bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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

[1] EDF Scientific Limited, Rathaha, Ladysbridge, Cork, Ireland.

[2] Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden.

[3] Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, USA.

[4] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden.

[5] National Institute of Aquatic Resources, Technical University of Denmark, Silkeborg, Denmark.

[6] Marine Scotland Science, 375 Victoria Rd, Aberdeen AB11 9DB, Scotland.

[7] Area 52 Research Group, School of Biology and Environmental Science/Earth Institute, Science Centre West, University College Dublin, Dublin, Ireland.

[8] Marine Institute, Rinville, Oranmore, Co. Galway, Ireland.

[9] Department of Biological Sciences, University of Bergen, Bergen, Norway.

[10] European Commission, Joint Research Centre (JRC), Ispra, Italy.

[11] National Institute of Aquatic Resources, Willemoesvej 2, Hovedbygning, 067, 9850 Hirtshals, Denmark.

[12] Scottish Pelagic Fishermen's Association, Heritage House, 135-139 Shore Street, Fraserburgh, Aberdeenshire, Scotland.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[13] Pelagic Freezer-trawler Association, Louis Braillelaan 80, 2719 EK Zoetermeer, The Netherlands.

Corresponding Author: edward.d.farrell@gmail.com, ORCID ID: orcid.org/0000-0002-0070-9154

Abstract

Atlantic herring in 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 geographic 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 SNPs, derived from whole genome sequencing, were used to genotype 3,480 baseline spawning samples (2014-2021). A temporally stable baseline comprising 2,316 herring from populations known to inhabit Division 6.a was used to develop a genetic assignment method, with a self-assignment accuracy >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 (1,514 individuals) from Division 6.a, 7.b-c indicated previously unrecognised 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.

Keywords: fisheries, management, genetic assignment, stock identification, Northwest herring, West of Scotland Herring.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Fish stock identification has been an important prerequisite for fisheries stock assessment throughout its history (Cadrin & Secor, 2009). 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 (FAO Major Fishing Area 27) where the European Union (EU) defines the term 'stock' as 'a marine biological resource that occurs in a given management area' and delineates and names stocks using ICES (International Council for Exploration of the Sea) Statistical Areas (Anon, 2014). 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 (Kerr *et al*., 2016) and that biological populations are more dynamic and complex (Reiss *et al*., 2009; Stephenson, 2002).

Whilst 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 (Hintzen *et al*., 2015). 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 (Casey *et al*., 2016; Hintzen *et al*., 2015) 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

There is a long history of research into the characterisation 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 Farrell *et al*., 2021; Hatfield *et al*., 2005; McQuinn, 1997). Whilst 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 (Saville and Bailey, 1980) 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 (Dickey-Collas *et al*., 2010; Simmonds, 2009). The definition of the western herring stocks has changed considerably over the last five decades (see Farrell *et al*., 2021; ICES, 2015) and the main stocks are currently recognised as: 6.a.N (*6aN_Aut*); 6.a.S, 7.b and 7.c (*6aS*); Division 7.a North of 52°30'N (Irish Sea/*IS*); Divisions 7.a South of 52°30'N, 7.g, 7.h, 7.j and 7.k (Irish Sea, Celtic Sea, and southwest of Ireland, *CS*) (Figure 1; ICES, 2014). The *6aN_Aut* herring spawn in Autumn (Sept/Oct) off Cape Wrath on the north coast of Scotland, the *6aS* herring spawn in winter (Nov-Feb) primarily off the coast of Donegal in the northwest of Ireland, *IS* herring spawn in Autumn (Sept/Oct) mainly on the Douglas Bank east of the Isle of Man in the Irish Sea and Celtic Sea herring spawn in winter (Nov-Feb) off the south coast of Ireland. Several groups of spring spawning (Feb-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 Farrell *et al*., 2021). 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (ICES, 2015). Though some geographic 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 Farrell *et al*., 2021). 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 (ICES, 1982). 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 affinity to this 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 (*IS* and *CS*) are believed to form mixed aggregations on common feeding grounds in Division 6.a during summer (Hatfield *et al.*, 2005). 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 (ICES, 2015), 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

119 commercial catches. If these conditions are not met, uncertainty of the status of each of the individual
120 stocks increases, as does the risk that one stock may sustain higher fishing mortality than the other.

121 Genetic assignment methods, which compare genetic data from individuals to genetic profiles of
122 reference samples from potential source populations to determine population of origin (Manel *et al*.,
123 2005), offer the potential to resolve these issues. However, the incorporation of genetic assignment
124 methods into regular fisheries data collection, assessment and management has been slow
125 (Bernatchez *et al*., 2017; Reiss *et al*., 2009; Waples *et al*., 2008), as many existing genetic studies have
126 been hampered by high cost, inadequate sampling coverage, low numbers of suitable molecular
127 markers and low power to detect genetic structure. The advent of high-throughput sequencing (HTS)
128 technologies fundamentally changed the way in which genetic sequence data are generated (see
129 Hemmer-Hansen *et al*., 2014; Davey *et al*., 2010). It is now possible to generate large genomic data
130 sets for non-model species, which facilitate the identification of genetic loci with high discriminatory
131 power for resolving specific population differentiation questions (Martínez Barrio *et al*., 2016; Nielsen
132 *et al*., 2012). There has also been a shift toward the analysis of sequence variation of functional,
133 adaptive significance rather than just neutral DNA sequence variation (Mariani & Bekkevold, 2013).
134 This approach focuses on identifying adaptive markers that are under diversifying selection and may
135 reflect distinctive features of local populations (Nielsen *et al*., 2012). Small panels of high-graded
136 markers may be selected to develop efficient and cost-effective genetic assignment tools for informing
137 marine fisheries assessment and management (Hemmer-Hansen *et al*., 2018).

138 Assignment methods that attempt to solve classification problems rely on computing a discriminant
139 function based on samples from potential source populations and then classify unknown individuals
140 to the group with the highest discriminant score (Manel *et al*., 2005). Genetic assignment methods
141 have traditionally relied on using the genotypic frequency distribution under the assumption of
142 Hardy–Weinberg equilibrium (HWE) and linkage equilibrium in each source population as their
143 discriminant function (Manel *et al*., 2005). These genetic assignment methods can be broadly divided

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

into Bayesian (Rannala and Mountain, 1997), frequency (Paetkau *et al*., 1995) and distance based (Cornuet *et al*, 1999) methods (Hauser *et al*., 2006). 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 (Cornuet *et al*., 1999). These methods are commonly implemented in the software GeneClass2 (Piry *et al*., 2004). 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 (Manel *et al*., 2005). However, these Bayesian clustering analyses such as that implemented in the software *Structure* (Pritchard *et al*., 2000) 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 (Jombart *et al*., 2009). 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 (Schrider and Kern, 2018). 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 James *et al*., 2013). SVM models can also be used to classify non-linear data through use of non-linear kernels (James *et al*., 2013) and can be optimised 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* (Chen *et al*., 2018) 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Recent studies utilising 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 strong population structure inferred to be associated with a high level of local ecological adaptation (Han *et al*, 2020; Lamichhaney *et al*., 2017; Martinez Barrio *et al*., 2016). 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 (Han *et al*., 2020). From the circa 10 million Single Nucleotide Polymorphisms (SNPs) identified by Han *et al*. (2020) a small subset of circa 800 SNPs, associated with ecological adaptation to different geographic areas and spawning conditions, were shown to be able to discriminate all the sampled populations of herring 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 (Han *et al*., 2020). 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 *CS*, *IS* 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:

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(i) validate a small 'toolbox' of genetic markers, from those identified by Han *et al*. (2020), 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.

(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.

Materials and Methods

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 (MSS) were maturity staged using the 9-point scale, those processed by the Irish Marine Institute (MI) were maturity staged using the 8-point scale and samples processed by the Wageningen University and Research (WUR) on behalf of the Dutch Pelagic Freezer Trawler Association (PFA) were maturity staged using the ICES 6-point scale (ICES, 2011). All maturity stages were converted to the ICES 6-point scale according to Mackinson *et al*. (2021). A 0.5 cm^3 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; Hatfield *et al*., 2005). Total genomic DNA (gDNA) was extracted from c.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). Extracted DNA was stored in 96 well PCR plates at -20°C until ready for genotyping.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

219

220 **Genetic marker identification**

221 The SNPs used in the current study (ESM Table 1) were identified during the GENSINC project (GENetic
222 adaptations underlying population Structure IN herring; Research Council of Norway project 254774)
223 and were derived from the analyses of WGS of pooled samples from herring populations across the
224 species distribution, which was undertaken to study the biological significance of the genetic variants
225 underlying ecological adaptation in the Atlantic herring (Han *et al*., 2020; Lamichhaney *et al*., 2012;
226 Martinez Barrio *et al*., 2016; Pettersson *et al*., 2019). The subset of SNPs was selected following testing
227 of candidate SNPs with the highest delta allele frequency (dAF) values from the major genomic regions
228 of divergence in the contrasts between populations around Ireland and Britain. The 45 SNPs selected
229 were distributed across 8 chromosomes and comprised 14 loci (ESM Table 1). Linked SNPs were
230 retained in the panel in order to add a level of redundancy and ensure that key genomic regions (loci)
231 were well represented even in the instance of missing genotype data from an individual.

232

233 **Genotyping**

234 The majority of samples were genotyped utilising a genotyping by sequencing approach (Vartia *et al*.,
235 2016) described in detail and validated in Farrell *et al*. (2016; 2021). In short, locus-specific forward
236 and reverse primers were designed for SNP loci with the Primer3 application (Rozen and Skaletsky,
237 2000) in Geneious® 7.0 (Kearse *et al*., 2012) with optimal primer length set at 20bp and product size
238 range at 120-180bp. Primers were designed to bind in conserved flanking regions to minimise the
239 possibility of null alleles and were cross-referenced with existing genome sequence data to identify
240 primers that annealed to multiple regions, which if detected were excluded. The forward and reverse
241 locus-specific primers were adapted, to facilitate combinatorial barcoding of amplicons, by adding
242 either an M13-R (5'-GGAAACAGCTATGACCAT-3') or CAG (5'-CAGTCGGGCGTCATCA-3') universal tail to

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the 5' end and were divided into multiplex panels in MultiPLX 2.1 (Kaplinski *et al*., 2005). A set of ninety-six 11bp combinatorial barcodes were used to identify individuals within pooled sequencing runs. An M13-R universal tail was added to the 3' end of forty-eight of the barcodes and a CAG universal tail to the 3' end of the remaining forty-eight barcodes, yielding 2,304 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*. (2016 and 2021). 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 standardised 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 (Babraham, 2016). Reads were sorted and grouped using a modified python script (Vartia *et al*., 2016) 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

267 SNP genotyping was automated by using a modified Perl script from the Genotyping-in-Thousands by
268 sequencing (GT-seq) approach (Campbell *et al*., 2015), which counts amplicon-specific sequences for
269 each allele, and uses allele ratios to determine the genotypes. The Perl scripts were modified in the
270 current project to use the output of the custom python scripts as the input. The default settings of the
271 GT-Seq Perl script designated allele ratios >10.0 to be called as homozygous for allele 1, ratios <0.1 to
272 be called as homozygous for allele 2, and ratios between 0.2 and 5.0 to be called as heterozygous
273 (Campbell *et al*., 2015). These ratios were optimised for the data and markers in the current study by
274 analysing each marker separately and plotting the genotyping calls from which new ratios were
275 calculated for each marker. The average designated allele ratios in the current study were >5.0 to be
276 called as homozygous for allele 1, ratios <0.2 to be called as homozygous for allele 2, ratios between
277 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
278 NA (NA = no genotype call made). Individuals with less than 10 reads at a particular locus were also
279 designated as NA. Only individuals with greater than 89% genotyping success (i.e. 40/45 genotypes)
280 were retained in the dataset.

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

288

289 **Baseline dataset analyses**

290 It should be noted that the aim of the current study was not to undertake an exhaustive population
291 genetics and demographic study of the herring populations around Ireland and Britain but was to

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

develop a genetic based method to separate the herring caught in putatively mixed survey and commercial catches in ICES Divisions 6.a, 7.b and 7.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 maximise 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 Hardy–Weinberg equilibrium (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 (Rousset, 2008). Microsatellite Analyzer (MSA) 4.05 (Dieringer and Schlötterer, 2003) was used, under default settings, to assess multi-locus pairwise F_{ST} with 1,000 bootstrap replications and 10,000 permutations. In all cases with multiple tests, significance levels were adjusted using the sequential Bonferroni technique (Rice, 1989). In order to visualise the pairwise F_{ST} results and to explore the relationships between the different samples, Principal Coordinate Analysis (PCoA) using the covariance standardised method was conducted in GenAlEx 6.51b2 (Peakall and Smouse, 2012).

Discriminant analysis of Principal Components (DAPC), from the *R* package *adegenet* (Jombart, 2008; Jombart *et al*., 2010), is a multivariate approach that transforms multi-locus genotype data using 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, Hardy–Weinberg equilibrium), 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

317 combined to form seven groups (*6aS*, *CS*, *IS*, *6aN_Aut*, *6aN_Sp*, *NS*, *DWN*) which represented the
318 putative populations in the study area. Following further analyses (see Results) a reduced *6a baseline*
319 *dataset* consisting only of samples from groups that are confirmed as being present in Division 6.a i.e.
320 *6aS*, *6aN_Aut* and *6aN_Sp,* was analysed using DAPC. In this instance DAPC was conducted as before
321 with prior definition of group membership and also following a second approach using the *find.clusters*
322 function to infer genetic clusters. This function transforms the data using principal component analysis
323 (PCA), then runs the *K*-means algorithm (function *kmeans* from the *stats* package) with increasing
324 values of *K* and computes Bayesian Information Criterion (BIC) to assess the best supported model.

325

326 **Assignment model development**

327 The *R* package *assignPOP* (Chen *et al*., 2018), which performs population assignment using a machine-
328 learning framework, was used to develop the assignment model. *assignPOP* uses Monte-Carlo cross-
329 validation (*assign.MC*) to divide the baseline data into a training dataset and test dataset. The
330 assignment model is developed with the training dataset and subsequently tested with the
331 independent test dataset, which avoids introducing 'high-grading bias' (see Anderson, 2010). As the
332 Monte-Carlo procedure samples random individuals each time, it does not guarantee that every
333 individual is sampled. Therefore, *assignPOP* can perform an additional method of *K*-fold cross-
334 validation (*assign.kfold*), which involves randomly dividing the individuals from each population into
335 *K* groups and then using one group from each population as test individuals and the remaining *K-1*
336 groups as the training individuals. Assignment tests are performed until every group and hence
337 individual is tested, resulting in *K* tests. *assignPOP* has a number of classification model options
338 including the *SVM* model from the *R* package *e1071* (Meyer *et al*., 2015). Based on the results of the
339 aforementioned baseline dataset analyses it was decided to develop the assignment model in
340 *assignPOP* trained on the reduced *6a baseline dataset* as the constituent populations had the highest
341 level of discrimination between them with the current marker panel. As per the DAPC analyses, the

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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. The assignment was conducted at two hierarchical levels based on the power to discriminate the different groups in the DAPC analyses. In *Level 1 6aS* and *6aN_Sp* were combined and tested against *6aN_Aut*. In *Level 2* the combined *6aS/6aN_Sp* group was split, and the individual groups tested against each other. *Approach 2* was also performed in a hierarchical manner as per *Approach 1*. However, *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.

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* (Meyer *et al*., 2015) 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 were specified and were 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, whilst 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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.

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* (Chen *et al*., 2018) 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 alternate outcome. This was deemed an acceptable level of confidence given the high level of self-assignment accuracy of the baseline datasets. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

392

393 **Exploratory analyses with contemporary non-baseline samples**

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

400

401 **Results**

402 **Sampling and genotyping success rate**

403 Due to the opportunistic nature of the sampling, the samples contained a significant mix of length
404 classes and maturity stages (ESM Tables 2 and 3). In total 92 contemporary samples were collected
405 (Figure 1 and ESM Table 2), comprising 6,591 individual herring of which 5,638 individuals (86%)
406 passed the genotyping threshold of 89% (40/45 SNPs genotyped). All 45 SNPs were successfully
407 genotyped in more than 92% of retained individuals. For the purposes of developing robust baselines
408 for genetic assignment, it is critical to avoid including individuals with uncertain origin. Therefore, only
409 samples with a significant number of maturity stage three (spawning) individuals (ICES, 2011), caught
410 in close proximity to known spawning grounds at recognised spawning times were selected to be
411 baseline samples. In order to further limit the potential for misclassification as a baseline spawning
412 sample, only individuals classified as maturity stage three were included in the baseline dataset. The
413 resulting contemporary *baseline dataset* contained 64 samples, comprising a total of 3,480 herring
414 (Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The 28 remaining samples (1,514 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.

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), *6aS_19c* (6/45). Samples *6aS_17d* and *6aS_17e* also displayed indications of a significant heterozygote deficiency in eight and thirteen loci, respectively, which was likely 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 ten, seven and seven loci, respectively. The significant indications of LD were in keeping with the loci already identified (ESM Table 1) from Han *et al*. (2020). All markers and all samples were retained in the *baseline dataset* for further analyses.

The analyses of multi-locus pairwise F_{ST} (ESM Table 4) indicated proportionately higher F_{ST}'s and significant differentiation between samples collected from the different putative populations except between the *6aN_Aut* and *NS* 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 *IS* or *DWN* population and there was little if any significant genetic differentiation between the temporal samples from the *CS* and *6aN_Sp*. There were some

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (ESM Figure 1). Therefore, the temporal samples were combined to form seven groups (*6aS*, *CS*, *IS*, *6aN_Aut*, *6aN_Sp*, *NS*, *DWN*), 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 *baseline dataset* indicated significant differentiation between all baseline population groups (Table 1). The lowest level of differentiation was between the *6aN_Aut* and *NS* 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 (ESM Table 4). 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 *DWN* group and the *CS* and *IS* groups, whilst the *DWN* group had a high level of differentiation from the *NS* group.

The DAPC results indicated the same pattern of structure as the F_{ST} analyses (Figure 3) and also as those observed in Han *et al*. (2020) 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 *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*, *CS* and *IS* groups overlapped,

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

indicating that the current marker panel cannot be used to distinguish these groups with a high level of accuracy. Therefore, the *CS* and *IS* 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 (Farrell *et al*., 2021). DAPC also indicated an overlap and an inability to distinguish between the *6aN_Aut* and *NS* 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 (Farrell *et al*., 2021), however this distinction was not the focus of the current study, as such the *NS* samples were removed from further analyses. The *DWN* group was confirmed to be distinct from the *NS* group though it could not be reliably discriminated from the *CS* and *IS* groups with the current panel of markers and as such the *DWN* 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*.

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 (ESM Table 5). DAPC of the *clustered 6a baseline dataset* indicated clear division between the clusters with minimal overlap (Figure 3), 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 Farrell *et al*., 2021), 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

considered, for the purposes of the assignment, to represent a mix of *6aS* and *6aN_Sp* herring. The resulting *clustered 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 $6aN\_Aut^{A2}$, $6aS^{A2L2}$ and $6aS/6aN\_Sp^{A2L2}$, respectively.

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 (RMSE) 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 parameter. The optimum model and kernel for the assignment model were the SVM model and the radial basis function (RBF) 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* (ESM Figures 2 & 3 and 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* (ESM Figure 3 and 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 $6aS/6aN\_Sp^{A2L2}$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The Approach 1-Level 1 assignment was more sensitive to the number of loci *than Approach 1-Level 2* (ESM Figure 4 and ESM Table 6). 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 twenty-three of the forty-five 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.

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 and ESM Table 7). The assignment of the 2003 *6aS* WESTHER sample (*6aS_03*) was more uncertain, with a quarter of individuals misassigned to the *6aN_Aut* groups (ESM Table 7). The 2004 *6aS* sample (*6aS_04*) had a higher level of correct assignment (>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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Approach 2-Level 2 assignment provided a more robust assignment with 67% of individuals assigned to *6aS/6aN_Sp*^{A2L2}. 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 (ESM Table 2).

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 (ESM Table 8) 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 alternate assignment (ESM Table 8). 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 *Level 2* assignments all had a probability greater than 0.9 (ESM Table 8). 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.

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 (SWC-IBTS) and comprised a number of small samples of herring of mixed maturity stages (Figure 5 and ESM Tables 2 & 3). 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 (ESM Table 7). 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

565 *6aS/6aN_Sp* groups. In the south Minch area there was a significant proportion of immature juvenile
566 (*Stage 1*) individuals in one sample (*6aN_Sp_19h*), which assigned primarily to *6aS*. The other two
567 hauls in this area (*6aN_Sp_19f* and *6aN_Sp_19g*) also had the majority of assignments to the
568 *6aS/6aN_Sp* groups.

569 The quarter 1 and 2 samples from Irish coastal waters were collected from bycatch of commercial
570 vessels and comprised four samples, one sample from Lough Foyle (*6aS_19b*), one from Lough Swilly
571 (*6aS_19a*) and two from Galway Bay (*6aS_18b* and *6aS_18c*) (Figure 5 and ESM Table 2). The Lough
572 Foyle sample contained mainly *Stage 2* individuals with a small proportion of *Stage 3* individuals and
573 the Lough Swilly sample, which was caught two days earlier, contained predominately *Stage 2*
574 individuals and some spent (*Stage 4*) fish. The assignment of both samples was quite similar with the
575 majority of individuals assigned to *6aS* and most of the remaining samples to *6aS/6aN_Sp*. The Galway
576 Bay samples were caught later in quarter 1 and 2 and had similar maturity staging with primarily *Stage*
577 *4* and *Stage 5* individuals. The assignments of the two samples were also quite consistent with each
578 other with the majority of individuals assigned to *6aS*.

579 The quarter 3 non-baseline samples all came from Division 6.a.N and comprised samples from acoustic
580 survey and monitoring fishery catches (see Mackinson *et al*, 2019a; 2019b; 2021). The maturity stages
581 and length frequencies of the samples were notably different to the quarter 1 samples as there was a
582 significant proportion of *Stage 1* and *2* fish (Figure 5 and ESM Tables 2 & 3). The three samples from
583 the Minch (*6aN_18a, 6aN_19h* and *6aN_19i*) primarily comprised *Stage 1* individuals, which assigned
584 mainly to *6aS* and *6aS/6aN_Sp*. The two samples from northwest of Cape Wrath (*6aN_18b* and
585 *6aN_19e*) were composed of mainly *Stage 2* individuals and comprised a mix of *6aN_Aut*, *6aS* and
586 *6aS/6aN_Sp*. The two samples from July (*6aN_19f* and *6aN_19g*), caught west of the Hebrides and
587 north of Scotland, had similar maturities with mainly *Stage 2* individuals and a smaller proportion of
588 *Stage 5* fish. The majority of individuals in the west of Hebrides sample were assigned to *6aS*.
589 Conversely the north of Scotland sample, which was caught adjacent to the 4°W line of longitude,

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

590 contained a significant proportion of *6aN_Aut* fish in addition to significant proportions of *6aS* and
591 6aS/*6aN_Sp* individuals.

592 The quarter 4 non-baseline samples all came from Divisions 6.a.S, 7.b and Lough Foyle and comprised
593 samples from monitoring fishery catches and bycatch. The nine samples, caught over 5 years,
594 contained a wide range of maturity stages, from *Stage 1* to *Stage 4* yet the assignments were relatively
595 consistent across all of the samples (Figure 5 and ESM Tables 2 & 7). The five Division 6.a.S samples
596 contained mainly a mixture of *Stage 2* and *Stage 4* individuals with a small number of *Stage 3* fish and
597 the majority of individuals were assigned to *6aS* in all samples (ESM Table 7). The Division 7.b sample
598 (*6aS_19d*) contained *Stage 1* and *Stage 2* fish and the assignments followed the same pattern as the
599 Division 6.a.S samples. The three Lough Foyle samples were collected in three different years and had
600 markedly different maturity stages, with one sample (*6aS_17f*) dominated by Stage 3 fish, one sample
601 (*6aS_18e*) predominately *Stage 2* and the third sample (*6aS_20b*) a mixture of *Stages 1-4* fish (Figure
602 5 and ESM Table 2). Regardless, the assignment outputs were broadly similar between the three
603 samples and with the Division 6.a.S, 7.b samples, apart from a higher proportion of *6aS/6aN_Sp* fish
604 in two of the samples (ESM Table 7).

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

614

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (>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 optimised 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 underway 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*^{A2L2} *(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*^{A2L2} 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 misassignments. In each assignment approach at least 60% of the 45 markers were

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 fourteen 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 eighteen spawning seasons, which is a temporally relevant time scale for the purposes of stock assessment. Whilst 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 Farrell *et al*., 2021; ICES, 2015) 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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

665 populations in Division 6.a. The Division 6.a.S and 7.b sample assignments were relatively consistent
666 across all quarters indicating stability in the composition of herring shoals in the area. In all samples a
667 minority proportion of individuals were assigned to *6aN_Aut* though this was mostly in keeping with
668 the expected error rate of the assignment model, which was higher for *6aS* and *6aS/6aN_Sp* than for
669 *6aN_Aut*. The samples from Lough Foyle were shown to be genetically and biologically the same as
670 the *6aS* samples underlining the inappropriateness of the existing classification of Lough Foyle as part
671 of the *6aN_Aut* stock.

672 Whilst there was consistency in the assignment of the samples collected in Division 6.a.S and 7.b, the
673 assignment of those from Division 6.a.N (excluding Lough Foyle) indicated a significant degree of
674 mixing of different populations. *6aS* and mixed *6aS/6aN_Sp* herring comprised a varying but significant
675 proportion of all samples and were far in excess of the expected error rate. The assignment of the
676 juvenile samples from the Minch primarily to *6aS*, *6aN_Sp*, *6aS/6aN_Sp* instead of *6aN_Aut* was not
677 unexpected given the existing knowledge about the larval drift in the area and the lack of
678 differentiation between the *6aN_Aut* and North Sea autumn spawning herring. It is well documented
679 that the larvae of autumn spawning herring off the northwest of Scotland are carried in easterly
680 flowing currents and spend their juvenile phase in the North Sea (Heath, 1989; Heath, 1990; Heath *et*
681 *al*., 1987; MacKenzie, 1985; Saville and Morrison, 1973). The mixed nature of the samples collected
682 off Cape Wrath during the spawning season for *6aN_Aut* herring indicated a need for ongoing
683 monitoring of survey and commercial catches in this area as any future fisheries in this area will likely
684 be mixed stock fisheries. The presence of *6aN_Sp* herring in the samples was also of interest as this
685 population used to be the dominant population in the region but was reported to have collapsed in
686 the 1950s (Baxter, 1958). Despite this, spring spawning herring were still known to comprise up to
687 38% of the catches off the north of Scotland, west of the 4°W line of longitude and in the North Minch
688 in the 1960's (Saville, 1970). However, as the autumn spawning component was more abundant, the
689 newly developing stock assessments at the time were restricted to that group and the spring spawning
690 herring were not distinguished, which over time led to them being merged with the autumn spawners

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

691 for assessment purposes. The results of the current study suggest the spring spawners are still present
692 in the area. It is not currently possible to separate them from the late spawning *6aS* herring so no
693 conclusions can be drawn about their relative abundance, but further efforts should be directed
694 towards improving the sampling of this population. The two samples collected in July west of the
695 Hebrides and North of Scotland also offer some insight into the future assignments of the MSHAS
696 samples that are collected during this period. The west of Hebrides sample comprised primarily *6aS*
697 individuals with a smaller proportion of *6aS/6aN_Sp* fish (Figure 5), whilst the north of Scotland
698 sample, which was caught adjacent to the 4°W line of longitude, contained a significant proportion of
699 *6aN_Aut* fish in addition to significant proportions of *6aS*, *6aN_Sp* and *6aS/6aN_Sp* individuals.
700 Therefore, the current approach of splitting the MSHAS data using the 56°N line of latitude and the
701 7°W line of longitude to delineate the Division 6.a stocks is inappropriate and should be replaced with
702 the genetic assignment approach.

703 One weakness of the assignment model in the current study is that it is solely based on the populations
704 empirically proven to occur within Division 6.a (i.e. *6aS*, *6aN_Aut* and *6aN_Sp*) and does not include
705 adjacent populations. The initial genetic analyses of the *full baseline dataset* in the current study and
706 those in Han *et al*. (2020) demonstrate that the Irish Sea herring and the Celtic Sea herring are distinct
707 from each other and from the other populations in Division 6.a. However, they are genetically closely
708 related to the herring in Division 6.a.S and as such it is difficult to distinguish them with a high degree
709 of certainty using the current marker panel. Inclusion of these populations in the baseline dataset
710 would increase the overall uncertainty of the assignments. Despite the assertions of the WESTHER
711 project (Hatfield *et al*., 2005) there is no definitive evidence that a significant abundance of herring
712 from either of these populations migrate to Division 6.a (see Farrell *et al*., 2021). Therefore, their
713 inclusion in the baseline datasets is not warranted at this time. The WESTHER project provided an
714 illustration of the dangers of including multiple populations in a baseline when the power of
715 discrimination between the populations is low. The inevitable outcome is that mixed samples will be
716 weakly assigned and will have a high rate of misassignment. This can lead to the incorrect conclusion

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

717 that mixed samples come from a larger number of source populations when the converse may be true.
718 In the current study there is the potential to misassign individuals from the Celtic Sea and Irish Sea
719 populations, if they were present in Division 6.a, however the assignment in its current form is still a
720 significant improvement on the existing method of splitting the stocks based on geographic
721 delineation, which is proven to be inappropriate. Efforts should be made to identify further population
722 specific genetic markers that may increase the discriminatory power between closely related
723 populations. For this reason, the current marker panel should be considered the best available at the
724 current time, but continued efforts should be made to develop it further.

725 The current study has also highlighted some of the potential stock identification issues that are
726 apparent with the North Sea herring. The lack of differentiation between the *6aN_Aut* herring and the
727 North Sea autumn spawning herring suggests that the 4°W line of longitude is also inappropriate as a
728 stock delineator. Though this has been recognised since its inception, as Saville and Bailey (1980)
729 noted, '*the dividing line between VIa and the North Sea (sub-area IV) at 4°W longitude was not chosen*
730 *on any criterion of herring stock differentiation but for convenience in statistics collection*'. Further,
731 the current study has demonstrated that *6aS* herring may be found up to at least the 4°W line of
732 longitude and Farrell *et al*. (2021) demonstrated the uncertainty in the composition of HERAS hauls in
733 close proximity east of this line. The winter spawning Downs herring have also been shown to be a
734 distinct and separate population to the North Sea autumn spawning herring and are relatively easily
735 distinguished with the genetic markers in the current panel. The extent of the distribution of Downs
736 herring in the North Sea area and their abundance in the HERAS or in the commercial catches in
737 Divisions 4.a and 4.b are currently unknown. There are also known and demonstrated issues of mixing
738 of the North Sea autumn spawning herring with Western Baltic herring to the east (Bekkevold et al.,
739 In preparation) and with Norwegian Spring Spawning herring to the North (Berg et al., 2017). What is
740 required now is a cohesive effort to study all of these stock identification issues, and those in the
741 current study, together rather than treating them all separately. The ideal scenario may be to develop
742 a universal marker panel that can discriminate all of the populations that could potentially be surveyed

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 likely not be informative for the populations west of Ireland and Britain (see Han *et al*., 2020). 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 geographic 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 standalone assignment models or in combined models with genetic data. This option may be attractive

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

769 to institutes that have long time series of morphometric and meristic based stock identification data
770 as the transition to genetic based methods can be made easier with direct comparison of the different
771 data sets that are capable of being conducted within the same analyses.

772 The SNP panel deployed in the current study was composed of adaptive markers that are known to
773 be under diversifying selection and proven to be associated with local ecological adaptation (Han *et*
774 *al*., 2020; Martinez Barrio *et al*., 2016). Genetic markers associated with loci under selection have been
775 proven to provide better resolution to distinguish population structure in herring than neutral genetic
776 markers (Bekkevold *et al*., 2016; Han *et al*., 2020). However, such high-graded adaptive markers may
777 undergo more rapid changes in allele frequencies within populations than putatively neutral genetic
778 markers, particularly in situations of dynamic environmental conditions (Jorde *et al*., 2018; Nielsen *et*
779 *al*., 2012). In the current study the contemporary baseline spawning samples collected from 2014 to
780 2021 (seven spawning seasons) indicated temporal stability of the genetic markers within the different
781 populations. Thus, these SNPs were appropriate for the purposes of stock identification in the current
782 study. However, it is advisable to continue to collect and analyse baseline spawning samples regularly
783 to monitor any changes in allele frequencies within the populations in the assignment model in order
784 to prevent erroneous assignments of mixed samples. This also raises the question of how long to retain
785 samples in the baseline dataset. As populations evolve and respond to changing conditions, older
786 samples may become less relevant as baseline samples and may not represent the populations in their
787 current state, which may have a negative impact on the assignments. Thus baseline samples should
788 perhaps be limited to a time scale that is relevant to the current population e.g. the longevity of the
789 species. Given the relatively short history of effective genetic stock identification this has not been an
790 issue to date but should be considered now as genetic stock identification starts to become an
791 important part of marine fish stock assessment.

792 To date there are few examples of genetic stock assignment being used for the regular assignment of
793 survey or catch data of marine fish into population of origin for the purposes of stock assessment.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (Reiss *et al*., 2009; Waples *et al*., 2008). Genetic stock identification methods have been most commonly used for salmonids, including Atlantic Salmon, *Salmo salar* (Gilbey *et al*., 2016) and species of pacific salmon including coho salmon, *Oncorhynchus kisutch* (Beacham *et al*., 2020). 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 (>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 *6aS* herring. The apparent lack of differentiation between the *6aN_Aut* 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 *6aN_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.

Data accessibility

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

817 All data are provided in the manuscript and extra supplementary tables (ESM Table 1 – ESM Table 8)
818 and figures (ESM Figure 1 – ESM Figure 4) are provided in the file
819 *Farrell_et_al_6a_herring_baseline_ESM_tables_figures.xlsx.*
820
821 **Research Ethics, Animal Ethics**
822 No ethical approval was required. All samples were collected opportunistically from fish caught during
823 fisheries surveys and commercial fisheries. No fish were killed for the purpose of the study and no
824 licences were required.
825
826 **Competing interests**
827 We have no competing interests.
828
829 **Author's contributions**
830 Conceptualisation: EDF, JC, MWC, SML, CN. Data Curation: EDF, SML, SM, CN, SO'C, MO'M, MP, EW.
831 Formal analysis: EDF, LA, DB, NC, JC, MWC, AE, AF, MG, SML, SM, CN, SO'C, MO'M, MP, MEP, EW.
832 Funding acquisition: EDF, NC, JC, MWC, SML, SM, CN, MP. Investigation: EDF, LA, DB, NC, JC, MWC,
833 AE, AF, MG, SML, SM, CN, SO'C, MO'M, MP, MEP, EW. Methodology: EDF, LA, DB, JC, AE, AF, MG,
834 SML, SM, CN, SO'C, MO'M, MP, MEP, EW. Project administration: EDF, NC, JC, MWC, MG, SML, SM,
835 CN, MP. Resources: EDF, LA, DB, MWC, SML, SM, CN, SO'C, MO'M, MP, MEP, EW. Validation: EDF,
836 LA, DB, NC, JC, MWC, AE, AF, MG, SML, SM, CN, SO'C, MO'M, MP, MEP, EW. Visualisation: EDF, EW.
837 Writing-original draft: EDF, EW. Writing-review & editing: EDF, LA, DB, NC, JC, MWC, AE, AF, MG,
838 SML, SM, CN, SO'C, MO'M, MP, MEP, EW.
839
840 **Acknowledgements**

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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). 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 turning 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

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

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.

Babraham, 2016. FastQC. (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc)

Baxter, I.G. 1958. The composition of the Minch herring stocks. Rapports et procès-verbaux des réunions / Conseil permanent international pour l'exploration de la mer, 143: 814-94.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

865 Beacham, T.D, Wallace, C., Jonsen, K., McIntosh, B., Candy, J.R., Rondeau, E.B., Moore, J-S.,
866 Bernatchez, L. and Withler, R.E. 2020. Accurate estimation of conservation unit contribution to
867 coho salmon mixed-stock fisheries in British Columbia, Canada, using direct DNA sequencing for
868 single nucleotide polymorphisms. Canadian Journal of Fisheries and Aquatic Sciences, 77: 1302–
869 1315

870 Bekkevold, D., Berg, F., Polte, P., Bartolino, V., Ojaveer, H., Mosegaard, H., Farrell, E.D., Fedotova, E.,
871 Hemmer-Hansen, J., Huwer, B., Trijoulet, V., Moesgaard Albertsen, C., Fuentes-Pardo, A.,
872 Gröhsler, T., Petterson, M., Janssen, T., Andersson, L. and Folkvord, A. (In preparation). Genetic
873 stock identification in Atlantic herring (*Clupea harengus*): implementation of new tools alters
874 stock concepts and traces complex migration dynamics across multiple management areas.

875 Bekkevold, D., Gross, R., Arula, T., Helyar, S.J. and Ojaveer, H. 2016. Outlier loci detect intraspecific
876 biodiversity amongst spring and autumn spawning herring across local scales. PLoS ONE11:
877 e0148499.

878 Berg, F., Slotte, A., Johannessen, A., Kvamme, C., Clausen, L.W. and Nash, R.D.M. 2017. Comparative
879 biology and population mixing among local, coastal and offshore Atlantic herring (*Clupea*
880 *harengus*) in the North Sea, Skagerrak, Kattegat and western Baltic. PLoS One 12:e0187374, doi:
881 10.1371/journal.pone.0187374.

882 Bernatchez, L. *et al*. 2017 Harnessing the Power of Genomics to Secure the Future of Seafood. Trends
883 in Ecology and Evolution, 32: 665–680.

884 Cadrin S.X. & Secor D.H. 2009. Accounting for spatial population structure in stock assessment: past,
885 present and future. In: Beamish R. J., Rothschild B. J., editors. The Future of Fishery Science in
886 North America. Dordrecht: Springer. p. 405-425. 736 pp.

887 Campbell, N.R., Harmon, S.A. and Narum, S.R. 2015. Genotyping-in-Thousands by sequencing (GT-
888 seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular
889 Ecology Resources, 15: 855–867.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

890 Casey, J., Jardim, E. and Martinsohn, J.T.H. 2016. The role of genetics in fisheries management under
891 the E.U. common fisheries policy. Journal of Fish Biology, 89: 2755–2767.
892 Chen, K., Marshall, E.A., Sovic, M.G., Fries, A.C., Gibbs, H.L., Ludsin, S.A. 2018. assignPOP: An r package
893 for population assignment using genetic, non-genetic, or integrated data in a machine-learning
894 framework. Methods in Ecology and Evolution, 9: 439-446.
895 Cornuet, J.-M., Piry, S., Luikart, G., Estoup, A. and Solignac, M. 1999. New methods employing
896 multilocus genotypes to select or exclude populations as origins of individuals. Genetics, 15:
897 1989-2000.
898 Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M. and Blaxter, M.L. 2011. Genome-
899 wide genetic marker discovery and genotyping using next-generation sequencing. Nature
900 Reviews Genetics, 12: 499–510.
901 Dickey-Collas, M., Nash, R.D.M., Brunel, T., van Damme, C.J.G., Marshall, C.T., Payne, M.R., Corten, A.,
902 *et al*. 2010 Lessons learned from stock collapse and recovery of North Sea herring: a review.
903 ICES Journal of Marine Science, 67: 1875-1886.
904 Dieringer, D. and Schlötterer, C. 2003 Microsatellite Analyser (MSA): a platform independent analysis
905 tool for large microsatellite data sets. Molecular Ecology Resources, 3: 167–169.
906 Farrell, E.D., Campbell, N., Carlsson, J., Egan, A., Gras, M., Lusseau, S.M., Nolan, C., O'Connell, S.,
907 O'Malley, M., and White, E. 2021 Herring in Divisions 6.a, 7.b and 7.c: Scientific Assessment of
908 the Identity of the Southern and Northern Stocks through Genetic and Morphometric Analysis.
909 Service Contract EASME/EMFF/2017/1.3.2.1/SI2.767459. DOI: 10.2826/208498
910 Farrell, E.D., Carlsson, J.E.L. and Carlsson, J. 2016. Next Gen Pop Gen: implementing a high-throughput
911 approach to population genetics in boarfish (Capros aper). Royal Society Open Science,
912 3:160651.
913 Gilbey, J., Cauweiler, E., Coulson, M.W., Stradmeyer, L., Sampayo, J.N., Armstrog, A., Verspoor, E.,
914 Corrigan, L. Shelley, J. and Middlemas, S. 2016. Accuracy of assignment of Atlantic Salmon

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

915 (Salmo salar L.) to rivers and regions in Scotland and Northeast England based on single
916 nucleotide polymorphism (SNP) markers. PLoS ONE 11(10): e0164327.
917 Hatfield, E.M.C. *et al*., 2005. (WESTHER, Q5RS-2002-01056): A multidisciplinary approach to the
918 identification of herring (Clupea harengus L.) stock components west of the British Isles using
919 biological tags and genetic markers. Final Report. 320pp.
920 Han, F., Jamsandekar, M., Pettersson, M.E., Su, L., Fuentes-Pardo, A., Davis, B., Bekkevold, D., Berg, F.,
921 Cassini, M., Dahle, G., Farrell, E.D., Folkvord, A. and Andersson, L. (2020). Ecological adaptation
922 in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. eLife,
923 2020;9:e61076 (DOI: 10.7554/eLife.61076)
924 Heath, M. 1989. Transport of larval herring (Clupea harengus L.) by the Scottish Coastal Current.
925 Rapports et procès-verbaux des réunions / Conseil permanent international pour l'exploration
926 de la mer, 191: 85-91.
927 Heath, M. 1990. Segregation of herring larvae from inshore and offshore spawning grounds in the
928 north-western North Sea – Implications for stock structure. Netherlands Journal of Sea
929 Research, 25: 267-278.
930 Heath, M.R., MacLachlan, P.M. and Martin, J.H. A. 1987. Inshore circulation and transport of herring
931 larvae off the north coast of Scotland. Marine Ecology Progress Series, 40: 11-23, 1987
932 Hemmer-Hansen, J., Therkildsen, N.O. and Pujolar, J.M. 2014. Population genomics of marine fishes:
933 Next-generation prospects and challenges. The Biological Bulletin, 227: 117-132.
934 Hintzen, N.T., Roel, B., Benden, D., Clarke, M., Egan, A., Nash, R.D.M., Rohlf, N. & Hatfield, E.M.C. 2015.
935 Managing a complex population structure: exploring the importance of information from
936 fisheries-independent sources. ICES Journal of Marine Science, 72: 528-542.
937 ICES, 1982. Report of the Herring Assessment Working Group for the Area South of 62ºN (HAWG).
938 ICES CM 1982/Assess:7, 132 pp.
939 ICES, 2011. Report of the Workshop on Sexual Maturity Staging of Herring and Sprat (WKMSHS), 20-
940 23 June 2011, Charlottenlund, Denmark. ICES CM 2011/ACOM:46. 143pp.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

941 ICES, 2014. Report of the Herring Assessment Working Group for the Area South of 62ºN (HAWG), 11-
942 20 March 2014, ICES HQ, Copenhagen, Denmark. ICES CM 2014/ACOM:06. 1257 pp.
943 ICES, 2015. Report of the Benchmark Workshop on West of Scotland herring (WKWEST 2015). 2–6
944 February 2015, Marine Institute, Dublin, Ireland.
945 James, G., Witten, D., Hastie, T. and Tibshirani, T. 2013. An introduction to statistical learning, volume
946 112. Springer, 2013.
947 Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers.
948 Bioinformatics, 24: 1403-1405.
949 Jombart, T., Pontier, D and Dufour, A-B. 2009. Genetic markers in the playground of multivariate
950 analysis. Heredity, 102: 330-341.
951 Jombart, T., Devillard, S. and Balloux, F. 2010. Discriminant analysis of principal components: a new
952 method for the analysis of genetically structured populations. BMC Genetics, 11: 94
953 Jorde, P.E., Synnes, A-E., Espeland, S.H., Sodeland, M. and Knutsen, H. 2018. Can we rely on selected
954 genetic markers for population identification? Evidence from coastal Atlantic cod. Ecology and
955 Evolution, 8:12547-12558.
956 Kaplinski, L., Andreson, R., Puurand, T., Remm, M. 2005 MultiPLX: automatic grouping and evaluation
957 of PCR primers. Bioinformatics, 21: 1701-1702.
958 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A.,
959 Markowitz, S., Duran, C., Thierer, T., Ashton, B., Mentjies, P. and Drummond, A. 2012. Geneious
960 Basic: an integrated and extendable desktop software platform for the organization and analysis
961 of sequence data. Bioinformatics, 28: 1647-1649.
962 Kerr, L.A., Hintzen, N.T., Cadrin, S.X., Worsøe Clausen, L., Dickey-Collas, M., Goethel, D.R., Hatfield,
963 E.M.C., Kritzer, J.P., & Nash, R.D.M. 2016. Lessons learned from practical approaches to
964 reconcile mismatches between biological population structure and stock units of marine fish.
965 ICES Journal of Marine Science, 74: 1708–1722.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

966 Lamichhaney, S., Martinez Barrio, A., Rafati, N., Sundström, G., Rubin, C.-J., *et al*., 2012. Population-
967 scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring.
968 PNAS, 109: 19345–19350.
969 MacKenzie, K. 1985 The use of parasites as biological tags in population studies of herring (Clupea
970 harengus L.) in the North Sea and to the north and west of Scotland. ICES Journal of Marine
971 Science, 42: 33-64.
972 Mackinson, S., Pastoors,M., O'Connell, S, Berges, B., Forbes-Birnie, J., McAulay, S., Brigden, K., Fraser,
973 S., O'Malley,M., Farrell, E.D. 2021. The 2020 industry-science survey of herring in the Western
974 British Isles (ICES div 6a, 7bc). 122 pp
975 Mackinson, S., Pastoors,M., Lusseau, S., O'Connell, S, Forbes-Birnie, J., Sakinan, S., Berges, B., Brigden,
976 K., O'Malley,M., Farrell, E.D. 2019a. The 2019 industry-science survey of herring in the Western
977 British Isles (ICES div 6a, 7bc). 96pp
978 Mackinson, S., Pastoors,M., Lusseau, S., Armstrong, E., O'Connell, S, Sakinan, S., Berges, B., McClean,
979 A, Langlands, B., Wiseman, A., O'Malley,M., Farrell, E.D. 2019b. The 2018 industry-science
980 survey of herring in the Western British Isles (ICES div 6a, 7bc). 99pp
981 Manel, S., Gaggioti, O.E., Waples, R.S. 2005. Assignment methods: matching biological questions with
982 appropriate techniques. Trends in Ecology and Evolution, 20: 136–142.
983 Mariani, S., and Bekkevold, D. 2013. The nuclear genome: neutral and adaptive markers in fisheries
984 science. In Stock Identification Methods, 2nd edn, pp. 397–428. Ed. by S. X. Cadrin, L. A. Kerr,
985 and S. Mariani. Elsevier, San Diego. 566 pp.
986 Martinez Barrio, A., Lamichhaney, S., Fan, G., Rafati, N., Pettersson, M., Zhang, H., Dainat, J., Ekman,
987 D., Höppner, M., Jern, P., Martin, M., Nystedt, B., Liu, X., Chen, W., Liang, X., Shi, C., Fu, Y., Ma,
988 K., Zhan, X., Feng, C., Gustafson, U., Rubin, C.J., Sällman Almén, M., Blass, M., Casini, M.,
989 Folkvord, A., Laikre, L., Ryman, N., Ming-Yuen, Lee S., Xu, X., Andersson, L. 2016 The genetic
990 basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife,
991 5:12081.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

McQuinn, I.H. 1997. Metapopulations and the Atlantic herring. Reviews in Fish Biology and Fisheries, 7: 297–329.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., and Lin, C.-C. 2015. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071). TU Wien.

Nielsen, E. E., A. Cariani, E. Aoidh, G. E. Maes, I. Milano, R. Ogden, M. Taylor *et al*. 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature Communications, 3: 851.

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

Peakall, R. and Smouse P.E. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update. Bioinformatics 28, 2537-2539.

Pettersson, M.E., Rochus, C.M., Han, F., Chen, J., Hill, J., Wallerman, O., Fan, G., Hong, X., Xu, Q., Zhang, H., Liu, S., Liu, X., Haggerty, L., Hunt, T., Martin, F.J., Flicek, P., Bunikis, I., Folkvord, A. and Andersson L. 2019. A chromosome-level assembly of the Atlantic herring genome-detection of a supergene and other signals of selection. Genome Research, 29: 1919-1928.

Piry, S., Alapetite, A., Cornuet, J-M., Paetkau, D., Baudouin, L., Estoup, A. 2004. geneclass2: a software for genetic assignment and first-generation migrant detection. Journal of Heredity, 95: 536–539.

Pritchard, J.K, Stephens, M. and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155, 945–959.

Rannala, B. and J. L. Mountain. 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. USA 94: 9197-9221.

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

Rice, W.R. 1989 Analyzing tables of statistical tests. Evolution. 43, 223–225.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1018 Rousset, F. 2008 Genepop'007: a complete reimplementation of the Genepop software for Windows
1019 and Linux. Molecular Ecology Resources, 8: 103-106.
1020 Rozen, S. and Skaletsky, H. 2000 Primer3 on the WWW for general users and for biologist
1021 programmers. Methods in Molecular Biology, 132: 365–386.
1022 Saville, A. 1970. The stock fished by the Scottish herring fishery to the west of Scotland. ICES Pelagic
1023 Fish (Northern) Committee, CM19/H: 29
1024 Saville, A. and Morrison, J.A. 1973. A reassessment of the herring stock to the west of Scotland. ICES
1025 Pelagic Fish (Northern) Committee, CM 1973/H:24.
1026 Saville, A. and Bailey, R.S. 1980. The assessment and management of the herring stocks in the North
1027 Sea and to the west of Scotland. Rapport et Proces-Verbaux des Reunions, 177: 7-36.
1028 Schrider, D.R. and Kern, A.D. 2018. Supervised machine learning for population genetics: a new
1029 paradigm. Trends in Genetics, 34:301-312.
1030 Simmonds, E.J. 2009. Evaluation of the quality of the North Sea herring assessment. ICES Journal of
1031 Marine Science, 66: 1814–1822.
1032 Stephenson, R.L. 2002. Stock structure management and structure: an ongoing challenge for ICES. ICES
1033 Marine Science Symposium, 215: 305-314.
1034 Vartia, S., Villanueva, J.L., Finarelli, J., Farrell, E.D., Collins, P.C., Hughes, G., Carlsson, J.E.L., Gauthier,
1035 D.T., McGinnity, P., Cross, T.F., FitzGerald, R.D., Mirimin, L., Cotter, P. and Carlsson, J. 2016. A
1036 novel method of microsatellite genotyping-by-sequencing using individual combinatorial
1037 barcoding. Royal Society Open Science, 3:150565
1038 Waples, R.S., Punt, A.E. and Cope, J.M. 2008. Integrating genetic data into management of marine
1039 resources: how can we do better? Fish and Fisheries, 9: 423–449.
1040

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1041 Table 1. Pairwise multi-locus F_{ST} (above the diagonal) for the *baseline dataset* and associated *P*-
1042 values (below the diagonal) with the temporal replicates condensed.

	6aS	CS	IS	6aN_Aut	6aN_Sp	NS	DWN
6aS		0.09	0.13	0.20	0.35	0.24	0.12
CS	0.0001		0.08	0.23	0.64	0.32	0.01
IS	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
NS	0.0001	0.0001	0.0001	0.0001	0.0001		0.33
DWN	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	

1043

1044 Table 2. Clustering analyses, using the *find.clusters* function in *adegenet*, of the *6a baseline dataset*.
1045 The percentage of each population group split by cluster and the percentage of each cluster split by
1046 population are shown.

	Populations by Cluster		
	1+3+5	**2**	**4+6**
6aS	3.5	13.8	82.7
6aN_Aut	93.6	0.4	6.0
6aN_Sp	2.0	95.1	2.9

	Clusters by Population		
	1+3+5	**2**	**4+6**
6aS	2.3	53.6	89.4
6aN_Aut	97.5	2.3	10.5
6aN_Sp	0.2	44.1	0.4

1047

1048

1049 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 ± SD	
								6aS/6aN_Sp	**6aN_Aut**
1	1	MC	1	800	-	6aS/6aN_Sp	956	0.94 ± 0.02	0.06 ± 0.02
						6aN_Aut	1360	0.08 ± 0.01	0.92 ± 0.01
1	1	K-fold	1	-	10	6aS/6aN_Sp	956	0.93 ± 0.03	0.07 ± 0.03
						6aN_Aut	1360	0.06 ± 0.03	0.94 ± 0.03
								6aS	**6aN_Sp**
1	2	MC	1	75	-	6aS	854	0.86 ± 0.01	0.14 ± 0.01
						6aN_Sp	102	0.06 ± 0.04	0.94 ± 0.04
1	2	K-fold	1	-	10	6aS	854	0.96 ± 0.02	0.04 ± 0.02
						6aN_Sp	102	0.53 ± 0.11	0.47 ± 0.11
								6aS/6aN_Sp^{A2L1}	**6aN_Aut^{A2}**
2	1	MC	1	800	-	6aS/6aN_Sp^{A2L1} (Clusters_2+4+6)	1011	0.96 ± 0.01	0.04 ± 0.01
						6aN_Aut^{A2} (Clusters_1+3+5)	1305	0.04 ± 0.01	0.96 ± 0.01
2	1	K-fold	1	-	10	6aS/6aN_Sp^{A2L1} (Clusters_2+4+6)	1011	0.95 ± 0.02	0.05 ± 0.02
						6aN_Aut^{A2} (Clusters_1+3+5)	1305	0.03 ± 0.01	0.97 ± 0.01
								6aS^{A2L2}	**6aS/6aN_Sp^{A2L2}**
2	2	MC	1	200	-	6aS^{A2L2} (Clusters_4+6)	791	0.99 ± 0.00	0.01 ± 0.00
						6aS/6aN_Sp^{A2L2} (Cluster_2)	220	0.00 ± 0.00	1.00 ± 0.00
2	2	K-fold	1	-	10	6aS^{A2L2} (Clusters_4+6)	791	1.00 ± 0.00	0.00 ± 0.00
						6aS/6aN_Sp^{A2L2} (Cluster_2)	220	0.01 ± 0.02	0.99 ± 0.02

1050

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	≥ 0.67	Assigned	6aN_Aut
1	1	6aS/6aN_Sp	≥ 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	≥ 0.67	Assigned	6aS
1	2	6aN_Sp	≥ 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	$6aN\_Aut^{A2}$ (Clusters_1+3+5)	≥ 0.67	Assigned	$6aN\_Aut^{A2}$
2	1	$6aS/6aN\_Sp^{A2L1}$ (Clusters_2+4+6)	≥ 0.67	Move to level 2	-
2	1	$6aN\_Aut^{A2}$ (Clusters_1+3+5)	<0.67	Not assigned	NA
2	1	$6aS/6aN\_Sp^{A2L1}$ (Clusters_2+4+6)	<0.67	Not assigned	NA
2	2	$6aS^{A2L2}$ (Clusters_4+6)	≥ 0.67	Assigned	$6aS^{A2L2}$
2	2	$6aS/6aN\_Sp^{A2L2}$ (Cluster_2)	≥ 0.67	Assigned	$6aS/6aN\_Sp^{A2L2}$
2	2	$6aS^{A2L2}$ (Clusters_4+6)	<0.67	Not assigned	$6aS/6aN\_Sp^{A2L1}$
2	2	$6aS/6aN\_Sp^{A2L2}$ (Cluster_2)	<0.67	Not assigned	$6aS/6aN\_Sp^{A2L1}$

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 ESM Table 8.

Type	No. samples	No. individuals	Approach	≤0.67		≤0.7		≤0.8		≤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

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

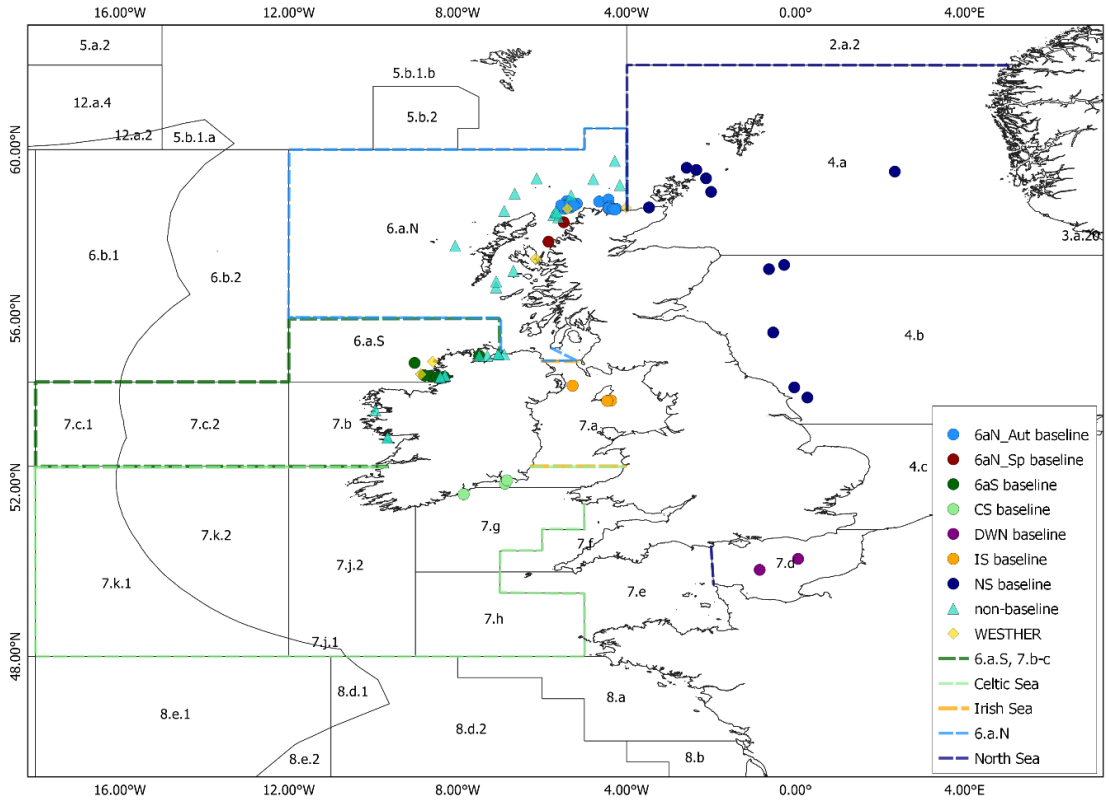


Figure 1. The distribution of herring samples collected and analysed in the current study. The sample type and current stock boundaries are indicated according to the legend.

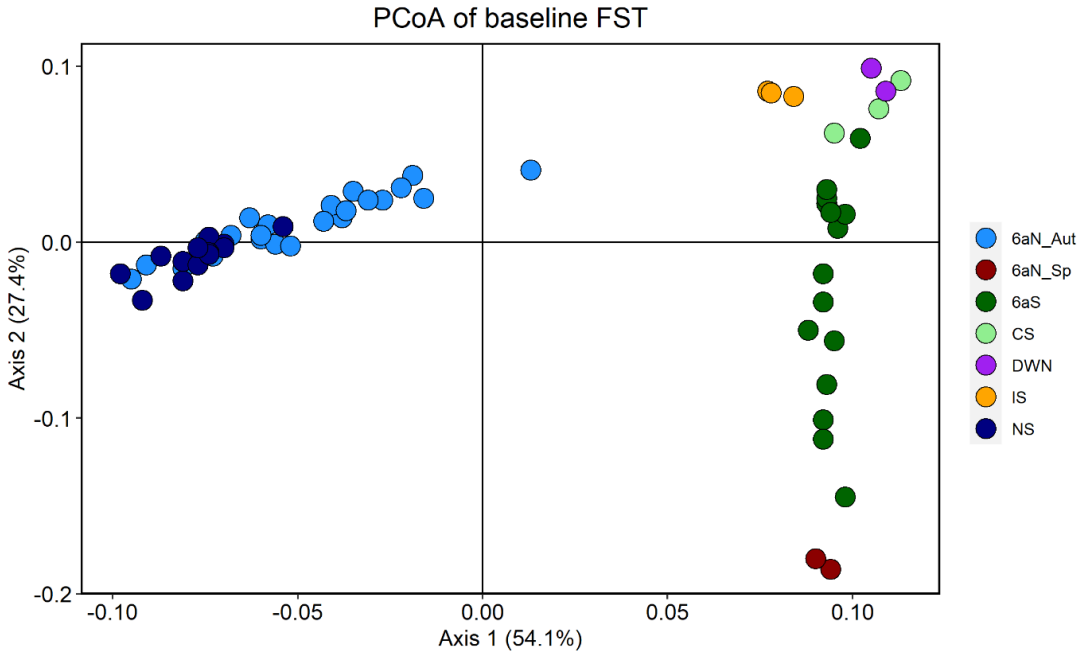


Figure 2. Principal Coordinate Analyses (PCoA) of the multi-locus pairwise F_{ST} analyses of the baseline samples.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

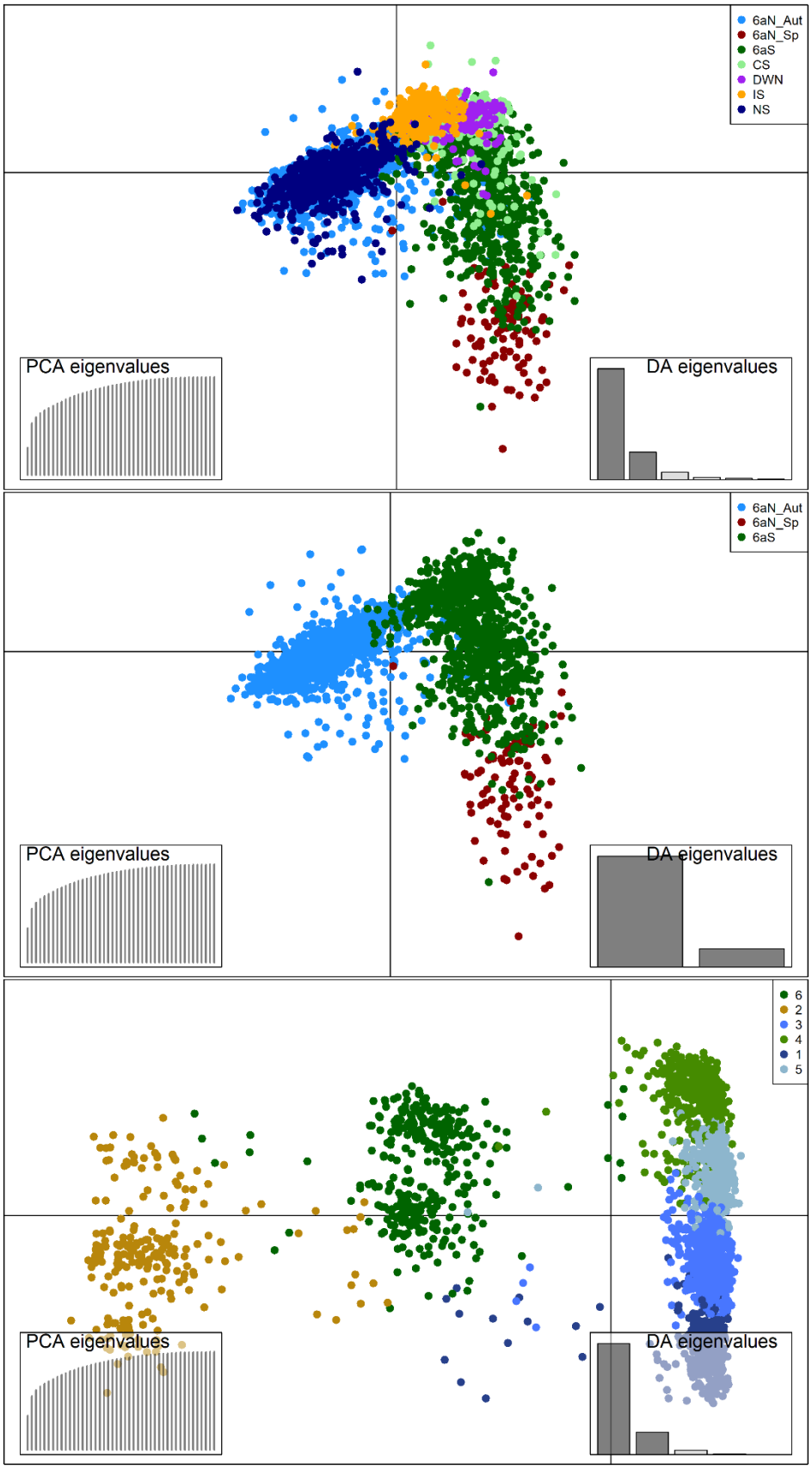


Figure 3. Discriminant Analysis of Principal Components (DAPC) of (top) the pooled *baseline dataset* (middle) the *6a baseline dataset* (bottom) the *clustered 6a baseline dataset*.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.07.487341; this version posted April 9, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

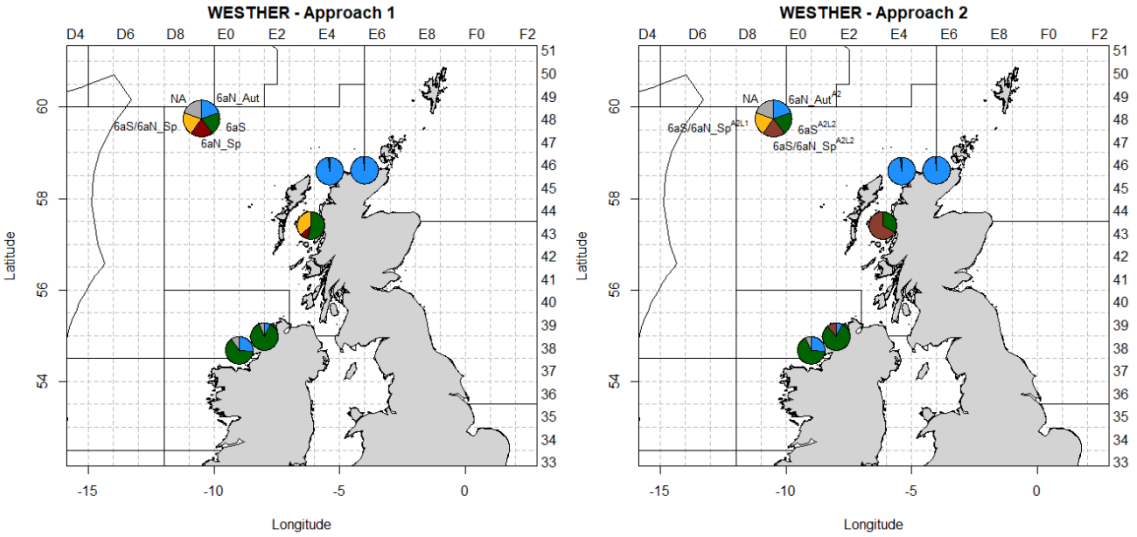
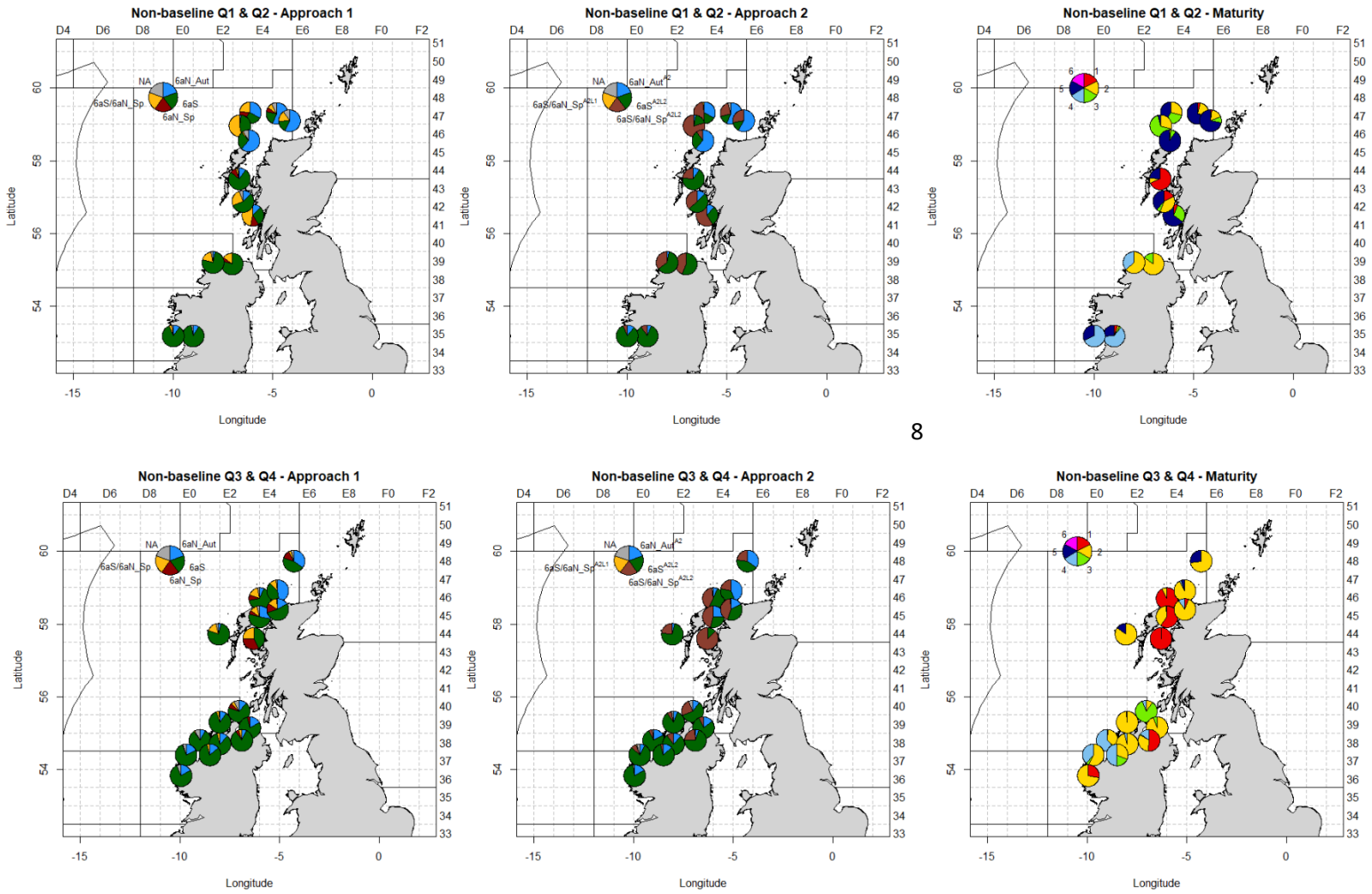


Figure 4. Assignment output of the archive samples from the WESTHER project following (left) Approach 1 and (right) Approach 2.



8

Figure 5. The assignment outputs and maturity stages of the contemporary non-baseline samples divided by quarter. Note the exact catch positions have been adjusted to minimise the overlap of the pie charts.