar

maturity staging with primarily *Stage 4* and *Stage 5* individuals. The assignments of the two samples were also quite consistent with each other, with the majority of individuals assigned to *6aS*.

The quarter 3 non-baseline samples all came from Division 6.a.N and comprised samples from acoustic survey and monitoring fishery catches (see [37,55,56]). The maturity stages and length frequencies of the samples were notably different to the quarter 1 samples as there was a significant proportion of *Stage 1* and *2* fish (figure 5; electronic supplementary material, tables S3 and S4). The three samples from the Minch (*6aN\_18a*, *6aN\_19h* and *6aN\_19i*) primarily comprised *Stage 1* individuals, which assigned mainly to *6aS* and *6aS/6aN\_Sp*. The two samples from northwest of Cape Wrath (*6aN\_18b* and *6aN\_19e*) were composed of mainly *Stage 2* individuals and comprised a mix of *6aN\_Aut*, *6aS* and *6aS/6aN\_Sp*. The two samples from July (*6aN\_19f* and *6aN\_19g*), caught west of the Hebrides and north of Scotland, had similar maturities with mainly *Stage 2* individuals and a smaller proportion of *Stage 5* fish. The majority of individuals in the west of Hebrides sample were assigned to *6aS*. Conversely, the north of Scotland sample, which was caught adjacent to the 4° W line of longitude, contained a significant proportion of *6aN\_Aut* fish in addition to significant proportions of *6aS* and *6aS/6aN\_Sp* individuals.

The quarter 4 non-baseline samples all came from Divisions 6.a.S, 7.b and Lough Foyle and comprised samples from monitoring fishery catches and by-catch. The nine samples, caught over 5 years, contained a wide range of maturity stages, from *Stage 1* to *Stage 4*, yet the assignments were relatively consistent across all of the samples (figure 5; electronic supplementary material, tables S3 and S8). The five Division 6.a.S samples contained mainly a mixture of *Stage 2* and *Stage 4* individuals with a small number of *Stage 3* fish and the majority of individuals were assigned to *6aS* in all samples (electronic supplementary material, table S8). The Division 7.b sample (*6aS\_19d*) contained *Stage 1* and *Stage 2* fish and the assignments followed the same pattern as the Division 6.a.S samples. The three Lough Foyle samples were collected in three different years and had markedly different maturity stages, with one sample (*6aS\_17f*) dominated by *Stage 3* fish, one sample (*6aS\_18e*) predominately *Stage 2* and the third sample (*6aS\_20b*) a mixture of *Stages 1–4* fish (figure 5; electronic supplementary material, table S3). Regardless, the assignment outputs were broadly similar between the three samples and with the Division 6.a.S, 7.b samples, apart from a higher proportion of *6aS/6aN\_Sp* fish in two of the samples (electronic supplementary material, table S8).

Similar to the WESTHER samples, the average proportion of unassigned individuals increased significantly with increasing assignment threshold for *Approach 1-Level 1* and in particular for *Approach 1-Level 2* (table 5). This was not as significant in *Approach 2*, where there was a minor increase in the proportion of unassigned individuals at increasing *Level 1* thresholds and almost no increase in the *Level 2* assignments. Analysis of the individual rather than average values (electronic supplementary material, table S9) showed differences between the individual samples. The non-baseline quarter 1 samples from Division 6.a.N displayed the highest level of unassigned individuals, particularly at the *Approach 1-Level 2* assignment. The *Approach 2* assignment of the majority of samples resulted in a lower incidence of unassigned individuals at all threshold levels.

## 4. Discussion

The genetic markers and assignment methods presented in the current study constitute a ‘tool box’ that can be used for the assignment of herring caught in mixed survey and commercial catches in Division 6.a into their population of origin with a high level of accuracy (greater than 90%). This will enable the population assignment of commercial catch and acoustic survey (e.g. MSHAS) samples, which will facilitate the development of separate stock assessments for the populations in this area.

Both assignment approaches had a high level of self-assignment accuracy, though it was notable that there was a higher power to discriminate between the groups in *Level 1* than the groups in *Level 2*. In *Approach 1-Level 2*, there was a weakness in the model in discriminating between the *6aN\_Sp* samples and the quarter 1 late-spawning *6aS* samples. This was due in part to the small number of samples of *6aN\_Sp* in particular and potentially also to the genetic markers used not being optimized for distinguishing between these groups. There are inherent difficulties in sampling these groups for which no specific fishery currently exists and which spawn in areas that are subject to unfavourable weather conditions for sampling at the time of spawning. Further sampling of both groups is required though, and efforts are also under way to conduct WGS on representative samples from these groups to identify more informative markers. In any case, *Approach 2* mitigated this issue by combining them into a single group (*6aS/6aN\_Sp*<sup>A2L2</sup> (*Cluster\_2*)) for assignment purposes, which resulted in higher

classification success and lower rate of unassigned individuals. In this case, the majority of *6aS* fish could be separated from the other two populations with a high level of accuracy and only the minority of *6aS* fish were left in an unsorted mix with *6aN\_Sp*. This is considered preferable to the outcome of *Approach 1*, as for the purposes of splitting catch and survey samples, it is better to be able to assign some level of grouping rather than have a high abundance of unassigned individuals. Mixed categories such as *6aS/6aN\_Sp*<sup>A2L2</sup> can be acknowledged in the overall abundance estimates but retained in a separate category that is not allocated to a single stock, which can also act as a precautionary buffer for any undetected mis-assignments. In each assignment approach, at least 60% of the 45 markers were required to ensure accurate self-assignment, which indicates that there is a level of redundancy built into the panel of markers as was expected given that the markers are distributed among 14 loci. This redundancy is an advantage when analysing unknown samples as it allows up to 40% missing data in the genotypes of individuals. Missing genotypes may occur when analysing suboptimal quality samples collected from commercial catches or when analysing older archive samples, which have not been stored under optimum conditions.

The genetic assignment of the archive WESTHER samples confirmed longer term temporal stability of the SNP panel in the Division 6.a populations over a period of at least 18 spawning seasons, which is a temporally relevant time scale for the purposes of stock assessment. While the 2003 and 2004 *6aN\_Aut* WESTHER samples assigned near perfectly to the *6aN\_Aut* groups, the assignments of the 2003 and 2004 *6aS* WESTHER samples were not as confident. There were a significant number of mis-assignments to *6aN\_Aut* groups, particularly in the 2003 *6aS* sample. This spawning sample was collected in October which is earlier than any of the contemporary samples in the *6a baseline dataset*, and it is possible that these misassigned individuals shared some genetic similarities related to spawning time with the *6aN\_Aut* autumn spawning herring. Historically, autumn spawning herring were abundant in Division 6.a.S and particularly in Division 7.b where they supported local fisheries (see [8,14]); however, no autumn spawning was observed or sampled in this area in the course of the current study (i.e. since 2014). In fact, no spawning baseline samples were collected in Division 7.b throughout the study despite repeated sampling attempts, suggesting that either spawning in that area is at a very low level or has not occurred in recent years. However, the non-spawning herring caught in Division 7.b, genetically assigned with a high probability to *6aS*. Continued efforts should be made to ensure any spawning activity in Division 7.b is sampled if it occurs and the data added to the baselines.

The assignment of the non-baseline samples also provided an additional layer of validation of the assignment model and an interesting exploratory analysis of potential mixing of the different populations in Division 6.a. The Division 6.a.S and 7.b sample assignments were relatively consistent across all quarters indicating stability in the composition of herring shoals in the area. In all samples, a minority proportion of individuals were assigned to *6aN\_Aut* though this was mostly in keeping with the expected error rate of the assignment model, which was higher for *6aS* and *6aS/6aN\_Sp* than for *6aN\_Aut*. The samples from Lough Foyle were shown to be genetically and biologically the same as the *6aS* samples underlining the inappropriateness of the existing classification of Lough Foyle as part of the *6aN\_Aut* stock.

While there was consistency in the assignment of the samples collected in Division 6.a.S and 7.b, the assignment of those from Division 6.a.N (excluding Lough Foyle) indicated a significant degree of mixing of different populations. *6aS* and mixed *6aS/6aN\_Sp* herring comprised a varying but significant proportion of all samples and were far in excess of the expected error rate. The assignment of the juvenile samples from the Minch primarily to *6aS*, *6aN\_Sp*, *6aS/6aN\_Sp* instead of *6aN\_Aut* was not unexpected given the existing knowledge about the larval drift in the area and the lack of differentiation between the *6aN\_Aut* and North Sea autumn spawning herring. It is well documented that the larvae of autumn spawning herring off the northwest of Scotland are carried in easterly flowing currents and spend their juvenile phase in the North Sea [57–61]. The mixed nature of the samples collected off Cape Wrath during the spawning season for *6aN\_Aut* herring indicate a need for ongoing monitoring of survey and commercial catches in this area as any future fisheries in this area will probably be mixed stock fisheries. The presence of *6aN\_Sp* herring in the samples was also of interest as this population used to be the dominant population in the region but was reported to have collapsed in the 1950s [62]. Despite this, spring spawning herring were still known to comprise up to 38% of the catches off the north of Scotland, west of the 4° W line of longitude and in the North Minch in the 1960s [63]. However, as the autumn spawning component was more abundant, the newly developing stock assessments at the time were restricted to that group and the spring spawning herring were not distinguished, which over time led to them being merged with the autumn spawners for assessment purposes. The results of the current study suggest the spring

spawners are still present in the area. It is not currently possible to separate them from the late-spawning *6aS* herring so no conclusions can be drawn about their relative abundance, but further efforts should be directed towards improving the sampling of this population. The two samples collected in July west of the Hebrides and North of Scotland also offer some insight into the future assignments of the MSHAS samples that are collected during this period. The west of Hebrides sample comprised primarily *6aS* individuals with a smaller proportion of *6aS/6aN\_Sp* fish (figure 5), while the north of Scotland sample, which was caught adjacent to the 4° W line of longitude, contained a significant proportion of *6aN\_Aut* fish in addition to significant proportions of *6aS*, *6aN\_Sp* and *6aS/6aN\_Sp* individuals. Therefore, the current approach of splitting the MSHAS data using the 56° N line of latitude and the 7° W line of longitude to delineate the Division 6.a stocks is inappropriate and should be replaced with the genetic assignment approach.

One weakness of the assignment model in the current study is that it is solely based on the populations empirically proven to occur within Division 6.a (i.e. *6aS*, *6aN\_Aut* and *6aN\_Sp*) and does not include adjacent populations. The initial genetic analyses of the *full baseline dataset* in the current study and those in Han *et al.* [35] demonstrate that the Irish Sea herring and the Celtic Sea herring are distinct from each other and from the other populations in Division 6.a. However, they are genetically closely related to the herring in Division 6.a.S and as such it is difficult to distinguish them with a high degree of certainty using the current marker panel. The inclusion of these populations in the baseline dataset would increase the overall uncertainty of the assignments. Despite the assertions of the WESTHER project [9], there is no definitive evidence that a significant abundance of herring from either of these populations migrate to Division 6.a (see [8]). Therefore, their inclusion in the baseline datasets is not warranted at this time. The WESTHER project provided an illustration of the dangers of including multiple populations in a baseline when the power of discrimination between the populations is low. The inevitable outcome is that mixed samples will be weakly assigned and will have a high rate of mis-assignment. This can lead to the incorrect conclusion that mixed samples come from a larger number of source populations when the converse may be true. In the current study, there is the potential to misassign individuals from the Celtic Sea and Irish Sea populations, if they were present in Division 6.a; however, the assignment in its current form is still a significant improvement on the existing method of splitting the stocks based on geographical delineation, which is proven to be inappropriate. Efforts should be made to identify further population-specific genetic markers that may increase the discriminatory power between closely related populations. For this reason, the current marker panel should be considered the best available at the current time, but continued efforts should be made to develop it further.

The current study has also highlighted some of the potential stock identification issues that are apparent with the North Sea herring. The lack of differentiation between the *6aN\_Aut* herring and the North Sea autumn spawning herring suggests that the 4° W line of longitude is also inappropriate as a stock delineator. Though this has been recognized since its inception, as Saville & Bailey [11] noted, '*the dividing line between VIa and the North Sea (sub-area IV) at 4° W longitude was not chosen on any criterion of herring stock differentiation but for convenience in statistics collection*'. Further, the current study has demonstrated that *6aS* herring may be found up to at least the 4°W line of longitude and Farrell *et al.* [8] demonstrated the uncertainty in the composition of HERAS hauls in close proximity east of this line. The winter spawning Downs herring have also been shown to be a distinct and separate population to the North Sea autumn spawning herring and are relatively easily distinguished with the genetic markers in the current panel. The extent of the distribution of Downs herring in the North Sea area and their abundance in the HERAS or in the commercial catches in Divisions 4.a and 4.b are currently unknown. There are also known and demonstrated issues of mixing of the North Sea autumn spawning herring with Western Baltic herring to the east [64] and with Norwegian spring spawning herring to the North [65]. What is required now is a cohesive effort to study all of these stock identification issues, and those in the current study, together rather than treating them all separately. The ideal scenario may be to develop a universal marker panel that can discriminate all of the populations that could potentially be surveyed or caught in the Northeast Atlantic area (FAO Major Fishing Area 27). In theory, this would solve some of the issues outlined above; it would, however, also create a cost-benefit issue relating to the use of more extensive and expensive panels. In order to differentiate a wider range of populations, including those in the Baltic Sea, the panel would certainly need to comprise a larger number of genetic markers. The markers that may be suitable for discriminating between some of the Baltic Sea populations would probably not be informative for the populations west of Ireland and Britain (see [35]). Therefore, using a universal panel of markers on a sample caught to the west of Ireland and Britain, which is highly unlikely to contain any Baltic

herring, would represent a degree of wasted resources. If the universal panel was used on a sample caught in the eastern North Sea, then the presence of the Baltic markers may actually be beneficial as there is the potential for some Baltic Sea populations to be present in this area. The difficulty arises in defining the cut-off points on where the use of the universal panel is justified and where it is wasteful. Such a definition is akin to delineating stocks based on geographical or statistical areas such as ICES Divisions and inevitably introduces an element of subjectivity that may bias the results. It also introduces issues concerning the temporal stability of such definitions in an era of changing environmental conditions and documented changes in species distributions. Therefore, the use of a universal or specific marker panel is a topic that requires very careful consideration and rigorous empirical testing.

Implementation of a universal marker panel would also necessitate the development and implementation of a standard assignment approach across multiple jurisdictions. The use of the SVM model in *assignPOP* in the current study was favoured over traditional methods that rely on genotypic frequency distribution, as it was not constrained by underlying assumptions of HWE and linkage equilibrium. Further, it enabled a transparent and reproducible approach that could be clearly understood by non-geneticists, which is important if genetic stock identification methods are to be accepted and widely implemented as part of standard fisheries data collection protocols. Though not used in the current study, *assignPOP* also allows for the use of non-genetic markers either as stand-alone assignment models or in combined models with genetic data. This option may be attractive to institutes that have long time series of morphometric and meristic-based stock identification data as the transition to genetic-based methods can be made easier with direct comparison of the different datasets that are capable of being conducted within the same analyses.

The SNP panel deployed in the current study was composed of adaptive markers that are known to be under diversifying selection and proven to be associated with local ecological adaptation [22,35]. Genetic markers associated with loci under selection have been proven to provide better resolution to distinguish population structure in herring than neutral genetic markers [35,66]. However, such high-graded adaptive markers may undergo more rapid changes in allele frequencies within populations than putatively neutral genetic markers, particularly in situations of dynamic environmental conditions [23,67]. In the current study, the contemporary baseline spawning samples collected from 2014 to 2021 (seven spawning seasons) indicated temporal stability of the genetic markers within the different populations. Thus, these SNPs were appropriate for the purposes of stock identification in the current study. However, it is advisable to continue to collect and analyse baseline spawning samples regularly to monitor any changes in allele frequencies within the populations in the assignment model in order to prevent erroneous assignments of mixed samples. This also raises the question of how long to retain samples in the baseline dataset. As populations evolve and respond to changing conditions, older samples may become less relevant as baseline samples and may not represent the populations in their current state, which may have a negative impact on the assignments. Thus, baseline samples should perhaps be limited to a time scale that is relevant to the current population e.g. the longevity of the species. Given the relatively short history of effective genetic stock identification this has not been an issue to date but should be considered now as genetic stock identification starts to become an important part of marine fish stock assessment.

To date, there are few examples of genetic stock assignment being used for the regular assignment of survey or catch data of marine fish into population of origin for the purposes of stock assessment. These methods have primarily been used for one-off studies, that at best have been used to inform management, but few have been developed for regular monitoring and data collection [4,19]. Genetic stock identification methods have been most commonly used for salmonids, including Atlantic salmon, *Salmo salar* [68] and species of Pacific salmon including coho salmon, *Oncorhynchus kisutch* [69]. In these studies, self-assignment accuracies of 70–80% were concluded to be acceptable levels of accuracy. The high level of self-assignment accuracy in the current study (greater than 90%) exceeds this and is maintained even in the event of a considerable number of missing genotypes per individual.

If implemented as part of regular data collection on the MSHAS, the genetic stock identification method in the current study will enable the splitting of the survey indices into their constituent Division 6.a populations, which has not previously been possible. As a result, it will be possible to develop a separate stock assessment for the Division 6.a.S, 7.b–c stock. Although it should be noted that as there were no spawning herring observed or sampled in Divisions 7.b and 7.c, it was not possible to test the assumption that the herring that spawn in these divisions are the same population as the 6.a.S herring. The apparent lack of differentiation between the 6.a.N<sub>Aut</sub> herring and the North Sea autumn spawning herring also raises the question of whether it is appropriate to conduct a stand-

alone assessment on the *6a*N\_Aut herring or whether it should be combined with the North Sea autumn spawning herring assessment, though it is beyond the remit of the current study to make this recommendation. What is clear is that the results of the current study have improved the capacity to delineate, survey and assess the herring stocks in Division 6.a, and there is a need now to translate this into improved management.

**Ethics.** The project was considered by the UCD Animal Research Ethics Committee. No official ethical approval was required as all samples were collected opportunistically from fish caught during national fisheries surveys and commercial fisheries. No fish were killed for the purpose of the study and no specific licences were required.

**Data accessibility.** All data are provided in the manuscript, electronic supplementary material [70] and in the Dryad dataset Farrell *et al.* [71], which also includes the genotyping scripts (<https://doi.org/10.5281/zenodo.6979477>). R functions used in the manuscript are available at <https://cran.r-project.org/web/packages/adeqenet/index.html>, <https://cran.r-project.org/web/packages/assignPOP/index.html> and <https://cran.r-project.org/web/packages/mapplots/index.html>.

**Authors' contributions.** E.D.F.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft and writing—review and editing; L.A.: formal analysis, investigation, methodology, resources, validation, visualization and writing—review and editing; D.B.: formal analysis, investigation, methodology, resources, validation, visualization and writing—review and editing; N.C.: formal analysis, funding acquisition, investigation, project administration, validation, visualization and writing—review and editing; J.C.: conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, validation, visualization and writing—review and editing; M.W.C.: conceptualization, formal analysis, funding acquisition, investigation, project administration, resources, validation, visualization and writing—review and editing; A.E.: formal analysis, investigation, methodology, validation, visualization and writing—review and editing; A.F.: formal analysis, investigation, methodology, validation, visualization and writing—review and editing; M.G.: formal analysis, investigation, methodology, project administration, validation, visualization and writing—review and editing; S.M.L.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, validation, visualization and writing—review and editing; S.M.: data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, validation, visualization and writing—review and editing; C.N.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, validation, visualization and writing—review and editing; S.O.: data curation, formal analysis, investigation, methodology, resources, validation, visualization and writing—review and editing; M.O.: data curation, formal analysis, investigation, methodology, resources, validation, visualization and writing—review and editing; M.E.P.: data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, validation, visualization and writing—review and editing; M.P.: formal analysis, investigation, methodology, resources, validation, visualization and writing—review and editing; E.W.: data curation, formal analysis, investigation, methodology, resources, validation, visualization, writing—original draft and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

**Conflict of interest declaration.** To ensure transparency we state that subsequent to the completion of the project and manuscript, E.D.F. has taken up a role as Chief Scientific and Sustainability Officer with the Killybegs Fishermen's Organisation, Ireland. M.P. and S.M. are also employed as scientists with fishing industry organizations.

**Funding.** The work in the current study was part funded (2018–2020) by the European Commission's Executive Agency for Small and Medium-sized Enterprises (EASME) under Service Contract EASME/EMFF/2017/1.3.2.1/SI2.767459. The authors would also like to thank the Killybegs Fishermen's Organisation, Pelagic Freezer Trawler Association, Scottish Pelagic Fishermen's Association, Marine Institute, Marine Scotland Science, Pelagic Advisory Council and the Northern Pelagic Working Group of the European Association of Fish Producers Organisations for continued financial and ancillary support throughout the duration of the project (2015–2022).

**Acknowledgements.** We are also grateful to all the scientists and crew of the RVs '*Celtic Explorer*', '*Scotia*' and '*Corystes*' and all of the commercial vessels and scientists that have helped with the collection of samples. We acknowledge the help and support provided by Kuan-Yu 'Alex' Chen with the tuning of the assignPOP model. Finally, we acknowledge the valuable contribution of the GENSINC project (GENetic adaptations underlying population Structure IN herring; Research Council of Norway project 254774), which identified the informative markers that made the current study possible.

## References

1. Cadrin SX, Secor DH. 2009 Accounting for spatial population structure in stock assessment: past, present and future. In *The future of fishery science in North America* (eds RJ Beamish, BJ Rothschild), pp. 405–425. Dordrecht, The Netherlands: Springer.
2. Anon 2014 *Proposal for a Council Regulation fixing for 2015 the fishing opportunities for certain fish stocks and groups of fish stocks, applicable in Union waters and, for Union vessels, in certain non-Union waters and repealing Council Regulation (EU)*

- No 779/2014. European Commission, Brussels, 28.10.2014, COM.
3. Kerr LA, Hintzen NT, Cadrin SX, Worsøe Clausen L, Dickey-Collas M, Goethel DR, Hatfield EMC, Kritzer JP, Nash RDM. 2016 Lessons learned from practical approaches to reconcile mismatches between biological population structure and stock units of marine fish. *ICES J. Mar. Sci.* **74**, 1708–1722. (doi:10.1093/icesjms/fsw188)
  4. Reiss H, Hoarau G, Dickey-Collas M, Wolff WJ. 2009 Genetic population structure of marine fish: mismatch between biological and fisheries management units. *Fish Fish.* **10**, 361–395. (doi:10.1111/j.1467-2979.2008.00324.x)
  5. Stephenson RL. 2002 Stock structure management and structure: an ongoing challenge for ICES. *ICES Mar. Sci. Symp.* **215**, 305–314.
  6. Hintzen NT, Roel B, Benden D, Clarke M, Egan A, Nash RDM, Rohlf N, Hatfield EMC. 2015 Managing a complex population structure: exploring the importance of information from fisheries-independent sources. *ICES J. Mar. Sci.* **72**, 528–542. (doi:10.1093/icesjms/fsv102)
  7. Casey J, Jardim E, Martinsohn JTH. 2016 The role of genetics in fisheries management under the E.U. common fisheries policy. *J. Fish Biol.* **89**, 2755–2767. (doi:10.1111/jfb.13151)
  8. Farrell ED *et al.* 2021 *Herring in Divisions 6.a, 7.b and 7.c: scientific assessment of the identity of the Southern and Northern stocks through genetic and morphometric analysis*. Service Contract EASME/EMFF/2017/1.3.2.1/SI2.767459. p. 262. ISBN 978-92-9460-583-2 (doi:10.2826/208498)
  9. Hatfield EMC *et al.* 2005 (*WESTHER, Q5RS-2002-01056*): a multidisciplinary approach to the identification of herring (*Clupea harengus L.*) stock components west of the British Isles using biological tags and genetic markers. Final Report, p. 320.
  10. McQuinn IH. 1997 Metapopulations and the Atlantic herring. *Rev. Fish Biol. Fish.* **7**, 297–329. (doi:10.1023/A:1018491828875)
  11. Saville A, Bailey RS. 1980 The assessment and management of the herring stocks in the North Sea and to the west of Scotland. *Rapport et Procès-Verbaux des Reunions* **177**, 7–36.
  12. Dickey-Collas M *et al.* 2010 Lessons learned from stock collapse and recovery of North Sea herring: a review. *ICES J. Mar. Sci.* **67**, 1875–1886. (doi:10.1093/icesjms/fsq033)
  13. Simmonds EJ. 2009 Evaluation of the quality of the North Sea herring assessment. *ICES J. Mar. Sci.* **66**, 1814–1822. (doi:10.1093/icesjms/fsp104)
  14. ICES. 2015 *Report of the Benchmark Workshop on West of Scotland herring (WKWEST 2015). 2–6 February 2015*, Marine Institute, Dublin, Ireland, p. 299.
  15. ICES. 2014 *Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG), 11–20 March 2014*, ICES HQ, Copenhagen, Denmark. ICES CM 2014/ACOM:06, p. 1257.
  16. ICES. 1982 *Report of the Herring Assessment Working Group for the Area South of 62°N (HAWG)*. ICES CM 1982/Assess:7, p. 132.
  17. Manel S, Gaggiotti OE, Waples RS. 2005 Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol. Evol.* **20**, 136–142. (doi:10.1016/j.tree.2004.12.004)
  18. Bernatchez L *et al.* 2017 Harnessing the power of genomics to secure the future of seafood. *Trends Ecol. Evol.* **32**, 665–680. (doi:10.1016/j.tree.2017.06.010)
  19. Waples RS, Punt AE, Cope JM. 2008 Integrating genetic data into management of marine resources: how can we do better? *Fish Fish.* **9**, 423–449. (doi:10.1111/j.1467-2979.2008.00303.x)
  20. Hemmer-Hansen J, Therikildsen NO, Pujolar JM. 2014 Population genomics of marine fishes: next-generation prospects and challenges. *Biol. Bull.* **227**, 117–132. (doi:10.1086/BBLv227n2p117)
  21. Davey JW, Hohenlohe PA, Etter PD, Boone JO, Catchen JM, Blaxter ML. 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* **12**, 499–510. (doi:10.1038/nrg3012)
  22. Martinez Barrio A *et al.* 2016 The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, 12081. (doi:10.7554/eLife.12081)
  23. Nielsen EE *et al.* 2012 Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nat. Commun.* **3**, 851. (doi:10.1038/ncomms1845)
  24. Mariani S, Bekkevold D. 2014 The nuclear genome: neutral and adaptive markers in fisheries science. In *Stock identification methods* (eds S. X. Cadrin, L. A. Kerr, S. Mariani), pp. 397–428, 2nd edn. San Diego, CA: Elsevier, p. 566.
  25. Hemmer-Hansen J *et al.* 2019 Genetic analyses reveal complex dynamics within a marine fish management area. *Evol. Appl.* **12**, 830–844. (doi:10.1111/eva.12760)
  26. Rannala B, Mountain JL. 1997 Detecting immigration by using multilocus genotypes. *Proc. Natl Acad. Sci. USA* **94**, 9197–9221. (doi:10.1073/pnas.94.17.9197)
  27. Paetkau D, Calvert W, Stirling I, Strobeck C. 1995 Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* **4**, 347–354. (doi:10.1111/j.1365-294X.1995.tb00227.x)
  28. Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M. 1999 New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **15**, 1989–2000. (doi:10.1093/genetics/153.4.1989)
  29. Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A. 2004 geneClass2: a software for genetic assignment and first-generation migrant detection. *J. Hered.* **95**, 536–539. (doi:10.1093/jhered/esh074)
  30. Pritchard JK, Stephens M, Donnelly P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959. (doi:10.1093/genetics/155.2.945)
  31. Jombart T, Pontier D, Dufour AB. 2009 Genetic markers in the playground of multivariate analysis. *Heredity* **102**, 330–341. (doi:10.1038/hdy.2008.130)
  32. Schrider DR, Kern AD. 2018 Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* **34**, 301–312. (doi:10.1016/j.tig.2017.12.005)
  33. James G, Witten D, Hastie T, Tibshirani T. 2013 *An introduction to statistical learning*, vol. 112. Berlin, Germany: Springer.
  34. Chen K, Marshall EA, Sovic MG, Fries AC, Gibbs HL, Ludsin SA. 2018 assignPOP: an R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. *Methods Ecol. Evol.* **9**, 439–446. (doi:10.1111/2041-210X.12897)
  35. Han F *et al.* 2020 Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife* **2020**, e61076. (doi:10.7554/eLife.61076)
  36. ICES. 2011 *Report of the Workshop on Sexual Maturity Staging of Herring and Sprat (WKMSHS), 20–23 June 2011, Charlottenlund, Denmark*. ICES CM 2011/ACOM:46, p. 143.
  37. Mackinson S *et al.* 2021 *The 2020 industry-science survey of herring in the Western British Isles (ICES div 6a, 7bc)*. Survey report, p. 122.
  38. Lamichhane S *et al.* 2012 Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc. Natl Acad. Sci. USA* **109**, 19–345–19 350. (doi:10.1073/pnas.1216128109)
  39. Pettersson ME *et al.* 2019 A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. *Genome Res.* **29**, 1919–1928. (doi:10.1101/gr.253435.119)
  40. Vartia S *et al.* 2016 A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. *R. Soc. Open Sci.* **3**, 150565. (doi:10.1098/rsos.150565)
  41. Farrell ED, Carlsson JEL, Carlsson J. 2016 Next Gen Pop Gen: implementing a high-throughput approach to population genetics in boarfish (*Capros aper*). *R. Soc. Open Sci.* **3**, 160651. (doi:10.1098/rsos.160651)
  42. Rozen S, Skaletsky H. 2000 Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386. (doi:10.1385/1-59259-192-2.365)
  43. Kearse M *et al.* 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649. (doi:10.1093/bioinformatics/bts199)
  44. Kaplinski L, Andreson R, Puurand T, Remm M. 2005 MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics* **21**, 1701–1702. (doi:10.1093/bioinformatics/bti219)
  45. Babraham. 2016 FastQC. See <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc>
  46. Campbell NR, Harmon SA, Narum SR. 2015 Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol. Ecol. Res.* **15**, 855–867. (doi:10.1111/1755-0998.13583)
  47. Rousset F. 2008 Genepop'007: a complete reimplementations of the Genepop software for Windows and Linux. *Mol. Ecol. Res.* **8**, 103–106. (doi:10.1111/j.1471-8286.2007.01931.x)

48. Dieringer D, Schlötterer C. 2003 Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Res.* **3**, 167–169. (doi:10.1046/j.1471-8286.2003.00351.x)
49. Rice WR. 1989 Analyzing tables of statistical tests. *Evolution* **43**, 223–225. (doi:10.1111/j.1558-5646.1989.tb04220.x)
50. Peakall R, Smouse PE. 2012 GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* **28**, 2537–2539. (doi:10.1093/bioinformatics/bts460)
51. Jombart T. 2008 adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403–1405. (doi:10.1093/bioinformatics/btn129)
52. Jombart T, Devillard S, Balloux F. 2010 Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94. (doi:10.1186/1471-2156-11-94)
53. Anderson EC. 2010 Assessing the power of informative subsets of loci for population assignment: standard methods are upwardly biased. *Mol. Ecol. Res.* **10**, 701–710. (doi:10.1111/j.1755-0998.2010.02846.x)
54. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang CC, Lin CC. 2015 *E1071: misc functions of the department of statistics, probability theory group (Formerly: E1071)*. TU Wien. See <https://rdrr.io/rforge/e1071/>.
55. Mackinson S *et al.* 2019 *The 2019 industry-science survey of herring in the Western British Isles (ICES div 6a, 7bc)*. Survey report. p. 96.
56. Mackinson S *et al.* 2019 *The 2018 industry-science survey of herring in the Western British Isles (ICES div 6a, 7bc)*. Survey report. p. 99.
57. Heath M. 1989 Transport of larval herring (*Clupea harengus* L.) by the Scottish Coastal Current. *Rapports et procès-verbaux des réunions / Conseil permanent international pour l'exploration de la mer* **191**, 85–91.
58. Heath M. 1990 Segregation of herring larvae from inshore and offshore spawning grounds in the north-western North Sea – implications for stock structure. *Neth. J. Sea Res.* **25**, 267–278. (doi:10.1016/0077-7579(90)90026-D)
59. Heath MR, MacLachlan PM, Martin JHA. 1987 Inshore circulation and transport of herring larvae off the north coast of Scotland. *Mar. Ecol. Prog. Ser.* **40**, 11–23.
60. MacKenzie K. 1985 The use of parasites as biological tags in population studies of herring (*Clupea harengus* L.) in the North Sea and to the north and west of Scotland. *ICES J. Mar. Sci.* **42**, 33–64. (doi:10.1093/icesjms/42.1.33)
61. Saville A, Morrison JA. 1973 *A reassessment of the herring stock to the west of Scotland*. ICES Pelagic Fish (Northern) Committee, CM 1973/H:24.
62. Baxter IG. 1958 The composition of the Minch herring stocks. *Rapports et procès-verbaux des réunions / Conseil permanent international pour l'exploration de la mer* **143**, 814–894.
63. Saville A. 1970 *The stock fished by the Scottish herring fishery to the west of Scotland*. ICES Pelagic Fish (Northern) Committee, CM19/H: 29.
64. Bekkevold D *et al.* Accepted. Genetic stock identification in Atlantic herring (*Clupea harengus*): genome analysis provides new tools for identifying management units and complex migration dynamics. *ICES J. Mar. Sci.*
65. Berg F, Slotte A, Johannessen A, Kvamme C, Clausen LW, Nash RDM. 2017 Comparative biology and population mixing among local, coastal and offshore Atlantic herring (*Clupea harengus*) in the North Sea, Skagerrak, Kattegat and western Baltic. *PLoS ONE* **12**, e0187374. (doi:10.1371/journal.pone.0187374)
66. Bekkevold D, Gross R, Arula T, Helyar SJ, Ojaveer H. 2016 Outlier loci detect intraspecific biodiversity amongst spring and autumn spawning herring across local scales. *PLoS ONE* **11**, e0148499. (doi:10.1371/journal.pone.0148499)
67. Jorde PE, Synnes AE, Espeland SH, Sodeland M, Knutsen H. 2018 Can we rely on selected genetic markers for population identification? Evidence from coastal Atlantic cod. *Ecol. Evol.* **8**, 12 547–12 558. (doi:10.1002/ece3.4648)
68. Gilbey J *et al.* 2016 Accuracy of assignment of Atlantic salmon (*Salmo salar* L.) to rivers and regions in Scotland and Northeast England based on single nucleotide polymorphism (SNP) markers. *PLoS ONE* **11**, e0164327. (doi:10.1371/journal.pone.0164327)
69. Beacham TD, Wallace C, Jonsen K, McIntosh B, Candy JR, Rondeau EB, Moore JS, Bernatchez L, Withler RE. 2020 Accurate estimation of conservation unit contribution to coho salmon mixed-stock fisheries in British Columbia, Canada, using direct DNA sequencing for single nucleotide polymorphisms. *Can. J. Fish. Aquat. Sci.* **77**, 1302–1315. (doi:10.1139/cjfas-2019-0339)
70. Farrell ED *et al.* 2022 A baseline for the genetic stock identification of Atlantic herring, *Clupea harengus*, in ICES divisions 6.a, 7.b–c. Figshare. (doi:10.6084/m9.figshare.c.6168324)
71. Farrell ED *et al.* 2022 A baseline for the genetic stock identification of Atlantic herring, *Clupea harengus*, in ICES Divisions 6.a, 7.b–c. Dryad Digital Repository. (doi:10.5061/dryad.bvq83bbkq)