DECLAN MORRISSEY

Population genomics of a marine gastropod with limited dispersal capabilities

Mestrado em Biologia Marinha

Supervisor

Dr Rita Castilho

Co-supervisor

Dr Michelle Taylor

UNIVERSIDADE DO ALGARVE

Faculdade de Ciências e Tecnologia

2019

Declaração de autoria de trabalho

Population genomics of a commercially exploited marine gastropod with limited dispersal capabilities

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

Declan Morrissey

Acknowledgements

Thank you to Dr Michelle Taylor and Dr Tom Cameron for having me at the University of Essex and for providing guidance and facilitating this research. It was a pleasure to be a part of your labs. Thank you to Dr Jake Goodall and Dr Zophonías Jónsson for hosting me in Iceland and for providing support every step of the way. Thank you, Dr Rita Castilho, for advising me and mentoring me during this research and during my whole time at the University of Algarve. I hope we can continue to work together in the future.

To Tania Maynard, our amazing molecular technician, I would still be extracting whelk DNA if it wasn't for your support. Thank you for always having my best interests at heart and being a sympathetic ear.

Thank you to my friends and lab mates at the University of Essex; Jess, Reem, Amy, Mark, Vincent, Matt, Mia, Beth, Phil, Alice M, Shawn, Gareth, Alice L, Howard, Tom, and Michelle. Thank you for the evening/day drinks, weekend/midweek barbeques, and overall support. Your friendship made it hard to leave.

Thanks to Dr Anna Trego for proofreading. This was probably the hardest task I asked anyone to do.

Finally, to Jamie Maxwell. Thank you for the love, patience and understanding when it was not easy to give. Thank you for both supporting my obsessive coding, and for reminding me to let science be a part of my life, not my whole life.

Resumo

A dispersão de organismos é um processo essencial que promove o fluxo genético e contraria a diferenciação das populações. A falta de barreiras aparentes no ambiente marinho significa que os organismls podem alcançar a panmixia em escalas espaciais muito menores do que as que se verificam no meio terrestre. Até recentemente, a genética de populações apoiava-se num conjunto relativamente pequeno de marcadores genéticos para fazer inferências sobre a conectividade de populações em grandes escalas espaciais. No entanto, recentes avanços nas técnicas moleculares produziram um conjunto de novos métodos de sequenciação de DNA, comumente referidos como Sequenciação de Nova Geração (NGS), que permitem a obtenção de milhares de marcadores do genoma, aumentando potencialmente a resolução dos dados. RADseq é um tipo de técnica NGS que emprega enzimas de restrição para cortar o genoma nos locais de restrição, produzindo milhares polimorfismos de nucleotídeo único (SNP, *single-nucleotide polymorphism*).

Buccinum undatum é um neogastropode amplamente distribuído em ambos os lados do Atlântico Norte, desde o Canadá até ao Golfo da Biscaia podendo atingir profundidades de 1000 m. Esta espécie é o maior gastrópode marinho comestível do Atlântico Norte, com um comprimento máximo de concha de 150 mm. Alimenta-se principalmente de bivalves e de pequenss crustáceos. *B. undatum* constitui uma pescaria relevante no Reino Unido desde 1922, só em 2017 foram desembarcadas 20.800 toneladas, correspondendo à sexta maior pesca de marisco no valor de 22,7 milhões de libras. Apesar de sua importância comercial, apenas três estudos se concentraram no fluxo genético e estrutura populacional de *B. undatum.* Todos esses estudos concordaram que o *B. undatum* pode ser geneticamente diferenciado ao longo de dezenas de quilómetros e que as populações são mais divergentes em baías e enseadas do que aquelas mais afastadas da costa. Esses estudos sugeriram um modelo de *stepping-stone* para uma grande população semi-contínua.

No total, 195 indivíduos foram coletados em 13 locais de amostragem do Mar do Norte, Canal da Mancha e Mar da Irlanda de Dezembro de 2018 a Fevereiro de 2019. Destes 191 indivíduos foram escolhidos para prosseguir com a preparação da biblioteca de DNA de dupla digestão (DdRAD) usando ApeKI e BamHI-HF. A sequenciação foi realizada no Illumina HiSeq X Ten. No total, obtiveram-se 1.427.813.991 *reads*. O controle de qualidade e a desmultiplexação foram realizados com o software STACKS e os parâmetros: m = 3, n = 3 e M = 3. 4. Do total de *reads*, 19% não tinham código de barras, 0.08% foram removidos devido à baixa qualidade, 4.4% não possuíam local de corte RAD, resultando em 1.303.931.081 *reads*. Depois da remoção dos indivíduos com menos de 20x de cobertura, permaneceram 141 indivíduos, o que corresponde a 6 a 15 indivíduos por local amostrado, representando 92.8% de todas as reads retidas (1.210.090.726). A cobertura obtida variou de 21.41x a 71.52x (média = 48.2x, S.D. = 13.2x). Os SNPS retidos correspondem ao que estavam simultaneamente presentes em 70% dos locais de amostragem, em 80% dos indivíduos, tinham uma frequência alélica menor maior que 5% e não excederam uma heterozigosidade máxima de 50%. Removeram-se ainda todos os SNPS que não estavam em Hardy Weinberg Equilibrium (HWE) e estavem em disequilibrio de ligação (*linkage disequilibrium*). No final da filtragem foram retidos 885 SNPs. Foram detectados 4 loci *outliers* sob seleção positiva putativa. No final foram utilizados para análise dois conjuntos de dados; (i) somente loci neutros, correspondendo a 881 loci e (ii) loci outlier - quatro loci outlier identificados como estando sob selecção positiva.

Os resultados baseados somente nos loci neutros (i) encontraram evidências de três grupos genéticos distintos com frequências genéticas semelhantes nos locais de amostragem. A comparação entre pares de locais foi significativa em 22% dos pares, que decresceu para 12% após correção para múltiplos testes. A subestrutura detectada foi de amplitude menor do que noutros estudos. No entanto, isso pode ser resultado do esforço de escala na amostragem, pois o presente estudo é de relativamente pequena escala geográfica. Os valores de Fst variaram de 0.0052 a 0.0133, indicando que subestrutura significativa estava presente em todo o âmbito geográfico do estudo. Detectou-se uma significativa tendência de isolamento por distância nos locais de amostragem do Mar do Norte (r = 0.51, p = 0.001) e uma correlação significativa entre a distância genética e geográfica ao usar todos os locais de amostragem (r = 0,54, p = 0,002). Com base na análise de agrupamentos bayesianos, detectam-se três agrupamentos genéticos distintos, um dos quais numa frequência um pouco maior nos locais de amostragem do Canal da Mancha e outro em frequências mais altas no Mar do Norte. O grupo restante tem uma mistura em proporções quase iguais em todos os locais de amostragem. Estes resultados concordam com estudos anteriores de que *B. undatum* consiste numa grande população isolada com subestrutura significativa. No entanto, a divergência populacional geral é suprimida pelo alto fluxo genético e um grande tamanho efetivo da população. Foi relatado anteriormente que *B. undatum* é mais diferenciado em baías e enseadas. No entanto, este estudo não encontrou nenhuma evidência nesse sentido, embora apenas um único local de amostragem se localizasse numa uma baía. Será necessário um esforço de amostragem mais intenso nas baías e enseadas para testar a robustez dessa tendência.

Os loci *outliers* não revelaram nenhuma estrutura populacional e, em vez disso, apoiam a existência de uma única população. Outliers sumetidos a *blast* do *National Center for Biotechnology Information* não revelaram qualquer correspondência significativa com genes anotados de outras espécies. Assim, na ausência de um genoma de *B. undatum* anotado ou dados ambientais à escala geográfica da amostragem, não é possível explicar os fatores subjacentes à selecção positiva detectada.

Este estudo contribui para a crescente literatura que usa RADseq para delinear a estrutura populacional de escala geográfica fina. *Buccinum undatum* tem baixo potencial de dispersão e os valores de Fst no presente estudo são muito inferiores ao que seria esperado. Submetemos a explicação possível que um grande tamanho populacional efectivo e uma população semi-contínua suprimem a divergência genómica desta espécie.

Palavras-chave: búzios, molusco, ddRAD, alto fluxo genético, genómica populacional

Abstract

Population genomics is important for understanding the degree of genetic connectivity and effective dispersal over geographic distances. Connectivity, or the constraint of it, influences both local and regional biodiversity and thus is of primary interest to both evolutionary and ecological studies. In recent years, previous assumptions regarding dispersal capabilities, and their function as a primary driver of expected genetic structure of populations have been challenged by Next-Generation Sequencing techniques. This study investigated the population connectivity of *Buccinum undatum* with Single Nucleotide Polymorphisms (SNPs) derived from double-digest Restriction associated DNA sequencing. In total, 191 individuals were sequenced from the Southern North Sea, English Channel, and Irish Sea, a geographic scope of 1165 km. After strict quality control and filtering, 885 biallelic SNPs and 141 individuals were retained. Outlier detection revealed 4 loci under putatively positive selection. Two datasets were analysed; a neutral loci dataset which contained 881 SNPS, and an outlier loci dataset that contained the 4 SNPs identified as outliers. Results from the neutral dataset advocated for a single large population with no overall structure but significant sub-structure. However, sub-structure was much less frequent than previously reported for the species. Individuals sampled within a bay were not more genetically differentiated than those outside of bay, a previously reported trait of *B. undatum*. There was significant isolation by distance observed across the majority of the geographic range. Outlier analysis did not reveal any hidden population structure, nor any isolation by distance. Overall, results presented within fundamentally agreed with previous studies that *B. undatum* consists of a single population, that is semi-continuous in nature, with sub-structure present. However, high gene flow and a large effective population size supressed overall population divergence.

Keywords: ddRAD, Mollusc, Population genomics, Population structure, Whelk

Table of Contents

Index of figures

Chapter 1. Introduction

Chapter 2: ddRAD reveals high levels of gene flow and no population structure in a marine gastropod with limited dispersal capabilities

Supplementary Material

Index of tables

Chapter 2: ddRAD reveals high levels of gene flow and no population structure in a marine gastropod with limited dispersal capabilities

Supplementary Material

List of abbreviations

16S - 16S ribosomal RNA gene

bp – base pair

CO1 – Cytochrome Oxidase Subunit 1

ddRAD – double digest Restriction site Associated DNA Sequencing

FDR – False Discovery Rate

gDNA – Genomic Deoxyribonucleic Acid

HWE – Hardy Weinberg Equilibrium

IBD – Isolation by Distance

IFCA – Inshore Fisheries and Conservation Authority

LD – Linkage Disequilibrium

LGM – Last Glacial Maxima MDS
– Multidimensional Scaling MLS –
Minimum Landing Size

NGS – Next-Generation Sequencing

PLD – Pelagic Larval Duration

RADseq – Restriction site Associated DNA Sequencing

SNPs – Single Nucleotide Polymorphisms

TBT – Tributyltin

1. Chapter 1: Introduction

Population connectivity in the marine environment

Dispersal is a key process that affects population growth, gene flow, and overall population persistence. For this reason, it is an important parameter to consider when discussing the evolution of species in natural systems. Dispersal drives population connectivity of a species over geographic scales, a process which influences local and regional biodiversity (Chust *et al.* 2016). Marine invertebrates display a large diversity of reproductive strategies. These strategies can be defined based on number of sexual reproductive events in a year and lifespan, sexual expression (gonochoric or hermaphroditic), or mode of reproduction (broadcast spawning of gametes or larvae, brooding, direct deposition of eggs) (Wangensteen *et al.* 2017). This large diversity of reproductive strategies has two extremes; entirely benthic and direct development of young in eggs, or entirely planktotrophic/pelagic (Riginos and Liggins 2013). Reproduction in marine organisms is complicated, and organisms commonly have a combination of benthic and pelagic stages in their life. e.g. Seagrasses sexually reproduce to produce seeds that do not disperse long distances; however, blades can detach due to high turbidity and disperse. Dispersal can be investigated by its effects on gene flow and the population structuring for which it is responsible. It is widely presumed that species that have a life history including the direct development of larvae without a larval planktonic stage, have more restrictive dispersal potential than those with a pelagic stage (Scheltema 1986).

In benthic species, dispersal often occurs at the earliest stage of an organism's life history, either as a larva or as gamete (Cowen and Sponaugle 2009). Factors that influence effective dispersal at this stage are biophysical, e.g. temperature and salinity tolerances (Cowen and Sponaugle 2009). In species that have a planktonic larval stage, the planktonic larval duration (PLD), in the absence of oceanic barriers, has been shown to have a direct influence on the gene flow of species over large spatial scales (Pascual *et al.* 2017). The longer larvae stay in the plankton, the further they can be transported along the ocean currents. However, challenges to these assumptions have been made in recent years with the advent of interdisciplinary genetics and ecological modelling research, called 'seascape genetics' (or now 'seascape genomics'). In the case where planktonic larvae develop into nekton, or other processes that alter our understanding of larva behaviour in the plankton, models can uncover unexpected dispersal trajectories (Leis 2010). By studying both the oceanographic currents present in the study region and the behaviour of the larvae it is possible to more accurately detect the dispersal limits and potential of marine species.

Species that do not have a planktonic larval stage generally exhibit high levels of population structure. For example, the population structure of four littorinid gastropods with differing life-history strategies were examined (Kyle and Boulding 2000). The gastropods were all intertidal and the sampling effort occurred over the same geographic scope. In that study, *Littorina subrotundata,* a direct developer, exhibited higher levels of population structure than both Littorina plena and *Littorina scutulata*, both of which go through a planktonic larval stage.

Finally, mixed development is an example of a life history strategy where there is a combination of direct development and a subsequent larval stage (Pechenik 1979). Larvae develop initially encapsulated in egg masses, before being released into the water column. This type of life history strategy is common amongst gastropods and polychaetes. Constructing an egg mass is metabolically costly, and thus must confer a distinct advantage. Larvae that spent a week in an egg mass before being released into the plankton had a 10% reduction in mortality when compared to larvae directly released to the plankton (Pechenik 1979). Pechenik (1979) hypothesised that egg capsules were involved in homeostasis of juveniles, preventing mortality from extreme changes in environmental conditions. Futher support for this concept indicated that encapsulation reduced mortality associated with large salinity changes by slowing the speed of osmotic transfer between the embryo and the environment (Pechenik 1982) and protected juveniles from ultraviolet radiation, desiccation, and the osmotic stress (Rawlings 1999).

Traditionally dispersal of larvae and juveniles has been through natural means; by water currents, rafting on natural/hitchhiking on other organisms. However, there is evidence of anthropogenic influence evidence of connectivity, primarily discussed in terms of non-native or invasive species. The vectors of introduction are numerous, but the shipping industry is the largest vector followed by the aquaculture trade (Molnar *et al.* 2008). These anthropogenic vectors can cause the establishment of new populations.

Neutral and adaptive genetic selection

Traditionally, population genetics used neutral markers, such as microsatellites that are in Hardy-Weinberg Equilibrium (HWE). The principal of HWE is that allele frequencies remain constant in a population over multiple generations in the absence of evolutionary forces. Populations that are in HWE must pass some basic assumptions. The first, that organisms are diploid. Secondly, that random mating must occur. This assumption can be broken in systems where breeding is based on either male of female preference (sexual selection) or where there is a hierarchical structure (a harem system). Inbreeding also violates non-random mating and will lead to an excess of homozygotes. Next, it is assumed allele frequencies must be equal between sexes, no sex-biased genes, and that populations are infinitely large and closed. Finally, the population must be free from other evolutionary processes. This assumes that natural selection is absent from the population. Natural selection can lead to the fixation of alleles that are associated with a trait that leads to a change in allele frequencies in the population. If a population can fulfil all the assumptions of HWE, it can be considered a truly neutral population, where allele frequencies between populations can be explained by isolation over time.

Many population genetics studies focus on mitochondrial DNA (mtDNA) or nuclear DNA (nDNA) in HWE. These neutral markers made it possible to delineate dispersal processes present in the environment that were independent of the selective drivers and evolutionary pressures experienced by organisms in their environment. Thus, neutral markers made it possible to observe the empirical allele frequencies with frequencies expected if that population was in HWE. HWE was the basis for statically testing of populations to determine if the frequency of genotypes present is in the expected proportions. For example, population stratification, or significant population structure, may be a cause of a deviation from HWE.

Natural selection is a term used to describe a process where an individual is more likely to survive and therefore reproduce, due to the presence of a gene or trait that provides an advantage to the organism. Thus natural selection may lead to the increased frequency of a gene over time through inheritance through descent. In populations that experience different environmental stressors, the genes that are associated with increased fitness may be different in one location than those in other locations. This can lead to genomic variation within the species. This type of variation is adaptive, or selective, selection. These are non-random processes that can affect genomic variation. Furthermore, these processes may lead to a change in an allele frequency based on its phenotypic effect. This change may be beneficial (positive selection) or disfavoured (purifying selection). In general, most mutations are disfavoured if they occur in protein-coding regions as a single mutation may alter the protein. Finally, balancing selection occurs where alleles that favour a heterozygote are advantage.

Adaptive selection may exist even in the presence of high gene flow. Genes with alleles that are crucial to fitness in one population may be transported into another population wherein they provide a distinct disadvantage. These genes may undergo a swift purifying selection. If population analysis was based solely on adaptive genes, the functional connectivity of the species would be masked. However, the same rhetoric applies to exclusively analysing neutral markers, the effective emigration rate may be lower due to the reduced survival of individuals with these genes. Neutral markers could show low differentiation due to the limited effects of genetic drift in large effective populations and not due to high dispersal capabilities. In such situations, adaptive divergence may occur which may not be reflected at all loci. This warrants further investigation into local adaption of species where traditionally high gene flow was thought to occur. It should also be considered in fine-scale distribution studies of species with low dispersal capabilities.

Next Generation Sequencing and population genomics.

Next-Generation Sequencing (NGS), also referred to as high-throughput sequencing is a suite of modern sequencing techniques that have revolutionised genomics. NGS allows cheaper sequencing of DNA and RNA (von Bubnoff 2008). A popular example of NGS technology is Illumina sequencing. Illumina sequencing produces short-read sequences usually between 100-150 bp. NGS produces large amounts of short reads, between 100-150 bp, creating sequencings that have many overlapping regions. These overlaps are intrinsic to Illumina sequencing and can help reduce biases in sequencing error when assembling genomes due to the reduced error of having a single nucleotide sequenced many times (the depth). In tandem to the advances made in sequencing capabilities, there have been major improvements to the software used and computational power required allowing for more rigorous and complex analyses than were previously available.

NGS can utilise a greater number of genetic markers from across the genome when compared to Sanger methods. RADseq (Restriction-site Associated DNA Sequencing) is a genomic technique that uses enzymes to periodically cut along the genome corresponding to an enzyme-specific restriction site. These fragments may contain variable regions referred to as single nucleotide polymorphisms (SNPs).

The use of RADseq in population genomics was first described in Davey and Blaxter (2010), a natural progression from the seminal study of Baird *et al.* (2008), a study on the use of RADseq to genotype individuals by discovering SNPs. RADseq and other NGS methods are quickly replacing traditional Sanger methods and markers. This is due to the higher number of comparative markers, often one or two orders of magnitude higher than traditional markers. This has revolutionised the field of population genomics, providing higher resolution insight into the gene flow between populations, that was previously masked due to a reduced number of genetic markers available; for example, 12-13 microsatellite markers replaced with 1000s of SNPs. For example, Emerson *et al.* (2010) used SNPs generated from RADseq to investigate the gene flow in the pitcher plant mosquito (*Wyeomyia smithii*). Previously, mtDNA had advocated for a single phylogeographic break. However, by using RADseq four distinct clusters were identified using SNPs. This higher insight was only possible by the large number of variable regions RADseq uses (thousands of SNPs), and thus resolves population structure at levels that microsatellites and mtDNA could not achieve.

The use of RADseq for delineating mollusc population structure

Microsatellite molecular makers have been difficult to develop in molluscs for reasons that are not fully understood, although some research indications that it is the presence of large numbers of transposable elements in their genome (McInerney *et al.* 2011). These transposable elements may play an important role in the reduction of inter-specific genome variation and prevention of microsatellite development. Next-generation sequencing (NGS) overcomes this issue.

RADseq has been used to delineate fine-scale population structure in molluscs. For example, Vendrami *et al*. (2017) conducted RADseq on the King Scallop (*Pectin Maximus*) around Northern Ireland. That study used double digest RADseq (ddRAD), a technical variant of RADseq, that uses two enzymes to simultaneously reduce the random shearing associated with RADseq (Peterson *et al.* 2012). Using this method 10,539 loci each containing a single biallelic SNP were retained. That study also compared their results with population structure inferred from 13 microsatellites. The microsatellite analysis identified weak population differentiation through pairwise F_{st} estimates. The microsatellite data advocated for the presence of a purely panmictic population in both Structure (Pritchard *et al.* 2000), a Bayesian clustering method, and by Principal Component Analysis (PCA). However, population structure analysis using SNPs found evidence of two distinct clusters. One cluster from Mulroy Bay, the other cluster contained eight other locations from the Northern Ireland and into the Irish Sea.

Another study used RADseq to reveal fine-scale population structure in Atlantic Deep Scallop (*Placopecten magellanicus*) on the coast of Newfoundland (Van Wyngaarden *et al.* 2017). That study analysed 245 individuals, from 12 locations using 7163 SNPs. They found evidence for two distinct clusters of populations; those North of Nova Scotia and those South of it.

The above studies demonstrated the applications of RADseq for the delineation of population structure over short geographic distances and additionally addressed the complications induced by microsatellite markers in molluscs. However, while NGS produced thousands of genetic markers, populations with large effective sizes or very high migration rates over small spatial scales may produce what appeared to be a panmictic population when neutral markers were exclusively used (Gagnaire *et al.* 2015).

Identifying potential loci under selection

As previously discussed, NGS can produce thousands of genetic markers. With a suite of markers potentially two to three orders of magnitude higher than traditional genetics it is possible to identify a panel of SNPs that fail to display HWE allele frequencies. Traditionally, these loci would have been removed from the population analysis. However, just analysing the neutral markers in the genome can mask the true population structure present by not 7 accounting for evolutionary forces such as adaption which can give rise to genomic variation. This variation may provide insight into the local forces that could potentially be causing populations to diverge even in the presence of high gene flow. To combat this problem, methods have been developed to use genetic markers that are outside of HWE and thus under the influence of selection. These loci are considered “outlier loci”.

Presently, there are a few methods to determine if loci are under selective forces. These methods are collectively referred to as genome scans as they rely on having a large panel of genetic markers that span the genome. One genome-scan method, and perhaps the most popular is to examine the locus-specific population Fst (Foll and Gaggiotti 2008). Fst is a fixation index which looks at the allele frequencies within a sample and compares that to the fixation of the allele across the whole sampling effort. The principle is that loci that are candidates for selection will have a higher or lower value of Fst than the Fst of neutral loci (Beaumont and Balding 2004). If the Fst is greater than the mean, this locus may be a candidate for adaptive selection. Those that fall below the mean may be experiencing purifying selection due to the decreased frequency at which they occur (Beaumont and Balding 2004).

Narum and Hess (2011) reviewed the most commonly applied software for the detection of outlier loci. One of the primary problems identified was the occurrence of false discoveries. False discoveries are, as suggested, an occurrence where the result is unreliable due to Type I (false positives) or Type II (false negatives) errors. In Narum and Hess (2011) BayeScan (Foll and Gaggiotti 2008) was identified to produce the fewest Type I and Type II false discoveries. In BayeScan, the user can specify a false discovery rate (FDR) to control the level of false positives. Commonly, FDR is set between 1-10%. The lower the FDR, the more conservative the identification of potential outliers. BayeScan uses a Bayesian approach to detect Fst outliers. It applies two components to its analysis, a population-specific Fst coefficient (beta), and a locus-specific Fst coefficient (alpha) shared by all populations. The software then implements a logistic regression and rejects a model of neutrality if the locus-specific component explains the observed variation. Thus, this method provides two models to explain the diversity at each locus. Subsequently, for each locus, BayeScan calculates a posterior probability, which is a statistical probability that with the currently available information, the hypothesis under examination is true. Due to the multiple testing implemented in BayeScan, a multiple testing correction is applied as q-values. The given q-values can then be compared to the FDR, where the user should select potential outliers with a q-value lower than the FDR.

In model organisms, the loci under selection can be mapped and attributed to a region of the genome. The mapped areas can provide insight into the underlying cause of the selection at the candidate loci. However, in non-model organisms without an annotated reference genome, loci putatively under selection may be identified but cannot be mapped to a region of the genome. Therefore, our understanding of the driving force of selection or the phenotype expressed in response to it is limited.

Buccinum undatum: a species summary

Ecology and reproduction

Buccinum undatum (Linnaeus 1798) (Figure 1) is a neogastropod that is widely distributed across the North Atlantic. In the Northeast Atlantic, its range extends from the Bay of Biscay to Iceland and it inhabits similar latitudes from Greenland to Canada. *Buccinum undatum* is the largest edible marine gastropod in the North Atlantic, with a maximum reported shell height of 150 mm (Nasution and Roberts 2004). It feeds primarily on molluscs, such as bivalves and barnacles, as well as small crustacea, but is also reported as a scavenger (Nielsen 1974; Taylor and Taylor 1977). It is a subtidal gastropod with a disputed maximum depth. Thomas and Himmelman (1988) indicated that 200 m is the maximum depth this species lives at, however Valentinsson *et al.* (1999) have reported, at least in Swedish waters, 600 m and, finally, Nielsen (1974) determines its depth as greater than 1000 m. Furthermore, Warén and Bouchet (2001) describe *B. undatum* found between 500 -1000 m as having specialised pigmented eyes that indicate it is not an occasional visitor to deep-water habitats. At the very least, *B. undatum* has a large depth range at which it lives, which extends from the subtidal to the deep-sea.

Buccinum undatum are sexually dimorphic. Females have higher shells and aperture length, a larger shell weight, and heavier tissue weight (Kenchington and Glass 1998). *Buccinum undatum* individuals reach sexual maturity at ~7 cm, with males achieving maturity at slightly smaller sizes than females (Martel *et al.* 1986). However, this is thought to vary widely by population (Haig *et al.* 2015).

The breeding season of *B. undatum* exhibits regional differences. In European waters, reproduction occurs October to February (Kideys *et al*. 1993) and in The Gulf of St

Lawrence, between Newfoundland and mainland Canada, it is reported to be from July to October (Martel *et al.* 1986). *Buccinum undatum* has a low fecundity, a direct-development reproductive strategy, no planktotrophic stage, and a sedentary lifestyle. Fertilisation is internal and female *B. undatum* deposits eggs in large masses on hard substrates (Figure 1) such as rocks and shells. These masses contain both embryonic eggs and nurse capsules. The larvae developed inside the embryonic eggs and consume the nurse capsules. These eggs hatch 3 to 8 months after being deposited (Valentinsson 2002).

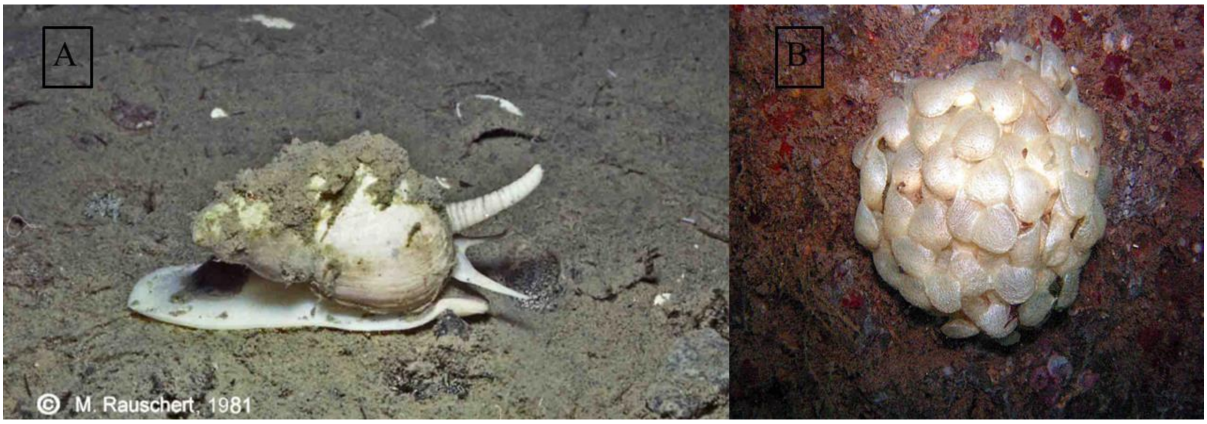

Figure 1.1 (A) Buccinum undatum *(Image credit: M. Rauschert, 1981. Available at: http://www.marinespecies.org/aphia.php?p=image&tid=138878&pic=10327) (B) a* Buccinumm undatumm *egg mass. (Image credit: Paul Newland. Available at (https://www.marlin.ac.uk/assets/images/marlin/species/web/o_bucund8.jpg)*

The early ontogeny of *B. undatum* in UK waters was investigated by Smith *et al.* (2013) where it was found that development was still successful at up to 18 ºC. The rate of development increased with increased temperature, however, the number of successful developments per egg mass decreased. This observed level of reduced reproductive success suggests that *B. undatum* may have reduced populations in shallow waters under future global warming scenarios. Furthermore, increased water temperatures may result in a range shift in the species as the UK populations are winter spawners and egg masses are laid when water is at its coolest (6 to 10 ºC). Increased sea temperatures have been proven to cause females to forgo reproduction (Smith *et al.* 2013). Thatje *et al.* (2019) provides anecdotal evidence that in warmer years a smaller number of egg capsules of *B. undatum* were collected. That study speculated that at the southern limit of *B. undatum's range*, increased temperatures may have a significant effect on reproductive output.

Fisheries of *Buccinum undatum*

In 1922, *B. undatum* first appeared in the UK sea fisheries annual statistics. In that year 33096 cwt. (1676 t) were landed in England and Wales. In, 2013 the total recorded landings for the UK were 13,700 tonnes valued at £9.1 m (9.6% the value of all shellfish fisheries) (Marine Management Organisation 2014). By 2016, 22,600 tonnes were being landed and the value of the fishery to the economy jumped to £22.9 m (7.2% of total shellfish fisheries) (Marine Management Organisation 2016). The most recent landing information for whelks is 2017, where 20,800 tonnes were landed in the UK. The valuation of whelks to the UK fisheries sector in 2017 was £22.7 m (6.6% of the income of shellfish fisheries) (Marine Management Organisation 2017).

Buccinum undatum is a non-quota species in the EU and so there are few restrictions regarding the total allowable catch. For this reason, it is seen as a displacement fishery as fishermen move away from more tightly regulated stocks (McIntyre *et al.* 2015) The lack of appropriate management for the whelk fishery has led to concerns about its sustainability (Nicholson and Evans 1997; McIntyre *et al.* 2015; Shrives *et al.* 2015). In 1997, there was already discussion about the overfishing of *B. undatum*, especially in Southeast England (Nicholson and Evans 1997). However formal stock assessments are not currently undertaken for *B. undatum*. In Fahy *et al*. (2005), it was concluded that the fisheries off Ireland suffered a major collapse in 2004. This collapse was thought to have occurred after two successive years of increased recruitment was followed by increased fishing effort. This highlighted the need for better management of *B. undatum* across the British Isles.

The minimum landing size (MLS) is the smallest size at which the government deems it safe to remove an individual from a stock without serious negative consequences such as stock collapse. The European Union designated MLS for *B. undatum* is at 45 mm shell height. Lawler (2014) reported on the large variation in MLS over a regional scale in the United Kingdom. For example, in Wales, the MLS is 55 mm, 70 mm in the Isle of Wight, and in the Shetland Islands, it is 75 mm. They also reported, is the large spatial variation in size at maturity between sites (44.8 mm in the Solent to 76.2 mm in the southern North Sea (Lawler 2014)). This large spatial variation in key life-history parameters demonstrates the need to manage whelks on a regional or local level.

There is evidence that *B. undatum* has the potential to form distinct populations over small spatial scales (Weetman *et al.* 2006; Mariani *et al.* 2012; Pálsson *et al.* 2014) (Refer to Section

4.3) Inshore Fisheries and Conservation Authorities (IFCAs) manage the exploitation of stocks within the 6nm from the Territorial Sea baselines. Within this 6 nm perimeter of England there exist ten IFCAs who each regionally manage the exploitation of inshore fisheries. Regarding whelks, this has led to regional differences in how MLS is enforced by the IFCAs. This regional approach is more aligned with scientific evidence and advice due to spatial variability in key life-history parameters (see above). IFCAs also have the authority to manage the number of baited pots per person/per vessel being set in the respective districts to ensure sustainable catch.

Population structure and connectivity of *B. undatum*

The life-history characteristics exhibited by *B. undatum*, direct development of larvae and a sedentary lifestyle, imply that the level of population structure should be evident, even at small spatial scales. These life-history strategies commonly make species vulnerable to overfishing and once impacted, recolonisation and recovery would be slow (McIntyre *et al.* 2015). Only three studies have focused on the gene flow and population structure of *B. undatum.* All used the same five microsatellites created for the species by Weetman *et al*. (2005).

Weetman *et al.* (2006) examined the population structure of *B undatum* at both the macro and microgeographic scales; this study encompassed much of *B. undatum*'s range including northern France, England, Scotland, Iceland, and Canada. By using a multidimensional scaling (MDS) and cluster analysis this study elucidated macrogeographic trends. Four distinct clusters were observed; The Icelandic, the Canadian, the Swedish, and the rest (containing samples from the European continental shelf). The Icelandic and Canadian were distinct clusters due to their isolation by distance. The Canadian and Icelandic samples contained markedly lower genetic diversity, usually indicative of a bottleneck. However, a definitive conclusion was impossible due to the low number of loci tested which provided low genomic resolution. The Swedish population was thought to be differentiated due to its low population density and its separation from the rest of the locations by the deep waters of the Norwegian trench and the Skaggerek. Individuals from the Solent – the strait separating mainland England from the Isle of White – formed a distinct genetic cluster. However, this was only supported by major shifts in the frequency of two alleles at one locus. Thus, the removal of this locus from the analysis collapsed the Solent group and it assimilated into the Shelf group This may be more reflective of selection acting on the whelks in this location. There were low global Fst values (0.011 ± 0.003 to 0.024 ± 0.007) which indicated low levels of differentiation. However, there was significant Isolation by Distance (IBD) along the British North Sea. The pairwise significant values ranged in geographic distance from 70-650 km. This would suggest that limited dispersal capabilities are more important to elucidating genetic structure rather than historical connectivity. Also, the low Fst value may be a result of rare long-distance or dispersal in a semi-continuous population.

On a microgeographic scale, Weetman *et al.* (2006) sampled multiple sites within three areas along the west coast of the UK. Areas had a pairwise geographic distance between 200-620 km while the pairwise distances between sites was 7-40 km. Statistical analysis suggested that genetic variation was more heavily partitioned between sampling sites rather than areas and thus that small-scale processes were more important to the overall genetic structure. Within each area, the most inshore site had the lowest genetic diversity and the highest level of differentiation, with asymmetrical migration into more offshore populations evident. Weetman *et al.* (2006) hypothesises that this may be a density-dependent reaction and that whelk population densities are greatest inshore, and this promotes emigration to areas with lower densities to reduce competition.

Pálsson *et al.* (2014) conducted a study on *B. undatum* from Britain to Iceland, and in Greenland, using the same microsatellite markers as above and mitochondrial DNA (mtDNA) genes, CO1 and 16S. In Iceland, there was significant genetic variation over the tens of kilometres using pairwise microsatellite differences. This was supported by the mtDNA differences which showed that there was also significant pairwise Fst values over small spatial scales. Pálsson *et al*. (2014) discussed the two sampling locations – Faroe Islands and British Channel (Isle of Wight) – as representative of Britain, although the sampling was far less than the more extensive study conducted by Weetman *et al.* (2006). MDS plots based on genetic distance using microsatellites showed that the British channel and Faroe Island clustered together. This agrees with Weetman *et al.* (2006) with low levels of Fst on the British North Sea coat. However, Pálsson *et al.* (2014) does not acknowledge that this may be an artefact of a semi-continuous population and there is in fact significant isolation by distance along the British North Sea Coast as previously reported by Weetman *et al*. (2006). 13

Pálsson *et al.* (2014) also found that genetic differentiation was greatest in nearshore inlets and genetic diversity was highest in deeper waters further offshore. This was in agreement with the findings of Weetman *et al.* (2006) on asymmetric migration from inshore to offshore populations. However, Pálsson *et al.* (2014), indicated that the direction of migration in that study was unknown. Pálsson *et al.* (2014) indicated limited evidence for a genetic bottleneck in southern England. Instead, that study inferred a stable population during the Last Glacial Maximum (LGM) as seen from the nucleotide mismatch analysis, the high nucleotide diversity, and low haplotype diversity observed in the British Channel using mtDNA. Further evidence for a stable population during the LGM was that Britain had the highest microsatellite variation in the study. Although this may have been a result of genetic admixture. The British Channel only opened after the LGM and may have been colonised from multiple populations. Alternatively, there may have been a glacial refugia present. Hurd's Deep is a deep trench located in the west British Channel. Studies on the red algae *Palmaria palmata* have revealed that there is genetic evidence suggesting the presence of a glacial refugia in the area (Provan *et al*. 2005). This glacial refugia may have been inhabited by *B. undatum*, reflected by the high genetic diversity seen in the British Channel *for B. undatum* This species has a eurybathic distribution (subtidal to greater than 1000 m), and historically evolved in the deep sea (Smith *et al.* 2013), which may have allowed for a persistent population in deeper waters in the refugia. Evidence of this adaption can also see in the specialised pigmented eyes in *B. undatum* sampled in deeper waters (Warén and Bouchet 2001). These historical adaptions may have allowed populations to remain stable. This large level of genetic diversity at Hurd's Deep has also been reported for *Raja clavata* (Thornback rays) (Chevolot *et al.* 2006). However, that study could not definitively attribute this to a refugia at Hurd's Deep. Hurd's Deep, therefore, could represent a major cryptic glacial refugia and should, therefore, be of increased interest to population genomic studies and phylogeographic studies in the future.

The inference of a stable population during the LGM was not in agreement with *Weetman et al*. (2006), who suggested that the British North Sea Coast would be under recent population expansion trends after a recent bottleneck. This is a common trend of marine invertebrate to expand after an Ice Age (Vermeij 2005). However, Weetman *et al*. (2006) did not explicitly provide data for his assumption that *B. undatum* followed the same pattern – only mtDNA would have given an evolutionary history deep enough for that inference. Weetman *et al*. (2006) inferred that population bottlenecks present *B. undatum* may have been anthropogenically induced and not a response to historic climatic conditions. Potential causes for this bottleneck included overfishing (see above) and Tributyltin (TBT) poisoning. TBT is a bioaccumulating toxin that was a popular biocide in the 1960s and caused imposex in gastropods. Imposex causes female gastropods to develop male sex organs. Nicholson and Evans (1997) quantitatively measured imposex in *B. undatum* in Europe. In this study, the Solent was the most heavily impacted region of 26 European sampling points. The Solent was also one of the regions Weetman *et al.* (2006) found evidence for a bottleneck.

However, there is evidence to contradict this theory of anthropogenically induced bottlenecks;

Hallers-Tjabbes *et al*. (1994) investigated levels of imposex in *B. undatum* due to TBT in the British North Sea and found high levels of imposex where Weetman *et al.* (2006) had not reported evidence of a bottleneck. This may have been due to the poor level of genomic resolution present in that study. As previously mentioned only five microsatellites were used thus any inference made about the presence of a bottleneck were to be interpreted cautiously.

Mariani *et al.* (2012) focused on the population connectivity of *B. undatum* populations on the east coast of Ireland in the Irish Sea. The molecular markers used were the same microsatellites created by Weetman *et al.* (2005). Mariani *et al.* (2012) found larger levels of population structure when compared to (Weetman *et al*. 2006) due to the presence of known oceanographic breaks. There were three large geographic discontinuities identified which were associated with population breaks, where populations on either side had significant population structure despite their close geographic proximity. These breaks were caused by oceanographic features present in the Irish Sea, e.g. the front between the Irish Sea and the Celtic Sea is a known break. This ocean front results in high velocities of 30 cm s-1 running parallel to the isopycnal contours separating the two seas (Simpson 1976) providing a strong barrier to dispersal. Overall, Mariani *et al.* (2012) stated that differentiation in the Irish sea was due to an Isolation by Distance (IBD) model. This was consistent with the findings of Weetman *et al.* (2006) and Pálsson *et al.* (2014). Mariani *et al.* (2012) found sampling locations that are geographically near to each other exhibit non-significant levels of population structure. Thus, the IBD model indicated that there were low levels of demographic connectivity and, moreover, there was a stepping-stone model of gene flow. This had been reported by Weetman *et al.* (2006) in the form of a semi-continuous population. Mariani *et al.* (2012) had a sampling location in an inlet that was significantly differentiated from its nearest sampling point despite it only being 30 km away. This finding was supported by both Weetman *et al.* (2006) and Pálsson *et al.* (2014). All three studies found increased levels of population isolation from samples that originated in an inlet rather than the ocean shelf. Perhaps there are barriers to dispersal present within these inlets and that dispersal is only capable through asymmetric migration as adults to populations further offshore.

Mariani *et al.* (2012) found no evidence of genetic bottlenecks in *B. undatum* in the Irish sea. Power and Keegan (2001) investigated TBT poisoning in a range of gastropods including *B. undatum* in the Irish Sea. *Buccinum undatum* accumulated TBT in their tissue at higher rates than the other gastropods, yet the conclusions of the study were that *B. undatum* was not significantly impacted by TBT compared to other gastropods. There was no correlation between the levels of TBT and the reported incidences of imposex in that study. While the levels of TBT varied between locations, the authors of that study used a sampling location ten kilometres offshore beside a port in Ireland heavy historical marine traffic, as an area heavily affected by TBT. This region is also within the *B. undatum* fisheries district in Ireland. This result contradicted the hypothesis proposed by Weetman *et al.* (2006) that TBT had affected the effective population size and led to the reported bottlenecks in that study.

All of these studies focused on the use of the same microsatellite's markers – although Pálsson *et al*. (2014) used additional mtDNA to elucidate phylogeographic processes. All the currently published studies agreed that *B. undatum* can be genetically differentiated over tens of kilometres and that inshore populations are more divergent than populations further offshore. These studies suggested a stepping-stone model of gene flow, which perhaps does not include the inshore populations explaining this differentiation. Evidence for population bottlenecks has been put forward by both Weetman *et al.* (2006) and Mariani *et al.* (2012) for the UK. However, Pálsson *et al*. (2014) disagrees with that assumption, and instead postulates that southern England populations have persisted during the LGM due to the high haplotype diversity in Britian. Evidence for a glacial refugia in the English Channel at Hurd's Deep has been put forward in Provan *et al.* (2005), and potentially Chevolot *et al*. (2006). While TBT has been put forward as a cause of a modern genetic bottleneck, findings between studies about the location of bottlenecks, and the correlation between TBT and the rate of imposex incident has been inconsistent

Aims and objectives

The overall aim of this study was to evaluate the population connectivity of *B. undatum* in the Southern North Sea and English Channel using genetic markers (SNPs) generated from ddRAD. This data will be the first to use NGS on *B. undatum*. This current study will provide further insight into the regional population structure of the species, building on the work of Weetman *et al.* (2006), Mariani *et al.* (2012), and Pálsson *et al.* (2014).

The objectives were:

- To investigate the population structure of *B. undatum* in the southern North Sea, English Channel and Irish Sea;
- To compare population structure derived from using SNPs with previously published studies on *B. undatum* that used microsatellites; and
- To investigate the presence of adaptive selection on the genome and to determine its contribution to population structure.

References

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A. (2008) 'Rapid SNP discovery and genetic mapping using sequenced RAD markers', PloS one, 3(10), e3376–e3376.

Beaumont, M.A., Balding, D.J. (2004) 'Identifying adaptive genetic divergence among populations from genome scans', Molecular Ecology, 13(4), 969–980.

von Bubnoff, A. (2008) 'Next-Generation Sequencing: The Race Is On', Cell, 132(5), 721–723.

Chevolot, M., Hoarau, G., Rijnsdorp, A.D., Stam, W.T., Olsen, J.L. (2006) 'Phylogeography and population structure of thornback rays (Raja clavata L., Rajidae)', Molecular Ecology, 15(12), 3693–3705.

Chust, G., Villarino, E., Chenuil, A., Irigoien, X., Bizsel, N., Bode, A., Broms, C., Claus, S., Fernández de Puelles, M.L., Fonda-Umani, S., Hoarau, G., Mazzocchi, M.G., Mozetič, P., Vandepitte, L., Veríssimo, H., Zervoudaki, S., Borja, A. (2016) 'Dispersal similarly shapes both population genetics and community patterns in the marine realm', Scientific Reports, 6, 28730.

Cowen, R.K., Sponaugle, S. (2009) 'Larval Dispersal and Marine Population Connectivity', Annual Review of Marine Science, 1(1), 443–466.

Davey, J.W., Blaxter, M.L. (2010) 'RADSeq: next-generation population genetics', Briefings in functional genomics, 9(5–6), 416–423.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., Holzapfel, C.M. (2010) 'Resolving postglacial phylogeography using high-throughput sequencing', Proceedings of the National Academy of Sciences of the United States of America, 107(37), 16196–16200.

Fahy, E., Carroll, J., Hother-Parkes, L., O'Toole, M., Barry, C. (2005) Fishery Associated Changes in the Whelk *Buccinum Undatum* Stock in the Southwest Irish Sea, 1995-2003, Marine Institute.

Foll, M., Gaggiotti, O. (2008) 'A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective', Genetics, 180(2), 977–993.

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(8), 769–786.

Haig, J.A., Pantin, J.R., Salomonsen, H., Murray, L.G., Kaiser, M.J. (2015) 'Temporal and spatial variation in size at maturity of the common whelk (*Buccinum undatum*)', ICES Journal of Marine Science, 72(9), 2707–2719.

Hallers-Tjabbes, C.C.T., Kemp, J.F., Boon, J.P. (1994) 'Imposex in whelks (*Buccinum undatum*) from the open North Sea: Relation to shipping traffic intensities', Marine Pollution Bulletin, 28(5), 311–313. 18

Hauser, L., Carvalho, G. (2008) 'Paradigm shifts in marine fisheries genetics: Ugly hypotheses slain by beautiful facts', Fish and Fisheries, 9, 333–362.

Kenchington, E., Glass, A. (1998) Local Adaptation and Sexual Dimorphism in the Waved Whelk (*Buccinu t Undatum*) in Atlantic Nova Scotia with Applications to Fisheries Management, Canadian Technical Report of Fisheries and Aquatic Science No. 2237.

Kideys, A.E., Nash, R.D.M., Hartnoll, R.G. (1993) 'Reproductive cycle and energetic cost of reproduction of the neogastropod *Buccinum undatum* L. in the Irish Sea.', Journal of the Marine Biological Association of United Kingdom, 73(2), 391–403.

Kyle, C.J., Boulding, E.G. (2000) 'Comparative population genetic structure of marine gastropods (Littorina spp.) with and without pelagic larval dispersal', Marine Biology, 137(5), 835–845.

Lawler, A. (2014) Determination of the Size of Maturity of the Whelk *Buccinum Undatum* in English Waters, Defra.

Leis, J.M. (2010) 'Behaviour of fish larvae as an essential input for modelling larval dispersal: behaviour, biogeography, hydrodynamics ontogeny , physiology and phylogeny meet hydrography', Marine Ecology Progress Series, 347, 185–193.

Mariani, S., Peijnenburg, K.T., Weetman, D. (2012) 'Independence of neutral and adaptive divergence in a low dispersal marine mollusc', Marine Ecology Progress Series, 446, 173–187.

Marine Management Organisation (2014) UK Sea Fisheries Annual Statistics Report 2013, Radford.

Marine Management Organisation (2016) UK Sea Fisheries Annual Statistics Report 2015, Newport.

Marine Management Organisation (2017) UK Sea Fisheries Annual Statistics Report 2016, London.

Martel, A., Larrivée, D.H., Klein, K.R., Himmelman, J.H. (1986) 'Reproductive cycle and seasonal feeding activity of the neogastropod *Buccinum undatum*', Marine Biology: International Journal on Life in Oceans and Coastal Waters, 92(2), 211–221.

McInerney, C.E., Allcock, A.L., Johnson, M.P., Bailie, D.A., Prodöhl, P.A. (2011) 'Comparative genomic analysis reveals species-dependent complexities that explain difficulties with microsatellite marker development in molluscs', Heredity, 106(1), 78.

McIntyre, R., Lawler, A., Masefield, R. (2015) 'Size of maturity of the common whelk, *Buccinum undatum*: Is the minimum landing size in England too low?', Fisheries Research, 162, 53–57.

Molnar, J.L., Gamboa, R.L., Revenga, C., Spalding, M.D. (2008) 'Assessing the global threat of invasive species to marine biodiversity', Frontiers in Ecology and the Environment, 6(9), 485–492.

Narum, S.R., Hess, J.E. (2011) 'Comparison of FST outlier tests for SNP loci under selection', Molecular Ecology Resources, 11(SUPPL. 1), 184–194. 19

Nasution, S., Roberts, D. (2004) 'Laboratory trials on the effects of different diets on growth and survival of the common whelk, *Buccinum undatum* L. 1758, as a candidate species for aquaculture', Aquaculture International, 12(6), 509–521.

Nicholson, G.J., Evans, S.M. (1997) 'Anthropogenic impacts on the stocks of the common whelk *Buccinum undatum* (L.)', Marine Environmental Research, 44(3), 305–314.

Nielsen, C. (1974) 'Observations on *Buccinum undatum* L. attacking bivalves and on prey responses, with a short review on attack methods of other prosobranchs', Ophelia, 13(1–2), 87–108.

Pálsson, S., Magnúsdóttir, H., Reynisdóttir, S., Jónsson, Z.O., Örnólfsdóttir, E.B. (2014) 'Divergence and molecular variation in common whelk *Buccinum undatum* (Gastropoda: Buccinidae) in Iceland: A trans-Atlantic comparison', Biological Journal of the Linnean Society, 111(1), 145–159.

Pascual, M., Rives, B., Schunter, C., Macpherson, E. (2017) 'Impact of life history traits on gene flow: A multispecies systematic review across oceanographic barriers in the Mediterranean Sea', PloS one, 12(5), e0176419–e0176419.

Pechenik, J.A. (1979) 'Role of Encapsulation in Invertebrate Life Histories', The American Naturalist, 114(6), 859–870.

Pechenik, J.A. (1982) 'Ability of some gastropod egg capsules to protect against low-salinity stress’, Journal of Experimental Marine Biology and Ecology, 63(3), 195–208.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E. (2012) ‘Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species’, PLOS ONE, 7(5), 1–11.

Power, A.J., Keegan, B.F. (2001) ‘The Significance of Imposex Levels and TBT Contamination in the Red Whelk, Neptunea antiqua (L.) from the Offshore Irish Sea’, Marine Pollution Bulletin, 42(9), 761–772.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000) ‘Inference of population structure using multilocus genotype data’, Genetics, 155(2), 945–959.

Provan, J., Wattier, R.A., Maggs, C.A. (2005) ‘Phylogeographic analysis of the red seaweed Palmaria palmata reveals a Pleistocene marine glacial refugium in the English Channel’, Molecular Ecology, 14(3), 793–803.

Rawlings, T.A. (1999) ‘Adaptations to Physical Stresses in the Intertidal Zone: The Egg Capsules of Neogastropod Molluscs’, American Zoologist, 39(2), 230–243.

Riginos, C., Liggins, L. (2013) ‘Seascape Genetics: Populations, Individuals, and Genes Marooned and Adrift’, Geography Compass, 7(3), 197–216.

Scheltema, R.S. (1986) ‘On dispersal and planktonic larvae of benthic invertebrates: an eclectic overview and summary of problems’, Bulletin of Marine Science, 39(2), 290–322.

Shrives, J.P., Pickup, S.E., Morel, G.M. (2015) ‘Whelk (*Buccinum undatum* L.) stocks around the Island of Jersey, Channel Islands: Reassessment and implications for sustainable management’, Fisheries research, 167, 236–242.

Simpson, J.H. (1976) ‘A boundary front in the summer regime of the Celtic Sea’, Estuarine and Coastal Marine Science, 4(1), 71–81. 20

Smith, K.E., Thatje, S., Hauton, C. (2013) ‘Thermal tolerance during early ontogeny in the common whelk *Buccinum undatum* (Linnaeus 1785): Bioenergetics, nurse egg partitioning and developmental success’, Journal of Sea Research, 79, 32–39.

Taylor, J.D., Taylor, C.N. (1977) ‘Latitudinal Distribution of Predatory Gastropods on the Eastern Atlantic Shelf’, Journal of Biogeography, 4(1), 73–81.

Thatje, S., Dunbar, C.G., Smith, K.E. (2019) ‘Temperature-driven inter-annual variability in reproductive investment in the common whelk *Buccinum undatum’*, Journal of Sea Research, 148, 17–22.

Thomas, M.L.H., Himmelman, J.H. (1988) ‘Influence of predation on shell morphology of *Buccinum undatum* L. on Atlantic coast of Canada’, Journal of Experimental Marine Biology and Ecology, 115(3), 221–236.

Valentinsson, D. (2002) ‘Reproductive cycle and maternal effects on offspring size and number in the neogastropod *Buccinum undatum* (L.)’, Marine Biology, 140(6), 1139–1147.

Valentinsson, D., Sjödin, F., R Jonsson, P., Nilsson, P., Wheatley, C. (1999) Appraisal of the Potential for a Future Fishery on Whelks (*Buccinum Undatum*) in Swedish Waters: CPUE and Biological Aspects, Fisheries Research.

Vendrami, D.L.J., Telesca, L., Weigand, H., Weiss, M., Fawcett, K., Lehman, K., Clark, M.S., Leese, F., McMinn, C., Moore, H., Hoffman, J.I. (2017) ‘RAD sequencing resolves fine-scale population structure in a benthic invertebrate: Implications for understanding phenotypic plasticity’, Royal Society Open Science, 4(2).

Vermeij, G.J. (2005) ‘From Europe to America: Pliocene to Recent trans-Atlantic expansion of cold-water North Atlantic molluscs’, Proceedings of the Royal Society B: Biological Sciences, 272(1580), 2545–2550.

Wangensteen, O.S., Turon, X., Palacín, C. (2017) ‘Reproductive strategies in marine invertebrates and the structuring of marine animal forests’, in Marine Animal Forests: The Ecology of Benthic Biodiversity Hotspots, 571–594.

Warén, A., Bouchet, P. (2001) ‘Gastropoda and Monoplacophora from hydrothermal vents and seeps; New taxa and records’, Veliger, 44, 116–231.

Weetman, D., Hauser, L., Bayes, M.K., Ellis, J.R., Shaw, P.W. (2006) ‘Genetic population structure across a range of geographic scales in the commercially exploited marine gastropod *Buccinum undatum’*, Marine Ecology Progress Series, 317, 157–169.

Weetman, D., Hauser, L., Shaw, P.W., Bayes, M.K. (2005) ‘Microsatellite markers for the whelk *Buccinum undatum’*, Molecular Ecology Notes, 5(2), 361–362.

Van Wyngaarden, M., Snelgrove, P.V.R., DiBacco, C., Hamilton, L.C., Rodríguez-Ezpeleta, N., Jeffery, N.W., Stanley, R.R.E., Bradbury, I.R. (2017) ‘Identifying patterns of dispersal, connectivity and selection in the sea scallop, Placopecten magellanicus, using RAD seq-derived SNP s’, Evolutionary applications, 10(1), 102–117. 21

2. Chapter 2: ddRAD reveals high levels of gene flow and no population structure in a marine gastropod with limited dispersal capabilities

Declan Morrissey[1*], Jake Goodall[2], Rita Castilho[1,3], Tom Cameron[4], and Michelle Taylor[4]

[1] Universidade do Algarve, Faro, Portugal.

[2] Faculty of Life and Environmental Sciences, University of Iceland, Reykjavík, Iceland

[3] Center of Marine Sciences (CCMAR), University of Algarve, Faro, Portugal

[4] School of Life Sciences, University of Essex, Colchester, UK

*Corresponding author: declanmorrissey4@gmail.com

Abstract

Population genomics is important for understanding the degree of genetic connectivity and effective dispersal over geographic distances. Connectivity, or the constraint of it, influences both local and regional biodiversity and thus is of primary interest to both evolutionary and ecological studies. In recent years, previous assumptions regarding dispersal capabilities, and their function as a primary driver of expected genetic structure of populations have been challenged by Next-Generation Sequencing techniques. This study investigated the population connectivity of *Buccinum undatum* with Single Nucleotide Polymorphisms (SNPs) derived from double-digest Restriction associated DNA sequencing. In total, 191 individuals were sequenced from the Southern North Sea, English Channel, and Irish Sea, a geographic scope of 1165 km. After strict quality control and filtering, 885 biallelic SNPs and 141 individuals were retained. Outlier detection revealed 4 loci under putatively positive selection. Two datasets were analysed; a neutral loci dataset which contained 881 SNPS, and an outlier loci dataset that contained the 4 SNPs identified as outliers. Results from the neutral dataset advocated for a single large population with no overall structure but significant sub-structure. However, sub-structure was much less frequent than previously reported for the species. Individuals sampled within a bay were not more genetically differentiated than those outside of bay, a previously reported trait of *B. undatum*. There was significant isolation by distance observed across the majority of the geographic range. Outlier analysis did not reveal any hidden population structure, nor any isolation by distance. Overall, results presented within fundamentally agreed with previous studies that *B. undatum* consists of a single population, that is semi-continuous in nature, with sub-structure present.

However, high gene flow and a large effective population size supressed overall population divergence.

Keywords: ddRAD, Mollusc, Population genomics, Population structure, Whelk

Introduction

Dispersal is a key process that affects population growth, gene flow, and overall population persistence. For this reason, it is an important parameter to consider when discussing the evolution of species in natural systems. However, dispersal is difficult to quantify directly in the marine environment as many marine species disperse as larvae (Cowen and Sponaugle 2009). The small size of these larvae makes direct observation or mark and recapture experiments both difficult and expensive. For this reason, genetics has been used to measure dispersal due to its effect on gene flow and population structure. Genetic approaches have refuted the traditional view that the majority of marine species have panmictic populations over large spatial scales due to the lack of obvious barriers to dispersal (Hauser and Carvalho 2008). Furthermore, old tropes of associating species with long pelagic larval stages with limited population structure have been challenged (Selkoe and Toonen 2011, and references within) and while some species do exhibit spatial structure based on reproductive life-history strategies (e.g. Kyle and Boulding 2000), a species by species approach is needed.

Buccinum undatum (Linnaeus 1798) is a neogastropod widely distributed across the North Atlantic. It feeds primarily on molluscs and small crustacea (Nielsen 1974; Taylor and Taylor 1977). It is subtidal with a maximum depth range greater than 1000 m Nielsen (1974). The breeding season of *B. undatum* exhibits regional differences. In European waters, reproduction occurs from October to February. *Buccinum undatum* has a low fecundity, a direct-development reproductive strategy, no planktotrophic stage, and a sedentary lifestyle.

Buccinum undatum is a non-quota species in the EU and so there are few restrictions regarding the total allowable catch. For this reason, it is seen as a displacement fishery as fishermen move away from more tightly regulated stocks (McIntyre *et al.* 2015) The lack of appropriate management for the whelk fishery has led to concerns about its sustainability (Nicholson and Evans 1997; McIntyre *et al.* 2015; Shrives *et al.* 2015). In 1997 there were concerns that there was overfishing of *B. undatum*, especially in Southeast England (Nicholson and Evans 1997). However formal stock assessments are not currently undertaken for *B. undatum. Whelks are the currently the sixth largest* in the UK and in 2017 20,800 t were landed valued at £22.7 m (6.6% of the income of shellfish fisheries) (Marine Management Organisation 2017).

Despite its commercial importance, only three studies have focused on the gene flow and population structure of *B. undatum* (Weetman *et al.* 2006; Mariani *et al.* 2012; Pálsson *et al.* 2014). All used the same five microsatellites created for the species by Weetman *et al.* (2005). All those studies agreed that *B. undatum* can be genetically differentiated over tens of kilometres and that populations are more divergent in bays and inlets than those further offshore. These studies suggested a stepping-stone model of gene flow, which perhaps does not include the inshore populations explaining this differentiation

RADseq and its technical variants have revolutionised population studies by providing a cheap and quick way to produce a much larger level of genetic markers (Davey and Blaxter 2010). These methods do not require species-specific primers to be developed, nor do they require a reference genome, and instead, use restriction enzymes to cut across the genome to discover single nucleotide polymorphisms (SNPs) (Baird *et al.* 2008). While individually microsatellites provide more insight into population structure, the large number of SNPs, usually hundreds to thousands, generated by RADSeq provides more genomic resolution than the small number of microsatellites commonly used in population genetics genetic studies (Liu *et al.* 2005). RADseq has found fine-population structure using SNPs where previously microsatellites did not find any (Benestan *et al.* 2015; Szulkin *et al.* 2016; Vendrami *et al.* 2017).

The aim of this study was to use double digest restriction-site associated DNA (ddRAD) (Peterson *et al.* 2012) sequencing to generate single biallelic SNPs to investigate the population structure of *B. undatum* in the Southern North Sea, English Channel, and Irish Sea and compare the results with previous studies that employed microsatellites.

Methods

A total of 195 individuals were collected from 13 sampling locations across the Southern North Sea, English Channel, and Irish Sea between December 2018 and February 2019 (Table 2.1, Figure 2.1) Subsamples of tissue were taken from the foot and stored in either 100% ethanol at -20 °C or frozen at -20 °C.

Table 2.1.Sampling information, ID, sea, coordinates (decimal degrees), depth, and number of individuals sequenced and retained. KEIFCA is the Kent and Essex Inshore Fisheries and Conservation Authority.

Location	ID	Sea	Longitude	Latitude	Depth (m)	*n* sequenced	*n* retained*
Norfolk (Outside the bay)	NOA	North Sea	0.501	53.026	-4	15	15
Norfolk (Inside the bay)	NOB	North Sea	0.321	52.971	-14	15	11
Suffolk	SUF	North Sea	1.833	52.225	-26	15	12
KEIFCA Outside Zone 1	K10	North Sea	1.361	51.709	-13	15	6
KEIFCA Inside Zone 1	K1N	North Sea	1.063	51.563	-7	15	9
KEIFCA Inside Zone 2	K2N	North Sea	1.210	51.415	-3	15	12
KEIFCA Outside Zone 2	K2O	North Sea	1.655	51.353	-28	14	12
KEIFCA Inside Zone 3	K3N	North Sea	1.415	51.242	-10	15	11
KEIFCA Inside Zone 4	K4N	English Channel	1.057	50.926	-32	15	13
Weymouth Bay	WEY	English Channel	-2.317	50.606	-15	13	13
Lyme Regis	LYM	English Channel	-2.540	50.573	-21	15	10
Jersey Island	JER	English Channel	-2.280	49.024	-17	15	9
South East of Ireland▲	IRE	Irish Sea	-5.985	52.956	-6	14	8

* Individuals were retained in the population's analysis using STACKS if the coverage of the individual was greater than 20x.

▲The exact coordinates and depth are unknown. These were obtained within ICES VIIa inside the Irish territorial sea. Given location is adjacent the port where the whelks were landed, and depth was taken from that point.

ddRAD library preparation and sequencing

Total genomic DNA (gDNA) was extracted using (i) a modified CTAB and proteinase-K digest followed by phenol-chloroform purification and ethanol precipitation (Herrera *et al.* 2015) from tissue stored in 100% ethanol and (ii) using the Omega Biotek E.N.Z.A Mollusc extraction kit as per the manufacturer's instructions. A total of 191 gDNA samples were selected from a combination of both methods to proceed to amplicon library preparation. The ddRAD libraries were constructed using 800 ng of gDNA from each individual and following the double digest RADseq protocol (Peterson *et al.* 2012) using the restriction enzymes *ApeKI* and *BamHI-HF.* A more detailed protocol is available in the supplementary material.

The libraries were sequenced using four lanes of Illumina HiSeq X ten, (paired-end, 2 x150 bp) with the addition of 10% PHIX to each library.

De novo assembly and data filtering

Raw reads were processed using STACKS v 2.4 (Catchen *et al.* 2011, 2013). Reads were demultiplexed and quality filtered using the "*process_radtags.pl*" pipeline in STACKS. Any individual read with a Phred score below 33, ambiguous barcodes, or an unrecognised cut site were removed. Loci were assembled *de novo*, as there was no reference genome available for *B. undatum*. *De novo* assembly was done using the *"de_novo_map.pl"*. The three main *de novo* parameters M and m were tested using the *"r80 method"* (Paris et al. 2017). The m parameter is the minimum number of reads required to make a putative allele and is implemented in the *ustacks* component. M is the number of mismatches allowed between putative alleles to form a putative locus in *ustacks.* Finally, n is the number of mismatches allowed between putative loci during construction of the catalogue in *cstacks.* This was fixed at n=M following the guidelines in Paris *et al.* (2017) for a single species study. After testing, the optimum parameters were m=3 M=3 n=3.

The "*populations*" script was run to filter loci that were present in 70% sampling sites (*p* = 9) in at least 80% of individuals (*r* = 0.8). Alleles with minor frequencies less than 5% were removed (*min_maf = 0.05*) and maximum heterozygosity was set to 50% (*max_obs_het* = 0.5) to remove potential homologs. Only individuals with coverage greater than 20x were used in the population script to ensure accurate genotyping. Only one SNP was used per locus *(write_random_snp)* to minimise the effects of linkage disequilibrium. SNPs were filtered in PLINK v 1.9 (Purcell *et al.* 2007) to remove SNPs that were not called in 90% of genotypes (*geno 0.1*). PLINK was used to identify loci out of HWE to a significance of p < 0.05 (*hardy 0.05*) and to identify loci in linkage disequilibrium (LD) using a sliding window of 50 loci and step size of 5 loci with a correlation cut off greater than 10% *(indep-pairwise 50 5 0.1).*

File conversions were either done directly in STACKS v 2., PLINK v. 1.9, or indirectly using PGDSpider v. 2.1.1.5 (Lischer and Excoffier 2012).

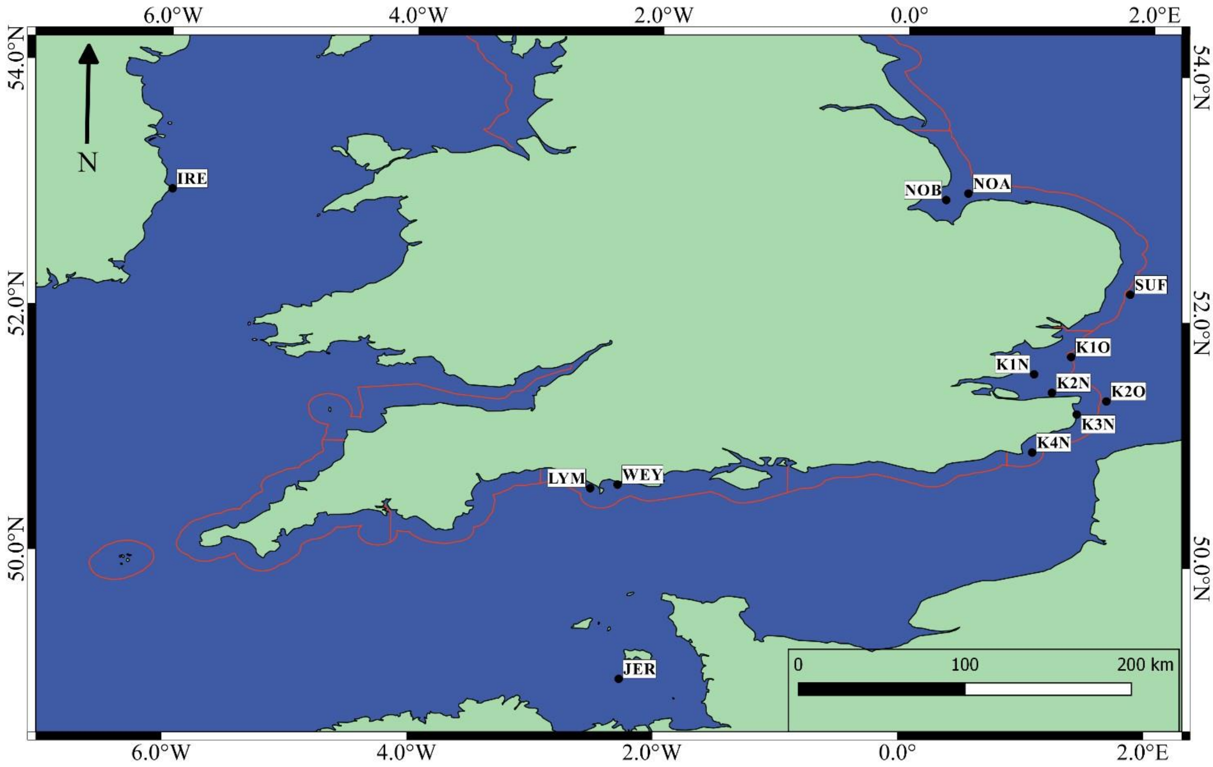

Figure 2.1. Sampling locations. Red line is the boundaries of the Inshore Fisheries and Conservation Authority Zones. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

Detection of outlier Loci

Outlier loci were detected using a genome-scan method implemented with BayeScan v 2.1 (Foll and Gaggiotti 2008) which uses a Bayesian method to estimate the posterior probability of selection acting on each locus by using a reversible-jump Monte Carlo Markov Chain process. BayeScan allows F_{st} coefficients to differ between populations which accommodates the different demographic history that may be present. A False Discovery Rate (FDR) (Storey 2003) of 0.1 was used. An FDR is the number false significant hypothesis that are expected when conducting multiple hypothesis testing. Prior odds of 10 were used in BayeScan, implying that the neutral model for each locus was 10 times more likely than the selective model. Twenty pilot runs were run with 10,000 iterations each, followed by a burnin of 50,000 followed by another 50,000 iterations. Identified outlier loci were blasted (blastn, Altschul *et al.* (1997)) through the National Center for Biotechnology Information (NCBI) nucleotide collection (nr/nt) to delineate the adaptive significance of the outliers based on similarity to existing gene and gene functions.

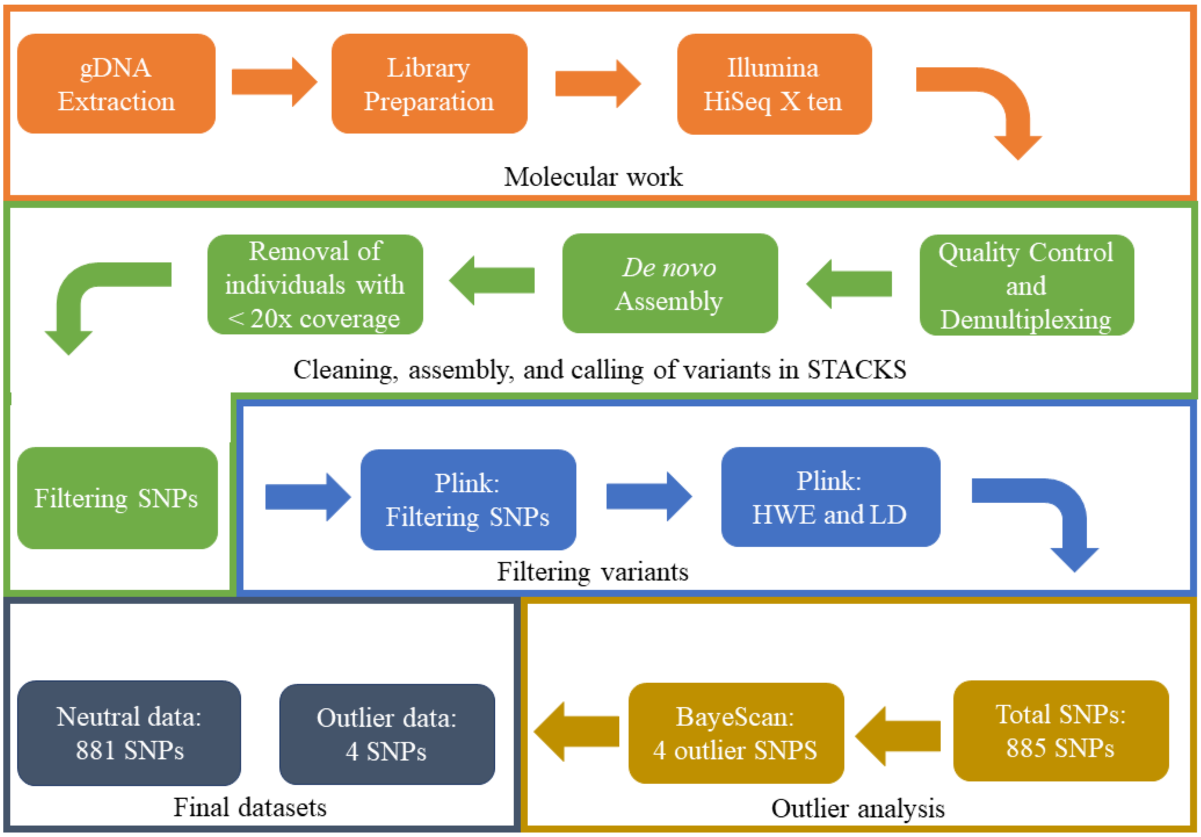

Figure 2.2 Workflow of the creation of the two datasets used. Abbreviations: (HWE) Hardy Weinberg Equilibrium (LD) Linkage Disequilibrium. SNPs outside of Hardy Weinberg Equilibrium and in Linkage Disequilibrium were removed.

Genetic analysis

A Bayesian clustering method implemented in Structure v 2.3.4 (Pritchard, Stephens, & Donnelly, 2000) was used to identify the presence of distinct genetic clusters (K). An admixture model with correlated allele frequencies without prior information on population membership was employed. A burnin of 500,000 and 500,000 further iterations was used. Structure Harvester Web v 0.6.94 (Earl 2012) was used to determine the optimal K based on ΔK, using the Evanno method (Evanno *et al.* 2005). Distruct plots were made using CLUMPAK (Kopelman *et al.* 2015). Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were carried out in R v 3.6 (R Core Team 2019) using the package *adegenet* v 2.1.1 (Jombart 2008). PCA loadings for each outlier were calculated in *adegenet***.** The number of Principal Components used for the DAPC were validated using the *Xvaldapc* function which returns the number of principal components with the lowest mean square error. Pairwise F_{st} (Weir and Clark Cockerham 1984), population specific F_{is} (Weir and Clark Cockerham 1984), H_e and H_o, were calculated in ARLEQUIN v 3.5 (Excoffier and Lischer 2010). Isolation by Distance (IBD) tests were carried out in ARLEQUIN v 3.5 using Linearized F_{st} and shortest distance by sea and significance was tested using a Mantel test (Mantel and Valand 1970) with 1000 permutations. A distance by sea matrix was created using the R package *marmap* (Pante and Simon-Bouhet 2013). Analysis of Molecular Variance (AMOVA) were calculated in ARLEQUIN v 3.5 to determine global F_{st} from the variance components with 1000 permutations to test for significance.

Results:

In total, 1,427,813,991 paired-end reads were generated across 191 samples. 4.19% had no barcode, 0.08% were removed due to low quality, 4.4% had no RAD cut site, retaining 1,303,931,081 paired-end reads. After removing individuals with less than 20x coverage 141 individuals remained with 6 to 15 individuals per site (Table 2.1) accounting for 1,210,090,726 (92.8% of all retained reads). Coverage varied from 21.41x to 71.52x (mean = 48.2x, S.D. = 13.2x) A total of 5,325 polymorphic loci and 141 individuals were retained after filtering in the *populations* script of STACKS of which 1,251 polymorphic loci were retained after removing loci with a call rate lower than 90% were removed. 181 and 185 loci were removed due to being identified as being in LD and outside of HWE respectively. Outlier analysis revealed four loci (0.45% of all loci) were identified as under putative positive selection. In total, 885 biallelic loci were retained.

To gain a better insight into the population structure present, the dataset was divided into two datasets (i) 881 loci in HWE excluding outliers identified by BayeScan, hereafter referred to as neutral loci, and (ii) four outlier loci identified as being under putative positive selection, hereafter referred to as outlier loci.

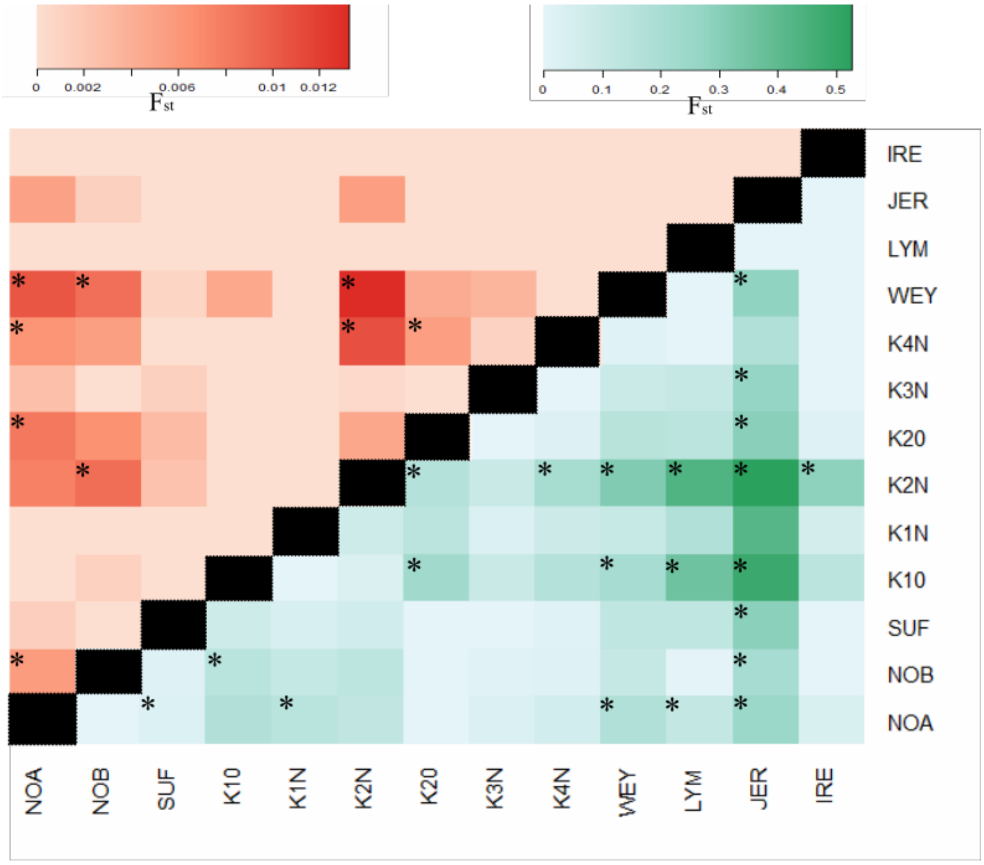

Figure 2.3. Heatmap of pairwise comparisons of F_{st} between sampling sites. Above the diagonal: Neutral loci. Below the Diagonal: Outlier loci. Asterisk denotes significance after Bonferroni correction. Refer to Table 1 for sampling location abbreviations. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

Neutral loci

Seventeen significant pairwise F_{st} comparisons between sampling locations, nine of which were significant after Bonferroni correction, were found (Figure 2.3, Table S2). Pairwise F_{st} ranged from 0.0052 (K2O vs K4N) to 0.0133 (K2N vs WEY). No significant sampling-location specific inbreeding (F_{is}) was observed (Table S4). There was no observed trend of IBD (r = -0.25, p = 0.882) when all sampling locations were considered. IBD trends were investigated further; by analysing the North Sea sampling locations exclusively (r = 0.51, p = 0.001) and solely analysing the English Channel-Irish Sea sampling locations (r = 0.26, p = 0.338). There was a strong and significant associated between genetic and geographic distance when using all sampling locations bar IRE, JER, and LYM (r = 0.54, p = 0.002). Based on the Bayesian cluster analysis (in Structure v 2.3.4) three distinct genetic clusters were present (ΔK = 2868.83), one of which was found at a slightly larger frequency in the English Channel sampling locations (Figure 2.4), and another at higher frequencies in the

North Sea. The remaining cluster was admixed at near equal proportions across all sampling sites. DAPC clustering found no evidence for genetic structure (Figure S1), and the PCA revealed large levels of admixture between sampling sites (Figure 2.4). Global F_{st} was -0.1453 and was not significant. Heterozygosity was moderate overall ranging between 0.2542±0.1517 at NOA to 0.2928±0.1704 at IRE (Figure 2.5, Table S4).

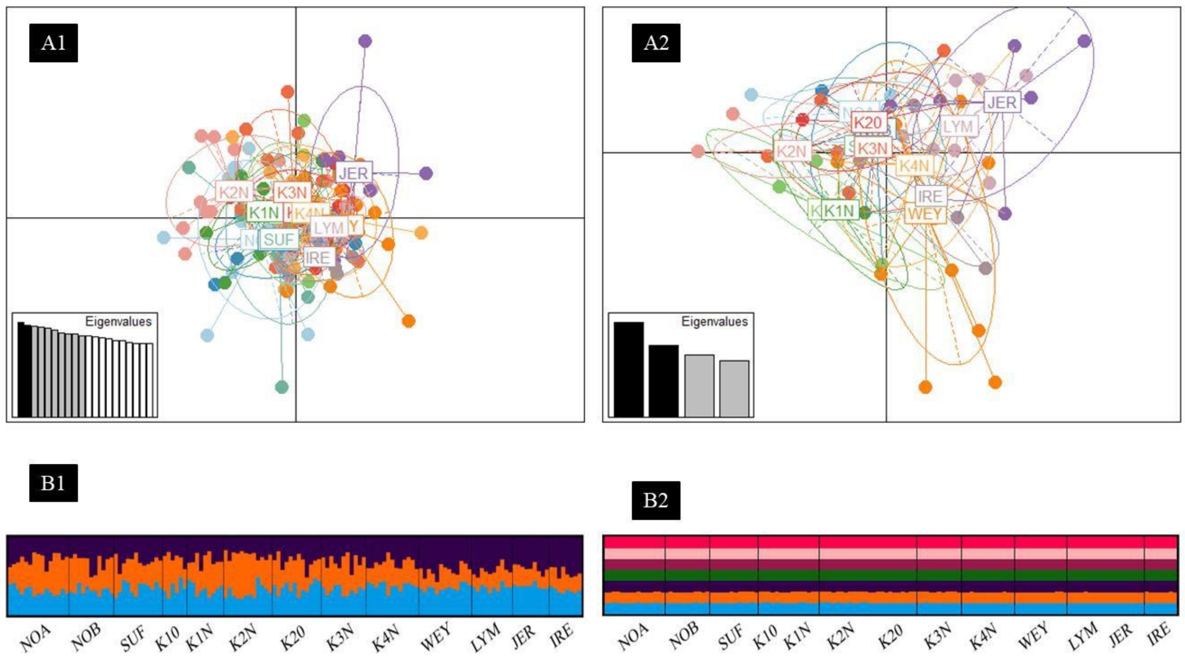

Figure 2.4.(A1) Principal Component Analysis using neutral loci (A2) Principal Component Analysis using outlier loci (B1) Distruct plot using neutral data where optimal K=3 (B2) Distruct plot using outlier data where optimal K=7. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

Outlier loci

Among the 13 sampling locations, 41 significant pairwise Fst comparisons between sampling locations were identified – 19 of which were significant after Bonferroni corrections was applied (Figure 2.3, Table S2). Pairwise F_{st} values ranged from 0.1144 (NOA vs LYM) to 0.5280 (K2N vs JER). There was no IDB present ($r = 0.03$, $b = < 0.001$, $p = 0.366$) when all sampling locations were considered. No IBD was observed when solely the North Sea sampling locations were analysed ($r = 0.05$, $p = 0.312$) or only the English Channel-Irish Sea sampling locations were analysed ($r = -0.25$, $p = 0.617$). Furthermore, unlike the neutral loci there was no IBD observed when IRE, JER, and LYM were omitted ($r = 0.28$, $p = 0.111$). No sampling-location specific inbreeding (F_{is}) was observed (Table S4). Seven distinct clusters were identified by Structure ($\Delta K = 5.44$), however, all were admixed at near equal proportions across all sampling locations (Figure 2.4) However, DAPC found four genetic clusters (Figure S1). These genetic clusters were not equally admixed but instead contained three admixed clusters and one distinct cluster. This distinct cluster contained individuals from NOA, NOB, SUF, K1O, K1N, K2N, K2O, and K3N, all of which were located in the North Sea (Table S3). Loadings of the four outlier loci indicated that the allele frequencies of a single SNP (1362698_271) contributed most to the variability of the principal components (Figure S2). There was variability in allele frequencies at this SNP until in the North Sea, until K4N. From here, the allele frequency of that SNP decreased, and one allele dominated WEY, LYM, JER, and IRE (Figure S2). The PCA displayed large levels of admixture between sampling locations (Figure 2.4). Global F_{st} was 0.1132 and was significant. Observed heterozygosity was moderate to very high and varied between sites ranging from 0.2523±0.1102 at K2N to 0.5279±0.0393 at LYM (Table S4). No significant alignments were identified in the NCBI nucleotide collection (Table S5)

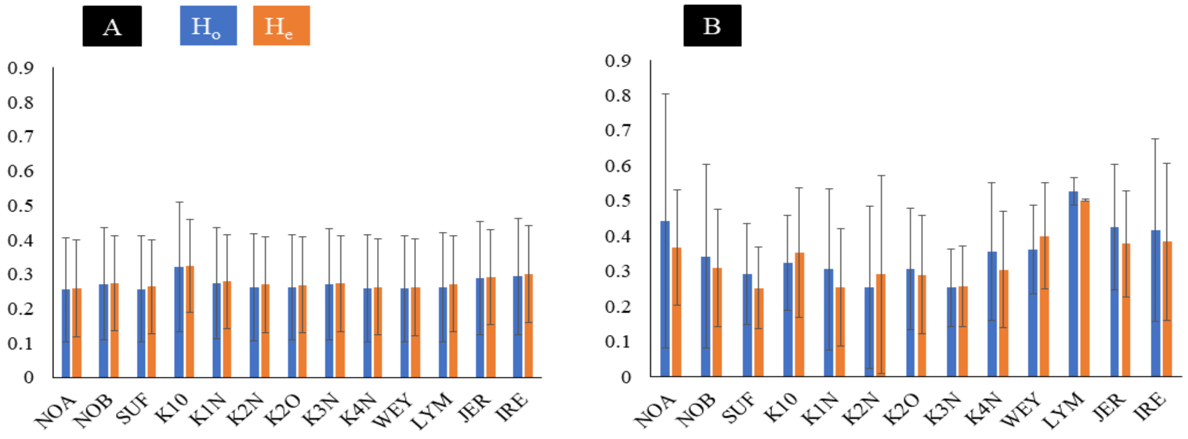

Figure 2.5. Sampling location specific H_e (maximum He for single bi-allelic SNPs is 0.5) and H_o using (A) neutral loci and (B) outlier loci. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

Discussion:

This present study generated genomic datasets for *B. undatum* in the Southern North Sea, English Channel and Irish Sea using bi-allelic SNPs derived from ddRAD. It represents the first study to investigate fine-scale population structure of the species using NGS. The results of this study agreed with previous studies that *B. undatum* consists of a large semi-continuous population over the Irish Sea, English Channel, and Southern North Sea. There was a strong and significant correlation between genetic and geographic distance when the North Sea sampling locations were analysed separately and when all sampling locations bar IRE, JER, and LYM, were analysed using neutral loci. No IBD observed in the English Channel-Irish Sea using neutral loci. No IBD was observed using outlier loci regardless of the exclusion of any sampling sites, or when analysing North Sea sampling locations or English Channel-Irish Sea samples independently. Outlier loci analyses did not reveal any locally adapted populations.

Genetic connectivity using neutral loci

The global F_{st} reported in the present study was -0.1453 and was not significant. This negligible value contradicts other studies that reported small but significant genetic differentiation along the British North Sea (global F_{st} = 0.010, Weetman *et al.* 2006) and Irish Sea (global F_{st} = 0.019, Mariani *et al.* 2012). The low global F_{st}, and admixed genetic clusters in this present study indicated that throughout the range there was one large population, however, the frequencies of the different clusters differed slightly between regions. Only 11.54% of pairwise F_{st} comparisons were significant (21.79% before Bonferroni correction), much lower than the 68.53% (98 out of 143, just UK samples removing the Solent as it is defined within that study as a distinct genetic cluster) reported in Weetman *et al.* (2006). This reduction may be due to the increased genomic resolution provided by using 881 SNPs over 5 microsatellites. However, it may be due to the spatial resolution in this current study compared to Weetman *et al.* (2006). That study included sampling locations further north in England, and into Scotland and included sampling locations in Wales and the English and Scottish side of the Irish Sea. A wider geographic scope is needed to directly compare this present study with Weetman *et al.* (2006) more accurately. The presence of significant pairwise differences in this present study may indicate that there was significant sub-structure within the population, but that gene flow via migration or semi-continuity appeared to be enough to prevent the population from becoming structured.

Significant IBD was observed along the Southern North Sea and English Channel as far as WEY (r = 0.54, p = 0.002). This trend of IBD has been previously reported (Weetman *et al.* 2006, Mariani *et al.* 2012) and was characteristic of the species. Furthermore, when all sampling locations were used IBD was not significant (r = -0.25, p = 0.882). This was similar to Weetman *et al.* (2006) where on a larger scale (Scotland, England, Sweeden, and France) there was no IBD observed.

Buccinum undatum has been reported to be more differentiated inside bays and inlets compared to more open sampling locations (Weetman *et al.* 2006; Mariani *et al.* 2012).

However, this current study did not provide evidence for this. This present study sampled both within a bay in Norfolk (NOB) and outside the same bay (NOA). The distance between these sampling sites was 14 km. While there was a significant pairwise F_{st} between both of these sampling sites, all comparisons with NOB accounted for 33% of significant pairwise differences, and comparisons involving NOA accounted for a further 33%. A possible explanation is NOB was sampled close to the mouth of the bay. This proximity to the mouth may mean that it would be better connected than individuals located more inshore. Hence, more thorough sampling within bays and inlets, and further within these locations, would be required to explore this trend.

Analysis of outlier loci

Bayesian clustering analysis revealed seven genetic clusters, equally admixed across all sampling locations and that there was overall a single large population. A larger number of pairwise F_{st} comparisons were observed using the outlier loci, 24.36% - 52.56% before Bonferroni correction was applied. Eight (38.09%) of significant pairwise F_{st} comparisons involved JER. JER was the most southerly sampling location in this study and had the largest pairwise comparison (JER vs K2N, $F_{st} = 0.5279$). The underlying cause of the elevated levels of genetic differentiation at JER is unknown, but it is the most southerly sampling location in this study and may represent a change in the environmental pressures at that latitude. However, no oceanographic data was analysed in this present study to test for correlation between environmental variables. Furthermore, results from blasting outlier consensus loci through the NCBI nucleotide collection did not reveal any significant matches with annotated genes that would provide insight into gene function. DAPC analysis revealed four genetic clusters, one of which was distinct from the other three. This distinct cluster contained some individuals (21.28%) located at all sampling locations within the North Sea. This may imply that local environmental pressures in the North Sea differ from those in the English Channel-Irish Sea leading to a distinct population to live sympatrically with the other. The loadings for the DAPC revealed that allele variations in one SNP contributed the most to the variation observed in the DAPC. This SNP had a single allele which dominated the English Channel-Irish Sea, and the frequencies between the two alleles became more variable in the North Sea. The increased frequency of significant pairwise F_{st} comparison between sampling locations and the DAPC loadings may indicate that selective divergence may occur on a local scale.

Large population size supresses genetic drift

The low level of genetic differentiation over the entire study area using the neutral loci could be the result of a large effective population size which was suppressing the effects of genetic drift. Genetic drift is the random loss of alleles over time. In a population of infinite size where random mating occurred, the frequency of alleles would remain stationary over time. Genetic drift leads to the fixation of alleles where populations are small or have mating systems that deviate from random. The low pairwise F_{st} values, combined with the similar frequencies of genetic clusters observed across the Southern North Sea, English Channel and Irish sea using the neutral loci indicate that there has not been a major fixation of alleles. Effective population size for *B. undatum* in the Irish Sea was previously calculated by Mariani *et al.* (2012). That study found that at most sampling locations in the Irish Sea, estimates of effective population size were infinite. Evidence for a large effective population size in the UK exists, as *B. undatum* have been reported as a fishery since 1922 and since then has grown to be the sixth-largest shellfish fishery in the UK (Marine Management Organisation 2017). For example, in the last 10 years, over 10,000 tonnes of *B. undatum* have been landed in the UK every year (Marine Management Organisation 2014, 2017).

Application of Next-Generation Sequencing and Genome-Scan outlier detection in population genomics

This study is the first to use RADSeq derived genetic markers for *B. undatum*. Previously, only microsatellites and mtDNA were used to infer population connectivity. Microsatellite markers have been difficult to develop in molluscs for reasons that are not fully understood, although some research indicated that it is the presence of large numbers of transposable elements in their genome (McInerney *et al.* 2011). These transposable elements may play an important part in reducing inter-specific genome variation and prevent microsatellite development (McInerney *et al.* 2011). Thus, RADseq represents a major technical breakthrough in molluscan population studies as it allows loci from non-model organisms, such as *B. undatum,* to be aligned without a reference genome (Davey and Blaxter 2010). RADseq can create thousands of genetic markers allowing for higher resolution population inferences. While no previous study exists using RADseq on *B. undatum*, other molluscan species have been examined using such *Pectan maximus* (Vendrami *et al.* 2017), *Pectan magellanicus* (Van Wyngaarden *et al.* 2017).

Vendrami *et al.* (2017) conducted a study on *Pectan maximus* across the Northern Ireland coast using both 15 microsatellites and 10, 539 SNPs. That study conclusively found that analyses involving SNPs revealed a distinct population within Mulroy Bay that was previously unidentified using microsatellites. While individual microsatellites are more informative that SNPs, the vast amount of SNPs allows for more robust inferences than a limited set of microsatellites (Liu *et al.* 2005). Thus, the population inference presented in the current study is more informative than previous studies.

RADseq and its technical variants use restriction enzymes to cut along the genome i.e. there is genome-wide representation of SNPs. With a suite of genetic markers across the genome, it is possible to identify a panel of SNPs that show abnormally large or small F_{st} compared to the mean. However, just analysing the neutral markers in the genome can mask the true population structure present by not accounting for evolutionary forces such as adaption which can give rise to genomic variation. Outlier analysis has revealed population structure that was masked by neutral loci e.g. European hake (Milano *et al.* 2014). This is because adaptive selection can cause heterogeneous genomic divergence even in the presence of high gene flow by only acting on specific loci and those linked to them (Nosil *et al.* 2009).

Conclusions

Analysis of the neutral loci indicated advocated for the presence of a large semi-continuous population with sub-structure present. Global F_{st} estimates were very low, and not significant. Furthermore, multiple significant pairwise F_{st} comparisons were found. However significant comparisons only accounted for 11.54|% of all comparison which suggested that sub-structure was small and that gene flow between sampling locations prevented overall significant population structure from forming. A strong and significant association was found between geographic and genetic distance when North Sea sampling locations were analysed and when analysing all sampling locations bar IRE, JER and LYM. *Buccinum undatum* may have a large effective population size that means genetic drift has not occurred in the geographic range of this study. Contrary to previous studies, this study found no evidence for individuals sampled within bays and inlets to be more differentiated than those sampled outside of them. However, this study only sampled inside one bay, and that was near the mouth which may have not been far enough into the bay to see an observable trend.

By using outlier detection methods implemented in BayeScan, 4 SNPs were identified as being potentially under selection (0.45% of total SNPs used). Much higher global F_{st} were found using outlier loci than neutral loci. These values were an order of magnitude higher and represented much greater genetic differentiation than reported in previous studies that used microsatellite markers and by using the neutral loci in this present study. Bayesian clustering methods determined the presence of a single admixed population, while DAPC clustering advocated for the presence of a distinct population in the North Sea, which was admixed with individuals from another population. Furthermore, blasting the outlier loci through NCBIs nucleotide collection did not produce any significant matches to previously annotated genes that may have provided insight into gene function. Without a reference genome or oceanographic data, the underlying causes of selection cannot be addressed.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic acids research,* 25(17), 3389–3402.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., Johnson, E.A. (2008) 'Rapid SNP discovery and genetic mapping using sequenced RAD markers', *PloS one,* 3(10), e3376–e3376.

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., Bernatchez, L. (2015) 'RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus)', *Molecular ecology,* 24(13), 3299–3315.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A. (2013) 'Stacks: an analysis tool set for population genomics', *Molecular ecology,* 22(11), 3124–3140.

Catchen, J.M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J.H. (2011) 'Stacks: building and genotyping loci de novo from short-read sequences', *G3: Genes, genomes, genetics,* 1(3), 171–182.

Cowen, R.K., Sponaugle, S. (2009) 'Larval Dispersal and Marine Population Connectivity', *Annual Review of Marine Science,* 1(1), 443–466.

Davey, J.W., Blaxter, M.L. (2010) 'RADSeq: next-generation population genetics', *Briefings in functional genomics,* 9(5–6), 416–423.

Earl, D.A. (2012) 'STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method', *Conservation genetics resources,* 4(2), 359–361.

Evanno, G., Regnaut, S., Goudet, J. (2005) 'Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study', *Molecular ecology,* 14(8), 2611–2620.

Excoffier, L., Lischer, H.E.L. (2010) 'Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows', *Molecular ecology resources,* 10(3), 564–567.

Foll, M., Gaggiotti, O. (2008) 'A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective', *Genetics*, 180(2), 977–993.

Hauser, L., Carvalho, G. (2008) 'Paradigm shifts in marine fisheries genetics: Ugly hypotheses slain by beautiful facts', *Fish and Fisheries*, 9, 333–362.

Herrera, S., Watanabe, H., Shank, T.M. (2015) ‘Evolutionary and biogeographical patterns of barnacles from deep-sea hydrothermal vents’, *Molecular ecology*, 24(3), 673–689.

Jombart, T. (2008) ‘adegenet: a R package for the multivariate analysis of genetic markers’, *Bioinformatics*, 24(11), 1403–1405.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I. (2015) ‘Clumpak: a program for identifying clustering modes and packaging population structure inferences across K’, *Molecular ecology resources*, 15(5), 1179–1191. 38

Kyle, C.J., Boulding, E.G. (2000) ‘Comparative population genetic structure of marine gastropods (Littorina spp.) with and without pelagic larval dispersal’, Marine Biology, 137(5), 835–845.

Lischer, H.E.L., Excoffier, L. (2012) ‘PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs’, Bioinformatics, 28(2), 298–299.

Liu, N., Chen, L., Wang, S., Oh, C., Zhao, H. (2005) ‘Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure’, in Bmc Genetics, BioMed Central, S26.

Mantel, N., Valand, R.S. (1970) ‘A Technique of Nonparametric Multivariate Analysis’, Biometrics, 26(3), 547.

Mariani, S., Peijnenburg, K.T., Weetman, D. (2012) ‘Independence of neutral and adaptive divergence in a low dispersal marine mollusc’, Marine Ecology Progress Series, 446, 173–187.

Marine Management Organisation (2014) UK Sea Fisheries Annual Statistics Report 2013, Radford.

Marine Management Organisation (2017) UK Sea Fisheries Annual Statistics Report 2016, London.

McInerney, C.E., Allcock, A.L., Johnson, M.P., Bailie, D.A., Prodöhl, P.A. (2011) ‘Comparative genomic analysis reveals species-dependent complexities that explain difficulties with microsatellite marker development in molluscs’, Heredity, 106(1), 78.

McIntyre, R., Lawler, A., Masefield, R. (2015) ‘Size of maturity of the common whelk, Buccinum undatum: Is the minimum landing size in England too low?’, Fisheries Research, 162, 53–57.

Milano, I., Babbucci, M., Cariani, A., Atanassova, M., Bekkevold, D., Carvalho, G.R., Espiñeira, M., Fiorentino, F., Garofalo, G., Geffen, A.J., Hansen, J.H., Helyar, S.J., Nielsen, E.E., Ogden, R., Patarnello, T., Stagioni, M., Consortium, F., Tinti, F., Bargelloni, L. (2014) ‘Outlier SNP markers reveal fine-scale genetic structuring across European hake populations (Merluccius merluccius)’, Molecular Ecology, 23(1), 118–135.

Nicholson, G.J., Evans, S.M. (1997) ‘Anthropogenic impacts on the stocks of the common whelk Buccinum undatum (L.)’, Marine Environmental Research, 44(3), 305–314.

Nielsen, C. (1974) ‘Observations on Buccinum undatum L. attacking bivalves and on prey responses, with a short review on attack methods of other prosobranchs’, Ophelia, 13(1–2), 87–108.

Nosil, P., Funk, D.J., Ortiz-Barrientos, D. (2009) ‘Divergent selection and heterogeneous genomic divergence’, Molecular ecology, 18(3), 375–402.

Pálsson, S., Magnúsdóttir, H., Reynisdóttir, S., Jónsson, Z.O., Örnólfsdóttir, E.B. (2014) ‘Divergence and molecular variation in common whelk Buccinum undatum (Gastropoda: Buccinidae) in Iceland: A trans-Atlantic comparison’, Biological Journal of the Linnean Society, 111(1), 145–159.

Pante, E., Simon-Bouhet, B. (2013) ‘marmap: A Package for Importing, Plotting and 39 Analyzing Bathymetric and Topographic Data in R’, PLOS ONE, 8(9), 1–4.

Paris, J.R., Stevens, J.R., Catchen, J.M. (2017) ‘Lost in parameter space: a road map for stacks’, Methods in Ecology and Evolution, 8(10), 1360–1373.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E. (2012) ‘Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species’, PLOS ONE, 7(5), 1–11.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J. (2007) ‘PLINK: a tool set for whole-genome association and population-based linkage analyses’, The American journal of human genetics, 81(3), 559–575.

R Core Team (2019) ‘A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2012’, URL https://www. R-project. org.

Selkoe, K., Toonen, R. (2011) ‘Marine connectivity: A new look at pelagic larval duration and genetic metrics of dispersal’, Marine Ecology Progress Series, 436, 291–305.

Shrives, J.P., Pickup, S.E., Morel, G.M. (2015) ‘Whelk (Buccinum undatum L.) stocks around the Island of Jersey, Channel Islands: Reassessment and implications for sustainable management’, Fisheries research, 167, 236–242.

Storey, J.D. (2003) ‘The positive false discovery rate: a Bayesian interpretation and the q -value’, Ann. Statist., 31(6), 2013–2035.

Szulkin, M., Gagnaire, P., Bierne, N., Charmantier, A. (2016) ‘Population genomic footprints of fine-scale differentiation between habitats in Mediterranean blue tits’, Molecular ecology, 25(2), 542–558.

Taylor, J.D., Taylor, C.N. (1977) ‘Latitudinal Distribution of Predatory Gastropods on the Eastern Atlantic Shelf’, Journal of Biogeography, 4(1), 73–81.

Vendrami, D.L.J., Telesca, L., Weigand, H., Weiss, M., Fawcett, K., Lehman, K., Clark, M.S., Leese, F., McMinn, C., Moore, H., Hoffman, J.I. (2017) 'RAD sequencing resolves fine-scale population structure in a benthic invertebrate: Implications for understanding phenotypic plasticity', Royal Society Open Science, 4(2).

Weetman, D., Hauser, L., Bayes, M.K., Ellis, J.R., Shaw, P.W. (2006) 'Genetic population structure across a range of geographic scales in the commercially exploited marine gastropod Buccinum undatum', Marine Ecology Progress Series, 317, 157–169.

Weetman, D., Hauser, L., Shaw, P.W., Bayes, M.K. (2005) 'Microsatellite markers for the whelk Buccinum undatum', Molecular Ecology Notes, 5(2), 361–362.

Weir, B., Clark Cockerham, C. (1984) 'Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38: 1358-1370', Evolution, 38, 1358–1370.

Van Wyngaarden, M., Snelgrove, P.V.R., DiBacco, C., Hamilton, L.C., Rodríguez-Ezpeleta, N., Jeffery, N.W., Stanley, R.R.E., Bradbury, I.R. (2017) 'Identifying patterns of dispersal, connectivity and selection in the sea scallop, Placopecten magellanicus, using RADseq-derived SNP s', Evolutionary applications, 10(1), 102–117.

Supplementary material

Methods

Genomic DNA extraction and library preparation methods

15 individuals were sampled from 13 sampling locations across England and Ireland. All samples were subsampled and stored in 100% ethanol at -20 °C and tissue was subsampled and stored at -20 °C. Genomic DNA was extracted from the ethanol preserved specimens using the CTAB phenol-chloroform method, and frozen tissue was extracted using the Omega Biotek E.N.Z.A Mollusc extraction kit as per the manufacturer's instructions.

191 individuals were selected from a combination of these methods to proceed to the library preparation. The ddRAD libraries were constructed using 800 ng of gDNA from each individual. gDNA was digested initially using *BamHI-HF* incubated at 37 °C for 4 hours. A second digestion using *ApeKI* was incubated at 75 °C for 4 hours.

The 191 individuals were then separated into two libraries, and each library was ligated with a combination of 96 barcodes. These barcodes were ligated in a 40 µl reaction volume containing 6.3 µl (100 ng) of gDNA, 0.3 µl of T4 ligase, 4 µl of 10x T4 ligase buffer, 23.4 µl of DNase free water, 3 µl (60 ng) of the respective Ape adaptor, and 3 µl of the respective Bam adaptor. The reaction was incubated at 21 °C for 12 hours and the T4 ligase was heat activated at 65 °C for 10 minutes.

After ligation the samples were pooled into the respective libraries. Each library was separated into 8 1.5ml Eppendorf tubes. In each tube the libraries were purified by adding equal volumes of Machery Nagel beads (480 µl) and allowed to stand on the magnetic stand for 10 minutes. The supernatant was removed, and the precipitate was washed with 80% EtOH three times and left to dry. 60 µl of 5mM Tris/HCl (pH 8.5) was added and the solution was incubated for a further 5 minutes. The tubes were allowed to stand on the magnetic stand for minutes. The elution buffers of each library were repooled (8x60 µl). Next each library was redistributed into 2 tubes and equal volumes of Machery Nagel beads were added (240 µl). The beads were mixed and incubated at room temperature for 5 minutes. The tubes were left to stand on a magnetic stand for a further 5 minutes. The supernatant was removed, and the precipitate was washed three times with 80% EtOH and allowed to dry. An aliquot of 30 µl of 5mM Tris-HCl (pH 8.5) buffer was added to each tube and incubated at room 41 temperature for 5 minutes. The solution was then allowed to stand on the magnetic stand for 5 minutes. The elution buffers from each library were repooled.

Each library was size selected using the Pippin Prep (Cassette marker L, 2% agarose) under “broad” settings with a target fragment length of 400 bp and a range of 360-440 bp. After size selection the following PCR method was used to amplify the libraries 72 °C for 3 mins; 11 cycles of 98 °C for 10 seconds, 65 °C for 30 seconds, and 72 °C for 45 seconds; followed by a final run of 72 °C for 5 minutes. The amplified libraries were purified using Nagel Machery magnetic beads and 60 µl of 5mM Tris-HCl (pH8.5) buffer. The library was quantified using a Qubit fluorometer and were compared against a reference library to ensure correct fragment lengths.

Figures and tables

Table S1. Minimum distance by sea distance matrix between sampling sites. Distances are in kilometres. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

	NOA	NOB	SUF	K1O	K1N	K2N	K2O	K3N	K4N	WEY	LYM	JER	IRE
NOA	0												
NOB	14	0											
SUF	151	164	0										
K1O	212	225	67	0									
K1N	236	249	92	27	0								
K2N	245	258	100	33	18	0							
K2O	247	260	98	44	49	34	0						
K3N	263	276	114	53	50	35	21	0					
K4N	308	321	159	98	95	80	65	45	0				
WEY	558	571	409	349	346	330	315	295	252	0			
LYM	575	588	426	365	362	347	332	312	269	18	0		
JER	649	661	499	439	436	421	406	386	343	176	175	0	
IRE	1152	1165	1003	943	940	924	909	890	847	597	579	612	0

Table S2. Pairwise F_{st} comparisons. Above the diagonal is the neutral loci. Below the diagonal is the outlier loci. Significant comparisons ($p <$ 0.05) are in bold. Significant comparisons after Bonferroni correction are highlighted in grey. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

NOA		NOB	SUF	K1O	K1N	K2N	K2O	K3N	K4N	WEY	LYM	JER	IRE
NOA		**0.0057**	0.0014	-0.0002	-0.0026	**0.0077**	**0.0083**	0.0027	**0.0063**	**0.0105**	-0.1130	**0.0052**	-0.0350
NOB	-0.0321		-0.0009	0.0011	-0.0031	**0.0090**	**0.0066**	-0.0010	**0.0054**	**0.0088**	-0.1105	0.0012	-0.0337
SUF	0.0261	0.0210		-0.0020	-0.0059	0.0025	0.0030	0.0013	-0.0001	0.0009	-0.1067	-0.0042	-0.0300
K1O	**0.1737**	**0.1466**	0.0829		-0.0125	-0.0050	-0.0067	-0.0051	-0.0021	0.0046	-0.1133	-0.0044	-0.0366
K1N	**0.1474**	**0.1085**	0.0431	-0.0399		-0.0037	-0.0069	-0.0058	-0.0037	-0.0016	-0.1293	-0.0088	-0.0398
K2N	**0.1206**	**0.1332**	0.0643	0.0351	0.0802		**0.0048**	0.0005	**0.0109**	**0.0133**	-0.0980	**0.0055**	-0.0335
K2O	-0.0139	-0.0253	-0.0055	**0.2314**	**0.1437**	**0.1643**		-0.0024	**0.0055**	**0.0044**	-0.1180	-0.0019	-0.0274
K3N	0.0294	0.0133	-0.0435	**0.0948**	0.0331	**0.0948**	-0.0075		0.0010	**0.0036**	-0.1149	-0.0077	-0.0333
K4N	**0.0618**	0.0210	0.0155	**0.1668**	**0.0854**	**0.2109**	0.0212	-0.0120		-0.0017	-0.1082	-0.0048	-0.0298
WEY	**0.1771**	**0.1017**	**0.1270**	**0.2154**	**0.0996**	**0.3201**	**0.1551**	**0.0884**	0.0142		-0.1133	-0.0065	-0.0321
LYM	**0.1144**	0.0032	**0.1258**	**0.3624**	**0.1762**	**0.4378**	**0.1413**	**0.1011**	-0.1378	-0.1639		-0.1248	-0.1596
JER	**0.2528**	**0.2128**	**0.2942**	**0.4856**	**0.4225**	**0.5279**	**0.2990**	**0.2704**	**0.1766**	**0.2796**	-0.1233		-0.0316
IRE	0.0358	-0.0338	-0.0190	0.1465	0.0567	**0.2871**	0.0208	-0.0789	-0.1565	-0.1453	-1.0398	-0.1033	

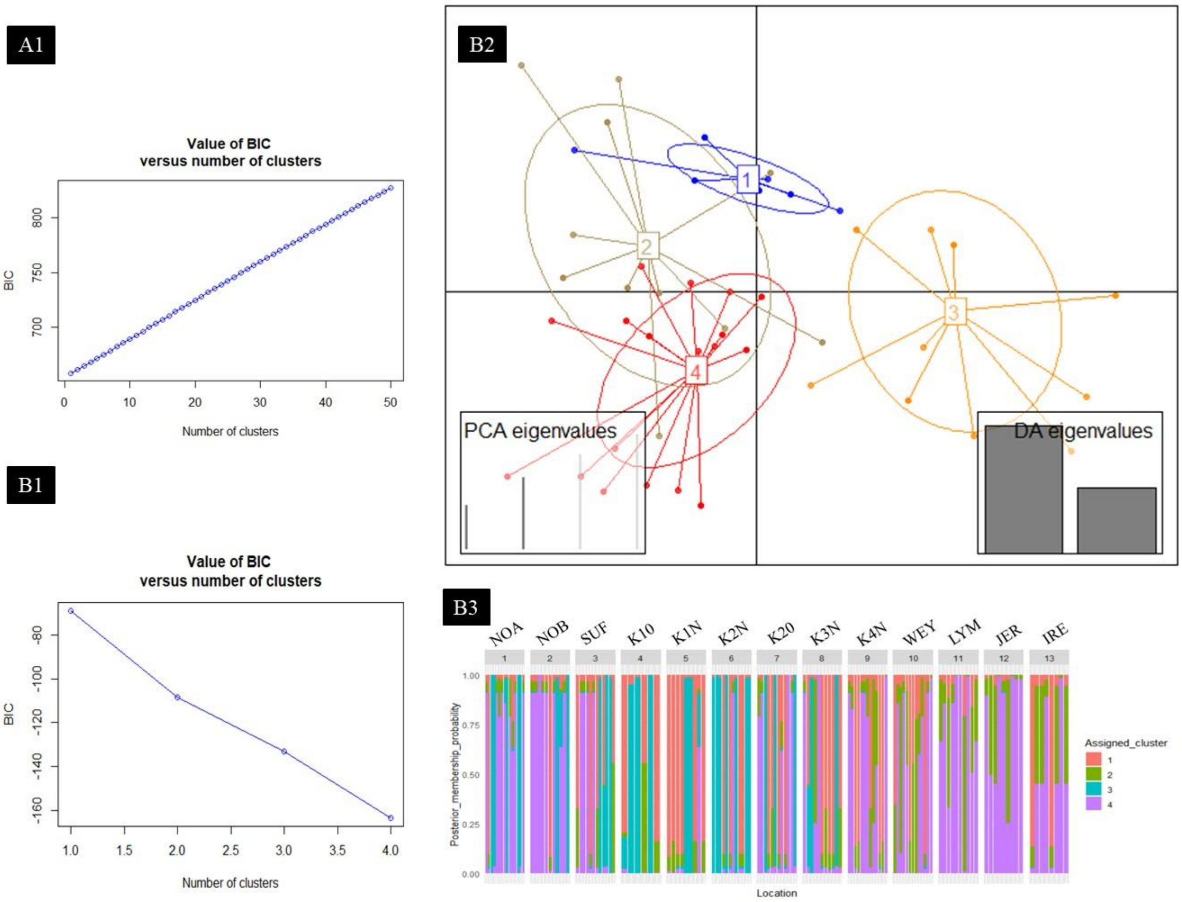

Figure S1. (A1) Bayesian Inference Criterion used to determine the optimal number of clusters using neutral loci. Optimal number of clusters was 1. (B1) Bayesian Inference Criterion used to determine the optimal number of clusters using outlier loci. Optimal number of clusters was 4. (B2) DAPC plot using outlier loci and (B3) Structure-like bar plot based on DAPC analysis of outlier loci. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

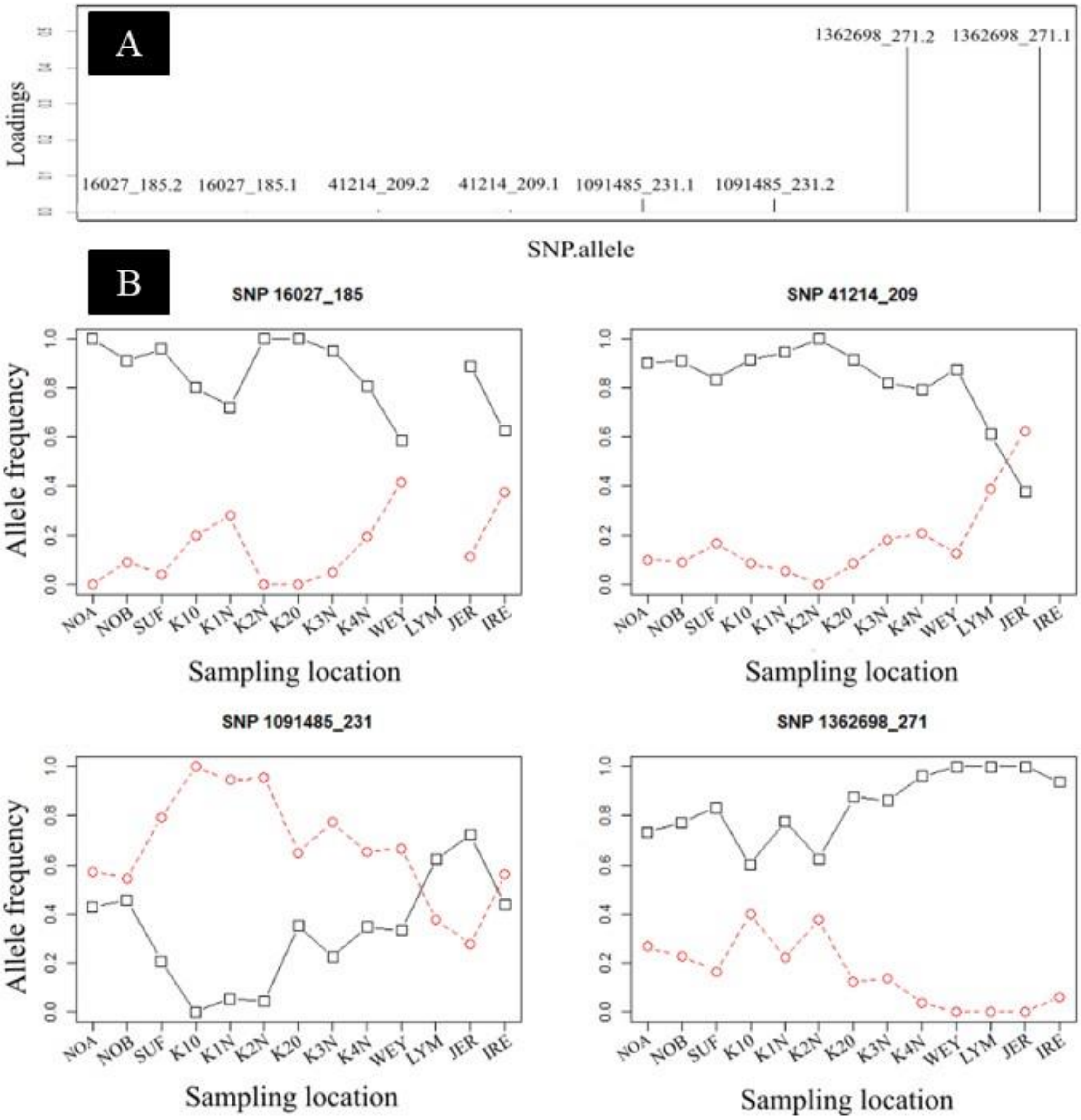

Figure S2 (A) Loadings plot indicating the contribution of each outlier SNP to the variability in the PCA of the outlier loci. (B) Change in allele frequencies of outlier SNPs between sampling locations. Circle and square represent a different allele in the biallelic loci. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

Table S3. Number of individuals from sampling locations assigned to clusters during DAPC analysis. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
NOA	1	0	6	8
NOB	0	2	2	7
SUF	3	1	4	4
K1O	2	1	3	0
K1N	5	2	2	0
K2N	4	0	7	1
K2O	4	0	3	5
K3N	5	1	3	2
K4N	5	3	0	5
WEY	2	5	0	6
LYM	1	0	0	9
JER	0	2	0	7
IRE	2	5	0	1

Table S4. H_e (± SD) (Maximum H_e for bi-allelic SNPs is 0.5), H_o (± SD), and F_{is} per population. Significant F_{is} ($p<0.05$) are in bold. Abbreviations: (NOA) Norfolk (Outside the bay), (NOB) Norfolk (Inside the bay), (SUF) Suffolk, (K1O) KEIFCA Outside Zone 1, (K1N) KEIFCA Inside Zone 1, (K2N) KEIFCA Inside Zone 2, (K2O)KEIFCA Outside Zone 2, (K3N) KEIFCA Inside Zone 3,(K4N) KEIFCA Inside Zone 4,(WEY) Weymouth Bay, (LYM) Lyme Regis, (JER) Jersey Island, (IRE) South East of Ireland.

	Neutral loci			Outlier loci		
	H_o	H_E	F_{is}	H_o	H_e	F_{is}
NOA	0.2542±0.1517	0.2589±0.1411	-0.0137	0.4413±0.3617	0.3663±0.1643	-0.2372
NOB	0.2716±0.1636	0.2726±0.1378	-0.0238	0.3409±0.2611	0.3084±0.1680	-0.1111
SUF	0.2571±0.1529	0.2631±0.1363	-0.0267	0.2917±0.1443	0.2518±0.1152	-0.1667
K1O	0.3220±0.1883	0.3231±0.1352	-0.0157	0.3222±0.1347	0.3519±0.1834	-0.0870
K1N	0.2728±0.1609	0.2787±0.1360	0.0027	0.3056±0.2291	0.2533±0.1659	-0.2222
K2N	0.2616±0.1566	0.2692±0.1389	0.0022	0.2538±0.2304	0.2900±0.2816	0.1200
K2O	0.2616±0.1533	0.2686±0.1391	-0.0100	0.3056±0.1735	0.2889±0.1682	-0.1702
K3N	0.2710±0.1620	0.2724±0.1392	-0.0300	0.2523±0.1102	0.2566±0.1156	0.0090
K4N	0.2590±0.1556	0.2627±0.1385	-0.0129	0.3542±0.1964	0.3037±0.1647	-0.2000
WEY	0.2577±0.1539	0.2628±0.1408	-0.0090	0.3611±0.1273	0.3998±0.1501	0.0189
LYM	0.2616±0.1587	0.2718±0.1387	-0.0499	0.5278±0.0393	0.5016±0.0023	-0.2656
JER	0.2877±0.1638	0.2907±0.1383	-0.0232	0.4259±0.1786	0.3780±0.1510	-0.2055
IRE	0.2928±0.1704	0.3012±0.1409	-0.0350	0.4167±0.2602	0.3833±0.2241	-0.0938

Table S5. Results of nucleotide BLAST search for the four outlier loci identified by BayeScan. No significant alignments were found (Searched 14-09-2019).

Locus	Sequence	Cover	e-value	identify
1091485	PREDICTED: Pomacea canaliculata zinc finger protein GLI2-like (LOC112556014), mRNA. Accession XM_025224617.1	21%	6x10^{-08}	84.13%
16027	PREDICTED: Rhinatrema bivittatum dynein axonemal heavy chain 6 (DNAH6), mRNA. Accession: XM_029602297.1	13%	0.200	87.18%
41214	Salarias fasciatus genome assembly, chromosome: 3. Accession: LR597438.1	97%	1x10^{-35}	73.41%
1362698	Salarias fasciatus genome assembly, chromosome: 14. Accession: LR597449.1	13%	0.056	91.43%