Optimisation of molecular tools for monitoring and population assessment of fish species in the North Celtic and Irish seas

Frances Clare Ratcliffe, BSc (hons), MSc

College of Science

Swansea University

Swansea University
Prifysgol Abertawe

This thesis is submitted for the degree of
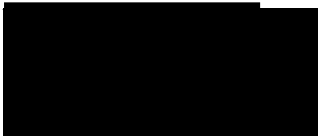
Doctor of Philosophy

2021

I

Summary

The North Celtic and Irish seas are extensively fished and, due to their shallow coastal nature, are particularly vulnerable to climate change stressors. Fish species assemblages in the region are characterised by high levels of diversity and include species of commercial and recreational importance, however poor larval recruitment and high fishing mortality have led to declines in abundance for some species. Monitoring of larval recruits and heavily exploited fish populations is essential for future sustainable fisheries management in a changing climate. This work aimed to develop and optimize molecular techniques for assessing fish communities in spawning grounds and fine scale population structure using adaptive molecular markers. Firstly, metabarcoding of bulk fish larvae homogenates was optimised by standardizing input material and using conserved priming sites, resulting in quantitative relative abundance estimates. This demonstrated that bulk larvae metabarcoding is a feasible alternative to traditional morphological assessment for assessing community diversity and composition. Secondly, species detections in spawning grounds from water sample and bulk larval sample metabarcoding were compared resulting in a 75% average agreement in detections across sample sites. Thirdly, a class I major histocompatibility complex (MHC) marker was developed and tested to assess sea bass population structure, allelic diversity and positive selection. Private alleles within the Celtic Shelf and Portuguese populations were identified. Finally, signals of positive selection and trans-species functional supertype structure in the MHC class II alpha and beta domains of the clade (series) Eupercaria were compared. Contrary to findings in other vertebrates, both domains exhibited similar levels of selection and should therefore be considered candidate regions for population structure studies in this clade. This thesis demonstrates that the molecular techniques demonstrated supplement, and in some cases improve on, existing monitoring and population assessment techniques and contribute to the sustainable management of fishes in the North Celtic and Irish seas.

Declaration and Statements

I, **Frances Ratcliffe**, certify that this work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.
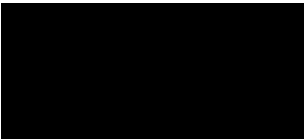
Signed   (candidate)

Date 17/07/2021

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).
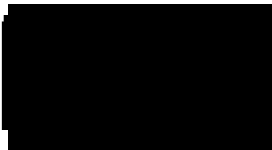
Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed   (candidate)
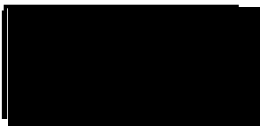
Date 28/07/2021

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed   (candidate)

Date 28/07/2021

STATEMENT 3

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed   (candidate)

Date 28/07/2021

Authorship Declaration

The following people and institutions contributed to the publication of work undertaken as part of this thesis:

	Name	Institution	Contribution chapter
Candidate	Frances Ratcliffe	Swansea University	Chapter 2, 3 and 4
Author 1	Sofia Consuegra	Swansea University	Chapter 2, 3 and 4
Author 2	Carlos Garcia de Leaniz	Swansea University	Chapter 2, 3 and 4
Author 3	Tamsyn Uren Webster	Swansea University	Chapter 2 and 3
Author 4	Deiene Rodriguez-Barreto	Swansea University	Chapter 3
Author 5	Richard O'Rorke	Swansea University	Chapter 3

Author details and their roles:

Paper 1, *Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples*, Chapter 2.

Candidate contributed to the study idea, designed the methodology, carried out the field and laboratory work, performed bioinformatic and statistical and wrote the manuscript. **Author 1** conceived the ideas and wrote and contributed to the manuscript. **Author 2** conceived the ideas and secured funding. **Author 3** contributed methodology design, laboratory work and bioinformatic analysis. **Author 4** contributed to the field sampling design and laboratory protocols. **Author 5** built the species database and contributed to bioinformatic analyses. All authors contributed critically to the drafts.

Paper 2, *A drop in the ocean: Monitoring fish communities in spawning areas using environmental DNA*, included in Chapter 3.

Candidate conceived the idea, carryout out the sampling, laboratory work, bioinformatic and statistical analyses and led the writing of the manuscript. **Author 1** conceived the idea and led the writing of the manuscript with the candidate. **Author 2** conceived the idea and secured funding. **Author 3** contributed to laboratory work and bioinformatic analyses. All authors contributed critically to the drafts and final version.

Paper 3 *MHC class I-alpha can reveal cryptic fine-scale population structure in a commercial fish, the European sea bass (Dicentrarchus labrax),* included in Chapter 4.

Candidate contributed to the idea, designed the primers, carried out the laboratory work, bioinformatic analyses and wrote the manuscript with Author 1. **Author 1** conceived the idea and led the writing of the manuscript and contributed to bioinformatic analyses. **Author 2** conceived the idea, contributed to statistical analysis and secured funding. All authors contributed critically to the drafts and final version.

We the undersigned agree with the above stated proportion of work undertaken for each of the above submitted peer-reviewed manuscripts contributing to this thesis:

Candidate _____

Author 1 _____

Author 2 _____

Author 3 _____

Author 4 _____

Author 5 _____

Contents

Contents

Acknowledgements

While a PhD thesis can be a somewhat lonely journey, when I look back at my time at Swansea University, it will be the people I met that will keep shining in my memory. I would like to thank my supervisors Professors Sonia Consuegra del Olmo and Carlos Garcia de Leaniz for giving me the opportunity to explore the world of genetic analyses and for all their help and support along the way. I am particularly grateful for the support I was given in the process of becoming a mum whilst finishing my PhD, during a global pandemic. The opportunity to learn from some wonderful postdocs, Dr Tamsyn Uren Webster, Dr Richard O'Rorke and Dr Deiene Rodriguez Barreto, was inspiring, thank you all. Special thanks to Tamsyn for wonderful friendship and positivity throughout this whole process. I would also like to thank my office mates who made this feel like a team effort despite us all working on different projects. They helped me laugh at myself (maybe a little too much), baked lots of cakes (maybe too many) and became great friends. I am especially grateful to my office and lab mate Jessica Minett for saving my sanity many times with R, formatting wizardry and blueberries.

With the birth of my daughter, this thesis became a family effort. Without the support of my parents who looked after my daughter whilst I worked, this process would have been very different. Whilst taking a toddler to feed the ducks might not be crucial to all thesis writing endeavours, knowing she was happy and cared for enabled me to focus on my work. Mum, you are amazing, thank you for the care you have given my daughter and me throughout this process. I want to thank my daughter too for focussing my mind, before her arrival and in the short hours when we were separated and for being such a wonderful distraction in all the other hours. Lastly, my deepest thanks goes to my partner Tom. This is the second PhD thesis we have weathered together, it's (hopefully) the last! We have worked side by side from home through the pandemic, there's not anyone I'd rather do that with. You've been an incredible emotional (and dare I say, technical and culinary) support and made the madness of finishing a thesis in our first years of parenting possible.

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
ANOSIM	Analysis of similarities
BIC	Bayesian Information Criterion
CPUF	Catch per unit filtered
bp	Base pair
DAPC	Discriminant Analysis of Principle Components
df	Degrees of freedom
DNA	Deoxyribonucleic acid
dN/dS	Ratio of non-synonymous to synonymous substitutions
eDNA	Environmental DNA
F	F-statistic
FEL	Fixed Effects Likelihood
FUBAR	Fast, Unconstrained Bayesian AppRoximation
K	Number of centroids
Km	Kilometres
ml	Millilitre
MHC	Major Histocompatibility Complex
MEME	Mixed Effects Model of Evolution
mg	Milligrams
mm	Millimetre
mt	Mitochondrial
N	Number
ng	Nanograms
No.	Number
NSS	Negatively selected sites
p	P-value
PBR	Peptide Binding Region
PCR	Polymerase chain reaction
PSS	Positively Selected Sites
r	Correlation coefficient
RPUF	Reads per unit filtered
SNP	Single Nucleotide Polymorphism
Spp.	Species
t	Test statistic
TTOL	Time tree of life
µl	Microliter
µm	Micrometre
µM	Micromolar
qPCR	Quantitative PCR
χ^2	Chi-squared statistic
°C	Degrees Celsius
%	Percentage
~	Approximately

Chapter 1   Introduction

Fishes play a central role in marine ecosystem dynamics and are often commercially exploited, and therefore management regulations are needed to prevent practices such as overfishing (Hernvann & Gascuel, 2020). For management interventions to be effective, accurate and current data on the distribution and population structure of larval and adult fishes is needed (Pauly & Zeller, 2003). However, monitoring in marine systems is complex: larval fish are difficult to capture and identify accurately (Ellis, Milligan, Readdy, Taylor, & Brown, 2012) and uncovering fine scale population structure of species that have wide dispersal ranges is challenging (Souche et al., 2015).

There are a wide range of genetic techniques that have the potential to compliment traditional fisheries monitoring and enhance data availability. However, these techniques still require optimisation, testing and ground-truthing before they can be incorporated into current practice. This thesis seeks to test and optimise the use of high throughput sequencing (metabarcoding and allele genotyping) for monitoring fish spawning grounds and uncovering fine scale population structure of Irish and North Celtic sea fish species.

1.1   The study region

The Irish and North Celtic seas (ICES divisions VIIa and VIIg) lie between Ireland to the west, and Wales and England to the east, linked by an area known as the St George's Channel (Brown et al., 2003). The Celtic sea is a transition zone from the continental shelf waters of the Atlantic ocean, to the coastal waters of the Irish Sea (Huthnance, Holt, & Wakelin, 2009). Both the Irish and North Celtic Seas are relatively shallow, (<130m in depth), with the exception of a basin known as the Celtic Deep, which stretches along St George's Channel, and into the southern Irish Sea (Sharples, Ellis, Nolan, & Scott, 2013). In summer months, a distinct tidal-mixing front, known as the Irish Sea Front extends across St Georges Channel from the Irish to the Welsh coast, limiting water mixing between the seas (Brown et al., 2003). To the north of this front, the Irish Sea remains mixed year round, due to shallower water and stronger tides, however, to the south, solar energy leads to warming of the surface waters of the Celtic sea sitting above a dense layer of higher salinity and lower temperature water (Brown et al., 2003; Lee, Nash, & Danilowicz, 2005).

Globally, mean sea surface temperatures have been increasing (Mieszkowska et al., 2006) and over the course of this century, climate change is expected to become the greatest driver of changes in biodiversity in marine ecosystems (Garciá Molinos et al., 2016). Due to their shallow and seasonally stratified nature, the Irish and North Celtic seas are likely to be disproportionately impacted by changes in climate, with changes not only to temperature profiles, but also to salinity and vertical mixing and

stratification (Holt, Wakelin, Lowe, & Tinker, 2010). This is likely to influence the distribution of water properties in the region, in turn affecting abundances and distributions of phytoplankton, and subsequently zooplankton (Holt et al., 2010). Changes in primary productivity has knock-on consequences for ecosystem functioning (Hawkins et al., 2009; Sharples et al., 2013), and this in turn will impact abundances of exploited organisms, such as fish and shellfish (Sumaila, Cheung, Lam, Pauly, & Herrick, 2011).
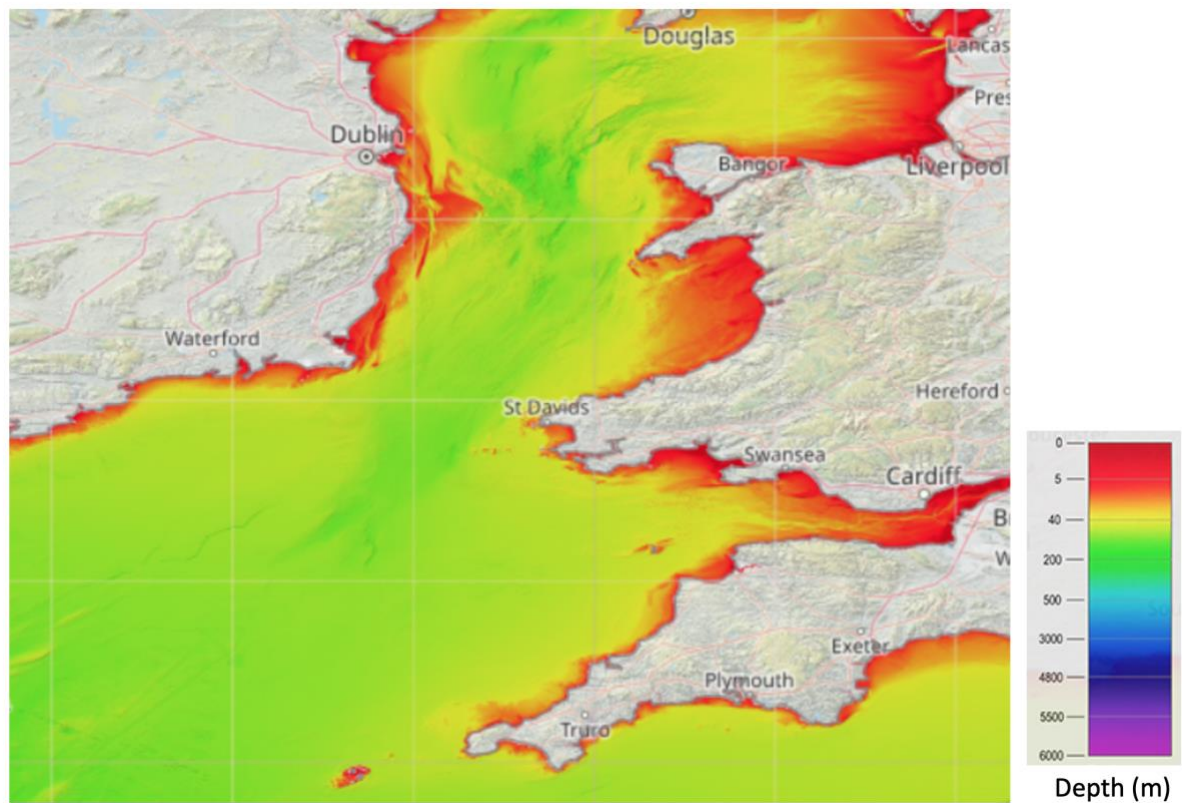
Figure 1.1 Bathymetric map of the study region (Irish and Celtic Seas), showing mean depths. Map generated using https://portal.emodnet-bathymetry.eu/

1.2 Fish ecology and fisheries in the Irish and North Celtic Seas

Both fish and invertebrate species assemblages in the region are characterised by high levels variability and diversity (Ellis, Rogers, & Freeman, 2000; Kaiser et al., 2004), driven by complex spatial organisation of habitats as well as the wide range of substrate types, depths and mixing fronts (Hernvann & Gascuel, 2020). The North Celtic sea in particular, which adjoins the continental shelf region, is an area of high primary productivity (Ruiz-Castillo, Sharples, Hopkins, & Woodward, 2019). Therefore, these seas contain many fish spawning grounds (Ellis, Milligan, Readdy, Taylor, & Brown, 2012) as well as economically important wild capture fisheries (Calderwood et al., 2020). Many species co-occur, both in spawning assemblages and fishing grounds (Ellis et al., 2012; Pinnegar, Jennings,

O'Brien, & Polunin, 2002) and fisheries in the region target a mixture of species (ICES 2020). In the Irish sea, while langoustine (*Nephrops norwegicus)* make up the highest proportion of landings (ICES 2019), finfish are also impacted due to high incidental catches and discard rates, particularly of gadoid species (Catchpole, Frid, & Gray, 2005). Haddock *(Melanogrammus aeglefinus),* is the second largest catch, but plaice (*Pleuronectes platessa*), anglerfish (*Lophius piscatorius*), hake (*Merluccius merluccius*), megrim (*Lepidorhombus whiffiagonis*), and cod (*Gadus morhua*) are also targeted (ICES 2020). To the south, in the North Celtic sea, pelagic species account for the largest catches, dominated by hake *(Merluccius merluccius)* catches but also targeting anglerfish, megrim, whiting *(Merlangius merlangus*), langoustine, haddock, cod, pollock (*Pollachius pollachius*), sole *(Solea solea),* ling *(Molva molva),* saithe *(Pollachius virens),* and plaice *(Pleuronectes platessa)* (ICES 2020). The status in terms of biomass and fishing mortality for the majority (60%) of the 106 fished stocks is unknown, although since the mid 1990s, fishing mortality has been in decline and the size of some stocks for which data is available, is increasing (ICES 2020). Notably, some species of socio-economic importance are considered outside of safe biological limits, including sea bass *(Dicentrarchus labrax)* (Walker et al., 2020)*,* cod (Bentley, Serpetti, Fox, Heymans, & Reid, 2020) and herring *(Clupea harengus)* (ICES 2019). Despite stringent management measures to reduce fishing mortality and therefore halt declines in abundance, recovery has been slow (Bentley et al., 2020). This is in part due environmental stressors, such as temperature fluctuations, which influence larval recruitment (Bentley et al., 2020; Walker et al., 2020). Therefore, understanding the impact multiple stressors on fish populations across the region is necessary to inform effective management of fish populations (Bentley et al., 2020).

1.3   *Monitoring larval and adult populations of fishes in Irish and Celtic Seas*

For fisheries and ecosystems to be managed effectively, data availability is of paramount importance (Pauly & Zeller, 2003) and in particular, the levels of larval recruitment and population genetic structure of stocks is needed (Ellis et al., 2012; Walker et al., 2020; Ward, 2000). Fish populations are defined as groups of individuals of a species or subspecies that are genetically, spatially, or demographically separated from other groups of conspecifics (Wells & Richmond 1995). Fish stocks, which are designated units for management purposes, ideally follow population boundaries, incorporating a degree of reproductive isolation from other stocks of the same species (Cadrin & Secor, 2009; Ward, 2000). The designation of stock units have, in some cases been arbitrarily statistically assigned because management practicality concerns and the political complexities of managing an essentially international resource (Reiss, Hoarau, Dickey-Collas, & Wolff, 2009). This has led to a mismatch between biologically important processes and management actions (Reiss et al., 2009). In addition to single species recruitment and population structure information, community composition data, which incorporates multiple species assessment is increasingly relied upon in an ecosystem based

approach to fisheries management (Bentley et al., 2020). For many species in the region, the status of the stock is unknown (ICES 2020), larval recruitment is largely unmonitored (Ellis et al., 2012) and stock units have been assigned without regard for population structure (Reiss et al., 2009). Over the coming decades, environmental stressors are likely to lead to complex changes in ecosystem function (Hernvann et al., 2020), and, combined with existing pressure from capture fisheries, the need for monitoring has never been so great.

1.4   *Traditional fish monitoring*

Traditional fish monitoring involves large scale, ship-borne surveys, such as groundfish surveys, which target multiple fishes (Moriarty, Greenstreet, & Rasmussen, 2017) or single species surveys, such as the mackerel egg surveys (Brunel, van Damme, Samson, & Dickey-Collas, 2018), and ICES stock assessments (ICES 2020). While ideally survey protocols, such as tow speed and duration, should be standardised between vessels to provide quantitative abundance estimates, in reality, differences between vessel behaviour and gear often occur (Moriarty et al., 2017). Traditional surveys rely on destructive sampling, and identification of specimens is traditionally carried out using morphological features, the accuracy of which dependent on taxonomic expertise of surveyors (Hansen, Bekkevold, Clausen, & Nielsen, 2018). These types of surveys are both costly and time consuming (Hansen et al., 2018).

Before genetic techniques became widespread, stock designation was based on geographic variation of phenotypic traits and closed migration circuits (Cadrin & Secor, 2009). To compensate for the challenge of obtaining data of the scale and accuracy required for stock assessment, statistical methods that integrate differing data collection methods (for example fishing gear type, season, or vessel size) are commonly used (Maunder & Punt, 2013). This has led to fisheries stock estimation methods to become increasingly elaborate and difficult to communicate to stakeholders (Cotter et al., 2004). Despite acknowledged limitations of current fisheries monitoring techniques, they underpin vast fisheries management legislation (Rätz, Dörner, Scott, & Barbas, 2010), and, therefore any new methodologies need to be developed with reference to existing protocols (Hansen et al., 2018). Traditional monitoring and stock designation techniques have now, in some cases, been supplemented by molecular techniques (Casey, Jardim, & Martinsohn, 2016; Reiss et al., 2009); however, newer techniques such as those based on high throughput sequencing still require optimisation and ground-truthing.

1.5 Molecular techniques for identifying species: barcoding and metabarcoding

Since the 1960s, DNA sequence assessment has been used for developing protocols to discriminate between species (Manwell & Baker 1963, Ward, Zemlak, Innes, Last, & Hebert, 2005). One of the most widely used methods is polymerase chain reaction (PCR) amplification of an area of the mitochondrial genome (the cytochrome oxidase-I- COI, cytochrome *b* and 16S-rRNA genes being amongst the most commonly targeted (Teletchea, 2009)). This technique uses primers that anneal to conserved (common) regions present between species and amplifies regions which differ between species, allowing taxonomic groups to be distinguished and is known as DNA barcoding (Adamowicz 2015; Hebert, Ratnasingham, & DeWaard, 2003). Barcoding has now become an established practice, supported by the initiative known as the International Barcode of Life (https://ibol.org/). Barcoding can provide reproducible species identifications, independent of taxonomic expertise (Ward et al., 2005), (it should be noted that discrepancies sometimes between taxonomical conclusions obtained from morphological and molecular identification methods (Teletchea, 2009)). Because DNA barcoding relies largely on Sanger sequencing, organisms need to separated from other specimens and the DNA extracted and amplified individually (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). Therefore, while possible, barcoding is not optimal for some applications such as multispecies samples, where individual sequencing would be costly and time consuming (Yu et al., 2012). With the widespread development of high-throughput-sequencing (HTS) platforms, a process allows for barcoding multiple species simultaneously has been developed, known as metabarcoding (Taberlet et al., 2012). Metabarcoding generates millions of short reads (typically 50-300 base pairs in length) simultaneously, and many multi-species samples can be multiplexed (sequenced) concurrently . This process enables the automated identification of 100s of individuals from multiple mixed species samples in a single sequencing run (Meyer & Kircher, 2010; Taberlet et al., 2012). Metabarcoding has the advantage of being cost effective for mixed species samples, particularly because individual organisms do not need to be separated out before DNA extraction, amplification and sequencing, allowing 100s of samples to be analysed together (Yu et al., 2012). In addition, because the process utilises short fragments of mitochondrial DNA which is in higher abundance than nucleic DNA, the process is applicable to degraded samples (Creer et al., 2016; Ficetola et al., 2015). This has led to the technique being used to identify taxa in a range of different sample types, from gut microbiomes (Escalas et al., 2021; Tarnecki, Burgos, Ray, & Arias, 2017), to homogenates of stomach samples in diet analyses (Siegenthaler, Wangensteen, Benvenuto, Campos, & Mariani, 2018; Thomas, Jarman, Haman, Trites, & Deagle, 2014), to the identification of fishes within a water sample (Cilleros et al., 2019; Jerde, Wilson, & Dressler, 2019; Thomsen et al., 2016).

1.6   Spawning grounds monitoring using metabarcoding

There are multiple fish spawning and nursery grounds across both the Irish and Celtic Seas (Ellis et al., 2012). However, despite their importance in terms of fish recruitment in the region, current maps of UK spawning grounds (Ellis et al., 2012) rely heavily on distributions from surveys in the 1990s (Coull et al., 1998) and, for many taxa regular surveys do not currently take place (CEFAS, 2020), in part due to the workload involved in identifying and processing samples. While descriptions of larvae exist for the study region (Russel 1976), no taxonomic keys have been developed and morphological assessment is time consuming, requiring specialist training to achieve accurate results (Brechon *et al.,* 2013). This has led to many samples sitting in storage, awaiting identification and quantification (Haberlin, Raine, McAllen, & Doyle, 2019). Molecular techniques such as metabarcoding provide an automated alternative to expert morphological identification, and the process can be standardised between laboratories, eliminating observer bias inherent in the morphological identification process (Harvey, Johnson, Fisher, Peterson, & Vrijenhoek, 2017). Because metabarcoding can be applied to different sample types, it offers the potential to assess not only the species present in a plankton tow sample, which are limited to specific sizes and therefore life stages, but also fish of other size classes whose DNA is present in water samples (Maruyama, Nakamura, Yamanaka, Kondoh, & Minamoto, 2014). This increases the amount of data that can be extracted from costly ship surveys. While metabarcoding is now commonly used in ecological surveys to garner species identification information, the relationship between relative abundance of reads in a bulk metabarcoded sample and relative abundance of organisms present can be variable (Deagle et al., 2019). Abundance estimates increase the utility and sensitivity of metrics such as community composition and diversity (Clarke, 1993) as well as stock assessment (Koslow & Wright, 2016). Therefore, for this technique to be more widely applied to fish larvae monitoring, reliability of relative abundance estimates derived from sequencing data require improvements that result in comparable estimates to non-molecular techniques. In addition, while metabarcoding of water samples is now widespread (Beng & Corlett, 2020), the relationship between species detected in a water sample and those present at a sampling location requires ground-truthing, especially in the marine environment (Collins et al., 2018; Hansen et al., 2018).

1.7   Genetic techniques using markers under selection to differentiate populations

Mitochondrial metabarcoding can provide insight into species community assemblages, enabling monitoring of vulnerable taxa and life stages; however one of the most powerful uses for genetic analyses is the uncovering of population structure information for single fish species management.

Since as early as the 1930s, the potential for genetic techniques in defining population structure has been appreciated and developed (Ward, 2000). Techniques to separate fish populations have undergone

many iterations from serological analysis, to protein electrophoretic variation (Ward, 2000), to microsatellites and single nucleotide polymorphisms (SNPs) (Vignal, Milan, SanCristobal, & Eggen, 2002). The majority of population genetic techniques that have been developed target areas of the genome under neutral selective pressure (Kirk & Freeland, 2011). However, there is growing awareness that these markers alone provide an incomplete insight into parameters such as local adaptation and genetic diversity of specific genes involved in fitness (Kirk & Freeland, 2011). In order to understand whether local adaptation is occurring, assessing areas of the genome that directly relate to fitness add valuable information to population genetic studies (Biedrzycka, Konopiński, Hoffman, Trujillo, & Zalewski, 2020; Larson, Lisi, Seeb, Seeb, & Schindler, 2016). Areas of the genome that code for proteins, thereby influencing fitness, such as immune genes, are ideal candidates to garner such information (Consuegra et al., 2005). Assessing coding regions gives an insight into a population's allelic richness (diversity), implying capacity functional differences, enabling adaptation to occur and enhance a species ability respond to stressors (Spurgin & Richardson, 2010). As the rate at which environmental change increases (Burrows et al., 2014), understanding a population's ability to adapt to environmental stressors could have strong implications for management purposes (Crozier & Hutchings, 2014).

1.8 Population differentiation using immune gene markers (Major Histocompatibility Complex)

Pathogens pose a major threat to wild fish populations (Grimholt, 2016) and the types of pathogenic challenge may change across regions and latitudinal gradients (Landry & Bernatchez, 2001). Marine vertebrates are therefore reliant on an effective immune response for their survival and may therefore show adaptations to localised parasites (Piertney & Oliver, 2006). At the centre of the adaptive vertebrate immune response is the Major Histocompatibility Complex (MHC) (Bernatchez & Landry, 2003). The major histocompatibility complex (MHC) is a multi-gene family, comprising of two classes, MHC class I and II. These genes code for receptor molecules that firstly recognise foreign peptides (antigens) and secondly bind these antigens in order to present them to immune cells, thereby triggering an immune response (Grimholt, 2016). The ability of the MHC to bind a particular pathogen derived peptide is dictated by the shape of the peptide binding regions (PBR). Different alleles (variants) of MHC genes may code for proteins that result in differing shapes of the PBR cleft, enabling different pathogens to be recognised (Doytchinova, Guan, & Flower, 2004). The need for recognition of diverse pathogens has therefore led to MHC genes becoming the most variable functional genes known in vertebrates (Piertney & Oliver, 2006).This area of the genome can evolve rapidly, because processes such as recombination and gene conversion can increase diversity (Consuegra et al., 2005). Because changes in MHC genes may be directly linked to environmental stressors, they are some of the best target genes to assess whether local adaptation is occurring and ascertain if fish from different regions

are responding to different stressors, and may be from different populations (Larson et al., 2016; Lighten et al., 2017).

Due to its extreme polymorphism however, the MHC molecules are not straightforward to amplify and analyse and questions such as how alleles are organised as loci remain unanswered for some species (Biedrzycka et al., 2020; Huang et al., 2019). While most of the variability is observed at PBR regions across all vertebrates, there are large differences in MHC between phyla, with some taxa lacking whole MHC classes (Malmstrøm et al., 2016). Therefore, information on MHC organisation specific to the clade of a species in question is required before targeting sequencing can be effective. In addition, there are multiple exons involved in shaping the binding cleft of class I and class II molecules (Malmstrøm et al., 2016), therefore choosing a region to target for sequencing is complex.

In cases where a species is in decline or possibly endangered and neutral genetic markers show little structuring, assessing diversity at the MHC may provide valuable insight into both local adaptation and potential resilience to environmental stressors (Talarico, Babik, Marta, & Mattoccia, 2019). An example of a species in decline and with shallow population structuring observed at neutral markers (Quéré et al., 2012; Souche et al., 2015) is the European sea bass *(Dicentrarchus labrax)*. From the 1970s until 2013, there was a high value commercial sea bass fishery in the region, however, due to declines in spawning stock biomass, poor recruitment and high fishing mortality, emergency measures to protect the stock have been introduced (Walker et al., 2020). Despite these limitations, no recovery has been observed and debate continues as to what measures will lead to an improvement (Walker et al., 2020). Therefore, insights into the diversity of MHC genes in this species and ascertaining if regional differentiation has occurred would potentially provide insight into how this *D. labrax* populations may respond to future environmental change and give indications for effective management measures.

1.9   Aims of thesis

The main aim of this thesis is to optimise molecular tools for the monitoring of fish spawning grounds and population analyses in North Celtic and Irish sea fish populations.

The aim of chapter 2 was to improve quantitative assessment of bulk fish larvae using metabarcoding. Larvae samples from the Irish and North Celtic Seas were identified morphologically and subsequently homogenised and assessed via metabarcoding, standardizing input material and using conserved primer binding sites to improve quantitative assessment relative abundance estimates. The sensitivity and accuracy of the molecular approach was compared with morphological identification, to assess whether metabarcoding can be used as alternative to traditional assessment of fish larvae in the region.

Ratcliffe, F. C., Uren Webster, T. M., Rodriguez-Barreto, D., O'Rorke, R., Garcia de Leaniz, C., & Consuegra, S. (2021). Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples. Ecological Applications, 31(3), e02284.

The aim of chapter 3 was to compare species detections between tissue and water samples from spawning grounds in the Irish and North Celtic seas. Water samples and fish larvae homogenate samples were processed using metabarcoding and an identical bioinformatics pipeline. Both sample types were then compared to assess the level of agreement between species detections, relative abundance, and community composition.

Ratcliffe, F. C., Uren Webster, T. M., Garcia de Leaniz, C., & Consuegra, S. (2021). A drop in the ocean: Monitoring fish communities in spawning areas using environmental DNA. Environmental DNA, 3(1), 43-54.

The aim of chapter 4 was to develop a marker to characterise MHC class I allelic diversity and selection signals for population differentiation assessment in sea bass (*Dicentrarchus labrax)*. An Illumina sequencing-based protocol to genotype the peptide binding region of the class I-alpha gene was developed, and its potential for detecting fine scale population structuring and signatures of local selection pressures was evaluated.

Ratcliffe, F. C., Garcia de Leaniz, C., & Consuegra, S., Allelic and supertype diversity of MHC class I-alpha reveal fine-scale population structure in European seabass (*Dicentrarchus labrax*). (Animal Genetics, under review).

The aim of chapter 5 was to assess the potential of the peptide binding region of the MHC class II alpha and beta genes as regions to develop markers to differentiate between populations for the clade (series) Eupercaria. Using alleles deposited in the National Center for Biotechnology Information nucleotide database, I assessed differences in positive selection signals, trans species functional supertype structure, and time-tree divergence times between alpha and beta chains.

1.10   References:

Adamowicz, S. J., & Scoles, G. J. (2015). International Barcode of Life: Evolution of a global research community. *Genome*, *58*(5), 151–162. https://doi.org/10.1139/gen-2015-0094

Beng, K. C., & Corlett, R. T. (2020). Applications of environmental DNA (eDNA) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and Conservation* (Vol. 29). https://doi.org/10.1007/s10531-020-01980-0

Bentley, J. W., Serpetti, N., Fox, C. J., Heymans, J. J., & Reid, D. G. (2020). Retrospective analysis of the influence of environmental drivers on commercial stocks and fishing opportunities in the Irish Sea. *Fisheries Oceanography*, *29*(5), 415–435. https://doi.org/10.1111/fog.12486

Bernatchez, L., & Landry, C. (2003). MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, *16*(3), 363–377. https://doi.org/10.1046/j.1420-9101.2003.00531.x

Biedrzycka, A., Konopiński, M., Hoffman, E., Trujillo, A., & Zalewski, A. (2020). Comparing raccoon major histocompatibility complex diversity in native and introduced ranges: Evidence for the importance of functional immune diversity for adaptation and survival in novel environments. *Evolutionary Applications*, *13*(4), 752–767. https://doi.org/10.1111/eva.12898

Brechon, A. L., Coombs, S. H., Sims, D. W., & Griffiths, A. M. (2013). Development of a rapid genetic technique for the identification of clupeid larvae in the Western English Channel and investigation of mislabelling in processed fish products. *ICES Journal of Marine Science*, *70*(2), 399–407. https://doi.org/10.1093/icesjms/fss178

Brown, J., Carrillo, L., Fernand, L., Horsburgh, K. J., Hill, A. E., Young, E. F., & Medler, K. J. (2003). Observations of the physical structure and seasonal jet-like circulation of the Celtic Sea and St. George's Channel of the Irish Sea. *Continental Shelf Research*, *23*(6), 533–561. https://doi.org/10.1016/S0278-4343(03)00008-6

Brunel, T., van Damme, C. J. G., Samson, M., & Dickey-Collas, M. (2018). Quantifying the influence of geography and environment on the northeast Atlantic mackerel spawning distribution. *Fisheries Oceanography*, *27*(2), 159–173. https://doi.org/10.1111/fog.12242

Burrows, M. T., Schoeman, D. S., Richardson, A. J., Molinos, J. G., Hoffmann, A., Buckley, L. B., … Poloczanska, E. S. (2014). Geographical limits to species-range shifts are suggested by climate velocity. *Nature*, *507*(7493), 492–495. https://doi.org/10.1038/nature12976

Cadrin, S. X., & Secor, D. H. (2009). The Future of Fisheries Science in North America. *The Future of Fisheries Science in North America*, 405-426. https://doi.org/10.1007/978-1-4020-9210-7

Calderwood, J., Robert, M., Pawlowski, L., Vermard, Y., Radford, Z., Catchpole, T. L., & Reid, D. G. (2020). Hotspot mapping in the Celtic Sea: An interactive tool using multinational data to optimise fishing practices. *Marine Policy*, *116* 103511. https://doi.org/10.1016/j.marpol.2019.103511

Casey, J., Jardim, E., & Martinsohn, J. T. H. (2016). The role of genetics in fisheries management under the E.U. common fisheries policy. *Journal of Fish Biology*, *89*(6), 2755–2767. https://doi.org/10.1111/jfb.13151

Catchpole, T. L., Frid, C. L. J., & Gray, T. S. (2005). Discarding in the English north-east coast Nephrops norvegicus fishery: The role of social and environmental factors. *Fisheries Research*, *72*(1), 45–54. https://doi.org/10.1016/j.fishres.2004.10.012

CEFAS. (2020). Cefas Data Portal Available at: https://data.cefas.co.uk/view/153 (Accessed: 15 July 2021).

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., … Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, *19*(1), 27–46. https://doi.org/10.1111/1755-0998.12900

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0192-6

Consuegra, S., Megens, H. J., Schaschl, H., Leon, K., Stet, R. J. M., & Jordan, W. C. (2005). Rapid evolution of the MH class I locus results in different allelic compositions in recently diverged populations of Atlantic salmon. *Molecular Biology and Evolution*, *22*(4), 1095–1106. https://doi.org/10.1093/molbev/msi096

Cotter, A. J. R., Burt, L., Paxton, C. G. M., Fernandez, C., Buckland, S. T., & Pan, J. X. (2004). Are stock assessment methods too complicated? *Fish and Fisheries*, *5*(3), 235–254. https://doi.org/10.1111/j.1467-2679.2004.00157.x

Coull, K. A., Johnstone, R., & Rogers, S. I. (1998). Fisheries sensitivity maps in British waters. Published and distributed by UKOOA Ltd, 9.

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., … Bik, H. M. (2016). The

ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Crozier, L. G., & Hutchings, J. A. (2014). Plastic and evolutionary responses to climate change in fish. *Evolutionary Applications*, *7*(1), 68–87. https://doi.org/10.1111/eva.12135

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., … Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. https://doi.org/10.1111/mec.14734

Doytchinova, I. A., Guan, P., & Flower, D. R. (2004). Identifiying Human MHC Supertypes Using Bioinformatic Methods. *The Journal of Immunology*, *172*(7), 4314–4323. https://doi.org/10.4049/jimmunol.172.7.4314

Ellis, J. R., Milligan, S. P., Readdy, L., Taylor, N., & Brown, M. J. (2012). Spawning and nursery grounds of selected fish species in UK waters. *Biological Conservation*, (147), 60. Retrieved from https://www.cefas.co.uk/publications/techrep/TechRep147.pdf

Ellis, J. R., Rogers, S. I., & Freeman, S. M. (2000). Demersal assemblages in the Irish Sea, St George's Channel and Bristol Channel. *Estuarine, Coastal and Shelf Science*, *51*(3), 299–315. https://doi.org/10.1006/ecss.2000.0677

Escalas, A., Auguet, J. C., Avouac, A., Seguin, R., Gradel, A., Borrossi, L., & Villéger, S. (2021). Ecological Specialization Within a Carnivorous Fish Family Is Supported by a Herbivorous Microbiome Shaped by a Combination of Gut Traits and Specific Diet. *Frontiers in Marine Science*, *8*. https://doi.org/10.3389/fmars.2021.622883

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., … Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Garciá Molinos, J., Halpern, B. S., Schoeman, D. S., Brown, C. J., Kiessling, W., Moore, P. J., … Burrows, M. T. (2016). Climate velocity and the future global redistribution of marine biodiversity. *Nature Climate Change*, *6*(1), 83–88. https://doi.org/10.1038/nclimate2769

Grimholt, U. (2016). MHC and evolution in teleosts. *Biology*, *5*(1). https://doi.org/10.3390/biology5010006

Haberlin, D., Raine, R., McAllen, R., & Doyle, T. K. (2019). Distinct gelatinous zooplankton communities across a dynamic shelf sea. *Limnology and Oceanography*, *64*(4), 1802–1818.

https://doi.org/10.1002/lno.11152

Hansen, B. K., Bekkevold, D., Clausen, L. W., & Nielsen, E. E. (2018). The sceptical optimist: challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish and Fisheries*, *19*(5), 751–768. https://doi.org/10.1111/faf.12286

Harvey, J. B. J., Johnson, S. B., Fisher, J. L., Peterson, W. T., & Vrijenhoek, R. C. (2017). Comparison of morphological and next generation DNA sequencing methods for assessing zooplankton assemblages. *Journal of Experimental Marine Biology and Ecology*, *487*, 113–126. https://doi.org/10.1016/j.jembe.2016.12.002

Hawkins, S. J., Sugden, H. E., Mieszkowska, N., Moore, P. J., Poloczanska, E., Leaper, R., … Burrows, M. T. (2009). Consequences of climate-driven biodiversity changes for ecosystem functioning of north European rocky shores. *Marine Ecology Progress Series*, *396*, 245–259. https://doi.org/10.3354/meps08378

Hebert, P. D. N., Ratnasingham, S., & DeWaard, J. R. (2003). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, *270*(SUPPL. 1), 96–99. https://doi.org/10.1098/rsbl.2003.0025

Hernvann, P. Y., & Gascuel, D. (2020). Exploring the impacts of fishing and environment on the Celtic Sea ecosystem since 1950. *Fisheries Research*, *225*(January), 105472. https://doi.org/10.1016/j.fishres.2019.105472

Hernvann, P. Y., Gascuel, D., Grüss, A., Druon, J. N., Kopp, D., Perez, I., … Robert, M. (2020). The Celtic Sea Through Time and Space: Ecosystem Modeling to Unravel Fishing and Climate Change Impacts on Food-Web Structure and Dynamics. *Frontiers in Marine Science*, *7*. https://doi.org/10.3389/fmars.2020.578717

Holt, J., Wakelin, S., Lowe, J., & Tinker, J. (2010). The potential impacts of climate change on the hydrography of the northwest European continental shelf. *Progress in Oceanography*, *86*(3–4), 361–379. https://doi.org/10.1016/j.pocean.2010.05.003

Huang, K., Zhang, P., Dunn, D. W., Wang, T., Mi, R., & Li, B. (2019). Assigning alleles to different loci in amplifications of duplicated loci. *Molecular Ecology Resources*, *19*(5), 1240–1253. https://doi.org/10.1111/1755-0998.13036

Huthnance, J. M., Holt, J. T., & Wakelin, S. L. (2009). Deep ocean exchange with west-European shelf seas. *Ocean Science*, *5*(4), 621–634. https://doi.org/10.5194/os-5-621-2009

ICES. 2019. Celtic Seas ecoregion – Fisheries overview, including mixed-fisheries considerations.

https://doi.org/10.17895/ices.advice.5708

ICES. 2020. Celtic Seas ecoregion – Fisheries overview, including mixed-fisheries considerations. https://doi.org/10.17895/ices.advice.7606

Jerde, C. L., Wilson, E. A., & Dressler, T. L. (2019). Measuring global fish species richness with eDNA metabarcoding. *Molecular Ecology Resources*, *19*(1), 19–22. https://doi.org/10.1111/1755-0998.12929

Kaiser, M. J., Bergmann, M., Hinz, H., Galanidi, M., Shucksmith, R., Rees, E. I. S., … Ramsay, K. (2004). Demersal fish and epifauna associated with sandbank habitats. *Estuarine, Coastal and Shelf Science*, *60*(3), 445–456. https://doi.org/10.1016/j.ecss.2004.02.005

Kirk, H., & Freeland, J. R. (2011). Applications and implications of neutral versus non-neutral markers in molecular ecology. *International Journal of Molecular Sciences*, *12*(6), 3966–3988. https://doi.org/10.3390/ijms12063966

Koslow, J. A., & Wright, M. (2016). Ichthyoplankton sampling design to monitor marine fish populations and communities. *Marine Policy*, *68*, 55–64. https://doi.org/10.1016/j.marpol.2016.02.011

Landry, C., & Bernatchez, L. (2001). Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon (Salmo salar). *Molecular Ecology*, *10*(10), 2525–2539. https://doi.org/10.1046/j.1365-294X.2001.01383.x

Larson, W. A., Lisi, P. J., Seeb, J. E., Seeb, L. W., & Schindler, D. E. (2016). Major histocompatibility complex diversity is positively associated with stream water temperatures in proximate populations of sockeye salmon. *Journal of Evolutionary Biology*, *29*(9), 1846–1859. https://doi.org/10.1111/jeb.12926

Lee, O., Nash, R. D. M., & Danilowicz, B. S. (2005). Small-scale spatio-temporal variability in ichthyoplankton and zooplankton distribution in relation to a tidal-mixing front in the Irish Sea. *ICES Journal of Marine Science*, *62*(6), 1021–1036. https://doi.org/10.1016/j.icesjms.2005.04.016

Lighten, J., Papadopulos, A. S. T., Mohammed, R. S., Ward, B. J., Paterson, G. I., Baillie, L., … Van Oosterhout, C. (2017). Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nature Communications*, *8*(1), 1–10. https://doi.org/10.1038/s41467-017-01183-2

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., … Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), 1204–1210. https://doi.org/10.1038/ng.3645

Manwell, C. & Baker, C. M. A. 1963 A sibling species of sea- cucumber discovered by starch-gel electrophoresis. Comp. Biochem. Physiol. 10, 39–53. (doi:10.1016/0010- 406X(63)90101-4.)

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS ONE*, *9*(12). https://doi.org/10.1371/journal.pone.0114639

Maunder, M. N., & Punt, A. E. (2013). A review of integrated analysis in fisheries stock assessment. *Fisheries Research*, *142*, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *5*(6). https://doi.org/10.1101/pdb.prot5448

Mieszkowska, N., Leaper, R., Moore, P., Kendall, M. A., Burrows, M. T., Lear, D., … Hawkins, S. J. (2006). Marine biodiversity and climate change: assessing and predicting the influence of climatic change using intertidal rocky shore biota. Commissioned Report No. 202 (ROAME No. F01AA402), *202*(20), 52. Retrieved from http://www.snh.org.uk/pdfs/publications/commissioned_reports/Report No202.pdf

Moriarty, M., Greenstreet, S. P. ., & Rasmussen, J. (2017). Derivation of Groundfish Survey Monitoring and Assessment Data Products for the Northeast Atlantic Area. Scottish Marine and Freshwater Science Report. *Scottish Marine and Freshwater Science*, *8*(16). https://doi.org/10.7489/1984-1

Pauly, D., & Zeller, D. (2003). The Global Fisheries Crisis as a Rationale for Improving the FAO's Database of Fisheries Statistics. *Fisheries Centre Research Reports*, *11*(6), 1–9.

Piertney, S. B., & Oliver, M. K. (2006). The evolutionary ecology of the major histocompatibility complex. *Heredity*, *96*(1), 7–21. https://doi.org/10.1038/sj.hdy.6800724

Pinnegar, J. K., Jennings, S., O'Brien, C. M., & Polunin, N. V. C. (2002). Long-term changes in the trophic level of the Celtic Sea fish community and fish market price distribution. *Journal of Applied Ecology*, *39*(3), 377–390. https://doi.org/10.1046/j.1365-2664.2002.00723.x

Quéré, N., Desmarais, E., Tsigenopoulos, C. S., Belkhir, K., Bonhomme, F., & Guinand, B. (2012). Gene flow at major transitional areas in sea bass (Dicentrarchus labrax) and the possible emergence of a hybrid swarm. *Ecology and Evolution*, *2*(12), 3061–3078.

https://doi.org/10.1002/ece3.406

Rätz, H. J., Dörner, H., Scott, R., & Barbas, T. (2010). Complementary roles of European and national institutions under the Common Fisheries Policy and the Marine Strategy Framework Directive. *Marine Policy*, *34*(5), 1028–1035. https://doi.org/10.1016/j.marpol.2010.03.001

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*(4), 361–395. https://doi.org/10.1111/j.1467-2979.2008.00324.x

Ruiz-Castillo, E., Sharples, J., Hopkins, J., & Woodward, M. (2019). Seasonality in the cross-shelf physical structure of a temperate shelf sea and the implications for nitrate supply. *Progress in Oceanography*, *177*(July 2018), 101985. https://doi.org/10.1016/j.pocean.2018.07.006

Sharples, J., Ellis, J. R., Nolan, G., & Scott, B. E. (2013). Fishing and the oceanography of a stratified shelf sea. *Progress in Oceanography*, *117*, 130–139. https://doi.org/10.1016/j.pocean.2013.06.014

Siegenthaler, A., Wangensteen, O. S., Benvenuto, C., Campos, J., & Mariani, S. (2018). DNA metabarcoding unveils multi-scale trophic variation in a widespread coastal opportunist. *Molecular Ecology*, (September), 1–18. https://doi.org/10.1111/mec.14886

Souche, E. L., Hellemans, B., Babbucci, M., MacAoidh, E., Guinand, B., Bargelloni, L., … Volckaert, F. A. M. (2015). Range-wide population structure of European sea bass Dicentrarchus labrax. *Biological Journal of the Linnean Society*, *116*(1), 86–105. https://doi.org/10.1111/bij.12572

Spurgin, L. G., & Richardson, D. S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1684), 979–988. https://doi.org/10.1098/rspb.2009.2084

Sumaila, U. R., Cheung, W. W. L., Lam, V. W. Y., Pauly, D., & Herrick, S. (2011). Climate change impacts on the biophysics and economics of world fisheries. *Nature Climate Change*, *1*(9), 449–456. https://doi.org/10.1038/nclimate1301

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Talarico, L., Babik, W., Marta, S., & Mattoccia, M. (2019). Genetic drift shaped MHC IIB diversity of an endangered anuran species within the Italian glacial refugium. *Journal of Zoology*, *307*(1), 61–70. https://doi.org/10.1111/jzo.12617

Tarnecki, A. M., Burgos, F. A., Ray, C. L., & Arias, C. R. (2017). Fish intestinal microbiome: diversity and symbiosis unravelled by metagenomics. *Journal of Applied Microbiology*, *123*(1), 2–17. https://doi.org/10.1111/jam.13415

Teletchea, F. (2009). Molecular identification methods of fish species : reassessment and possible applications, *Reviews in Fish Biology and Fisheries,* 19(3), 265–293. https://doi.org/10.1007/s11160-009-9107-4

Thomas, A. C., Jarman, S. N., Haman, K. H., Trites, A. W., & Deagle, B. E. (2014). Improving accuracy of DNA diet estimates using food tissue control materials and an evaluation of proxies for digestion bias. *Molecular Ecology*, *23*(15), 3706–3718. https://doi.org/10.1111/mec.12523

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.0165252

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, *34*(3), 275–305. https://doi.org/10.1051/gse:2002009

Walker, N. D., Boyd, R., Watson, J., Kotz, M., Radford, Z., Readdy, L., … Hyder, K. (2020). A spatially explicit individual-based model to support management of commercial and recreational fisheries for European sea bass *Dicentrarchus labrax*. *Ecological Modelling*, *431*(June), 109179. https://doi.org/10.1016/j.ecolmodel.2020.109179

Ward, R. D. (2000). Genetics in fisheries management. *Hydrobiologia*, *420*(1), 191–201. https://doi.org/10.1023/A:1003928327503

Ward, Robert D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1847–1857. https://doi.org/10.1098/rstb.2005.1716

Wells, J. V., & Richmond, M. E. (1995). Populations, metapopulations, and species populations: what are they and who should care?. *Wildlife Society Bulletin*, 458-462.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., & Ding, Z. (2012). Biodiversity soup: Metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, *3*(4), 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x

Chapter 2   Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples

This work was published as: Ratcliffe, F. C., Uren Webster, T. M., Rodriguez-Barreto, D., O'Rorke, R., Garcia de Leaniz, C., & Consuegra, S. (2021). Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples. Ecological Applications, 31(3), e02284.

2.1 Abstract

Accurate assessment of larval community composition in spawning areas is essential for fisheries management and conservation but is often hampered by the cryptic nature of many larvae, which renders them difficult to identify morphologically. Metabarcoding is a rapid and cost-effective method to monitor early life-stages for management and environmental impact assessment purposes but its quantitative capability is under discussion. We compared metabarcoding with traditional morphological identification to evaluate taxonomic precision and reliability of abundance estimates, using 332 fish larvae from multinet hauls (0-50m depth) collected at 14 offshore sampling sites in the Irish and Celtic seas. To improve quantification accuracy (relative abundance estimates), the amount of tissue for each specimen was standardised and mitochondrial primers with conserved binding sites were used. Relative family abundance estimated from metabarcoding reads and morphological assessment were positively correlated, as well as taxon richness (Rs=0.81, P=0.007) and diversity (Rs=0.90, P=0.002). Spatial patterns of community composition did not differ significantly between metabarcoding and morphological assessments. Our results show that DNA metabarcoding of bulk tissue samples can be used to monitor changes in fish larvae abundance and community composition. This represents a feasible, efficient and faster alternative to morphological methods that can be applied to terrestrial and aquatic habitats.

2.2 Introduction

Assessing larval community composition is needed to provide accurate information about spawning areas for fisheries management and conservation, but the location and dispersal of larval stages are largely unknown aspects of many fish life-cycles (Legrand *et al.,* 2019). Early life stages of organisms are particularly sensitive to abiotic stressors (Radchuk, Turlure, & Schtickzelle, 2013) and, for fish, understanding the quantitative relationship between environmental quality and population dynamics remains challenging (Rose 2000). Thus, larval monitoring provides critical information about population changes over time (Asch, 2015) to inform conservation and policy (Ellis, Milligan, Readdy, Taylor, & Brown, 2012; Borja, Elliott, Uyarra, Carstensen, & Mea, 2017), but its application is often hampered by the cryptic morphology of early life-stage organisms (Sigut et al., 2017, Brechon, Coombs, Sims, & Griffiths, 2013; Kimmerling et al., 2018).

Traditional fish larvae monitoring involves identifying each individual using a light microscope, counting myotomes, assessing pigmentation patterns and jaw morphology (Russel, 1976). Yet, identification keys are incomplete for many parts of the world (Becker, Sales, Santos, Santos, & Carvalho, 2015) and, where descriptions are available, morphological assessment is time consuming and requires specialist training (Brechon *et al.,* 2013). Morphological taxonomy also relies on the identifying features remaining intact for species level assignment (Russel, 1976), but damage is common during sampling (e.g. when using continuous plankton recorders), leading to misidentification and loss of valuable information (Richardson et al., 2006).

In cases where morphological identification is unfeasible, DNA sequencing technologies may be used to identify organisms, as long as their sequences are in the databases (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012). The development of high-throughput sequencing technology allows amplicon-based sequencing (metabarcoding) of multiple individuals of various species concurrently (i.e. bulk samples), providing a relatively quick method of processing many samples to obtain taxonomical information (Taberlet et al., 2012) and estimate biodiversity (Dopheide *et al.,* 2019). However, obtaining accurate absolute abundance (number of individuals) estimates through relative read abundance (RRA) from amplicon sequence data has remained challenging (Deagle et al., 2019; Lamb et al., 2019). This is because biases in RRA estimations can be introduced at different stages of the metabarcoding protocol, for example, cell and DNA quantity, mitochondrial copy number, extraction success and PCR amplification rates can vary between tissue type and species (Lamb et al., 2019; Piñol, Senar, & Symondson, 2019), leading to inaccurate estimates. Another source of bias can arise from unequal body size of individuals pooled within a bulk sample, which can be mitigated by size fractioning of organisms prior to extraction, increasing the reliability of RRA estimates (Elbrecht, Peinert, & Leese, 2017). The choice of primers and target region may introduce further bias (Deagle, Jarman, Coissac, Pompanon, & Taberlet, 2014). These biases have led to designing costly and

bioinformatically challenging metagenomic approaches (Tang et al., 2015; Kimmerling et al., 2018) or to the use of multiple loci (Richardson et al., 2015) to identify particular species and estimate their abundance.

Improving the reliability of abundance estimates is thus needed to make metabarcoding more useful for biodiversity monitoring, calculation of metrics such as diversity indices, as well as detection of natural shifts in multispecies community composition (Bohmann et al., 2014). Different approaches have been proposed to improve abundance estimates based on RRA, whilst still using a cost effective, single marker PCR approach (Thomas, Deagle, Eveson, Harsch, & Trites, 2016; Elbrecht, Peinert, & Leese, 2017). For example, using primers with widely conserved priming sites may reduce taxa specific biases (Krehenwinkel et al., 2017), although taxonomic resolution can be reduced due to highly similar sequences within a family (Thomsen et al., 2016).

Here, using a single mitochondrial marker (12S ribosomal RNA, considered highly specific in fish), we have refined the reliability of DNA metabarcoding abundance estimates by standardizing input material and choosing conserved primer binding sites. Using bulk fish larvae samples from the Irish and Celtic Seas, we compared the sensitivity and accuracy of this approach with traditional morphological identification, to assess whether metabarcoding can be a feasible and rapid alternative to traditional assessment for estimating fish larvae richness, diversity and community composition metrics.

2.3 Methods

Field sampling

Sampling was carried out onboard the RV Celtic Voyager between May 17th and May 26th 2018. Fish larvae (3-30mm) from 14 hauls (1 per site) were sampled using a MultiNet plankton sampler (Hydro-Bios, Kiel, Germany). Sites 1-8 and 12 were sampled with 1 oblique haul to 50m depth per site, filtering a mean volume of 215 ± 55 m^3 of water. Hauls 9-14 (with the exception of haul 12) consisted of two vertical hauls from the surface to 50m, filtering a mean volume of 38 ± 6m^3, which were pooled for each site. Fish larvae from each haul were separated from other zooplankton species and preserved in RNAlater (Qiagen) at room temperature for 24hrs, then refrigerated at 4°C until morphological identification.

Morphological Identification

Fish larvae ranged from 2mm-30mm total length. For morphological identification all larvae were first separated into major groupings based on body shape following the classification by Russel (1976) and subsequently assigned to family level. Assignment to genus and species where possible, was then carried out. Assignments were checked against the species descriptions first in Russel (1976), and, where possible, double checked against the description by Rodriguez, Alemany & Garcia (2017). For taxa which could not be confidently morphologically identified, DNA was extracted from 1 or more representative individuals (34 individuals of 16 taxa across the survey, Appendix S1: Table A1.1) using the Qiagen DNeasy Blood and Tissue kit (GmbH, Hilden, Germany) following the manufacturer's instructions. Extracted DNA was then amplified using the 12S V5 primers (Riaz, Shehzad, & Viari, 2011), cleaned using a Sodium Acetate/EtOH solution, resuspended in 10µl HiDi Formamide (Applied Biosystems) and analysed using Sanger Sequencing on an ABI 3730 DNA Analyser (Applied Biosystems). Resulting sequences were aligned in BioEdit (v 7.2.5), and input to BLAST and BOLD databases to confirm species identity. When 12S barcoding did not resolve taxonomic identification to species level, due to database limitations or synonymous sequences, the barcoding region of ~650 bp of the CO1 gene (F1, R1, Ward, Zemlak, Innes, Last, & Hebert, 2005) was used to update taxonomic assignment to the lowest possible taxonomic level, resulting in six additional 12S reference sequences not present in the NCBI nucleotide database (Appendix S1: Table A1.1; Genbank accession numbers: MN539950, MN539961, MN539952, MN539964, MN539965, MN539966). Taxonomy of Sanger sequenced individuals was assigned to the lowest possible level using the MegaBLAST algorithm (Morgulis et al., 2008) against the National Center for Biotechnology Information (NCBI) GenBank nucleotide database (accessed November, 2018). To estimate accuracy and repeatability of taxonomic

assignments, a group of 15 specimens were also sent to an experienced taxonomist and verified by CO1 barcoding (Morphological taxonomic assignment concordance test).

DNA extraction

After taxonomic identification, bulk tissue samples from all larvae of each haul were prepared for DNA extraction as follows: 2-8 mg of tissue were cut from the area anterior to the tail of each juvenile fish (for individuals <5mg, the entire larva was used, n=88) and placed in a Falcon tube on ice. Buffer ATL and proteinase K (Qiagen DNeasy Blood and Tissue kit, GmbH, Hilden, Germany) were then added to the pooled tissue sample in a ratio of 180μl of ATL and 20μl proteinase K for 15mg of tissue. Each falcon tube (representing a haul) was vortexed thoroughly and incubated overnight to digest at 56°C, shaking at 65 rpm. Samples were visually inspected for tissue remnants, vortexed and re-incubated until all tissue dissolved. Digestions from each haul were then vortexed for 45 seconds to ensure thorough mixing of digested products and divided in three sub-samples of 200ul that were extracted using the Qiagen DNeasy Blood and Tissue kit, following the manufacturer's instructions. Extraction blanks were carried through each step of the process.

Library preparation and sequencing

A 106 bp fragment of the 12S mitochondrial gene was amplified with the 12S V5 primers (Riaz et al., 2011) using Phusion® High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Loughborough, UK), with an annealing temperature of 52°C, in 3 extraction replicates per haul. Libraries were prepared using a 2-step PCR approach, based on the Illumina 16S Metagenomic Sequencing Library preparation guidelines (Illumina, Inc., San Diego, CA, USA), with following adaptations: in the first PCR step, each extraction replicate was amplified in triplicate in order to increase detection of rare species (Alberdi, Aizpurua, Gilbert, & Bohmann, 2018). Subsequently, 10ul from each triplicate were pooled prior to first cleanup. Cleanups were performed using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA), using a 1.2 x volume of beads to PCR product. Amplicons were indexed using Nextera XT Index Kit v2 Set C (Illumina, Inc., San Diego, CA, USA), and DNA concentration of each reaction was quantified via Qubit dsDNA HS Assay (Invitrogen, Carlsbad, CA USA) and pooled in equal molar concentrations. PCR and extraction blanks (using molecular grade water instead of template) were subjected to all steps of the library preparation process. In addition, a sequencing/tag jumping blank, where no sample was added prior to sequencing, was used. Pair-end sequencing was carried out at Swansea University using an Illumina MiSeq platform (Illumina, San Diego, CA, USA) (2x300bp reads), including 5% PhiX.

Bioinformatics- sequence processing

De-multiplexed samples containing raw pair end sequences were processed using Qiime2 (version 2019.1, Bolyen et al., 2019). Initially, raw sequences were quality checked using interactive quality plots, in order to obtain values for sequence trimming and truncation. De-noising was carried out using DADA2 (Callahan, McMurdie, & Holmes, 2017) where the first 10 bp of each sequence were trimmed to remove adaptors and all sequences truncated to 100 bp in length based on quality scores. Default DADA2 settings within Qiime2, were used to detect and, where possible, correct sequencing errors and filter out phiX reads, and chimeric sequences, join pair end reads and de-replicated sequences. The amplicon sequence variant (ASV) approach was chosen because it provides a higher resolution than a traditional OTU approach, enabling detection of single nucleotide differences (Callahan et al., 2017). After de-noising, the ASV and BIOME tables were exported for taxonomic assignment.

Database construction and taxonomic assignment

A custom database was constructed using *in silico* PCR against the NCBI database (downloaded February 2019): 12S V5 primers were allowed to have 3 base mismatches *in silico* (search_PCR command, Edgar, 2010) and a corresponding taxonomy file was constructed using the obiannotate tool (OBITools, Boyer et al., 2016). All sequences were trimmed to the target region. A list of all marine fish species encountered in the British Isles, including non-native fish (366 species) (Fish Base: accessed 31/3/2019) was then used to filter the main database to fish species present in the study region, of which 207 were available. The six 12S Sanger sequences (generated with the 12S V5 primers) missing from NCBI database and verified using CO1 barcoding from this study were added to the database (Appendix S1: Table A1.1) which also included marine mammals, bacteria and other contaminants (such as *Homo sapiens*) that might be amplified by the primers.

Initially, ASVs were classified using the KNN method in Mothur (Schloss et al., 2009) using parameter 'numwanted=1' (Findley et al., 2013), against the custom database. Because this parameter may lead to false positive assignments, KNN assignments were then verified using NCBI megaBLAST, with max-target sequences =10. The top 10 assignments were screened for UK species (Fish Base) on a case by case basis. Where the percentage of UK species match fell below 98%, or where multiple UK species matched above a 98% match, MEGAN (6.15.1) was used to assign species to the lowest common ancestor (Huson, Auch, Qi, & Schuster, 2007). ASVs for which there were no vertebrate matches were discarded from downstream analysis.

Tag jumping/cross contamination (Schnell, Bohmann, & Gilbert, 2015) was removed on the following basis: a taxa was removed from a haul if it had less than 115 reads (maximum reads for a single species in tag jumping control sample) or did not appear in all 3 replicates.

For spatial analysis, numbers of individuals of each taxon in a haul were estimated from the proportion of reads in the corresponding sample, as follows (equation 1):

$$A_i = N \times P_i$$

Where A_i is the abundance (number of individuals) of the taxon of interest (i) in a given haul, N is total number of individuals in the haul, P_i is the proportion of that taxon in the haul amplicon pool.

Statistical analysis

The accuracy of estimates of RRA and diversity indices derived from metabarcoding was assessed against results from morphological taxonomy using Spearman's rank correlation analysis performed in R (version 3.5.2). Diversity indices (Shannon Weiner Index, Simpson's Diversity) and richness were estimated based on RRA and morphological relative abundances using the Vegan package (R version 3.5.2) for both lowest possible taxonomic and family level taxon identifications. For spatial analysis, the survey area was divided into 3 locations along a temperature gradient: Loc 1 (North of the Celtic/Irish sea front, 9-10.99°C), Loc 2 (Channel spawning grounds, 11-12.99°C), Loc 3 (Western Celtic Sea, 13-14°C) (Figure 2.1). The number of individuals (assessed morphologically)and estimated from reads (equation 1), of a given taxon (mean of the 3 technical replicates per site) were divided by the volume of water filtered in the corresponding haul (Canfield & Jones, 1996) to obtain catch per unit filtered (CPUF) or estimated number of individuals from reads per unit filtered (RPUF) values respectively. This analysis was carried out at both lowest possible taxonomic level and family level. All 14 hauls surveyed were included in this analysis, where only 1 individual was present in a haul this was divided by the volume of water filtered and included in the both the CPUF and RPUF datasets. The family Ammodytidae was excluded from this analysis, because not all individuals were retained in haul 4. CPUF and RPUF values were square-root transformed, and composition similarity calculated by hierarchical clustering using a Bray-Curtis resemblance matrix. Subsequently, pairwise analysis of similarities (ANOSIM) were used to test whether there was a significant difference in community composition between locations (Clarke, 1993), using both the CPUF and RPUF methods. Where significant differences were detected, SIMPER analysis (Clarke, 1993) was used to ascertain which taxa accounted for the differences observed. Diversity indices calculations and multivariate spatial analyses were performed using Primer-v7 (Clarke & Gorley, 2015).
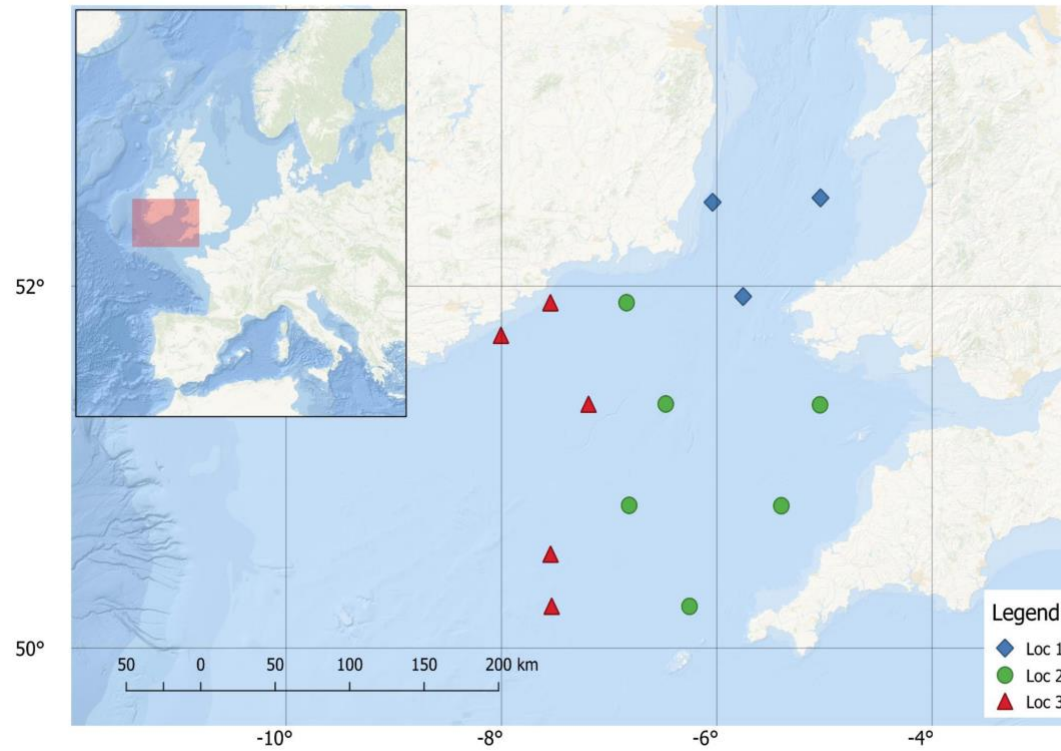
Figure 2.1. Multinet haul locations in the Irish and Celtic seas. Locations for spatial analysis, based on SST, are indicated as Loc 1 (Above the Celtic/Irish sea front: 9-10.99°C), Loc 2 (Channel spawning grounds: 11-12.99°C), Loc 3 (Western Celtic Sea: 13-14°C).

2.4   Results

Morphological assessment

A total of 332 fish larvae were caught in 11 of the 14 hauls in the survey. No larvae were encountered in hauls 10, 11, and 14 and only one in hauls 1 and 6, therefore 9 of the 14 hauls were used in metabarcoding. The maximum number of individuals per haul was 63 (haul 2) (Appendix 1: Table A1.2). Morphological identification assigned 324 (98%) of individuals to family level. It was not possible to assign the families of the remaining 8 larvae, due to damaged identifying features. Of those specimens assigned to family level, 255 (77%) were assigned to a genus and 100 (30%) to a species. Sanger sequencing to check morphological assignment comprising of 34 individuals across 9 hauls), contained 15 taxa (Appendix 1: Table A1.3). In the morphological taxonomic assignment test of the 15 individuals identified by two independent observers and subsequently checked by Sanger sequencing, 100% and 93% were correctly assigned to family level by the first and second observer respectively, 86.3% and 53.3% to genus and 40% to species level in both cases.

Based on morphology alone, before verification with CO1 barcoding, taxa within Ammodytidae and Clupeidae could not be assigned further than family level. Most clupeids did not amplify with the CO1 primers and those that did were assigned to *S. sprattus*; for Callionymus there was no *C. reticulatus* sequence to compare with. Incorrect morphological assignments occurred in the cases of *Micromesistius poutassou* (Sanger seq: *M. merlangius*), *Aphia minuta* (Sanger seq: *C. harengus/ S. sprattus*) and *Mugilidae* (Sanger seq: *L. bergylta* and *C. mustela*) (Figure 2.2, Appendix 1: Table A1.3). In addition, *Sardina pilchardus, Labrus mixtus/bergylta, Molva molva,* and a taxon belonging to the Gobiidae family were only detected using sequencing. In contrast, *M. merlangus,* and *Pollachius virens/pollachius,* were identifiable through morphology, but not resolved to species level by 12S metabarcoding due to lack of variability of the 12S fragment. *C. harengus* and *S. sprattus* could not be separated by morphology or 12S metabarcoding.
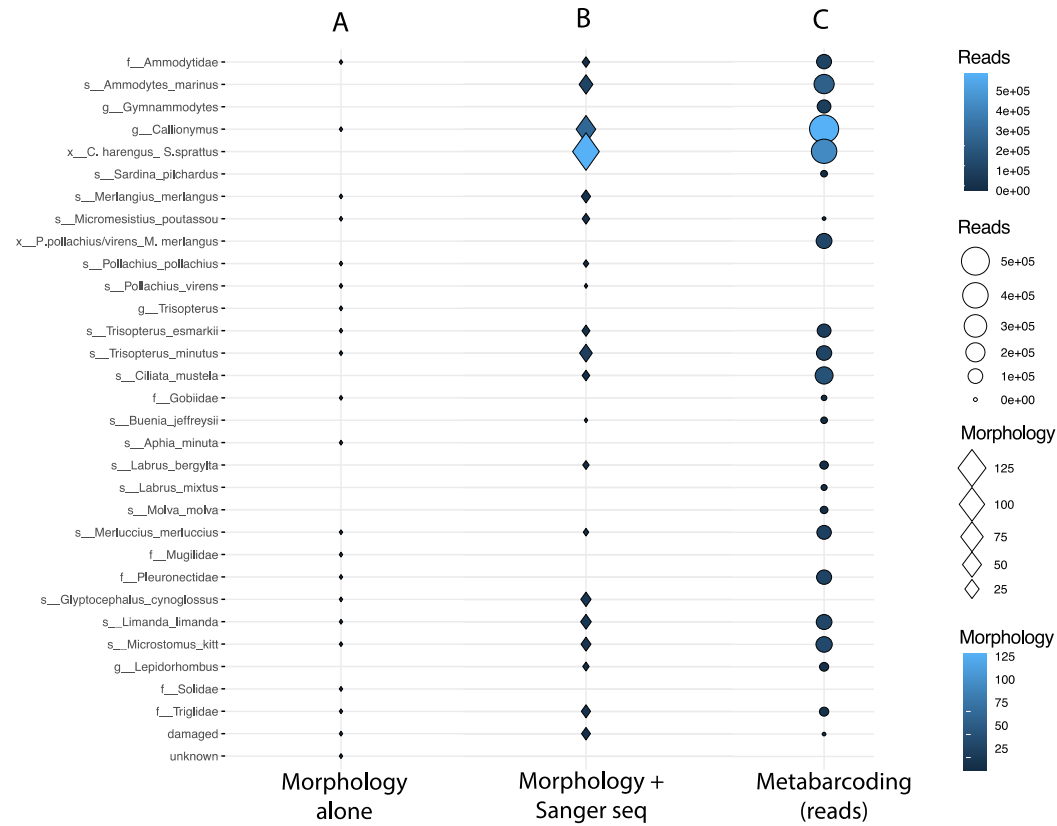
Figure 2.2 Overview of larval detections during the survey. Panel A: Taxonomic assignments using morphology alone (presence/absence). Panel B: Morphological taxonomic assignments updated with Sanger sequencing, diamonds represent total number of larvae of a taxa observed during the survey. Panel C: Metabarcoding taxonomic assignments, circles represent total number of reads obtained for each taxa, post-filtering.
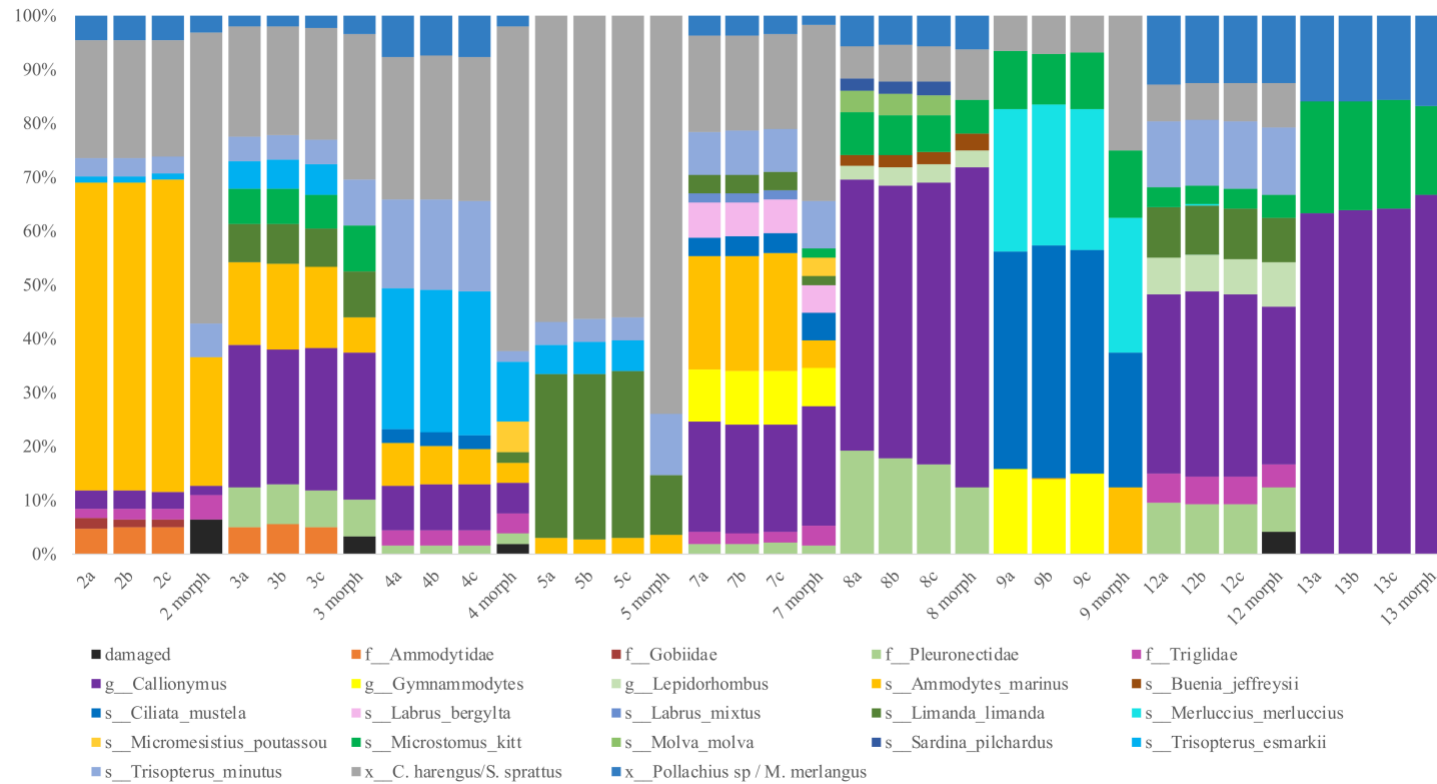
Figure 2.3 Comparison of relative read abundances (3 replicates per haul, 'a', 'b', 'c' samples) and morphological taxonomic assignments, corrected by Sanger sequencing (1 per haul, 'morph samples'). f__indicates family level assignment, s__ indicates species level and x__ indicates 2-3 possible species assignments. Morphological assignments of *P. pollachius/virens, M. merlangus* were grouped and morphologically assigned *Glyptocephaus cycnoglossus* has been re-assigned to Pleuronectidae to match metabarcoding assignments to aid visual interpretation of abundances.

Metabarcoding assessment

A total of 3,398,391 raw 300 bp pair end reads were generated for this study. After Qiime2 DADA denoising, a total of 2,675,140 reads remained for downstream analysis. Once the taxonomic assignment was complete, reads likely present due to tag jumping from concurrent sample sequencing (*Solidae* 274 reads, *Scomber scombrus*, 10 reads, *Salmo salar*, 3 reads), and human reads (2,338) were removed from downstream analysis. A total of 49 fish ASVs remained for downstream analysis. Samples contained a mean of 93,223 reads (standard deviation = 31,866) post filtering, the tag jumping blank contained 146 reads, the PCR blank 64 reads and extraction blanks 116 and 71 reads, respectively. Tag jumping read removal resulted in 0.046% of reads being excluded from downstream analysis across the samples in the study. Post filtering, taxa distribution was concordant among the 3 haul replicates in all 9 hauls (Figure 2.3).

Comparison of abundance estimates by morphology and metabarcoding

The relative abundance (%) of individuals identified morphologically in a sample and the corresponding RRAs were positively correlated for all families assessed (Spearman's rank: Ammodytidae R_s =0.93, P<0.001, Callionymidae R_s = 0.99, P<0.001, Clupeidae R_s = 0.97, P<0.001, Gadidae R_s =0.95, P<0.001, Pleuronectidae R_s = 0.68, P = 0.05, Triglidae R_s = 0.88, P = 0.002, Appendix 1: Figure A1.1). In addition, comparable levels of diversity and taxon richness were detected between the relative abundance of morphological assignments and RRA assignments at either lowest possible taxonomic level or family level, across hauls (lowest possible taxonomic level, Spearman's Rank: richness: R_s = 0.84, P=0.005, Shannon Index: R_s = 0.90, P=0.002, Simpson's Diversity: R_s = 0.90, P= 0.002. Family level, Spearman's Rank: richness: R_s = 0.93, P<0.001, Shannon Index: R_s 0.91, P=0.001, Simpson's Diversity: R_s = 0.80, P= 0.01, Figure 2.4).
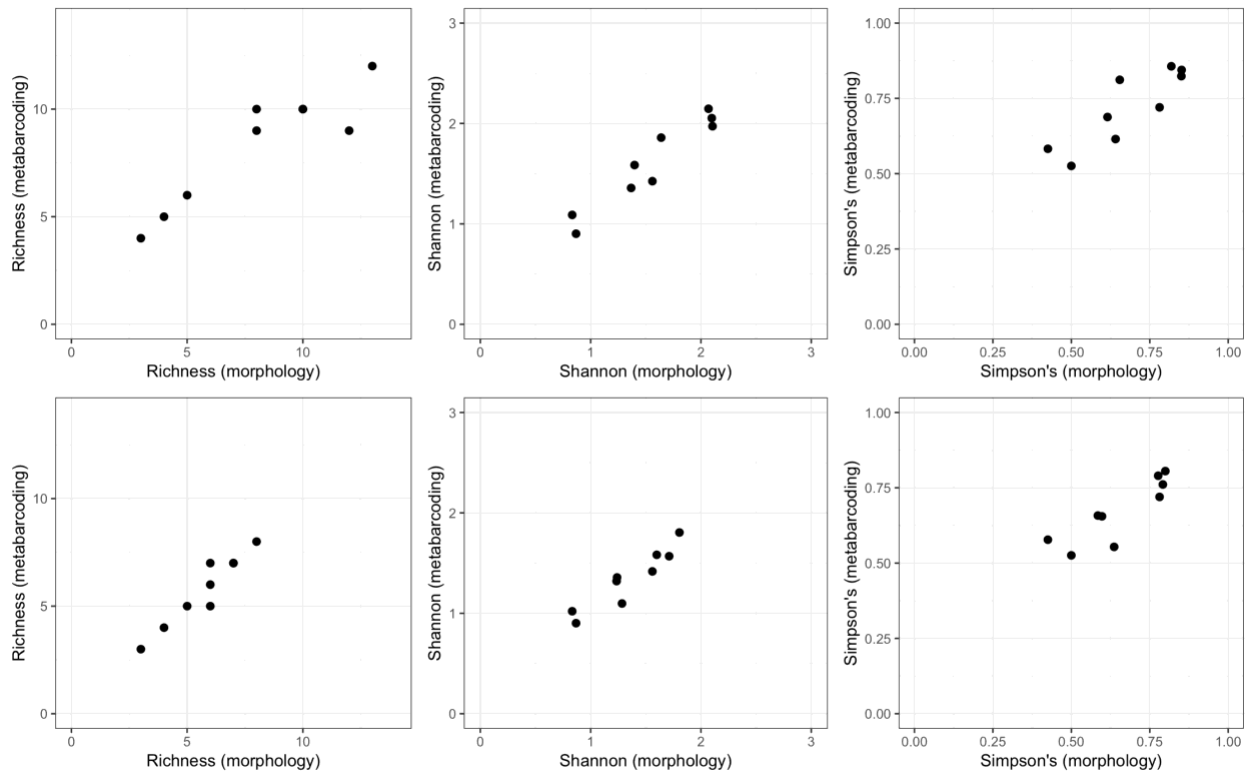
Figure 2.4. Consistency of diversity metrics between relative abundances of morphological assignments and relative read abundance assignments, post bioinformatic filtering (mean of 3 technical replicates per site, for 9 sites in the study where >1 larvae was found), for a) species richness, b) Shannon Wiener diversity index and c) Simpson's diversity (1-lamda). Rs = Spearman's rank Rho values.

Spatial distribution of larvae assessed by both methods

Assessment of patterns in community composition yielded comparable results from morphological and metabarcoding assessment at both lowest possible taxonomic level and family level. Catch per unit filtered (CPUF) and back-estimated reads per unit filtered (RPUF), were no different between locations 1 and 2, and 1 and 3, although locations 2 and 3 differed in composition (lowest possible taxonomic level: ANOSIM CPUF R=0.233, P = 0.039, RPUF R=0.209, P = 0.045. Family level: ANOSIM CPUF R=0.22, P = 0.041, RPUF R=0.205, P = 0.048, Table 2.1). SIMPER analysis (% cumulative dissimilarity contribution) attributed 48.39% (CPUF) and 42.82% (RPUF) of the difference in composition between locations 2 and 3 to three taxa: *C. harengus/S. sprattus* (CPUF: 21.42%, RPUF: 15.74%), *Triglidae* (CPUF: 14.37%, RPUF: 14.43%) and *Callionymus* (CPUF: 12.61%, RPUF: 12.65%) (Appendix 1: Table A1.4, Figure A1.2). The greatest difference observed in dissimilarity contributions for the remaining, less abundant taxa was 2.7% (*C. mustela*). This pattern was repeated at family level (Appendix 1: Table A1.4; Figure A1.3).

Table 2.1 ANOSIM matrix, showing R values of pairwise comparisons of community composition between 3 locations in the Irish/Celtic seas, using morphological taxonomic assignments and abundances (CPUF) and metabarcoding taxonomic assignments and back-estimated abundances (RPUF). * indicates significant difference in community composition between 2 locations.

Lowest Possible taxonomic Level

		RPUF		
		Loc 1	Loc 2	Loc 3
	Loc 1		-0.123	-0.013
CPUF	Loc 2	-0.111		0.209*
	Loc 3	-0.053	0.233*	

Family Level

		RPUF		
		Loc 1	Loc 2	Loc 3
	Loc 1		-0.130	-0.073
CPUF	Loc 2	-0.123		0.205*
	Loc 3	-0.087	0.220*	

2.5   Discussion

Here we demonstrate that metabarcoding is a reliable and practical alternative to traditional morphological assessment. We show that RRA estimates can be achieved by standardising the amount of tissue analysed per specimen and choosing primers with conserved binding sites. These estimates can then be used to successfully calculate diversity and community composition metrics needed to monitor changes over time. Although more costly in terms of consumables and sequencing, metabarcoding involved considerably less time than morphological identification, particularly for those cases which required additional barcoding to refine the morphological identification. We could use all the individuals collected for metabarcoding, irrespective of their preservation state, while the presence of damaged or poorly preserved specimens made difficult or even impossible their morphological identification, particularly at the species level.

There is considerable debate over whether amplicon sequencing can deliver reliable quantitative data (Deagle et al., 2019). The reliability of abundance estimates from metabarcoding varies considerably between studies, with some showing only a weak correlation between RRA and abundance (Lamb et al., 2019; Piñol, Senar, & Symondson, 2019). Still, information from RRA tends to be more informative than presence/absence assessments (Deagle et al., 2019). In contrast, metagenomic approaches that do not require PCR amplification can successfully estimate abundance (e.g. Kimmerling et al., 2018), although the costs and bioinformatic complexity of this approach may be prohibitive in many contexts (Porter & Hajibabaei, 2018). As shown here, amplicon sequencing can be used to estimate abundance and we suggest that further refinements in metabarcoding abundance estimates will enable wider application of amplicon sequencing.

We have shown that the use of approximately equal weights of tissue per individual can improve RRA and diversity estimates. Approaches based on photographically assessing the surface area of taxa and modelling biomass might also eliminate the need for weighing tissue (Kimmerling et al., 2018), although not necessarily reducing time and costs. There are, however, several factors that can bias RRA estimates. For instance, mitochondrial copy number can vary, not only between different species (Piñol, Mir, Gomez-Polo & Agustí 2015), but between different tissue types (Wiesner, Rüegg & Morano, 1992). We mainly used the region anterior to the tail of each larva to account for one of these biases (tissue type) as much as possible but interspecific biases remain a challenge to the quantitative capabilities of metabarcoding techniques (Deagle et al., 2019). Where they are consistent for a given taxon across all samples within a study, correction factors may be applied (Thomas et al., 2016; Krehenwinkel et al., 2017). While we found differences in relative abundance between morphological assessment and metabarcoding, they did not impact the calculation of diversity and community metrics. For example, estimates of the number of individuals of Ammodytidae differed by 7 individuals (SD 9.32) compared to those assessed

morphologically but this was not sufficient to influence community composition. However, for applications where exact numbers of individuals are needed (e.g. census of particular species) these differences may require consideration.

While there is no perfect marker for all studies (Deagle et al., 2014), we have shown here the benefits of using primers with well conserved binding sites, particularly for RRA estimates. Whereas the CO1 marker has extensive sequence databases as well as a strong capability to discriminate between species, it also carries an increased risk of amplification bias due to the lack of conserved binding sites across a broad range of taxa (Deagle et al., 2014). This can result in false negatives where taxa known to be present in a sample do not amplify (Collins et al., 2019; Nobile et al., 2019). Using more conserved priming sites, such as the 12S marker, may reduce taxa specific biases (Krehenwinkel et al., 2017), although it has been argued that taxonomic resolution may be reduced due to lack of sequence variability within families (Thomsen et al., 2016), and the completeness of reference databases also influences the resolution to species level (Miya et al., 2015). Here, using 12S primers, 40% of the taxa identified with metabarcoding could be assigned to species level, with the rest being assigned to family or genus level. In comparison to morphological identification without the assistance of CO1 Sanger sequencing, 12S metabarcoding achieved higher taxonomic resolution and more accurate identifications to family level. Morphologically assessed groupings supported by barcoding with Sanger sequencing achieved a similar level of assignment at the family level to metabarcoding across the study. Yet, while short reads can struggle to resolve some families to species level (Thomsen et al., 2016), hindering species level data interpretation, we found that the use of metabarcoding improved taxonomic assignment overall. Morphology only performed better than sequencing in the case of *Glyptocephalus cynoglossus*, due to distinct morphological characteristics, and in a few cases due to lack of information or sequence variation at the 12S region. In general, synonymous sequences at the target region resulted in just two (e.g. *C. harengus/S. sprattus*) or three species (*M. merlangus/P. pollachius/P. virens*) not being distinguished from each other. For studies requiring species level identification taxa affected by lack of marker sequence information or variability, a targeted qPCR approach (Brechon et al., 2013), similar to those carried out to detect particular species using eDNA (Robinson et al,. 2019) or a family specific, multi primer approach (Riaz et al., 2011) could be easily used to refine metabarcoding assignments. Combining different markers, as we have done here with the 12S metabarcoding and the CO1 barcoding, can be used to refine the databases by adding novel sequences and by separating species which cannot be identified based solely on small fragments. The completeness of the database used as a reference is critical for the accuracy of the taxonomic assignments and, while databases are continuously increasing in size for the most common metabarcoding makers, given the large diversity of fish and the increasingly lower cost of sequencing, focusing on full mitochondrial genomes may have

wider relevance (Collins et al., 2019). We found that in some cases metabarcoding could only not detect species level, for some applications, genus level analysis provides similar diversity and community composition information than species level and would be appropriate, for example to detect responses to environmental change (Hernandez, Carassou, Graham, & Powers, 2013). In some other cases, family level analysis has been deemed sufficient to detect broadscale changes, e.g. after major environmental disturbance (Hernandez et al., 2013). Therefore, dependent on hypothesis, a single 12S analysis or an additional qPCR can be performed.

Spatial patterns detected in community composition remained the same, independent of whether they were assessed using morphological (CPUF) or metabarcoding (RPUF). The small differences in abundance of rare taxa (Appendix 1: Table A1.4), were mainly the result of miss-identification of *C. mustela* during morphological identification, indicating that metabarcoding of bulk samples may be used as a viable alternative to morphological identification of samples, particularly when the latter proves difficult.

All taxa detected in the survey were known to spawn in the survey area (Acevedo, Dwane, & Fives, 2002; Ellis et al., 2012) and for the family Ammodytidae, difficult to survey and data limited due to its cryptic morphology (Ellis et al., 2012), metabarcoding identified *A. marinus* and a species of the genus *Gymnammodytes,* further illustrating its potential for detecting cryptic species.

Conclusions

We have shown that using a single marker (12S), equal amounts of tissue per sample and estimation of number of individuals from RRA, metabarcoding can provide quantitative abundance estimates for the calculation of alpha and beta diversities. This method could be applied to bulk samples from different terrestrial and marine habitats to improve abundance estimates. Specifically, we recommend the use of markers with highly conserved binding sites and using a small, equally sized pieces of tissue from each specimen to minimise biases and handling steps. This provides a rapid, community level assessment method, that could be used to further understand responses to disturbance and inter-annual or seasonal variability and monitor biodiversity in a changing global climate.

Ethics approval

Sampling has been conducted following Home Office regulations and approved by Swansea University Ethics Committees under approval No. 181019/1996.

2.6 References

Acevedo, S., Dwane, O., & Fives, J. M. (2002). The community structure of larval fish populations in an area of the Celtic Sea in 1998. Journal of the Marine Biological Association of the United Kingdom, 82(4), 641–648. doi:10.1017/S0025315402006008

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. Methods in Ecology and Evolution, 9(1), 134–147. doi:10.1111/2041-210X.12849

Asch, R. G. (2015). Climate change and decadal shifts in the phenology of larval fishes in the California Current ecosystem. Proceedings of the National Academy of Sciences, 112(30), E4065–E4074. doi:10.1073/pnas.1421946112

Becker, R. A., Sales, N. G., Santos, G. M., Santos, G. B., & Carvalho, D. C. (2015). DNA barcoding and morphological identification of neotropical ichthyoplankton from the Upper Paraná and São Francisco. Journal of Fish Biology, 87(1), 159–168. doi:10.1111/jfb.12707

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., … de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. Trends in Ecology and Evolution, 29(6), 358–367. doi:10.1016/j.tree.2014.04.003

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857. doi:10.1038/s41587-019-0209-9

Borja, A., Elliott, M., Uyarra, M. C., Carstensen, J., & Mea, M. (2017). Editorial: Bridging the Gap between Policy and Science in Assessing the Health Status of Marine Ecosystems. Frontiers in Marine Science, 4(September). doi:10.3389/fmars.2017.00032

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. Molecular ecology resources, 16(1), 176-182.

Brechon, A. L., Coombs, S. H., Sims, D. W., & Griffiths, A. M. (2013). Development of a rapid genetic technique for the identification of clupeid larvae in the Western English Channel and investigation of mislabelling in processed fish products. ICES Journal of Marine Science, 70(2), 399–407. doi:10.1093/icesjms/fss178

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal, 11(12), 2639–2643. doi:10.1038/ismej.2017.119

Canfield, T. J., & Jones, J. R. (1996). Zooplankton abundance, biomass, and size-distribution in selected midwestern waterbodies and relation with trophic state. Journal of Freshwater Ecology, 11(2), 171–181. doi:10.1080/02705060.1996.9663476

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18(1), 117–143. doi:10.1111/j.1442-9993.1993.tb00438.x

Clarke, K. R., & Gorley, R. N. (2015). Getting started with PRIMER v7. PRIMER-E: Plymouth, Plymouth Marine Laboratory, 20.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., … Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. Methods in Ecology and Evolution, 2041–210X.13276. doi:10.1111/2041-210X.13276

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. Biology Letters, 10(September), 2–5. doi:10.1098/rsbl.2014.0562

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., … Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? Molecular Ecology, 28(2), 391–406. doi:10.1111/mec.14734

Demir, N., Southward, A. J., & Dando, P. R. (1985). Comparative notes on postlarvae and pelagic juveniles of the rocklings gaidropsaurus mediterraneus, rhinonemus cimbrius, ciliata mustela and c. septentrionalis. Journal of the Marine Biological Association of the United Kingdom, 65(3), 801–839. doi:10.1017/S0025315400052607

Dopheide, A., L. K. Tooman, S. Grosser, B. Agabiti, B. Rhode, D. Xie, M. I. Stevens, N. Nelson, T. R. Buckley, A. J. Drummond, and R. D. Newcomb. 2019. Estimating the biodiversity of terrestrial invertebrates on a forested island using DNA barcodes and metabarcoding data. Ecological Applications 29:e01877. doi.org/10.1111/2041-210X.13086

Edgar, R. (2010). Usearch. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Elbrecht, V., Peinert, B., & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. Ecology and Evolution, 7(17), 6918–6926. doi:10.1002/ece3.3192

Ellis, J. R., Milligan, S. P., Readdy, L., Taylor, N., & Brown, M. J. (2012). Spawning and nursery grounds of selected fish species in UK waters. Retrieved from https://www.cefas.co.uk/publications/techrep/TechRep147.pdf

Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J. A., … Segre, J. A. (2013). Topographic diversity of fungal and bacterial communities in human skin. Nature, 498(7454), 367–370. doi:10.1038/nature12171

Gillet, B., Cottet, M., Destanque, T., Kue, K., Descloux, S., Chanudet, V., & Hughes, S. (2018). Direct fishing and eDNA metabarcoding for biomonitoring during a 3-year survey significantly improves number of fish detected around a South East Asian reservoir. PLoS ONE, 13(12), 1–25. doi:10.1371/journal.pone.0208592

Hernandez, F. J., Carassou, L., Graham, W. M., & Powers, S. P. (2013). Evaluation of the taxonomic sufficiency approach for ichthyoplankton community analysis. Marine Ecology Progress Series, 491(Fahay 2007), 77–90. doi:10.3354/meps10475

Hunt, B. P. V, & Hosie, G. W. (2003). The Continuous Plankton Recorder in the Southern Ocean: A comparative analysis of zooplankton communities sampled by the CPR and vertical net hauls along 140°E. Journal of Plankton Research, 25(12), 1561–1579. doi:10.1093/plankt/fbg108

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome Research, 17(3), 377–386. doi:10.1101/gr.5969107

Kimmerling, N., Zuqert, O., Amitai, G., Gurevich, T., Armoza-Zvuloni, R., Kolesnikov, I., … Sorek, R. (2018). Quantitative species-level ecology of reef fish larvae via metabarcoding. Nature Ecology & Evolution, 2(2), 306–316. doi:10.1038/s41559-017-0413-2

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. Scientific Reports, 7(1). doi:10.1038/s41598-017-17333-x

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. Molecular Ecology, 28(2), 420–430. doi:10.1111/mec.14920

Legrand, T., A. Di Franco, E. Ser-Giacomi, A. Calo, and V. Rossi. 2019. A multidisciplinary analytical framework to delineate spawning areas and quantify larval dispersal in coastal fish. Marine Environmental Research 151:104761. doi.org/10.1016/j.marenvres.2019.104761

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., ... & Kondoh, M. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA fromfishes: detection of more than 230 subtropical marine species. Royal Society open science, 2(7), 150088.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. Bioinformatics, 24(16), 1757–1764. doi:10.1093/bioinformatics/btn322

Nobile, A. B., Freitas-Souza, D., Ruiz-Ruano, F. J., Nobile, M. L. M. O., Costa, G. O., de Lima, F. P., … Oliveira, C. (2019). DNA metabarcoding of Neotropical ichthyoplankton: Enabling high accuracy with lower cost. Metabarcoding and Metagenomics, 3, 69–76. doi:10.3897/mbmg.3.35060

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular Ecology Resources, 15(4), 819–830. doi:10.1111/1755-0998.12355

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. Molecular Ecology, 28(2), 407–419. doi:10.1111/mec.14776

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology, 27(2), 313–338. doi:10.1111/mec.14478

Radchuk, V., Turlure, C., & Schtickzelle, N. (2013). Each life stage matters: The importance of assessing the response to climate change over the complete life cycle in butterflies. Journal of Animal Ecology, 82(1), 275–285. doi:10.1111/j.1365-2656.2012.02029.x

Riaz, T., Shehzad, W., & Viari, A. (2011). ecoPrimers : inference of new DNA barcode markers from whole genome sequence analysis, 39(21), 1–11. doi:10.1093/nar/gkr732

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). EcoPrimers: Inference of new DNA barcode markers from whole genome sequence analysis. Nucleic Acids Research, 39(21), 1–11. doi:10.1093/nar/gkr732

Richardson, A. J., Walne, A. W., John, A. W. G., Jonas, T. D., Lindley, J. A., Sims, D. W., … Witt, M. (2006). Using continuous plankton recorder data. Progress in Oceanography, 68(1), 27–74. doi:10.1016/j.pocean.2005.09.011

Richardson, R. T., Lin, C., Quijia, J. O., Riusech, N. S., Goodell, K., & Johnson, R. M. (2015). Rank-based characterization of pollen assemblages collected by honey bees using a multi-locus

metabarcoding approach. Applications in Plant Sciences, 3(11), 1500043. doi:10.3732/apps.1500043

Robinson, C.V., Webster, T.M.U., Cable, J., James, J. & Consuegra, S. (2018). Simultaneous detection of invasive signal crayfish, endangered white-clawed crayfish and the crayfish plague pathogen using environmental DNA. Biological Conservation, 222, 241-252. doi: 10.1016/j.biocon.2018.04.009

Rodriguez, J. M., Alemany, F., & Garcia, A. (2017). A guide to the eggs and larvae of 100 common Western Mediterranean Sea bony fish species. Retrieved from https://www.agris.fao.org

Rose, K. A. 2000. Why are quantitative relationships between environmental quality and fish populations so elusive? Ecological Applications 10:367-385. doi.org/10.1890/1051-0761(2000)010[0367:WAQRBE]2.0.CO;2

Russell, F.S. 1976. The eggs and planktonic stages of British marine fishes. Academic Press, London

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology, 75(23), 7537–7541. doi:10.1128/AEM.01541-09

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular Ecology Resources, 15(6), 1289–1303. doi:10.1111/1755-0998.12402

Sigut, M., Kostovćik, M., Sigutova, metabarcoding, standard barcoding, and morphological approach in the identification of host-parasitoid interactions. PLoS ONE, 12(12), 1–18. doi:10.1371/journal.pone.0187803

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology, 21(8), 2045–2050. doi:10.1111/j.1365-294X.2012.05470.x

Tang, M., Hardman, C. J., Ji, Y., Meng, G., Liu, S., Tan, M., … Yu, D. W. (2015). High-throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods in Ecology and Evolution, 6(9), 1034–1043. doi:10.1111/2041-210X.12416

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. Molecular Ecology Resources, 16(3), 714–726. doi:10.1111/1755-

0998.12490

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. PLoS ONE, 11(11). doi:10.1371/journal.pone.0165252

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1462), 1847–1857. doi:10.1098/rstb.2005.1716

Wiesner, R. J., Rüegg, J. C., & Morano, I. (1992). Counting target molecules by exponential polymerase chain reaction: copy number of mitochondrial DNA in rat tissues. Biochemical and Biophysical Research Communications, 183(2), 553-559. doi.org/10.1016/0006-291X(92)90517-O

Data Accessibility Statement

Sequences from metabarcoding have been deposited in the NCBI under accession reference BioProject PRJNA576002. Sanger sequences for the reference collection have been deposited in GenBank under accession numbers MN539918-MN539945 (CO-I) and MN539946- MN539976 (12S).

Chapter 3   A drop in the ocean: monitoring fish communities in spawning areas using environmental DNA

3.1   Abstract

Early life stages of aquatic organisms are particularly vulnerable to climatic stressors, however, they are difficult to monitor due to challenges in sampling and morphological identification. Environmental DNA (eDNA) from water samples represents an opportunity for rapid, non-destructive monitoring of aquatic community composition as well as single species monitoring. eDNA can also detect spawning events, although has not been yet tested in offshore spawning grounds. Here, we used metabarcoding of water samples to detect the presence of key fish taxa in spawning areas that are difficult to monitor using traditional means. We analysed DNA from water samples and fish larvae samples at 14 offshore sites, using 12S mitochondrial metabarcoding and compared taxa detections, diversity and community structure estimated by both sample types. Species richness and diversity did not differ between water and larvae samples. Both sample types detected a core of 12 taxa across the survey, with an average agreement in detections of 75% at sampling site level. Water samples detected two of the three most abundant taxa, the sandeel, *Ammodytes marinus,* and clupeids, *Clupea harengus/Sprattus sprattus*, at 31% and 38% more sites than larvae samples respectively, while *Callionymus* sp was more prevalent in larvae samples. Mackerel (*Scomber scombrus*) and blue whiting (*Micromestius poutassou*) were only detected in water samples despite sampling taking place at peak spawning times for these species. Our results demonstrate that eDNA metabarcoding provides a rapid and feasible monitoring method for the management of key taxa, such as sandeel, that cannot be easily monitored using traditional capture surveys.

3.2   Introduction

As climate stressors increase globally (IPCC, 2018), spatial and temporal monitoring of biodiversity is required to detect changes in both species composition (Poloczanska et al., 2013) and geographic range (Burrows et al., 2011). This information is critical to informing policy decisions and assessing the efficacy of conservation/management interventions (Douvere & Ehler, 2011; Geijzendorffer et al., 2016). In the marine environment, monitoring is sometimes complicated by inaccessibility (Bicknell, Godley, Sheehan, Votier, & Witt, 2016) or limited resources (Costello et al., 2010). In addition, the velocity of climate change and seasonal shifts in timing of temperature changes is, in some cases, greater in the ocean than on land (Burrows et al., 2011), therefore development of rapid (and feasible) monitoring methods are crucial to detecting the magnitude of these changes (Costello et al., 2010; Thomsen et al., 2012).

Monitoring spawning grounds and fish recruitment is essential in the face of these global pressures as larval and juvenile developmental stages of fishes are highly sensitive to environmental stressors (Pitois, Lynam, Jansen, Halliday, & Edwards, 2012). Temperature changes (Lee, Nash, & Danilowicz, 2005), prey availability (Régnier, Gibb, & Wright, 2017) and offshore construction (Cordes et al., 2016) may all impact recruitment. Globally, spawning areas are often protected by policy measures (Pastoors, Rijnsdorp, & Van Beek, 2000), such as MPAs (Christie et al., 2010) or restrictions in offshore development such as pile driving or oil drilling (La Védrine, 2014). However, current data or time series information on larval distribution and spawning aggregations within these sensitive areas tends to be limited (Greve, Prinage, Zidowitz, Nast, & Reiners, 2005; Kimmerling et al., 2018). Traditional larvae monitoring methods involve deploying a plankton net from a research vessel and morphologically identifying individual larvae (Habtes, Muller-Karger, Roffer, Lamkin, & Muhling, 2014). Morphological identification is challenging and time consuming (Brechon, Coombs, Sims, & Griffiths, 2013) and in, some cases, only accurate to family level (Ellis, Milligan, Readdy, Taylor, & Brown, 2012). Monitoring economic and ecological costs can be, therefore, high (Koslow & Wright, 2016) and as a consequence data on spawning distributions is globally sparse (Ellis et al., 2012; Kimmerling et al., 2018; Maggia et al., 2017). Furthermore, information on key taxa that are difficult to capture and identify using traditional trawls (e.g. sandeel), is currently lacking, even in areas that are surveyed more regularly (Ellis et al., 2012; Lynam et al., 2013).

Environmental DNA (eDNA) extracted from water samples is potentially a rapid, cost effective tool for monitoring species distributions (Lodge et al., 2012), thus reducing the need for destructive sampling (Bylemans et al., 2017). Metabarcoding analysis of eDNA can be useful for whole community/broad range assessment (Bohmann et al., 2014; Lacoursière-Roussel et al., 2018; Thomsen et al., 2016) and also perform well for target species monitoring (Harper et al., 2018), offering a potential alternative to traditional monitoring of spawning grounds. The use of eDNA has shown potential for detecting pulses

of spawning in freshwater systems where mass release of gametes result in sudden increases of concentration of mitochondrial DNA in the water column (Bylemans et al., 2017). In some cases, increases in eDNA concentration reveal the movement of adults toward a spawning area, rather than the release of gametes (Erickson et al., 2016), but its usefulness to detect spawning in the marine environment remains uncertain.

Although the non-destructive nature of eDNA would make it ideal for use in sensitive environments (Stat et al., 2019), or for the monitoring of rare or threatened species (Bylemans et al., 2017), understanding how eDNA sampling reflects or differs from traditional monitoring techniques can be challenging (Hansen, Bekkevold, Clausen, & Nielsen, 2018). For example, eDNA cannot distinguish among year classes, which in turn can be specifically targeted by physical sampling (Maruyama, Nakamura, Yamanaka, Kondoh, & Minamoto, 2014). In addition, eDNA may show differing dispersal patterns in the open ocean to those physically exhibited by the fish larvae themselves (Goldberg et al., 2016). In fact, comparisons of simultaneous water and visual/physical sampling in the marine environment have found variable level of agreement in detections between eDNA analyses and morphological taxonomy (Leduc et al., 2019; Stat et al., 2019; Thomsen et al., 2016). In order to address these discrepancies, we assessed both eDNA (water samples) and ichthyoplankton physical samples (larvae) using metabarcoding and an identical bioinformatics pipeline. We compared species detection levels, relative abundance and community composition in both sample types, taken simultaneously at the same locations from the Celtic and Irish seas.

3.3   Methods

Field sampling

Water and larvae samples were taken onboard RV Celtic Voyager in the Irish and Celtic seas at 14 sites in known spawning areas for Ammodytidae, Clupeidae, Gadidae, Scombridae and Pleuronectidae (Ellis et al., 2012) (Figure 3.1, Table A2.1). Sampling was carried out in May (17th to 26th, 2018), during or shortly after the spawning season for many of the fish species in the sampling area (Table A2.2). Filtration of water samples was carried out in an area physically separated from the processing of larvae samples. Gloves were changed between samples and surfaces cleaned using a 10% bleach solution, before and after each sampling event. Sea surface water samples were taken by rope and bucket (bleach sterilized and swilled in sea water from the same site). At each of the 14 sites, 400 ml from three replicate buckets of sea water (true biological replicates) and one control (de-ionized water) were filtered using a syringe (Terumo, Tokyo, Japan, 50ml), through polycarbonate filter holders, (25 mm, Cole-Parmer, IL, USA) containing a 0.22µm hydrophilic polyethersulfone filter (Merck Millipore, MA, USA). Filters were left in the holders to minimize contamination and holders were then filled with ethanol and stored at -20°C until extraction. After water samples had been taken, a larvae haul was conducted starting from the same coordinates, using MultiNet plankton sampler (Hydro-Bios, Kiel, Germany) that continuously filtered water from the surface to 50m depth and back to the surface (for volumes filtered, see Appendix 2: Table A2.1). Ichthyoplankton were separated from other zooplankton species and preserved in RNAlater at room temperature for 24hrs, then refrigerated at 4°C until lab processing. Temperature, salinity, and density were recorded at each sampling site (Appendix 2: Table A2.1).
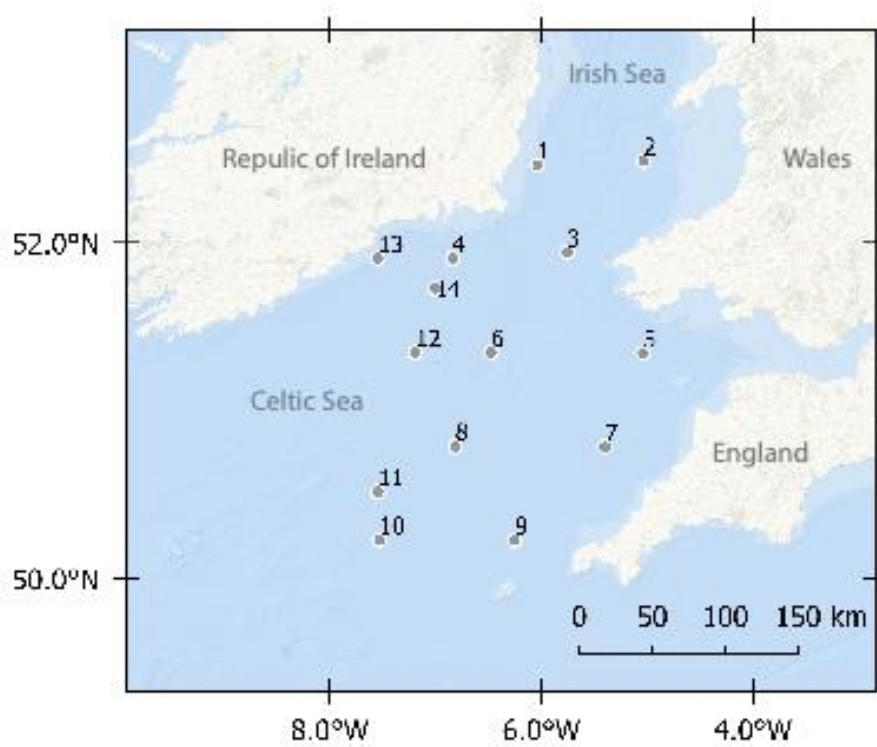
Figure 3.1 Map of the sampling locations (14 sites) across the Irish and Celtic sea region. For coordinates of each sampling location, see Appendix 2: Table A2.3.

Lab processing- DNA extraction and library preparation

All water samples were extracted using the QIAGEN PowerSoil kit, using a homogenization step (a Precellys 24 tissue homogenizer, Bertin Instruments, Montigny-le-Bretonneux, France). Extracted DNA was stored at -80°C until library preparation. Extraction blanks (where no filter was added), were carried through all steps of the library preparation and bioinformatic analysis.

Pooled homogenates of all fish larvae present in the MultiNet haul corresponding to each sampling site were extracted in bulk, by taking a 5mg (+/-3mg) of tissue anterior to the tail (or the complete larva for smaller specimens) from all individuals in a haul (Ratcliffe et al., 2020).

Environmental DNA libraries were prepared using 12S V5 primers (Riaz, Shehzad, & Viari, 2011) which amplify a 106bp fragment of the 12S mitochondrial gene. A nested PCR approach was optimized, using Platinum™ Hot Start PCR Master Mix (2X) (Thermo Fisher Scientific, Loughborough, U.K.). Initially, all samples were amplified in triplicate, in 12ul PCR reactions, with 2µl of template and 0.5µl of 12S primers, for 25 cycles, with an annealing temperature of 52°C. Subsequently, 5µl of this first reaction was used as a template for a second reaction, using the 12S primers with the addition of overhang adaptors for subsequent Nextera indexing, using identical PCR conditions, for 10 cycles. PCR triplicates were then pooled, using 10µl of each triplicate from the nested PCR reaction and purified using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA). Subsequently, amplicons were indexed using the Nextera XT Index Kit v2 (Illumina, Inc., San Diego, California, USA), and DNA concentration of each reaction was quantified via Qubit dsDNA HS Assay (Invitrogen, Thermo Fisher Scientific, Loughborough, U.K) and pooled in equal molar concentrations. Filtration blanks for each site, extraction blanks for each round of extractions and PCR blanks were carried through all steps of library preparation, including triplication in the first PCR, and bioinformatics processing.

Larvae samples (and associated extraction and PCR blanks) were directly amplified using the 12S primers with Nextera overhang adapters attached, using Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific, Loughborough, U.K) as described in Ratcliffe et al. (2020).

The effect (if any) of the different DNA polymerases and the nested and non-nested PCR protocols was assessed using mock communities, constructed from known molar concentrations of DNA extracted individually from larvae in the survey. Community 1 contained 12.5% of the following taxa: *Ammodytes marinus, Callionymus sp., Ciliata mustela, Lepidorhombus sp, Merlangus merlangius, Merluccius merluccius, Sprattus sprattus, Trisopterus minutus*. Community 2 contained *A. marinus* (1.83%), *Callionymus sp.,* (23.77%), *C. mustela* (2.64%), *L. boscii* (18.78%), *M. merlangius* (26.92%), *M. merluccius* (22.10%), *S. sprattus* (1.66%) *T. minutus* (2.30%). Water and larvae/mock community libraries were prepared separately to avoid contamination and were pooled after indexing and sequenced in the same Illumina Miseq run to avoid any sequencing bias. A tag jumping/sequencing control, where

62

no sample was added, was used to monitor any effect of concurrent sequencing of water and larvae samples.

Bioinformatic analyses

Water, larvae and mock community and all blanks were subjected to the same bioinformatics pipeline and processed simultaneously. The amplicon sequence variant (ASV) approach was used because it enables detection of single nucleotide differences (Callahan, McMurdie, & Holmes, 2017) and therefore provides a higher resolution than a traditional OTU approach.

Qiime2 (version 2019.1, Bolyen et al., 2019) was used to process de-multiplexed paired end sequences. DADA2 (Callahan et al., 2016) within Qiime2 was used for de-noising steps. Based on read quality scores, the first 10bp of each sequence were trimmed and all sequences truncated to 100bp in length. Subsequently, sequencing errors were corrected where possible, and chimeric sequences removed, paired end reads were joined and sequences de-replicated using the default DADA2 settings in Qiime2.

Taxonomic assignment was conducted using custom databases (Ratcliffe et al., 2020). Initially, reads were classified against a full database that included all taxa available on NCBI amplified *in silico* with the 12S V5 primers, using the KNN method in mothur (Schloss et al., 2009). A second screening was then carried out using a smaller database that included all available sequences of fish encountered (native and non-native) in the British Isles (Fish Base: accessed 31/3/2019) as well as outgroups known to be present from the classification against the full database, again using the KNN method, and the parameter 'numwanted=1' (Findley et al., 2013). Due to the potential for false positive assignments using this parameter, these assignments were verified using NCBI megablast (Morgulis et al., 2008), where the top 10 hits were screened on a case by case basis, for the highest match of a UK fish taxon. Assignments below 98% identity were assigned to genus or family level using the lowest common ancestor algorithm (Huson, Auch, Qi, & Schuster, 2007) in MEGAN (6.15.1).

Once taxonomic assignment was complete, non-fish ASVs were removed from downstream analysis. To remove contaminant ASVs (false positives) from the data, filtration blank read counts (subjected to the same workflow at all steps of the process as field samples) were subtracted from each field replicate in the corresponding site, (Grey et al., 2018). This was carried out before any downstream data analysis (Andruszkiewicz et al., 2017).

Statistical analysis

All statistical analysis was performed using R (version 3.5.3, R Core Team, 2017). For comparison purposes, as only one larvae haul was sampled per site, the reads from the three water biological replicates and three larvae technical replicates were pooled for each site. Reads were then converted to relative abundance data (proportion) to account for unequal sequencing depths between samples. Subsequently, to remove any reads present due to tag jumping (Schnell, Bohmann, & Gilbert, 2015) taxa that accounted for less than 0.05% (set using the tag jumping blank) of the relative abundance of a sample were removed from that site for downstream analysis.

Mock communities were analysed using a chi square (goodness of fit) test to ascertain if there was any difference between relative abundance of genomic DNA in the sample (expected) and relative abundance of reads after sequence processing (observed). Species richness was calculated for each sample using 'specnumber' in R and Shannon Wiener diversity was calculated using the 'diversity' function with 'method = 'Shannon'. Wilcoxon signed rank tests (paired samples) were used to ascertain if medians differed between water and larvae samples and variances within each treatment were calculated as Median Absolute Deviation (MAD), using r function 'mad'. After checking for normal distribution using 'skewness' (Moments package, an F-test was used to check for significant differences in variance between the sample types. Log likelihood ratio (G-test) test of independence with Williams' correction were used to test the effect of sample site and taxon on detections between the two methods.

To assess differences in community composition between the two methods, a dummy number of species of 1 was added to all samples in order compute Brae-Curtis dissimilarity for sites where no larvae were captured (Clarke, Somerfield, & Chapman, 2006). Reads (relative abundance) were square-root transformed, a Bray Curtis dissimilarity matrix was generated and PERMDISP (to test for homogeneity of dispersion) and subsequent PERMANOVA analysis was used (Anderson, 2014) (Appendix 2: Figure A2.1). SIMPER analysis was then carried out on untransformed relative abundance data to ascertain which species were driving the differences observed between the two sampling methods (Clarke, 1993).

3.4   Results

A total of 42 water biological replicate samples from 14 sites (three per site) were collected in the survey. At nine of the sites, the larvae hauls contained multiple larvae suitable for metabarcoding, however at two sites the larvae hauls contained only one individual and three sites the hauls did not contain larvae.

A total of 13,149,751 raw paired end reads were generated from the water and larvae samples. After DADA denoising, 7,379,309 reads remained for downstream analysis (Table 3.1). A total of 209 ASVs were generated across all samples in the study. Of these, 95 matched to fishes, 71 had no vertebrate match, 36 matched to family Hominidae, three to Delphinidae, two to Felidae and one to Laridae and Phasianidae respectively. The primary contaminant observed in filtration blanks was *Homo sapiens,* however a proportionally small amount (3.5% of blank reads) mapped to fish (*Salmo salar*, not found in any field replicates, *A. marinus, C. harengus/S. Sprattus* and *Pollacchius sp/M. merlangus*) and were used to set a cut-off below which reads were subtracted from each field replicate in the corresponding site. Filtration blank read removal resulted in the removal of 0.36% of water reads across the study. After pooling site replicates for analysis, the mean number of water sample reads per site was 223,745. Site 11 contained only 108 reads (no larvae were physically captured at this location) and was therefore discarded from downstream analysis. The mean number of larvae reads per site (nine sites) was 279,667.

Table 3.1 Number of reads remaining and removed at each step of the denoising process. Denoising was carried out using DADA2 (Callahan et al., 2016) within Qiime2 (version 2019.1, Bolyen et al., 2019). Removal of non-vertebrate and non-fish reads was conducted after taxonomy had been assigned.

	Reads remaining	Reads removed
Raw reads	13096645	
Filtered	7961080	5135565
Denoised	7961080	0
Merged	7653885	307195
Non-chimeric	7371118	282767
Vertebrate	6030566	1340552
Fish	6010983	19583

Table 3.2 Agreement in detections between larvae and water samples at 13 sites. Both = number of sites where the taxon was detected using both sample types. Larvae = number of sites where taxon was detected in larvae samples. Water = number of sites where taxon was detected in water samples. Neither = number of sites where a taxon was not detected by either sample type. % Agreement = Sum of 'Both' and 'Neither' / total sites * 100.

Taxa	Both	Neither	Larvae	Water	% Agreement
C. harengus / S. sprattus	7	0	8	12	53.8
Ammodytes marinus	8	0	8	13	61.5
Callionymus sp	1	3	7	4	30.8
Limanda limanda	5	7	6	5	92.3
Triglidae	1	7	5	2	61.5
Scomber scombrus	0	6	0	7	46.2
Ciliata mustela	2	7	3	5	69.2
P. pollachius/virens / M. merlangus	2	4	7	4	46.2
Trisopterus minutus	1	6	6	2	53.8
Microstomus kitt	0	8	5	0	61.5
Pleuronectidae	1	7	5	2	61.5
Trisopterus esmarkii	1	9	4	1	76.9
Merluccius merluccius	0	11	1	1	84.6
Gymnammodytes sp	0	11	2	0	84.6
Soleidae	0	9	0	4	69.2

Bothidae	0	11	0	2	84.6
Ammodytidae	1	7	2	4	61.5
Solea solea	0	12	0	1	92.3
Lepidorhombus sp	0	11	2	0	84.6
Micromesistius poutassou	0	12	0	1	92.3
Labrus bergylta	0	12	1	0	92.3
Molva molva	0	12	1	0	92.3
Sardina pilchardus	0	12	1	0	92.3
Buenia jeffreysii	0	12	1	0	92.3
Gobiidae	0	12	1	0	92.3
Labrus mixtus	0	12	1	0	92.3
Clupeidae	0	12	0	1	92.3
Actinopterigii	0	12	0	1	92.3

Each of the eight species added to both mock communities were detected using both Phusion (Thermo Fisher) and Platinum (Thermo Fisher) Taq polymerases. Only reads assigned to the input DNA taxa were observed in mock community reads, except for 23 and 19 reads in mock community 1 assigned to *Trisopterus esmarkii* when amplifying with Phusion and Platinum, respectively. For mock community 1, which contained equal concentrations of DNA from each of the 8 taxa, the relative quantity of DNA inputted (expected) and relative proportion of reads (observed) differed significantly (Phusion, Chi-square: $X^2 = 14.59$, df = 7, *P* = 0.041, Platinum $X^2 = 18.26$, df = 7, *P* = 0.011), mainly due to an excess and deficit of *A. marinus* reads and *S. sprattus* respectively. However, in Mock 2, where input molar concentrations varied, there was no difference observed between the relative input of DNA and the observed proportion of reads (Phusion, $X^2 = 11.39$, df = 7, *P* = 0.123, Platinum $X^2 = 8.11$, df = 7, *P* = 0.323) (Appendix 2: Figure A2.2).

In both water and larvae samples we successfully detected 12 of the same taxa (Figure 3.2) and a similar number of taxa (19 taxa in water, 20 taxa in larvae) were detected overall. At a site level there was an average of 75% agreement in taxa detection between the sample types (Table 3.2; Figure 3.3). The number of taxa detected in the two sample types depended on the sampling site (G = 31.43, df = 12, *P* = 0.002) and the taxon considered (G = 42.80, df = 27, *P* = 0.027, Figure 3.3, Table 3.2). In general, the more abundant taxa were detected by both sampling methods while less abundant taxa exhibited much greater variance, with some taxa being detected in one sample type. Of these, 10 taxa were detected at only one site, 60% of which were observed in larvae samples (Figure 3.3, Table 3.2).
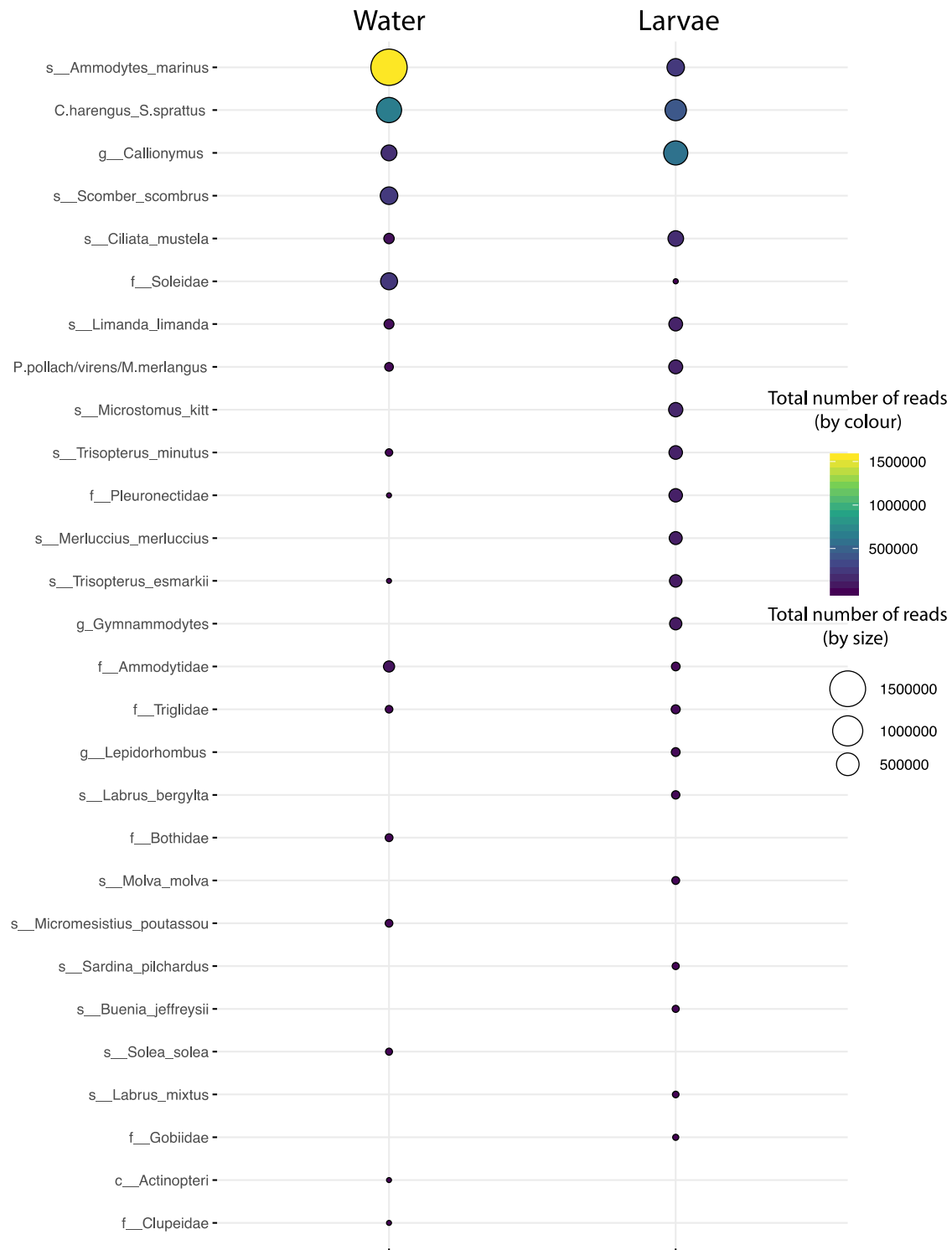
Figure 3.2. Overview of the total number of raw reads per taxon in the two sample types: water samples = Water and larvae samples = Larvae. Taxa are identified to lowest possible taxonomic level. s__ = species level, g__= genus level, f__ = family level.
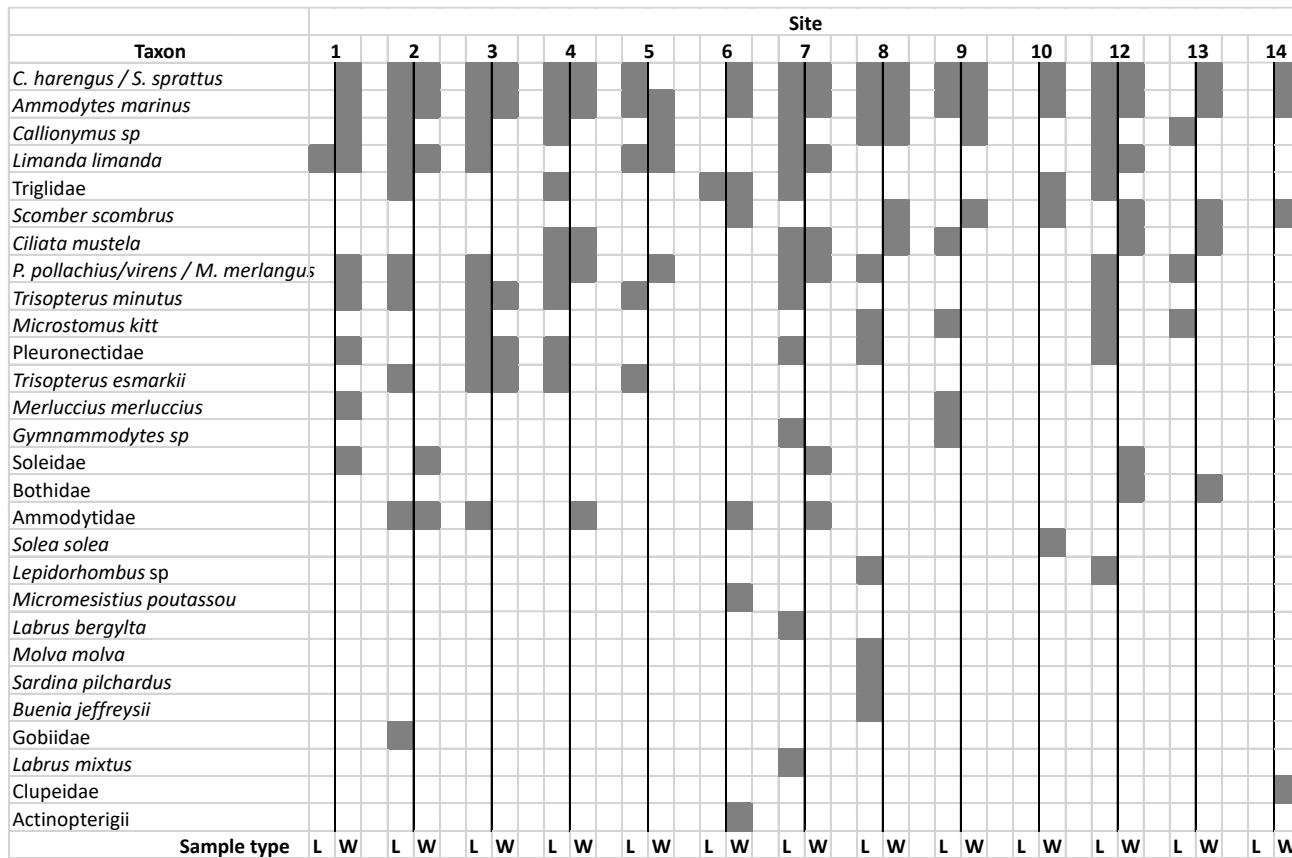
Figure 3.3.Site by site detections (presence (grey)/absence) between the two different sampling methods, larvae sampling 'L' and water sampling, 'W'. Taxa are identified to lowest taxonomic level possible and ordered by overall abundance in the survey.

Patterns of relative abundance broadly followed those observed in the number of detections. *C. harengus/ S. sprattus,* and *A. marinus* were detected in higher relative abundance in water than in larvae samples (Figure 3.2), *C. harengus / S. sprattus:* W = 37.5, p-value = 0.016. *A. marinus:* W = 32, *P* = 0.008). In contrast, the third most abundant taxa, *Callionymus sp* was more frequently detected in the larvae samples, however there was no difference in relative abundance between sample types (W = 113, p-value = 0.110). Comparisons of relative abundance estimates between the sample types were only significantly correlated for *L. limanda* (S = 76.44, P<0.001 *rho* = 0.79) (Appendix 2: Table A2.4).

Community composition, differed between water and larvae samples (PERMANOVA df: 1,24, ps-F$_=$ 4.107, R^2 = 0.146, *P*(perm) = 0.001, permutations: 999). These differences were driven by the pattern of higher abundances of *A. marinus, C. harengus / S. sprattus* and *S. scombrus* in water samples, in contrast to higher abundances of *L. limanda* and *Callionymus* sp. in larvae samples, together contributing to 70.64% of the differences observed between sample types (SIMPER analysis, Appendix 2: Table A2.5). Taxon richness and alpha diversity did not differ between the two sampling methods (richness: V = 42, *P* = 0.83, Shannon-Weiner Diversity: V = 29, *P* = 0.27). Larvae samples did, however, exhibit greater variance than water samples in terms of species richness (richness Larvae MAD=4, Water MAD=1, F- test: df =12, F = 8.85, *P* < 0.001. Shannon-Weiner Diversity Bulk MAD= 0.62, water MAD=0.2, F- test: df =12, F = 3.04, *P* = 0.065) (Figure 3.4).
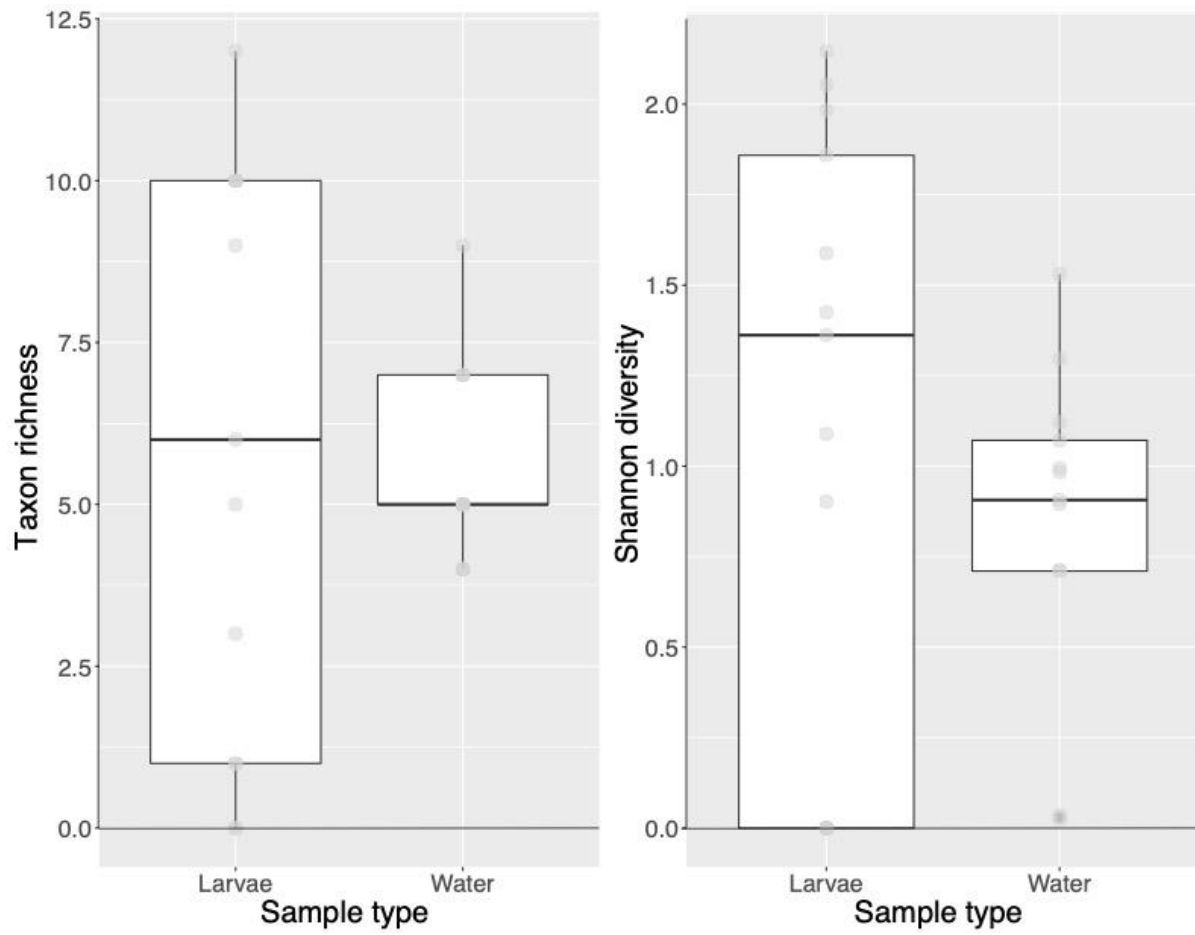
Figure 3.4 Median taxon richness and diversity (lowest possible taxonomic level) for eDNA and bulk samples across 14 sites in the survey. The median is represented by the horizontal line within each box, boxes define the 25th and 75th percentiles and the most extreme data point which is no more than 1.5 times the interquartile range from the corresponding box.

3.5 Discussion

We applied water eDNA metabarcoding to detect fish in marine spawning areas and demonstrated that water samples not only broadly reflect larvae samples, with 75% average agreement on site by site basis but can also be more sensitive in the case of particular taxa, such as sandeels and herring/sprat. While rare species, those detected at one site only, were more likely to be found in bulk samples, some taxa expected to be part of the spawning assemblage (e.g mackerel *S. scombrus* and blue whiting, *M. poutassou)* were only detected in water samples, highlighting the potential of this tool to complement traditional sampling.

Previous studies that compare water sampling to visual or capture surveys, have found varying levels of agreement between the taxa detected (Cilleros et al., 2019; Stat et al., 2019; Thomsen et al., 2016). Thomsen et al., (2016) found correlations between capture biomass and numbers of water sample reads, while Stat et al., (2019) found fish assemblages differed between visual (BRUV) and water samples. Differences in detection between traditional methods and eDNA are influenced by eDNA dispersal range, year class and detection sensitivity of the particular methods compared. Importantly, all monitoring techniques are subject to biases e.g. trawl types may also differ in the species captured due to gear selectivity biases (Hansen et al., 2018). Hence, while water samples may not always perfectly reflect capture/visual samples, they represent a rapid and feasible and non-destructive way to efficiently assess fish community assemblages (Cilleros et al., 2019; Thomsen et al., 2016).

Here, we compared surface water samples to larvae sampled between the surface and 50m depth. Fish eggs and larvae are most abundant in depths shallower than 50m (Conway, Coombs & Smith, 1997, Sabatés, 2004). In the Irish Sea, densities of fish eggs increase with decreasing depth and peak at the surface, while larval density increases towards a peak at of 10-15m, with little difference observed between species, including the families most abundant in this study (Clupeidae, Ammodytidae, Callionymidae) (Coombs et al., 1997). For some taxa, however, this general pattern may not apply, for instance hake larvae (*Merluccius merluccius*) have been shown aggregate at maximum density at depths of 60-80m (Sabatés, 2004). This taxon was not detected in water samples, despite being detected in one larvae haul, therefore, while sampling depth is unlikely to have affected the majority of taxa, it may be a reason for some of the discrepancies in the detection of rare species. Thus, as eDNA exhibits sensitivity to vertical zonation in stratified water (Jeunen et al., 2020), multiple sampling depths may be advisable, depending on the taxa and life stages of interest.

For sensitive taxa such as sandeels that are hard to monitor using traditional means (Ellis et al., 2012), this survey demonstrates the potential of eDNA metabarcoding as a monitoring tool. Sandeels are a key prey species, consumed by fish, seabirds and marine mammals, however, due to their short life cycle, and the reliance of their stocks on larval recruitment, these taxa are difficult to sample and the stocks are therefore difficult to manage (Lynam et al., 2013). Neither otter nor beam trawls are effective

methods for assessing their abundance, particularly at early life stages (Ellis et al., 2012). In addition, morphological identification is often unreliable (Thiel & Knebelsberger, 2016), limiting the assessment to family level only (Ellis et al., 2012). Our survey encountered the lesser sandeel, *A. marinus,* a taxon of the genus *Gymnammodytes* and a further taxon identified to family level in both water and larvae samples, distributed in areas where sandeels are known to spawn in the Irish Sea and the Bristol channel (Ellis et al., 2012; Lynam et al., 2013). *A. marinus* was always detected in water samples where the larvae were also encountered. Therefore, while water sample metabarcoding alone cannot determine the age class of organisms encountered, it can give a picture of the distribution of these taxa during spawning events, to a higher taxonomic resolution than is often available through traditional means (Ellis et al., 2012).

Herring/sprat (*C. harengus/ S. sprattus*) were also frequently encountered using both sample types at the same sites, with water samples displaying higher sensitivity. Most *C. harengus* spawning in the Irish Sea occurs in September/October, and some can spawn as late as March (Brophy & Danilowicz, 2002). However, due to difficulties in morphologically separating these two species, spring surveys tend to assume that larvae caught in this period are *S. sprattus* (Fox et al., 1995), which spawn from March to August (de Silva, 1973). While the primers used in this study cannot separate *S. sprattus* and *C. harengus*, a qPCR approach could be used in the samples where the presence of one or both species is identified by metabarcoding, to rapidly separate these species in water samples (e.g Brechon et al., 2013). *S. sprattus* is relatively under-assessed within the Celtic Seas ecoregion and is considered data limited (Moore et al., 2019), therefore information obtained from a water sampling approach has the potential to add to traditional methods of assessment.

Relative abundances were correlated for *L. limanda* between the two sample types. However, relative abundances estimated from metabarcoding should be treated with caution (Lamb et al., 2019; Thomas, Deagle, Eveson, Harsch, & Trites, 2016) due to amplification bias. In fact, the sequencing of mock community 1 indicated that, with the primers used in this study, the relative abundance of *A. marinus* reads could be, on average 2 times higher than the relative abundance of input material, while *Lepidorhombus sp* reads were 0.6 times as abundant as their known inputs. In addition, DNA shedding may differ between organisms (Sassoubre, Yamahara, Gardner, Block, & Boehm, 2016) and sizes of fish (Maruyama et al., 2014), and can further complicate abundance signals (Hansen et al., 2018). However, using a presence/absence approach can also overestimate the importance of rare taxa and relative abundance estimates may provide useful information (Deagle et al., 2019; Lamb et al., 2019) for comparative studies. In this case, *C. harengus/ S. sprattus* and *A. marinus* were more frequently detected, and had a higher relative abundance, in water samples, while *Callionymus* sp was more frequently detected in bulk samples, but no difference in relative abundance was observed between sample types. This indicates that, in some contexts, measures of relative abundance can provide useful information.

For rarer species, detections by both methods were more sporadic. For example, taxa such as mackerel and blue whiting, which are known to undergo peak spawning in May (Ellis et al., 2012), were not encountered in larvae samples, but were detected in water samples. While it is not possible to know if spawning had occurred and was missed by the larvae sampling, or whether only adults were present, this indicates that taxa potentially missed using traditional means can still be detected in water samples.

eDNA metabarcoding detections are sensitive to computational filtering thresholds (Evans et al., 2017). In this case, blank filtering (Grey et al., 2018) and discarding of taxa with low reads to account for index hopping (Schnell et al., 2015) resulted in less water sample than larvae sample detections for some of the less common/abundant taxa e.g. *M. merluccius*, and *L. limanda* (Figure 3.3). This reflects the trade-off between stringency and uncertainty when applying thresholds of detection to metabacoding data, as while more stringent filtering can underestimate taxa richness, it also reduces the risk of false positives due to tag jumping (Schnell et al., 2015) or cross contamination. Potential solutions include using combinations of markers (Evans et al., 2017) or specific qPCR assays (Harper et al., 2018; Schneider et al., 2016) in conjunction with metabarcoding (Deiner et al., 2017).

While water and larvae samples did not differ overall in richness and alpha diversity measures, larvae samples exhibited greater variance in species richness, community composition and detection of rare taxa differed between sample types, demonstrating how these two sampling strategies may complement each other. When considering how to interpret eDNA data, the ecology of the eDNA molecules should be considered (Barnes & Turner, 2016). eDNA transport in offshore areas has not been studied extensively (Collins et al., 2018), however, in freshwater systems, eDNA signals may travel up to 10 km (Deiner & Altermatt, 2014). In the marine environment, tidal currents, seasonal stratification, pH, and temperature (Lacoursière-Roussel et al., 2018) may all influence the distribution of eDNA in the marine environment (Hansen et al., 2018). In coastal environments, with a small tidal amplitude, eDNA has been shown to have a limited dispersion area, <1000m, and may only remain detectable for as little as an hour after the source has been removed, providing a snapshot of the organisms present (Murakami et al., 2019). In offshore environments however, eDNA may degrade more slowly than in coastal areas (Collins et al., 2018). Environmental factors can, therefore, lead to widely variable dispersal of eDNA particles, dependent on oceanographic, biological and chemical parameters (Hansen et al., 2018).

Conclusions

Water and larval sampling both coincided in the detection of the most common taxa, which constituted 63% and 60% of the taxa in the water and larvae samples respectively. On a site by site basis, there was a 75% agreement in detection between sample types. Sandeels were detected more frequently in water samples, which improves upon traditional methods that often are unable to capture or identify this

family. Herring/sprat were also more frequently detected in water samples, indicating that, metabarcoding combined with a targeted approach such as qPCR, could also provide higher sensitivity distributions for these taxa. While eDNA still requires an extensive sampling effort, its non-invasive and rapid nature renders it particularly suitable for use in spawning and protected areas and for fisheries management applications.

Ethics approval

Sampling has been conducted following Home Office regulations and approved by Swansea University Ethics Committees under approval No. 181019/1996.

Data accessibility

Metabarcoding sequences from larvae samples (BioProject PRJNA576002) and water samples (PRJNA596623) have been deposited in the NCBI.

3.6   References

Anderson, M. J. (2014). Permutational multivariate analysis of variance (PERMANOVA). *Wiley StatsRef: Statistics Reference Online*, 1-15. https://doi.org/10.1002/9781118445112.stat07841

Andruszkiewicz, E. A., Starks, H. A., Chavez, F. P., Sassoubre, L. M., Block, B. A., & Boehm, A. B. (2017). Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. *PLoS ONE*, *12*(4), 1–20. https://doi.org/10.1371/journal.pone.0176343

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bicknell, A. W. J., Godley, B. J., Sheehan, E. V., Votier, S. C., & Witt, M. J. (2016). Camera technology for monitoring marine biodiversity and human impact. *Frontiers in Ecology and the Environment*, *14*(8), 424–432. https://doi.org/10.1002/fee.1322

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., … de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology and Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Brechon, A. L., Coombs, S. H., Sims, D. W., & Griffiths, A. M. (2013). Development of a rapid genetic technique for the identification of clupeid larvae in the Western English Channel and investigation of mislabelling in processed fish products. *ICES Journal of Marine Science*, *70*(2), 399–407. https://doi.org/10.1093/icesjms/fss178

Brophy, D., & Danilowicz, B. S. (2002). Tracing populations of Atlantic herring (Clupea harengus L.) in the Irish and Celtic Seas using otolith microstructure. *ICES Journal of Marine Science*, *59*(6), 1305–1313. https://doi.org/10.1006/jmsc.2002.1304

Burrows, M. T., Schoeman, D. S., Buckley, L. B., Moore, P., Poloczanska, E. S., Brander, K. M., … Richardson, A. J. (2011). The Pace of Shifting Climate in Marine and Terrestrial Ecosystems. *Science*, *334*(6056), 652–655. https://doi.org/10.1126/science.1210288

Bylemans, J., Furlan, E. M., Hardy, C. M., McGuffie, P., Lintermans, M., & Gleeson, D. M. (2017). An environmental DNA-based method for monitoring spawning activity: a case study, using the

endangered Macquarie perch (Macquaria australasica). *Methods in Ecology and Evolution*, *8*(5), 646–655. https://doi.org/10.1111/2041-210X.12709

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Christie, M. R., Tissot, B. N., Albins, M. A., Beets, J. P., Jia, Y., Ortiz, D. M., … Hixon, M. A. (2010). Larval connectivity in an effective network of marine protected areas. *PLoS ONE*, *5*(12), 1–10. https://doi.org/10.1371/journal.pone.0015715

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., … Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, *19*(1), 27–46. https://doi.org/10.1111/1755-0998.12900

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*(1), 117–143. https://doi.org/10.1111/j.1442-9993.1993.tb00438.x

Clarke, K. R., Somerfield, P. J., & Chapman, M. G. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, *330*(1), 55–80. https://doi.org/10.1016/j.jembe.2005.12.017

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, *1*(1). https://doi.org/10.1038/s42003-018-0192-6

Conway, D. V. P., Coombs, S. H., & Smith, C. (1997). Vertical distribution of fish eggs and larvae in the Irish Sea and southern North Sea. ICES Journal of Marine Science, 54(1), 136-147. https://doi.org/10.1006/jmsc.1996.0176

Cordes, E. E., Jones, D. O. B., Schlacher, T. A., Amon, D. J., Bernardino, A. F., Brooke, S., … Witte, U. (2016). Environmental impacts of the deep-water oil and gas industry: A review to guide management strategies. *Frontiers in Environmental Science*, *4*(SEP). https://doi.org/10.3389/fenvs.2016.00058

Costello, M. J., Coll, M., Danovaro, R., Halpin, P., Ojaveer, H., & Miloslavich, P. (2010). A census of marine biodiversity knowledge, resources, and future challenges. *PLoS ONE*, *5*(8). https://doi.org/10.1371/journal.pone.0012110

de Silva, S. S. (1973). Aspects of the reproductive biology of the sprat, Sprattus sprattus (L.) in inshore waters of the west coast of Scotland. *Journal of Fish Biology*, *5*(6), 689–705. https://doi.org/10.1111/j.1095-8649.1973.tb04505.x

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., … Eveson, J. P. (2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. https://doi.org/10.1111/mec.14734

Deiner, K., & Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a natural river. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0088786

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., … Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Douvere, F., & Ehler, C. N. (2011). The importance of monitoring and evaluation in adaptive maritime spatial planning. *Journal of Coastal Conservation*, *15*(2), 305–311. https://doi.org/10.1007/s11852-010-0100-9

Ellis, J. R., Milligan, S. P., Readdy, L., Taylor, N., & Brown, M. J. (2012). Spawning and nursery grounds of selected fish species in UK waters. Retrieved from https://www.cefas.co.uk/publications/techrep/TechRep147.pdf

Erickson, R. A., Rees, C. B., Coulter, A. A., Merkes, C. M., McCalla, S. G., Touzinsky, K. F., … Amberg, J. J. (2016). Detecting the movement and spawning activity of bigheaded carps with environmental DNA. *Molecular Ecology Resources*, *16*(4), 957–965. https://doi.org/10.1111/1755-0998.12533

Evans, N. T., Li, Y., Renshaw, M. A., Olds, B. P., Deiner, K., Turner, C. R., … Pfrender, M. E. (2017). Fish community assessment with eDNA metabarcoding: Effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(9), 1362–1374. https://doi.org/10.1139/cjfas-2016-0306

Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J. A., ... & Becker, J. (2013). Topographic diversity of fungal and bacterial communities in human skin. *Nature, 498*(7454), 367.

Fox, C. J., Dickey-Collas, M., & Winpenny, A. J. (1997). Spring plankton surveys of the Irish Sea in 1995: the distribution of fish eggs and larvae. *Fisheries Research Technical Report - UK Ministry of Agriculture, Fisheries and Food, Directorate of Fisheries Research*, *104*(104). doi:10.1016/S0967-0653(98)80494-8

Geijzendorffer, I. R., Regan, E. C., Pereira, H. M., Brotons, L., Brummitt, N., Gavish, Y., … Walters, M. (2016). Bridging the gap between biodiversity data and policy reporting needs: An Essential Biodiversity Variables perspective. *Journal of Applied Ecology*, *53*(5), 1341–1350. https://doi.org/10.1111/1365-2664.12417

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., … Taberlet, P. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

Greve, W., Prinage, S., Zidowitz, H., Nast, J., & Reiners, F. (2005). On the phenology of North Sea ichthyoplankton. *ICES Journal of Marine Science*, *62*(7), 1216–1223. https://doi.org/10.1016/j.icesjms.2005.03.011

Grey, E. K., Bernatchez, L., Cassey, P., Deiner, K., Deveney, M., Howland, K. L., … Lodge, D. M. (2018). Effects of sampling effort on biodiversity patterns estimated from environmental DNA metabarcoding surveys. *Scientific Reports*, *8*(1), 2–11. https://doi.org/10.1038/s41598-018-27048-2

Habtes, S., Muller-Karger, F. E., Roffer, M. A., Lamkin, J. T., & Muhling, B. A. (2014). A comparison of sampling methods for larvae of medium and large epipelagic fish species during spring SEAMAP ichthyoplankton surveys in the Gulf of Mexico. *Limnology and Oceanography: Methods*, *12*(2), 86–101. https://doi.org/10.4319/lom.2014.12.86

Hansen, B. K., Bekkevold, D., Clausen, L. W., & Nielsen, E. E. (2018). The sceptical optimist: challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish and Fisheries*, *19*(5), 751–768. https://doi.org/10.1111/faf.12286

Harper, L. R., Lawson Handley, L., Hahn, C., Boonham, N., Rees, H. C., Gough, K. C., … Hänfling, B. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (Triturus cristatus). *Ecology and Evolution*, *8*(12), 6330–6341. https://doi.org/10.1002/ece3.4013

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data.

Genome research, 17(3), 377-386. https://doi.org/10.1101/gr.5969107

Jeunen, G. J., Lamare, M. D., Knapp, M., Spencer, H. G., Taylor, H. R., Stat, M., ... & Gemmell, N. J. (2020). Water stratification in the marine biome restricts vertical environmental DNA (eDNA) signal dispersal. Environmental DNA, 2(1), 99-111.

Kimmerling, N., Zuqert, O., Amitai, G., Gurevich, T., Armoza-Zvuloni, R., Kolesnikov, I., … Sorek, R. (2018). Quantitative species-level ecology of reef fish larvae via metabarcoding. *Nature Ecology & Evolution*, *2*(2), 306–316. https://doi.org/10.1038/s41559-017-0413-2

Koslow, J. A., & Wright, M. (2016). Ichthyoplankton sampling design to monitor marine fish populations and communities. *Marine Policy*, *68*, 55–64. https://doi.org/10.1016/j.marpol.2016.02.011

La Védrine, Maximillian. (2014). Oil and Gas Fisheries Risk Assessment Advice, Updated Cefas Recommendations for Spawning Finfish – English & Welsh Blocks. https://doi.org/10.13140/RG.2.1.3543.3126.

Lacoursière-Roussel, Anaïs, Kimberly Howland, Eric Normandeau, Erin K. Grey, Philippe Archambault, Kristy Deiner, David M. Lodge, Cecilia Hernandez, Noémie Leduc, and Louis Bernatchez. "eDNA metabarcoding as a new surveillance approach for coastal Arctic biodiversity." *Ecology and evolution 8, no. 16* (2018): 7763-7777. https://doi.org/10.1002/ece3.4213

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430. https://doi.org/10.1111/mec.14920

Leduc, N., Lacoursière-Roussel, A., Howland, K. L., Archambault, P., Sevellec, M., Normandeau, E., ... & Bernatchez, L. (2019). Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. Environmental DNA, 1(4), 342-358. https://doi.org/10.1002/edn3.35

Lee, O., Nash, R. D. M., & Danilowicz, B. S. (2005). Small-scale spatio-temporal variability in ichthyoplankton and zooplankton distribution in relation to a tidal-mixing front in the Irish Sea. *ICES Journal of Marine Science*, *62*(6), 1021–1036. https://doi.org/10.1016/j.icesjms.2005.04.016

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., … Pfrender, M. E. (2012). Conservation in a cup of water: Estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, *21*(11), 2555–2558. https://doi.org/10.1111/j.1365-

294X.2012.05600.x

Lynam, C. P., Halliday, N. C., Höffle, H., Wright, P. J., Van Damme, C. J. G., Edwards, M., & Pitois, S. G. (2013). Spatial patterns and trends in abundance of larval sandeels in the North Sea: 1950-2005. *ICES Journal of Marine Science*, *70*(3), 540–553. https://doi.org/10.1093/icesjms/fst006

Maggia, M. E., Vigouroux, Y., Renno, J. F., Duponchelle, F., Desmarais, E., Nunez, J., … Mariac, C. (2017). DNA metabarcoding of amazonian ichthyoplankton swarms. *PLoS ONE*, *12*(1), 1–14. https://doi.org/10.1371/journal.pone.0170009

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The release rate of environmental DNA from juvenile and adult fish. *PLoS ONE*, *9*(12). https://doi.org/10.1371/journal.pone.0114639

Moore, C., Lynch, D., Clarke, M., Officer, R., Mills, J., Louis-Defour, J., & Brophy, D. (2019). Age verification of north Atlantic sprat. *Fisheries Research*, *213*(August 2018), 144–150.

Moore, C., Lynch, D., Clarke, M., Officer, R., Mills, J., Louis-Defour, J., & Brophy, D. (2019). Age verification of north Atlantic sprat. *Fisheries Research, 213,* 144-150. https://doi.org/10.1016/j.fishres.2019.01.018

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757–1764. https://doi.org/10.1093/bioinformatics/btn322

Murakami, H., Yoon, S., Kasai, A., Minamoto, T., Yamamoto, S., Sakata, M. K., ... & Masuda, R. (2019). Dispersion and degradation of environmental DNA from caged fish in a marine environment. Fisheries science, 85(2), 327-337. https://doi.org/10.1007/s12562-018-1282-6

Pastoors, M., Rijnsdorp, A. D., & Van Beek, F. A. (2000). Effects of a partially closed area in the North Sea ("plaice box") on stock development of plaice. *ICES Journal of Marine Science*, *57*(4), 1014–1022. https://doi.org/10.1006/jmsc.2000.0586

Pitois, S. G., Lynam, C. P., Jansen, T., Halliday, N., & Edwards, M. (2012). Bottom-up effects of climate on fish populations: Data from the continuous plankton recorder. *Marine Ecology Progress Series*, *456*, 169–186. https://doi.org/10.3354/meps09710

Poloczanska, E. S., Brown, C. J., Sydeman, W. J., Kiessling, W., Schoeman, D. S., Moore, P. J., … Richardson, A. J. (2013). Global imprint of climate change on marine life. *Nature Climate Change*, *3*(10), 919–925. https://doi.org/10.1038/nclimate1958

Ratcliffe, F. C., Uren Webster, T. M., Rodriguez-Barreto, D., O'Rorke, R., de Leaniz, C. G., &

Consuegra, S. (2020). Rapid quantitative assessment of fish larvae community composition using metabarcoding. *BioRxiv,* 2020.01.01.884544. doi:10.1101/2020.01.01.884544

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Régnier, T., Gibb, F. M., & Wright, P. J. (2017). Importance of trophic mismatch in a winterhatching species: Evidence from lesser sandeel. *Marine Ecology Progress Series*, *567*, 185–197. https://doi.org/10.3354/meps12061

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic acids research*, *39*(21), e145-e145. https://doi.org/10.1093/nar/gkr732.

Sabatés, A. (2004). Diel vertical distribution of fish larvae during the winter-mixing period in the Northwestern Mediterranean. ICES Journal of Marine Science, 61(8), 1243-1252. https://doi.org/10.1016/j.icesjms.2004.07.022

Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A., & Boehm, A. B. (2016). Quantification of Environmental DNA (eDNA) Shedding and Decay Rates for Three Marine Fish. *Environmental Science and Technology*, *50*(19), 10456–10464. https://doi.org/10.1021/acs.est.6b03114

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Sahl, J. W. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.,* 75(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09.

Schneider, J., Valentini, A., Dejean, T., Montarsi, F., Taberlet, P., Glaizot, O., & Fumagalli, L. (2016). Detection of invasive mosquito vectors using environmental DNA (eDNA) from water samples. *PLoS ONE*, *11*(9), 1–18. https://doi.org/10.1371/journal.pone.0162493

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated - reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conservation Biology*, *33*(1), 196–205. https://doi.org/10.1111/cobi.13183

Thiel, R., & Knebelsberger, T. (2016). How reliably can northeast Atlantic sand lances of the genera

Ammodytes and Hyperoplus be distinguished? A comparative application of morphological and molecular methods. *ZooKeys*, *2016*(617), 139–164 https://doi.org/10.3897/zookeys.617.8866

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived from control material. *Molecular ecology resources, 16*(3), 714-726. https://doi.org/10.1111/1755-0998.12490

Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. *PLoS ONE*, *7*(8), 1–9. https://doi.org/10.1371/journal.pone.0041732

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS ONE*, *11*(11). https://doi.org/10.1371/journal.pone.016525

Chapter 4　MHC class 1-alpha can reveal cryptic fine-scale population structure in a commercial fish, the European sea bass (*Dicentrarchus labrax)*

4.1   *Abstract*

Identifying population structuring in highly fecund marine species with high dispersal rates is challenging, but critical for conservation and stock delimitation for fisheries management. European sea bass (*Dicentrarchus labrax*) stocks are declining in the North Atlantic despite management measures to protect them and identifying their fine population structure is needed for managing their exploitation. As for other marine fishes, neutral genetic markers indicate that Eastern Atlantic sea bass form a panmictic population and is currently managed as arbitrarily divided stocks. The genes of the major histocompatibility complex (MHC) are key components of the adaptive immune system and ideal candidates to assess fine structuring arising from local selective pressures. We used Illumina sequencing to characterise allelic composition and signatures of selection at the MHC class I-alpha region of 6 *D. labrax* populations across the Atlantic range. We found high allelic diversity driven by positive selection, corresponding to moderate supertype diversity, with 131 alleles clustering into 4 to 8 supertypes, and a mean number of 13 alleles per individual. Individual loci could not be identified, but private alleles allowed us to detect regional genetic structuring not found previously using neutral markers. Our results suggest that MHC markers can be used to detect cryptic population structuring in marine species where neutral markers fail to identify differentiation. This is particularly critical for fisheries management, and of potential use for selective breeding or identifying escapes from sea farms.

4.2   Introduction

With the decline of many traditional fisheries, accurate fish stock management has never been so critical (Beddington et al. 2007). Yet, identifying population structuring and genetic differentiation in marine populations using molecular markers can be challenging, particularly for highly fecund species with high dispersal rates and complex life cycles (Hedgecock et al. 2007). The use of large population genomic data sets and markers influenced by selection are increasingly revealing fine population structure and patterns of reproductive isolation even in highly dispersive marine species, with important implications for their conservation and management (Gagnaire et al. 2015). Yet, in many cases, lack of information about population structure results in a pragmatic approach to management (ICES 2021) and a mismatch between biological and management units (Reiss et al. 2009) which can exacerbate the decline of stocks. Commercial fisheries are already contributing to the genetic homogenization of marine fish species (Gandra et al. 2021) and ignoring fine scale local adaptation can further result in the loss of functionally important biodiversity (Limborg et al. 2012).

European seabass (*Dicentrarchus labrax*) is a traditionally important species in terms of commercial and recreational fisheries, and has more recently become a key species for aquaculture (Vandeputte et al. 2019). The extent of population structuring in European seabass is unclear, and in the Atlantic area has been arbitrarily divided by the ICES into four stocks: (1) the English Channel, Celtic and Irish Seas, and North Sea, (2) Bay of Biscay, (3) Iberian waters and (4) West Ireland/West Scotland (Drogou et al. 2014; De Pontual et al. 2019). The seabass stock found in the English Channel, Celtic and Irish Seas and North Sea is characterised by slow growth and late maturation and its productivity is greatly influenced by sea water temperature (Walker et al. 2020). This stock has declined markedly since 2010, despite stringent quotas and fishing bans (Walker *et al.* 2020), mainly due to overfishing and low recruitment since 2008, but also due to the influence of long-term climate change and local environmental factors that affect mostly the early developmental stages (Bento et al. 2016). Tagging studies indicate that there is strong site fidelity (Pawson et al. 2008) and frequent migrations between the southern area and the Bay of Biscay, and between the North Sea and the English Channel, highlighting the need for more accurate information on fine population structuring, necessary for correct stock management (Pawson et al. 2007). For example, the Irish stock is exploited separately from the Atlantic stocks although its distribution spans more than one ICES division (ICES 2021). Existing studies based on mitochondrial DNA suggest that there is population structuring, which is not consistently detected by neutral markers. Mitochondrial DNA indicates that Atlantic and Mediterranean seabass populations are differentiated as a consequence of their isolation during the Pleistocene, that resulted in divergence in haplotype frequencies (Lemaire et al. 2005). However, a post-glacial secondary contact is thought to have eroded this divergence at the level of nuclear markers (Duranton

et al. 2018), while differential introgression resulted in the creation of genomic islands of differentiation (Tine et al. 2014). Within the Atlantic seabass stock, there are three mitochondrial DNA lineages distributed predominantly in the Bay of Biscay (Atlantic 1; from where the Mediterranean lineage could have originated), European coast (Atlantic 2) and the British Islands and Norway (Atlantic 3) (Coscia & Mariani 2011). In contrast, there is limited regional structuring based on nuclear markers (microsatellites and SNPs), with the exception of the populations in the south-eastern range (Portugal and Morocco) that display some Mediterranean influence in their SNP allelic composition (Souche et al. 2015). Markers under selection, like allozymes or the somatolactin gene allele distribution have identified some differentiation between the Bay of Biscay and the southern North Sea (Quéré et al. 2010) and at the local scale (Castilho & McAndrew 1998). This suggests that markers under selection may prove better candidates to identify finer scale structuring for this species than neutral markers.

The genes of the Major Histocompatibility Complex (MHC) are some the most studied and highly polymorphic genes in vertebrates (Edwards & Hedrick 1998). MHC genes encode for proteins that present antigens to T-cells, triggering the adaptive immune response (Janeway et al. 2004) and most of their diversity is concentrated in the region that binds antigens from pathogens, the peptide binding region (PBR) (Hedrick & Kim 2000). Polymorphism within the PBR determines which pathogens an individual can respond to (Radwan et al. 2020). PBR diversity is maintained by pathogen selection (Slade & McCallum 1992; Eizaguirre et al. 2012) and in some species by mate choice (Milinski 2006; Consuegra & Garcia de Leaniz 2008), and can be influenced (directly or indirectly) by local environmental factors like temperature (Dionne et al. 2007). We developed an Illumina sequencing-based protocol to genotype the PBR of the European seabass MHC class I-alpha gene, and investigated its potential for detecting fine scale population structuring and signatures of local selection pressures across the species' Atlantic range.

4.3   *Methods*

Genomic DNA was extracted using the Qiagen DNeasy Blood & Tissue Kit from 62 wild seabass individuals (caught between 2013 and 2015) from six areas of the North Eastern Atlantic (Biscay: ICES rectangle 20 E8, n=6, Dover: ICES 30 F1, n=7, Irish Sea ICES 34E4, n=5, Celtic Shelf: ICES 28 E1, n=22, North Sea: ICES 38 E9, n=3, and Portugal: ICES 4 E1, n=19). Fish were genotyped using a custom designed primer pair: sbMHC_F2 5' CTGGAGTCCCAAACTTCCC 3', sbMHC_R2 5' AGGTGGACACCTCCAGTTTG 3', amplifying a 241bp fragment of the protein coding sequence region of the MHC class I-alpha gene. Primers were designed based on existing *Dicentrarchus labrax* MHC-I sequences (NCBI accession numbers: HQ290103.1- HQ290126.1 and JX171686.1-JX171696.1, (Pinto et al. 2013)). Primer sbMHC_F2 starts at nucleotide position 180 in the alignment from Pinto *et al.* (2013), and sbMHC_R2 ends at position 421. This region corresponds to the α1 helical region and the beginning of the α2 helical region that together form the PBR.

Libraries were prepared in a two-step PCR protocol, using Platinum™ Hot Start PCR Master Mix (2×) (Thermo Fisher Scientific). Initially, 2.5 µl of DNA per individual was amplified in a 25µl reaction (12.5µl taq, 0.5 µl each of forward and reverse primer, 9µl water), using the above primers with Illumina overhang adapter sequences added (Illumina, Inc., 16S Metagenomic protocol). PCR amplification consisted of the following: 2 min at 94°C, 28 cycles of 30s at 94°C, 30s at 55°C (locus specific annealing temperature), 1 min 72°C (extension). Subsequently, 2µl of product from the initial reaction was used as template in a second PCR (12.5µl taq, 1.25 µl each of a sample-specific combination of Nextera XT adapters (Index Kit v2, Illumina, Inc) and 10µl water) and 8 cycles of this second PCR was run with identical conditions to the first. 5µl of each indexed PCR product was pooled and the final pool was cleaned using Agencourt AMPure XP beads (Beckman Coulter), using a 1:1 ratio of beads to pooled template. The pool was then quantified using qPCR (NEBNext® Library Quant Kit for Illumina®, protocol according to manufacturer's instructions). Blanks (where molecular grade water was added in the place of template) were used to check for any cross-contamination between samples. Pair-end sequencing was carried out on the MiSeq platform (Illumina, Inc.).

Bioinformatics and data processing

De-multiplexed raw pair end sequences were processed using the ampliSAT suite of tools (Sebastian et al. 2016) (available at: http:// evobiolab.biol.amu.edu.pl/amplisat/). This online analysis suite, based on stepwise clustering thresholds (Stutz & Bolnick, 2014), has been shown to improve on previous MHC genotyping methods and is suitable for de novo genotyping of multi-gene families (Sebastian *et al.* 2016; Biedrzycka et al. 2017). Any samples with less than 1000 reads coverage were discarded from downstream analysis. To optimise parameters and ensure maximum reliability of genotyping for the

entire dataset, initially only the 6 replicate samples (6 individuals, 12 separate PCR reactions), were processed to establish the parameters that would give the most repeatable (appearing in both replicates) genotype between replicate samples. Pair-end reads were merged using the AmpliMERGE tool, using default settings. AmpliCLEAN, was used to remove low quality and anomalous long/short reads, using a minimum Phred score of 30 and a maximum number of reads per amplicon of 5000. AmpliCHECK, was used with default parameters for an initial exploration of the dataset to check read length, coverage and frequency of variants and screen for potential PCR/sequencing artefacts. Parameters for AmpliSAS genotyping were chosen based on the AmpliCHECK results, and subsequent stepwise parameter optimisation trials. Filtering per amplicon frequency (PAF) was set at 1% (the minimum PAF for variants that appeared in both replicates). The minimum dominant frequency was set at 10%, in order to keep true similar variants, whilst removing high frequency motive specific errors, meaning that only sequences with a frequency below that threshold were clustered with a parental sequence (Sebastian *et al.* 2016; Biedrzycka *et al.* 2017). Illumina sequencing was also specified, as was the discarding of non-coding sequences (i.e., those which contained a stop codon in the major reading frame). A maximum of 16 alleles per individual was considered, based on the *D. labrax* MHC class I alleles identified by (Pinto *et al.* 2013) from cDNA cloning, which indicated a minimum of 6 and maximum of 8 loci per individual. The following manual filtering steps were then applied to the AmpliSAS output: all singletons (variants that appeared in one fish only) and any sequences with less than 10 reads overall were removed from downstream analysis (Migalska et al. 2019). In addition, any variants less than 200bp in length were also removed. Repeatability between replicates was then calculated following (Biedrzycka *et al.* 2017), where the number of identical variants are divided by the total number of alleles called in both replicates in a given individual to ascertain a repeatability proportion. Once these optimum processing parameters were established, the entire dataset was processed using AmpliMERGE, AmpliCLEAN and AmpliSAS with the above settings and then subjected to the same manual filtering steps.

Polymorphism analysis

Variants (nucleotide sequences) identified with the AmpliSAT pipeline were aligned using ClustalW in MEGAX (Kumar et al. 2018). Initial screening of aligned variants was carried out in DnaSP (Rozas et al. 2017), using the 'polymorphism data' tool. The Codon-based Z-Test of Selection (MEGAX), using the Nei-Gojobori (p-distance) method, was used to ascertain if there was evidence of positive selection in the variants overall. The seabass amino acid sequences were then aligned with other teleost MHC class I sequences of the U linage described in Grimholt et al. (2015), and a neighbour-joining tree based on p-distance and pairwise deletion was constructed using 1000 bootstrapping iterations using MEGA X to assess their relationship.

Recombination and Selection analysis

Recombination was assessed using SBP (Single Break Point) analysis using the '010021' model of nucleotide substitution, identified using the automatic model selection tool (datamonkey.org). Positive and negative selection were assessed using the HyPhy package (datamonkey.org) using 3 selection models. To detect pervasive positive/diversifying selection, FEL (Fixed Effects Likelihood, (Kosakovsky Pond & Frost 2005)) and FUBAR (Fast, Unconstrained Bayesian AppRoximation, (Murrell et al. 2013)) were used. FEL infers non-synoymous (dN) and synonymous (dS) substitution rates on a per-site basis, using maximum-likelihood approach and fixed branch length estimates (Kosakovsky Pond & Frost 2005). FUBAR uses a Bayesian approach and similarly to FEL to infer non-synoymous (dN) and synonymous (dS) substitution rates on a per-site basis. Both methods assume that the selection pressure for each site is constant along the entire phylogeny and detect both positive pervasive/diversifying selection as well as negative/purifying selection (pervasive: FEL, episodic: FUBAR). To detect episodic positive/diversifying selection, MEME (Mixed Effects Model of Evolution, (Pond et al. 2006)) was used. MEME uses a mixed-effects maximum likelihood approach to detect sites subject to positive selection under a proportion of branches (Kosakovsky Pond et al. 2006). The results of the three models were combined to identify positively selected sites for downstream analysis. A site was only considered to be under positive selection if it was selected by all three models. Recombination sites were compared with positively selected sites to avoid confounding effects, and, as they did not overlap, positively selected sites were analysed without removal of recombination sites.

In addition, alleles were analysed for conservative or radical amino-acid changes using TreeSAAP (Woolley et al. 2003), using a neighbour joining tree with 1000 bootstrap replicates (amino acid sequences). TreeSAAP compares the expected distribution of 31 amino acid properties under neutral conditions to the observed pattern of replacement (McClelland et al. 2011). A sliding window value of 1 was used to obtain codon level magnitude of change scores. Significant differences in the top 2 magnitude categories (7 and 8) were considered as 'radical' changes.

Supertype identification

The amino acids of positively selected sites (PSS) as identified by FEL, FUBAR and MEME were assigned 5 z-descriptors (Doytchinova & Flower 2005) : z1 (hydrophobicity), z2 (steric bulk), z3 (polarity), z4 and z5 (electronic effects) (Sandberg et al., 1998) and translated into a mathematical matrix. Gaps in the alignment of PSS were assigned zeros in the matrix. This matrix was then used to

identify 'supertypes' which are groups of molecules (alleles) that share overlapping peptide binding specificity and are therefore functionally similar (Doytchinova & Flower 2005).

Supertype clustering was performed in discriminant analysis of principle components (DAPC) with the 'adegenet' package (Jombart 2008) in R (Version 4.0.4, (R Core Team 2019)). Initially, the 'find.clusters' function was used to identify the number of clusters. The optimal number of clusters was chosen using BIC values. Cluster values between the minimum number of clusters after which BIC decreased by a negligible amount, and the lowest BIC value were then evaluated using stepwise DAPC runs, plotted and inspected visually. The number of PCs was chosen using the function 'optim.a.score' and kept consistent for each DAPC analysis.

Loci identification via allele clustering

Phylogenetic trees were generated using the following methods: (1) neighbour joining (codons), using P-distance with 1000 bootstrap support (MEGAX, Kumar et al., 2018), (2) Ward's (Euclidian distance) using the z-matrix (Doytchinova & Flower 2005), (3) UPGMA using the z-matrix, with 3 distances: Euclidian, Cosine and Pearson Correlation algorithms (PAST 4, (Hammer et al. 2001)). In addition, a Neighbor-Net (1000 bootstrap support) was generated in SplitsTree4 (Huson & Bryant 2006), because networks may be more appropriate than trees for identifying relationships between alleles where duplication, recombination and gene conversion are common (Biedrzycka *et al.* 2017).

The performance of all methods in identifying loci was evaluated by testing whether any given individual exhibited more than 2 variants/allele per cluster. This was based on the assumption that, as a diploid organism (Felip et al. 2001), no individual should exhibit more than 2 alleles at a given locus.

Population analysis

Fish from the two regions with the largest sample size, Portugal (n= 19, ICES rectangle 4 E1) and Celtic Shelf (n=22, ICES rectangle 28 E1) were examined to identify private alleles to each region and potential population structuring. To further explore population structure between these two populations, we used DAPC incorporating 'optim.a.score' (Jombart 2008; Jombart & Collins 2015), using a presence/absence matrix where each column represented an allele, and each row an individual fish. Presence/absence of alleles was then recoded as frequencies, where alleles were represented as a proportion of the total alleles for each individual. This matrix was then used to generate a 'genind' object for further analysis. Pairwise Jost's D (pairwise D_{est}, (Jost 2008)) was calculated using the 'mmod' package in R (Winter 2012), because this metric is appropriate for genes that exist in multiple

copies with unknown locus affiliation of alleles (Jost 2008; Lighten et al. 2017). We used the function 'confusionMatrix' in the 'caret' package (Kuhn 2009) in R, to assess the accuracy of the assignment of individuals to their stock based on DAPC.

4.4   *Results*

MHC class I alpha polymorphism in seabass

A total of 62 individual fish were sequenced with sufficient depth for downstream analysis (>1000 reads). Samples were capped at a maximum of 5000 reads for AmpliSAT processing reasons and, after processing (prior to removal of singletons), had an average coverage of 3140 reads (STDEV= 1346 reads). Singleton removal reduced the number of variants/alleles from 332 to 133, and removal of two variants <200bp in length resulted in 131 variants retained in the dataset for selection analysis. None of the variants were an exact match to already published alleles (Pinto et al., 2013), however SBmhc1_111 has only one nucleotide difference to HQ290120.1/1-825 clone_15 and HQ290116.1/1-825 clone_1 (which are identical in the target region). Repeatability between the 6 PCR replicated samples was estimated as 0.89 (89% of the alleles were present in both replicates, based on those samples as processed with the whole dataset). This level of repeatability is to be expected, given the number of reads per sample, and repeatability of 0.9 is considered sufficient for downstream analysis (Biedrzycka *et al.* 2017). Individual copy number varied between 7 and 16 alleles (mean = 12.6, mode = 13).

The 131 different nucleotide sequences identified corresponded to 123 unique amino acid sequences (Figure 4.1). An indel of 2 amino acids resulted in sequences of two lengths, 240 bp and 234 bp. From the 240 sites, 137 were variable and there were a total 228 mutations (Eta). Nucleotide diversity per site (Pi) was: 0.13216. The codon-based Z-test of selection (based on all alleles in the dataset) indicated the presence of positive selection (P= 0.03, Z = 1.85).

MHC class 1-alpha can reveal cryptic fine-scale population structure in a commercial fish, the European sea bass (Dicentrarchus labrax)

```
                        10        20        30        40        50  * *  *  60     *  70        80
                 ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
SBmhc1_1/1-77    GVPNFPEFVTVGLVDEFQTVHYDSNTRRAEPKQDWMSRVTAEDPQYWERETQACLGKQ-QAFKANIETAKQRFNQTGG--
SBmhc1_2/1-77    ..........V.ID......NK........................Q...S....-R.........G..........--
SBmhc1_3/1-77    ..........V.IDY...............K................T...F-.I....D..............--
SBmhc1_4/1-77    ........S...DV.MF..............D..........EVSVRA.-RT..T....G..............--
SBmhc1_5/1-77    ..........V.ID......................Q...S....-R.........G..........--
SBmhc1_6/1-77    ..........V.ID.C......................L.M....SVRA.-RT.....................--
SBmhc1_7/1-77    ..........V.ID...............K................T...F-.I........G..........--
SBmhc1_8/1-77    ..........V.IDY..............K................T...F-.I........G..........--
SBmhc1_9/1-77    ........S...V.IDY..........................GQ..T...S-.T........I..........--
SBmhc1_10/1-77   .....................S...................EVSVRA.-RT..T...................--
SBmhc1_11/1-77   ..........G.IDY.............INA..............SIV.A.-.V......I.....S....--
SBmhc1_12/1-77   ......A..F..G.I.............T--KVV.......N.GNF.N.-.T..V...I..........VH
SBmhc1_13/1-77   ...........................I.............EVSVRA.-RT..T...................--
SBmhc1_14/1-77   ..........V.IDY...............................T...F-.I........G..........--
SBmhc1_15/1-77   ........S...V.MF............I.A..D........DV.T.H.-.RH.DS..YI.............--
SBmhc1_16/1-77   ..........DG.IDY............INA..............SIV.A.-.V......I.....S....--
SBmhc1_17/1-77   ........S...DV.MF.............D.........M..EVSVRA.-RT..T....G..............--
SBmhc1_18/1-77   ..........V.ID....................Q...S....-.T........G..........--
SBmhc1_19/1-77   ......ALS...DV.MNY.........N...E.......Q..IR..A.-.T..D...I.....K.....--
SBmhc1_20/1-77   ..........V.ID.C..........................M....SVRA.-RT................--
SBmhc1_21/1-77   ..................S.................EVSVRA.-RT.................--
SBmhc1_22/1-77   .....QY.SF...DV.IS.C......N.......N.A........G.........S.-.S...........--
SBmhc1_23/1-77   ........S...DG.IDY.............INA..............SIV.A.-.V......I.....S....--
SBmhc1_24/1-77   ..........................................T...-................--
SBmhc1_25/1-77   ........I.....A.MI......G..........D...L.TQ...A..A.-.T......IL.P......--
SBmhc1_26/1-77   ........S.....V.IY...................I.........L....E..KRH.-.RH.VS..AL......--
SBmhc1_27/1-77   ........S...F..G.INY.........K.....VP--K.V..F......IL..T-PT........V.......VH
SBmhc1_28/1-77   ..........V.ID.........................L.M....SVRA.-RT...............--
SBmhc1_29/1-77   ......Y.S....DV.IS.C......N...........EGKAD...YS..II..A.-.T..G...IL.......--
SBmhc1_30/1-77   ..........DG.IDY.......FV.....T..............SIV.S.-.V......I.........--
SBmhc1_31/1-77   ........A.....GV.IS.........EV...G..N--K.V...F.K.N..FS..S.-.N.......IL.......VH
SBmhc1_32/1-77   ........A.....V.IDY...........K................T...F-.I........G..........--
SBmhc1_34/1-77   ........I.....A.IEY........FV.....VN--K.V..........KS..A.-.I......IL.P......VH
SBmhc1_35/1-77   ........I.....A.IES.....K...........EGKAD...YS...A..A.-.N......I...........--
SBmhc1_36/1-77   ........I.....A.IES........FV.....VN--K.V..........KS..A.-.N......I.........VH
SBmhc1_37/1-77   ........S.....V.IY...................I.........L....E..KRH.-RRH.DS..AL........--
SBmhc1_38/1-77   ......Y.S.....AEIY...........................T...S-.I..G....I.........--
SBmhc1_39/1-77   ........A.....VPMNY.......T.........AGDRAD..DWQ..IIV.T.-.T...G...IL........--
SBmhc1_40/1-77   ........S.....V.MF............I.A..D........DV.T.H.-.RRH.DS..YI.............--
SBmhc1_41/1-77   ........A.....IPMNY.......FIS.............L..Y..VF..T.-.S...............--
SBmhc1_42/1-77   ........S.....V.IY...........I.A........L....DV..KRH.-RRH.VS..AL..........--
SBmhc1_43/1-77   ...............................................-........G.......--
SBmhc1_44/1-77   ........I.....DG.IDY.............DRAD..DW...KR..A.-.V.........--
SBmhc1_45/1-77   ......F..V.INY.........F.....VI.T.E......D.Y..VF..N.-.T.......I.........--
SBmhc1_46/1-77   ......Y.S....DI.IS.C....M.N.A.................T...A.-.S.............--
SBmhc1_48/1-77   ..........V.IDY......N....................Q...S....-R.........G..........--
SBmhc1_49/1-77   ........S.....V.INY...........I...D....L....DV.M.H.-.RH.VS..AL..........--
SBmhc1_50/1-77   ........S.....A.MI...D....................G...GLI.A.-.W......I..........--
SBmhc1_51/1-77   ..............A.MI....................L..Q..GLI.A.-.V...............--
SBmhc1_52/1-77   ..............A.MI....................L..Q..GLI.T.-.N...............--
SBmhc1_53/1-77   ........A.....VPMNY.......T.........AGDRAD..DWQ..IIV.T.-.T........I.........--
SBmhc1_54/1-77   ..............DV.MI.............F.E.....L..N..IY.DD.-.W..G...........--
SBmhc1_55/1-77   ........S.....DV.MF............D.......M..EVSVRA.-RT..T....GR.........--
SBmhc1_56/1-77   ..............A.IS................S...GV..IF..S.-PT..D...IL.P........--
SBmhc1_57/1-77   ........A.....V.MY........K.......VN--K.A........N.GN..AA.-.S..V...........VH
SBmhc1_58/1-77   ........I.....A.IES.....M.F.......EGKAD...YS...S..T.-.I......IV.........--
SBmhc1_59/1-77   ........S.....DG.IDY......DV......INA........N..T...H.-.N..G...IL.........--
SBmhc1_60/1-77   ........S.....A.IDY................A.DRAD...YS..LA..S.-.T..G...I..........--
SBmhc1_61/1-77   ......ALA....V.INY........V.....VT--K.V........N..IR..H.-.N.......I..........VH
SBmhc1_62/1-77   ......Y.A.....G.MYY...................L....KS..A.-.T......IL.P.......--
SBmhc1_63/1-77   ........A..F..A.IES........VT--K.V..........KS..H.-.S..V....I.........VH
SBmhc1_64/1-77   ........I.....V.MY.............EGKAD...Y..GNG..T.-.I..G....V.........--
SBmhc1_65/1-77   ..............A.MI........K........DRAD..DWQ.DIFA.T.-.V..G...........--
SBmhc1_67/1-77   ........A.....G.MYY..................L..N..IS..H.-.T...S..NV.P.......--
SBmhc1_68/1-77   ........A.....G.MYY...................L....KS..A.-.T......IL.P.......--
SBmhc1_69/1-77   ......V......V.IES........V.......EGKAD..DWQ..LA..S.-.T..G....V.........--
SBmhc1_70/1-77   ..............G.IDY.......................Q..IY.SA.-TL.....V.P.........--
SBmhc1_71/1-77   ........A..F..G.I.............T--KVV........N.GNF.VN.-.W..V...I..........VH
SBmhc1_72/1-77   .....QY.S....DV.IS.C....M.N............Q..L..N.-.W......I.........--
SBmhc1_73/1-77   ........A.....DV.MYY..............D...L.GN.N...S.-.T..V....V...L.......--
SBmhc1_74/1-77   .............R....S......................T...-................--
SBmhc1_75/1-77   ........A.....IPMNY.......FIS.............L..Y..VF..T.-.S......I.........--
SBmhc1_76/1-77   .....QY.S....DV.IS.C......N.......N.A........N..T...H.-.N......I...........--
```

```
SBmhc1_77/1-77    ................DV.MI......G...L.......F.E....L..N..IY.DD.-.W..G................--
SBmhc1_79/1-77    .......Y.A......A.MI................D....L..Q..IS..H.-.S.....IV.............--
SBmhc1_80/1-77    ........ALA...V.INY.......V.................Q..IIV.T.-.T..V....V...L.....--
SBmhc1_81/1-77    ........A.....DV.MYY................D....L.GN..N..S.-.T......IL.P.......--
SBmhc1_82/1-77    ........I.....A.IES......V..........EGKAD...YS..A..A.-.N.....I.........--
SBmhc1_83/1-77    ........S.....V.IY.............I.......L....DV.KRH.-RRH.VS..AL.........--
SBmhc1_84/1-77    ........S.....G.IDY.....KDV........TD......G..KR..N.-.W..V.............--
SBmhc1_85/1-77    ..............V.MYY......T..........D.......DV.MSH.-.RH.VS..RL.........--
SBmhc1_86/1-77    ........I.....A.MI.....K...........EGKAD..DWQ..A..T.-.V...G...V........--
SBmhc1_87/1-77    ........A.....IPMNY.......FIS..........L..Y..VF..N.-.V.....IL.P.......--
SBmhc1_88/1-77    ........I..F..A.IEY.....T.FV....VT--K.V...F...N..IA..T.-.T...S..NV.P......VH
SBmhc1_89/1-77    ..............G.IDY.............INA...........SIV.S.-.T......I.....S...--
SBmhc1_90/1-77    ........A.....V.IDY......T..........EGKAD...YS...S..A.-.V..G..IL.......--
SBmhc1_91/1-77    ..............V.MF......NK...............Q...S...-R........GR.......--
SBmhc1_92/1-77    .....QY.S.....DV.IS.C....M.N..............GQ..L..H.-.W.....I........--
SBmhc1_93/1-77    ........I.....A.MI......G..............L.....KS..A.-.T.....IL.P.......--
SBmhc1_95/1-77    ........................S..............-....-.............-.......--
SBmhc1_96/1-77    ........AA...F..G.IEY......T.....I....D...L..H..IF..T.-PT...G..IL.P....--
SBmhc1_97/1-77    ........S.....G.IDY.......DV........DRAD...G...KR..H.-.T......IL......--
SBmhc1_98/1-77    ........S......MF...........I...D....L...DV.M.H.-.RH.VS..AL.........--
SBmhc1_99/1-77    ..............A.MI.................EGKAD...Q..MWV.D.-.N.....I..........--
SBmhc1_100/1-77   .......Y.A.....A.MI................D....L..H..IF..T.-.S.....I.........--
SBmhc1_101/1-77   ..............DG.IDY............INA...........SIV.S.-.V.....I.....S...--
SBmhc1_102/1-77   ........S.....G.IDY...........V.......N.GN..T.-.T...........L.......--
SBmhc1_103/1-77   ........A.....DV.MY................D....L.TQ..L..H.-.N......V........--
SBmhc1_104/1-77   ........I.....A.MI......K..........EGKAD..DYQ..MWV.T.-.I..G...IL.P....--
SBmhc1_105/1-77   ........S.....V.IY................I.A..D......DV.KRH.-.RH.VS..AL..HL....--
SBmhc1_106/1-77   ........VA.V...ILSGYC.N.KNIV......VNKLFED..HLQWY.TK.SDDEPHF..TTLDNL.......---
SBmhc1_107/1-77   ..............A.IS...............S....GV..IF..S.-.T..G..IL.P....--
SBmhc1_108/1-77   .......F......V.IDY.....................T...F.-.I........G......--
SBmhc1_110/1-77   .......Y.A.....A.MI...................L....KS..T.-.N......V.........--
SBmhc1_111/1-77   ........A..F..A.IES...........VT--K.V..........KS..H.-.T..V...I........VH
SBmhc1_112/1-77   ........S.....V.MF.............I.A..D......DV.T.H.-.RH.DG..YI........--
SBmhc1_113/1-77   ........I.....V.MY.................EGKAD...Y..GNG..T.-.I..G.........--
SBmhc1_114/1-77   ........S.....A.MI....D............G..GLI.A.-.W...........--
SBmhc1_115/1-77   ........I.....A.IEY.....KFV.....N--K.V......N..KS..A.-.I..V..DIV........VH
SBmhc1_116/1-77   ..............A.MI....D...........L..Q..GLI.A.-.V............--
SBmhc1_118/1-77   ........S.....DV.MF................D........SVRA.-RT..T....G.......--
SBmhc1_119/1-77   ........I.....A.IES......F.....N..EGKAD...YS...S..T.-.T.....IV........--
SBmhc1_120/1-77   ...............................EVSVRA.-RT..T........--
SBmhc1_121/1-77   ........ALA...V.INY............INA..D......N..IR..A.-.T..G...I........--
SBmhc1_122/1-77   ........VAVV...ILCGYC.N.KNI......VKNIFED...HL.FY.RE.SEY.PNF...T.HHL.......---
SBmhc1_123/1-77   ........S.....DV.MF.................EVSVRA.-RT..T....G.......--
SBmhc1_125/1-77   ..............DG.IDY.......FV.....T...........SIV.A.-.V.....I.........--
SBmhc1_126/1-77   ..............DV.MF......M.....I...D....NGQ..TS.HS.-RT......IV.........--
SBmhc1_127/1-77   ..............V.MF......G...........DRAD...TQ..MLA.A.-.N..G...IV......--
SBmhc1_128/1-77   ........I.....V.MY......T...........F.D....L.TQ..L...A.-.T..G...I.......--
SBmhc1_129/1-77   .......Y.A.....A.MI................D....L..H..IF..T.-.S......V........--
SBmhc1_130/1-77   .......Y.A.....A.MI...................N..L...S.-.V.....I..........--
SBmhc1_131/1-77   ........ALS...DV.MNY......TV.....INA..D......TQ..IR..A.-.V..D...I.........--
```

Figure 4.1 Codon alignment (Clustal W) of all MHC class 1 alleles sequenced in this study. Yellow highlighting indicates pervasive/episodic positive/diversifying selection detected by MEME, FEL and FUBAR. Blue highlighting indicates negative episodic negative/purifying selection detected by both FEL and FUBAR. Pink highlighting indicates breakpoint (nucleotide 109) detected by SBP (all analyses performed using datamonkey.org tools). Asterisks represent codons which align to residues under selection at the PBR of the human HLA-A2 gene (Grimholt et al, 2015).

Recombination and Selection analysis

Strong support for recombination was found at position 109, inferred from both AIC and BIC with 100% support. MEME analysis indicated that there was possible episodic positive/diversifying selection at 23 codon sites (Table 4.1, Figure 4.1). FEL analysis indicated that there was pervasive positive/diversifying selection at 19 sites and pervasive negative/purifying selection at 6 sites (Table 4.1). FUBAR found evidence of episodic positive/diversifying selection at 20 sites exhibited episodic negative/purifying selection at 8 sites. In total, 14 sites were identified as being under positive selection by all 3 models (Table 4.1, Figure 4.1) and a further 5 sites that were identified as under negative selection by both FUBAR and FEL (Table 4.1). Of the positively selected sites, seven coincided with residues contributing to the peptide binding pockets of the human HLA-A2 gene, based on the alignment of the sequences with those from (Grimholt *et al.* 2015, Figure 4.1).

Table 4.1 Selection models used to identify codons under positive and negative selection. Stars indicate codons where all models identified either positive (Shared +, MEME, FEL, FUBAR) or negative selection (Shared –, FEL and FUBAR) at a given codon site.

Model	FEL	MEME	FEL	FUBAR		
Codon Site	Omega	*P*	*P*	Bayes Factor [$\alpha<\beta$]	Shared +	Shared -
1	NaN	1	1	0.211		
2	NaN	1	1	0.164		
3	NaN	1	1	0.158		
4	NaN	1	1	0.133		
5	NaN	1	1	0.08		
6	0	0.67	0.114	0.045		
7	Infinity	0.48	0.577	0.651		
8	Infinity	0.15	0.201	1.87		

9	0.063	0.67	0	0.045		*
10	27.753	0	0	3371286.466	*	
11	Infinity	0.1	0.081	6.158		
12	0.149	0.67	0.002	0.005		*
13	0.071	0.67	0.003	0.005		*
14	1.574	0.49	0.578	1.54		
15	0	0.67	0.09	0.025		*
16	2.785	0.33	0.337	7.023		
17	8.854	0	0	11966.017	*	
18	Infinity	0.24	0.216	2.351		
19	2.076	0.41	0.455	7.788		
20	Infinity	0	0	65634.153	*	
21	Infinity	0	0.002	100.014	*	
22	1.363	0.59	0.778	0.883		
23	NaN	1	1	0.131		
24	0.271	0.67	0.373	0.109		
25	Infinity	0.2	0.307	1.245		
26	0.38	0.26	0.486	0.258		
27	Infinity	0.05	0.035	16.774		
28	0.813	0.67	0.807	0.819		

29	0.72	0	0.349	0	
30	Infinity	0.03	0.026	21.043	*
31	0.685	0.55	0.685	0.407	
32	NaN	1	1	0.14	
33	NaN	1	1	0.125	
34	Infinity	0.42	0.47	0.828	
35	NaN	1	1	0.055	
36	Infinity	0.39	0.422	1.603	
37	6.715	0.04	0.024	30.418	*
38	Infinity	0	0.041	14.206	*
39	3.163	0.09	0.073	16.993	
40	0.898	0.17	0.902	0.907	
41	1.171	0.59	0.785	2.165	
42	1.669	0.35	0.364	6.452	
43	Infinity	0	0.064	6.245	
44	0.301	0.67	0.152	0.087	
45	Infinity	0.01	0.403	1.001	
46	0.379	0.67	0.229	0.184	
47	Infinity	0.19	0.169	5.896	
48	Infinity	0.09	0.099	7.816	

49	2.885	0.14	0.113	79.843		
50	Infinity	0	0	53191.85	*	
51	NaN	1	1	0.194		
52	3.604	0	0.186	11.382		
53	Infinity	0	0	35392422.56	*	
54	Infinity	0	0	124144.576	*	
55	0.973	0.67	0.954	1.156		
56	7.542	0.01	0.012	30.295	*	
57	13.127	0	0	1100889888	*	
58	0.04	0.67	0.001	0.004		*
59	1.004	0.67	0.996	1.779		
60	Infinity	0.21	0.19	29.038		
61	NaN	1	1	86932370.99		
62	0.533	0.07	0.608	0.229		
63	NaN	1	1	0.14		
64	Infinity	0	0	1620987.456	*	
65	2.105	0	0.286	5.236		
66	Infinity	0.39	0.419	1.167		
67	Infinity	0.24	0.22	2.406		
68	0.443	0.67	0.01	0		*

69	9.46	0	0	20430.617	*	
70	Infinity	0.23	0.438	0.944		
71	0.446	0.6	0.277	0.451		
72	0.491	0.67	0.62	0.26		
73	Infinity	0.18	0.232	1.633		
74	Infinity	0.3	0.308	1.239		
75	NaN	1	1	0.142		
76	NaN	1	1	0.194		
77	NaN	1	1	0.211		
78	NaN	1	1	0.206		
79	NaN	1	1	0.555		
80	NaN	1	1	0.464		0

Table 4.2 Magnitude of amino acid property change at nonsynonymous residues of the entire fragment for the categories (magnitude classes) of greatest change (7 and 8). Stars indicate significance level.

Property	Category	Z Value	Significance level (P value)		
			.05	.01	.001
Equilibrium constant (ionization of COOH)	8	2.029	*		
Normalized consensus hydrophobicity	8	-1.779	*		
Power to be at the C-terminal	8	-3.429	*	*	*
Power to be at the N-terminal	8	-2.155	*		
Thermodynamic transfer hydrohphobicity	8	-2.18	*		
Buriedness	7	-2.143	*		
Chromatographic index	7	-1.83	*		
Compressibility	7	-3.252	*	*	*
Helical contact area	7	-1.758	*		
Hydropathy	7	-3.016	*	*	
Isoelectric point	7	-2.299	*		
Mean r.m.s. fluctuation displacement	7	-3.297	*	*	*
Polarity	7	-2.878	*	*	
Power to be at the C-terminal	7	-4.165	*	*	*
Power to be at the middle of alpha-helix	7	5.872	*	*	*
Power to be at the N-terminal	7	-2.108	*		

Refractive index	7	-2.484	*	*	
Short and medium range non-bonded energy	7	-2.55	*	*	
Solvent accessible reduction ratio	7	3.667	*	*	*
Thermodynamic transfer hydrohphobicity	7	4.262	*	*	*
Total non-bonded energy	7	-3.01	*	*	

MHC class 1-alpha can reveal cryptic fine-scale population structure in a commercial fish, the European sea bass (Dicentrarchus labrax)

104

Amino acid properties under selection (TreeSAAP)

TreeSAAP analysis identified significant deviations from neutral expectations (categories 7 and 8, Table 4.2, Appendix 3: Figure A3.1), within the whole sequenced fragment, for the following physicochemical amino acid properties: 'Power to be at the C-terminus' (energy potential of the C-terminus of an alpha helix to interact with other residues), 'Power to be at the middle of the α-Helix' (energy potential of the middle of an alpha helix to interact with other residues), 'Mean r.m.s. fluctuation displacement' (the ability of a residue to change position in three- dimensional space), 'compressibility' (the contribution of a residue to the local density of protein secondary structures), 'Solvent accessible reduction ratio' (the reduction ratio of cross peak intensity when residues are or are not irradiated with aliphatic protons) and 'Thermodynamic transfer hydrophobicity' (the difference in the solubility of amino acids in water and ethanol) (Appendix 3: Figure A3.1, Table 4.2). This indicates that selection pressure on these properties have driven changes at this region (Woolley et al., 2003).

Supertype and loci clustering

BIC values indicated that there were four main super-type clusters (Appendix 3: Figure A2.2A-C) before BIC values decreased by a negligible amount (Appendix 3: Figure A3.3). We also explored a finer scale clustering, up to eight clusters, which represented the lowest BIC value (Appendix 3: Figure A3.3). Assignment of potential alleles to individuals indicated that there could be between 9 and 18 loci, depending on the method used (Appendix 3: Figures A3.4, A3.6-A3.10, Table A3:2) but none of the clustering methods resulted in all individuals in the dataset having only 2 alleles in a cluster (Appendix 3: Table A3.2), even when fine clustering (18 clusters) was applied. Seabass sequences clustered closer to class I-alpha lineage I than to the other lineages (Appendix 3: Figure A3.5). Two sequences, SBmhc1_106 and SBmhc1_122, clustered with lineage I sequences from salmonids (Atlantic salmon, brown trout and rainbow trout), and close to Medaka sequences, also from lineage I.

Population analysis

We found 30 unique alleles (representing 29 unique amino acid sequences) in the Celtic Shelf population that were not represented in the Portuguese population and 22 alleles (21 unique amino acid sequences) unique to the Portuguese population (Figure 4.2A), including the 2 alleles with an insertion at codon 59 in the alignment (SBmhc1_106 and SBmhc1_122). Private alleles were present across the phylogenetic tree, but Celtic shelf alleles were more frequent in recently diverged clusters than those from Portugal (Figure 4.2A). Most unique alleles appeared in only one or two individuals in the respective populations (Appendix 3: Table A3.4), with the exception of allele SBmhc1_25 and

SBmhc1_38, which appeared in six and three Portuguese individuals respectively, and SBmhc1_54, present in 3 Celtic Shelf individuals. There was no difference in the number of alleles per fish between the two populations (mean of Celtic population =12.64 alleles per individual, mean of Portuguese population = 12.79 alleles per individual; Welch Two Sample t-test t = -0.26442, df = 38.999, p-value = 0.7928). DAPC analysis indicated low differentiation in allele frequencies as a whole (Figure 4.2B) between the two regions (D_{est} = 0.05), with most of the variance described by the first principal component. Based on all the alleles, we were able to assign 75% of the fish (31/41) to their region of origin (Celtic Shelf or Portugal). The accuracy of the assignment was significantly greater than chance (Accuracy = 0.756; 95% CI: 0.597, 0.8764; P= 0.003).

A
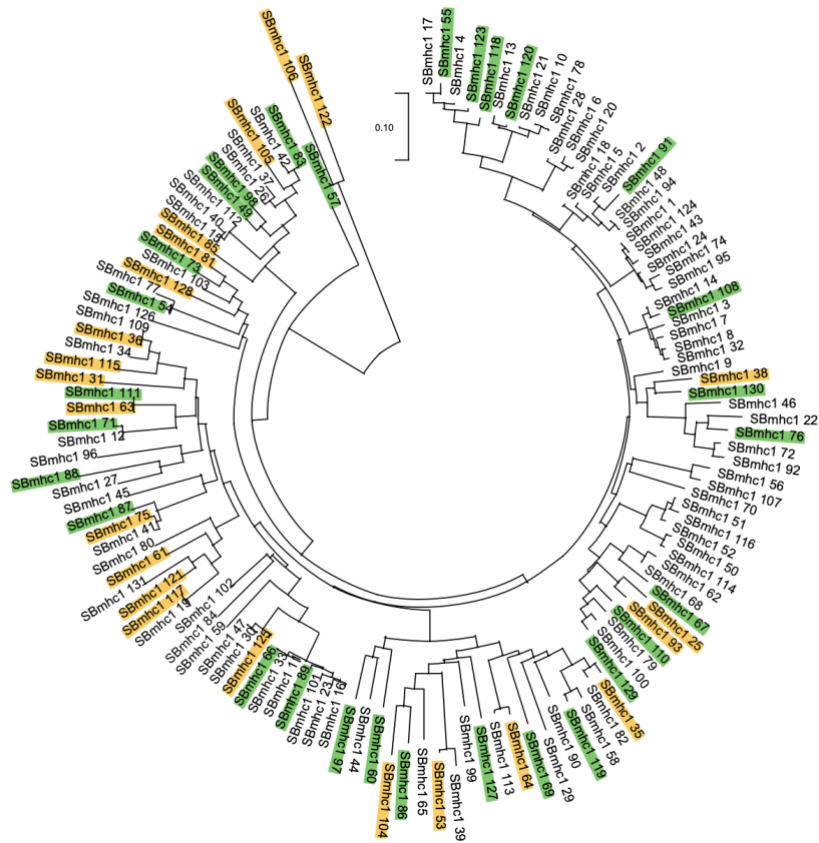
B
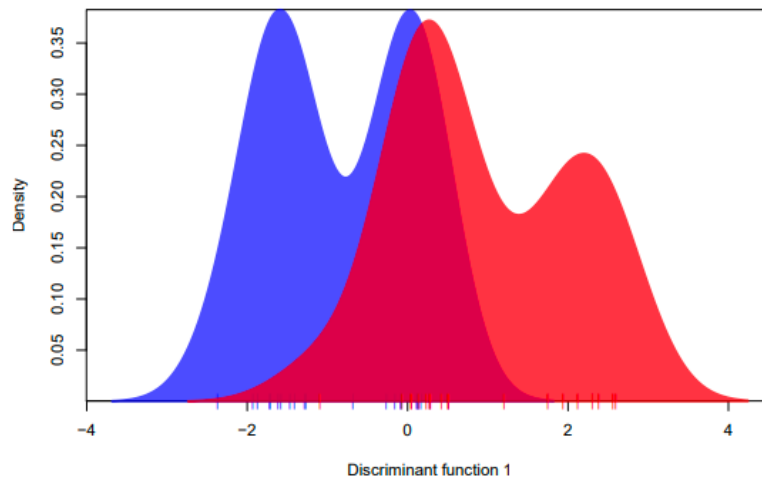
Figure 4.2 **A.** Neighbour-joining phylogenetic tree of the *Dicentrarchus labrax* MHC class I-alpha alleles identified in this study. Celtic Shelf population alleles highlighted in green and Portugal population alleles highlighted in yellow. **B.** Individual densities of allele frequencies (presence/absence for each individual) for the Portuguese and Celtic Shelf populations, plotted against the first discriminant component used in analysis of principal components (DAPC).

4.5   Discussion

The high variability of MHC genes is known to be driven by natural and sexual selection and, for this reason, they represent good candidates to assess fine scale structuring potentially derived from local selective pressures (Larson et al. 2019). We found that the MHC class I alpha gene in the Atlantic seabass displays high allelic diversity, evidence for selection and, importantly, potential for detecting fine scale structuring within stocks largely considered panmictic (ICES 2021). Our study represents the first population study of MHC class I in seabass and indicates that, as for other fishes, the *D. labrax* MHC class I-alpha locus has undergone strong positive selection and recombination, while maintaining intermediate copy numbers and relatively few supertypes. This is similar to the diversity of MHC class II in guppies, where supertypes are maintained by balancing selection and consist of groups of alleles with similar functionality that are subject to positive selection (Lighten *et al.* 2017).

Most seabass MHC-I sequences clustered close to the lineage I of the U-type classical alleles from different fish species, two of them grouping with salmonid alleles (Grimholt *et al.* 2015). Close clustering between seabass and salmonid alleles might indicate trans-species polymorphism, typical of MHC genes (Klein et al. 1998), but this must be interpreted with caution as it is based on the alpha-1 domain only and the bootstrapping support was relatively low. We found an average of 13 copies per individual in *D. labrax*, which is similar to other teleost species with U lineage genes, such as the African moonfish, *Selene dorsalis* (mean copy number 18) and Sand roller *Percopsis transmontana* (mean copy number 13) (Malmstrøm et al. 2016). In general, teleost MHC class I genes are highly polymorphic, with copy number varying between families and species (Malmstrøm *et al.* 2016). For instance, Atlantic salmon (*Salmo salar*) has a single MHC I locus (Grimholt et al. 2003), while three spined stickleback, *Gasterosteus aculeatus* possess between 3 and 9 alleles per individual, which are difficult to assign to loci probably due to high allele similarity (Aeschlimann et al. 2003). In contrast, most gadoids have lost classical MHC II genes and possess a largely expanded number of MHC I copies, for example Atlantic cod (*Gadus morhua*) which has approximately 100 (Star et al. 2011). The observed diversity in *D. labrax* was, therefore, within the range expected for a species with a classical MHC II and fitted well with the previously predicted 6-8+ MHC-I loci for the species (Pinto *et al.* 2013). Yet, as our analysis was based on DNA sequences and not on expressed alleles (mRNA), we could not identify the exact number of loci and it is possible that some of the sequences could represent pseudogenes.

Positive selection was identified at 14 polymorphic sites across the 80 codons in this study, some of them coinciding with PBR sites previously defined in fish based on the alignment with human sequences (Grimholt *et al.* 2015). This level of selection is comparable to other fishes (Wegner 2008) for example sockeye salmon (*Oncorhynchus nerka*) where 4 codons are under positive selection in a 32 codon fragment (McClelland *et al.* 2011). High rates of positive selection are almost always observed in MHC alleles, driven by parasite mediated balancing selection or sexual selection (Bernatchez & Landry 2003). In addition, recombination can also act as a significant force to rapidly increase diversity in MHC alleles (Consuegra et al. 2005), and we found strong support for a breakpoint at position 109bp, possibly contributing to the observed allelic diversity.

While codon based selection models give an indication of the areas of the gene that are under selection, the ability to bind antigens from specific pathogens is determined by changes in the shape of the cleft of the peptide binding region (PBR) (Dionne *et al.* 2007). Nucleotide differences, or even amino acid differences, may not directly translate to protein binding differences, and additional alleles may not necessarily confer the ability to respond to a greater range of pathogens, unless the changes in the protein also change the binding properties of that allele (Ellison et al. 2012; Lighten *et al.* 2017). The analysis of functional differences based on binding properties analysis indicated that the energy potential of the C-terminus, the position of amino acids at the middle of an alpha helix (where they can interact with other residues), and the ability for a residue to change position in three-dimensional space were under selection, supporting the role of selective pressures (Doytchinova & Flower 2005; Lighten *et al.* 2017). Alleles within a supertype are predicted to bind similar antigenic 'supermotifs' (Phillips et al. 2018). We identified between 4 and 8 clusters, although the precise number of clusters is not clear. Clustering can be challenging when loci share identical alleles, there are null alleles or, as found here, copy number varies between haplotypes (Huang et al. 2019). Even with an upper estimate of 8 clusters, we found relatively low supertype diversity in comparison to other fishes, despite a relatively diverse allele pool of 131 alleles. For instance, in guppies, 66 alleles were clustered to 13 supertypes (Smallbone et al. 2021). This could be the result of convergent evolution resulting in a relatively small number of overlapping peptide-binding motifs, as for the highly diverse human class I alleles that may be clustered into as few as 9 supertypes (Sidney et al. 2008).

Despite loci remaining undefined, we observed allelic differentiation in the two seabass populations we compared, confirmed by DAPC analysis, although private alleles were detected in relatively few individuals within each region. Overall, the differentiation between regions was small, but based on the MHC sequences identified here, we were able to assign 75% of the fish to their region of origin (Celtic Shelf or Portugal) with high confidence. Previous studies of Eastern Atlantic seabass using

microsatellite and SNP (Single Nucleotide Polymorphism) analysis failed to identify consistent population structuring (Souche *et al.* 2015), and tagging studies have also shown that *D. labrax* are capable of swimming large distances, e.g. 1200km within 2 months of tags being deployed, and migrate considerable distances offshore to spawn (Pawson *et al.* 2007). It had therefore been assumed that Eastern Atlantic seabass formed one single panmictic population. Seabass seems to display a shallow population structuring based on neutral markers but, as for other marine species (Milano et al. 2014), a finer level of structuring resulting from local selective pressures can be detected when markers under selection are used. The difference in private alleles between the Celtic shelf and Portuguese uncovered population structuring not observed using neutral markers (Souche *et al.* 2015), and indicates the potential for MHC-I to detect fine-scale population variation. Using genome-wide high-throughput sequencing can help identify markers under selection (Carreras et al. 2017), but genome-wide scans can be time-consuming and expensive. However, markers within the MHC region can be particularly good for showing fine scale differentiation, as changes can accumulate relatively quickly and are affected by local selection (Consuegra *et al.* 2005). Our study provides the first detailed analysis of MHC class I in seabass, a species that supports important commercial and sport fisheries, as well as aquaculture. Our approach provides a quick tool to screen the highly variable MHC region of Atlantic seabass for fisheries management purposes, that can be adapted to other marine species. Moreover, the same approach can be used for genotyping farmed species and could potentially be used for selective breeding for disease resistance (Pawluk et al. 2019) and to monitor the incidence of escapees from fish farms (Monzón-Argüello et al. 2013).

Environmental complexity has the potential to create refugia for marine species that may result in local adaptation and create cryptic population structuring, essential for the long term persistence of exploited stocks (Midway et al. 2018). In many marine species with high fecundity and dispersal rates, this fine population structure can only be identified by markers under natural selection (André et al. 2011; Lamichhaney et al. 2012; Jorde et al. 2018). Based on our initial results from sea bass, we propose that MHC markers can be used for the management of marine species with cryptic population structure, for which preserving their fine population structuring is needed to maintain their functional biodiversity (Limborg *et al.* 2012).

Data Accessibility

Sequences have been submitted to GenBank and are accessible under Accession numbers MZ466411-MZ466541.

Ethics approval

All work was carried out with approval from Swansea University Ethics Committee (Reference number SU-Ethics-Student-280621/4326).

4.6   references

Aeschlimann P.B., Haberli M.A., Reusch T.B.H., Boehm T. & Milinski M. (2003) Female Sticklebacks Gasterosteus aculeatus Use Self-Reference to Optimize MHC Allele Number during Mate Selection. Behavioral Ecology and Sociobiology 54, 119-26.

André C., Larsson L.C., Laikre L., Bekkevold D., Brigham J., Carvalho G., Dahlgren T., Hutchinson W., Mariani S. & Mudde K. (2011) Detecting population structure in a high gene-flow species, Atlantic herring (Clupea harengus): direct, simultaneous evaluation of neutral vs putatively selected loci. Heredity 106, 270-80.

Beddington J.R., Agnew D.J. & Clark C.W. (2007) Current problems in the management of marine fisheries. Science 316, 1713-6.

Bento E.G., Grilo T.F., Nyitrai D., Dolbeth M., Pardal M.Â. & Martinho F. (2016) Climate influence on juvenile European sea bass (Dicentrarchus labrax, L.) populations in an estuarine nursery: a decadal overview. Marine Environmental Research 122, 93-104.

Bernatchez L. & Landry C. (2003) MHC studies in nonmodel vertebrates: what have we learned about natural selection in 15 years? Journal of Evolutionary Biology 16, 363-77.

Biedrzycka A., O'Connor E., Sebastian A., Migalska M., Radwan J., Zając T., Bielański W., Solarz W., Ćmiel A. & Westerdahl H. (2017) Extreme MHC class I diversity in the sedge warbler (Acrocephalus schoenobaenus); selection patterns and allelic divergence suggest that different genes have different functions. BMC Evolutionary Biology 17, 1-12.

Carreras C., Ordóñez V., Zane L., Kruschel C., Nasto I., Macpherson E. & Pascual M. (2017) Population genomics of an endemic Mediterranean fish: differentiation by fine scale dispersal and adaptation. Scientific Reports 7, 1-12.

Castilho R. & McAndrew B. (1998) Population structure of seabass in Portugal: evidence from allozymes. Journal of Fish Biology 53, 1038-49.

Consuegra S. & Garcia de Leaniz C. (2008) MHC-mediated mate choice increases parasite resistance in salmon. Proceedings of the Royal Society Biological Sciences Series B 275, 1397-403.

Consuegra S., Megens H.J., Schaschl H., Leon K., Stet R.J.M. & Jordan W.C. (2005) Rapid Evolution of the MH Class I Locus Results in Different Allelic Compositions in Recently Diverged Populations of Atlantic Salmon. Molecular Biology and Evolution 22, 1095-106.

Coscia I. & Mariani S. (2011) Phylogeography and population structure of European sea bass in the north-east Atlantic. Biological Journal of the Linnean Society 104, 364-77.

De Pontual H., Lalire M., Fablet R., Laspougeas C., Garren F., Martin S., Drogou M. & Woillez M. (2019) New insights into behavioural ecology of European seabass off the West Coast of France: implications at local and population scales. ICES Journal of Marine Science 76, 501-15.

Dionne M., Miller K.M., Dodson J.J., Caron F. & Bernatchez L. (2007) Clinal variation in MHC diversity with temperature: evidence for the role of host–pathogen interaction on local adaptation in Atlantic salmon. Evolution 61, 2154-64.

Doytchinova I.A. & Flower D.R. (2005) In silico identification of supertypes for class II MHCs. Journal of Immunology 174, 7085-95.

Drogou M., Biseau A. & Le Goff R. (2014) European sea bass stocks and management measures within the European Union. In: Policy Department B: Structural and Cohesion Policies (ed. by R. Centenera), Brussels.

Duranton M., Allal F., Fraïsse C., Bierne N., Bonhomme F. & Gagnaire P.-A. (2018) The origin and remolding of genomic islands of differentiation in the European sea bass. Nature communications 9, 1-11.

Edwards S.V. & Hedrick P.W. (1998) Evolution and ecology of MHC molecules: from genomics to sexual selection. Trends in Ecology & Evolution 13, 305-11.

Eizaguirre C., Lenz T.L., Kalbe M. & Milinski M. (2012) Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. Nat Commun 3, 621.

Ellison A., Allainguillaume J., Girdwood S., Pachebat J., Peat K.M., Wright P. & Consuegra S. (2012) Maintaining functional major histocompatibility complex diversity under inbreeding: the case of a selfing vertebrate. Proceedings of the Royal Society B: Biological Sciences 279, 5004-13.

Felip A., Piferrer F., Carrillo M. & Zanuy S. (2001) Comparison of the gonadal development and plasma levels of sex steroid hormones in diploid and triploid sea bass, Dicentrarchus labrax L. Journal of Experimental Zoology 290, 384-95.

Gagnaire P.A., Broquet T., Aurelle D., Viard F., Souissi A., Bonhomme F., Arnaud-Haond S. & Bierne N. (2015) Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era. Evolutionary applications 8, 769-86.

Gandra M., Assis J., Martins M.R. & Abecasis D. (2021) Reduced global genetic differentiation of exploited marine fish species. Molecular Biology and Evolution 38, 1402-12.

Grimholt U., Larsen S., Nordmo R., Midtlyng P., Kjoeglum S., Storset A., Saebo S. & Stet R.J.M. (2003) MHC polymorphism and disease resistance in Atlantic salmon (Salmo salar); facing pathogens with single expressed major histocompatibility class I and class II loci. Immunogenetics 55, 210 - 9.

Grimholt U., Tsukamoto K., Azuma T., Leong J., Koop B.F. & Dijkstra J.M. (2015) A comprehensive analysis of teleost MHC class I sequences. BMC Evolutionary Biology 15, 1-17.

Hammer O., Harper D. & Ryan P. (2001) PAST: Paleontological statistics software package for education and data analysis. Paleontologia Electronica 4, 9.

Hedgecock D., Barber P.H. & Edmands S. (2007) Genetic approaches to measuring connectivity. Oceanography 20, 70-9.

Hedrick P. & Kim T. (2000) Genetics of complex polymorphisms: parasites and maintenance of MHC varation. Harvard University Press, Cambridge, MA.

Huang K., Zhang P., Dunn D.W., Wang T., Mi R. & Li B. (2019) Assigning alleles to different loci in amplifications of duplicated loci. Molecular ecology resources 19, 1240-53.

Huson D.H. & Bryant D. (2006) Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23, 254-67.

ICES (2021) Working Group for the Celtic Seas Ecoregion (WGCSE). In: ICES Scientific Reports (eds. by M. Lundy & S. Nimmegeers), p. 1461.

Janeway C.A., Travers P., Walport D. & Shlomchik M.J. (2004) Immunobiology: The Immune System in Health and Disease.

Jombart T. (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403-5.

Jombart T. & Collins C. (2015) A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0. 0. London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling.

Jorde P.E., Synnes A.E., Espeland S.H., Sodeland M. & Knutsen H. (2018) Can we rely on selected genetic markers for population identification? Evidence from coastal Atlantic cod. Ecology and Evolution 8, 12547-58.

Jost L. (2008) GST and its relatives do not measure differentiation. Molecular Ecology 17, 4015-26.

Klein J., Sato A., Nagl S. & O'hUigín C. (1998) Molecular trans-species polymorphism. Annual Review of Ecology and Systematics 29, 1-21.

Kosakovsky Pond S. & Frost S. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution 22, 1208-22.

Kosakovsky Pond S.L., Posada D., Gravenor M.B., Woelk C.H. & Frost S.D. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Molecular Biology and Evolution 23, 1891-901.

Kuhn M. (2009) The caret package. Journal of Statistical Software 28.

Kumar S., Stecher G., Li M., Knyaz C. & Tamura K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35, 1547.

Lamichhaney S., Barrio A.M., Rafati N., Sundström G., Rubin C.-J., Gilbert E.R., Berglund J., Wetterbom A., Laikre L. & Webster M.T. (2012) Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences 109, 19345-50.

Larson W.A., Dann T.H., Limborg M.T., McKinney G.J., Seeb J.E. & Seeb L.W. (2019) Parallel signatures of selection at genomic islands of divergence and the major histocompatibility complex in ecotypes of sockeye salmon across Alaska. Molecular Ecology 28, 2254-71.

Lemaire C., Versini J.J. & Bonhomme F. (2005) Maintenance of genetic differentiation across a transition zone in the sea: discordance between nuclear and cytoplasmic markers. Journal of Evolutionary Biology 18, 70-80.

Lighten J., Papadopulos A.S., Mohammed R.S., Ward B.J., Paterson I.G., Baillie L., Bradbury I.R., Hendry A.P., Bentzen P. & Van Oosterhout C. (2017) Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. Nature communications 8, 1-11.

Limborg M.T., Helyar S.J., De Bruyn M., Taylor M.I., Nielsen E.E., Ogden R., Carvalho G.R., Consortium F. & Bekkevold D. (2012) Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (Clupea harengus). Molecular Ecology 21, 3686-703.

Malmstrøm M., Matschiner M., Tørresen O.K., Star B., Snipen L.G., Hansen T.F., Baalsrud H.T., Nederbragt A.J., Hanel R. & Salzburger W. (2016) Evolution of the immune system influences speciation rates in teleost fishes. Nature Genetics 48, 1204-10.

McClelland E.K., Ming T.J., Tabata A. & Miller K.M. (2011) Sequence analysis of MHC class I α2 from sockeye salmon (Oncorhynchus nerka). Fish & Shellfish Immunology 31, 507-10.

Midway S.R., White J.W. & Scharf F.S. (2018) The potential for cryptic population structure to sustain a heavily exploited marine flatfish stock. Marine and Coastal Fisheries 10, 411-23.

Migalska M., Sebastian A. & Radwan J. (2019) Major histocompatibility complex class I diversity limits the repertoire of T cell receptors. Proceedings of the National Academy of Sciences 116, 5021-6.

Milano I., Babbucci M., Cariani A., Atanassova M., Bekkevold D., Carvalho G.R., Espiñeira M., Fiorentino F., Garofalo G. & Geffen A.J. (2014) Outlier SNP markers reveal fine-scale genetic structuring across E uropean hake populations (M erluccius merluccius). Molecular Ecology 23, 118-35.

Milinski M. (2006) The major histocompatibility comlex, sexual selection, and mate choice. Annual Review of Ecology, Evolution, and Systematics 37, 159 - 86.

Monzón-Argüello C., de Leaniz C.G., Gajardo G. & Consuegra S. (2013) Less can be more: loss of MHC functional diversity can reflect adaptation to novel conditions during fish invasions. Ecology and Evolution 3, 3359-68.

Murrell B., Moola S., Mabona A., Weighill T., Sheward D., Kosakovsky Pond S.L. & Scheffler K. (2013) FUBAR : A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Molecular Biology and Evolution.

Pawluk R.J., Garcia de Leaniz C. & Consuegra S. (2019) Sea lice loads correlate with the diversity at the Major Histocompatibility Complex-related loci in farmed Atlantic salmon, Salmo salar. Journal of Fish Diseases 42, 1091.

Pawson M., Brown M., Leballeur J. & Pickett G. (2008) Will philopatry in sea bass, Dicentrarchus labrax, facilitate the use of catch-restricted areas for management of recreational fisheries? Fisheries Research 93, 240-3.

Pawson M., Pickett G., Leballeur J., Brown M. & Fritsch M. (2007) Migrations, fishery interactions, and management units of sea bass (Dicentrarchus labrax) in Northwest Europe. ICES Journal of Marine Science 64, 332-45.

Phillips K.P., Cable J., Mohammed R.S., Herdegen-Radwan M., Raubic J., Przesmycka K.J., Van Oosterhout C. & Radwan J. (2018) Immunogenetic novelty confers a selective advantage in host–pathogen coevolution. Proceedings of the National Academy of Sciences 115, 1552-7.

Pinto R.D., Randelli E., Buonocore F., Pereira P.J. & dos Santos N.M. (2013) Molecular cloning and characterization of sea bass (Dicentrarchus labrax, L.) MHC class I heavy chain and β2-microglobulin. Developmental & Comparative Immunology 39, 234-54.

Pond S.L.K., Posada D., Gravenor M.B., Woelk C.H. & Frost S.D.W. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Molecular Biology and Evolution 23, 1891 - 901.

Quéré N., Guinand B., Kuhl H., Reinhardt R., Bonhomme F. & Desmarais E. (2010) Genomic sequences and genetic differentiation at associated tandem repeat markers in growth hormone, somatolactin and insulin-like growth factor-1 genes of the sea bass, Dicentrarchus labrax. Aquatic Living Resources 23, 285-96.

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radwan J., Babik W., Kaufman J., Lenz T.L. & Winternitz J. (2020) Advances in the evolutionary understanding of MHC polymorphism. Trends in Genetics 36, 298-311.

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, 361-95.

Rozas J., Ferrer-Mata A., Sánchez-DelBarrio J.C., Guirao-Rico S., Librado P., Ramos-Onsins S.E. & Sánchez-Gracia A. (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Molecular Biology and Evolution 34, 3299-302.

Sebastian A., Herdegen M., Migalska M. & Radwan J. (2016) amplisas: A web server for multilocus genotyping using next-generation amplicon sequencing data. Molecular Ecology Resources 16, 498-510.

Sidney J., Peters B., Frahm N., Brander C. & Sette A. (2008) HLA class I supertypes: a revised and updated classification. BMC Immunology 9, 1-15.

Slade R.W. & McCallum H.I. (1992) Overdominant vs. frequency-dependent selection at MHC loci. Genetics 132, 861-4.

Smallbone W., Ellison A., Poulton S., van Oosterhout C. & Cable J. (2021) Depletion of MHC supertype during domestication can compromise immunocompetence. Molecular Ecology 30, 736-46.

Souche E.L., Hellemans B., Babbucci M., MacAoidh E., Guinand B., Bargelloni L., Chistiakov D.A., Patarnello T., Bonhomme F. & Martinsohn J.T. (2015) Range-wide population structure of European sea bass Dicentrarchus labrax. Biological Journal of the Linnean Society 116, 86-105.

Star B., Nederbragt A.J., Jentoft S., Grimholt U., Malmstrøm M., Gregers T.F., Rounge T.B., Paulsen J., Solbakken M.H. & Sharma A. (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207-10.

Tine M., Kuhl H., Gagnaire P.-A., Louro B., Desmarais E., Martins R.S., Hecht J., Knaust F., Belkhir K. & Klages S. (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nature communications 5, 1-10.

Vandeputte M., Gagnaire P.A. & Allal F. (2019) The European sea bass: a key marine fish model in the wild and in aquaculture. Animal Genetics 50, 195-206.

Walker N.D., Boyd R., Watson J., Kotz M., Radford Z., Readdy L., Sibly R., Roy S. & Hyder K. (2020) A spatially explicit individual-based model to support management of commercial and recreational fisheries for European sea bass Dicentrarchus labrax. Ecological Modelling 431, 109179.

Wegner K. (2008) Historical and contemporary selection of teleost MHC genes: did we leave the past behind? Journal of Fish Biology 73, 2110-32.

Winter D.J. (2012) MMOD: an R library for the calculation of population differentiation statistics. Molecular ecology resources 12, 1158-60.

Woolley S., Johnson J., Smith M., Crandall K. & McClellan D. (2003) TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. Bioinformatics 19, 671 - 2.

Chapter 5　Implications of consistent selection rates between alpha (α1) and beta (ß1) MHC class II domains in the clade (series) Eupercaria on molecular marker choice

5.1   Abstract

Population genetic analyses can provide a basis for stock designation for fisheries management. Molecular markers in genes that are subjected to strong selective pressures can illuminate fine scale population structure. Immune genes, which are crucial in triggering an immune response to pathogens, such as major histocompatibility complex (MHC) genes, are particularly suited for this purpose. In the MHC class II, there are two candidate regions which are directly involved with binding peptides derived from pathogens: the $\alpha 1$ and ß1 domains. Many model vertebrates exhibit higher rates of selection in the ß1 domain in comparison to the $\alpha 1$, therefore ß1 is often targeted for population differentiation analyses. Here, I assess the rates of positive and negative selection, trans-species functional supertype structure, and time-tree species divergence times between 300 base-pair fragments of the $\alpha 1$ and ß1 domains for 18 species (95 alleles per domain) within a group of bony fishes, the clade (series) Eupercaria.

Contrary to expectations of higher levels of selection at the ß1 domain, both domains had similar levels of positive (19 sites) and negative selection (21 and 30 respectively). In addition, 6 $\alpha 1$ and 5 ß1 trans-species functional supertype clusters were identified, further indicating similar selective pressures at both domains. Time-tree analysis indicated that species divergence times based on $\alpha 1$ and ß1 did not differ. However, $\alpha 1$ domain species divergence times were significantly more recent when compared to a calibrated (neutral) time tree. The $\alpha 1$ domain can therefore provide valuable insight into the selective pressures acting on the MHC class II for this clade, and it is therefore recommended that both domains be targeted when designing markers to assess inter and intraspecific differentiation within this clade.

5.2   Introduction

Sustainable management and exploitation of marine fishes requires an understanding of population structure in order to assign stock units (Reiss et al., 2009), however, for many exploited species genetic population structure data is either lacking or not incorporated when designating management units (Casey et al., 2016). Traditionally, population structure is assessed using non-coding markers such as SNPs and/or Microsatellites (Brumfield et al., 2003; Narum et al., 2008; Vignal et al., 2002), which are assumed to be subject to neutral evolutionary processes (Helyar et al., 2011). In populations that have recently diverged, or show little structure at neutral markers, using markers under selection can provide useful discriminatory power between populations (Bernatchez et al., 2003; Consuegra et al., 2005), and thereby could provide a basis for stock designation and management decisions (Reiss et al., 2009). Because Major Histocompatibility Complex (MHC) genes are at the centre of the vertebrate adaptive immune response and therefore subjected to strong selective pressures (Sommer, 2005), they are some of the best candidate genes for designing non-neutral protein coding markers (Consuegra et al., 2005; Minias et al., 2019). However, assessing MHC diversity between populations is complex due to difficulties in assigning loci (Biedrzycka et al., 2020; Lighten et al., 2017), copy number variation and the presence of pseudogenes (Huang et al., 2019).

The MHC complex is a multigene family which codes for receptors that bind fragments of proteins (peptides) derived from pathogens and transport them to cell membrane surfaces for T cell recognition, which then triggers the immune response (Bernatchez et al., 2003). Most variability between MHC alleles is located at the Peptide Binding Regions (PBR), because the shape of the PBR dictates which pathogen motifs can be bound (Grimholt, 2016). There are several MHC genes directly involved with peptide binding and therefore there are multiple exons that could be targeted for population level analysis (Bernatchez et al., 2003). In fishes, the MHC complex usually (but not always, e.g. Gadoids (Malmstrøm et al., 2016)) comprises of two classical type of genes, class I and class II, which, in contrast to humans, are unlinked (Grimholt, 2016). Class I genes recognise peptides derived from the processing of intracellular pathogens and class II genes recognise those from extracellular processing (Minias et al., 2019). Class I tends to be more variable, containing a greater number of lineages (Grimholt, 2016), which renders it attractive for assessing inter-population variability but, because of large introns of multiple thousand base-pairs present in some lineages (Johannes Martinus Dijkstra et al., 2007; Grimholt, 2016), this variability comes at the cost of difficulty in aligning sequences and difficulty designing primers. In contrast, MHC class II molecules are less variable, and do not possess a large intron (Dijkstra, Grimholt, Leong, Koop, & Hashimoto, 2013; Grimholt, 2016) therefore having an advantage of being more easily amplified. The presence of conserved domains within the MHC II (Dijkstra et al., 2013) renders it more suitable for primer design. There are two candidate regions for targeted sequencing in the MHC class II, the α1 and ß1 domains which together form the peptide binding region (PBR). In many vertebrates, such as humans (Reche et al., 2003) and chickens (Kaufman

et al., 1999), the beta chain exhibits higher rates of positive selection and therefore variability, than the alpha chain (Gómez et al., 2010). Therefore, many population level studies focus on the beta chain (e.g. Dionne, Miller, Dodson, Caron, & Bernatchez, 2007; Fraser, Ramnarine, & Neff, 2009; Landry & Bernatchez, 2001; Lighten, van Oosterhout, Paterson, Mcmullan, & Bentzen, 2014; Talarico, Babik, Marta, & Mattoccia, 2019). However, in some fishes, such as Atlantic salmon, this pattern is not observed, and α1 diversity is augmented, potentially to compensate for limited number of class I loci, and therefore variation of alleles (Gómez et al., 2010). With the diminishing costs of high throughput (Next Generation) sequencing (Schwarze et al., 2020) sequencing and possibilities for extending read length (Besser et al., 2018) the possibilities for combining the sequencing of multiple domains within the same study could increase the discriminatory ability of MHC markers, and render them more powerful for population genetic studies.

In the clade (series) Eupercaria (a clade of bony, ray-finned fishes that include basses, perches and sticklebacks), some species are estimated to have an intermediate copy number of MHC class I alleles (e.g. Stickleback (Aeschlimann et al., 2003), sea bass (Chapter 3, Pinto, Randelli, Buonocore, Pereira, & dos Santos, 2013)) and could therefore be expected to follow the general vertebrate pattern of lower α1 domain diversity, remaining more conserved between species and exhibiting less potential for inter and intra-specific discrimination. However, if the α1 region does display evidence of positive selection, sequencing both regions would increase the discriminatory power of MHC class II markers, whilst retaining advantages such as conserved primer binding sites and lack of lengthy introns.

Here, we assess the potential class II alpha and beta genes as a marker for the clade (series) Eupercaria to ascertain whether focussing on the Beta or Alpha chain or both would have the greatest potential to differentiate between populations of species within this clade. To test this, I assessed differences in positive selection signals, trans species functional supertype structure, and time-tree divergence times, between the alpha and beta chain PBR domains using sequences from 18 species within the clade (series) Eupercaria. Within species variation was not assessed due to a lack of reliable sequence datasets.

5.3 Methods

Data compilation

MHC class II sequences were downloaded from the publicly available databases of the National Center for Biotechnology Information (NCBI, Bathesda, MD, USA). Initially, nucleotide sequences for the alpha and beta chains of sea bass (*Dicentrarchus labrax)* (Alpha DQ821106.1 Beta: AM113471.1) were chosen as templates to Blast-search (Morgulis et al., 2008) the NCBI nucleotide database which also incorporates annotated sequences from the Genome database. The search was limited to the clade (series) Eupercaria, requesting the maximum number of sequences (5000) per search and using default search parameters. To avoid unbalanced sample sizes, a maximum of 10 sequences per species were chosen and the same number of sequences (n= 95) were compiled for each domain (alpha and beta). Where more than 10 sequences per species were available, a subset of sequences were chosen by ensuring an even number of sequences from all available studies and then selecting sequences at random from each study.

Sequence quality filtering and alignment

Sequences were then aligned using multiple alignment using fast Fourier transform (MAFFT version 7.486, (Katoh et al., 2019)), with default settings. Sequences were then translated to codons and manually filtered to remove sequences that did not translate correctly. Both alignments (alpha and beta) were then trimmed to a length of 300bp, starting at the beginning of the α1 domain (Silva et al., 2007) and ß1 domain (Buonocore et al., 2007).

Recombination and selection inference

Recombination was characterised on codon alignments using Single Break Point (SBP) analysis (datamonkey.org), using '012223' model of nucleotide substitution. Positive and negative selection was then analysed using the Hyphy package (datamonkey.org), with the following models: Fast, Unconstrained Bayesian AppRoximation, (FUBAR, Murrell et al., 2013) Fixed Effects Likelihood, (FEL, Kosakovsky Pond and Frost, 2005) and MEME Mixed Effects Model of Evolution, (MEME, Pond et al, 2006). These models infer the strength of natural selection using the dN/dS metric (for full details see Chapter 3). Only sites selected by all 3 models were considered under positive selection or negative for downstream analysis (shared sites). If a positively selected site (PSS), coincided with a breakpoint, it was discarded from downstream analysis.

Phylogenetic tree analysis

Neighbor-joining (NJ) trees were constructed with all alleles in the database for both domains using 1000 bootstrap replicates and pairwise deletion for gaps/missing data, as used by Dijkstra et al., (2013).

The Time Tree of life (TTOL) time-tree was generated by uploading a custom species list to http://www.timetree.org/ (Hedges et al., 2015) for all species in the current study (with the exception of *Etheostoma spectabile* which was substituted for the closely related *Etheostoma caeruleum). Cyprinus carpio* was included as an outgroup.

Timetree.org uses a hierarchical average linkage method of estimating timings of speciation events synthesising molecular time-tree data from 2,274 publications representing 50,632 species to generate speciation times (Hedges et al., 2015). Briefly, this method searches for a given pair taxa in the Time-tree of life, then identifies the most recent common ancestor. If taxa are missing from the TTOL, NCBI taxonomy is scanned for the closest relative, which is then used as a proxy to find the most recent common ancestor and retrieve divergence times from the Time-tree datebase (Hedges et al., 2015). Trees are then constructed using maximum likelihood. This synthesis of published time-trees has identified clock-like change in speciation and diversification, indicating dominance of random processes (Hedges et al., 2015) and is therefore assumed to represent neutral evolution for comparisons to MHC alleles in this study.

The $\alpha 1$ and ß1 time-trees were created by choosing the most common allele for each species in the dataset where possible. Where common allele information was lacking, an allele was chosen at random. An allele from *Cyprinus carpio* (accessions: $\alpha 1$ JX466840, β1 XM_037113271) was used as an outgroup. The maximum likelihood method (default settings) in MEGAX following the protocol of Hall, (2013) was used in order to be comparable to maximum likelihood trees generated by http://www.timetree.org/. MHC allele time-trees were then generated as specified by Mello (2018). Divergence times for calibrating the time tree were generated using the 'Get Divergence Time for a Pair of Taxa' field within the web interface at http://www.timetree.org/, using the two most distantly related taxa in each domain ($\alpha 1$ and ß1) tree. These times were then used as a calibration constraint when generating time-trees using the RelTime method and default settings within MEGAX (Kumar et al., 2018).

Time-tree comparisons

Time-trees were compared in R (version 4.0.5). The function 'treedist' in the phangorn package was used to calculate 3 distance metrics between trees. Robinson-Foulds distance (RF distance) (Robinson et al., 1981) uses only the tree topology and is a count of the number of branches in a given tree which define a split (partition) absent from a second tree, plus the number of branches in the second tree which define a split absent from the first. Branch score difference incorporates branch length information, calculated as the sum of squares of the differences between each branch's length in the first and second

trees (Kuhner et al., 1994). Clade composition and branching times of clades in common between trees were compared between the original time-tree and the α1 and ß1 time-trees using the 'comparePhylo' function in the Ape package. Branching times for clades present in both trees were then compared statistically using a paired T test.

Supertype cluster analysis

The amino acids of the PSS were translated into a mathematical matrix (Doytchinova et al., 2005). For full methodology see methods section of Chapter 4, page 94. Gaps in the PSS, introduced by indels, were assigned '0's in the matrix (Kawashima & Kanehisa, 2000). Supertype clustering was carried out using the 'adegenet' package(R version 4.0.5). The find.clusters function was used to define the number of clusters for α1 and ß1 domains respectively (Jombart, 2012). BIC values for k were used to define the number of supertype clusters per domain (Jombart, 2008). The number of clusters was chosen by selecting the by lowest BIC value before an increase in values. Discriminant analysis of principal components (DACP), was conducted once the optimum number of principal components had been chosen using the 'optim.a.score' function. The number of supertypes was compared between α1 and β1 domains.

5.4   Results

Recombination and selection

Evidence of recombination was identified in both domains, with single break point at codon 74 (α1) and 81 (ß1). For the beta domain , this breakpoint coincided with a shared positively selected site, therefore, this site was removed from downstream PSS analysis.

There were similar numbers of positively selected sites identified within the α1 and ß1 domains by all three models used (FUBAR, FEL and MEME). While ß1 had a slightly greater number of sites identified by FUBAR and MEME, the α1 domain had a greater number identified by FEL. Both domains exhibited greater numbers of sites subjected to episodic selection as identified by MEME. In terms of sites considered under positive selection by all 3 models, results did not differ between domains. There were 19 shared positively selected sites identified in the α1 domain, and 20 in the β1(Figure 5.1, Table 5.1), however, because one site in the ß1 coincided with a breakpoint as assessed by SBP, they were considered to have an equal number of PSS (Figure 5.1, Table 1). There was a greater number of negatively selected sites (NSS) than PSS identified in both domains, and 10 more shared negatively selected sites in the ß1 domain than the α1 domain (Table 1).

Table 5.1 Positive and negative selection as assessed by 3 models: FUBAR Fast, Unconstrained Bayesian AppRoximation; FEL Fixed Effects Likelihood; MEME Mixed Effects Model of Evolution. Shared PSS = Positively Selected Sites under all 3 models. Shared NSS = Negatively Selected Sites under both models.

	Number of residues						
	Pervasive + selection		Episodic + selection		Negative selection		
Region	FUBAR	FEL	MEME	Shared PSS	FUBAR	FEL	Shared NSS
Alpha (α1)	20	23	39	19	25	23	21
Beta (β1)	26	21	46	20	35	33	30
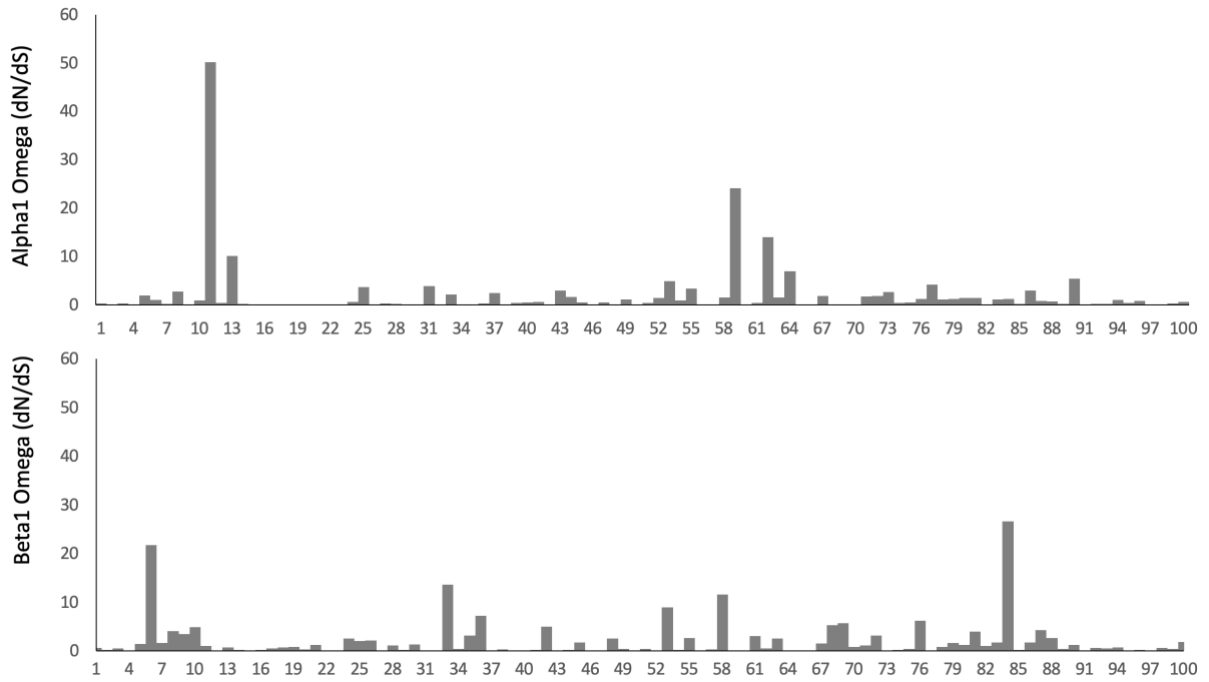
Figure 5.1 Omega values, which represent the dN/dS. Ratio (non-synonymous to synonymous substitutions) for each codon in the MHC class II alpha-1 and beta-1 alignments of 18 species belonging to the clade (series) Eupercaria

NJ trees and supertypes

The $\alpha 1$ and ß1 clustered into 6 and 5 groups respectively (Figure 5.2). Within the $\alpha 1$ NJ tree (based on entire allele sequence of 300bp), species largely grouped together with conspecifics or closely related species (Figure 5.3). Alleles from $\alpha 1$ supertype groups 1, 5 and 6 also clustered together in the $\alpha 1$ NJ tree, while groups 2 and 4 were split between locations on the tree. While ß1 alleles largely clustered by species, or closely related species, in the NJ tree, there were a few exceptions, for example L33965.1 *M. saxatilis* and XM 038731622.1 *M. salmoides* which did not cluster with conspecific alleles (Figure 5.4). Supertype groups were larger for the ß1 domain, with the exception of XM030413186.1 *S. aurata*, which was designated as a single allele supertype. The 4 major supertype groups were also more dispersed/intermixed within the NJ tree.
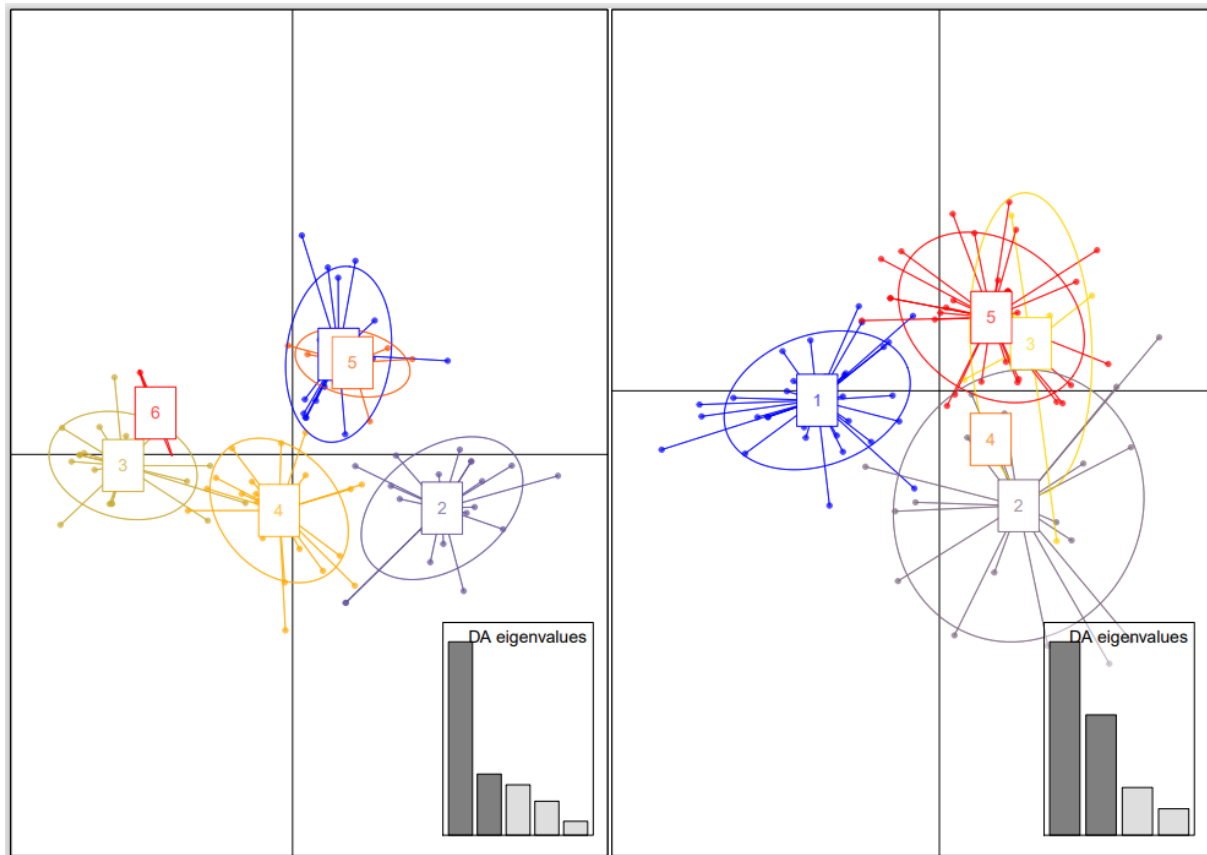
Figure 5.2 Clustering of MHC class II α1 (left) and ß1 (right) alleles in 18 species of the clade Eupercaria, as assessed using the physicochemical properties of translated amino acids of the PBR using discriminant analysis of principal components (DAPC). DAPC inferred the presence 6 alpha supertypes and 5 beta supertypes. Point represent the positioning of each MHC allele within the discriminant functions (see inset) and circles represent MHC supertypes.

Figure 5.3 Neighbor-Joining tree of 95 MHC class II α1 alleles from 18 species of the clade (series) Eupercaria. Each allele is labelled with NCBI accession number and species identity. Colours represent functional supertypes as assessed by DAPC analysis (Figure 5.2) Bootstrap values in percentage from 1000 trials are shown.

Figure 5.4 Neighbor-Joining tree of 95 MHC class II ß1 alleles from 18 species of the clade (series) Eupercaria. Each allele is labelled with NCBI accession number and species identity. Colours represent functional supertypes as assessed by DAPC analysis (Figure 5.2). Bootstrap values in percentage from 1000 trials are shown.

Time-tree analysis

The α1 and ß1 time-trees, based on the most common allele available, had 4 clades that differed between them and were the most similar in terms of RF distance and branch score difference (Table 5.2, Figure 5.5). However, these were largely organisational, for example in the case of the family Percidae, (represented by *Perca fluviatalis, Perca flavescens, Sander lucioperca* and *Etheostoma spectibile),* all 4 species were closely grouped in both trees, but resulted in slightly differing clades.

The MHC allele time-trees and the TTOL time-tree again exhibited organisational differences, that resulted in differing clades (Figure 5.5). While many species segregated in a similar way to α1 and ß1, differences here were more marked. For example, the family Moronidae (*D. labrax* and *M. saxatilis*) which clustered with Percidae species in both the α1 and ß1 time-trees, with an estimated diversion time of ~42 MYA α1 and ~41 MYA ß1, were more distantly related in the TTOL time-tree and had an estimated divergence time of ~110 MYA (pairwise estimate, timetree.org). In addition, pairwise divergence time estimates for *S. argus* and *M. salmoides* were ~110 MYA (timetree.org), whereas in the α1 and ß1 time-trees divergence times were estimated to be much more recent, at ~29 MYA (Figure 5.5). While the greatest distances in terms of RF and branch scores were observed between the Time-tree and the ß1 tree, branching times of common clades did not differ between α1 and ß1 trees, or ß1 and Time-tree (Table 5.2). Branching times did, however, differ significantly between α1 tree and Time-tree, indicating that potentially differences in alleles at the α1 region have arisen more recently than those based on the TTOL time-tree.
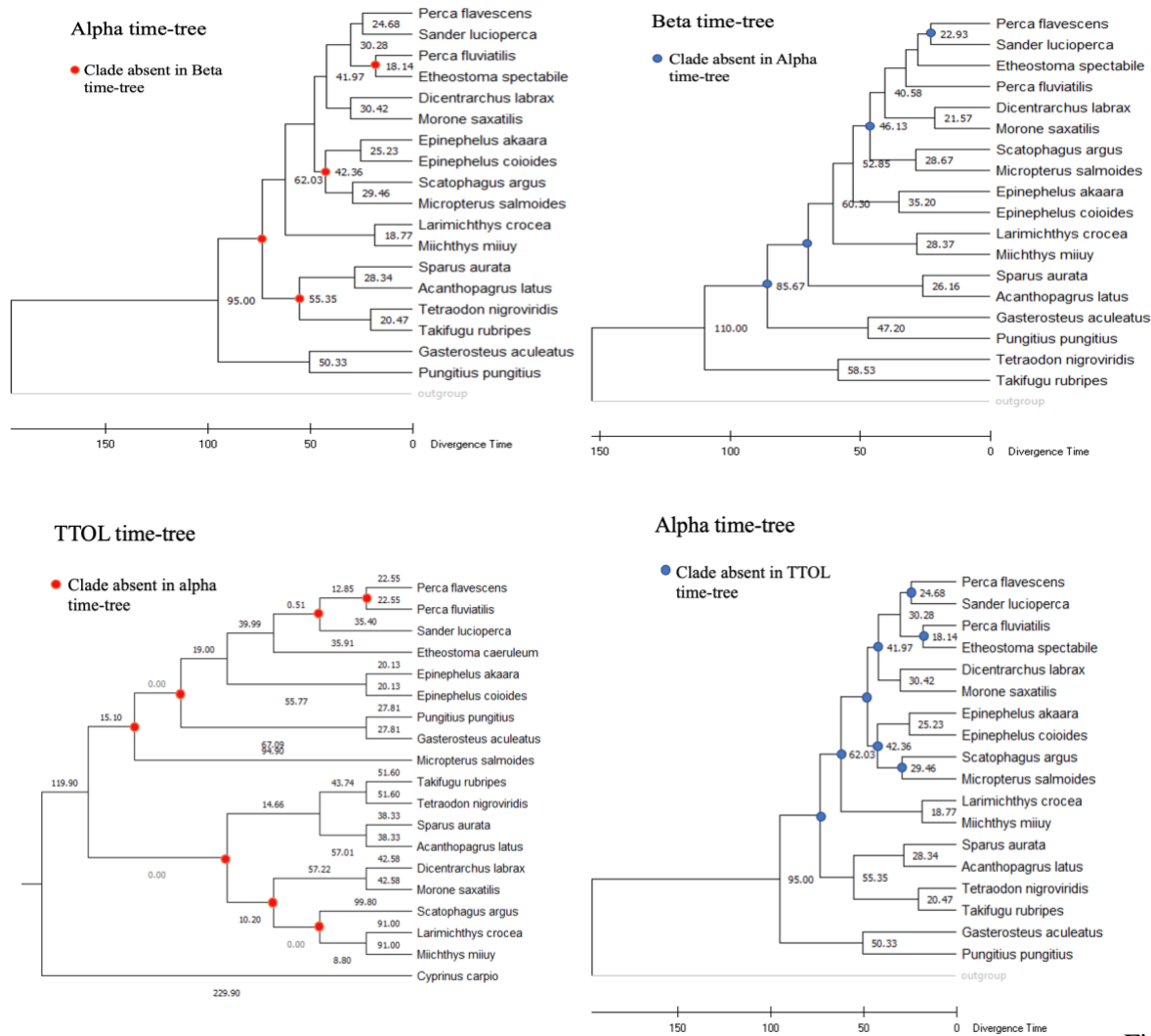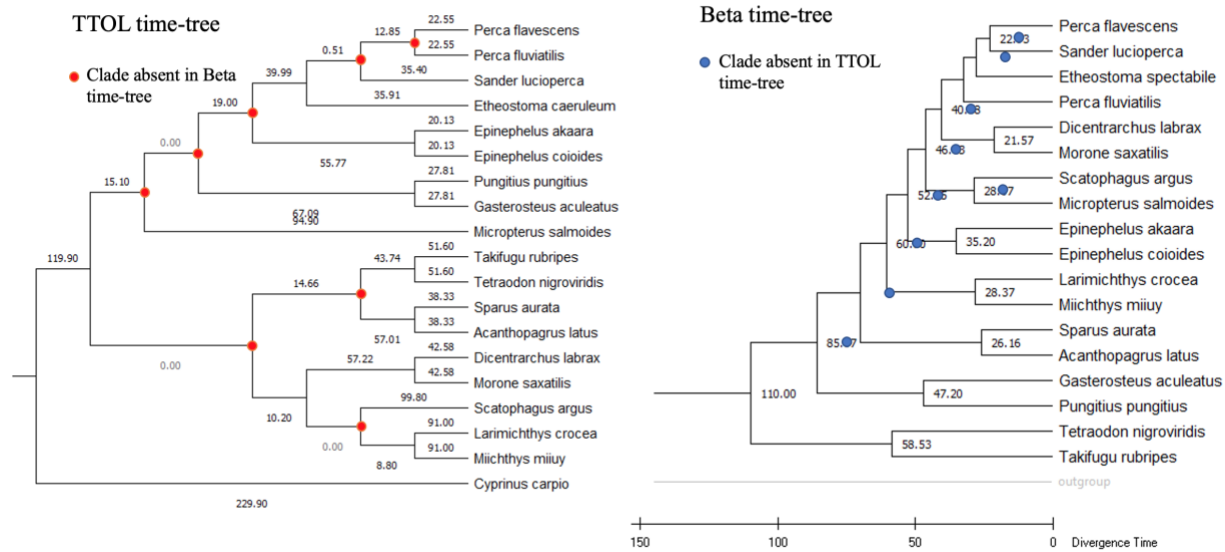
Figure 5.5

Figure 5.5. Time-trees constructed for the α1 and ß1 class II MHC domain for 18 species of the clade (series) Eupercaria. TTOL time-tree = time-tree as constructed by timetree.org. Red and Blue dots represent clades absent in the corresponding tree. Divergence times is in units of Million Years Ago (MYA).

Table 5.2. Summary of distance metrics to compare time-trees. No. diff. clades = number of different clades between two respective trees; RF symmetric diff. = Robinson Foulds distance between trees (Robinson et al., 1981); Branch score diff. = Branch score difference between trees (Steel et al., 1993). T tests compare branching times for all clades present in respective trees.

	Summary of differences			Difference in branching times of common clades		
	No. diff. clades	RF symmetric diff.	Branch score diff.	t	df	*P*
Alpha vs Beta	4	8	128.90	-0.27	13	0.79
Timetree vs Alpha	8	16	188.57	2.3	9	0.047*
Timetree vs Beta	9	18	227.33	1.33	8	0.22

5.5   *Discussion*

Selection and recombination

In contrast to other vertebrates, within the clade (series) Eupercaria, both the $\alpha 1$ and ß1 MHC class II domains exhibit similar levels of positive selection and recombination. Therefore, both domains have the potential to be used for population differentiation as well as increasing our understanding of evolution of the teleost MHC itself. In many vertebrates, such as humans the class II ß1 domain is considered to be far more polymorphic than the alpha loci (Grimholt, 2016), however, this study highlights that, for this clade of teleosts, almost identical levels of positive selection were observed. This pattern has previously been observed in salmonids, and, because they only possess one class I locus, it was proposed that the additional $\alpha 1$ variation could compensate for the apparent paucity of class I alleles (Gómez et al., 2010). In Eupercaria however, intermediate copy numbers, which are likely to confer a fitness advantage, have been observed at MHC I loci (Malmstrøm et al., 2016). Therefore, it is unlikely that the rates of $\alpha 1$ positive selection observed here are specifically to compensate for low class I allelic diversity. The alleles in this study are from a small subset of species of this clade (Hughes et al., 2018) and do not represent the whole diversity of alleles for those species. However, because inter-specific differentiation is larger than intra-specific differences, a pattern of reduced $\alpha 1$ diversity would have been observed despite the limitations of data availability. In addition, indels at 8 codons were present in the $\alpha 1$ but absent from the ß1 domains, indicating that substantial selective pressures have acted on the $\alpha 1$ domain (Minias et al., 2018). From this analysis, it is clear that both domains are involved in the adaptive response to pathogens in this clade, and because together they form the cleft class II PBR (Sato et al., 2012), the variability observed is likely to augment the number of pathogen motifs that can bind.

NJ trees and supertypes

NJ trees indicated that there could be greater inter-specific similarities in alleles at the ß1 domain than in the $\alpha 1$, as indicated by a small number of alleles not grouping with conspecific or closely related species alleles. This was not observed in the $\alpha 1$ NJ tree, where alleles grouped consistently with conspecific and closely related species. Both domains had a similar number of supertype groupings. In addition, the $\alpha 1$ domain exhibited clearer supertype definition (characterised by clearer DAPC clusters (Figure 5.2), which could indicate that alleles at this locus have diverged to a greater extent than at the ß1 domain.

The $\alpha 1$ supertypes clusters followed the structure of the NJ tree more closely than those of the ß1 domain, although no monophyletic (within the NJ tree) supertypes were observed at either domain. It has been proposed that balancing selection maintains functionally divergent MHC supertypes, while alleles within

these supertypes are subjected to a fast rate of renewal (Lighten et al., 2017). Contrasting allelic structure based on amino acids/nucleotide sequences with that obtained by functional supertype clustering, has previously been used to give an indication of the level of trans species polymorphism (or lack thereof) present at MHC domains (Ejsmond et al., 2018; Lighten et al., 2017). Trans species polymorphism occurs when multiple allelic clusters that originated in a common ancestral species are retained in descendant species, evidenced by monophyletic groups of alleles or supertypes present across multiple species (Ejsmond et al., 2018). With the caveats of limited data availability, this pattern was not observed at either region, as evidenced by the lack of monophyletic functional supertypes. In contrast to the theory of trans species polymorphism, this could potentially be the result of different allelic clades of a supertype evolving independently in different species and becoming fixed under strong selective pressures (Ejsmond et al., 2018). Whilst defining the mechanism underpinning the observed pattern is beyond the scope of this study, it is interesting to note that the α1 and ß1 exhibited differing patterns between supertypes and NJ trees and indicates that both domains should be considered when characterising the class II PBR.

Time-tree analysis

Divergence times between the two MHC II domains were also not significantly different, indicating that, for this subset of alleles, they are evolving at a similar rate. This tallies with the similar levels of positive and negative selection observed between the domains. The α1 time-tree however, had significantly different divergence times to those in the TTOL time-tree (Figure 5.5). It could be expected that, if the MHC domains were under strong levels of selection in contrast to neutral markers, then divergence times would be more recent for both domains, which together respond to pathogens. The ß1 and TTOL time-trees did not show any difference in divergence times, which contrasts with strong selective pressures observed in many vertebrates (Eizaguirre et al., 2012).

Conclusion

This study shoes that, contrary to expectations of higher levels of selection at the ß1 regions, both domains provide valuable insight into the selective pressures acting on the MHC class II PBR for the clade (series) Eupercaria. In addition, supertype and time-tree analysis were not congruent between the two domains, with indications that the α1 domain could have evolved at a different rate to that of the TTOL time-tree. It is therefore recommended that both domains be targeted when designing markers to assess inter and intraspecific differentiation within this clade.

5.6   References

Aeschlimann, P. B., Häberli, M. A., Reusch, T. B. H., Boehm, T., & Milinski, M. (2003). Female sticklebacks *Gasterosteus aculeatus* use self-reference to optimize MHC allele number during mate selection. *Behavioral Ecology and Sociobiology*, *54*(2), 119–126. http://doi.org/10.1007/s00265-003-0611-6

Bernatchez, L., & Landry, C. (2003). MHC studies in nonmodel vertebrates: What have we learned about natural selection in 15 years? *Journal of Evolutionary Biology*, *16*(3), 363–377. http://doi.org/10.1046/j.1420-9101.2003.00531.x

Besser, J., Carleton, H. A., Gerner-smidt, P., Lindsey, R. L., & Trees, E. (2018). NGS and their Application to the Study and Control of Bacterial Infections. *Clinical Microbiology and Infection*, *24*(4), 335–341. http://doi.org/10.1016/j.cmi.2017.10.013.Next-Generation

Biedrzycka, A., Konopiński, M., Hoffman, E., Trujillo, A., & Zalewski, A. (2020). Comparing raccoon major histocompatibility complex diversity in native and introduced ranges: Evidence for the importance of functional immune diversity for adaptation and survival in novel environments. *Evolutionary Applications*, *13*(4), 752–767. http://doi.org/10.1111/eva.12898

Brumfield, R. T., Beerli, P., Nickerson, D. A., & Edwards, S. V. (2003). The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology and Evolution*, *18*(5), 249–256. http://doi.org/10.1016/S0169-5347(03)00018-1

Buonocore, F., Randelli, E., Casani, D., Costantini, S., Facchiano, A., & Scapigliati, G. (2007). Molecular cloning , differential expression and 3D structural analysis of the MHC class-II b chain from sea bass (*Dicentrarchus labrax* L .), *23*. http://doi.org/10.1016/j.fsi.2007.03.013

Casey, J., Jardim, E., & Martinsohn, J. T. H. (2016). The role of genetics in fisheries management under the E.U. common fisheries policy. *Journal of Fish Biology*, *89*(6), 2755–2767. http://doi.org/10.1111/jfb.13151

Consuegra, S., Megens, H. J., Schaschl, H., Leon, K., Stet, R. J. M., & Jordan, W. C. (2005). Rapid evolution of the MH class I locus results in different allelic compositions in recently diverged populations of Atlantic salmon. *Molecular Biology and Evolution*, *22*(4), 1095–1106. http://doi.org/10.1093/molbev/msi096

Dijkstra, Johannes M., Grimholt, U., Leong, J., Koop, B. F., & Hashimoto, K. (2013). Comprehensive analysis of MHC class II genes in teleost fish genomes reveals dispensability of the peptide-loading

DM system in a large part of vertebrates. *BMC Evolutionary Biology*, *13*(1), 1–14. http://doi.org/10.1186/1471-2148-13-260

Dijkstra, Johannes Martinus, Katagiri, T., Hosomichi, K., Yanagiya, K., Inoko, H., Ototake, M., … Shiina, T. (2007). A third broad lineage of major histocompatibility complex (MHC) class I in teleost fish; MHC class II linkage and processed genes. *Immunogenetics*, *59*(4), 305–321. http://doi.org/10.1007/s00251-007-0198-6

Dionne, M., Miller, K. M., Dodson, J. J., Caron, F., & Bernatchez, L. (2007). Clinal variation in MHC diversity with temperature: Evidence for the role of host-pathogen interaction on local adaptation in Atlantic salmon. *Evolution*, *61*(9), 2154–2164. http://doi.org/10.1111/j.1558-5646.2007.00178.x

Doytchinova, I. A., & Flower, D. R. (2005). In Silico Identification of Supertypes for Class II MHCs. *The Journal of Immunology*, *174*(11), 7085–7095. http://doi.org/10.4049/jimmunol.174.11.7085

Eizaguirre, C., Lenz, T. L., Kalbe, M., & Milinski, M. (2012). Rapid and adaptive evolution of MHC genes under parasite selection in experimental vertebrate populations. *Nature Communications*, *3*, 1–6. http://doi.org/10.1038/ncomms1632

Ejsmond, M. J., Phillips, K. P., Babik, W., & Radwan, J. (2018). The role of MHC supertypes in promoting trans-species polymorphism remains an open question. *Nature Communications*, *9*(1), 1–4. http://doi.org/10.1038/s41467-018-06821-x

Fraser, B. A., Ramnarine, I. W., & Neff, B. D. (2009). Selection at the MHC class IIB locus across guppy (Poecilia reticulata) populations. *Heredity*, *104*(2), 155–167. http://doi.org/10.1038/hdy.2009.99

Gómez, D., Conejeros, P., Marshall, S. H., & Consuegra, S. (2010). MHC evolution in three salmonid species: A comparison between class II alpha and beta genes. *Immunogenetics*, *62*(8), 531–542. http://doi.org/10.1007/s00251-010-0456-x

Grimholt, U. (2016). MHC and evolution in teleosts. *Biology*, *5*(1). http://doi.org/10.3390/biology5010006

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution*, *30*(5), 1229–1235. http://doi.org/10.1093/molbev/mst012

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, *32*(4), 835–845. http://doi.org/10.1093/molbev/msv037

Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., … Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, *11*(SUPPL. 1), 123–136. http://doi.org/10.1111/j.1755-0998.2010.02943.x

Huang, K., Zhang, P., Dunn, D. W., Wang, T., Mi, R., & Li, B. (2019). Assigning alleles to different loci in amplifications of duplicated loci. *Molecular Ecology Resources*, *19*(5), 1240–1253. http://doi.org/10.1111/1755-0998.13036

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., … Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6249–6254. http://doi.org/10.1073/pnas.1719358115

Jombart, T. (2012). A tutorial for Discriminant Analysis of Principal Components (DAPC) using adegenet 1.3-4, 1–43.

Jombart, T., Lyon, D., & Biome, L. De. (2008). adegenet : a R package for the multivariate analysis of genetic markers, *24*(11), 1403–1405. http://doi.org/10.1093/bioinformatics/btn129

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service : multiple sequence alignment , interactive sequence choice and visualization, *20*(September 2017), 1160–1166. http://doi.org/10.1093/bib/bbx108

Kaufman, J., Milne, S., Göbel, T. W. F., Walker, B. A., Jacob, J. P., Auffray, C., … Beck, S. (1999). The chicken B locus is a minimal essential major histocompatibility complex. *Nature*, *401*(6756), 923–925. http://doi.org/10.1038/44856

Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, *11*(3), 459–468. http://doi.org/10.1093/oxfordjournals.molbev.a040126

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. http://doi.org/10.1093/molbev/msy096

Landry, C., & Bernatchez, L. (2001). Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic salmon

(Salmo salar). *Molecular Ecology*, *10*(10), 2525–2539. http://doi.org/10.1046/j.1365-294X.2001.01383.x

Lighten, J., Papadopulos, A. S. T., Mohammed, R. S., Ward, B. J., Paterson, G. I., Baillie, L., … Van Oosterhout, C. (2017). Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nature Communications*, *8*(1), 1–10. http://doi.org/10.1038/s41467-017-01183-2

Lighten, J., van Oosterhout, C., Paterson, I. G., Mcmullan, M., & Bentzen, P. (2014). Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (Poecilia reticulata). *Molecular Ecology Resources*, *14*(4), 753–767. http://doi.org/10.1111/1755-0998.12225

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., … Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), 1204–1210. http://doi.org/10.1038/ng.3645

Mello, B. (2018). Estimating timetrees with MEGA and the timetree resource. *Molecular Biology and Evolution*, *35*(9), 2334–2342. http://doi.org/10.1093/molbev/msy133

Minias, P., Pikus, E., & Anderwald, D. (2019). Allelic diversity and selection at the MHC class i and class II in a bottlenecked bird of prey, the White-tailed Eagle. *BMC Evolutionary Biology*, *19*(1), 1–13. http://doi.org/10.1186/s12862-018-1338-3

Minias, P., Pikus, E., Whittingham, L. A., & Dunn, P. O. (2018). A global analysis of selection at the avian MHC. http://doi.org/10.1111/evo.13490

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R., & Schäffer, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics*, *24*(16), 1757–1764. http://doi.org/10.1093/bioinformatics/btn322

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205. http://doi.org/10.1093/molbev/mst030

Narum, S. R., Banks, M., Beacham, T. D., Bellinger, M. R., Campbell, M. R., Dekoning, J., … Garza, J. C. (2008). Differentiating salmon populations at broad and fine geographical scales with microsatellites and single nucleotide polymorphisms. *Molecular Ecology*, *17*(15), 3464–3477. http://doi.org/10.1111/j.1365-294X.2008.03851.x

Pinto, R. D., Randelli, E., Buonocore, F., Pereira, P. J. B., & dos Santos, N. M. S. (2013). Molecular cloning and characterization of sea bass (Dicentrarchus labrax, L.) MHC class I heavy chain and β2-microglobulin. *Developmental & Comparative Immunology*, *39*(3), 234–254. http://doi.org/10.1016/j.dci.2012.10.002

Reche, P. A., & Reinherz, E. L. (2003). Sequence variability analysis of human class I and class II MHC molecules: Functional and structural correlates of amino acid polymorphisms. *Journal of Molecular Biology*, *331*(3), 623–641. http://doi.org/10.1016/S0022-2836(03)00750-2

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*(4), 361–395. http://doi.org/10.1111/j.1467-2979.2008.00324.x

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147. http://doi.org/10.1016/0025-5564(81)90043-2

Sato, A., Dongak, R., Hao, L., Shintani, S., & Sato, T. (2012). Organization of Mhc class II A and B genes in the tilapiine fish Oreochromis. *Immunogenetics*, *64*(9), 679–690. http://doi.org/10.1007/s00251-012-0618-0

Schwarze, K., Buchanan, J., Fermont, J. M., Dreau, H., Tilley, M. W., Taylor, J. M., … Wordsworth, S. (2020). The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine*, *22*(1), 85–94. http://doi.org/10.1038/s41436-019-0618-7

Silva, D. S. P., Reis, M. I. R., Nascimento, D. S., Vale, A. do, Pereira, P. J. B., & dos Santos, N. M. S. (2007). Sea bass (Dicentrarchus labrax) invariant chain and class II major histocompatibility complex: Sequencing and structural analysis using 3D homology modelling. *Molecular Immunology*, *44*(15), 3758–3776. http://doi.org/10.1016/j.molimm.2007.03.025

Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Frontiers in Zoology*, *2*, 1–18. http://doi.org/10.1186/1742-9994-2-16

Steel, M. A., & Penny, D. (1993). Distributions of tree comparison metrics some new results. *Systematic Biology*, *42*(2), 126–141. http://doi.org/10.1093/sysbio/42.2.126

Talarico, L., Babik, W., Marta, S., & Mattoccia, M. (2019). Genetic drift shaped MHC IIB diversity of an endangered anuran species within the Italian glacial refugium. *Journal of Zoology*, *307*(1), 61–70.

http://doi.org/10.1111/jzo.12617

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics Selection Evolution*, *34*(3), 275–305. http://doi.org/10.1051/gse:2002009

Yamaguchi, T., & Dijkstra, J. M. (2019). Major Histocompatibility Complex (MHC) Genes and Disease Resistance in Fish, 1–31.

Chapter 6   General discussion

6.1   General summary

This thesis as a whole set out to optimise molecular tools for monitoring and population analyses of the North Celtic and Irish sea fish populations. As an extensively fished ecosystem (Calderwood et al., 2020) that is vulnerable to climate change stressors (Hernvann et al., 2020), improved monitoring and population analysis is key for future sustainable fisheries management.

Fish larvae are particularly vulnerable to climate stressors (Radchuk et al., 2013), however difficulties with morphological identification render them challenging to monitor (Brechon *et al.*, 2013; Kimmerling *et al.*, 2018). Chapters 1 and 2 assessed the potential of metabarcoding as an alternative to traditional techniques for assessing fish assemblages in the numerous spawning grounds across the North Celtic and Irish seas. Chapter 1 addressed the need for improving the quantitative capabilities of bulk metabarcoding of larval organisms. Relative abundances could be improved with simple interventions (standardising the amount of tissue, using conserved primer sites) that are relatively fast to implement. Chapter 2 compared fish detections between bulk larval samples and water samples using metabarcoding. Good levels of agreement were found between water and larval samples and, for some taxa, water samples were more sensitive. With water samples alone however, it is not possible to detect the size class of fishes, therefore the comparison of larval fish to DNA in water samples adds to a growing body of research into the use of environmental DNA in spawning areas. Together, metabarcoding of both sample types demonstrated the potential of the technique for community level assessment of fish within this ecosystem, providing a potential pathway to update species distribution maps and monitor spawning. However, for management applications, species level identification is often required (Hernandez et al., 2013) and therefore this technique, independent of sample type, would either need to be supplemented with qPCR, or additional metabarcoding loci, for taxa that could not be resolved to species level with the primers used, and while this would add to the time and cost of sample processing, it is still likely to be more efficient and repeatable than traditional methods.

Understanding fine-scale population structure and local adaptation of over-fished species that show limited recovery when stringent management measures are implemented could result in more directed management practices to aid recovery. Chapters 3 and 4 explored the use of MHC markers for assessing population structure and local adaptation. Primers were successfully developed for the sea bass MHC class I gene, and private alleles were identified between the Celtic Shelf and Portuguese populations. Because the

identification of loci is particularly challenging in MHC genes (Biedrzycka et al., 2020; Lighten et al., 2017), conventional tests for population structuring using programs designed for diploid organisms, such as GENEPOP (Rousset et al., 2008) or Arlequin (Schneider et al., 2010), cannot be used (Huang et al., 2020). This limits the discriminatory power of these markers, as evidenced in Chapter 3, therefore while this marker showed potential, as demonstrated by assignment to populations based on allele frequencies, statistical power was hampered. Chapter 4 indicated that, for species of the clade (series) Eupercaria, MHC both the alpha and beta chains may be good candidates for assessing allele frequency and diversity between populations. When methodological advances allow, the identification of loci would enable further analyses of these loci. However, using a combination of multiple genes to capture allelic diversity and selection pressure across the peptide binding regions of both MHC classes as well as comparisons to neutral loci within a single study may also increase the utility of MHC markers for population discrimination.

6.2 *Incorporating this research into fisheries management.*

Molecular techniques have many advantages including repeatability, lack of observer bias and the ability to garner information from a huge range of sample types, from fish scales to stomach contents to water samples (Deagle et al., 2019; Taberlet et al., 2012; Ward, 2000). Insights that would not be possible to grasp using other methods, such as population structuring (Baltazar-Soares et al., 2018; Ward, 2000) and identification of damaged organisms (Brechon et al., 2013) are rendered feasible using these techniques. The advantages of incorporating these methods into fisheries management practices are great, however, although these techniques are not new, there are still many instances where stock designations contrast with biological separations, despite evidence based on genetic analyses (Kerr et al., 2017; Reiss et al., 2009). Marine species typically display low genetic population differentiation (Ward, Woodwark, & Skibinski, 1994), and studies are sometimes conflicting in their findings, depending on the markers used (Quéré et al., 2012; Reiss et al., 2009; Souche et al., 2015). Therefore managers are sometimes faced with findings that are complex and spatial stock units are not updated as a result (Kerr et al., 2017; Reiss et al., 2009). Markers such as the MHC could provide useful information for management purposes, however for this to be taken up, standardizing the level of genetic separation needed to be considered separate enough for stock differentiation is necessary and may require multiple markers and comparisons to neutral markers such as SNPs. In addition, with increases in capacity for genome sequencing, using genomic scanning techniques can increase the number of markers available from dozens to hundreds of thousands within a study (Baltazar-Soares et al., 2018). Using genomic scanning would enable comparisons between neutral and adaptive markers as well as integrate the quantification of adaptive genetic variation in marine fisheries monitoring (Baltazar-Soares et al., 2018; Carreras et al., 2017).

Techniques such as metabarcoding are yet to be routinely incorporated into management practices. This is in part due to the controversy over the potential for quantitative data from the technique (Deagle et al., 2019; Hansen et al., 2018; Lamb et al., 2019). While information neither traditional nor metabarcoding approaches provide true species richness and abundance information (Hansen et al., 2018), they are subjected to different biases. Metabarcoding biases include primer/template mismatches and mitochondrial copy number differences (Piñol et al., 2015). In water, additional biases exist such as the rate at which an organism sheds DNA, the rate at which that DNA degrades and where and how it is then transported (Collins et al., 2018; Deiner et al., 2014; Hansen et al., 2018). While studies demonstrate that water and bulk sample analysis perform equally well or better than capture/visual assessment based techniques (Thomas et al., 2016; Thomsen et al., 2016) it does not provide data on the number of fish present in a location, their size, weight or fecundity (Hansen et al., 2018). In marine systems, which are more biodiverse and hydrographically complex than freshwater, meta analyses indicate that detections similarities between traditional and environmental DNA (water sample) methods are more variable/sporadic but this limitation could be overcome by the use of multiple markers (McElroy et al., 2020). However, the majority of water metabarcoding studies have focused on freshwater environments (McElroy et al., 2020), and while water sample metabarcoding has had more focus than bulk, the number of bulk sample metabarcoding studies is rapidly increasing (van der Loos et al., 2020). As the number of marine based studies that compare these techniques to traditional assessments increases, so too does the potential for incorporating them into management decisions.

Based on this thesis, and the current literature, choice of circumstance is critical when deciding whether a technique could be used for management purposes. For instance, bulk larval metabarcoding has clear advantages for damaged organisms, while water metabarcoding has clear advantages for species that frequently evade capture such as sandeels. Therefore, these techniques initially could be used to supplement monitoring in areas where traditional techniques do not perform well or are prohibitive in terms of time and cost, rather than aim for a wholesale replacement of established methods.

6.3 Future directions

The sites in the North Celtic and Irish seas assessed by metabarcoding indicated that fish larvae community composition differed across the region, despite the number of sites (14) being small and only one time-point sampled. There are many larval distribution questions that bulk metabacoding could assist in addressing. In the study region, features such as the Irish sea front may be a barrier to larval transport (Lee et al., 2005) and further investigation, in the form of a more dense and targeting grid of sampling sites or

time-series sampling could address the front's impact on fish larvae distribution and community composition. In addition, because the technique can process many samples quickly, vertical stratification of larvae between mixed and stratified waters could be investigated. Sampling over multiple years would give an indication of how these sensitive life stages respond to shifts not only in temperature but also in hydrographic structuring and stratification.

While the metabarcoding primers used showed good quantitative potential, improvement in taxon assignment for some species would be needed. The 12S region is a compromise between conserved priming sites, which enable more even amplification and therefore more reliable quantitative signals and variability which allows for species level identification (Collins et al., 2019). Primers such as MiFish (Miya et al., 2015), while similar in terms of overall species level discrimination, may be able to separate different species to Riaz (2011) primers used here. Testing the quantitative capabilities of other 12S primers would be advantageous, so that the modifications of the protocol in this thesis could be applied more widely and to other regions with different target species. In addition, targeted surveys for sand eel assessment, along with ground-truthing of technique against existing stock assessment methods could improve existing monitoring for these taxa.

There is also the potential for markers in different areas of the mitochondrial and nucleic genome to be developed, which may allow for improved species discrimination and quantitative assessment. Sequencing methods are rapidly improving and hence long read and metagenomic approaches are likely to become more common place (Ye et al., 2019). This could vastly increase the potential for analysis of bulk samples, incorporating populations genetics and functional information such as the presence/absence of genes, structural variants and copy number (Lam et al., 2015).

The use of MHC markers for sea bass population structuring has shown potential, with only one marker and two populations with relatively small sample sizes being assessed. As indicated, for the clade (series) Eupercaria to which sea bass belong, assessing both a1 and b1 diversity would be beneficial, therefore combining markers for these 3 regions is likely to increase discriminatory power. Atlantic sea bass have a wide geographic range, from the Azores to Norway (Souche et al., 2015), and have already shown some differences in coding genes e.g. somatolactin insulin-like growth factor-1 genes between the Bay of Biscay and the North Sea (Quéré et al., 2010). Therefore, there is potential for further sea bass MHC markers for fish from different areas of Atlantic. MHC markers and genomic scans that focus on adaptive markers may be beneficial as early signs that local adaptation is occurring, showing differentiation before neutral markers, due to the speed at which changes can accumulate (Baltazar-Soares et al., 2018; Consuegra et al.,

2005). Monitoring MHC diversity over time or potentially using archival genetic material (Valenzuela-Quinonez, 2016) as well as current day sampling, whilst climatic parameters change in Eastern Atlantic regions might provide a warning sign for management purposes. In addition, assessing allelic diversity at these markers between farmed and wild sea bass or different lineages of farmed stocks, could provide insight into how MHC allelic structure relates to fitness in this species and how inbreeding changes allelic composition. As sequencing platforms develop, long reads and metagenomic approaches may enhance MHC analyses, enabling feasible identification of loci and the linking of allele composition to function. Transcriptomic approaches (Lowe et al., 2017) could also be applied to link MHC allele composition to fitness and enhance monitoring capacity.

6.4   Thesis contributions and conclusions

This thesis demonstrates the potential for using metabarcoding and MHC analysis for ecological and fisheries monitoring in the Celtic and Irish seas, with wider application for ecosystem monitoring in other regions. Improvements to the quantitative capacities of bulk metabarcoding enhanced possibilities to investigate the impact of climatic drivers on fish larvae. Comparisons of water and larval bulk fish metabarcoding indicated that, for some key species water sampling could be a valuable supplement to traditional spawning ground monitoring. Analysis of the MHC class 1-alpha region demonstrated the potential of this marker for incorporating data on fine population structuring needed to maintain their functional biodiversity, into management practices. Finally, clade level analysis of MHC class II markers demonstrated that, for species in Eupercaria, both membrane-spanning chains involved in peptide binding should be incorporated into selection and population analysis. Molecular techniques have the potential to supplement, and in some cases improve on, existing monitoring methods and as sequencing technologies develop so too will the range of situations in which their application is appropriate.

6.5   References

Baltazar-Soares, M., Hinrichsen, H. H., & Eizaguirre, C. (2018). Integrating population genomics and biophysical models towards evolutionary-based fisheries management. *ICES Journal of Marine Science*, *75*(4), 1245–1257. http://doi.org/10.1093/icesjms/fsx244

Biedrzycka, A., Konopiński, M., Hoffman, E., Trujillo, A., & Zalewski, A. (2020). Comparing raccoon major histocompatibility complex diversity in native and introduced ranges: Evidence for the importance of functional immune diversity for adaptation and survival in novel environments. *Evolutionary Applications*, *13*(4), 752–767. http://doi.org/10.1111/eva.12898

Brechon, A. L., Coombs, S. H., Sims, D. W., & Griffiths, A. M. (2013). Development of a rapid genetic technique for the identification of clupeid larvae in the Western English Channel and investigation of mislabelling in processed fish products. *ICES Journal of Marine Science*, *70*(2), 399–407. http://doi.org/10.1093/icesjms/fss178

Calderwood, J., Robert, M., Pawlowski, L., Vermard, Y., Radford, Z., Catchpole, T. L., & Reid, D. G. (2020). Hotspot mapping in the Celtic Sea: An interactive tool using multinational data to optimise fishing practices. *Marine Policy*, *116*(October 2019), 103511. http://doi.org/10.1016/j.marpol.2019.103511

Carreras, C., Ordóñez, V., Zane, L., Krusc, C., Nasto, I., Macpherson, E., & Pascual, M. (2017). Population genomics of an endemic Mediterranean fish : differentiation by fine scale dispersal and adaptation, (March), 1–12. http://doi.org/10.1038/srep43417

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., … Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 2041–210X.13276. http://doi.org/10.1111/2041-210X.13276

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, *1*(1). http://doi.org/10.1038/s42003-018-0192-6

Consuegra, S., Megens, H. J., Schaschl, H., Leon, K., Stet, R. J. M., & Jordan, W. C. (2005). Rapid evolution of the MH class I locus results in different allelic compositions in recently diverged populations of Atlantic salmon. *Molecular Biology and Evolution*, *22*(4), 1095–1106. http://doi.org/10.1093/molbev/msi096

Deagle, B. E., Thomas, A. C., McInnes, J. C., Clarke, L. J., Vesterinen, E. J., Clare, E. L., … Eveson, J. P.

(2019). Counting with DNA in metabarcoding studies: How should we convert sequence reads to dietary data? *Molecular Ecology*, *28*(2), 391–406. http://doi.org/10.1111/mec.14734

Deiner, K., & Altermatt, F. (2014). Transport distance of invertebrate environmental DNA in a natural river. *PLoS ONE*, *9*(2). http://doi.org/10.1371/journal.pone.0088786

Hansen, B. K., Bekkevold, D., Clausen, L. W., & Nielsen, E. E. (2018). The sceptical optimist: challenges and perspectives for the application of environmental DNA in marine fisheries. *Fish and Fisheries*, *19*(5), 751–768. http://doi.org/10.1111/faf.12286

Hernandez, F. J., Carassou, L., Graham, W. M., & Powers, S. P. (2013). Evaluation of the taxonomic sufficiency approach for ichthyoplankton community analysis. *Marine Ecology Progress Series*, *491*(Fahay 2007), 77–90. http://doi.org/10.3354/meps10475

Hernvann, P. Y., Gascuel, D., Grüss, A., Druon, J. N., Kopp, D., Perez, I., … Robert, M. (2020). The Celtic Sea Through Time and Space: Ecosystem Modeling to Unravel Fishing and Climate Change Impacts on Food-Web Structure and Dynamics. *Frontiers in Marine Science*, *7*. http://doi.org/10.3389/fmars.2020.578717

Huang, K., Dunn, D. W., Ritland, K., & Li, B. (2020). polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution*, *11*(3), 448–456. http://doi.org/10.1111/2041-210X.13338

Kerr, L. A., Hintzen, N. T., Cadrin, S. X., Clausen, L. W., Dickey-Collas, M., Goethel, D. R., … Nash, R. D. M. (2017). Lessons learned from practical approaches to reconcile mismatches between biological population structure and stock units of marine fish. *ICES Journal of Marine Science*, *74*(6), 1708–1722. http://doi.org/10.1093/icesjms/fsw188

Kimmerling, N., Zuqert, O., Amitai, G., Gurevich, T., Armoza-Zvuloni, R., Kolesnikov, I., … Sorek, R. (2018). Quantitative species-level ecology of reef fish larvae via metabarcoding. *Nature Ecology and Evolution*, *2*(2), 306–316. http://doi.org/10.1038/s41559-017-0413-2

Lam, K. N., Cheng, J., Engel, K., Neufeld, J. D., & Charles, T. C. (2015). Current and future resources for functional metagenomics. *Frontiers in Microbiology*, *6*(OCT), 1–8. http://doi.org/10.3389/fmicb.2015.01196

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430.

http://doi.org/10.1111/mec.14920

Lee, O., Nash, R. D. M., & Danilowicz, B. S. (2005). Small-scale spatio-temporal variability in ichthyoplankton and zooplankton distribution in relation to a tidal-mixing front in the Irish Sea. *ICES Journal of Marine Science*, *62*(6), 1021–1036. http://doi.org/10.1016/j.icesjms.2005.04.016

Lighten, J., Papadopulos, A. S. T., Mohammed, R. S., Ward, B. J., Paterson, G. I., Baillie, L., … Van Oosterhout, C. (2017). Evolutionary genetics of immunological supertypes reveals two faces of the Red Queen. *Nature Communications*, *8*(1), 1–10. http://doi.org/10.1038/s41467-017-01183-2

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5), 1–23. http://doi.org/10.1371/journal.pcbi.1005457

McElroy, M. E., Dressler, T. L., Titcomb, G. C., Wilson, E. A., Deiner, K., Dudley, T. L., … Jerde, C. L. (2020). Calibrating Environmental DNA Metabarcoding to Conventional Surveys for Measuring Fish Species Richness. *Frontiers in Ecology and Evolution*, *8*, 0–12. http://doi.org/10.3389/fevo.2020.00276

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., … Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088. http://doi.org/10.1098/rsos.150088

Piñol, J., Mir, G., Gomez-Polo, P., & Agustí, N. (2015). Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. *Molecular Ecology Resources*, *15*(4), 819–830. http://doi.org/10.1111/1755-0998.12355

Quéré, N., Desmarais, E., Tsigenopoulos, C. S., Belkhir, K., Bonhomme, F., & Guinand, B. (2012). Gene flow at major transitional areas in sea bass (Dicentrarchus labrax) and the possible emergence of a hybrid swarm. *Ecology and Evolution*, *2*(12), 3061–3078. http://doi.org/10.1002/ece3.406

Quéré, N., Guinand, B., Kuhl, H., Reinhardt, R., Bonhomme, F., & Desmarais, E. (2010). Genomic sequences and genetic differentiation at associated tandem repeat markers in growth hormone , somatolactin and insulin-like growth factor-1 genes of the sea bass , Dicentrarchus labrax, *296*, 285–296.

Radchuk, V., Turlure, C., & Schtickzelle, N. (2013). Each life stage matters: The importance of assessing the response to climate change over the complete life cycle in butterflies. *Journal of Animal Ecology*,

82(1), 275–285. http://doi.org/10.1111/j.1365-2656.2012.02029.x

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*(4), 361–395. http://doi.org/10.1111/j.1467-2979.2008.00324.x

Rousset, F., De, S., Montpellier, U., & Bataillon, P. E. (2008). GENEPOP ' 007 : a complete re-implementation of the GENEPOP software for Windows and Linux, 103–106. http://doi.org/10.1111/j.1471-8286.2007.01931.x

Schneider, S., Excoffier, L., & Laval, G. (2010). Arlequin suite ver 3 . 5 : a new series of programs to perform population genetics analyses under Linux and Windows, (January 2017). http://doi.org/10.1111/j.1755-0998.2010.02847.x

Souche, E. L., Hellemans, B., Babbucci, M., MacAoidh, E., Guinand, B., Bargelloni, L., … Volckaert, F. A. M. (2015). Range-wide population structure of European sea bass *Dicentrarchus labrax*. *Biological Journal of the Linnean Society*, *116*(1), 86–105. http://doi.org/10.1111/bij.12572

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. http://doi.org/10.1111/j.1365-294X.2012.05470.x

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726. http://doi.org/10.1111/1755-0998.12490

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes. *PLOS ONE*, *11*(11), e0165252. http://doi.org/10.1371/journal.pone.0165252

Valenzuela-Quiñonez, F. (2016). How fisheries management can benefit from genomics?. Briefings in Functional Genomics, 15(5), 352-357.http://doi.org/10.1093/bfgp/elw006

van der Loos, L. M., & Nijland, R. (2020). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, (July 2020), 3270–3288. http://doi.org/10.1111/mec.15592

Ward, R. D. (2000). Genetics in fisheries management. *Hydrobiologia*, *420*(1), 191–201.

http://doi.org/10.1023/A:1003928327503

Ward, R. D., Woodwark, M., & Skibinski, D. O. F. (1994). A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. Journal of fish biology, 44(2), 213-232.

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. http://doi.org/10.1016/j.cell.2019.07.010.

.

Appendix 1: Supplementary material Chapter 2

Table A1.1 Individuals barcoded with 12S V5 primers (Riaz et al., 2011 and CO1 primers F1, R1 (Ward et al., 2005), * indicates taxa for which the 12S barcode was added to the reference database, once lowest possible taxonomic level identification had been achieved using the CO1 barcode

12S Taxon assignment	COI taxon assignment
Ammodytidae (unresolved)	*Ammodytes marinus**
Clupea harengus/Sprattus sprattus	*Sprattus sprattus*
Trisopterus minutus	
Pleuronectidae (unresolved)	*Limanda limanda**
Ciliata mustela	
Trisopterus esmarkii	
Trisopterus minutus	
Labrus merula/Labrus bergylta	*Labrus bergylta*
Ammodytidae (unresolved)	*Gymnammodytes sp**
Callionymus sp	*Callionymus sp**
Pleuronectidae (unresolved)	*Microstomus kitt**
Buenia affinis	*Buenia jeffreysii**
Lepidorhombus sp	

Merluccius merluccius

Merlangius merlangus   *Merlangius merlangus*

Lepidorhombus sp

Table A1.2. Number of individual larvae captured in each haul in the survey.

Haul number	Total individuals in haul
1	1
2	63
3	59
4	53
5	27
6	1
7	58
8	32
9	8
10	0
11	0

12	24
13	6
14	0

Table A1.3. Overview of ichthyoplankton taxonomic assignment. Detection (presence/absence): families assigned by morphology alone, morphology corrected by a subsample of Sanger CO1 sequencing, and 12S metabarcoding. 'x' indicates where a method achieved lowest taxonomic classification. NA indicates where a taxon was unidentified by a method. Abundance: the total number of individuals detected by Sanger updated morphology and total number of bulk reads post filtering. * denotes taxa that were identified morphologically only.

Taxonomic classification		Detection			Abundance	
Family	Lowest classification	Morphology alone	Sanger Corrected morphology	Metabarcoding	Sanger Corrected morphology	No. reads (post filtering)
f__Ammodytidae	f__Ammodytidae	Ammodytidae	x	x	5	24429
	s__*Ammodytes_marinus*		x	x	26	247248
	g__Gymnammodytes		x	x		90543
f__Callionymidae	g__Callionymus	Callionymidae	x	x	62	586077
f__Clupeidae	x__C. harengus_ S.sprattus	Cluepeidae	Clupeidae	*C. harengus/S. sprattus*	128	427286
	s__*Sardina_pilchardus*	*NA*	*NA*	x	0	8784

f__Gadidae	s__*Merlangius_merlangus*	x	x	*Pollaccius* *sp* */M.* *merlangus*	9	0
	s__*Micromesistius_poutassou**	x	x	*NA*	5	0
	x__*P.pollachius/virens_M.* *merlangus*	*NA*	*NA*	*Pollaccius* *sp* */M.* *merlangus*	0	131238
	s__*Pollachius_pollachius*	x	x	*Pollaccius* *sp* */M.* *merlangus*	2	0
	s__*Pollachius_virens*	x	x	*Pollaccius* *sp* */M.* *merlangus*	1	0
	g__Tricopterus	x	*NA*	*NA*	0	0
	s__*Trisopterus_esmarkii*	x	x	x	6	93440
	s__*Trisopterus_minutus*	x	x	x	21	122247
f__Lotidae	s__*Ciliata_mustela*	*NA*	x	x	5	183976
f__Gobiidae	f__Gobiidae	Gobiidae	*NA*	x	0	3168
	s__*Buenia_jeffreysii*	*NA*	x	x	1	8306

	s__*Aphia_minuta**	*Aphia minuta*	*NA*	*NA*	0	0
f__Labridae	s__*Labrus_bergylta*	*NA*	x	x	3	20578
	s__*Labrus_mixtus*	*NA*	*NA*	x	0	5172
f__Lotidae	s__*Molva_molva*	*NA*	*NA*	x	0	14595
f__Merlucciidae	s__*Merluccius_merluccius*	x	x	x	2	104615
f__Mugilidae	f__*Mugilidae**	x	*NA*	*NA*	0	0
f__Pleuronectidae	f__*Pleuronectidae*	x	x	x	0	117330
	s__*Glyptocephalus_cynoglossus**	x	x	Pleuronectidae	12	0
	s__*Limanda_limanda*	x	x	x	13	129542
	s__*Microstomus_kitt*	x	x	x	11	142793
f__Scophthalmidae	g__*Lepidorhombus_sp*	*NA*	x	x	3	26632
f__Solidae	f__*Solidae**	x	*NA*	*NA*	0	0
f__Triglidae	f__Triglidae	*NA*	x	x	9	28714

damaged	damaged	*NA*	x	0	8	0
unknown	unknown	x	*NA*	*NA*	0	0

Table A1.4. SIMPER analysis showing average abundances (Av.Abund), average dissimilarity between locations (Av.Diss), the contribution dissimilarity between locations (Contrib%) and the cumulative contributions dissimilarity (Cum.%) for each of the 7 taxa which contribute the most to between group dissimilarity between locations 2 and 3, using morphology (CPUF) and metabarcoding (RPUF). s__ denotes species level classification, g__ genus, f__ family and x__ two possible species (synonymous sequences).

Lowest possible taxonomic level

Taxon	Location 2 Av.Abund	Location 3 Av.Abund	Av.Diss	Contrib%	Cum.%
CPUF					
x__*C. harengus/S. sprattus*	0.23	0.02	18.76	21.42	21.42
f__Triglidae	0.05	0.02	12.58	14.37	35.78
g__Callionymus	0.11	0.11	11.04	12.61	48.39
s__*Microstomus_kitt*	0.06	0.05	5.91	6.75	55.15
s__*Trisopterus_minutus*	0.05	0.03	5.45	6.22	61.37
s__*Ciliata_mustela*	0.06	0.00	4.81	5.49	66.86
s__*Merlangius _merlangus*	0.02	0.06	4.63	5.29	72.14
RPUF					
x__*C. harengus_S. sprattus*	0.16	0.02	13.63	15.74	15.74
f__Triglidae	0.04	0.02	12.5	14.43	30.17

g__Callionymus	0.11	0.11	10.96	12.65	42.82
s__*Ciliata_mustela*	0.09	0.00	7.05	8.14	50.95
s__*Trisopterus_minutus*	0.07	0.03	5.94	6.85	57.8
s__*Microstomus_kitt*	0.05	0.05	5.7	6.58	64.38
s__*Limanda_limanda*	0.05	0.03	5.63	6.5	70.88

Family level

	Location 2	Location 3			
Family	Av.Abund	Av.Abund	Av.Diss	Contrib%	Cum.%
CPUF					
f__Clupeidae	0.17	0.02	13.89	16.85	16.85
f__Triglidae	0.04	0.02	12.58	15.25	32.1
f__Callionymidae	0.11	0.11	11.63	14.11	46.21
f__Gadidae	0.13	0.07	11.02	13.36	59.57
f__Pleuronectidae	0.12	0.07	10.6	12.85	72.42
RPUF					
f__Clupeidae	0.19	0.02	21.45	25.64	25.64
f__Triglidae	0.04	0.01	13.43	16.05	41.69
f__Callionymidae	0.11	0.06	12.59	15.05	56.74
f__Gadidae	0.10	0.05	10.77	12.87	69.61

f__Pleuronectidae	0.09	0.04	10.66	12.74	82.35
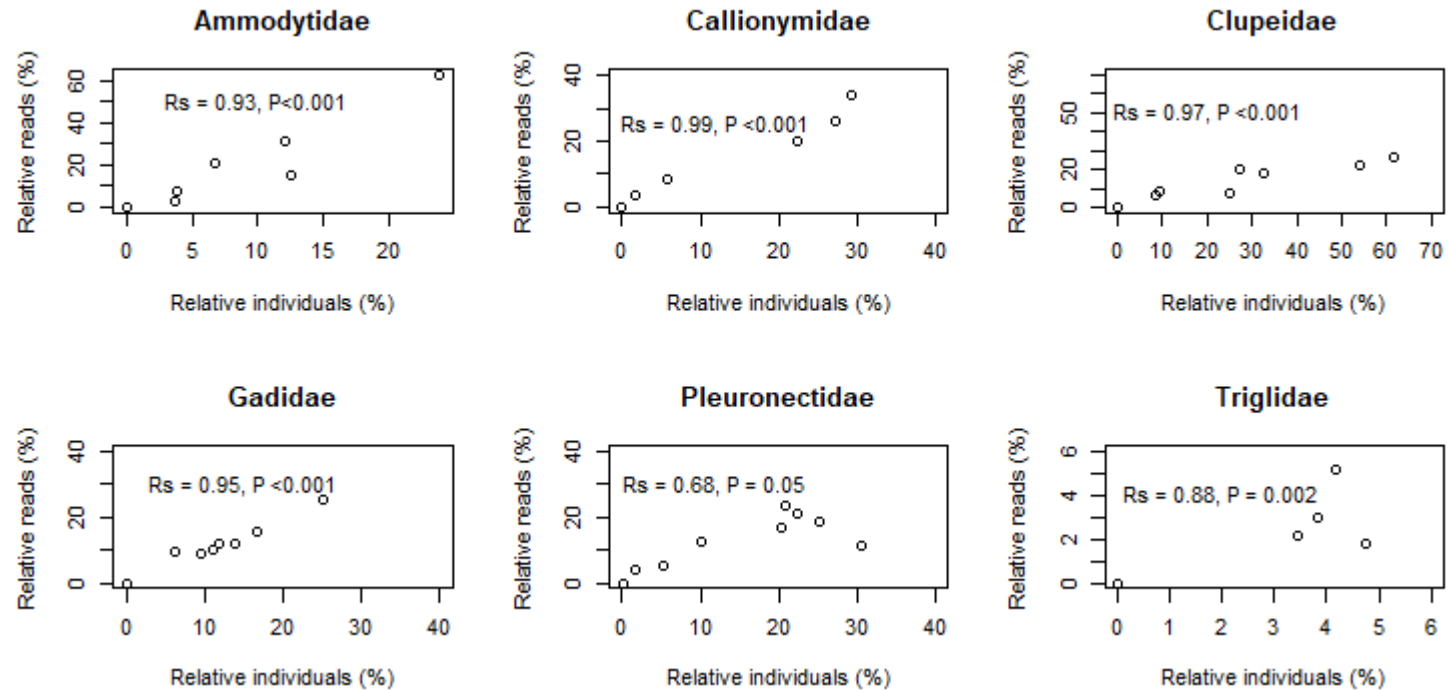
Figure A1.1 Relationship between relative number of individuals (%) within a taxonomic family in each haul, and relative number of reads post filtering (%) in the corresponding sample. Family level Spearman's Rho correlation were calculated across all hauls in the survey.
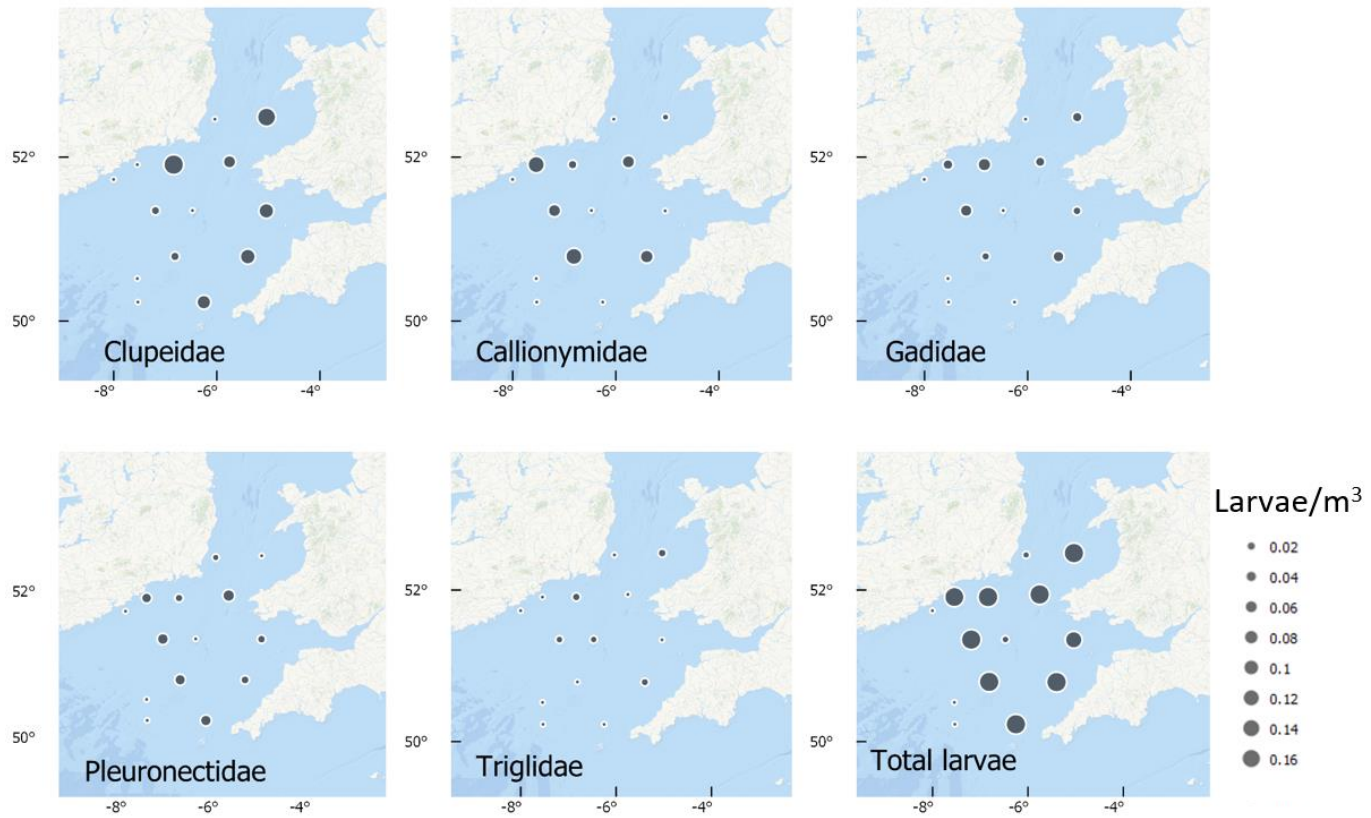
Figure A1.2. Abundances (number individuals) of families per M^3 of water filtered within survey, assessed by morphology.
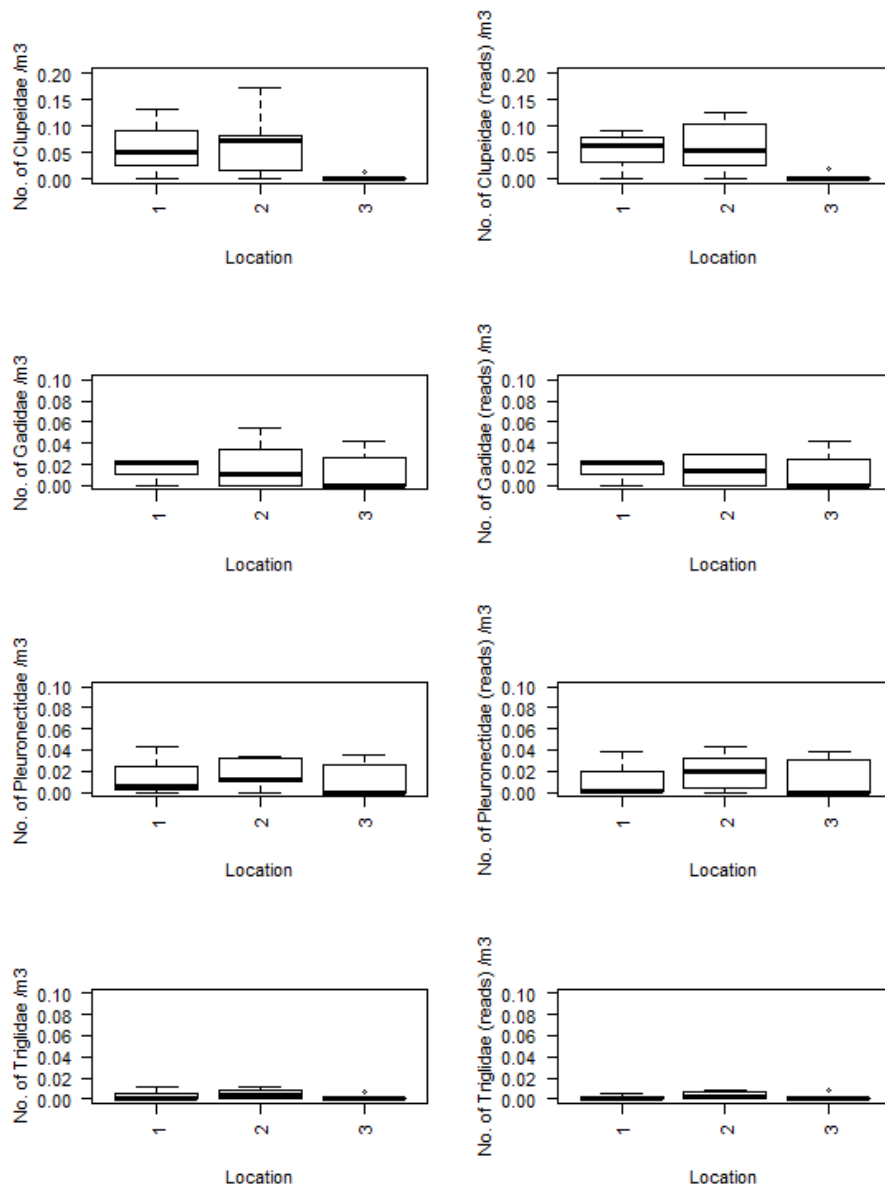
Figure A1.3. Abundance of individuals of a given family per m3 in each of the 3 locations, based on morphology (left, CPUF) and back-estimated reads (right, RPUF).

Appendix 2: Supplementary material Chapter 3

Figure A2.1. PCoA of PERMDISP showing dispersion of community composition (Bray Curtis matrix) for 1. Larvae samples, 2. Water samples.

Figure A2.2. Metabarcoding relative read abundance in two mock communities. Mock 1 (M1) was constructed using equal masses of Sanger-barcoded genomic DNA. Mock 2 (M2) was constructed from varying concentrations of genomic DNA. M1 Platinum/M2 platinum show the relative abundance of outputted reads using Platinum taq polymerase and nested PCR approach and M1 Phusion/M2 Phusion show the relative abundance of outputted reads using Phusion taq polymerase, without a nested PCR. Taxa were identified to the lowest possible taxonomic level. s__ = species level, g__= genus level.

Table A2.1. Metadata for each site in the survey. Btm depth (m)= bottom depth at site, Temp 1m/15m/btm = temperature (°C) at depth, Salinity 1m/15m/btm = salinity (PSU) at depth. Density 1m/15m/btm = density (kg/m^3).

Site name	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Haul number	1	4	12	15	21	25	34	38	45	52	55	61	64	66
Vol. Filtered (MultiNet)	200	256	322	186	252	160	231	186	30	45	33	142	39	42
Btm depth (m)	56	82	105	70	68	123	74	102	86	107	107	90	54	66
Temp 1m	9.865	9.789	10.013	12.051	12.378	12.329	12.622	12.603	11.848	13.508	13.508	13.389	13.025	13.971
Salinity 1m	34.5	34.685	34.767	34.885	34.975	34.915	35.562	35.393	35.157	35.481	35.481	34.88	34.646	34.679
Density 1m	1026.59	1026.75	1026.77	1026.49	1026.5	1026.46	1026.91	1026.78	1026.75	1026.67	1026.67	1026.23	1026.12	1025.95
Temp 15m	9.569	9.552	9.633	11.318	10.25	11.437	11.029	11.414	11.745	12.634	12.634	12.774	12.082	13.064
Salinity 15m	34.546	34.703	34.831	35.001	35.31	35.047	35.47	35.38	35.27	35.415	35.415	34.955	34.847	34.544
Density_15m	1026.74	1026.86	1026.95	1026.79	1027.22	1026.8	1027.2	1027.06	1026.92	1026.86	1026.86	1026.47	1026.52	1026.09
Temp btm	9.506	9.496	9.589	9.354	10.206	8.915	10.988	9.271	10.727	9.993	9.993	9.238	9.049	8.707

Salinity btm	34.551	34.708	34.844	35.262	35.296	35.242	35.463	35.454	35.335	35.367	35.367	35.262	35.19	35.164
Density btm	1026.91	1027.03	1027.13	1027.49	1027.37	1027.55	1027.36	1027.66	1027.31	1027.49	1027.49	1027.53	1027.44	1027.53

Table A2.2: Spawning times of taxa encountered in samples for this study in the for the Irish/Celtic Sea (where information is available). Grey fill indicates spawning season, * indicates peak spawning. Sampling for this study was conducted in May (17th to 26th, 2018, outlined). Adapted from Ellis et al., 2012.

Taxa	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Reference
Ammodytidae	▓	▓	▓								▓	▓	Coull et al., 1998
Callionymidae	▓	▓	▓	▓	▓	▓							Nichols et al., 1993
Clupea harengus (SW Ireland)	▓	▓	▓	▓	▓								Coull et al., 1998
Merlangius merlangus		▓	▓	▓	▓	▓							Coull et al., 1998
Micromesistius poutassou			▓	*	*	▓							Coull et al., 1998
Molva molva			▓	▓	▓	▓							Genner et al., 2010
Merluccius merluccius	▓	*	*	▓	▓	▓							Korta et al., 2010; Domínguez-Petit et al., 2010;
Scomber scombrus			▓	▓	*	*	▓						Coull et al., 1998
Solea solea				*	▓	▓							Coull et al., 1998
Sprattus sprattus				▓	*	▓							Coombs et al., 1992

Table A2.3. Position of sites in the survey (Decimal degrees).

Site	Latitude	Longitude
1	52.45083333	-6.038611111
2	52.47444444	-5.036388889
3	51.94333333	-5.755277778
4	51.91	-6.837777778
5	51.35305556	-5.040833333
6	51.35777778	-6.473611111
7	50.79611111	-5.400277778
8	50.79833333	-6.813333333
9	50.23416667	-6.252777778
10	50.235	-7.533055556
11	50.52666667	-7.543333333
12	51.35638889	-7.190833333
13	51.91	-7.544722222
14	51.73277778	-7.003888889

Table A2.4: Spearman's rank correlations of site by site relative abundance of eDNA samples and bulk samples.

Taxon	S	P	rho
Ammodytes marinus	497.73	0.22	-0.37
C. harengus/S. sprattus	76.44	0.22	0.79
Callionymus sp	423.1	0.6	-0.16
Ciliata mustela	296.92	0.55	0.18
Limanda limanda	76.44	0	0.79
Merluccius merluccius	394.33	0.79	-0.08
P. pollachius/ M. merlangus	389.69	0.82	-0.07
Trisopterus esmarkii	226.8	0.21	0.38
Trisopterus minutus	353.63	0.93	0.03

Table A2.5. SIMPER analysis showing average abundances (mean larvae, mean water), between larvae and water samples and the cumulative contributions to dissimilarity between sample types (% cumulative contribution).

	Mean larvae	Mean water	% contribution	Cumulative
Ammodytes marinus	8.07	36.07	23.74	
C. harengus / S. sprattus	12.55	36.14	46.21	
Callionymus sp	15.91	3.65	56.82	
Limanda limanda	11.59	0.30	63.84	
Scomber scombrus	0.00	9.85	70.64	
Triglidae	8.62	1.65	76.7	
Ciliata mustela	3.69	3.20	80.7	
P. pollachius / M. merlangus	4.00	1.98	83.77	
Trisopterus minutus	3.79	1.39	86.48	
Microstomus kitt	3.67	0.00	88.68	

References:

Coombs, S. H., Nichols, J. H., Conway, D. V. P., Milligan, S., & Halliday, N. C. (1992). Food availability for sprat larvae in the Irish Sea. Journal of the Marine Biological Association of the United Kingdom, 72(4), 821-834. a1). https://doi.org/10.1017/S0025315400060069

Coull, K.A., Johnstone, R., and Rogers, S.I. 1998. Fisheries Sensitivity Maps in British Waters. Published and distributed by UKOOA Ltd., v + 58 pp.

Domínguez-Petit, R., Saborido-Rey, F. and Medina, I. 2010. Changes of proximate composition, energy storage and condition of European hake (Merluccius merluccius, L. 1758) through the spawning season. Fisheries Research, 104: 73–82. https://doi.org/10.1016/j.fishres.2009.05.016

Ellis, J. R., Milligan, S. P., Readdy, L., Taylor, N., & Brown, M. J. (2012). Spawning and nursery grounds of selected fish species in UK waters. Retrieved from https://www.cefas.co.uk/publications/techrep/TechRep147.pdf

Genner, M.J., Halliday, N.C., Simpson, S.D., Southward, A.J., Hawkins, S.J. and Sims, D.W. 2010. Temperature- driven phenological changes within a marine larval fish assemblage. Journal of Plankton Research, 32: 699–708. https://doi.org/10.1093/plankt/fbp082

Korta, M., Domínguez-Petit, R., Murua, H. and Saborido-Rey, F. 2010. Regional variability in reproductive traits of European hake Merluccius merluccius L. populations. Fisheries Research, 104: 64–72. DOI: 10.1016/j.fishres.2009.03.007

Nichols, J.H., Haynes, G.M., Fox, C.J., Milligan, S.P., Brander, K.M. and Chapman, R.J. 1993. Spring plankton surveys of the Irish Sea in 1982, 1985, 1987, 1988 and 1989: hydrography and the distribution of fish eggs and larvae. Fisheries Research Technical Report, Ministry of Agriculture, Fisheries and Food, 95: 111 pp

Appendix 3: Supplementary material Chapter 4

Supporting information S3

MHC class I-alpha can reveal cryptic fine-scale population structure in a commercial fish, the European sea bass (*Dicentrarchus labrax*)

Table A3.1**.** Clustering methods for *D. labrax* MHC class 1a alleles, detailing numbers of clusters identified and whether they conformed to the expectation of only 2 or less alleles from a cluster being observed in all individuals in the study. For visual representations of each method, see Figures S2-S7.

Method	Number of Clusters Identified	2 or more alleles per fish?
Neighbour Joining Tree	9	Y
NeighborNet	11	Y
Ward (Euclidean distance)	10	Y
UPGMA (Euclidean distance)	18	Y
UPGMA (Correlation)	18	Y
UPGMA (Cosine)	18	Y

Table A3.2. Comparison of clustering based on positively selected sites (PSS) as identified by FEL, MEME and FUBAR, which were subsequently translated to a mathematical matrix of Z-scores. The matrix was then clustered based on 2 algorithms and 3 distances.

Allele	Ward Euclidian	UPGMA Euclidian	UPGMA cosine	UPGMA correlation
SBmhc1_27	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_39	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_53	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_60	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_70	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_80	Cluster1	Cluster 13	Cluster 14	Cluster 15
SBmhc1_41	Cluster1	Cluster 15	Cluster 15	Cluster 15
SBmhc1_45	Cluster1	Cluster 15	Cluster 15	Cluster 15
SBmhc1_87	Cluster1	Cluster 15	Cluster 15	Cluster 15
SBmhc1_107	Cluster1	Cluster 9	Cluster 12	Cluster 13
SBmhc1_96	Cluster1	Cluster 9	Cluster 14	Cluster 15
SBmhc1_56	Cluster1	Cluster 9	Cluster 12	Cluster 13
SBmhc1_127	Cluster1	Cluster 9	Cluster 11	Cluster 13
SBmhc1_29	Cluster1	Cluster 9	Cluster 11	Cluster 13
SBmhc1_11	Cluster 2	Cluster 14	Cluster 13	Cluster 14
SBmhc1_125	Cluster 2	Cluster 14	Cluster 13	Cluster 14

SBmhc1_23	Cluster 2	Cluster 14	Cluster 13	Cluster 14
SBmhc1_30	Cluster 2	Cluster 14	Cluster 13	Cluster 14
SBmhc1_89	Cluster 2	Cluster 14	Cluster 13	Cluster 14
SBmhc1_121	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_131	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_19	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_44	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_61	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_84	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_97	Cluster 2	Cluster 16	Cluster 16	Cluster 16
SBmhc1_10	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_118	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_13	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_21	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_4	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_6	Cluster 3	Cluster 10	Cluster 3	Cluster 3
SBmhc1_126	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_15	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_26	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_37	Cluster 4	Cluster 2	Cluster 2	Cluster 2

SBmhc1_42	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_85	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_98	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_49	Cluster 4	Cluster 2	Cluster 2	Cluster 2
SBmhc1_103	Cluster 4	Cluster 5	Cluster 10	Cluster 12
SBmhc1_67	Cluster 4	Cluster 5	Cluster 10	Cluster 12
SBmhc1_79	Cluster 4	Cluster 5	Cluster 10	Cluster 12
SBmhc1_110	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_113	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_25	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_64	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_86	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_93	Cluster 5	Cluster 6	Cluster 4	Cluster 5
SBmhc1_119	Cluster 5	Cluster 7	Cluster 4	Cluster 6
SBmhc1_35	Cluster 5	Cluster 7	Cluster 4	Cluster 6
SBmhc1_58	Cluster 5	Cluster 7	Cluster 4	Cluster 6
SBmhc1_69	Cluster 5	Cluster 7	Cluster 4	Cluster 6
SBmhc1_12	Cluster 6	Cluster 18	Cluster 18	Cluster 18
SBmhc1_71	Cluster 6	Cluster 18	Cluster 18	Cluster 18
SBmhc1_50	Cluster 6	Cluster 6	Cluster 11	Cluster 13

SBmhc1_51	Cluster 6	Cluster 6	Cluster 11	Cluster 13
SBmhc1_52	Cluster 6	Cluster 6	Cluster 11	Cluster 13
SBmhc1_100	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_104	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_129	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_54	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_65	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_99	Cluster 6	Cluster 9	Cluster 11	Cluster 13
SBmhc1_115	Cluster 7	Cluster 11	Cluster 17	Cluster 17
SBmhc1_31	Cluster 7	Cluster 11	Cluster 17	Cluster 17
SBmhc1_34	Cluster 7	Cluster 11	Cluster 17	Cluster 17
SBmhc1_36	Cluster 7	Cluster 11	Cluster 17	Cluster 17
SBmhc1_88	Cluster 7	Cluster 11	Cluster 17	Cluster 17
SBmhc1_63	Cluster 7	Cluster 11	Cluster 18	Cluster 18
SBmhc1_59	Cluster 7	Cluster 17	Cluster 13	Cluster 14
SBmhc1_106	Cluster 8	Cluster 1	Cluster 1	Cluster 1
SBmhc1_122	Cluster 8	Cluster 1	Cluster 1	Cluster 1
SBmhc1_9	Cluster 8	Cluster 12	Cluster 7	Cluster 9
SBmhc1_90	Cluster 8	Cluster 12	Cluster 7	Cluster 9
SBmhc1_102	Cluster 8	Cluster 12	Cluster 7	Cluster 9

SBmhc1_57	Cluster 8	Cluster 8	Cluster 5	Cluster 7
SBmhc1_62	Cluster 8	Cluster 8	Cluster 5	Cluster 7
SBmhc1_73	Cluster 8	Cluster 8	Cluster 5	Cluster 7
SBmhc1_81	Cluster 8	Cluster 8	Cluster 5	Cluster 7
SBmhc1_22	Cluster 9	Cluster 3	Cluster 8	Cluster 10
SBmhc1_46	Cluster 9	Cluster 3	Cluster 8	Cluster 10
SBmhc1_72	Cluster 9	Cluster 3	Cluster 8	Cluster 10
SBmhc1_76	Cluster 9	Cluster 3	Cluster 8	Cluster 10
SBmhc1_92	Cluster 9	Cluster 3	Cluster 8	Cluster 10
SBmhc1_24	Cluster 9	Cluster 4	Cluster 9	Cluster 11
SBmhc1_43	Cluster 9	Cluster 4	Cluster 9	Cluster 11
SBmhc1_74	Cluster 9	Cluster 4	Cluster 9	Cluster 11
SBmhc1_1	Cluster 9	Cluster 4	Cluster 9	Cluster 11
SBmhc1_128	Cluster 9	Cluster 5	Cluster 10	Cluster 12
SBmhc1_130	Cluster 9	Cluster 5	Cluster 10	Cluster 12
SBmhc1_38	Cluster 9	Cluster 5	Cluster 10	Cluster 12
SBmhc1_3	Cluster 10	Cluster 12	Cluster 6	Cluster 8
SBmhc1_32	Cluster 10	Cluster 12	Cluster 6	Cluster 8
SBmhc1_7	Cluster 10	Cluster 12	Cluster 6	Cluster 8
SBmhc1_8	Cluster 10	Cluster 12	Cluster 6	Cluster 8

SBmhc1_2	Cluster 10	Cluster 3	Cluster 8	Cluster 10
SBmhc1_48	Cluster 10	Cluster 3	Cluster 8	Cluster 10
SBmhc1_5	Cluster 10	Cluster 3	Cluster 8	Cluster 10
SBmhc1_91	Cluster 10	Cluster 3	Cluster 8	Cluster 10

Table A3.3. Common and private alleles in the Portuguese and Celtic Shelf populations. Highlighting indicates an allele absent from a given population.

Allele name	Count Celtic	Count Portugal
SBmhc1_1	22	18
SBmhc1_2	20	12
SBmhc1_3	16	13
SBmhc1_4	17	12
SBmhc1_5	12	12
SBmhc1_6	11	6
SBmhc1_7	10	9
SBmhc1_8	12	5
SBmhc1_9	13	4
SBmhc1_10	3	9
SBmhc1_11	2	3
SBmhc1_12	3	3
SBmhc1_13	4	6
SBmhc1_14	3	6
SBmhc1_15	6	6
SBmhc1_16	4	6

SBmhc1_17	6	4
SBmhc1_18	5	3
SBmhc1_19	5	4
SBmhc1_20	1	4
SBmhc1_21	4	2
SBmhc1_22	2	4
SBmhc1_23	4	2
SBmhc1_24	1	3
SBmhc1_25	0	6
SBmhc1_26	4	3
SBmhc1_27	2	3
SBmhc1_28	3	2
SBmhc1_29	2	2
SBmhc1_30	2	2
SBmhc1_31	0	2
SBmhc1_32	3	4
SBmhc1_33	0	0
SBmhc1_34	3	1
SBmhc1_35	0	2
SBmhc1_36	0	2

|		
SBmhc1_37	4	2
SBmhc1_38	0	3
SBmhc1_39	3	2
SBmhc1_40	2	3
SBmhc1_41	2	2
SBmhc1_42	5	1
SBmhc1_43	2	1
SBmhc1_44	2	1
SBmhc1_45	3	1
SBmhc1_46	2	1
SBmhc1_47	1	2
SBmhc1_48	2	1
SBmhc1_49	1	0
SBmhc1_50	0	0
SBmhc1_51	1	2
SBmhc1_52	2	1
SBmhc1_53	0	1
SBmhc1_54	3	0
SBmhc1_55	2	0
SBmhc1_56	1	1

SBmhc1_57	1	0
SBmhc1_58	0	0
SBmhc1_59	0	0
SBmhc1_60	1	0
SBmhc1_61	0	1
SBmhc1_62	1	1
SBmhc1_63	0	2
SBmhc1_64	0	2
SBmhc1_65	2	2
SBmhc1_66	2	0
SBmhc1_67	1	0
SBmhc1_68	0	0
SBmhc1_69	3	0
SBmhc1_70	1	2
SBmhc1_71	1	0
SBmhc1_72	0	0
SBmhc1_73	2	0
SBmhc1_74	2	2
SBmhc1_75	0	2
SBmhc1_76	1	0

SBmhc1_77	1	1
SBmhc1_78	1	1
SBmhc1_79	1	1
SBmhc1_80	1	1
SBmhc1_81	0	1
SBmhc1_82	2	1
SBmhc1_83	1	0
SBmhc1_84	1	1
SBmhc1_85	0	1
SBmhc1_86	1	0
SBmhc1_87	1	0
SBmhc1_88	2	0
SBmhc1_89	1	0
SBmhc1_90	0	0
SBmhc1_91	2	0
SBmhc1_92	1	1
SBmhc1_93	0	2
SBmhc1_94	0	0
SBmhc1_95	1	0
SBmhc1_96	1	1

SBmhc1_97	1	0
SBmhc1_98	1	0
SBmhc1_99	1	1
SBmhc1_100	0	0
SBmhc1_101	1	1
SBmhc1_102	1	1
SBmhc1_103	1	1
SBmhc1_104	0	2
SBmhc1_105	0	2
SBmhc1_106	0	1
SBmhc1_107	0	0
SBmhc1_108	2	0
SBmhc1_109	1	2
SBmhc1_110	1	0
SBmhc1_111	1	0
SBmhc1_112	0	0
SBmhc1_113	1	1
SBmhc1_114	1	1
SBmhc1_115	0	1
SBmhc1_116	1	1

SBmhc1_117	0	1
SBmhc1_118	2	0
SBmhc1_119	1	0
SBmhc1_120	1	0
SBmhc1_121	0	1
SBmhc1_122	0	1
SBmhc1_123	2	0
SBmhc1_124	1	1
SBmhc1_125	0	1
SBmhc1_126	0	0
SBmhc1_127	2	0
SBmhc1_128	0	1
SBmhc1_129	2	0
SBmhc1_130	1	0
SBmhc1_131	1	1
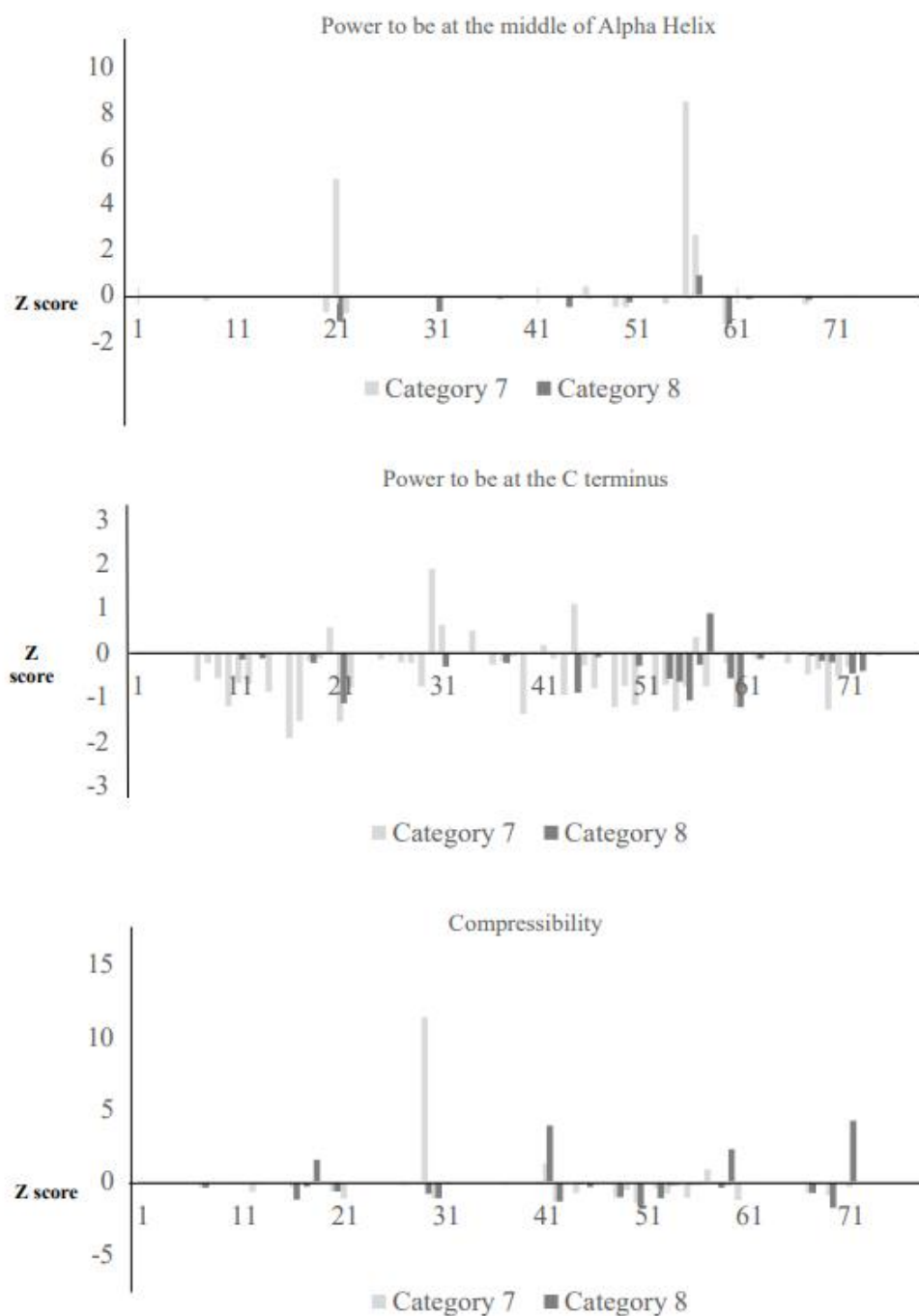
Figure A3.1. Z-scores relating to phenotypic amino acid properties/traits (Power to be at the middle of the alpha-helix, power to be at the c-terminus and compressibility). Categories represent magnitude of amino acid property change at nonsynonymous residues on a site by site basis. Class 8 represents changes of the highest magnitude, followed by class 7. Low and 0 z score values represent areas of conservation in terms of amino acid properties, in contrast to high values which represent areas of selection.
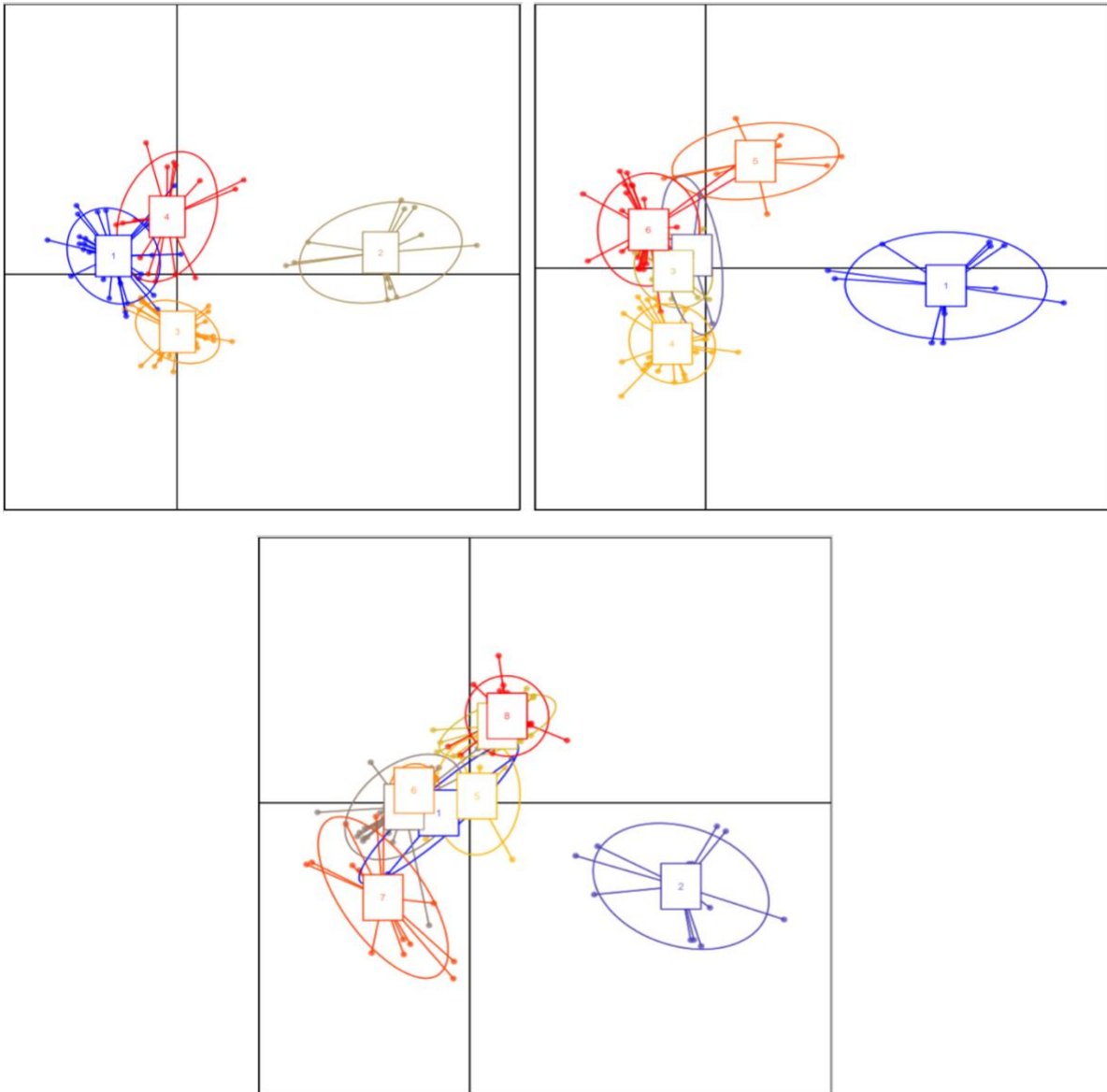
Figure A3.2. Clustering of major histocompatibility complex (MHC) alleles in *Dicentrarchus labrax*, based on the physicochemical properties of amino acids (Z scores) of positively selected sites, using discriminant analysis of principal components (DAPC).
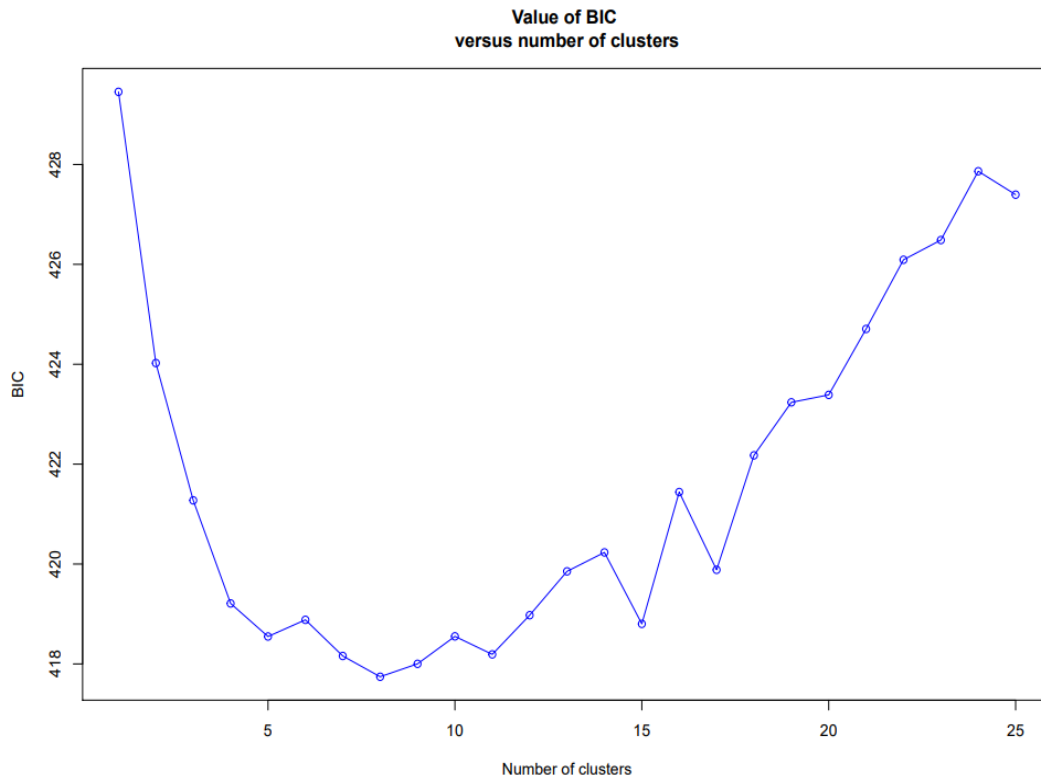
Figure A3.3. Bayesian information criterion (BIC), used to identify the optimal number of clusters.
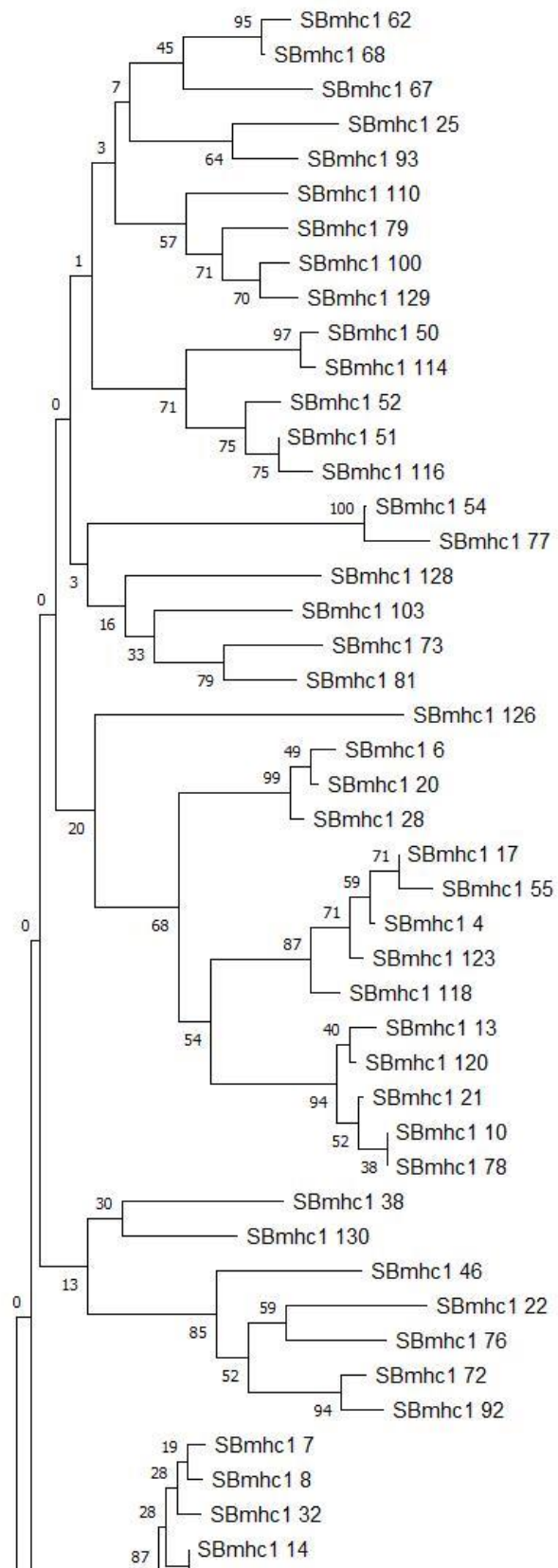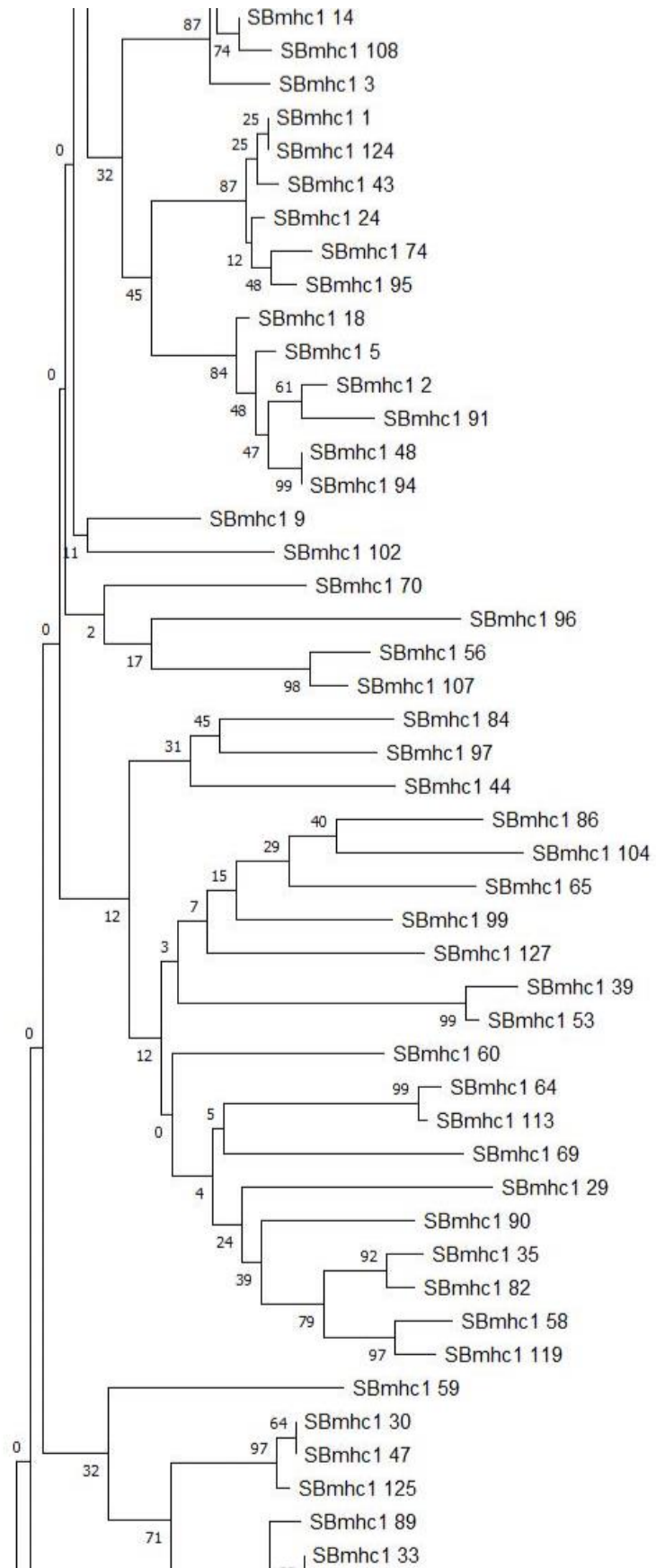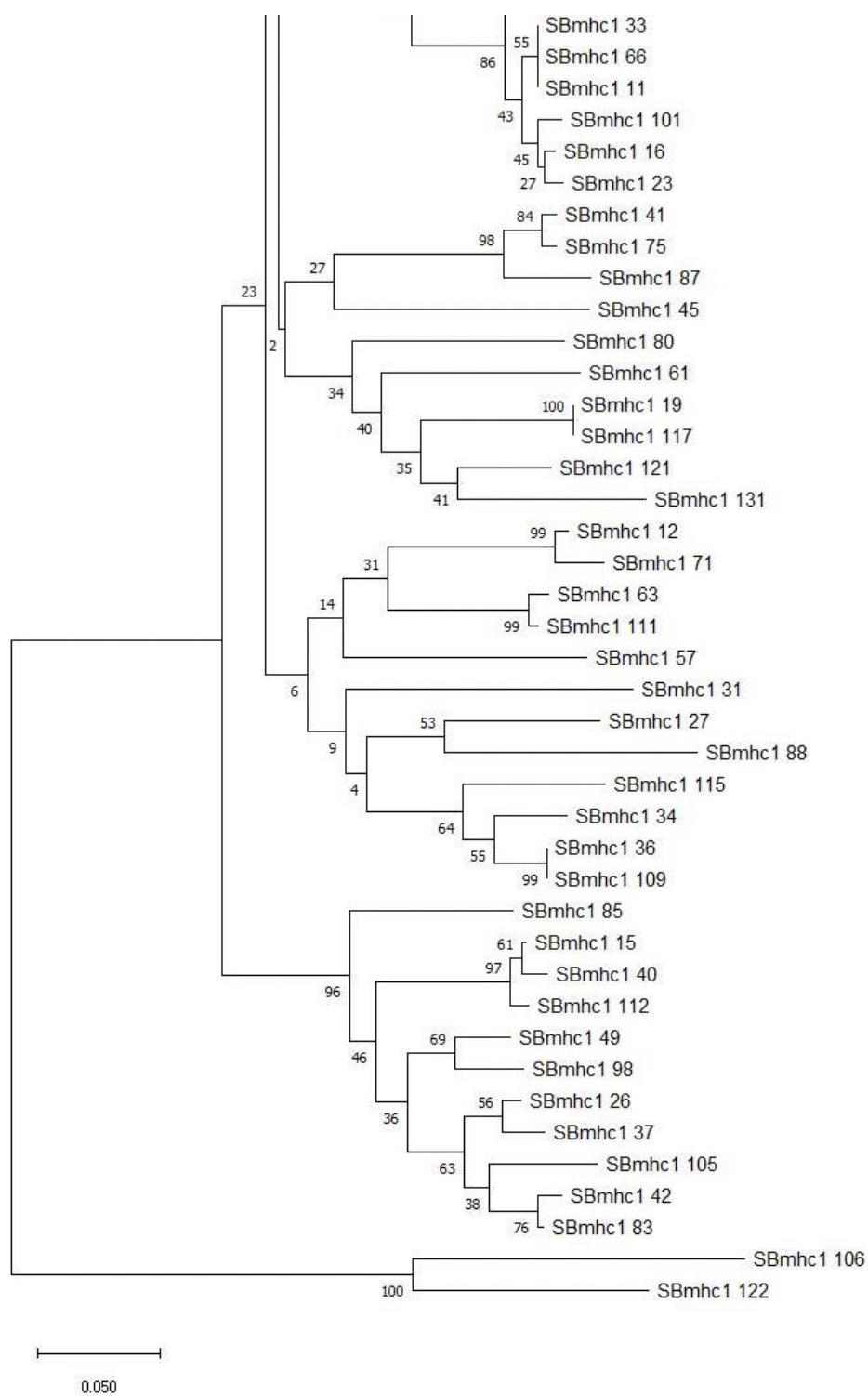
Figure A3.4. Neighbour Joining Tree based on codon alignment of sea bass alleles in this study (amino acid sequences), using ClustalW in MEGAX. 1000 bootstrap replicates were used.
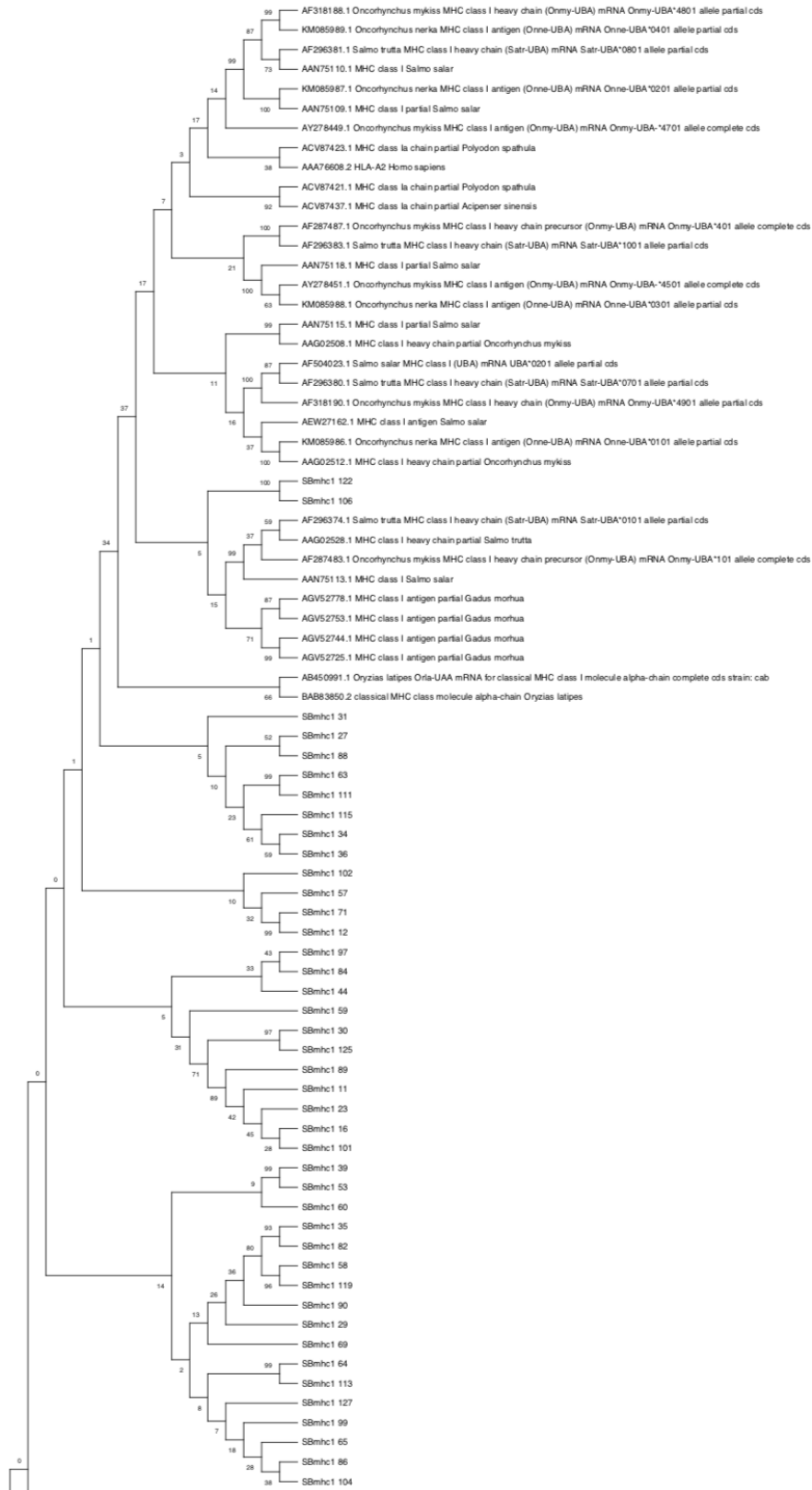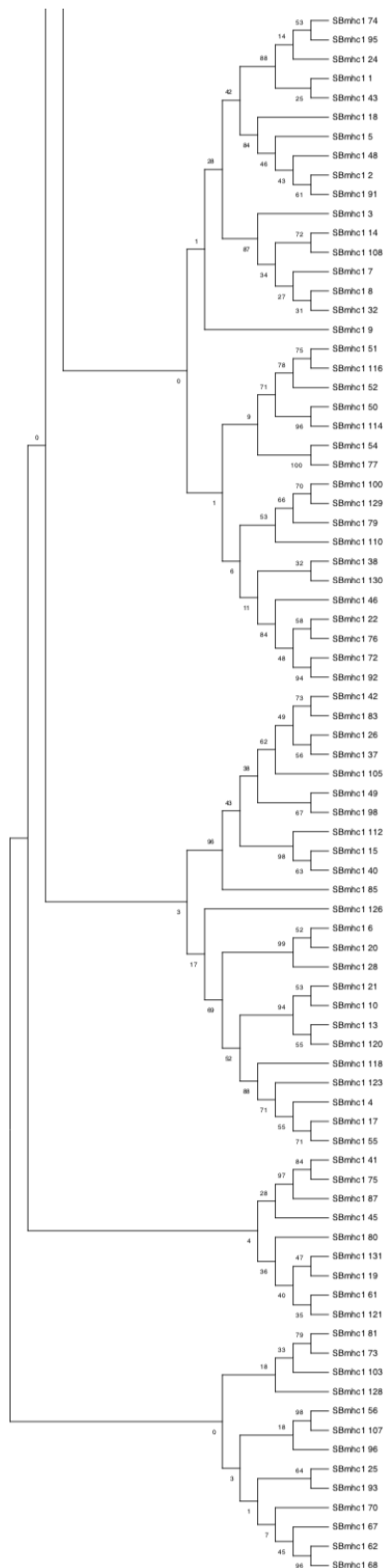
Figure A3.5.   Neighbour Joining Tree based on codon alignment of sea bass alleles in this study (amino acid sequences) and 32 MHC class 1 fish sequences (Grimholt et al., 2015) using ClustalW in MEGAX. 1000 bootstrap replicates were used.
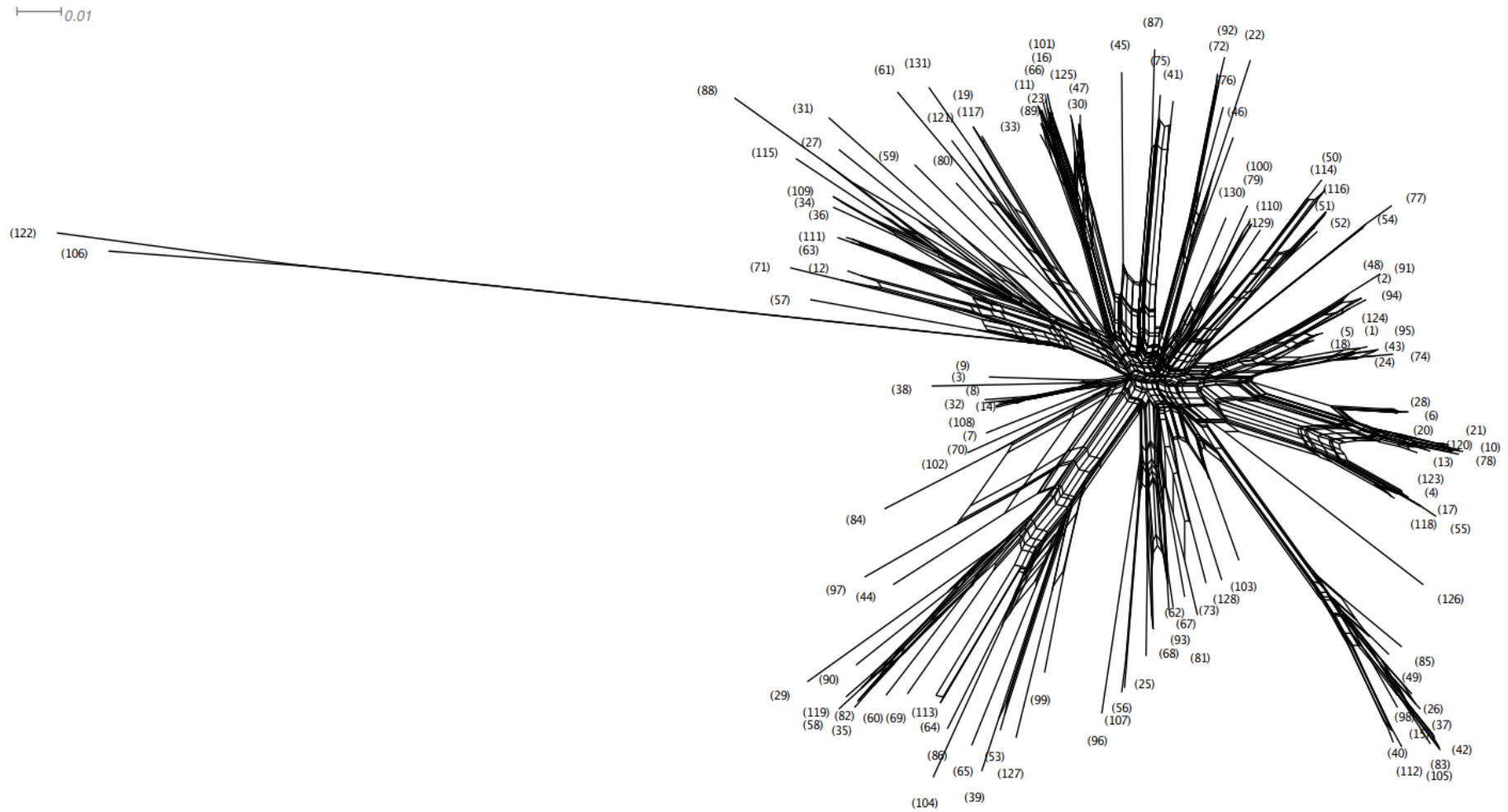
Figure A3.6. NeighborNet (SplitsTree), based on codons/amino acid sequence using p-distance, numbering represents allele numbers.
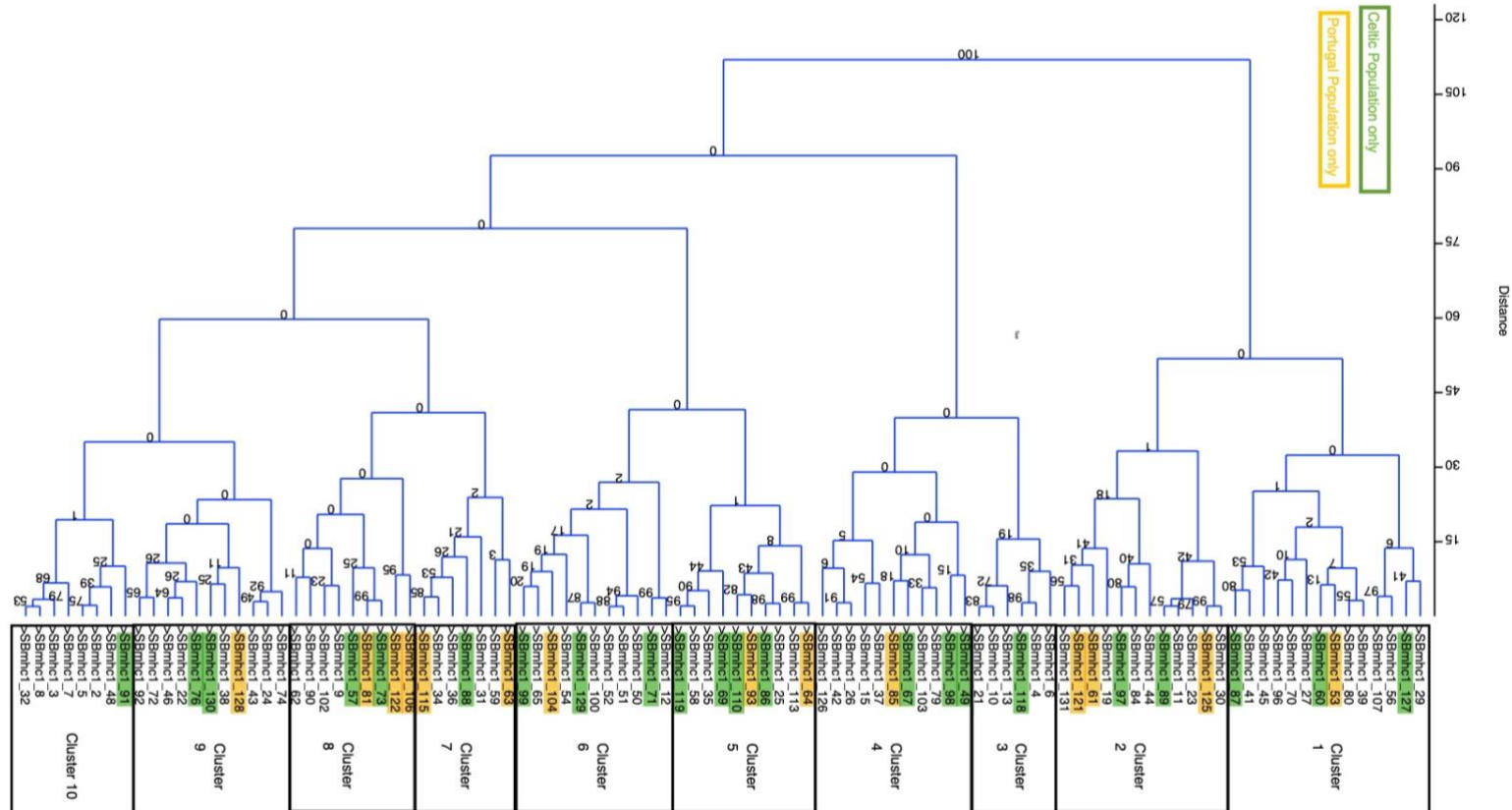
Figure A3.7. Phylogenetic tree constructed using Ward's method with Euclidian distance (PAST, 1000 bootstrap), based on Z scores of amino acids for positively selected sites (PSS) (Doytchinova & Flower, 2005). Only non-redundant sequences at PSS level are represented here. Green and orange highlighting denotes alleles present in only the Celtic Shelf and Portuguese populations respectively. Clustering has been picked by eye.
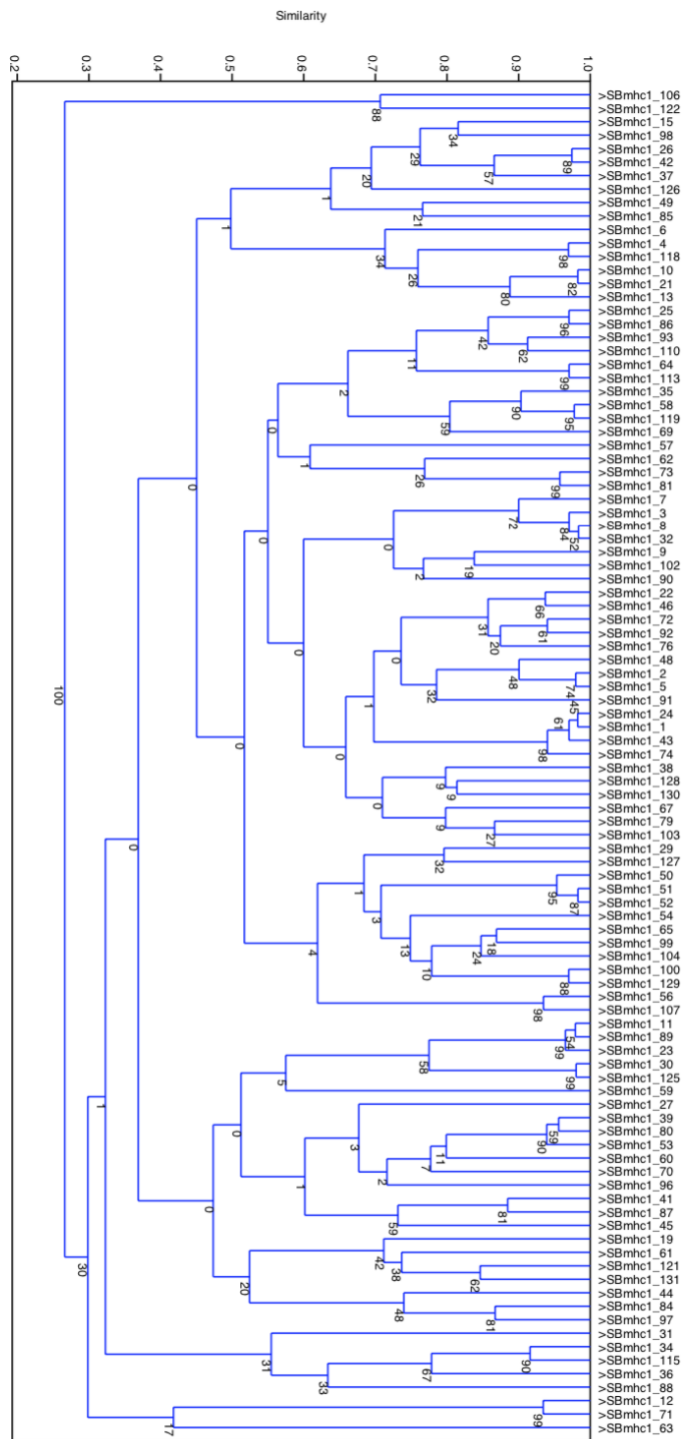
Figure A3.8. Clustering based on Positively Selected Sites (amino acids, PSS), using the UPGMA algorithm and correlation as distance, generated using PAST4.
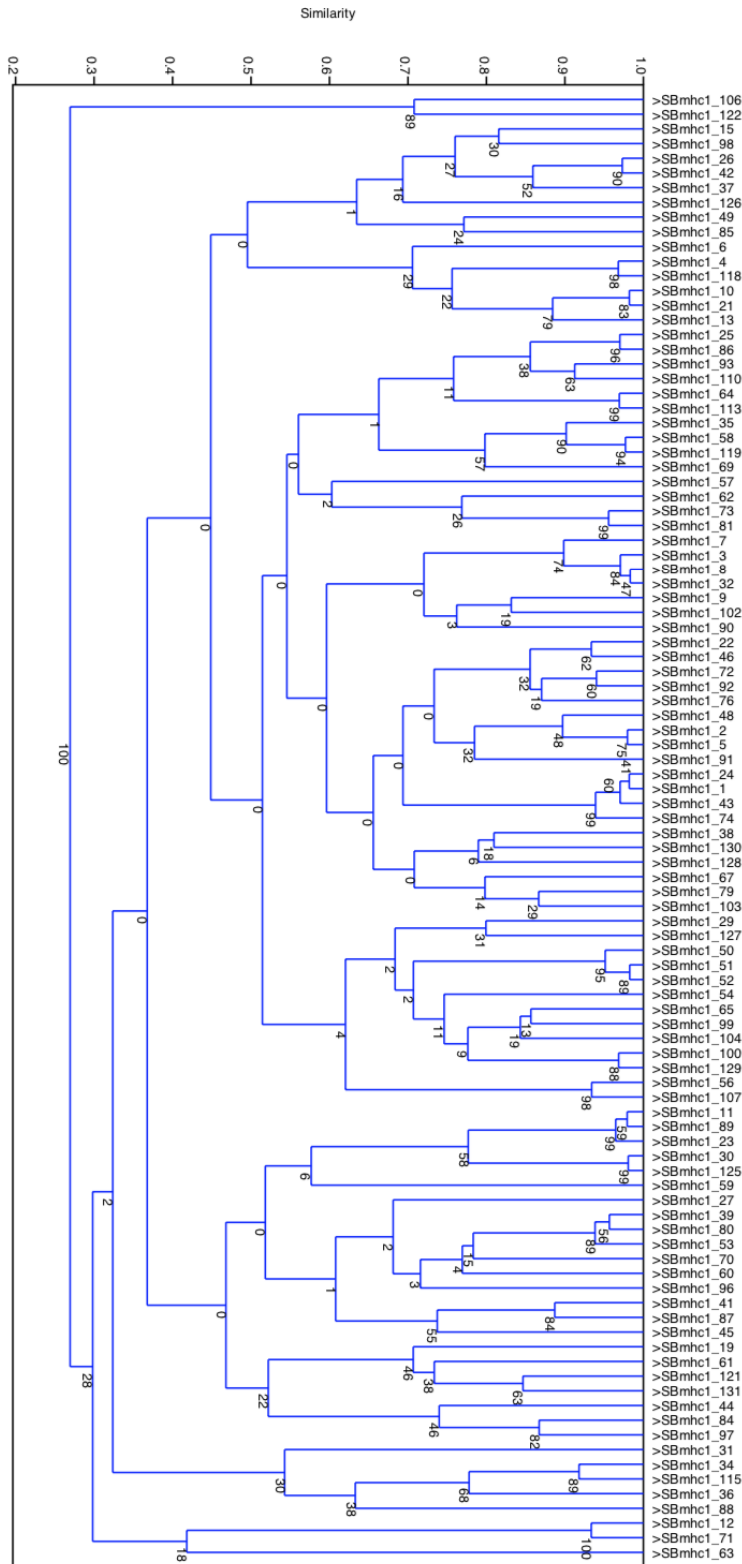
Figure A3.9. Clustering based on Positively Selected Sites (amino acids, PSS), using the UPGMA algorithm and cosine as distance, generated using PAST4.
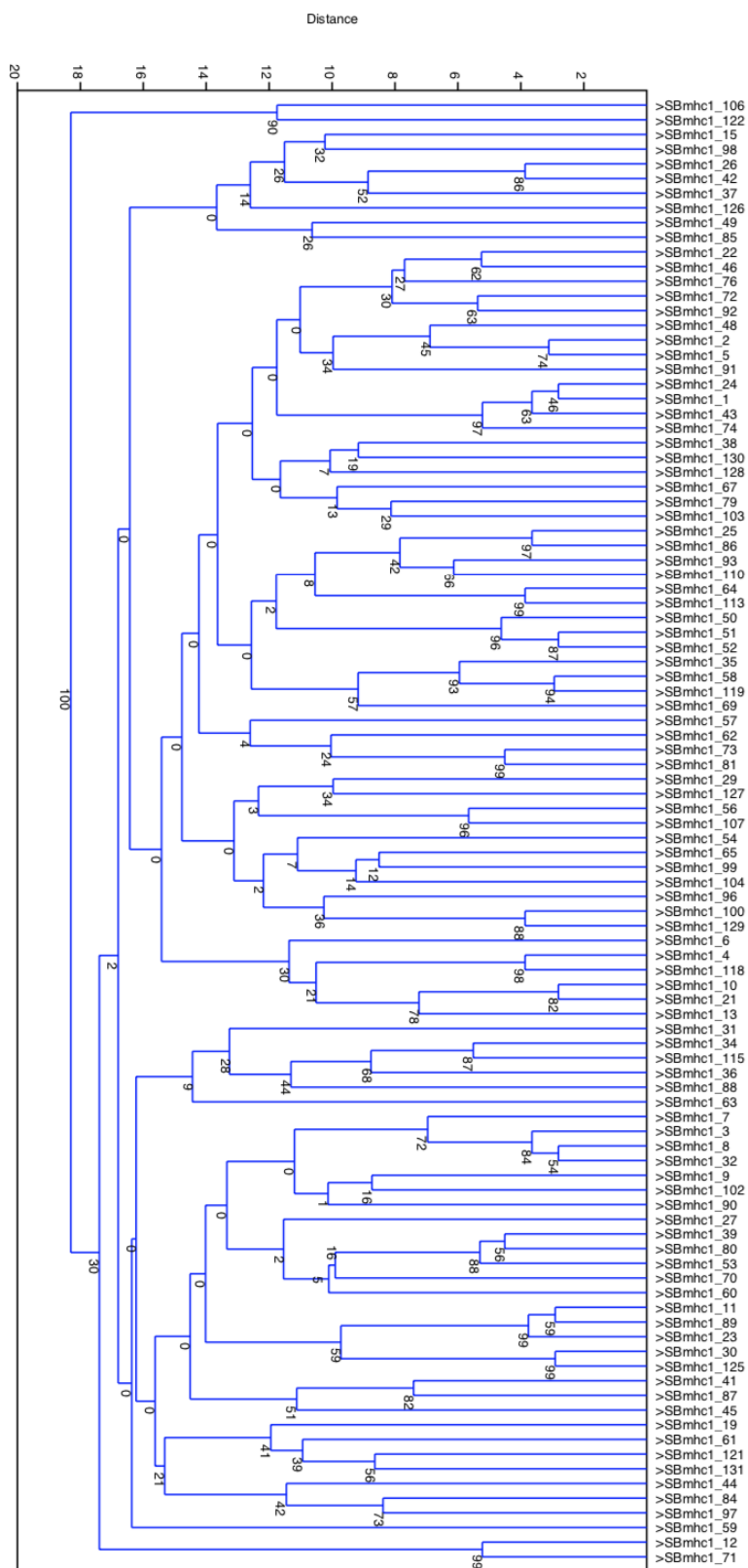
Figure A3.10. Clustering based on Positively Selected Sites (amino acids, PSS), using the UPGMA algorithm and Euclidian distance, generated using PAST.

Appendix 4: Ethical approval

Project Ethics Assessment Confirmation|Cadarnhad o Asesiad Moeseg Prosiect

C cosethics@swansea.ac.uk

Fri 18/10/2019 09:10

To: RATCLIFFE F.

Cc: Consuegra

This is an automated confirmation email for the following project. The Ethics Assessment status of this project is: APPROVED

Applicant Name: Frances Ratcliffe
Project Title: Rapid quantitative assessment of fish larvae community composition using metabarcoding
Project Start Date: 15/07/2018
Project Duration: 18 months
Approval No: SU-Ethics-Student-181019/1996

NOTE: This notice of ethical approval does not cover aspects relating to Health and Safety. Please complete any relevant risk assessments prior to commencing with your project.

Project Ethics Assessment Confirmation|Cadarnhad o Asesiad Moeseg Prosiect

C

cosethics@swansea.ac.uk
Mon 28/06/2021 11:06

👍 ↩ ↞ → ⋯

To: RATCLIFFE F. ▮▮▮▮▮▮
Cc: Consuegra ▮

This is an automated confirmation email for the following project. The Ethics Assessment status of this project is: APPROVED

Applicant Name: Frances Ratcliffe
Project Title: Allelic and supertype diversity of MHC class I-alpha reveal fine-scale population structure in European seabass (Dicentrarchus labrax)
Project Start Date: 01/02/2020
Project Duration: 3 months
Approval No: SU-Ethics-Student-280621/4326

NOTE: This notice of ethical approval does not cover aspects relating to Health and Safety. Please complete any relevant risk assessments prior to commencing with your project.