INTROGRESSION PATTERNS IN SCOTTISH BLUE MUSSEL (*Mytilus edulis*) POPULATIONS

Joanna Wilson

UNIVERSITY of STIRLING

Thesis submitted for the degree of Doctor of Philosophy

Faculty of Natural Sciences

Institute of Aquaculture

30th September 2016

Abstract

Background: The blue mussel, *Mytilus edulis* L., is an important contributor to the shellfish sector of Scottish aquaculture, with 7,270 tonnes worth £8.8 million being produced for the year 2015. Since 2010, production values have fluctuated as a result of inconsistent spat settlement, several business closures, and heightened levels of marine toxins in some areas. On Scotland's west coast, some farms (most notably Loch Etive) have suffered production losses from the appearance of non-marketable mussels with particularly fragile shells and poor quality meat.

Recent research has demonstrated that these undesirable traits have a genetic factor, linked to the presence of a non-native but related species *Mytilus trossulus* (Gould, 1850) and often its hybrids with the native *M. edulis. M. trossulus* has been classed as a commercially damaging species under Scottish law, but there is insufficient data on hybridisation and introgression patterns in Scottish mussel populations to evaluate any possible impacts this could have on production. Existing research has focused on single locus genotyping to identify *Mytilus* spp. and their hybrids in Scotland. By instead utilising multilocus genotyping, introgression could be identified and a better understanding of population structure could be gained, with implications for management to maintain productivity and profitability.

The aim of the research presented here was to develop and validate a suite of new species diagnostic markers for multilocus genotyping of field populations of Scottish mussels, thereby establishing a more complete picture of the taxonomic relationships between species than previous studies have permitted.

Results: Analysis of SNPs identified with RADseq confirmed the presence of three genetically distinct *Mytilus* species in Scotland: *M. edulis*, *M. galloprovincialis* and *M. trossulus.* RADseq and KASP genotyping technology successfully identified and validated a suite of 12 highly robust diagnostic SNP markers for multilocus genotyping of *Mytilus* mussel populations. These markers permitted more comprehensive genotyping than previous studies had, allowing presumed pure

species individuals to be distinguished from first generation (F1) hybrids and introgressed (FX) genotypes in reference populations, and subsequently presented the possibility of exploring introgression in a wider scale study.

Multilocus genotyping of mussel populations from around Scotland revealed widespread introgression of *M. edulis* with both *M. galloprovincialis* and *M. trossulus.* No pure *M. galloprovincialis* was identified and pure *M. trossulus* was restricted to a single site in Loch Etive, possibly part of a relict population. F1 hybrids between *M. edulis* and *M. trossulus* were identified in Loch Etive and in Loch Fyne on the west coast. This was evidence of ongoing hybridisation and suggested an active hybrid zone existed in Scotland, something that previous single locus genotyping studies had not acknowledged. A link between shell fragility and *M. trossulus* introgression was recognised at a single site outside of Loch Etive, but this was not apparent anywhere else and the actual causes of shell fragility remain unevaluated. There was a clear difference between the genetics of most farmed stock and wild populations, which indicated an anthropogenic effect on introgression and subsequent species composition, and had implications for future farm site selection and broodstock sourcing.

Temporal species composition in Loch Etive differed over a short time period, but high proportions of *M. trossulus* alleles were observable some 25 months after a major fallowing event had taken place. Pure *M. trossulus* was also identifiable, which was consistent with the presence of an established population of *M. trossulus* existing in this area.

Conclusion: Multilocus genotyping has produced a more in depth picture of species diversity in Scottish mussel populations. SNP assays revealed widespread introgression between three genetically distinct species – *M. edulis, M. galloprovincialis* and *M. trossulus* – and furthermore recognised that, to date, single locus genotyping has overestimated the abundance of pure *Mytilus* mussels in Scottish waters. However, this hitherto unidentified genetic complexity does not appear disadvantageous to mussel production, despite the prevalence of *M. trossulus*

introgression among farmed populations, and it is somewhat unlikely that genetics are the sole cause of undesirable shell characteristics among *Mytilus* spp. mussels.

Declaration

I, Joanna Wilson, hereby certify that this thesis is a unique piece of work that has been composed entirely by myself; which embodies the results of my own research; and which has not been submitted for any other higher degree or qualification. All sources of information have been suitably acknowledged in the text.

Acknowledgements

A special thanks to my parents and sisters, and to all of my friends (old and new; Cricetid and Canid pets included) for their invaluable support during my PhD. I am grateful for your patience and perseverance with this emotional wreck during what has felt like a tremendously long slog: I do not take it for granted how important you have been in helping me stay focused and motivated, and however clichéd it sounds I know that I could not have done this without you!

Thanks to my supervisors for their guidance and encouragement throughout the study, and thanks to MASTS and Marine Scotland Science for their financial support. Additional thanks to all farmers and researchers who provided me with specimens: your insights into the shellfish industry and the associated science have been most illuminating to someone who, before starting this PhD, knew absolutely nothing about mussels and had only worked with small, furry mice.

The list of other people who have helped me out in one way or another, and the list of things that they have helped me with, would fill another thesis in itself. To save myself that pain, I shall instead conclude by simply thanking everyone else who has helped me with even the most mundane of tasks during my PhD, and let you know that I have appreciated your assistance in whatever form it took.

Table of Contents

List of Figures i
List of Tables iv
List of Equations vi
Abbreviations and Acronyms vii

CHAPTER 1 – GENERAL INTRODUCTION
1.1. BACKGROUND 1
1.1.1. Shellfish aquaculture in Scotland 1
1.1.2. *Mytilus* spp. life history 4
1.2. HYBRIDISATION AND INTROGRESSION 7
1.2.1. Hybrid zones 8
1.2.2. Barriers to hybridisation and introgression 10
1.3. GENOTYPING FOR SPECIES IDENTIFICATION 11
1.3.1. Genotyping the *Mytilus edulis* species complex 12
1.3.1.1. Allozymes 12
1.3.1.2. Single locus DNA-based markers 13
1.3.1.3. Microsatellites and SNPs 14
1.3.1.4. Mitochondrial DNA markers 15
1.3.2. Multilocus genotyping 16
1.4. DNA SEQUENCING AND GENOTYPING 17
1.5. METHODS FOR POPULATION GENETIC ANALYSIS 19
1.6. THESIS AIMS AND OBJECTIVES 20

CHAPTER 2 – GENERAL MATERIALS AND METHODS
2.1. SAMPLE COLLECTION 22
2.2. DNA EXTRACTION, QUALITY ASSESSMENT AND CLEANUP 25
2.2.1. Automated DNA extraction 25
2.2.2. Manual DNA extraction 26
2.2.2.1. SSTNE/SDS DNA extraction 27
2.2.2.2. RealPure DNA extraction 28
2.2.3. DNA quantification and quality assessment 28
2.2.3.1. Spectrophotometry 28
2.2.3.2. Gel electrophoresis 28
2.2.3.3. Additional cleanup 29
2.3. INITIAL TAXONOMIC GENOTYPING 29
2.3.1. Me15/16 PCR 29
2.3.2. Glu5’ PCR 30
2.4. RAD LIBRARY CONSTRUCTION AND SEQUENCING 31
2.4.1. Complete protocol for building RAD libraries 31

2.5. SEQUENCE ANALYSIS 39
2.5.1. *de novo* genome assembly with *Stacks* 39
2.5.2. Multivariate data analysis 44
2.6. SNP ASSAYS 45
2.6.1. KASP genotyping technology 45
2.6.2. Genotyping with SNP assays 46
2.6.3. Analysis of genotyping data 49
2.6.3.1. Allele frequencies and relative proportion of diagnostic alleles 49
2.6.3.2. Inferring population structure with STRUCTURE 50
2.6.3.3. Inferring population structure with NEWHYBRIDS 52
2.6.3.4. Phylogenetic analysis 53

CHAPTER 3 - USE OF RAD SEQUENCING TO IDENTIFY SPECIES-DIAGNOSTIC SNPS WITHIN THE "*MYTILUS EDULIS* SPECIES COMPLEX"
3.1. INTRODUCTION 55
3.2. METHODS 60
3.2.1. Sample collection 60
3.2.2. DNA extraction and preliminary PCR 62
3.2.3. RAD library preparation and sequencing 62
3.2.4. Data analysis 63
3.2.4.1. *de novo* genome assembly 63
3.2.4.2. Phylogenetic analysis 64
3.2.4.3. PCA, DAPC and allele frequencies 65
3.2.5. Selection of SNPs for assay design 65
3.3. RESULTS 66
3.3.1. RAD library preparation and sequencing 66
3.3.2. Sequence analysis 67
3.3.2.1. Number of assembled loci 67
3.3.2.2. Identifying loci for marker design 67
3.3.2.3. F_{ST} values 72
3.3.2.4. Other potentially informative loci 73
3.3.3. SNP assay optimisation and validation 73
3.3.4. SNP genotyping 75
3.3.4.1. PCR conditions 75
3.3.4.2. Genotype class and Type 75
3.3.5. Genotypes per site 77
3.3.5.1. Individual Type: pure species or introgressed 77
3.3.5.2. Genotypes with Me15/16 and SNP assays 78
3.3.6. Analysis of genotyping data 80
3.3.6.1. PCA and DAPC analysis 80
3.3.6.2. Allele frequencies 81

3.3.6.3. Genotype compositions per population sample 83
3.4. DISCUSSION 84
3.5. CONCLUSIONS AND SUMMARY 90

CHAPTER 4 – ASSESSING LEVELS OF GENETIC ADMIXTURE OF *MYTILUS EDULIS* WITH CONGENERIC SPECIES *M. GALLOPROVINCIALIS* AND *M. TROSSULUS*
4.1. INTRODUCTION 91
4.2. METHODS 95
4.2.1. Sample collection 95
4.2.2. Genotyping 97
4.2.2.1. Me15/16 PCR 97
4.2.2.2. SNP assays 97
4.2.3. Analysis of genotyping data 97
4.2.3.1. Inferring population structure 97
4.2.3.2. PCA, DAPC and allele frequencies 99
4.3. RESULTS 101
4.3.1. Genotypes per site 101
4.3.1.1. Individual Type: pure species or introgressed 101
4.3.1.2. Genotypes with Me15/16 and SNP assays 102
4.3.2. All genotypes 104
4.3.3. Admixture analysis 106
4.3.3.1. STRUCTURE 106
4.3.3.2. NEWHYBRIDS 111
4.3.4. *M. galloprovincialis* and *M. trossulus* introgression in Scotland 112
4.3.4.1. Relative proportion of diagnostic alleles 112
4.3.4.2. Shell fragility and *M. trossulus*. 114
4.4. DISCUSSION 116
4.5. CONCLUSIONS AND SUMMARY 125

CHAPTER 5 – TEMPORAL DISTRIBUTION OF *MYTILUS* SPP. MUSSELS IN A SCOTTISH LOCH
5.1. INTRODUCTION 127
5.2. METHODS 130
5.2.1. Sample collection and genotyping 130
5.2.2. Inferring population structure with STRUCTURE 131
5.2.3. PCA and DAPC analysis 131
5.2.4. *M. trossulus* genetic contribution 132
5.3. RESULTS 132
5.3.1. Single and multilocus genotyping 132
5.3.1.1. Me15/16 PCR 132
5.3.1.2. SNP assays 133

5.3.1.3. Individual Type: pure species or introgressed 134
5.3.2. Inferring population structure with STRUCTURE 134
5.3.3. DAPC analysis 136
5.3.4. *M. trossulus* genetic contribution 136
5.4. DISCUSSION 137
5.5 CONCLUSIONS AND SUMMARY 140

CHAPTER 6 – GENERAL DISCUSSION AND CONCLUSIONS
6.1. OVERALL CONCLUSIONS 142
6.2. DIRECTIONS FOR FUTURE RESEARCH 149

BIBLIOGRAPHY 152

APPENDIX 1 172

APPENDIX 2 173

APPENDIX 3 175

APPENDIX 4 182

APPENDIX 5 187

APPENDIX 6 189

APPENDIX 7 196

APPENDIX 8 203

List of Figures

CHAPTER 1
FIGURE 1.1 – Diagram of *M. edulis* lifecycle 5

CHAPTER 2
FIGURE 2.1 - Map of all named sampling sites for this study 23

FIGURE 2.2 - Diagram of a dissected mussel and the regions where tissue samples were taken for manual DNA extractions 26

FIGURE 2.3 - 0.8% 0.5X TAE agarose gel showing *Mytilus* DNA samples of high quality and high molecular weight 29

FIGURE 2.4 - A 2% 0.5X TAE agarose gel image showing results of PCR with the Me15/16 primer set 30

FIGURE 2.5 – 1.5% 0.5X TAE vagarose gel showing shearing results of four libraries (L1 – L4) 34

FIGURE 2.6 – Size selection of DNA fragments from libraries on 1.1% 0.5X TAE agarose gel 36

FIGURE 2.7 – 1.1% 0.5X TAE agarose gel showing amplified PCR products comprising the final libraries (L1 – L4) sent for outsource sequencing 38

FIGURE 2.8 - Stylised representation of the "Catalog view" in the *Stacks* web interface 43

FIGURE 2.9 – Example of a diagnostic biallelic *Mytilus* species locus with a single T/C SNP 43

FIGURE 2.10 – Model Cartesian plot of fluorescence values generated by SNP assay 47

FIGURE 2.11 - Possible allelic combinations at a single biallelic locus in pure individuals, F1 hybrids and introgressed (FX) hybrids 48

CHAPTER 3
FIGURE 3.1 - Map of sampling sites chosen as sources of presumed pure specimens, based on historical morphological and genetic evidence 61

FIGURE 3.2 - 2% 0.5X TAE agarose gel image showing results of PCR with the Me15/16 primer set when pure species and their hybrids are present 63

FIGURE 3.3 – Radial phylogenetic tree of 40 individuals constructed using composite genotypes of 362 SNPs at 349 biallelic RAD loci 68

FIGURE 3.4 – DAPC scatterplot of clusters generated by PCA of composite genotypes of 362 SNPs at 349 biallelic RAD loci 69

FIGURE 3.5 - DAPC scatterplot of clusters generated by PCA of composite genotypes of 38 individuals at 18 loci (RAD data) 70

FIGURE 3.6 – DAPC scatterplot of clusters generated by PCA of composite genotypes of 40 individuals with 12 SNP assays 74

FIGURE 3.7 – The proportions of putatively pure and introgressed individuals, detected with multilocus genotyping using 12 SNP assays 78

FIGURE 3.8 – Genotypes of *Mytilus* individuals generated from (A) single locus analysis with Me15/16 and (B) multilocus genotyping with 12 SNP markers..... 79

FIGURE 3.9 – DAPC scatterplot of clusters generated by PCA of composite genotypes of 228 individuals with 12 SNP assays 80

CHAPTER 4

FIGURE 4.1 – Map of sampling site locations in Scotland 96

FIGURE 4.2 – The proportions of presumed pure, F1 hybrids and introgressed individuals at 22 Scottish sites, detected with multilocus genotyping using 12 SNP assays 102

FIGURE 4.3 – Genotype classes per site, with data from (A) single locus genotyping with Me15/16 and (B) multilocus genotyping with 12 SNP assays . 103

FIGURE 4.4 - Total numbers of all genotype classes identified among 22 Scottish population samples (*n*=991) after multilocus genotyping with 12 SNP assays .. 104

FIGURE 4.5 - DAPC scatterplot of clusters generated by PCA of composite genotypes of 991 individuals at 12 biallelic loci 105

FIGURE 4.6 - Structure plots constructed using the Admixture Ancestry Model with independent allele frequencies per population [*K*=4 (Δ*K* = 64.706, determined from 100 iterations using Evanno's method (2005)), burnin = 10,000, reps = 10,000], showing the genetic composition of each site sampled 108

FIGURE 4.7 – Map of sampling sites colour-coded according to group 109

FIGURE 4.8 – NEWHYBRIDS classifications for (A) *M. edulis, M. galloprovincialis* and their hybrids, and (B) *M. edulis, M. trossulus* and their hybrids, with posterior probabilities for each given genotype class 112

FIGURE 4.9 – Bar graphs showing (A) species composition and (B) The relative proportions of *M. edulis*, *M. galloprovincialis* and *M. trossulus* alleles amongst strong and fragile shelled mussels from Site X in Scotland 115

CHAPTER 5

FIGURE 5.1 – Map showing the location of Loch Etive in Scotland. The sampling site at Achnacloich is labelled and marked with a "*" symbol 130

FIGURE 5.2 – Bar graph showing species composition by sampling date after genotyping with (A) Me15/16 and (B) SNP assays..... 133

FIGURE 5.3 – Bar graph showing proportions of pure individuals [*M. edulis* (*Me*) and *M. trossulus* (*Mt*)] and different types of hybrid [F1 and introgressed (FX)] at each sampling date..... 134

FIGURE 5.4 – Structure plots constructed using the Admixture Ancestry Model with independent allele frequencies per population [*K*=2 (Δ*K* = 37.974, determined from 100 iterations using Evanno's method (2005)), burnin = 10,000, reps = 10,000], showing the genetic composition of temporal samples from Loch Etive..... 135

FIGURE 5.5 – DAPC scatterplot of temporal genetic data from Loch Etive 136

List of Tables

CHAPTER 2

TABLE 2.1 - Details of all sites sampled in this study 24

TABLE 2.2 - Adapter key for RAD library construction 32

TABLE 2.3 – Mastermix volumes for step 2 (restriction enzyme digestion) and step 3 (P1 adapter ligation) of RAD library construction protocol 33

TABLE 2.4 – Volumes of reagents used in bulk PCR mastermix preparation, corresponding to the number of samples in the RAD library 37

TABLE 2.5 – Summary of parameters used in scripts for building *Stacks* pipelines 42

TABLE 2.6 - Summary of parameters used in the *Stacks* script *export_sql.pl* 44

TABLE 2.7 - Genotype frequency classes assumed for a NEWHYBRIDS model with two generations of potential inbreeding 52

CHAPTER 3

TABLE 3.1 – Details of sampling sites and the numbers of *Mytilus* individuals (*n*) collected for diagnostic marker development and validation 62

TABLE 3.2 – Results of preliminary single locus genotyping with the Me15/16 primer set 66

TABLE 3.3 – Average number of loci per species generated through *de novo* assembly of RAD tags 67

TABLE 3.4 – Primer sequences for SNP assays, corresponding to 18 biallelic loci identified with RADseq 71

TABLE 3.5 – F_{ST} values for 18 loci chosen for marker design, generated using the *Stacks* Population (Pop) module 72

TABLE 3.6 - Percentages of matching RAD and KASP genotyping calls per locus 73

TABLE 3.7 – Genotypes of presumed pure individuals, F1 hybrids and introgressed individuals after genotyping with 12 diagnostic SNP assays 76

TABLE 3.8 – Allelic frequencies per locus per population, calculated with GENALEX 82

TABLE 3.9 – p values generated per locus per population with Hardy Weinberg exact tests with GENEPOP 83

TABLE 3.10 – Numbers of unique and shared composite genotypes in populations used for marker development and validation 83

CHAPTER 4

TABLE 4.1 – Sampling site details 95

TABLE 4.2 - Genotype frequency classes assumed for a NEWHYBRIDS model for *M. edulis, M. galloprovincialis* and their hybrids 99

TABLE 4.3 - Genotype frequency classes assumed for a NEWHYBRIDS model for *M. edulis, M. trossulus* and their hybrids 99

TABLE 4.4 - Types and Genotype classes (Pure and hybrid) according to genotyping results with 12 SNP assays 101

TABLE 4.5 – Average membership proportion (q) of each pre-defined population in each of the four clusters assigned by STRUCTURE 107

TABLE 4.6 - Allele frequency divergence (net nucleotide distance) among *K* clusters, calculated using STRUCTURE 110

TABLE 4.7 - Comparison of two methods measuring the levels of *M. galloprovincialis* and *M. trossulus* introgression at each sampling site 113

CHAPTER 5

TABLE 5.1 – Average membership proportion (q) of temporal samples in each of the 2 clusters assigned by Structure software 135

TABLE 5.2 - Comparison of two methods measuring the levels of *M. trossulus* introgression in temporal samples from Loch Etive 137

List of Equations

CHAPTER 2

EQUATION 1 – Equation for calculating genotype frequency 49

EQUATION 2 – Equation for calculating allele frequency 49

EQUATION 3 – Equation for estimating the proportion of introgression within a population sample ..... *50* (100)* 132

*[EQUATION 3 is used in CHAPTER 2, CHAPTER 4 and CHAPTER 5 of the thesis. The page number corresponding to CHAPTER 2 is in *italics;* the page number corresponding to CHAPTER 4 is in (round brackets); the page number corresponding to CHAPTER 5 is underlined]

Abbreviations and Acronyms

Below is a list of the most commonly used abbreviations and acronyms in the text

F1 hybrid	First generation hybrid
F1 *MeMg*	First generation hybrid of *M. edulis* and *M. galloprovincialis*
F1 *MeMt*	First generation hybrid of *M. edulis* and *M. trossulus*
F1 *MgMt*	First generation hybrid of *M. galloprovincialis* and *M. trossulus*
FX hybrid	Introgressed hybrid (second generation or beyond)
HTS	High Throughput Sequencing
KASP	Kompetitive Allele Specific PCR
Me	Individual with the *M. edulis* genotype class
Mg	Individual with the *M. galloprovincialis* genotype class
Mt	Individual with the *M. trossulus* genotype class
MeMg	Introgressed hybrid of *M. edulis* and *M. galloprovincialis*
MeMt	Introgressed hybrid of *M. edulis* and *M. trossulus*
MgMt	Introgressed hybrid of *M. galloprovincialis* and *M. trossulus*
MeMgMt	Introgressed hybrid of *M. edulis*, *M. galloprovincialis* and *M. trossulus*
HXE	Introgressed hybrid with a confirmed allelic contribution from *M. edulis* only
HXG	Introgressed hybrid with a confirmed allelic contribution from *M. galloprovincialis* only
HXT	Introgressed hybrid with a confirmed allelic contribution from *M. trossulus* only
mtDNA	Mitochondrial DNA
RAD	Restriction Site Associated DNA
RADseq	Restriction Site Associated DNA sequencing
SNP	Single Nucleotide Polymorphism
UPW	Ultrapure water

Chapter 1
General Introduction

1.1. BACKGROUND

1.1.1. Shellfish aquaculture in Scotland

On a global scale, aquaculture contributes to approximately 50% of the world's fish food supply and is valued in excess of £63 billion (Bostock *et al.,* 2010). Around 1% (£650 million) of this total revenue comes from shellfish aquaculture, which produces approximately 12 million tonnes per annum (Pawiro, 2010). Annual production of shellfish in the UK was approximately 27,100 tonnes for the year 2011, and was valued at £19.1 million. *Mytilus edulis* (blue mussel) (Linneaus, 1758) and *Crassostrea gigas* (Pacific oyster) (Thunberg, 1973) are the two main contributors to UK shellfish production that are grown in Scotland, England, Wales and Northern Ireland. *Ostrea edulis* (native oyster) (Linneaus, 1758) is grown in Scotland, England and Wales, while *Pecten maximus* (king scallop) (Linneaus, 1758) and *Aequipecten opercularis* (queen scallop) (Linneaus, 1758) target smaller niche markets in Scotland and Northern Ireland. Aquaculture of *Cerastoderma edule* (common cockle) (Linneaus, 1758) is restricted to England only (SEAFISH, 2012).

Total revenue from Scottish shellfish farming was estimated at £10.1 million for the year 2015, the bulk of which was generated from *M. edulis* production: 7,270 tonnes of mussel worth £8.8 million was produced for the table, and 77% of all mussel farming took place in The Shetland Isles. Despite high production for 2015, this was actually a 5% decrease from production in 2014, which saw Scottish mussel production reach an all-time high of 7,863 tonnes. Indeed, yearly production values in Scotland have fluctuated over the last five years due to inconsistent yields and variable market values. In 2010, production reached a high point of 7,199 tonnes but, by 2012, had suffered a 13% decline to 6,277 tonnes (Munro and Wallace, 2016). The higher tonnage values in 2013 (6,757 tonnes) and 2014 suggested the industry was recovering following this earlier decline. However, understanding the causes of observed production declines will be crucial for updating management practices to ensure sustainable growth of the Scottish shellfish industry in future.

A range of approaches to mussel aquaculture are carried out worldwide. These are generally split into two categories: bottom culture, where mussels are grown directly on the seabed; and off-bottom culture, where mussels are constantly submerged and suspended in the water column. Mussel farming in Scotland is generally off-bottom and depends on the settlement of mussel larvae (spat) on ropes suspended from moored rafts (McKindsey *et al.,* 2011). Spat can settle naturally on outdoor growing systems or ropes can be imported, either from another farm with outdoor growing systems or from a hatchery (FAO, 2012a). In Scotland, low levels of spat settlement with temporal and environmental variation, and importing of poor quality spat have been partly responsible for production fluctuations since 2010. A number of business closures throughout the country and heightened levels of marine toxins in some areas have contributed to production declines (Mayes and Fraser, 2012). Production losses in some areas of the west coast (most notably in Loch Etive) have also arisen from the appearance of undesirable, non-commercial traits amongst mussels: thin, fragile shells and poor quality meat (Gubbins *et al.,* 2012). Although the exact causes of shell fragility and poor quality meat have not been ascertained, previous research has demonstrated such undesirable traits have a genetic factor, arising from the presence of a non-native, related species *Mytilus trossulus* (the Baltic mussel) (Gould, 1850), and sometimes from its hybrids with the native *M. edulis* (Beaumont *et al.,* 2008; Dias *et al.,* 2008; Dias *et al.,* 2011a; Gubbins *et al.,* 2012). Undesirable traits arising from hybridisation have also been recorded in finfish aquaculture; for instance, hybrids often have deformed gill rakers which makes feeding less efficient, and ultimately results in reduced growth and overall lower meat yield. Gill malformations from interspecies hybridisation has been observed in several hybrid cyprinids, such as hybrids between *Hypophthalmichthys molitrix* (silver carp) (Valenciennes, 1844) and *Aristichthys nobilis* (bighead carp) (Richardson, 1845) (Bartley *et al.,* 2001; Battonyai *et al.,* 2015); hybrids of *A. nobilis* with *Ctenopharyngodon idellus* (grass carp) (Valenciennes, 1844) (Berry and Low, 1970); and hybrids of *Campostoma anomalum* (central stoneroller) (Rafinesque, 1820) with *Luxilus* spp. (commonly known as highscale shiners) (Poly, 1997).

Despite the negative effects it can bring to aquaculture, managing hybridisation in an open marine environment presents a huge challenge for industry stakeholders.

Once a species and its hybrids have become established in an area, ocean currents have the potential to transport larvae far and wide. Even if the source area is targeted, it is impossible to predict where larvae may have spread and which other areas could potentially be at risk from undesirable species and their hybrids (Thresher and Kuris, 2004). In Scotland, attempts have been made to manage undesirable *M. trossulus*. In 2013, *M. trossulus* was classed as a commercially damaging species under "The Aquaculture and Fisheries (Scotland) Act (2013)", with the aim of reducing its spread and mitigating the damage that it could bring to Scottish mussel farming. The legislation stipulates that any suspected or confirmed presence of *M. trossulus* on a farm should be reported to the relevant authorities and dealt with accordingly. However, there are no guidelines for the management and control of *M. trossulus* hybrids because this is substantially more difficult to regulate due to a lack of information on the potential impact of hybrids outside of studies in Loch Etive (e.g., Beaumont *et al.*, 2008; Dias *et al.*, 2009a; Dias *et al.*, 2009b; Zbawicka *et al.*, 2010). Additionally, as observed in other bivalve species (e.g., Lydeard *et al.*, 1996; Roe and Lydeard, 1998; Baker *et al.*, 2003; Huff *et al.*, 2004), phenotypic plasticity often hinders the positive identification of hybrid forms. To date, studies of *Mytilus* populations in Scotland have focused on identifying *M. edulis, M. trossulus* and their hybrids through genotyping at a single locus (Beaumont *et al.*, 2008; Dias *et al.*, 2011a). These studies have not, however, examined the extent of introgression (i.e., genetic mixing from repeated backcrosses) between the two species and, subsequently, the impact this could have on Scottish aquaculture has not been properly evaluated. The spread of *M. trossulus* and its hybrids, and any undesirable traits associated with them, could be problematic for farmers if it contributed to further production decreases. It is therefore essential that patterns of genetic mixing amongst Scottish mussel populations, via hybridisation and introgression, be analysed in greater detail.

A third species of *Mytilus* mussel, *Mytilus galloprovincialis* (Mediterannean mussel) (Lamarck, 1819), has also been identified in Scottish waters. *M. galloprovincialis* is widely cultured, particularly in the Mediterranean; it frequently hybridises with *M. edulis* but hybrids between *M. galloprovincialis* and *M. trossulus* are considered much rarer (Dias *et al.*, 2011b). Hybrid *M. galloprovincialis* forms are

not known to cause any problems for aquaculture (Beaumont *et al.*, 2007; Dias *et al.*, 2009a).

1.1.2. Mytilus *spp. life history*

Due to their morphological and genetic similarities, *M. edulis*, *M. galloprovincialis* and *M. trossulus* are often grouped together in the "*Mytilus edulis* species complex" (e.g., Gardner, 1996; Rawson *et al.*, 1996; Brooks, 2000; Gardeström *et al.*, 2008). Among molluscs, species complexes have also been identified in gastropods such as *Littorina* spp. snails (Warwick *et al.*, 1990) and *Crepidula* spp. slipper shells (Collin, 2000), and in bivalves including *Crassostrea* spp. oysters (Ren *et al.*, 2016); *Brachidontes exustus* (the scorched mussel) (Linneaus, 1758) (Lee and Foighil, 2004); and *Vesicomya* spp. clams (Goffredi *et al.*, 2003). Mussels in the *M. edulis* species complex are eurytopic and occupy a wide range of intertidal habitats, ranging from sheltered to exposed, gravelly to rocky substrates (Hepper, 1957). Distributions are largely influenced by temperature and salinity: generally, *M. edulis* and *M. galloprovincialis* prefer warm waters with high salinities, whereas *M. trossulus* can tolerate much lower salinities in cooler waters (Brooks, 2000). A widespread distribution across habitats with extensive spatial and temporal variation affects both phenotype and genotype. Shell morphology is highly plastic and subject to the influence of multiple factors, including temperature; salinity; water depth and patterns of water flow; degree of wave exposure and shelter; food availability; oxygen availability and optical density; and pollution levels (Hepper, 1957; Seed, 1968; Widdows and Johnson, 1988; Daguin *et al.*, 2001). Despite extensive phenotypic overlap, each species has some distinguishing morphological characteristics that can sometimes aid species identification among adult forms. Typically, *M. edulis* has a finely lined and triangular blue, black or brown shell that reaches up to 10cm in length (Newell, 1989); *M. galloprovincialis* has a roughly lined and triangular grey, blue or black shell that reaches up to 15cm in length (FAO, 2012b); while *M. trossulus* has a smoother, elongated black or brown shell that is between 7–11cm in length (Cowles, 2005). However, using the size or shape of shell for species identification is unreliable because mussel growth depends on many factors, including age or density of mussels within a bed (Seed, 1968). The

overall ecosystem quality also influences growth: faster growth rates and larger sizes tend to be associated with favourable conditions, whereas less favourable conditions promote slower growth rates and smaller sizes (Widdows and Johnson, 1988).

Mytilus spp. mussels have a high reproductive capacity and free-living, motile larvae (FAO, 2012a) (FIGURE 1.1). The settlement of mussel larvae (spat) is of

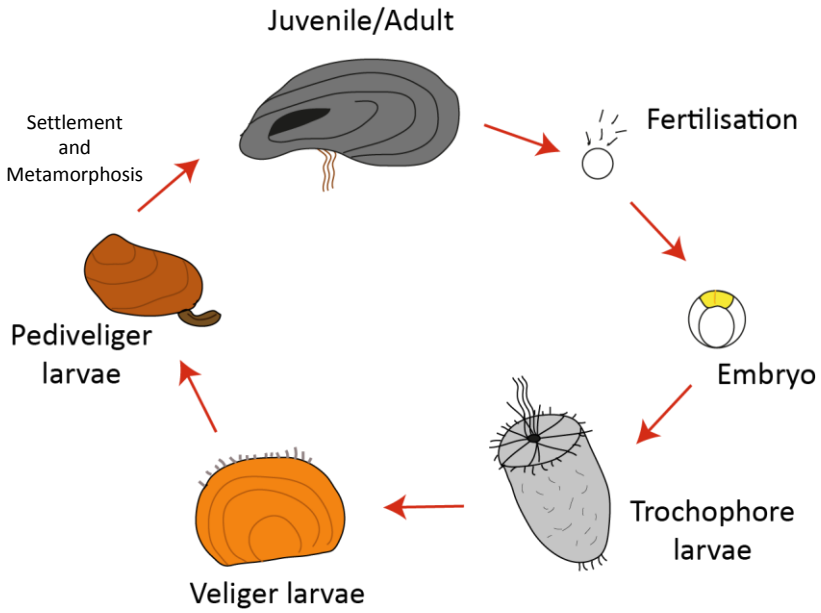

FIGURE 1.1 – Diagram of *M. edulis* lifecycle, showing development from fertilisation to metamorphosis. Together, all larval stages typically last 3-4 weeks before settlement takes place. Adult mussels can attain a marketable size anywhere between 18-36 months, depending on conditions

critical importance in aquaculture systems: natural spat is used most often as a source because it can be in high abundance and is easily translocated (Dare and Davies, 1975), but it is also possible to use spat produced in hatcheries. Hatcheries can supply farms with spat throughout the year, often much earlier than natural spat is available, and can supply spat that has been selectively bred to, for instance, grow more rapidly. In spite of these potential benefits, however, the use of hatchery reared spat is often limited by cost (Helm *et al.*, 2004), and it is yet to be exploited by farms in Scotland.

Mussels in the *M. edulis* species complex are dioecious. Although it may be possible to sex mussels by the colour of their gonadal tissue during the breeding season (pink-orange for females and white for males), this is otherwise impossible because there are no obvious signs of sexual dimorphism (Seed, 1969). Synchronous

spawning into the water column releases gametes which enables external fertilisation (Maloy, 2001). Spawning can take place in any month of the year and, most often, occurs once annually, but mussels can spawn multiple times if environmental conditions (e.g., water temperature and food availability) are favourable (Brooks, 2000). In non-hybridising populations, *M. edulis, M. galloprovincialis* and *M. trossulus* tend to have different peak spawning times: *M. edulis* tends to begin spawning in late April-May; *M. galloprovincialis* in July-August; and *M. trossulus* in July-October (Chipperfield, 1953; Seed, 1971; Toro *et al.,* 2002). There is often less difference in the peak spawning times of sympatric populations, leading to an overlap in gamete release and facilitating interspecies hybridisation (Maloy *et al.*, 2003).

Larval development begins after successful fertilisation, comprising two fully motile (non-feeding trochophore and feeding veliger) and one partially-motile (pediveliger) stage. The fully-motile larval stages typically last three-four weeks, during which time the major body parts (i.e., foot, digestive gland and gonad) begin to develop. After three-four weeks, veliger larvae are fully developed pediveligers that are ready to settle, a reversible stage of the mussel lifecycle that precedes metamorphosis. Pediveliger larvae drop out of the water column and onto a substrate, testing the surface with their sensory foot (Helm *et al.,* 2004). Pediveliger larvae can undergo a two-step settlement process, whereby they initially settle on filamentous substrates (e.g., *Polysiphonia* spp. algae or the byssus threads of adult mussels) and then detach and drift in the water column until they find adult beds, or they settle directly into adult beds. Primary settlement may avoid competition with adult mussels for food, or being inhaled by suspension-feeding adults. In the absence of a suitable substrate or conditions pediveligers can delay settlement; it is not uncommon for planktonic life to extend beyond a two-month period (Bayne, 1965), but larvae do become less selective of conditions the longer settlement is delayed (MarLIN, 2006). Once permanently settled, larvae begin metamorphosis. The factors triggering metamorphosis are poorly understood but it is thought to involve interactions between physical, chemical and biological cues (Helm *et al.,* 2004).

The minimum market size for mussels in the UK is 50 mm; this can be reached in less than 18 months for individuals grown in deeper water, but takes longer (24-36

months) if grown higher on the shore (Dare, 1980). Rope-cultured mussels are harvested through the use of a hydraulic powered system, and held in purified water for around 42 hours to remove any possible contamination. Automatic equipment in processing plants is then used to wash, separate and remove byssus threads from individuals before they are packed and sold to the appropriate distributor (Dare and Davies, 1975; Karayücel and Karayücel, 1999).

1.2. HYBRIDISATION AND INTROGRESSION

Natural hybridisation can be defined as the interbreeding of individuals from two distinct populations, which are distinguishable on the basis of at least one heritable characteristic (Harrison, 1990). Traditional "species boundaries" are based on genetic uniqueness and reproductive isolation of a particular population (Dobzhansky, 1935; Mayr, 1942). Thus, from a taxonomic perspective, hybridisation can be controversial because it raises doubts about what actually defines a "species" and, practically, makes identification very challenging (Barton and Hewitt, 1985). Due to extensive hybridisation between sympatric populations of *Mytilus* spp. mussels, there has historically been debate among taxonomists about the actual classification of *M. edulis*, *M. galloprovincialis* and *M. trossulus.* It has been suggested by some researchers that *M. galloprovincialis* and *M. trossulus* are instead subspecies, ecotypes or varieties of *M. edulis* (Seed, 1971; Gosling, 1984), rather than discrete species**.** However, other studies have shown that, regardless of their exact taxonomic status, each *Mytilus* type does exhibit some unique genetic characteristics that allow it to be considered a distinct entity (Koehn, 1991; McDonald *et al.*, 1991; Brooks, 2000; Riginos and Cunningham, 2005). Previous studies of Scottish *Mytilus* spp. populations have considered *M. edulis*, *M. galloprovincialis* and *M. trossulus* to be separate species according to results from single locus genotyping (Beaumont *et al.*, 2008; Zbawicka *et al.*, 2010; Dias *et al.*, 2011a), each with a different geographical origin as verified from mitochondrial DNA genotyping (Zbawicka *et al.*, 2010). Historically, *M. edulis* has an Atlantic origin (Vermeij *et al.*, 1991) and is thus considered a native species in the UK, whereas *M. galloprovincialis* and *M. trossulus* are considered to be non-native species with Mediterranean (Riginos and Cunningham, 2005) and Pacific (Beaumont

et al., 2008) origins respectively. The present study thus retained this convention for ease of comparison with historical data where appropriate, and to investigate the levels of genetic introgression within population samples.

Hybridisation is not uncommon in nature, estimated to take place in up to 10% of animal species and 25% of plant species (Mallet, 2005; Schwenk *et al.*, 2008; Twyford and Ennos, 2012). Hybridisation is a topic of interest among geneticists because of the potential impacts it can have on conservation and population management (Anderson and Thompson, 2002), and in some cases it has been suggested that hybridisation plays an important role in adaptation and speciation (Arnold, 1997). In nature, hybridisation can be most readily recognised when the ranges of previously allopatric populations come together in secondary contact, resulting in a new, sympatric population of interbreeding individuals (Harrison and Larson, 2014). Interbreeding between sympatric populations often results in the formation of a hybrid zone, wherein a complex mixture of parental genotypes, first generation (F1) hybrids, second generation (F2) and later hybrids, and backcross genotypes co-exist in varying proportions (Barton and Hewitt, 1985; Koehn, 1991). Natural hybrid zones have an enormous array of genotypes that have arisen from multiple (potentially hundreds to thousands) generations of genetic recombination. Natural hybrid zones are extremely valuable to population studies because such genetic diversity is not easily obtained from artificial crosses (Harrison and Larson, 2014). Introgression can be described as the stable integration of alleles from one species into the gene pool of a second, diverged species, usually via repeated backcrossing of fertile hybrids (Anderson and Hubricht, 1938; Rieseberg and Wendel, 1993).

1.2.1. Hybrid zones

Wherever the ranges of genetically distinct populations overlap and interbreeding produces offspring of mixed ancestry, a hybrid zone occurs (Barton and Hewitt, 1985). Hybrid zone structure and stability depends on the degree of genetic mixing via hybridisation and introgression that takes place between species. This is affected by environmental factors, dispersal distance of the species involved, and any selection acting on the genes that are exchanged (Harrison and Larson, 2014).

The mechanisms of selection in hybrid zones can be either endogenous (i.e., related to the relative fitness of parent species and their hybrid offspring) or exogenous (i.e., related to environmental adaptation), or a mixture of both (Hatfield and Schluter, 1999; Hilbish *et al.,* 2003). In a tension hybrid zone, hybrid offspring have reduced fecundity and viability in comparison to their parents, giving them very limited opportunity for spread. In this case, hybrid genotypes will largely be selected against while parental genotypes are favoured (Key, 1968). In a clinal hybrid zone, hybrid offspring have fitness equal to or exceeding that of their parents, but occupy their own niche separate from the parents (Endler, 1977). In a mosaic hybrid zone, parents and their hybrid offspring co-exist because neither displays an advantage over the other, resulting in equal selection pressures for both parent and hybrid individuals. Mosaic hybrid zones are widespread in nature and have been recognised in both terrestrial and marine environments [e.g., between species of field cricket in the genus *Gryllus* (Larson *et al.,* 2013); and between sympatric species of Cyprinid fish in the genus *Chondrostoma* (Costedoat *et al.,* 2005)]. In the marine environment, mosaic hybrid zones are likely to arise from differential larval settlement and adaptation in patchy coastal environments (Gilg and Hilbish, 2003; Smietanka *et al.,* 2004). Adaptation to patchy environments could explain why mosaic hybrid zones have been observed among wild populations of *Mytilus* mussels, because *M edulis, M. galloprovincialis* and *M. trossulus* all have wide, overlapping geographical ranges in a variety of habitats open to local selection (Varvio *et al.,* 1988; Riginos and Cunningham, 2005; Sousa *et al.,* 2013). There is evidence of mosaic hybrid zones between *M. edulis* and *M. galloprovincialis* along the Atlantic coast of Europe (Bierne *et al.,* 2003; Daguin *et al.,* 2001; Varela *et al.,* 2007) and on Irish coasts (Coghlan and Gosling, 2007; Gosling *et al.,* 2008; Doherty *et al.,* 2009); between *M. edulis* and *M. trossulus* in the North Atlantic (Innes and Bates, 1999; Toro *et al.,* 2003; Miranda *et al.,* 2010); and between *M. galloprovincialis* and *M. trossulus* along the Pacific coast of North America (Rawson and Hilbish, 1995; Rawson *et al.,* 1999).

1.2.2. Barriers to hybridisation and introgression

Introgression and hybridisation can be limited in hybrid zones due to various reproductive barriers that are largely based on environmental conditions (Toro *et al.*, 2002; Bierne *et al.*, 2006; Doherty *et al.*, 2009; Monteiro *et al.*, 2012).

Pre-zygotic reproductive isolation mechanisms include spatial and temporal separation of populations by ecological barriers; species-specific mating behaviour; and gamete incompatibility (Toro *et al.*, 2002). Reproduction by external fertilisation is common in the sea, presenting many more opportunities for hybridisation than on land (Bierne *et al.*, 2003; Miranda *et al.*, 2010). It is more difficult for ecological barriers to arise in a marine environment and, as such, species-specific mating behaviour (assortative mating and asynchronous spawning) and gamete incompatibility tend to occur more frequently than ecological barriers between marine organisms (Monteiro *et al.*, 2012).

Assortative (non-random) mating ensures that the genes of one species will not be mixed with another, and in mixed populations of sympatric species, asynchronous spawning aids in promoting assortative mating. Gamete incompatibility refers to the failure of sperm from one individual to fertilise the eggs of another, whether of the same or a different species (Rawson *et al.*, 2003). Successful fertilisation is dependent upon the ability of the egg and sperm to bind and fuse. In *Mytilus* species, it has been proposed that surface proteins on the egg and sperm are involved in gamete recognition. The sperm acrosome produces proteins (lysins) with species-specific sequences for gamete recognition. Lysins enable sperm to bind to and dissolve the protein coat surrounding the egg, thereby initiating fertilisation. Wherever gametes are notably different from each other, failure of the proteins on the sperm to recognise those on the egg could act as a barrier to fertilisation, thus limiting hybridisation and introgression between related species (McCartney and Lima, 2011). Depending on the environmental conditions, post-zygotic isolation mechanisms may be fully or partially effective, or ineffective. Although *Mytilus* species tend to spawn at different times of the year (temporal separation), there may be early or late spawners of each species that could potentially allow some hybridisation to occur (Brooks, 2000). In sympatric populations of different species, like those in the *M. edulis* species complex, gamete incompatibility is perhaps a more

likely cause of pre-zygotic isolation because the differences in the peak spawning times of different taxa are often only slight (Rawson *et al.,* 2003; Doherty *et al.,* 2009). In sympatric, hybridising populations, genetic compatibilities will be affected by habitat conditions and environmental influences, which will in turn affect how much hybridisation is prevented or facilitated (Miranda *et al.,* 2010).

Post-zygotic reproductive isolation takes place after successful fertilisation, typically through increased mortality during larval stages, reduced survival of later stage hybrids, or hybrid sterility (Doherty *et al.,* 2009). Hybrid larvae may have high mortality rates from genetic incompatibilities activated by changes in allele expression during development (Bierne *et al.,* 2003). Larvae that survive into adulthood could have a reduced chance of survival because they exhibit characteristics placing them at a competitive disadvantage against non-hybrids. Studies into the fitness of hybrid larvae tend to be conducted *in vitro* because it would be highly challenging to monitor the growth and development of individuals in the wild. For example, a study by Miranda *et al* (2010) compared the development and survival of pure *M. edulis* and pure *M. trossulus* larvae with hybrid larvae. Overall, hybrid crosses were found to have a lower proportion of normal growth in the first 72 hours after fertilisation, and after 10 days, larvae from hybrid crosses had 20% higher mortality rates than crosses of pure species. Abnormal growth in hybrid larvae could prevent effective resource exploitation and would thus be attributable to a decrease in fitness and survival, which would not be favoured by natural selection. While *in vitro* models for hybrid larval fitness are important in predicting the trends of wild populations, they should be treated with caution because controlled, artificial conditions will not completely reflect a dynamic natural environment (Toro *et al.,* 2012).

1.3. GENOTYPING FOR SPECIES IDENTIFICATION

Although *M. edulis*, *M. galloprovincialis* and *M. edulis* each have some distinguishing morphological characteristics that can occasionally be used for identification (see SECTION 1.1.2), such extensive phenotypic overlap with environmental variation and widespread hybridisation makes morphology a poor indicator of species ID (Koehn, 1991). DNA-based techniques are a far more reliable

and specific method of species identification than studying morphology alone (Knowlton, 2000; Capote *et al.*, 2012). A genetic marker is any heritable polymorphism in individuals or populations that, when measured, allows multiple important questions in population genetics and evolution to be answered (Davey *et al.*, 2011). Studies of genetic markers in various bivalve mollusc species [for instance, oysters in the genus *Crassostrea* (Wang *et al.*, 2014); and freshwater mussels of the families Unionoida (Vannarattanarat *et al.*, 2013) and Dreissenidae (Therriault *et al.*, 2004)], have been carried out to better define species boundaries in populations displaying extensive phenotypic overlap, with applications to both commercial and conservational issues.

1.3.1. Genotyping the Mytilus edulis *species complex*

Over the past three decades a range of species diagnostic markers (detailed below) have been developed to study *Mytilus* mussels, which have helped in resolving taxonomic issues where morphology has been unable to.

1.3.1.1. Allozymes

Allozymes are protein variants in enzymes that can be distinguished by gel electrophoresis, according to differences in size and charge that have arisen from amino-acid substitutions. Numerous different allozyme markers have been developed for *Mytilus* spp. genotyping and have, historically, been widely used in research (e.g., Hvilsom and Theisen, 1984; Varvio *et al.*, 1988; Koehn, 1991; McDonald *et al.*, 1991). Allozyme genotyping is inexpensive and can rapidly genotype large populations. However, some expression patterns can vary due to environmental effects and, because they are protein-based markers, may be insensitive in recognising DNA variation at the species level, which is a key disadvantage in their use (Schlötterer, 2004). Additionally, allozyme genotyping requires fresh tissue, which somewhat limits its flexibility for large scale experiments and surveys (Weising *et al.*, 2005; Zbawicka *et al.*, 2012).

1.3.1.2. Single locus DNA-based markers

DNA-based markers are more specific than allozymes because they detect variation at the genetic level, rather than differences in the proteins encoded by DNA, and also allow the number of mutations between different alleles to be quantified (Schlötterer, 2004). Several different nuclear DNA markers are available for genotyping *Mytilus* spp. mussels at a single locus. These are either diagnostic for one species [ITS (Heath *et al.,* 1995), diagnostic for *M. trossulus* only]; two species [Glu3' (Rawson *et al.,* 1996), distinguishes *M. edulis* from *M. galloprovincialis*]; or have a unique sequence for each of the three species in the *M. edulis* species complex [Me15/16 (Inoue *et al.,* 1995); Glu5' (Rawson *et al.,* 1996); Efbis (Bierne *et al.,* 2002)]. Single locus markers are named according to the locus or genomic region they correspond to: specific details are available in the referenced texts. Of all available single locus DNA-based markers, Me15/16, located in the gene encoding a polyphenolic foot protein, is the most routinely used marker for discrimination between *Mytilus* species and their hybrids. Many studies have used Me15/16, either alone or alongside allozymes and other single locus DNA-based markers, to genotype *Mytilus* populations worldwide. Species of the *M. edulis* species complex and their hybrids have been identified at various locations within Europe [e.g., Scotland (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a); Ireland (Gosling *et al.,* 2008); France (Bierne *et al.,* 2003), Iceland, The Netherlands, Poland (Smietanka *et al.,* 2004), Italy and Norway (Zbawicka *et al.,* 2012)], and outside of Europe [e.g., Canada (Smietanka *et al.,* 2013), New Zealand, Chile (Westfall *et al.,* 2010), and Japan (Inoue *et al.,* 1997)]. Although more informative to the point of species identification than morphological studies are, these studies remain very limited in resolving genetic structure within populations – i.e., they are unable to recognise whether or not introgression has taken place – because they have focussed only on a single locus or very low numbers of loci in the genome. Over generations, continued crosses between introgressed individuals leads to a series of complex genotypes, some of which can be unexpected or which remain challenging to identify (Patel *et al.,* 2015). Thus, assuming that introgression has taken place in *Mytilus* populations, screening of only a single marker will not have sufficient power to

recognise a possible multitude of different hybrid genotypes (Twyford and Ennos, 2012).

1.3.1.3. Microsatellites and SNPs

Microsatellite DNA is characterised by a variable number of short, repeating sequences at a given locus. The exact number or unique range of tandem repeats at a given locus may enable species to be distinguished from each other. Microsatellites are more widespread in a genome and are more species-specific than "traditional" single locus DNA-based markers, increasing their capacity for population genetic and mapping studies. However, microsatellites do tend to produce non-specific products which can lead to issues with automated scoring of alleles, potentially limiting their capacity for large scale genotyping (Schlötterer, 2004). The high level of polymorphism exhibited by many microsatellite loci often leads to alleleic overlap among species, reducing their potential as diagnostic markers. Some microsatellite markers for *Mytilus* spp. mussels have been developed [e.g., *M. edulis* (Lallias *et al.*, 2009), *M. galloprovincialis* (Li *et al.*, 2011); *M. trossulus* (Gardeström *et al.*, 2007)], but no evidence is available in external literature to indicate that these have been widely used in genotyping studies.

Single Nucleotide Polymorphisms (SNPs) are single base changes in a DNA sequence. SNPs are widespread throughout the genomes of all species (Vera *et al.*, 2010; Sharma *et al.*, 2012), present in both coding and noncoding regions of the genome. Their ubiquitous nature makes them easy to discover and thus, they are ideal for studying a non-model organism about which limited genetic information is available. SNPs can, in theory, have up to four alleles, each comprising a different base substitution (A, C, G or T) (Liu and Cordes, 2004), but they are most often biallelic in nature. Compared to microsatellites, SNPs are more suitable for large scale, automated genotyping because they have a lower mutation rate, making it easier to define a specific marker for species identification (Schlötterer, 2004). Some SNPs have been identified for genotyping *Mytilus* spp. mussels: e.g., Vera *et al* (2010) identified SNPs diagnostic for *M. galloprovincialis;* Zbawicka *et al* (2012) identified SNPs diagnostic for *M. edulis, M. galloprovincialis* and *M. trossulus;* and Wenne *et al* (2016) identified SNPs diagnostic to *M. edulis* and *M. trossulus*. Each of

these studies used a multicapillary electrophoresis approach to genotyping: Zbawicka *et al* (2012) and Wenne *et al* (2016) used a Sequenom MassARRAY iPLEX genotyping platform, capable of multiplexing up to 36 SNPs [see Gabriel *et al* (2009) for full details of SNP genotyping protocol]; whereas Vera *et al* (2010) used a SNaPshot multiplex kit (PE Applied Biosystems, Foster City, CA, USA), capable of multiplexing up to 10 SNPs.

Vera *et al* (2010) and Wenne *et al* (2016) focused on genotyping samples from Spain and Greenland respectively. Vera *et al* (2010) used *M. galloprovincialis* SNPs for parentage assignment in mussel breeding programmes; and Wenne *et al* (2016) used *M. edulis* and *M. trossulus* SNPs to identify hybridisation between the two species which, previously, had not been recognised. The study by Zbawicka *et al* (2012) was on a wider scale and genotyped mussels from 23 European locations, including individuals from Loch Etive in Scotland. Large scale genotyping revealed introgression that previous single locus genotyping studies had been unable to recognise. This was acknowledged as an important step forward in understanding *Mytilus* spp. distribution, and furthermore and important step in population management. However, there is no evidence indicating that these SNP markers have been used for further research. If the premise of the large scale study were to be continued, with focus restricted solely on Scottish population samples, multilocus SNP genotyping of *Mytilus* spp. individuals would provide a more detailed overview of genetic structure than existing studies have permitted: this would subsequently benefit Scottish aquaculture by allowing the potential impacts of *M. trossulus* and its hybrids on production to be assessed in greater detail.

1.3.1.4. Mitochondrial DNA markers

In addition to the use of nuclear DNA markers to establish species and hybrid identity, studies utilising mitochondrial DNA (mtDNA) markers have been carried out to establish the evolutionary origins of *Mytilus* species. Historically, *M. edulis* has an Atlantic origin (Vermeij *et al.,* 1991); *M. galloprovincialis* has a Mediterranean origin (Riginos and Cunningham, 2005); and *M. trossulus* has a Pacific origin (Beaumont *et al.,* 2008). MtDNA inheritance in *Mytilus* mussels is doubly uniparental, an unusual form of mtDNA inheritance with separate maternal

and paternal mitochondrial genomes (Rawson *et al.,* 1996). Females are homoplasmic, with a single maternally inherited mitochondrial genome; and males are heteroplasmic, with two mitochondrial genomes (female and male) inherited from their mother and father respectively (Wood *et al.,* 2003; Kenchington *et al.,* 2009). The exact mechanisms behind doubly uniparental inheritance in bivalves have not been fully evaluated because the rates of gene exchange in the female and male mitochondrial genomes differ (Riginos *et al.,* 2004; Kenchington *et al.,* 2009) and it is difficult to identify reliable markers for species identification (Smietanka *et al.,* 2004). However, as with nuclear DNA, there is evidence to suggest that hybridisation and introgression do occur between different *Mytilus* mitochondrial genomes (Kijewski *et al.,* 2006; Zbawicka *et al.,* 2007).

1.3.2. Multilocus genotyping

It is well documented that multilocus genotyping provides more detailed information about population structure (i.e., the degree of hybridisation and introgression) than single locus genotyping does (e.g., Storey *et al.,* 2005; Hayden *et al.,* 2008; Linnen and Hoekstra, 2009; Davey and Blaxter, 2010; Zuo *et al.,* 2014). In a commercial setting, understanding the extent of introgression is important for both population management and maintaining production efficiency: this has been widely studied in, for instance, salmonid aquaculture. Understanding the degree of introgression between farmed and wild salmon is crucial in studies of survival and development, and subsequently essential for population management (Utter *et al.,* 2001; Hansen *et al.,* 2001; Kruse *et al.,* 2000). In seven US states, there are now guidelines for the management of cutthroat trout populations, divided into categories based on degrees of introgression: those with "low" (< 10%) introgression are managed as native species, while those with "high" (>10%) introgression are managed as non-native species (Pritchard *et al.,* 2007).

Some SNP markers have been identified in *Mytilus* spp.: Vera *et al.,* 2010 (n=12); Zbawicka *et al.,* 2012 (n=21); Wenne *et al.,* 2016 (n=54). These, plus the majority of previously identified markers for studying the *Mytilus* genome, allow the simultaneous analysis of small panels of loci through gel-based methods [e.g., allozymes (McDonald *et al.,* 1991)], or through the generation of limited sequencing

data [e.g., microsatellites (Lallias *et al.,* 2009) with capillary electrophoresis (Jorgensen and Lukacs, 1983)]. However, such methods can become highly time consuming and laborious with increasing sample size: for instance, microsatellites often have different annealing temperatures which makes them difficult to multiplex, meaning multiple assays have to be run separately. Additionally, they may be applicable only to certain populations, and therefore unreliable for studying wild or highly diverged populations (Davey *et al.,* 2011). Use of a different marker that is more reliable for large scale genotyping and which can be easily multiplexed (such as a SNP) is preferred for modern day genomic studies.

1.4. DNA SEQUENCING AND GENOTYPING

As detailed in SECTION 1.3.1.3, there is some evidence of multilocus SNP genotyping in *Mytilus* spp. population studies in Europe (Vera *et al.,* 2010; Zbawicka *et al.,* 2012; Wenne *et al.,* 2016). All of these studies utilised capillary electrophoresis, a methodology developed by Jorgensen and Lukacs (1983) that offers an automated alternative to traditional gel electrophoresis (Karger and Guttman, 2009). DNA fragments are amplified with a fluorescent primer; PCR products are then loaded into individual capillaries and scanned with a laser to detect omitted fluorescence, which corresponds to fragment size. Although capillary electrophoresis has been widely used in genotyping studies (e.g., Vignal *et al.,* 2002; Goffredi *et al.,* 2003; Tanguy *et al.,* 2008; Lallias *et al.,* 2009), it can be limiting in terms of throughput and cost as the demand for greater volumes of genetic data increases (Chen and Sullivan, 2003; Morozova and Marra, 2008; Claesson *et al.,* 2010). High Throughput Sequencing (HTS) refers to rapid sequencing of large genomes, which can uncover hundreds of thousands of polymorphic markers in an individual even when little or no existing genetic information is available. HTS enables millions of DNA strands to be sequenced simultaneously and cost-effectively, and has become a routine part of modern biological research since its inception in the mid-2000s (Goodwin *et al.,* 2015). It has a wide range of applications, from whole genome sequencing to RNA expression profiling, and can be easily applied to wild populations which benefits conservation, genetics and ecology (Morozova and Marra, 2008; Davey *et al.,* 2011).

The earliest HTS technologies involved relatively short reads (<300 bp), but some technologies could sequence larger fragments (i.e., 454 pyrosequencing generates reads up to 1kb in length). As HTS technologies evolve, new platforms, such as PacBio and MinION sequencers, are addressing the need for longer read lengths (>50kb) in sequencing, but these are not yet capable of sequencing at the same depth as shorter read sequencers. Currently, Illumina has the largest market for short read sequencing platforms and its massively parallel sequencing by synthesis technology is the most commonly used worldwide. Illumina has a range of sequencing platforms, from low throughput benchtop sequencers (e.g., MiSeq) to ultra-high throughput instruments for whole genome sequencing (e.g., HiSeq X, which produces reads of up to 300 bp long). Although they have an overall high accuracy rate of > 99.5% (Goodwin *et al.,* 2015), Illumina platforms are, much like any sequencing technology, not completely infallible. Some of their main problems arise from an under-representation of AT and GC rich regions (Harismendy *et al.,* 2009), and a tendency towards calling false positives from sequencing errors (Minoche *et al.,* 2011).

Restriction Site Associated DNA (RAD) tags are short fragments of DNA situated on each side of a restriction enzyme site. RAD tags were first screened with microarray technology (Miller *et al.,* 2007), and were later adapted for simpler sequencing using HTS by Baird *et al* (2008). Multiplex sequencing of RAD tags with RADseq (Restriction Site Associated DNA sequencing) allowed for parallel screening of thousands of polymorphic markers and high throughput genotyping of large populations (Baird *et al.,* 2008). RADseq is ideal for large wild populations without a reference genome because it allows hundreds of thousands of markers to be scored accurately in most individuals, precisely estimating population parameters (Davey *et al.,* 2011). RADseq allows the identification of thousands of Single Nucleotide Polymorphisms (SNPs) in a genome (Catchen *et al.,* 2011; Davey *et al.,* 2011; Etter *et al.,* 2011). RADseq has been widely used in aquaculture to genotype species of commercial importance [e.g., identifying hybridisation in rainbow and westslope cutthroat trout (*Oncorhynchus* spp.) (Hohenlohe *et al.,* 2011); identifying Quantitative Trait Loci in Atlantic salmon (*Salmo salar*) (Linneaus, 1758) (Houston *et al.,* 2012); and linkage mapping of Nile tilapia (*Oreochromis niloticus*) (Linneaus,

1758) (Palaiokostas *et al.*, 2013a); and Atlantic halibut (*Hippoglossus hippoglossus*) (Linneaus, 1758) (Palaiokostas *et al.*, 2013b)], and there is some evidence of RADseq use in *Mytilus* species mussels (Peñaloza *et al.*, 2014; Araneda *et al.*, 2016).

1.5. METHODS FOR POPULATION GENETIC ANALYSIS

Genotypes, gene frequencies and DNA sequence polymorphisms in individuals and populations are the result of multiple, random influences and can therefore be challenging to study fully without the use of probabilistic models (Beaumont and Rannala, 2004). Statistical inference through Bayesian clustering methods is widely applied to population genetic data because it enables fairly straightforward implementation of complex models with multiple parameters (Beaumont and Rannala, 2004; Drummond and Rambaut, 2007), thereby allowing characteristic features of a given dataset to be examined in more detail: e.g., levels of admixture in a population sample [STRUCTURE (Pritchard *et al.*, 2000); and BAPS: Bayesian Analysis of Population Structure (Corander and Marttinen, 2006)]; classification of hybrid types [NEWHYBRIDS (Anderson and Thompson, 2002)]; variable migration rates between groups (Wilson and Rannala, 2003); and phylogenetic inference [BEAST: Bayesian Evolutionary Analysis by Sampling Trees (Drummond and Rambaut, 2007)]. Specifically in this study, STRUCTURE and NEWHYBRIDS were chosen for analysis.

STRUCTURE (Pritchard *et al.*, 2000) is the most commonly used statistical method for multilocus genetic analysis, and has been widely applied to genotyping studies of both model and real life datasets (e.g., Matsuoka *et al.*, 2002; Rosenberg *et al.*, 2002; Evanno *et al.*, 2005; Vähä and Primmer, 2006; Falush *et al.*, 2007; Hubisz *et al.*, 2009; Heled and Drummond, 2010; Marie *et al.*, 2012). The model underlying STRUCTURE was one of the first that allowed individuals to have admixed ancestry, with proportions of their genome originating from multiple subpopulations (Anderson, 2008). STRUCTURE assigns individuals probabilistically to subpopulations based on their genetic composition across multiple loci with dominant markers (e.g., SNP, microsatellite or RFLP markers) (Pritchard *et al.*, 2000). BAPS (Corander and Marttinen, 2006) also allows individuals to be of mixed ancestry and proportionately assigns individuals to genetic subpopulations (Latch *et al.*, 2006). BAPS uses a

slightly different algorithm than STRUCTURE that requires less computational resources and subsequently runs faster simulations (Corander and Marttinen, 2006). It has been demonstrated that both STRUCTURE and BAPS have similar statistical powers in detecting genetically differentiated groups in datasets (Latch *et al.*, 2006; Wilkinson *et al.*, 2011); however, the simpler algorithm employed by BAPS, which does not include the assumption that subpopulations are in Hardy Weinberg equilibrium, may affect the accuracy of cluster assignment when smaller numbers of loci are genotyped, in comparison to STRUCTURE (Rodriguez-Ramilo *et al.*, 2009).

NEWHYBRIDS (Anderson and Thompson, 2002) also allows individuals to have a mixed ancestry but, rather than focusing on the overall population sample like STRUCTURE and BAPS, instead focuses on individuals and estimates the probability of each belonging to distinct hybrid or purebred classes (Anderson and Thompson, 2002; Vähä and Primmer, 2006). NEWHYBRIDS has been widely used to analyse both model and real life datasets (e.g., Anderson and Thompson, 2002; Dudu *et al.*, 2011; Marie *et al.*, 2011; Cullingham *et al.*, 2012; Pujolar *et al.*, 2014; Kovach *et al.*, 2015; Patel *et al.*, 2015; Mckean *et al.*, 2016). NEWHYBRIDS is useful in a situation where populations are known to consist of pure individuals and recent hybrids of two species (Anderson and Thompson, 2002); and when looking to accurately assess the numbers of hybrids in a sample (Vähä and Primmer, 2006; Marie *et al.*, 2011).

Each Bayesian clustering method has a slightly different approach to modelling population genetic data, but there is no evidence to indicate that one consistently outperforms the other (Latch *et al.*, 2006; Vähä and Primmer, 2006; Marie *et al.*, 2011). Each approach is capable of providing relevant statistics to help answer biologically relevant questions depending on the aims of the study (Manel *et al.*, 2005; Rodriguez-Ramilo, 2009), which will ultimately influence model choice (Marie *et al.*, 2011).

1.6. THESIS AIMS AND OBJECTIVES

Within the last decade, single locus genotyping studies of Scottish mussel populations have identified the presence of the *M. edulis* species complex, and have acknowledged the presence of interspecies hybrids between *M. edulis*, *M. galloprovincialis* and *M. trossulus* (Dias *et al.*, 2008; Beaumont *et al.*, 2008;

Zbawicka *et al.,* 2010). However, there have been no studies examining genetic admixture (introgression) outside of Loch Etive (Zbawicka *et al.,* 2012) and, subsequently, the extent of hybridisation in Scotland remains unknown. Previous evidence indicates that pure *M. trossulus* negatively affects production by causing undesirable traits to appear in farmed individuals, and it has also been recognised that hybrids of *M. edulis* with *M. trossulus* can exhibit the same phenotype (Beaumont *et al.,* 2008; Dias *et al.,* 2011b). However, the actual effect of hybridisation on phenotype, and the subsequent effects on production, remains unevaluated, and there is no detailed information about species composition or population genetic structure in Scotland. Thus, a need has arisen to apply multilocus genotyping to better understand firstly, the extent of hybridisation in *Mytilus* spp. mussels in Scotland, and secondly how this could potentially influence future production of the industry by affecting broodstock sourcing. The three main aims of the research presented here were as follows:

1. To identify and validate a new set of species diagnostic (SNP) markers, capable of identifying introgression in *Mytilus* spp. mussels, based on RAD sequencing and KASP assay genotyping

2. To apply new species diagnostic markers to field samples of Scottish *Mytilus* spp. mussels to assess levels of introgression throughout the country

3. To measure temporal genetic patterns in Loch Etive

Chapter 2
General Materials and Methods

All practical work for the study was funded by Marine Alliance for Science and Technology for Scotland (MASTS) and Marine Scotland Science.

2.1. SAMPLE COLLECTION

Samples for this study were collected from a total of 25 named sites and a single unnamed Scottish site. These comprised a mixture of shoreline locations (wild and bottom grown aquaculture) and rope-sourced aquaculture sites (see FIGURE 2.1 for an overview of named site locations; and TABLE 2.1 for site names; a more detailed map of site locations and names is provided in CHAPTER 4, FIGURE 4.1). All wild mussels were collected from within the intertidal zone at low tide; distances from the shoreline were not recorded. Sites within Scotland were selected so as to cover as much of the coastline as possible. For commercial sensitivity issues, both rope and bottom grown aquaculture sites were chosen from areas with multiple businesses to keep actual farm locations anonymous, and the location of the farm at Site X was left undisclosed. The exception to this was the site at Loch Etive. No information about the depth that rope grown samples were taken from was provided. Wild shoreline sites were selected from areas without mussel farms, and two were chosen as a likely sources for pure, native *M. edulis*. Three sites outside of Scotland were selected as sources of pure, non-native species [*M. galloprovincialis* (Slovenia) and *M. trossulus* (North America and Canada)]. Between November 2012 and May 2016, live adult mussels (measuring at least 40 mm in length, as in Beaumont *et al.*, 2008) were obtained from 22 sites in Scotland [Dornoch Firth, Ferryness, Flotta, Kylesku, Loch Ailort, Loch Eireasort, Loch Fyne, Loch Laxford, Loch Linnhe, Loch Long, Loch Roag, Loch Ryan, Loch Spelve, Lunderston Bay, Montrose, Northside, Rascarrel Bay, Scapa Beach, Shetland BR, Shetland BX, St Andrews and Site X], one site in Slovenia (Bay of Piran; donated by Andreja Ramšak at NIB, Ljubljana) and one site in Canada (Bras d'Or Lake; donated by Barry MacDonald, Bedford Institute of Oceanography). In April 2013, September 2013 and November 2014, juvenile

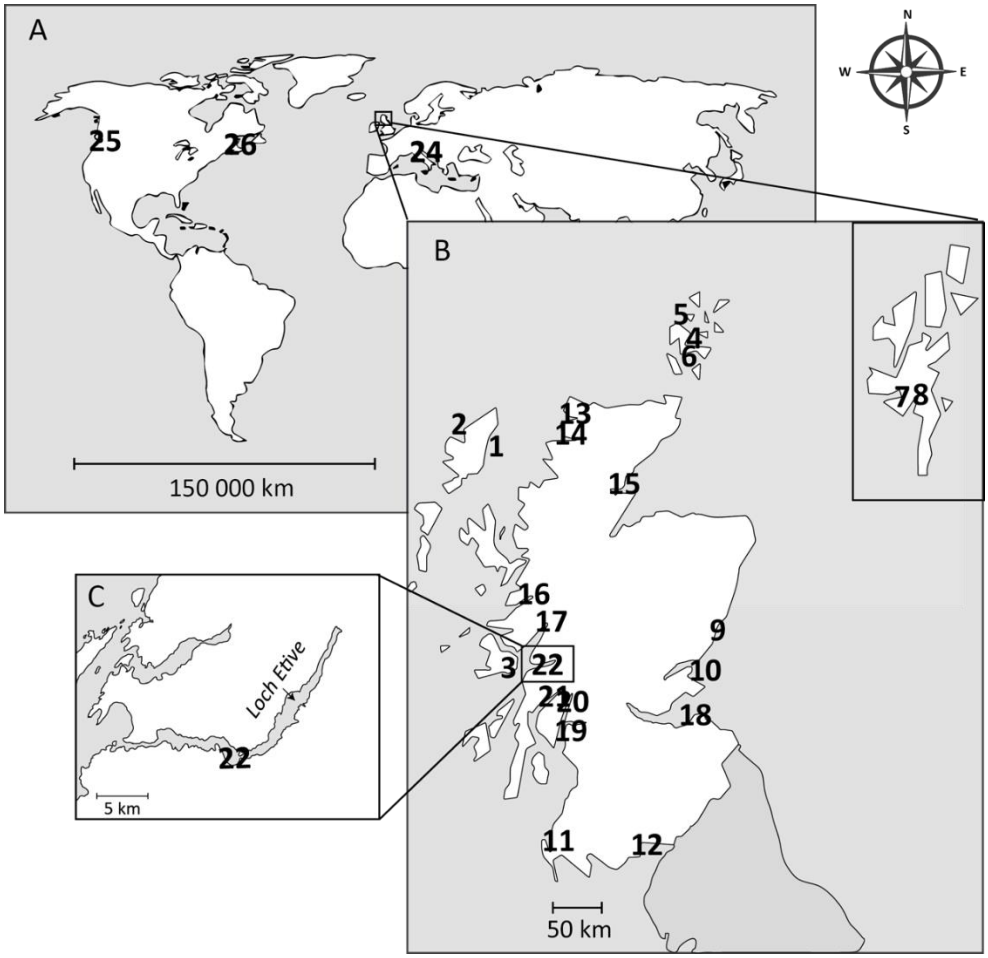

FIGURE 2.1 – Map of all named sampling sites for this study: A. sites outside of Scotland; B. sites within Scotland; C. a close-up of the site at Loch Etive. The names of each numbered site are as follows: 1. Loch Eireasort; 2. Loch Roag; 3. Loch Spelve; 4. Scapa Beach; 5. Northside; 6. Flotta; 7. Shetland BX; 8. Shetland BR; 9. Montrose; 10. St Andrews; 11. Loch Ryan; 12. Rascarrel Bay; 13. Loch Laxford; 14. Kylesku; 15. Dornoch Firth; 16. Loch Ailort; 17. Loch Linnhe; 18. Ferryness; 19. Lunderston Bay; 20. Loch Long; 21. Loch Fyne; 22. Loch Etive; 24. Bay of Piran; 25. Penn Cove; 26. Bras d'Or Lake

mussels (approximately 15 months old) were collected from one site in Scotland (Loch Etive). In November 2013, preserved gill/mantle and adductor muscle tissue, initially collected in July 2005 from one site in North America (Penn Cove), was obtained from a collection at the Senckenberg Natural History Museum, Germany (donated by Heiko Stuckas). TABLE 2.1 details the numbers of mussels collected at each site and the collection dates. The samples from Scotland comprised 10 named shoreline sites (one bottom grown aquaculture site and nine wild), 12 named rope-

TABLE 2.1 - Details of all sites sampled in this study: R = rope; S = shoreline; *n* = number of individuals collected. Only coordinates of wild shoreline locations and the aquaculture site at Loch Etive are provided. The unnamed aquaculture site is referred to as Site X. The "uses" of each site are abbreviated as follows: marker development (MD); marker validation (MV); coastline survey (CS); time series (TS); fragility vs genotype test (FG). Site numbers (Nº) correspond with the map in FIGURE 2.1 (excluding Site X). Samples from Loch Etive were taken at three different time points. Those from Jan 2012, marked with *, were used for marker validation, coastline survey and time series work; those from Jul 2012 and Aug 2013, marked with ^, were used for time series work only

Nº	Site location	GPS coordinates	Source	*n*	Date sampled	Date received	Use
1	Loch Eireasort	-	R	49	June 2014	June 2014	CS
2	Loch Roag	-	R	50	June 2014	June 2014	CS
3	Loch Spelve	-	R	50	Aug 2014	Aug 2014	CS
4	Scapa Beach	58°56'47.00"N 2°59'13.27"W	S	10	Oct 2013	Oct 2013	CS
5	Northside	59°09'25.37"N 3°12'50.53"W	S	10	Nov 2013	Nov 2013	CS
6	Flotta	-	R	45	Dec 2012	Dec 2012	CS
7	Shetland BX	-	R	45	Nov 2012	Nov 2012	CS
8	Shetland BR	-	R	45	Nov 2012	Nov 2012	CS
9	Montrose	56°42'16.31"N 2°28'13.71"W	S	49	Feb 2014	Feb 2014	CS
10	St Andrews	56°20'07.67"N 2°48'23.28"W	S	50	Feb 2014	Feb 2014	CS
11	Loch Ryan	54°56'06.83"N 5°03'38.69"W	S	50	Feb 2013	Feb 2013	MD CS
12	Rascarrel Bay	54°48'53.11"N 3°51'22.74"W	S	50	Feb 2013	Feb 2013	MD CS
13	Loch Laxford	-	R	30	Sept 2014	Sept 2014	CS
14	Kylesku	-	R	28	Sept 2014	Sept 2014	CS
15	Dornoch Firth	-	S	40	Oct 2014	Oct 2014	CS
16	Loch Ailort	-	R	50	Sept 2014	Sept 2014	CS
17	Loch Linnhe	-	R	28	Nov 2012	Nov 2012	CS
18	Ferryness	55°58'56.78"N 2°54'40.65"W	S	50	March 2014	March 2014	CS
19	Lunderston Bay	55°55'31.49"N 4°52'51.19"W	S	45	Feb 2013	Feb 2013	CS
20	Loch Long	56°02'09.51"N 4°53'14.41"W	S	47	Feb 2013	Feb 2013	CS
21	Loch Fyne	-	R	92	Nov 2012	Nov 2012	CS
22	Loch Etive	56°27'05.53"N 5°19'13.32"W	R	80* 80^ 150^	April 2013 Nov 2013 Nov 2014	June 2014 Nov 2015 Nov 2015	MV* CS* TS*^
23	Site X	-	R	39	May 2016	June 2016	FG
24	Bay of Piran	45°30'11.10"N 13°33'44.75"E	S	50	Nov 2013	Nov 2013	MD
25	Penn Cove	-	?	8	July 2005	Nov 2013	MD
26	Bras d'Or Lake	45°59'55.37"N 60°43'30.97"W	S	50	Dec 2013	Dec 2013	MV

sourced aquaculture sites, and one unnamed rope grown aquaculture site. Adult mussels were dissected as described in SECTION 2.2. Tissue samples were stored in 99% ethanol at -20°C. Gill/mantle and mantle edge tissue samples were taken from live mussels collected from shorelines in Slovenia and Canada, and these were stored

in 99% ethanol. These samples were stored at -20°C upon their arrival at Stirling University. Gill/mantle and adductor muscle tissue samples from North America were all stored in 99% ethanol and stored at -20°C upon their arrival in Stirling. However, it is not known if these tissue samples came from live specimens, nor if the mussels were collected from the shoreline or ropes. The shells from adult mussels from Site X were retained for further analysis; all other shells were discarded. The proportions of body tissue in juvenile mussels were very small, so all tissue types were removed together and stored in 99% ethanol.

2.2. DNA EXTRACTION, QUALITY ASSESSMENT AND CLEANUP

Two different DNA extraction methods were undertaken: automated DNA extraction utilising magnetic-particle technology was carried out with all tissue samples processed at Marine Scotland Science (Aberdeen); and manual DNA extraction, based on salt precipitation of proteins, was carried out with all tissue samples processed at the Institute of Aquaculture (Stirling). High quality and high molecular weight DNA, largely free from RNA contamination and with little degradation, was desired for optimal sequencing results, and was obtained with both automated and manual technologies.

2.2.1. Automated DNA extraction

Automated DNA extraction had previously been used by Dias *et al* (2011a) to extract DNA from *Mytilus* spp. tissues. Similar technology was applied during the course of this project to extract DNA from adult mussels from Flotta, Loch Fyne, Loch Linnhe, Loch Long, Loch Ryan, Lunderston Bay, Rascarrel Bay, Shetland BR and Shetland BX, and juvenile mussels from Loch Etive. Approximately 50 mg of gill and mantle tissue from adults, and all soft tissue from juveniles, was taken using disposable sterile scalpels and fixed in 99% ethanol (Sigma). Fixed tissue was homogenised using a TissueLyser (Qiagen) machine, and DNA was subsequently extracted from a volume of homogenate expected to yield approximately 5 mg of DNA, using a QIAsymphony automated extraction system (Qiagen) and QIAsymphony DNA Mini Kit (Qiagen) according to the manufacturer's instructions.

2.2.2. Manual DNA extraction

For initial tests of manual DNA extraction protocols, three tissue samples (approximately 10 mg each) were taken from each mussel: gill/mantle tissue; adductor muscle tissue; and mantle edge tissue (FIGURE 2.2). Using fine-tipped

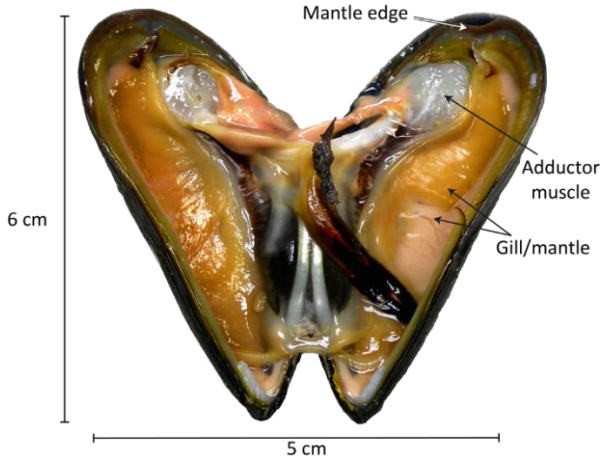

FIGURE 2.2 - Diagram of a dissected mussel and the regions where tissue samples were taken for manual DNA extractions: gill/mantle; adductor muscle and mantle edge

forceps, the tissue samples were transferred to separate 1.5 mL microcentrifuge tubes containing 1 mL of 99% ethanol (Fisher Scientific), before being stored at -20°C. Two manual DNA extraction protocols were trialled, both based on salt precipitation of proteins, followed by isopropanol precipitation of DNA. The "Spermine, Spermidine, Tris, NaCl, EDTA" (SSTNE) protocol used reagents prepared in house, while the second method involved using a commercially available genomic DNA extraction kit (RealPure; Durviz SL, ref RBMEG03). Both manual protocols were rapid (i.e., executable in 48 hours) and yielded DNA of high quality and high molecular weight from the three tissue types that were tested. Gill/mantle tissue consistently yielded the best quality DNA (i.e., high molecular weight and quality) for experimental work. The SSTNE method was significantly cheaper compared to the RealPure kit (<£1 per 100 samples compared to approximately £49 per 100 samples), and thus was routinely used for experimental work. The SSTNE extraction protocol was applied to gill/mantle tissue samples from Bay of Piran, Bras d'Or Lake, Dornoch Firth, Ferryness, Kylesku, Loch Ailort, Loch Eireasort, Loch Laxford, Loch Roag, Loch Spelve, Montrose, Northside, Penn Cove, Scapa Beach, St Andrews and Site X.

2.2.2.1. SSTNE/SDS DNA extraction

See APPENDIX 1A for SSTNE extraction buffer recipe, with details of chemical suppliers included. This method of DNA extraction was based on the salt extraction protocol described by Aljanabi and Martinez (1997), with an additional RNase digestion step and resuspension in 5 mM Tris solution (steps 5 and 16 respectively, marked with a * symbol):

1. Add a tissue sample to 200 µL of SSTNE + 1 % SDS mixture
2. Add 5 µL 10 mg/mL Proteinase K
3. Incubate for at least 3 hours at 55°C. Check progress every hour and mix until all tissue is dissolved
4. Incubate for 15 minutes at 70°C to inactivate Proteinase K
5. *Cool to 37°C and add 5 µL 2 mg/mL RNase A
6. Mix and incubate for 60 minutes at 37°C
7. Add 168 µL (0.8 X vol) 5 M NaCl, mix well – leave on ice for 10 minutes to precipitate protein
8. Centrifuge at high speed (> 12,000 *g*) for 10 minutes to spin down precipitated proteins
9. Remove and retain at least 100 µL supernatant, and add an equal volume of room temperature isopropanol
10. Mix by 5-6 sharp (rapid and abrupt) inversions
11. Leave on ice for 5 minutes then centrifuge at high speed (> 12,000 *g*) for 10 minutes to produce a pellet
12. Remove supernatant
13. Add 1 mL 70% ethanol; place tubes in rotator and spin at 5 rpm overnight to wash at room temperature
14. Centrifuge at high speed (> 12,000 *g*) for 5 min
15. Remove all excess ethanol by pipette, and heat to 60°C for at least 5 minutes to completely evaporate any remaining ethanol
16. *Suspend DNA pellet in 40 µl 5 mM Tris pH 8.0; leave on heat block at 60°C until visible pellet has dissolved. Allow to cool at room temperature before refrigeration at 4°C

17. Measure DNA concentration after a minimum of 24 h to allow pellet to fully dissolve

2.2.2.2. RealPure DNA extraction

DNA was extracted using a RealPure Genomic DNA Extraction Kit, according to the manufacturer's instructions. DNA was re-suspended in 40 µL 5 mM Tris pH 8.0 and kept overnight at 4°C to allow the DNA pellet to fully dissolve.

2.2.3. DNA quantification and quality assessment

2.2.3.1. Spectrophotometry

After successful DNA extraction from tissue, all DNA samples were quantified and quality assessed using Nanodrop spectrophotometry. DNA samples with an A260/A280 ratio between 1.7-2.0 and an A260/A230 ratio greater than 1.5 were considered to be of sufficient purity for RAD library construction and sequencing. It was preferred that samples used in preliminary PCR and KASP assays were also of high quality, but in cases where the A260/A280 and A260/A230 ratios fell outside of the desired range, the DNA samples could still be used for experimental work. Quality assessment of DNA through spectrophotometry alone was sufficient for samples used in preliminary PCR and KASP assays because partially degraded DNA had minimal effect on reaction efficiency.

2.2.3.2. Gel electrophoresis

To more reliably identify any residual RNA contamination and to examine DNA integrity, the quality of all DNA samples chosen for RAD library construction was further assessed by agarose gel electrophoresis (0.8% gel made with 0.5X TAE buffer and stained with 100 ng/mL ethidium bromide (EtBr)). Samples were run at 60 volts for 40 minutes before visualisation using the InGenius gel imaging system (Syngene) and accompanying GeneSnap gel acquisition software**.** FIGURE 2.3 shows an example of high quality and high molecular weight *Mytilus* DNA samples that were used in RAD library construction.

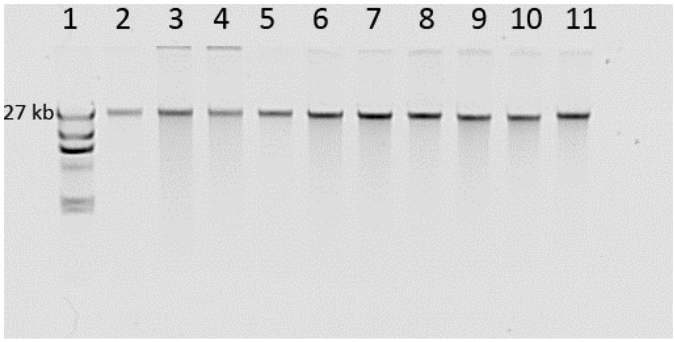

FIGURE 2.3 - 0.8% 0.5 X TAE agarose gel showing *Mytilus* DNA samples of high quality and high molecular weight (wells 2 – 11) from the Bay of Piran. This DNA was extracted from gill/mantle tissue using the SSTNE method, and was used in RAD library construction. λ*Hind*III digested DNA marker is in well 1 (highest molecular weight band = 27 kb; New England Biolabs). All wells were loaded with 5 µL UPW plus 1.2 µL 6X loading dye (Thermo Scientific); well 1 included 1 µL λ*Hind*III; wells 2 – 11 included 1 µL DNA.

2.2.3.3. Additional cleanup

Some DNA samples that had been extracted with automated technology (Qiagen) needed additional cleanup to remove residual contamination. In these cases, DNA was re-precipitated into a smaller volume of 5 mM Tris, using a standardised protocol with 0.1 volumes of sodium acetate (pH 5.8) and 2.2 volumes of 100% ethanol. This procedure preferentially precipitated the DNA fraction of a sample.

2.3 INITIAL TAXONOMIC GENOTYPING

Preliminary PCR was necessary to provisionally validate reference material for each of the three *Mytilus* species for RAD library construction: pure *M. edulis* (Loch Ryan and Rascarrel Bay), *M. galloprovincialis* (Bay of Piran) or *M. trossulus* (Penn Cove). All presumed pure individuals were genotyped with two primer sets: Me15/16 [(Me15: CCAGTATACAAACCTGTGAAGA; Me16: GTTGTCTTAATAGGTTTGTAAGA (Inoue *et al.,* 1995)] and Glu5' [(JH-5: GTAGGAACAAAGCATGAACC; JH-54: GGGGGGATAAGTTTTCTTAG (Rawson *et al.,* 1996)]. Each primer set amplified a potentially diagnostic locus in the *Mytilus* genome.

2.3.1. Me15/16 PCR

Each 6 µL PCR reaction comprised 3 µL 2X MyTaq^{TM} mix (Bioline); 0.4 µL 10 µM forward and reverse primer; 0.5 µL template DNA (5-50 ng/µL)]; and 1.7 µL

ultrapure water (UPW). PCR conditions on a Biometra TGradient Thermocycler were 95°C for 1 min, [95°C for 15 s, 56°C for 15 s, 72°C for 30 s] x 35 cycles, and 72°C for 2 min. PCR products (1 µL) were run at 60V for 40 mins on a 2% agarose gel (0.5X TAE, stained with 100 ng/µL EtBr). Visualisation on a UV transilluminator showed PCR products of the following sizes: 180 bp (*M. edulis);* 168 bp (*M. trossulus);* 126 bp (*M. galloprovincialis);* or a combination of diagnostic bands in a hybrid individual (FIGURE 2.4). PCR with Me15/16 performed consistently for species ID and provided sufficient information for preliminary species genotyping.

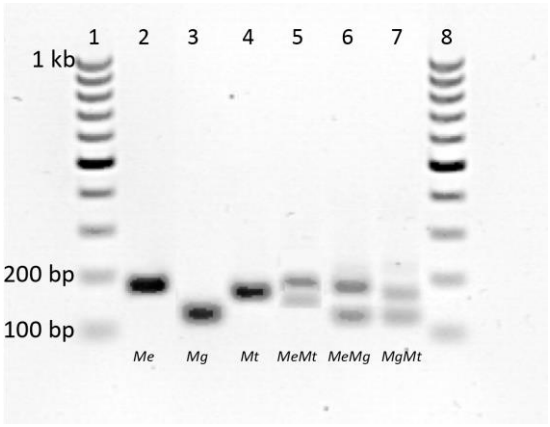

FIGURE 2.4 - A 2% 0.5X TAE agarose gel image showing results of PCR with the Me15/16 primer set: samples are in wells 2 – 7 [*M. edulis (Me;* well 2)*; M. galloprovincialis* (*Mg;* well 3)*; M. trossulus* (*Mt;* well 4)*; M. edulis/M. trossulus* hybrid (*MeMt;* well 5); *M. edulis/M. galloprovincialis* hybrid (*MeMg;* well 6); *M. galloprovincialis/M. trossulus* hybrid (*MgMt;* well 7)]; and a 100 bp marker (Fermentas) is in wells 1 and 8 (highest molecular weight band is 1kb, each division is 100 bp in size). All species diagnostic bands are in the range 100 – 200 bp. All wells were loaded with 5 µL UPW plus 1.2 µL 6X loading dye (Thermo Scientific); wells 1 and 8 included 1 µL 100 bp marker; wells 2 – 7 included 1 µL PCR product.

2.3.2. Glu5' PCR

Each 6 µL PCR reaction comprised 3 µL 2X MyTaq^{TM} mix (Bioline); 0.4 µL 10 µM forward and reverse primer; 0.5 µL template DNA (5-50 ng/µL)]; and 1.7 µL UPW. Expected band sizes for Glu5' were 350/380bp (*M. edulis*), 300/500bp (*M. galloprovincialis*), and 240bp (*M. trossulus*). PCR conditions on a Biometra TGradient Thermocycler were 95°C for 1 min, [95°C for 15 s, 56°C for 15 s, 72°C for 30 s] x 35 cycles, and 72°C for 2 min. Initial PCR reactions with Glu5' produced smearing and nonspecific bands, and temperature gradient optimisation (T_a = 53°C;

56°C; 58°C; 60°C) had no effect on improving the reaction. No additional optimisation steps with Glu5' were attempted.

2.4. RAD LIBRARY CONSTRUCTION AND SEQUENCING

RAD library construction took place over a period of five days in November 2013 at the Institute of Aquaculture, University of Stirling, and at the The Roslin Institute, University of Edinburgh. Using the results from preliminary genotyping at a single locus with Me15/16, a total of 40 (presumed pure species) individuals (21 *M. edulis,* 15 *M. galloprovincialis* and four *M. trossulus*) were identified for construction of four DNA pools (RAD libraries). Each library comprised 10 individuals: Library 1 comprised 6 *M. edulis,* 3 *M. galloprovincialis* and 1 *M. trossulus;* and Libraries 2, 3 and 4 each comprised 5 *M. edulis*, 4 *M. galloprovincialis* and 1 *M. trossulus* (TABLE 2.2). The small sample size of *M. trossulus* was a result of no other pure *M. trossulus* individuals being available at the time of library construction, and because high quality DNA could not be extracted from all Penn Cove samples available. Additional *M. trossulus,* from Bras d'Or Lake and Loch Etive, were not received until after library construction had taken place (December 2013 and June 2014 respectively). An additional 70 of these individuals (50 from Bras d'Or Lake and 20 from Loch Etive) were used for verification of the *M. trossulus* diagnostic markers once assay development had been completed (detailed in SECTION 3.3.6).

2.4.1 Complete protocol for building RAD libraries

This protocol is based on the methodology originally described in Baird *et al* (2008) and comprehensively detailed in Etter *et al* (2011). Steps marked with a * symbol and a number refer to the corresponding sections in the protocol from Etter *et al* (2011). The RAD specific P1 & P2 adapters, designed for Illumina based sequencing technology were made as outlined in Baxter *et al* (2011). In-line combinatorial barcodes on P1 & P2 adapters (TABLE 2.2) were used to identify individuals. Between days, samples were stored on ice overnight at 4°C.

TABLE 2.2 - Adapter key for RAD library construction (corresponds with steps 3 and 8 of this protocol). Each of the four DNA pools (libraries) comprised 10 samples. P1 and P2 indices are sequences of P1 and P2 adapters

Library	Tube Number	Sample	Genotype	P1 Index	P2 Index
L1	1	LR_01	*M. edulis*	TCAGA	TAGCA
	2	LR_02	*M. edulis*	GATCG	
	3	RB_01	*M. edulis*	CATGA	
	4	RB_02	*M. edulis*	ATCGA	
	5	RB_03	*M. edulis*	TCGAG	
	6	BP_01	*M. galloprovincialis*	TGCAACA	
	7	BP_02	*M. galloprovincialis*	CGTATCA	
	8	BP_03	*M. galloprovincialis*	TCTCTCA	
	9	PC_01	*M. edulis*	GTACACA	
	10	PC_02	*M. trossulus*	CTCTTCA	
L2	11	BP_04	*M. galloprovincialis*	GCATT	AGCTGA
	12	BP_05	*M. galloprovincialis*	ACGTA	
	13	BP_06	*M. galloprovincialis*	AGAGT	
	14	BP_07	*M. galloprovincialis*	ATGCT	
	15	PC_03	*M. trossulus*	GACTA	
	16	LR_03	*M. edulis*	TGCAACA	
	17	LR_04	*M. edulis*	CGTATCA	
	18	LR_05	*M. edulis*	CTCTTCA	
	19	RB_04	*M. edulis*	ACACGCA	
	20	RB_05	*M. edulis*	GCTAACA	
L3	21	LR_06	*M. edulis*	TCAGA	AGTCA
	22	LR_07	*M. edulis*	GATCG	
	23	LR_08	*M. edulis*	CATGA	
	24	RB_06	*M. edulis*	GCATT	
	25	RB_07	*M. edulis*	ACGTA	
	26	BP_08	*M. galloprovincialis*	TGCAACA	
	27	BP_09	*M. galloprovincialis*	CGTATCA	
	28	BP_10	*M. galloprovincialis*	TCTCTCA	
	29	BP_11	*M. galloprovincialis*	ACACGCA	
	30	PC_04	*M. trossulus*	GCTAACA	
L4	31	BP_12	*M. galloprovincialis*	ATCGA	TACGTC
	32	BP_13	*M. galloprovincialis*	TCGAG	
	33	BP_14	*M. galloprovincialis*	AGAGT	
	34	BP_15	*M. galloprovincialis*	ATGCT	
	35	PC_05	*M. trossulus*	GACTA	
	36	LR_09	*M. edulis*	GTACACA	
	37	LR_10	*M. edulis*	CTCTTCA	
	38	RB_08	*M. edulis*	TGCAACA	
	39	RB_09	*M. edulis*	ACACGCA	
	40	RB_10	*M. edulis*	GCTAACA	

Day 1

1. DNA extraction and preparation

DNA extraction, quantification and quality assessment were carried out as per the steps outlined in SECTION 2.2. A total of 40 samples were selected: each was diluted to 40 ng/µl in 5 mM Tris pH 8.5.

2. Restriction Enzyme Digestion (*3.2)

Each sample was digested at 37°C for 30 minutes with *Pst*I high fidelity restriction enzyme [(5'-C^TGCA|G-3') New England Biolabs (NEB)], using 2.5U

*Pst*I in 1X Reaction Buffer 4 (NEB) in a 12.5 µL reaction volume (6.25 µL master mix plus 6.25 µL (250 ng) of 40 ng/µL DNA). Excess volumes of mastermixes (i.e., for 48 reactions) were prepared to account for potential pipetting errors; this also applies to step 3 (TABLE 2.3). After digestion, the reactions were heat inactivated at 70°C for 30 minutes.

TABLE 2.3 – Mastermix volumes for steps 2 and 3 of RAD library construction protocol. 1X reaction refers to the volume of each reagent required per reaction; 48X mastermix refers to the volume of reagents added to a mastermix for 48 reactions, to minimise pipetting error in a 40X reaction; Aliquot volume for 1 reaction refers to the volume of mastermix used in each of 40 reactions in this experiment. Adding the total volumes of each step (in **bold**) gave the total volume of *Pst*I digested sample (15 µL)

Step of protocol	Reagent	1X rxn (µL)	48X mastermix (µL)	Aliquot vol 1 rxn (µL)
2. Restriction enzyme digestion	40 ng/µL DNA	6.25	/	6.25
	10X Reaction Buffer 4 (NEB)	1.25	60	
	2.5 U *Pst*I restriction enzyme (NEB)	0.13	6	
	UPW	4.87	234	
	TOTAL	**12.5**	300	
3. P1 adapter ligation	10X Reaction Buffer 4 (NEB)	0.25	12	1
	100 mM rATP (Promega)	0.15	7.2	
	UPW	0.6	28.8	
	TOTAL	**1**	48	
	mix			
	100 nM P1 adapter	**0.6**	/	
	mix, incubate at room temperature for 15 mins			
	2 MU/mL T4 ligase (NEB)	0.125	6.0	0.9
	10X Reaction Buffer 4 (NEB)	0.075	3.6	
	UPW	0.68	32.4	
	TOTAL	**0.9**	42	
	TOTAL FINAL VOLUME	**15**		

3. P1 Adapter ligation (*3.3 and *3.4)

Individual specific P1 adapters, each with a unique 5 or 7 base barcode (TABLE 2.2), were ligated to the *Pst*I digested DNA. Firstly 0.15 µL 100 mM rATP (Promega), 0.25 µL 10X Reaction Buffer 2 (NEB) and 0.6 µL UPW was added to the digested DNA, followed by 0.6 µL 100 nM P1 adapter. This mixture was incubated at room temperature (22ºC) for 15 minutes before the addition of

0.125 µL 2 MU/mL T4 ligase (NEB), 0.075 µL 10X Reaction Buffer 2 (NEB) and 0.68 µL UPW, giving a total volume of 15 µL (TABLE 2.3). Addition of the ligation mix after the barcodes increased the chance of barcodes ligating to the digested DNA rather than themselves or each other (TABLE 2.3).

Ligation reactions were heat inactivated at 65ºC for 20 minutes before being combined in appropriate multiplex DNA pools (libraries), each comprising 10 individuals (TABLE 2.2).

Day 2

4. DNA shearing (*3.5)

For each library, 120 µL (approximately. 2 µg digested DNA) was sheared to a 100-800 bp size range by sonication (Covaris) at the Roslin Institute, University of Edinburgh. To check shearing results were within the desired size range, 2.5 µL of each library was loaded on to a 1.5% agarose gel (0.5X TAE, stained with 100 ng/µL EtBr) and run at 60 V for 30 min (FIGURE 2.5).

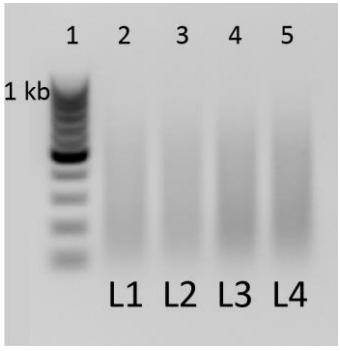

FIGURE 2.5 – 1.5% 0.5X TAE agarose gel showing shearing results of four libraries (L1 – L4). This is an example of good shearing/separation of fragments within a 100-800 bp size range, with the main bulk of fragments sized between 100-600 bp. Libraries are in wells 2 – 5. 100 bp marker (Fermentas) is in well 1 (highest molecular weight band is 1kb, each division is 100 bp in size). Well 1 was loaded with 5 µL UPW plus 1.2 µL 6X loading dye (Thermo Scientific) and 1 µL 100 bp marker; wells 2-5 were loaded with 2.5 µL UPW, 1.2 µL 6X loading dye and 2.5 µL PCR product.

The sheared DNA was column purified using a MinElute PCR Purification Kit, Qiagen (according to the manufacturer's instructions), and eluted into a final volume of 36 µL EB buffer (Qiagen). Aliquots of EB buffer were kept at 55°C to increase the efficiency of recovery from the spin column membrane (also applicable to steps 5, 6, 7 and 8 of this protocol).

5. Size selection and agarose gel extraction (*3.6)

Each of the four libraries was then size selected (to between 250-500 bp, optimal for Illumina sequencing) by gel electrophoresis (0.5X TAE; 1.1% gel). Gels were run in ice-cold buffer with a gradually increased voltage (40 V for 3 min; 60 V for 3 min; 80 V for 3min; 100 V for 1h) to minimise small fragment diffusion. Libraries were excised from the gel and temporarily stored at 4°C (FIGURE 2.6A). The remainder of the gel was quickly stained with ethidium bromide (500 ng/µL), viewed under UV, and the appropriate size range was flagged by nicking the marker lanes (100 bp ladder) (FIGURE 2.6B). The gel was then reassembled and the identified size selected band was excised using a clean scalpel blade (one per library). In this way, the size-selected DNA was not exposed to ethidium bromide or UV radiation, helping to maintain its integrity. The size selected DNA in each gel slice was column purified using a MinElute Gel Extraction Kit, Qiagen (according to the manufacturer's instructions), and eluted into a final volume of 20 µL EB buffer (Qiagen) (FIGURE 2.6C).

Day 3

6. End repair (*3.7)

Size-selected samples were processed using the Quick Blunting Kit (NEB) to repair DNA fragment ends. To each of the four size selected libraries, 2.5 µL 10X blunting buffer, 2.5 µL 1 mM dNTP mix and 1 µL blunting enzyme mix was added and incubated at 22°C for 40 mins. Each reaction was column purified using a Qiagen MinElute PCR Purification Kit (according to the manufacturer's instructions), and eluted into a final volume of 45 µL EB buffer.

7. P2 adapter ligation (*3.9)

Specific barcoded P2 adapters were added to each of the four size-selected and cleaned libraries. To each library eluate from step 7, 1 µL unique P2 adapter (10 µM), 0.5 µL 100 mM rATP and 0.5 µL 2 MU/mL T4 DNA ligase (NEB) were added and incubated at 37°C for 40 mins. See TABLE 2.2 for P2 adapter sequences. Each reaction was purified using Ampure magnetic bead (Beckman)

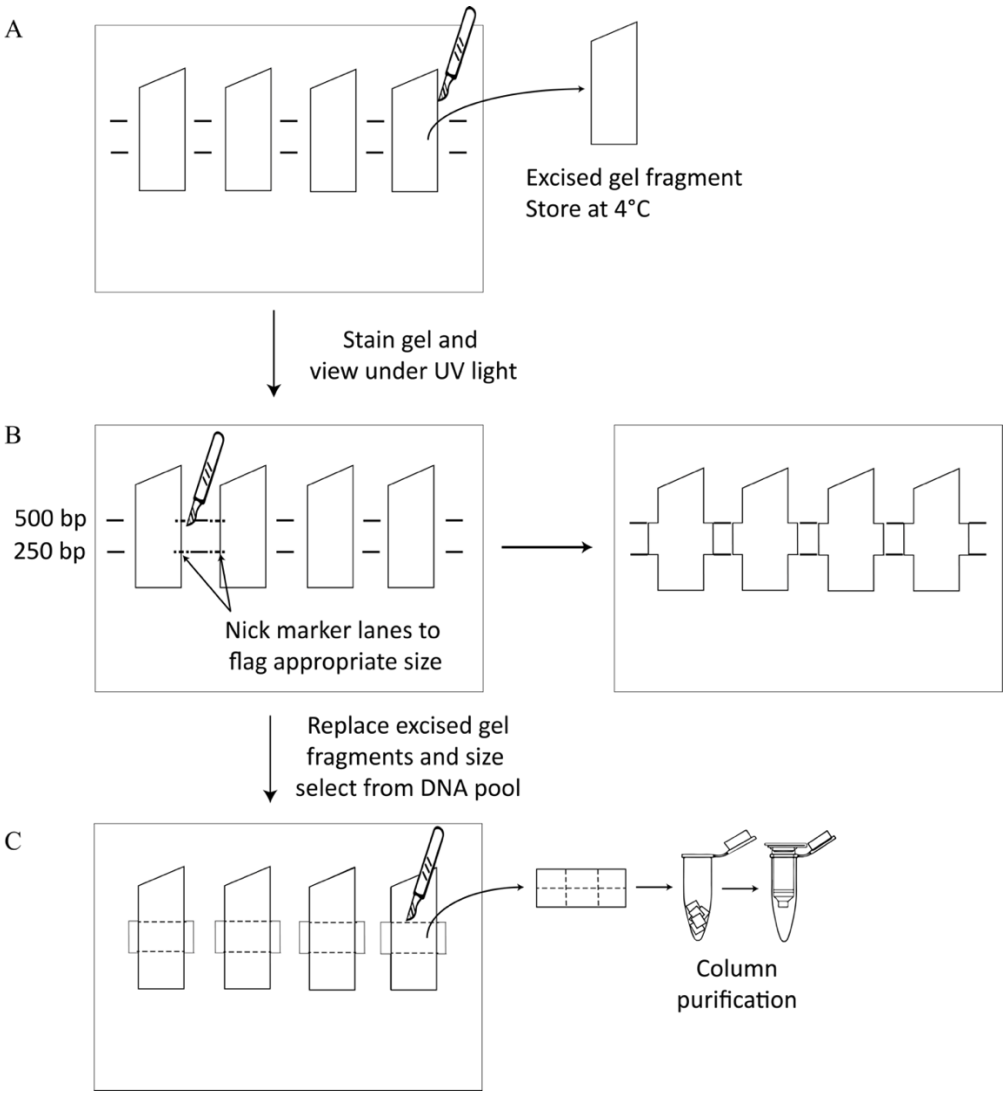

FIGURE 2.6 – Size selection of DNA fragments from libraries on 1.1% 0.5X TAE agarose gel, showing: (A) excision of libraries before gel staining; (B) flagging of desired fragment size by nicking bands of 100 bp marker; (C) replacement of libraries and excision of desired fragment size before column purification

cleanup, according to the manufacturer's instructions and using a 1:1 ratio of beads to library template. The final elution volume was 60 µL EB buffer. Four library template samples had now been produced for the final amplification and enrichment step.

Day 4

8. RAD tag amplification/enrichment (*3.10)

PCR of library templates simultaneously enriched and amplified fragments that contained both P1 and P2 adapters, suitable for paired-end sequencing (Illumina). The primers required were as follows: P1: 5'-AATGATACGGCGACCACCGA-3'; P2: 5'-CAAGCAGAAGACGGCATACGA-3'.

In order to establish the minimum number of PCR cycles required to produce sufficient product for sequencing, an initial low volume test amplification (12.5 µL) was performed with each library. Each reaction comprised 6.25 µL 2X Q5 Taq polymerase mastermix (NEB), 0.35 µL primer mix (10 µM each of P1 and P2), 0.5 µL library template, and 5.4 µL UPW. PCR conditions on a Biometra TGradient Thermocycler were: 98°C for 30s, [98 °C for 10s, 65 °C for 30 s; 72 °C for 30 s] x18, and 72°C for 5 min. To check PCR results, 5 µL of each reaction was loaded on a 1.1% agarose gel (0.5X TAE, stained with 100 ng/µL EtBr), run for 40 min at 60 V. All libraries had amplified well and to a similar extent, each showing a clear smear. The next test PCR amplification used four times the volume of library template (2 µL) and the number of cycles was reduced to 13, so that PCR conditions became as follows: 98°C for 30s, [98 °C for 10s, 65 °C for 30 s; 72 °C for 30 s] x13, and 72°C for 5 min. Libraries also amplified well under these conditions. These conditions were used for subsequent bulk PCR amplifications.

Bulk PCR preparations were made for each of the four libraries (250 µL final volume), comprising 125 µL 2X Q5 Taq polymerase mastermix (NEB), 4 µL primer mix (10 µM each of P1 and P2), 40 µL library template, and 81 µL UPW (TABLE 2.4).

TABLE 2.4 – Volumes of reagents used in bulk PCR mastermix preparation, corresponding to the number of samples in the RAD library

Reagent	Single (1X) (µL)	Bulk (10X) (µL)	
2X Q5 mastermix (NEB)	12.5	125	Mastermix
10 µM primer mix	0.4	4	
UPW	8.1	81	
Library template	4	40	
	25	**250**	**TOTAL**

Each bulk PCR preparation was then split into 16 x 15.5 µL reactions on a 96 well PCR plate for amplification. PCR conditions on a Biometra TGradient Thermocycler were 98°C for 30s, [98 °C for 10s, 65 °C for 30 s; 72 °C for 30 s] x13, and 72°C for 5 min. After PCR, the 16 aliquots from each library were recombined and purified using a Qiagen MinElute PCR Purification Kit (according to the manufacturer's instructions), and each library was eluted into a final volume of 50 µL EB buffer. A second purification was performed with Ampure magnetic bead (Beckman) cleanup, using a 1:1 ratio of beads to PCR product and following the manufacturer's instructions. Each library was eluted into a final volume of 25 µL EB buffer, and quality checked by running 1 µL on a 1.5% agarose gel (0.5X TAE, stained with 100 ng/µL EtBr) at 60 V for 40 mins (FIGURE 2.7).

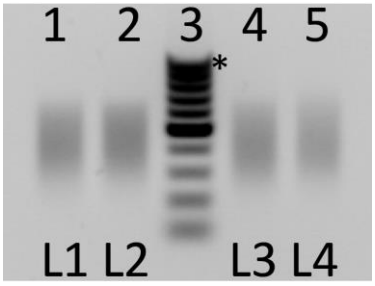

FIGURE 2.7 – 1.1% 0.5X TAE agarose gel showing amplified PCR products comprising the final libraries (L1 – L4; in lanes 1, 2, 4 and 5**) sent for outsource sequencing.** 100 bp marker (Fermentas) is in well 3 (highest molecular weight band is 1kb, marked with a * symbol, and each division is 100 bp in size). All wells were loaded with 5 µL UPW plus 1.2 µL 6X loading dye (Thermo Scientific); well 3 included 1 µL 100 bp marker; wells 1, 2, 4 and 5 included 1 µL PCR product.

9. Sequencing

Libraries were outsourced to the BMR Genomics facility at the University of Padua, Italy, for sequencing. Following accurate quantification by fluorimetry and quality checking by electrophoresis (Bioanalyser, Agilent) equimolar volumes of the four RAD libraries were combined and run on a single sequencing lane (Illumina HiSeq 2000; 100 base paired-end reads). Additional sequencing was performed in house using an Illumina MiSeq. To assess the suitability of generated data for further analysis, two quality checks were performed after sequencing was complete. Raw reads were subject to an initial quality check using in-built Illumina software [HiSeq Control Software (HCS);

MiSeq Control Software (MCS)], before additional quality checks with external software FastQC (version 0.11.3, Andrews, 2010). Low quality sequencing reads were discarded. Reads were considered low quality if they had missing restriction sites, unclear barcodes, or a Phred score below 30 (i.e., base calling with a less than 99.9% accuracy). All retained reads were assembled *de novo* into loci, then genotyped using *Stacks* software (version 1.13) (Catchen *et al.,* 2011).

2.5. SEQUENCE ANALYSIS

2.5.1. de novo *genome assembly with* Stacks

Following the completion of all initial quality checks, all retained sequencing reads (RAD tags) were assembled *de novo* into loci. Firstly, the sequencing data from four HiSeq and two MiSeq runs was demultiplexed. Demultiplexing separates RAD tags according to their unique barcode and assigns them to their sample of origin (Renaud *et al.*, 2015), thereby enabling an individual's genotype at multiple loci to be identified. Demultiplexed reads from both HiSeq and MiSeq were then merged and converted into a format suitable for assembly with *Stacks* software (version 1.13) (Catchen *et al.*, 2011). Demultiplexing and merging were performed with shell scripts: "build_samples.sh" (APPENDIX 2A) and "merge_samples.sh" (APPENDIX 2B). Both scripts were written by Michaël Bekaert (personal communication).

Stacks is a software pipeline for building loci from short read (e.g., Illumina) sequences, either *de novo* or from a reference genome, which calls genotypes using a maximum likelihood statistical model. *Stacks* was developed to work with restriction enzyme based data, such as that from RADseq and ddRAD, in order to build genetic maps, and conduct population genomic and phylogeography analysis (Catchen *et al.*, 2011). *Stacks* is implemented by component programs written in C++ and Perl. The *Stacks* web interface is implemented in PHP and both stores and retrieves data from a MySQL database (Catchen *et al.*, 2013). *Stacks* has an inbuilt "Populations" program for calculating core population genetic statistics, including F_{IS} (the inbreeding coefficient of individuals relative to the subpopulation) and F_{ST} (a measure of population differentiation) (Catchen *et al.*, 2013), which are useful when identifying species ("population") diagnostic markers.

Open source *Stacks* software and the official *Stacks* tutorial can be accessed via the *Stacks* website (available at http://catchenlab.life.illinois.edu/stacks/). Presented here is the protocol used for *de novo* genome assembly of *Mytilus* RAD tags, with instructions for running the *Stacks* pipeline; viewing the data; and exporting the results.

1. Running the *Stacks* pipeline

Before the *Stacks* pipeline could be run, a MySQL database (named "mussel_radtags"), was created on the server to hold the sequencing results.

```
% mysql –e "CREATE DATABASE mussel_radtags"
% mysql mussel_radtags
```

Within the *mussel_radtags* database there were four separate *Stacks* pipelines [i.*stacks.denovo;* ii.*stacks.edilus;* iii.*stacks.galloprovincialis;* iv.*stacks.trocssilus,* all detailed below).

i. Sequencing results for all 40 samples (*stacks.denovo*)

```
mkdir stacks.denovo
denovo_map.pl -m 5 -M 2 -n 1 -T 32 -B mussel_radtags -b 4 -a 2014-06-04 -D "mussel RAD1" -o
stacks.denovo -O population.txt -s ./samples/LR12.1.fq -s ./samples/LR13.1.fq -s ./samples/RB06.1.fq
-s ./samples/RB08.1.fq -s ./samples/RB13.1.fq -s ./samples/SLO1.1.fq -s ./samples/SLO2.1.fq -s
./samples/SLO4.1.fq -s ./samples/PC1.1.fq -s ./samples/PC4.1.fq -s ./samples/SLO5.1.fq -s
./samples/SLO8.1.fq -s ./samples/SLO9.1.fq -s ./samples/SLO12.1.fq -s ./samples/PC5.1.fq -s
./samples/LR14.1.fq -s ./samples/LR18.1.fq -s ./samples/LR20.1.fq -s ./samples/RB17.1.fq -s
./samples/RB20.1.fq -s ./samples/LR22.1.fq -s ./samples/LR23.1.fq -s ./samples/LR27.1.fq -s
./samples/RB21.1.fq -s ./samples/RB22.1.fq -s ./samples/SLO13.1.fq -s ./samples/SLO14.1.fq -s
./samples/SLO15.1.fq -s ./samples/SLO18.1.fq -s ./samples/PC6.1.fq -s ./samples/SLO20.1.fq -s
./samples/SLO21.1.fq -s ./samples/SLO24.1.fq -s ./samples/SLO26.1.fq -s ./samples/PC7.1.fq -s
./samples/LR30.1.fq -s ./samples/LR33.1.fq -s ./samples/RB23.1.fq -s ./samples/RB26.1.fq -s
./samples/RB27.1.fq
7z a -t7z -m0=lzma -mx=9 -mfb=128 -md=64m -ms=on stacks.denovo.7z stacks.denovo
rm -rf stacks.denovo
```

ii. *M. edulis* sequencing results only (*stacks.edilus)*

```
#!/bin/bash
mkdir stacks.edilus
denovo_map.pl -m 5 -M 2 -n 1 -T 32 -B mussel_radtags -b 1 -a 2014-06-04 -D "mussel RAD1
[edilus]" -o stacks.edilus -s ./samples/LR12.1.fq  -s ./samples/LR13.1.fq  -s ./samples/RB06.1.fq  -s
./samples/RB08.1.fq  -s ./samples/RB13.1.fq  -s ./samples/LR14.1.fq  -s ./samples/LR18.1.fq  -s
./samples/LR20.1.fq  -s ./samples/RB17.1.fq  -s ./samples/RB20.1.fq  -s ./samples/LR22.1.fq  -s
./samples/LR23.1.fq  -s ./samples/LR27.1.fq  -s ./samples/RB21.1.fq  -s ./samples/RB22.1.fq  -s
```

```
./samples/LR30.1.fq  -s ./samples/LR33.1.fq  -s ./samples/RB23.1.fq  -s ./samples/RB26.1.fq  -s
./samples/RB27.1.fq
7z a -t7z -m0=lzma -mx=9 -mfb=128 -md=64m -ms=on stacks.edilus.7z stacks.edilus
rm -rf stacks.edilus
```

iii. *M. galloprovincialis* sequencing results only (*stacks.galloprovincialis*)

```
mkdir stacks.galloprovincialis
denovo_map.pl -m 5 -M 2 -n 1 -T 32 -B mussel_radtags -b 2 -a 2014-06-04 -D "mussel RAD1
[galloprovincialis]" -o stacks.galloprovincialis -s ./samples/SLO1.1.fq  -s ./samples/SLO2.1.fq  -s
./samples/SLO4.1.fq  -s ./samples/SLO5.1.fq  -s ./samples/SLO8.1.fq  -s ./samples/SLO9.1.fq  -s
./samples/SLO12.1.fq  -s ./samples/SLO13.1.fq  -s ./samples/SLO14.1.fq  -s ./samples/SLO15.1.fq  -s

./samples/SLO18.1.fq  -s ./samples/SLO20.1.fq  -s ./samples/SLO21.1.fq  -s ./samples/SLO24.1.fq  -s
./samples/SLO26.1.fq
7z a -t7z -m0=lzma -mx=9 -mfb=128 -md=64m -ms=on stacks.galloprovincialis.7z
stacks.galloprovincialis
rm -rf stacks.galloprovincialis
```

iv. *M. trossulus* sequencing results only (*stacks.trocssilus*)

```
mkdir stacks.trocssilus
denovo_map.pl -m 5 -M 2 -n 1 -T 32 -B mussel_radtags -b 3 -a 2014-06-04 -D "mussel RAD1
[trocssilus]" -o stacks.trocssilus -s ./samples/PC1.1.fq  -s ./samples/PC5.1.fq  -s ./samples/PC4.1.fq -s
./samples/PC6.1.fq  -s ./samples/PC7.1.fq
7z a -t7z -m0=lzma -mx=9 -mfb=128 -md=64m -ms=on stacks.trocssilus.7z stacks.trocssilus
rm -rf stacks.trocssilus
```

The shell scripts for running these pipelines (collectively named “Build_tags.sh”, Michaël Bekaert, personal communication) is detailed below; the relevant command line parameters are detailed in TABLE 2.5.

2. Viewing the data

The *Stacks* pipeline output was captured in a log file, and could be viewed in the web interface once the pipeline had completed execution by using the URL of the University of Stirling’s web server:

```
stacks% more denovo_map.log
http://127.0.0.1/stacks/
```

This brings up a web interface wherein the entire catalogue of samples can be viewed

TABLE 2.5 – Summary of parameters used in scripts for building *Stacks* pipelines

Parameter	Function
denovo_map.pl	Runs each of the *Stacks* components: UNIQUE STACKS (*ustacks*) builds loci and calls SNPs in each; CATALOGUE STACKS (*cstacks*) merges loci from *ustacks* into a catalogue of loci in the population sample with matching SNPs; SEARCH STACKS (*sstacks*) matches each individual against the catalogue to identify its genotype/parentage
-m	Minimum stack depth. This parameter is passed to *ustacks* and controls the minimum number of exactly matching reads that must be found to create a stack in an individual. The minimum stack depth in all cases here was 5
-M	Maximum number of mismatches between stacks. This parameter is passed to *ustacks* and controls the maximum number of mismatches between stacks for them to be merged into putative loci in an individual. The maximum number of mismatches here was 2
-n	Number of mismatches between alleles. This parameter is passed to *cstacks* and enables fuzzy catalogue matching. If a locus is, for example, homozygous in one species but heterozygous in another, this parameter allows for mismatches between tags when constructing the catalogue. This is useful in a case where a dataset contains two distant populations or is composed of hybrid individuals
-t	Remove, or break up, highly repetitive RAD tags
-b	Batch ID. This specifies a batch ID for the database. *Stacks* can be run multiple times on the same dataset and the results can be stored in the same database by specifying different batch IDs. In this case, there were 4 different batch IDs corresponding to the four different *Stacks* pipelines

together (select "Catalog" view), or information about individual samples can be viewed (select "Samples" view). Catalog view has several different options to allow filtering of the data (FIGURE 2.8). In our case, data was filtered to include loci that had a maximum of three alleles and three SNPs. This allowed for the situation where, if three separate *Mytilus* species were present in the population sample, each could have a unique diagnostic allele recognisable by a unique SNP at a given locus.

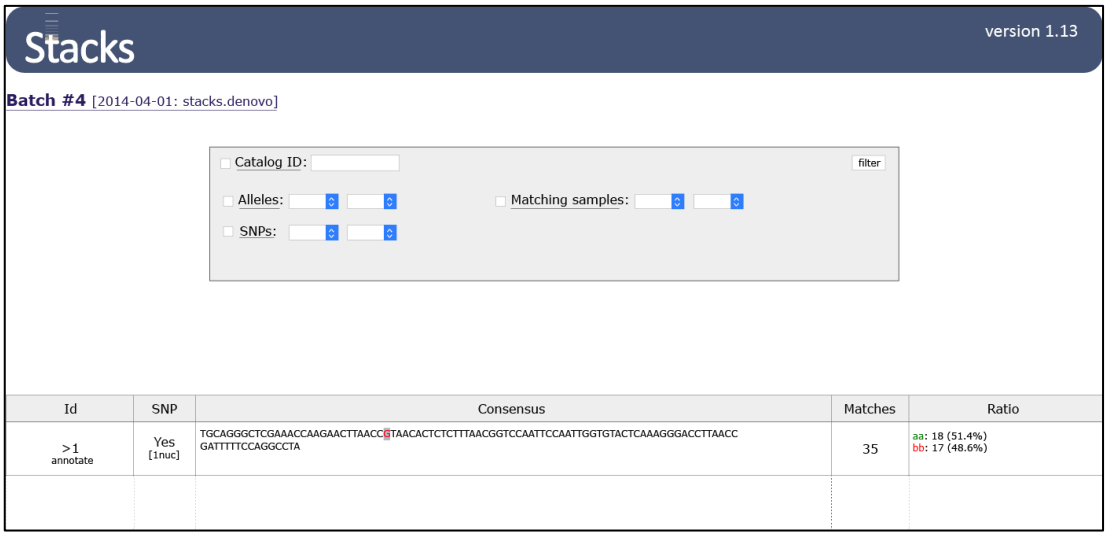

FIGURE 2.8 – Stylised representation of the "Catalog view" in the *Stacks* web interface, showing the sequence of an example locus with a single SNP (highlighted in red) identified in 35 individuals. The proportion of individuals homozygous for the diagnostic (aa) and non-diagnostic (bb) SNP are listed in the "Ratio" column. In a complete dataset, subsequent loci would be displayed in place of the dotted lines. Selecting "annotate" underneath the locus ID brings up the genotype of individual samples at this locus.

Ideally, however, loci with two alleles recognisable by a single SNP were desired for further analysis (i.e., loci with a diagnostic SNP (A) present in Species X, and a shared non-diagnostic SNP (B) present in Species Y and Species Z) (FIGURE 2.9).

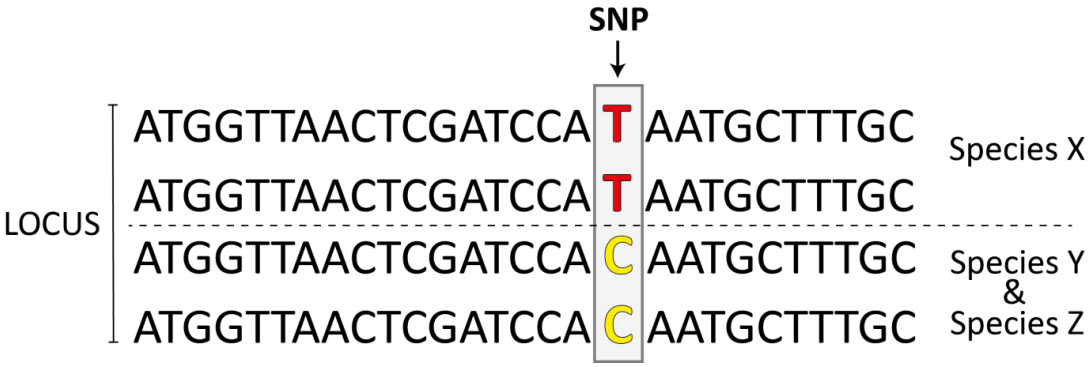

FIGURE 2.9 – Example of a diagnostic biallelic *Mytilus* species locus with a single T/C SNP. The "T" allele is diagnostic to Species X, while the "C" allele is shared by Species Y and Z.

3. Exporting results

Data was exported in a compact form from *Stacks* in a .tsv file, using the script *export_sql.pl.*

% export_sql.pl –D mussel_radtags –b 1 –f haplo.denovo.all.tsv – o tsv –F snps=1

For each locus, this reported the consensus sequences and the number of individuals that *Stacks* could find; the number of SNPs found at the locus; a listing of the individual SNPs and the observed alleles; followed by the particular allele observed in each individual. The parameters used in this script are detailed in TABLE 2.6.

TABLE 2.6 – Summary of parameters used in the *Stacks* script *export_sql.pl*

Parameter	Function
D	Specifies which database to export data from
b	Specifies batch ID of the dataset to export
f	Specifies name of output file that data will be exported to
o	Specifies type of data to export: ".tsv" or ".xls"
F	Specifies one or more filters to be used in the format name=value

Loci that were ideal candidates for subsequent marker design (as detailed in FIGURE 2.9) were filtered from the exported dataset using a custom Perl script *find.pattern.pl* (Michaël Bekaert, personal communication) (APPENDIX 3).

2.5.2. Multivariate data analysis

Principal Component Analysis (PCA) (Pearson, 1901; Hotelling, 1933) is a multivariate data analysis technique widely used throughout scientific disciplines (Abdi and Williams, 2010). PCA transforms a data matrix of related variables into a smaller number of unrelated variables that retain enough variation between them to represent the total variation in the dataset (Jackson, 1991). In the context of population genetics, PCA aims to summarise the overall variability among individuals, including both the genetic divergence between groups (structured genetic variability), and the variation occurring within groups (random genetic variability) (Jombart *et al.,* 2010). Discriminant Analysis (DA) (Fisher, 1936; Lachenbruch and Goldstein, 1979) has a slightly different approach to data transformation which achieves better discrimination of individuals into groups than PCA does. It defines a model that maximises variation between groups and minimises the variation within groups (Back *et al.,* 1996): i.e., in population genetic studies, DA attempts to summarise the overall genetic variation between groups

while overlooking the smaller, random genetic variation within groups. DA can, however, have limitations when applied to multilocus (SNP) genotyping datasets that have information pertaining to large numbers of individuals, particularly where the number of surveyed loci is less than the number of individuals (Jombart *et al.*, 2010). Discriminant Analysis of Principal Components (DAPC), introduced by Jombart *et al* (2010), is a slightly more flexible approach to modelling genomic datasets that isn't constrained by large numbers of loci or large numbers of individuals. DAPC relies on data transformation using PCA prior to DA, thereby ensuring that the variables submitted to DA are perfectly uncorrelated and that their number is less than that of the individuals in a given dataset. This transformation does not lose any genetic information and allows DA to be applied to any genetic dataset. Wherever group identities are unknown, DAPC uses K-means to cluster principal components and identify groups of individuals (Fraley and Raftery, 1998; Lee *et al.*, 2009). As with DA models, K-means splits genetic variation into a between groups and within group component, and attempts to find clustering arrangements that minimise variation within groups. DAPC uses the Bayesian Information Criterion (BIC) to identify which model best supports a given dataset, and therefore identifies the optimum number of clusters (Jombart *et al.*, 2010). Full details of the rationale behind DAPC can be found in Jombart *et al* (2010).

PCA and DAPC were carried out with the *adegenet* package (version 1.4-1; Jombart, 2008) for R (version 3.1.0) (R Core Team, 2014). Both PCA and DAPC helped to verify the diagnostic properties of potential markers, thereby allowing the most suitable loci for SNP assay design to be identified. DAPC was also used to group individuals after multilocus SNP genotyping into clusters of like individuals, showing the possible relationships between pure and admixed genotypes. All of the scripts used for PCA and DAPC analysis can be found in APPENDIX 4.

2.6. SNP ASSAYS

2.6.1. KASP genotyping technology

A total of 18 biallelic loci, each containing a *Mytilus* spp. diagnostic SNP, were selected for SNP assay design. SNP assays were designed for use with Kompetitive

Allele Specific PCR (KASP) genotyping technology, a uniplex system based on the extension of allele-specific oligos and transfer of fluorescent energy (FRET) for signal generation that can accommodate up to 1,536 samples in a single run (Semagn *et al.*, 2013). All SNP assays were designed and manufactured by LGC Genomics Limited, who offer users two services for SNP assay development: KASP By Design (KBD), whereby SNP assays are designed and sent to the customer for verification and optimisation; and KASP On Demand (KOD), whereby SNP assays are designed, verified and optimised by LGC Genomics before being sent to the customer. LGC Genomics supplied all components of the SNP assay, including *Mytilus* allele-specific primers, which were as follows:

- 2X KASP Master Mix (containing FRET cassettes, Taq polymerase, free nucleotides and $MgCl_2$)
- KASP Assay Mix (containing both *Mytilus* allele-specific forward primers, labelled with either FAM or HEX fluorescent dye, and a common reverse primer)

In October 2014, 13 SNP assays were developed by KBD and trialled in house over a period of six months. Additional SNP assays were desired for genotyping but, due to time constraints, in house trial and optimisation would have been challenging; thus, in April 2015, the KOD service was instead used for the development and optimisation of a further five SNP assays.

2.6.2. Genotyping with SNP assays

SNP assays were named according to the species they were potentially specific to: assays E1 – E6 were diagnostic for *M. edulis;* assays G1 – G7 were diagnostic for *M. galloprovincialis;* and assays T1 – T5 were diagnostic for *M. trossulus.*

All KBD assays were trialled with an initial PCR reaction under standard KASP touchdown conditions (on a Biometra TGradient Thermocycler): 94°C for 15 mins; [94°C for 20 s, 61-55°C for 60 s (0.6°C drop per cycle)] x 10; and [94°C for 20 s, 55°C for 60 s] x 40. Each 5 µL reaction comprised 2.5 µL 2X KASP Master Mix; 0.07 µL KASP Assay Mix; 0.4 µL template DNA (minimum concentration of 5 ng/ µL); plus 2.1 µL ultrapure water (UPW). Wherever KBD assays produced inconclusive results (i.e., no fluorescence or failure to form tight genotyping clusters)

optimisation reactions followed. Optimisation trials of KBD assays included the addition of DMSO (5% and 10%) and betaine (1 M and 2 M), and altered PCR conditions with an increased extension step: 94°C for 15 mins; [94°C for 20 s, 61-55°C for 120 s (0.6°C drop per cycle)] x 10; and [94°C for 20 s, 55°C for 120 s] x 40. After PCR, fluorescent signals from the end-point assays were detected on a Techne Quantica Real Time PCR Thermal Cycler, using Quansoft software to visualise and score the genotypic assays. The relative fluorescence levels of two allele-specific dyes (FAM and HEX) resolved the genotypic score for the locus. Each locus had two alleles, recognisable by a single SNP with the bases being generically termed "A" and "B": "A" was always the species-specific base and "B" was shared by the other two species. Individuals could be either homozygous for allele 1 ("AA"), homozygous for allele 2 ("BB"), or heterozygous ("AB") (FIGURE 2.10).

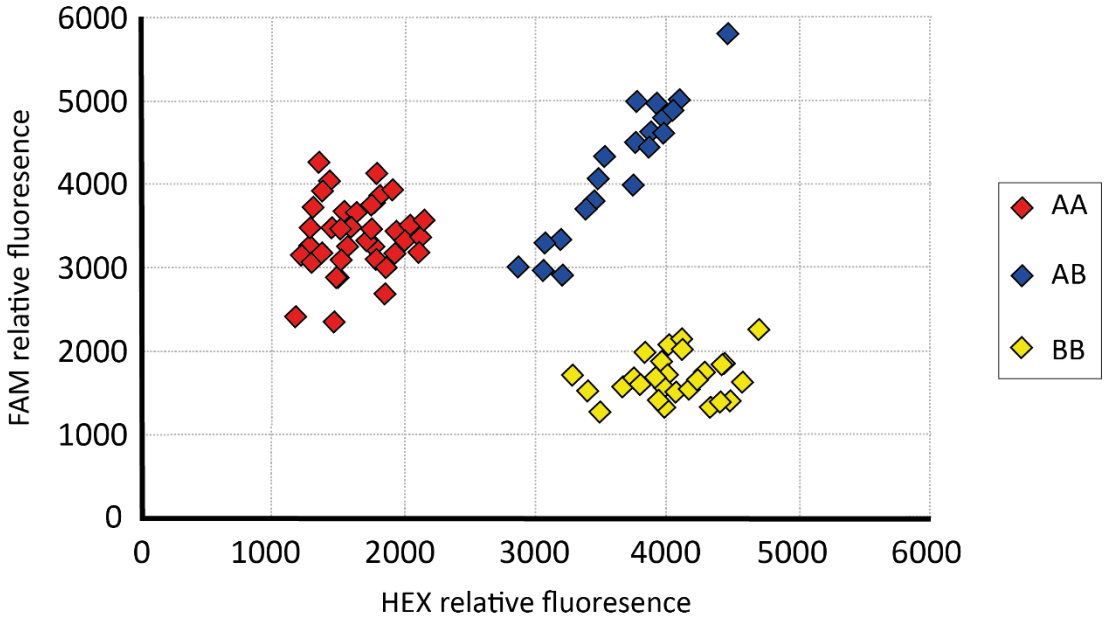

FIGURE 2.10 – Model Cartesian plot of fluorescence values generated by SNP assay, visualised using Quansoft software, with the FAM allele (recognisable by base "A" of a single SNP) on the y axis and the HEX allele (recognisable by alternate base "B" of a single SNP) on the x axis. Homozygous individuals ("AA" and "BB") are represented by red and yellow diamonds along either of the axis; heterozygous individuals ("AB") are represented by blue diamonds

Determining whether a hybrid was F1 or introgressed (FX) was based on the NEWHYBRIDS genotype model (Anderson and Thompson, 2002; Anderson, 2008). NEWHYBRIDS estimates the probability that individuals are either purebred or hybrid, and belong to one of a series of genotype classes depending on the arrangement of founders in a pedigree (FIGURE 2.11). NEWHYBRIDS itself was used for a portion of genotyping analysis; details are given in SECTION 2.6.3.3.

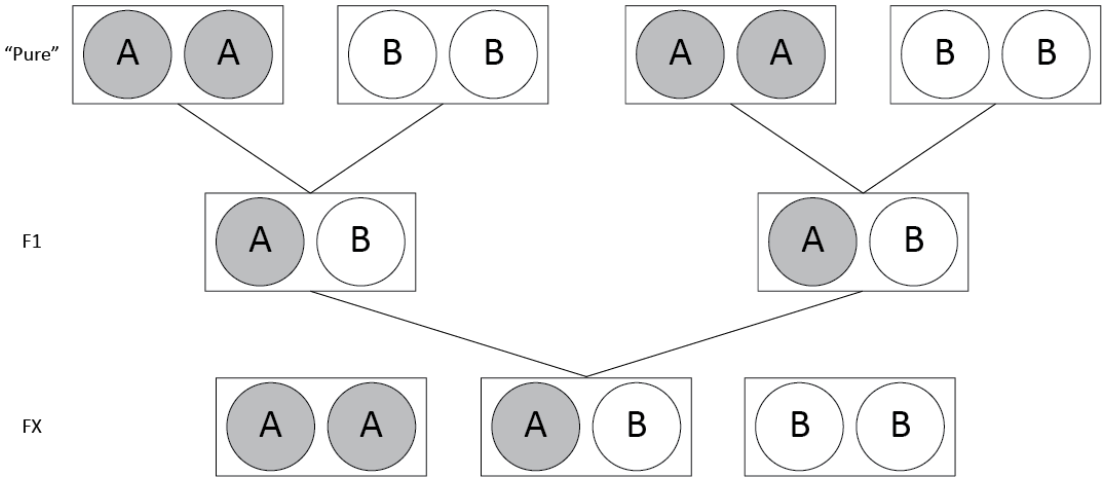

FIGURE 2.11 – Possible allelic combinations at a single biallelic locus in pure individuals, F1 hybrids and introgressed (FX) hybrids (based upon Figure 1f in Anderson and Thompson, 2002). Each box represents a locus and each circle represents an allele. Grey "A" alleles are diagnostic for species X; white "B" alleles are shared by species Y and species Z.

Individuals with 100% homozygous diagnostic alleles for species X, and 100% homozygous non-diagnostic alleles for species Y and Z, were assumed to be either pure *M. edulis*, pure *M. galloprovincialis* or pure *M. trossulus*. Individuals with diagnostic alleles for multiple species were considered as hybrids. The ratios of loci with homo- and heterozygous genotypes in a hybrid individual determined hybrid type: i.e., whether an individual was a first generation hybrid (F1) or an introgressed (second generation and beyond) hybrid (FX). F1 hybrids were 100% heterozygous at species diagnostic loci (i.e., all loci were "AB"). FX hybrids had a mixture of homo- and heterozygous genotypes ("AA", "AB" and "BB" in varying proportions) at species diagnostic loci.

2.6.3. Analysis of genotyping data

2.6.3.1. Allele frequencies and relative proportion of diagnostic alleles

Calculating genotype frequencies and allele frequencies indicate which genotypes and alleles are most or least prevalent in a population sample, providing basic information useful for management of real populations (Hartl and Clark, 2007). The equations for calculating genotype and allele frequencies are derived from the Hardy Weinberg equilibrium equation, which assumes that allele and genotype frequencies in a population will remain constant across generations in the absence of significant evolutionary influences (Andrews, 2010). Allele frequency calculations have been used widely in studies of *Mytilus* spp. to assess patterns of genetic variation according to, for instance, spatial and temporal variation (some examples include Hilbish, 1985; Daguin *et al.*, 2001; Hilbish *et al.*, 2002; Coghlan and Gosling, 2007; Jensen and Patursson, 2011) .

The equations for genotype frequency (1) and allele frequency (2) are detailed below. In each case, "A" and "B" refer to two alleles at a biallelic locus: "AA" and "BB" are homozygous individuals; "AB" is heterozygous individuals; *n* is the number of individuals with a particular genotype; *f* is the frequency; *T* is the total number of individuals in a population sample. For the allelic frequency calculation, *n*(AA), *n*(BB) and *T* values are multiplied by 2 because all individuals are assumed to be diploid, and homozygous individuals have two copies of a given allele at a locus.

$$T = n\mathrm{AA} + n\mathrm{AB} + n\mathrm{BB}$$

$$f(\mathrm{AA}) = \frac{n\mathrm{AA}}{T}$$

$$f(\mathrm{AB}) = \frac{n\mathrm{AB}}{T}$$

$$f(\mathrm{BB}) = \frac{n\mathrm{BB}}{T} \qquad (1)$$

$$f(\mathrm{A}) = \frac{[2(n\mathrm{AA}) + (n\mathrm{AB})]}{2T}$$

$$f(\mathrm{B}) = \frac{[2(n\mathrm{BB}) + (n\mathrm{AB})]}{2T} \qquad (2)$$

The frequencies of alleles per locus per population sample were calculated using the GENALEX package for Microsoft Excel (version 6.5; Peakall and Smouse, 2006, 2012). GENALEX was used to export data to GENEPOP, version 1.2 (Raymond and Rousset, 1995) for Hardy Weinberg exact tests per population sample.

A new equation (3), based on Equations 1 and 2, was designed for application to the present study, as a means of estimating the proportion of introgression within a population sample. This estimated only the proportion of diagnostic alleles in a population sample relative to the number of diagnostic markers used for genotyping. In the equation, R(A) is the relative proportion of diagnostic alleles; *n* represents the number of individuals; and *T* represents the total number of individuals in a population sample. "A" and "B" refer to two alleles at a biallelic locus: "AA" is homozygous diagnostic and "AB" is heterozygous; homozygous non-diagnostic "BB" individuals were excluded. *M* represents the number of diagnostic markers, which differed according to species: *M*=3 for *M. edulis; M*=4 for *M. galloprovincialis; M*=5 for *M. trossulus.* As in EQUATION 2, values for *n*(AA) and *M* were multiplied by 2 to represent diploid organisms with two alleles at each locus.

$$\text{R(A) per individual} = \frac{[2(n\text{AA}) + n\text{AB}]}{2M}$$

$$\text{R(A) per population} = \frac{\sum^{T} \left(\frac{[2(n\text{AA}) + n\text{AB}]}{2M} \right)}{T} \qquad (3)$$

2.6.3.2. Inferring population structure with STRUCTURE

STRUCTURE (version 2.3.4, Pritchard *et al.*, 2000) is a modelling software for population genetic data (available for free download at http://web.stanford.edu/group/pritchardlab/structure.html). STRUCTURE works by identifying genetic subpopulations and assigns individuals probabilistically to these subpopulations based on their genetic composition across multiple loci with dominant markers (e.g., SNP, microsatellite or RFLP markers) (Pritchard *et al.*, 2000). It is useful in a situation where two or more species have hybridised to produce admixed offspring, particularly if this has taken place over many generations

(Pritchard *et al.,* 2000; Falush *et al.,* 2003) and the origins of population structure are unknown (Anderson and Thompson, 2002).

STRUCTURE implements a Bayesian clustering approach [Markov chain Monte Carlo (MCMC)] to assume a model in which there are *K* subpopulations, each of which is characterised by a set of allele frequencies at the loci tested. Admixed individuals can have genetic makeup drawn from more than one of the *K* subpopulations. STRUCTURE assumes markers are at unlinked loci and are at linkage equilibrium, each providing independent information on an individual's ancestry (Pritchard *et al.,* 2000; Falush *et al.,* 2003). STRUCTURE also assumes that populations are in Hardy Weinberg equilibrium. All models attempt to find population groupings that are, as far as possible, not in disequilibrium (Pritchard *et al.,* 2000).

Specific model assumptions in STRUCTURE, and other algorithms used for inferring *K*, depend on whether individuals are assumed to have originated in a single population (no admixture) or from multiple populations (admixture). The "no admixture" model has parameters *N* (number of diploid individuals); *L* (number of loci); *K* (number of populations); *X* (observed genotypes); *Z* (population origin); and *P* (allele frequencies). The "admixture" model includes each of these parameters and also introduces a new vector, *Q*, which is the admixture proportion for each individual. In both the "no admixture" and "admixture" models, STRUCTURE implements the Markov chain Monte Carlo (MCMC) scheme to cluster individuals into *K* populations, estimating the probability of membership (no admixture) or proportion of membership (admixture) in each population for each individual (Falush *et al.,* 2003). A detailed explanation of all model parameters and their functions is available in Pritchard *et al* (2000), and expanded upon by Falush *et al* (2003).

Identifying the optimum value of population clusters from large numbers of runs per *K* value becomes highly challenging because stochastic STRUCTURE models are prone to producing different outcomes for replicate runs, even when identical parameters are chosen (Kopelman *et al.,* 2015). Models with a range of *K* values, each with replicated runs, were trialled to find the value that best represented genetic diversity within and between *Mytilus* spp. populations using CLUMPAK (Cluster Markov Packager Across *K*) software (Kopelman *et al.,* 2015; available free online

and accessible at http://clumpak.tau.ac.il/). CLUMPAK identifies the best *K* value from multiple iterations of each *K* value, according to the method outlined by Evanno *et al* (2005). This uses an ad hoc statistic Δ*K*, which is based on the rate of change in the log probability of data between successive *K* values. The optimal value of *K* was determined from STRUCTURE results generated from 100 iterations of each *K* value.

2.6.3.3. Inferring population structure with NEWHYBRIDS

NEWHYBRIDS (Anderson and Thompson, 2002) is a modelling software for genetic data that implements a Bayesian method for analysis of structured populations, similar to the approach used by STRUCTURE (Pritchard *et al.*, 2000). However, rather than looking at overall levels of admixture in population samples, NEWHYBRIDS instead focuses on individuals and assesses whether they are likely to belong to a parental or hybrid class based on their genotype (Nielsen *et al.*, 2006). NEWHYBRIDS software is available to download for free at http://ib.berkeley.edu/labs/slatkin/eriq/software/software.htm.

NEWHYBRIDS implements a model that, based on observed and unobserved data, computes the posterior probability that each individual in a dataset belongs to a specific hybrid category (genotype frequency class): e.g., in two generations of potential inbreeding, a total of six genotype frequency classes – Pure, F1 hybrid, F2 hybrid and backcross – would be assumed for the genotype model (TABLE 2.7).

TABLE 2.7 – Genotype frequency classes assumed for a NEWHYBRIDS model with two generations of potential inbreeding. Table is based on TABLE 1 in Anderson and Thompson, 2002. "spp" refers to species; F1 refers to first generation hybrid; F2 refers to second generation hybrid; "BX" refers to backcross

Genotype frequency class	Q	Frequency of AA	Frequency of AB/BA	Frequency of BB
Pure_spp1	1.00	1.00	0.00	0.00
Pure_spp2	0.00	0.00	0.00	1.00
F1	0.5	0.00	1.00	0.00
F2	0.5	0.25	0.50	0.25
BX_spp1	0.75	0.50	0.50	0.00
BX_spp2	0.25	0.00	0.50	0.50

The number of genotype frequency classes can be increased or decreased depending on the levels of admixture (and subsequent generations of potential inbreeding) expected in a population sample. An extensive description of the parameters used by

the NEWHYBRIDS model is available in Anderson and Thompson (2002), and a brief overview is available in Anderson (2008).

NEWHYBRIDS simulations were trialled with a range of genotype frequency classes to identify which combination best represented the dataset: the number of genotype frequency classes that produced the best model varied depending on the number of individuals; the number of loci genotyped; and the total number of composite hybrid genotypes present in a population sample. All simulations had a burnin and MCMC length of 100,000 for a total of five chains, as per the simulation parameters used in Anderson and Thompson (2002).

2.6.3.4. Phylogenetic analysis

RAxML (Randomised Axelerated Maximum Likelihood) is a sequential and parallel program for inference of maximum likelihood phylogenies from a nucleotide substitution model with bootstrap support values (Stamatakis, 2006). The basic principles behind nucleotide substitution models and bootstrapping are provided here; full and comprehensive details of the rationale are given in Lemey *et al* (2009). Nucleotide substitution models provide a measure of sequence divergence from a common ancestor. A measure of sequence divergence is referred to as genetic distance: smaller values denote species that are less genetically distinct, and larger values denote species that are more genetically distinct from each other (Strimmer and von Haeseler, 2009). Bootstrapping is a widely used sampling technique for estimating statistical error in situations where sampling distribution is unknown or difficult to derive analytically (Efron and Gong, 1983). Bootstrapping was first applied to estimating confidence intervals for phylogenies inferred from sequencing data by Felsenstein (1985). Firstly, new alignments are obtained by randomly selecting columns from the original sequencing data. Each column in the alignment can be selected multiple times or not at all until a new set of sequences – a bootstrap replicate – the same length as the original sequence has been constructed. Secondly, a tree is constructed for each of these artificial datasets and the proportion of each clade among all the bootstrap replicates in computed, thereby providing a statistical confidence interval for supporting the monophyly of the subset (Van de Peer, 2009).

Phylogenetic trees were constructed with RAxML (Randomised Axelerated Maximum Likelihood), Version 8 (Stamatakis, 2014), using the RAxML BlackBox online web server (Stamatakis *et al.*, 2008) (accessible at http://embnet.vital-it.ch/raxml-bb/). Maximum-likelihood phylogenetic trees were inferred using the GTR+CAT nucleotide substitution model (Lartillot and Philippe, 2004) and bootstrap support values estimated from 100 replicate searches of randomly generated trees. Completed phylogenetic trees were visualised and annotated using the graphical viewing software FigTree, version 1.4.2 (Rambaut, 2007), available to download for free at http://tree.bio.ed.ac.uk/software/figtree/.

Chapter 3

Use of RAD sequencing to identify species-diagnostic SNPs within the "*Mytilus edulis* species complex"

3.1. INTRODUCTION

Mytilus edulis (blue mussel) is an important contributor to shellfish aquaculture in Scotland, with 7,270 tonnes of mussels produced for the table in 2015 (Munro and Wallace, 2016). *M. edulis* and two closely related species, *Mytilus galloprovincialis* and *Mytilus trossulus* belong to the "*M. edulis* species complex" (Fly and Hilbish, 2013). Hybridisation between these species is not uncommon and has been observed across the world [(e.g., in the Pacific Ocean (Suchanek *et al.,* 1997; Rawson *et al.,* 1999; Wonham *et al.,* 2004) and in the Irish Sea (Coghlan and Gosling, 2007; Gosling *et al.,* 2008; Doherty *et al.,* 2009)], including Scotland where all combinations of species hybrids have been identified (Beaumont *et al.,* 2008; Dias *et al.,* 2011a; Zbawicka *et al.,* 2010; Zbawicka *et al.,* 2012). *M. trossulus* is considered undesirable for aquaculture because it displays a lower meat yield, thinner shell and reduced shelf life compared to *M. edulis* (Dias *et al.,* 2008; Gubbins et al., 2012). The same may be true of hybrids between *M. edulis* and *M. trossulus,* but undesirable traits are not always observed (Beaumont *et al.,* 2008). *M. trossulus* tends to be more abundant in the sheltered conditions found on, for instance, aquaculture ropes (Dias *et al.,* 2008). In Scotland, *M. trossulus* has posed a particular problem over the last decade for mussel farming in Loch Etive (Dias *et al.,* 2011a), contributing to production declines and, ultimately, a cessation of farming at this historically important aquaculture site (Walter Speirs, personal communication, June 2nd, 2016). To try and reduce the spread of *M. trossulus,* and thereby to protect additional sites from suffering production declines, *M. trossulus* is now recognised as a commercially damaging species and its presence is reportable to relevant authorities under "The Aquaculture and Fisheries (Scotland) Act 2013". Existing studies of Scottish mussel populations have acknowledged that three *Mytilus* species and their hybrids are present (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a). Beaumont *et al* (2008) used a combination of genetic markers, morphometric

and strength analyses to compare genotype with shell characteristics of mussels from Loch Etive. Generally, they found that individuals with *M. edulis* or *M. galloprovincialis* alleles, but without *M. trossulus* alleles, had smooth, strong shells, whereas individuals with *M. trossulus* alleles tended to have weaker, ridged shells. Zbawicka *et al* (2010) expanded the work by Beaumont *et al* (2008) to study species composition and the taxonomic origin of *M. trossulus* in Loch Etive, using a combination of nuclear and mitochondrial genotyping. Mitochondrial genotyping indicated that *M. trossulus* had a Pacific origin; although this did suggest *M. trossulus* in Loch Etive is part of a long-established population that exists naturally, rather than it being a recent introduction from the Baltic, this view remains disputed by farmers in Loch Etive and, subsequently, the origin of *M. trossulus* in Loch Etive is open to debate (Gubbins *et al.*, 2012). Dias *et al* (2011a) conducted a large scale study of mussels from around the Scottish coastline, which revealed hybridisation between *M. edulis* and *M. trossulus* at six sites on the west coast, including Loch Etive. Although useful in providing preliminary information about the distribution of *Mytilus* spp. mussels in Scotland, these studies were nevertheless limited in resolving population genetic structure because they only used small numbers of genetic markers. Thus, the extent to which hybridisation between *M. edulis, M. galloprovincialis* and *M. trossulus* has affected the genetic makeup of Scottish mussels remains unknown.

Hybridisation and introgression between species are fairly common evolutionary phenomena (Mallet, 2005; Schwenk *et al.*, 2008; Twyford and Ennos, 2012). Introgression arises from repeated backcrossing with fertile hybrids, allowing stable integration of genomic material of one species into the genome of another species without a significant deleterious effect on fitness (Rieseberg and Wendel, 1993). Accurate identification of pure *Mytilus* species is important from both commercial and research viewpoints, but this can become very challenging with occurrences of hybridisation and introgression. There are some distinguishing morphological features for each *Mytilus* species that can occasionally aid species identification. However, a range of biotic and abiotic factors (e.g., tidal flow and water temperature) can influence, amongst other characteristics, shell colour, shape, texture and size (Hepper, 1957; Seed, 1968; Widdows and Johnson, 1988). This induces widespread

phenotypic plasticity amongst the *M. edulis* species complex and makes morphology highly unreliable for species identification (Koehn, 1991). Recently, the analysis of differences in nuclear and mitochondrial DNA has been used to resolve taxonomic issues. This is considered a more specific and sensitive approach that is more reliable than studies focused on morphological traits alone (Knowlton, 2000; Capote *et al.*, 2012). Over the past three decades, a range of species diagnostic markers have been developed for single locus genotyping of *Mytilus* mussels, the most routinely used of which is the nuclear DNA marker Me15/16 (Inoue *et al.*, 1995). Me15/16 is favoured due to its simple methodology, which amplifies a size-specific species-diagnostic region of the genome (Dias *et al.*, 2008). Single locus genotyping delivers more accurate identification of *Mytilus* spp. than studying morphology does, but is limited in its scope for analysing patterns of hybridisation or genome introgression (Twyford and Ennos, 2012). Multilocus genotyping, in comparison, allows for a far better understanding of introgression (e.g., Storey *et al.*, 2005; Hayden *et al.*, 2008; Linnen and Hoekstra, 2009; Davey and Blaxter, 2010; Zuo *et al.*, 2014). Multilocus genotyping of the *M. edulis* species complex would have the power to recognise first and second generation (and beyond) hybrids in field populations, subsequently improving our knowledge of species distribution and population structure (Zbawicka *et al.*, 2012) and thereby aiding in management strategies: e.g., sourcing broodstock, eradicating invasive species and improving sustainability (Vercaemer, 2006). Some evidence is available of multilocus genotyping being applied to *Mytilus* spp. mussel populations. Wenne *et al* (2016) used 54 SNP markers to genotype mussels from a Greenlandic fjord. The authors acknowledge this is the first study to identify *M. trossulus* and its hybrids with *M. edulis* in this location, and thus the importance such information will have in environmental management and monitoring of the area. Zbawicka *et al* (2012) used 21 SNP markers to assess the genotypic composition of mussels from various European locations. This included a sample from Loch Etive in Scotland, in which *M. edulis*, *M. galloprovincialis* and *M. trossulus*, plus their hybrids, were identified. Other than in Loch Etive, however, there is no evidence of multilocus genotyping studies in Scotland. Such data would be a clear benefit in investigating the genetic integrity of mussel populations in Scottish aquaculture, and in assessing any potential threat to production from *M. trossulus* or its hybrids.

High Throughput Sequencing (HTS) technologies enable researchers to discover and characterise large panels of variable and diagnostic loci in large numbers of individuals [i.e., tens to hundreds of thousands of markers in hundreds of individuals (Davey and Blaxter, 2010; Peterson *et al.,* 2012)]. HTS is faster and cheaper than older sequencing technologies (Kircher and Kelso, 2010), and can be applied to non-model species for which little or no existing genetic data is available (Miller *et al.,* 2007; Ekblom and Galindo, 2011). DNA is prepared for HTS by enzymatic or mechanical fragmentation, followed by the addition of platform-specific adapter molecules which facilitate amplification of the DNA sequence by acting as primer sites (Di Bella *et al.,* 2013). Restriction Site Associated DNA sequencing (RADseq; Baird *et al.,* 2008) and Double-Digest RAD (ddRAD; Peterson *et al.,* 2012) both involve high coverage (accurate) sequencing of DNA adjacent to specific restriction enzyme sites throughout the genome. Both approaches provide an opportunity to assess polymorphism at thousands of loci for reasonable expenditure and manpower. RADseq (utilised in the present study for genotyping *Mytilus* individuals) uses a single restriction enzyme digestion followed by random fragmentation to produce larger numbers of fragments for analysis than ddRAD, the latter using size selection of DNA digested with two restriction enzymes to enable sequencing of fewer, more targeted sites (Kai *et al.,* 2014). In both RADseq and ddRAD, fragments are sequenced at high coverage to generate a series of robust RAD tags (Davey and Blaxter, 2010). RAD tags are then assembled into loci using custom software (e.g., *Stacks;* Catchen *et al.,* 2011), either by alignment with a reference genome or by *de novo* assembly if no reference genome is available (Willing *et al.,* 2011), allowing the identification of thousands of Single Nucleotide Polymorphisms (SNPs) in a genome (Catchen *et al.,* 2011; Davey *et al.,* 2011; Etter *et al.,* 2011). SNPs are (most often) biallelic markers that are present in coding and noncoding genomic regions (Liu and Cordes, 2004; Vera *et al.,* 2010; Sharma *et al.,* 2012). SNPs are now relatively inexpensive to identify and genotype, and can be rapidly assayed with reasonably low genotyping error rates (Rafalski, 2002; Morin *et al.,* 2004; Schlötterer, 2004). Additionally, biallelic markers exhibit lower rates of mutation and allelic dropout when compared to multiallelic markers (such as microsatellites), which potentially makes defining a specific marker for species identification more

straightforward (Vignal *et al.,* 2002; Morin *et al.,* 2004; Zbawicka *et al.,* 2012). A stable diagnostic marker with a low mutation rate would be particularly desirable for genotyping studies of *Mytilus* species, which have wide, overlapping geographical ranges in a variety of habitats open to local selection (Varvio *et al.,* 1988; Smietanka *et al.,* 2004; Riginos and Cunningham, 2005; Sousa *et al.,* 2013).

The progression and development of SNP discovery has led to a variety of uniplex and multiplex genotyping platforms being developed for easy, rapid assaying of SNPs (Gut, 2001; Syvanen, 2001; Chen and Sullivan, 2003; Sobrino *et al.,* 2005). These utilise various allelic discrimination techniques, detection methods and reaction formats for visualisation and analysis of data. The number of markers generated per run and the number of individuals being assayed influences which SNP genotyping platform should be selected. Multiplexed SNP chip-based technology is currently the genotyping platform with the highest throughput, capable of generating over one million SNPs in a single run. Multiplexing platforms tend to be better suited for genotyping large numbers of SNPs (thousands to millions) in few or individual samples (Low *et al.,* 2006). Uniplex SNP genotyping platforms, on the other hand, tend to be ideal for genotyping fewer numbers of SNPs in greater numbers of samples (Semagn *et al.,* 2013). A widely used uniplex platform for genotyping SNPs is the Kompetitive Allele Specific PCR (KASP) assay, a homogeneous, fluorescence-based genotyping technology that is based on the extension of allele-specific oligos and transfer of fluorescent energy (FRET) for signal generation. KASP genotyping can be carried out in 96-, 384- and 1,536-well plate format. It is available both as a product (i.e., non-validated and validated primer sets) and as a genotyping service by LGC Genomics service labs in North America and Europe (Semagn *et al.,* 2013). KASP is a powerful genotyping tool that can be applied to a range of studies, including SNP validation as a starting point in Quantitative Trait Loci mapping, and quality control in selective breeding (Miles and Wayne, 2008; Semagn *et al.,* 2013). KASP has been used in aquaculture to investigate species of commercial importance. For example, Palaiokostas *et al.* used KASP assays for the validation of SNPs associated with the sex determining region in the Nile tilapia (2013a) and Atlantic halibut (2013b); and Gonen *et al* (2014) used KASP assays to validate SNPs associated with the sex determining region in Atlantic salmon. Amish

et al (2012) used KASP technology to validate species diagnostic SNPs in sympatric, hybridising populations of trout (*Oncorhynchus* spp.). As in *Mytilus* mussels, hybridisation between trout species from the genus *Oncorhynchus* has the potential to threaten production. KASP allowed for a rapid, cost-effective investigation into levels of hybridisation and introgression, with the ultimate goal of improving production.

Hybridisation patterns in closely related *Mytilus* species are largely unevaluated because the majority of studies to date have focused on genotyping at a single locus, primarily with the genetic marker Me15/16. Patterns of hybridisation and introgression in *Mytilus* mussels can be better evaluated by multilocus genotyping with new genetic markers. The aims of this study are as follows:

1. To conduct an analysis of reference putatively pure specimens of *M. edulis*, *M. galloprovincialis* and *M. trossulus*, using RADseq technology to identify potentially diagnostic SNP markers for *M. edulis*, *M. galloprovincialis* and *M. trossulus;*

2. To validate a panel of novel, diagnostic SNP markers for multilocus genotyping, using KASP technology;

3. To obtain a more detailed overview of population structure by recognising hybridisation and potential interspecies introgression where it has occurred, through distinguishing first generation (F1) hybrids from second generation and beyond (FX) hybrids.

3.2. METHODS

3.2.1. Sample collection

Adult mussels (at least 40 mm in length) were collected from regions where pure *Mytilus* species were expected to occur, based on historical genetic analysis or morphological evidence (Suchanek *et al.*, 1997; Tremblay, 2002; Beaumont *et al.*, 2008; Dias *et al.*, 2011a; Žižek *et al.*, 2012) (FIGURE 3.1).

FIGURE 3.1 – Map of sampling sites chosen as sources of presumed pure specimens, based on historical morphological and genetic evidence. Site names are abbreviated as follows: Penn Cove (PC); Bras d'Or Lake (BDL); Bay of Piran (BP); Loch Etive (LET); Loch Ryan (LR); Rascarrel Bay (RB).

Specimens of *M. edulis* were collected from two shoreline locations in southwest Scotland [Loch Ryan (LR) and Rascarrel Bay (RB)]; *M. galloprovincialis* from the shoreline in Slovenia [Bay of Piran (BP)]; and *M. trossulus* from [Penn Cove (PC), North America] and [Bras d'Or Lake (BDL), Canada]. Bras d'Or Lake was a shoreline site; it is unknown whether the samples from Penn Cove were collected from the shoreline or ropes. Juvenile mussels (approximately 15 months old) from a rope in Loch Etive (LET), Scotland, previously genotyped as *M. trossulus* (Marine Scotland Science, unpublished data), were also obtained. Sample sizes varied depending on the availability of material, and the subsequent use of samples (marker development or marker validation) depended on the date of receipt (TABLE 3.1). Tissue samples (gill/mantle from adults; all body tissues from juveniles) were taken and stored in 99% ethanol at -20°C.

TABLE 3.1 – Details of sampling sites and the numbers of *Mytilus* individuals (*n)* collected for diagnostic marker development and validation. Only samples that were received by November 2013, listed above the dotted line, were used for RAD library construction [marker development (MD)], totalling 40 individuals (individual species numbers in brackets). All 178 individuals, including those used in RAD library construction and those collected after November 2013, were used for marker validation (MV).

Site location	Site name	GPS coordinates	Species	*n*	Date received	Use
Loch Ryan	LR	54°56'06.83"N 5°03'38.69"W	*M. edulis*	(10) 50*	Feb 2013	(MD) *MV
Rascarrel Bay	RB	54°48'53.11"N 3°51'22.74"W	*M. edulis*	(10) 50*	Feb 2013	(MD) *MV
Bay of Piran	BP	45°30'11.10"N 13°33'44.75"E	*M. galloprovincialis*	(15) 50*	Nov 2013	(MD) *MV
Penn Cove	PC	-	*M. edulis* *M. trossulus*	(1) (4) 8*	Nov 2013	(MD) *MV
Loch Etive	LET	56°27'05.53"N 5°19'13.32"W	*M. trossulus*	20	June 2014	MV
Bras d'Or Lake	BDL	45°59'55.37"N 60°43'30.97"W	*M. trossulus*	50	Dec 2013	MV

3.2.2. DNA extraction and preliminary PCR

DNA was extracted from gill/mantle tissue of adults and all body tissues of juveniles using the automated and manual methods as described in SECTIONS 2.2.1 and 2.2.2.1. PCR was carried out at a single locus with the Me15/16 primer set (Inoue *et al.*, 1995). This preliminarily confirmed the species status of reference material for RAD library construction. Both reaction volumes and PCR conditions were as described in SECTION 2.3.1. Agarose gel electrophoresis (2% 0.5X TAE) of PCR products showed diagnostic bands of the following sizes in presumed pure (homozygous) individuals: 180 bp (*M. edulis),* 168 bp (*M. trossulus),* and 126 bp (*M. galloprovincialis).* In hybrid (heterozygous) individuals, a combination of diagnostic bands was observable (FIGURE 3.2).

3.2.3. RAD library preparation and sequencing

A total of 40 presumed pure specimens (21 *M. edulis*, 15 *M. galloprovincialis* and 4 *M. trossulus*) were chosen for species reference library construction. Limited *M. trossulus* material was available at the time of library construction, accounting for the

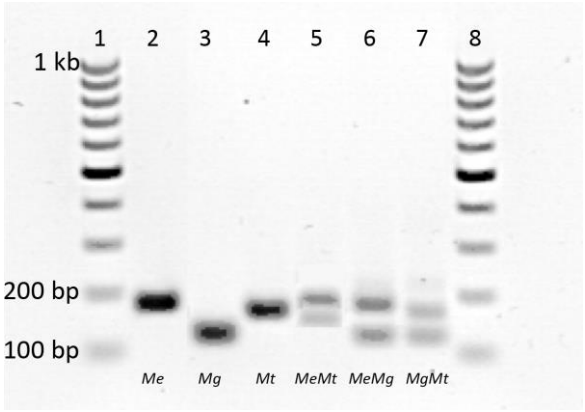

FIGURE 3.2 - 2% 0.5X TAE agarose gel image showing results of PCR with the Me15/16 primer set when pure species and their hybrids are present. Samples are in wells 2 – 7 [*M. edulis (Me;* well 2)*; M. galloprovincialis* (*Mg;* well 3)*; M. trossulus* (*Mt;* well 4)*; M. edulis/M. trossulus* hybrid (*MeMt;* well 5); *M. edulis/M. galloprovincialis* hybrid (*MeMg;* well 6); *M. galloprovincialis/M. trossulus* hybrid (*MgMt;* well 7)]; and a 100 bp marker (Fermentas) is in wells 1 and 8 (highest molecular weight band is 1kb, each division is 100 bp in size). All wells were loaded with 5 µL UPW plus 1.2 µL 6X loading dye (Thermo Scientific); wells 1 and 8 included 1 µL 100 bp marker; wells 2 – 7 included 1 µL PCR product.

small sample size used here: only individuals from Penn Cove were used in library construction, with additional *M. trossulus* samples from Bras d'Or Lake and Loch Etive being used after sequencing for marker validation. Libraries were constructed according to the protocol outlined in Etter *et al* (2011), with some modifications. The final libraries were sequenced on an Illumina HiSeq 2000 platform, outsourced to BMR Genomics, Padua, Italy, and an Illumina MiSeq platform (Institute of Aquaculture, University of Stirling, Scotland) (See SECTION 2.4 for complete RAD library construction protocol). Sequences were submitted to the European Bioinformatics Institute (EBI) and can be found in the Sequence Read Archive (SRA) under the accession number PRJEB7210.

3.2.4. Data analysis

3.2.4.1. de novo *genome assembly*

Low quality sequencing reads with a Phred score under 30, missing restriction sites or unclear barcodes were discarded; all retained reads were assembled *de novo* into loci and genotyped with *Stacks* software (version 1.13) (Catchen *et al.*, 2011). *Stacks* assigns loci based on nucleotide positions in RAD tags using a likelihood-based algorithm (Hohenlohe *et al.*, 2010) to separate actual SNPs from SNPs likely to have arisen from sequencing error. Using the default parameters for *de novo*

assembly, a minimum stack depth of 5 and a maximum of 2 mismatches were allowed per locus in an individual, with no more than 1 mismatch between alleles. The scripts used in genome assembly can be found in APPENDIX 2. A custom PERL script *"find.pattern.pl"* (Michaël Bekaert, personal communication; APPENDIX 3) was used to identify diagnostic fixed alleles among the three *Mytilus* species. This filtered all assembled loci [from RAD tags 93 or 95 bases in length (100 bp read minus 5 or 7 bp barcode)] to retain only those with a maximum of three SNPs and three alleles that were present in all three species. *Stacks* software also provided F_{ST} values, statistics that act as a measure of population differentiation (Catchen *et al.,* 2013). In this case, a "population" referred to a group of individuals belonging to a single species: *M. edulis, M. galloprovincialis* or *M. trossulus.* Generally, higher F_{ST} values (closer to one) indicate more differentiation between species, while lower values (closer to zero) indicate less differentiation. The thresholds used for F_{ST} values were taken from Hartl and Clark (1997) as follows: F_{ST} <0.05 (little genetic differentiation); F_{ST} = 0.05-0.15 (moderate genetic differentiation); F_{ST} = 0.15-0.25 (great genetic differentiation); F_{ST} >0.25 (very great genetic differentiation).

3.2.4.2. Phylogenetic analysis

Sequencing data from filtered polymorphic loci was combined into a single alignment of alleles (composite genotype) for a total of 40 individuals used in RAD library construction. Phylogenetic trees were constructed with RAxML (Randomised Axelerated Maximum Likelihood), Version 8 (Stamatakis, 2014), using the RAxML BlackBox online web server (Stamatakis *et al.,* 2008) (accessible at http://embnet.vital-it.ch/raxml-bb/). Maximum-likelihood phylogenetic trees were inferred using the GTR+CAT nucleotide substitution model (Lartillot and Philippe, 2004) and bootstrap support values estimated from 100 replicate searches of randomly generated trees (see SECTION 2.6.3.4 for further details). Completed phylogenetic trees were visualised and annotated using the graphical viewing software FigTree, version 1.4.2 (Rambaut, 2007), available to download for free at http://tree.bio.ed.ac.uk/software/figtree/.

3.2.4.3. PCA, DAPC and allele frequencies

Multivariate data analysis was carried out using R (version 3.1.0) (R Core Team, 2014) and an associated package *adegenet* (version 1.4-1; Jombart, 2008) for Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC). PCA creates simplified models of the total variation within the dataset (Jackson, 1991), and DAPC identifies clusters of genetically related individuals (Jombart *et al.,* 2010) within the most statistically likely PCA model, determined using the Bayesian Information Criterion (Schwarz, 1978). DAPC also sorts loci by the strength of their association with species identity, designated by a theoretical "loading value". All scripts used in DAPC analysis can be found in APPENDIX 4A.

The frequencies of alleles per locus per population sample were calculated using the GENALEX package for Microsoft Excel (version 6.5; Peakall and Smouse, 2006, 2012). GENALEX was used to export data to GENEPOP, version 1.2 (Raymond and Rousset, 1995) for Hardy Weinberg exact tests per population sample.

3.2.5. Selection of SNPs for assay design

Each potentially informative locus contained two alleles that were identifiable by the presence of a SNP. One allele was diagnostic for a single species, while the other allele was shared by the remaining two species. For primer design to be feasible, the SNP of interest at a given locus needed to be at least 20 base pairs from the start [excluding 5 or 7 bp P1 adapter sequence from RAD library construction] and 20 base pairs from the end of a given sequence. This allowed sufficient sequence for compatible primers to be designed. SNP assays were designed and manufactured for use with Kompetitive Allele Specific PCR (KASP) genotyping technology by LGC Genomics Limited, who also supplied all components of the SNP assay. Assay components were as follows: 2X KASP Master Mix (containing FRET cassettes, Taq polymerase, free nucleotides and $MgCl_2$); and KASP Assay Mix [containing both *Mytilus* allele-specific forward primers (labelled with either FAM or HEX fluorescent dye) and a common reverse primer]. SNP genotyping conditions are detailed in SECTION 3.3.4.

3.3. RESULTS

3.3.1. RAD library preparation and sequencing

Preliminary PCR at a single locus with the Me15/16 primer set confirmed the presence of pure species in all population samples [*M. edulis* (Loch Ryan and Rascarrel Bay); *M. galloprovincialis* (Bay of Piran); and *M. trossulus* (Penn Cove, Bras d'Or Lake and Loch Etive). Such individuals were suitable for RAD library construction and marker validation (TABLE 3.2; APPENDIX 6A).

TABLE 3.2 – Results of preliminary single locus genotyping with the Me15/16 primer set. Genotypes are as follows: *M. edulis (Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis (MeMg);* hybrid of *M. edulis* and *M. trossulus (MeMt);* hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt).* Site names are abbreviated as detailed in TABLE 3.1

	Genotype					
Site	*Me*	*Mg*	*Mt*	*MeMg*	*MeMt*	*MgMt*
LR	47	0	0	3	0	0
RB	50	0	0	0	0	0
BP	0	50	0	0	0	0
PC	1	0	6	0	0	1
BDL	0	0	40	0	10	0
LET	0	0	20	0	0	0

A total of 40 individuals, genotyped as pure (homozygous) with Me15/16 were chosen for RAD library construction: these comprised 21 *M. edulis* (10 each from Loch Ryan and Rascarrel Bay and a single individual from Penn Cove); 15 *M. galloprovincialis;* and four *M. trossulus* from Penn Cove. *M. trossulus* had such a small sample size because only limited material from Penn Cove was available at the time of library construction. The reasons for this small sample size were two-fold: firstly, all DNA samples had to be of high molecular weight and have a concentration of at least 40 ng/ µL for optimal sequencing results, and it was only possible to extract DNA of high enough quality from five Penn Cove individuals. Secondly, one mussel was originally wrongly assigned to *M. trossulus* and eventually reassigned as *M. edulis*, further reducing the *M. trossulus* sample size to four. The number of *M. edulis* used in library construction was subsequently increased to 21. High throughput sequencing of these 40 individuals produced 574,728,488 raw reads in total [490,811,956 HiSeq reads (combined P1 and P2 reads from three sequencing runs); and 83,916,532 MiSeq reads (combined P1 and P2 reads from two sequencing runs)]. After the removal of low-quality and incomplete reads, 71.9% of the total raw

reads were retained (413,377,018 reads). From these, a total of 3,254,022 RAD tags were detected, of which 38,420 were polymorphic and shared by at least 75% of individuals. Two sequencing platforms were used because of a delay in output from HiSeq technology. MiSeq technology was used to generate some preliminary data which were later combined with the HiSeq data.

3.3.2. Sequence analysis

3.3.2.1. Number of assembled loci

The number of RAD tag loci detected per individual *M. edulis* and *M. galloprovincialis* was relatively consistent, ranging from 131,000 – 313,000 loci. There were two exceptions among *M. edulis* individuals (RB_01, which had 18,220 loci, and PC_01, which had 5,459 loci); these were possibly low quality samples that had failed to digest or ligate efficiently during RAD library construction. *M. trossulus* values ranged from around 59,000 – 268,000 loci (APPENDIX 6B). On average, numbers of loci per species exceeded 150,000, but values for *M. edulis* and *M. galloprovincialis* were higher than those for *M. trossulus* (TABLE 3.3).

TABLE 3.3 – Average number of loci per species generated through *de novo* assembly of RAD tags. *n* = number of individuals

Species	Average number of loci	*n*
M. edulis	231,119	21
M. galloprovincialis	277,236	15
M. trossulus	152,959	4

3.3.2.2. Identifying loci for marker design

To identify robust genetic markers and to minimise the proportion of erroneous data, all potentially informative loci were filtered to show only those with a maximum of three alleles and/or three SNPs, and which were detected in all three species. This increased the likelihood of identifying a "true" SNP [i.e., a polymorphism occurring at a frequency of >1 % (Perkel, 2008)], rather than a false SNP generated from *de novo* assembly of non-homologous loci. A total of 362 SNPs spread across 349 RAD loci were identified (some loci had more than one SNP), and were used in subsequent analyses to identify the most suitable candidate loci for primer design.

The phylogenetic tree constructed from the composite genotypes of 349 shared alleles in 40 individuals (recognisable by 362 SNPs) showed three distinct clusters, accurately delineating the three species that were used for library construction (FIGURE 3.3). DNA sequences change over time because of selection;

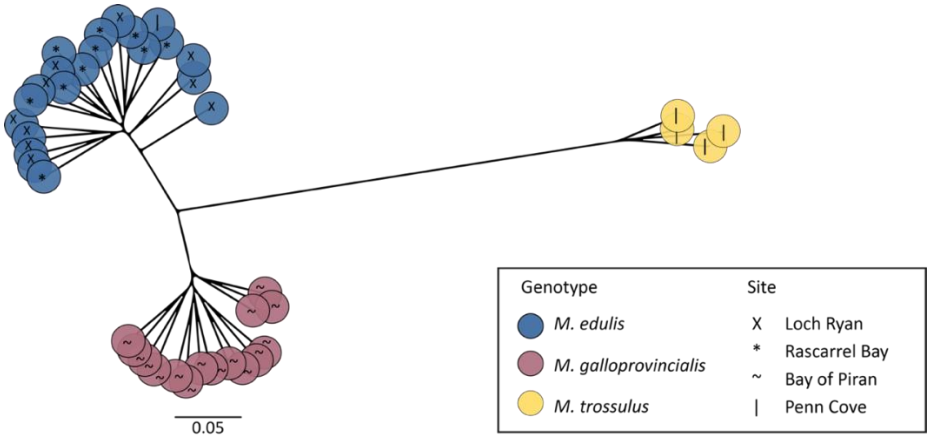

FIGURE 3.3 – Radial phylogenetic tree of 40 individuals constructed using composite genotypes of 362 SNPs at 349 biallelic RAD loci, showing separation by genotypes from preliminary PCR with Me15/16 and by site. The scale indicates the number of nucleotide substitutions per site

thus, any two sequences that derive from the same common ancestor and evolve independently will eventually diverge from each other. A measure of this divergence is called a genetic distance, which measures the genetic dissimilarity between DNA sequences (Lemey *et al.*, 2009). Smaller values denote species that are less genetically distinct, and larger values denote species that are more genetically distinct from each other. In this case, genetic distances were calculated from the changes in composite genotypes (i.e., the rate of nucleotide substitutions) between individuals. *M. edulis* and *M. galloprovincialis* were the most similar (a genetic distance of 0.1 nucleotide substitutions) with *M. trossulus* more distant (a genetic distance of 0.365 nucleotide substitutions). The *M. edulis* from Penn Cove was grouped with the *M. edulis* from Loch Ryan and Rascarrel Bay, confirming its identity as *M. edulis* rather than an *M. trossulus* individual with *M. edulis* alleles. Several PCA models were trialled to identify that which best represented variation in the RAD genotype data from 349 potentially informative loci. Accordingly, the PCA model representing 80% of the cumulative variance within the dataset, and with the smallest Bayesian Information Criterion (BIC) value, was selected. This model had

three clusters representing three groups in the RADseq dataset, comprising 21, 15 and 4 individuals. DAPC of PCA clusters (retaining 90% of cumulative variance) (FIGURE 3.4) confirmed that these three groups were distinct from each other,

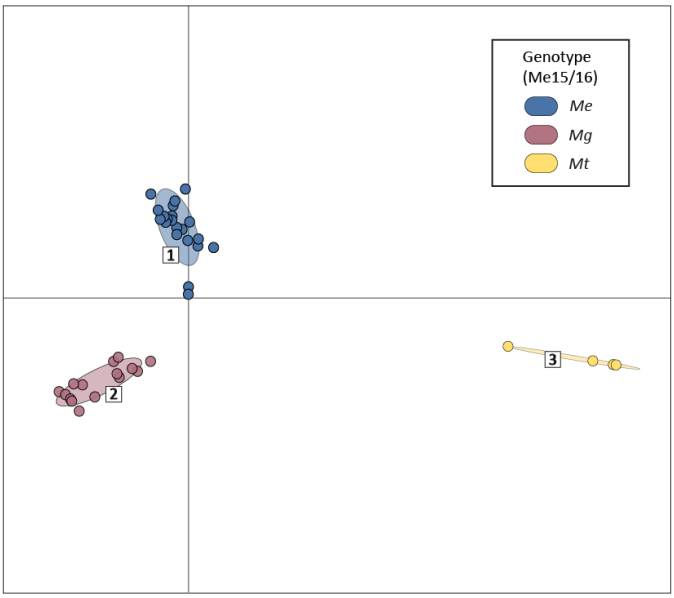

FIGURE 3.4 – DAPC scatterplot of clusters generated by PCA of composite genotypes of 362 SNPs at 349 biallelic RAD loci. PCA = 80% cumulative variance; DAPC = 90% cumulative variance**.** Individuals are represented by dots, which are colour coded according to their genotype from preliminary PCR with Me15/16 (*Me = M. edulis; Mg = M. galloprovincialis; Mt = M.trossulus*), and groups are represented by ellipses.

although it should also be acknowledged that this discreteness could have arisen in part from the small number of population samples that were examined. DAPC sorted potentially species diagnostic loci by their "loading values". "Loading values" are theoretical values based on the sequencing depth per sample at each locus (Jombart, 2014): they offer a guideline into which loci are most strongly associated with species ID and, subsequently, which loci best contribute to separating DAPC clusters (Kothera *et al.,* 2013). Loci with the highest sequencing depths had the highest "loading values" and, thus, were assumed to have the strongest relative association with species identity, improving their reliability as potential diagnostic markers. A full explanation of "loading values" can be found in Jombart *et al* (2010). All potentially diagnostic loci were identified at least 75% (n=30) of individuals used for sequencing. From these potentially diagnostic loci, loci with the highest "loading values" were preferred; however, in the case where the SNP of interest was less than 20 bp from the start or end of a given sequence, making it unsuitable for SNP assay

development, loci with lower loading values were instead selected (APPENDIX 5). Thus, a total of 18 biallelic loci with at least 20 bp on either side of the SNP of interest, corresponding to two diagnostic alleles, were chosen for SNP assay development. DAPC of 40 individuals at 18 selected loci showed similar results to FIGURE 3.4. This model had three clusters representing three groups, comprising 21 *M. edulis*, 15 *M. galloprovincialis* and 4 *M. trossulus* as per preliminary Me15/16 genotypes (FIGURE 3.5). Cumulative variance represented by PCA (100%)

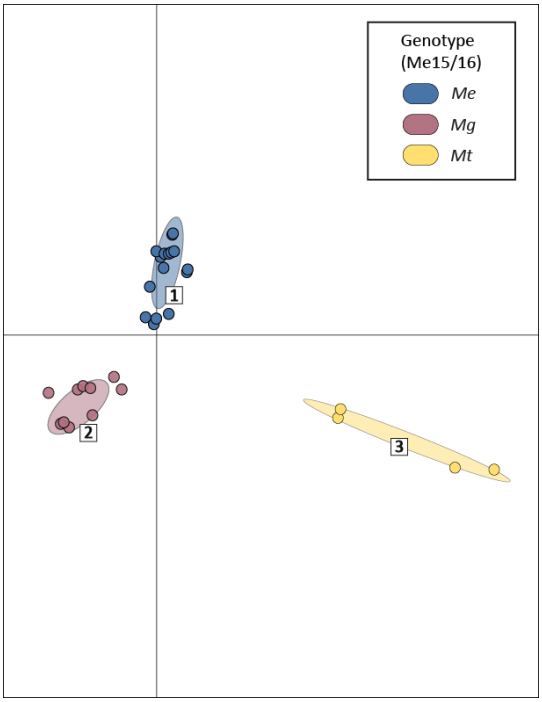

FIGURE 3.5 - DAPC scatterplot of clusters generated by PCA of composite genotypes of 38 individuals at 18 loci (RAD data) (PCA = 100% cumulative variance; DAPC = 97.5% cumulative variance). Individuals are represented by dots, which are colour coded according to their genotype from preliminary PCR with Me15/16 (*Me = M. edulis; Mg = M. galloprovincialis; Mt = M.trossulus*), and groups are represented by ellipses.

and DAPC (97.5%) were higher than previously: this could have indicated these 18 loci were most strongly associated with species ID out of the original 349 loci, but was probably a reflection of the smaller dataset being slightly easier to model. This was nevertheless a good indicator that the chosen loci were reliable for separating individuals according to their genotype, and were thus suitable for marker design. The names of selected loci, and their associated assays, were as follows: *M. edulis* *n*=6 (E1, E2, E3, E4, E5 and E6)*; M. galloprovincialis n*=7 (G1, G2, G3, G4, G5, G6 and G7)*; M. trossulus n*=5 (T1, T2, T3, T4 and T5) (TABLE 3.4).

TABLE 3.4 – Primer sequences for SNP assays, corresponding to 18 biallelic loci identified with RADseq. Additional SNPs that were present in primers but which were not diagnostic are represented by IUPAC codes (A/C = M; A/G = R; A/T = W; C/G = S; C/T = Y; G/T = K). Primer sets with underlined names were amplified under standard KASP thermal cycling conditions, and those in *italics* required optimisation by increasing the extension time of the thermal cycling conditions and addition of 5% DMSO. All sequences below the dotted line could not be optimised

Assay name	Allele X (HEX) primer sequence	Allele Y (FAM) primer sequence	Common reverse primer sequence
E1	TTAACATTTTGCGCGACCAACAAATTAT	AACATTTTGCGCGACCAACAAATTAC	TGCAGTTTACCGATTTGGAAGCGGT
E2	ACCTGATATTTACCACAAATTTATTCCATC	ACCTGATATTTACCACAAATTTATTCCATA	TAAGATGGGTAAAGTGKCTCAAGTGATATA
E3	CAGGCCAAAGTGTTTCTTCCTGATA	CAGGCCAAAGTGTTTCTTCCTGATT	GCAAGAGATTACAKATTGGTCACCATATAA
G1	GAGAATGTGTCAAATCAATATAACTGCCT	GAATGTGTCAAATCAATATAACTGCCG	AGAGCCCTAGCAGAAAGAGGAGAAA
G2	AAGGGATTTTATTTTATAAWAGATAAAGATACC	AAGGGATTTTATTTTATAAWAGATAAAGATACA	GCAGATTTAAAGTTGATAAAACTCAACCTA
G3	AATACGTTTGTAACAGTTCTCATCCGT	ACGTTTGTAACAGTTCTCATCCGC	GCAGTYGTAGGGAATCTGTTAGTCATA
G4	AAATGTTGTTTTGTGACAGCCATCTTG	AAATGTTGTTTTGTGACAGCCATCTTC	AACAGCAGCAAACCTTTCATCCTTATCAT
T1	CAAAAAGGAATCTGGTTTATTCGATTCAA	CAAAAAGGAATCTGGTTTATTCGATTCAG	CATAGCAGTCATATAGTAGGGGTAACATT
T2	AAAACAAAATTAATTAGGGATGTTGTGTGC	GAAAACAAAATTAATTAGGGATGTTGTGTGA	CTTCTAAATGTGGATGCCACACAAAGATA
T3	GTCATTTGCGTTAAATTAGCAGTATCG	GTCATTTGCGTTAAATTAGCAGTATCA	CTTCCTTTGCCGCYTCCATTGCAA
T4	AATATTGGCAGGTTGTAGAGGAGGA	AATATTGGCAGGTTGTAGAGGAGGT	AGGGCTAGCAGTGTAAGACCCAATA
T5	GTAAAGGTTGTAATAACCTTGTACAC	CTGTAAAGGTTGTAATAACCTTGTACAT	CAGCATTATACAAAGGATGCTGATGGTTT
E4	AGGAAAAGGAGGACCCACGG	CTAGGAAAAGGAGGACCCACGA	CTGGATTKACTGCTGGGGGCGA
E5	GCACTATTTCAGAGAAACCAATTTG	GCTGCACTATTTCAGAGAAACCAATTTT	AGTTGGCCTGGCAGTATGCTAACTA
E6	CAGCATACCCAAACATAAATGATGAGA	AGCATACCCAAACATAAATGATGAGG	GCATGGTTTTTCATTAGTTGCCCTCATTA
G5	CCATTGTTCGTGTGCAATCCTGA	CCATTGTTCGTGTGCAATCCTGT	TTAATAAGACATTMCTTGTTTTTCATGCTA
G6	AACCACCCCCACCACAACAGTA	AACCACCCCCACCACAACAGTT	CTGCTTCTTGTGGTGGTGGTGGTT
G7	ACTGTGAGCAATGTGGCGAC	GCTACTGTGAGCAATGTGGCGAT	GTGCAACCTTAATTTCCCATACTCCATAA

The sequences of all 18 loci selected for primer design were checked against sequences in the NCBI NR database using BlastX (Altschul *et al.,* 1990) for *M. edulis, M. galloprovincialis* and *M. trossulus* (using a "megablast" search parameter for each), but no matches to any known genes were identified (full sequences of loci are detailed in APPENDIX 5).

3.3.2.3. F_{ST} values

Stacks generated F_{ST} values for 18 potentially informative loci. These were expressed as comparisons of "Species 1" (*M. edulis*), "Species 2" (*M. galloprovincialis)* and "Species 3" (*M. trossulus*) (TABLE 3.5). For the majority of loci, F_{ST} >0.25 which indicated "very great genetic differentiation" according to the threshold values stated by Hartl and Clark (1997). Three loci (E5, G3 and G4) had negative F_{ST} values, suggesting these loci did not show differentiation between species. Comparisons to "Species 3" were not available for four loci (G1, G2, G3 and G4) because these had no *M. trossulus* reads.

TABLE 3.5 - F_{ST} values for 18 loci chosen for marker design, generated using the *Stacks* Population (Pop) module. Sp = species: Sp 1 (*M. edulis*); Sp 2 (*M. galloprovincialis);* Sp 3 (*M. trossulus*)

	Sp 1	Sp 2	Sp 3	Locus
Sp 1		1	1	E1
		0.94	0.81	E2
		1	1	E3
		0.93	0.80	E4
		-0.19	-0.16	E5
		0.87	0.58	E6
Sp 2	1		1	G1
	1		1	G2
	-0.15		/	G3
	-0.17		/	G4
	1		1	G5
	1		1	G6
	1		/	G7
Sp 3	1	1		T1
	1	1		T2
	0.86	0.53		T3
	1	1		T4
	1	1		T5

3.3.2.4. Other potentially informative loci

The majority of loci identified between individuals were biallelic. Triallelic loci (n=5) were also identified that, potentially, had a different diagnostic allele for each of the three *Mytilus* species of interest, but these were not considered further because of their unsuitability for use with biallelic KASP assays. Other RAD loci were identified that were present in only one of the three *Mytilus* species (*M. edulis* n=7; *M. galloprovincialis* n=16; *M. trossulus* n=117). These loci may have reflected species-specific differences at restriction enzyme sites. However, without the availability of a robust reference genome, there is insufficient information about the sequences flanking restriction sites to enable further exploration and assaying of these potential markers.

3.3.3. SNP assay optimisation and validation

All SNP assays were designed for use with KASP genotyping technology by LGC Genomics Limited. Assay optimisation was carried out with the 40 samples used in RAD library construction, according to the protocol described in SECTION 2.6.2. A total of 12 SNP assays from 18 were successfully optimised: E1, E2 and E3 (*M. edulis*); G1, G2, G3 and G4 (*M. galloprovincialis*); and T1, T2, T3, T4 and T5 (*M. trossulus*). SNP genotyping results were obtainable at all 12 loci for each of the 40 samples. To validate the diagnostic properties of SNP markers and, therefore, their suitability for use in genotyping additional samples, these results were compared to RADseq genotyping calls where possible (TABLE 3.6).

TABLE 3.6 – Percentages of matching RAD and KASP genotyping calls per locus

Locus	Number of RADseq calls	% matching KASP
E1	32	90.6
E2	31	96.8
E3	32	100
G1	30	100
G2	31	96.8
G3	32	100
G4	31	100
T1	35	94.3
T2	33	100
T3	32	100
T4	32	96.9
T5	35	100

Each locus had RAD genotyping calls in at least 30 individuals. At all 12 loci, SNP assay and RAD genotyping calls were identical in over 90% of individuals, indicating consistency in detecting species-diagnostic polymorphisms. The reliability of validated SNP assays in detecting species-diagnostic polymorphisms was explored further through DAPC analysis (FIGURE 3.6). The best-fitting DAPC model

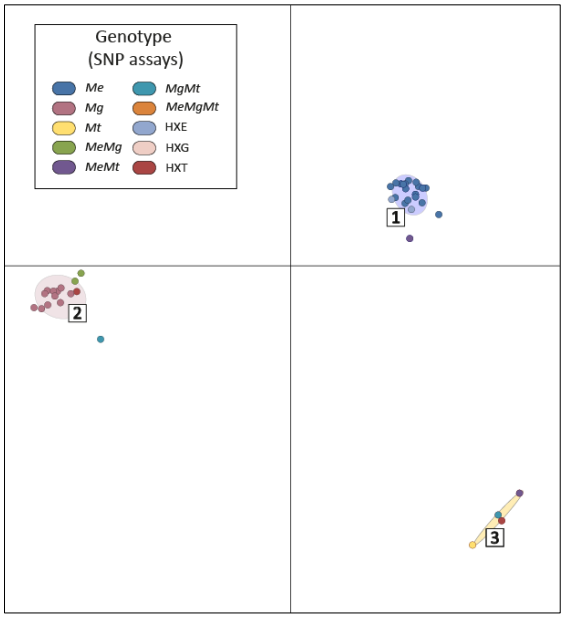

FIGURE 3.6 – DAPC scatterplot of clusters generated by PCA of composite genotypes of 40 individuals with 12 SNP assays (PCA = 95% cumulative variance; DAPC = 97.5% cumulative variance)**.** Genotypes are as follows: *M. edulis* (*Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis* (*MeMg*); hybrid of *M. edulis* and *M. trossulus* (*MeMt*); hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt*); hybrid of all three species (*MeMgMt*); hybrid with confirmed allelic contribution from one species only, with heterozygous loci (HXE = *M. edulis;* HXT = *M. trossulus*)

represented 97.5% of the cumulative variance: this was higher than the variation represented by the best-fit DAPC model of all potentially informative loci (FIGURE 3.4) and equal to the variation represented by the best-fit DAPC model of 18 potentially informative loci chosen for assay design (FIGURE 3.5). Individuals were grouped into three clusters: mostly, these clusters comprised pure specimens (*M. edulis*, *M. galloprovincialis* and *M. trossulus*), as in FIGURE 3.4, but the detection of hybrid individuals was seen to alter this composition slightly and possibly hinted at some introgression. However, the 12 optimised markers were still considered reliable tools for detecting species-specific polymorphisms due to: the notable separation of

different species according to their genotype with SNP assays; the high percentages of matching SNP assay and RAD genotyping calls; and successful genotyping of all 40 samples at all 12 loci. The suite of 12 optimised markers was deemed suitable for use with additional samples.

The remaining six assays could not be optimised. Two failed to amplify any product during PCR (E4 and E6). Four assays (E5, G5, G6 and G7) did produce fluorescent product during PCR, but this product did not consistently group into specific genotypic clusters. These six assays were excluded from further analysis.

3.3.4. SNP genotyping

3.3.4.1. PCR conditions

For primer sets E1, E2, E3, G1, G2, G3, T1, T2 and T3, each 5 µL reaction comprised 2.5 µL 2x KASP Master Mix, 0.07 µL KASP Assay Mix, 0.4 µL template DNA, and 2.1 µL UPW. Reactions were carried out under standard KASP thermal cycling conditions on a Biometra TGradient Thermocycler (as in SECTION 2.6.2). For primer sets G4, T4 and T5, each 5 µL reaction comprised 2.5 µL 2x KASP Master Mix, 0.07 µL KASP Assay Mix, 0.4 µL template DNA, 1.8 µL UPW and 0.25 µL 100% DMSO, and reactions were carried out under extended KASP thermal cycling conditions (as in SECTION 2.6.2). With all assays, two positive controls for each genotype (*M. edulis*, *M. galloprovincialis* and *M. trossulus*) were included to verify that the assays were working as expected, alongside negative controls (no template DNA) for fluorescent calibration and potential contamination identification. Fluorescent signals from the end-point assays were detected on a Techne Quantica Real Time PCR Thermal Cycler, using accompanying Quansoft software to visualise and score the genotypic assays.

3.3.4.2. Genotype class and Type

Genotyping individuals with all 12 diagnostic SNP markers distinguished pure individuals and introgressed (FX) individuals. Examples of the possible outcomes from SNP genotyping are detailed in TABLE 3.7. Each population could be divided into a series of Types (pure, hybrid and introgressed) and genotype classes (genotype based on alleles identified with SNP assays).

TABLE 3.7 – Genotypes of presumed pure individuals and F1 hybrids, and example genotypes of introgressed individuals (FX) after genotyping with 12 diagnostic SNP assays, where **D** – homozygous for diagnostic SNP allele; H – heterozygous genotype; n – homozygous for non-diagnostic SNP allele.

		***Me* loci**			***Mg* loci**				***Mt* loci**					
Type	**Genotype class**	E1	E2	E3	G1	G2	G3	G4	T1	T2	T3	T4	T5	**Notes**
Pure	*Me*	**D**	**D**	**D**	n	n	n	n	n	n	n	n	n	Only *Me* diagnostic alleles
	Mg	n	n	n	**D**	**D**	**D**	**D**	n	n	n	n	n	Only *Mg* diagnostic alleles
	Mt	n	n	n	n	n	n	n	**D**	**D**	**D**	**D**	**D**	Only *Mt* diagnostic alleles
F1 hybrid	F1 *MeMg*	H	H	H	H	H	H	H	n	n	n	n	n	100% heterozygous *Me* and *Mg*
	F1 *MeMt*	H	H	H	n	n	n	n	H	H	H	H	H	100% heterozygous *Me* and *Mt*
	F1 *MgMt*	n	n	n	H	H	H	H	H	H	H	H	H	100% heterozygous *Mg* and *Mt*
Introgressed hybrids (FX)	*MeMg*	**D**	**D**	**D**	n	n	H	n	n	n	n	n	n	*Me* and *Mg* diagnostic alleles
	MeMt	n	H	H	n	n	n	n	n	H	**D**	**D**	n	*Me* and *Mt* diagnostic alleles
	MgMt	n	n	n	n	n	H	H	n	**D**	H	n	n	*Mg* and *Mt* diagnostic alleles
	MeMgMt	n	H	n	n	H	n	n	n	H	n	n	H	*Me*, *Mg* and *Mt* diagnostic alleles
	HXE	**D**	H	**D**	n	n	n	n	n	n	n	n	n	*Me* contribution confirmed
	HXT	n	n	n	n	n	n	n	**D**	**D**	H	H	**D**	*Mt* contribution confirmed
	HXG	n	n	n	**D**	**D**	**D**	H	n	n	n	n	n	*Mg* contribution confirmed

Individuals that were 100% homozygous for the diagnostic allele at all species diagnostic loci in a single species, and 100% homozygous for the non-diagnostic allele in other species, would be considered pure species. Individuals that were 100% heterozygous at diagnostic loci of two species would be classed as F1 hybrids. Individuals heterozygous at diagnostic loci or with diagnostic alleles of multiple species, in any other proportion, would be classed as introgressed FX hybrids (i.e., F2 and beyond). Genotype classes could be one of 13 possibilities depending on the composite genotype of an individual (see APPENDIX 6D for the list of all composite genotypes). *Me, Mg* and *Mt* referred to pure *M. edulis, M. galloprovincialis* and *M. trossulus* respectively. Their hybrids are named according to the combination of alleles identified in each. All F1 hybrids had allelic contributions from two species in a 50:50 ratio (F1 *MeMg,* F1 *MeMt* and F1 *MgMt* hybrids). Introgressed FX hybrids had allelic contributions from two species (*MeMg*, *MeMt* and *MgMt* hybrids); confirmed allelic contributions from three species (*MeMgMt* hybrids); and confirmed allelic contribution from one species only, with one or more heterozygous diagnostic locus (HXE = *M. edulis;* HXG = *M. galloprovincialis;* HXT = *M. trossulus*) (TABLE 3.7).

3.3.5. Genotypes per site

3.3.5.1. Individual Type: pure species or introgressed

It was possible to successfully genotype all 228 individuals at all 12 loci. Rascarrel Bay (*M. edulis*) and Bay of Piran (*M. galloprovincialis*) had the highest proportions of pure individuals (both 76%), followed by Loch Ryan (64% *M. edulis*), Penn Cove (25% *M. trossulus*; 12.5% *M. edulis*) Bras d'Or Lake (6% *M. trossulus*) and Loch Etive (5% *M. trossulus*). The remaining individuals in all population samples were introgressed hybrids (FX): the highest percentages were in Loch Etive (95%) and Bras d'Or Lake (94%), followed by Penn Cove (62.5%), Loch Ryan (36%), and Bay of Piran and Rascarrel Bay (24%). No F1 hybrids were identified in any of the population samples (FIGURE 3.7).

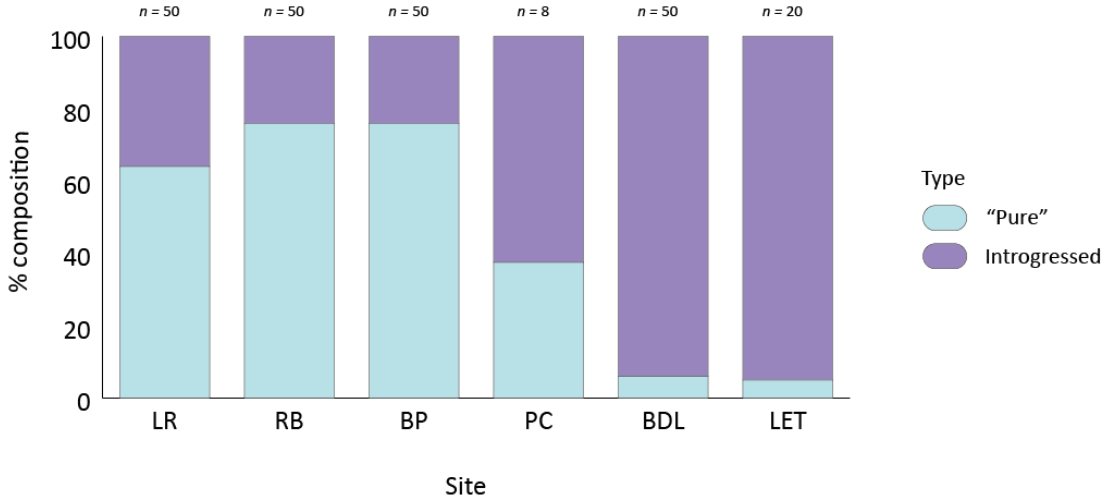

FIGURE 3.7 – The proportions of pure and introgressed individuals, detected with multilocus genotyping using 12 SNP assays. No F1 hybrids were identified. Site names are abbreviated as follows: Loch Ryan (LR); Rascarrel Bay (RB); Bay of Piran (BP); Penn Cove (PC); Bras d'Or Lake (BDL); Loch Etive (LET)

3.3.5.2. Genotypes with Me15/16 and SNP assays

Single locus genotyping with Me15/16 identified only pure (homozygous) individuals at Rascarrel Bay, Bay of Piran and Loch Etive. These comprised 50 *M. edulis*, 50 *M. galloprovincialis* and 20 *M. trossulus* respectively. The samples from Loch Ryan, Penn Cove and Bras d'Or Lake each comprised a mixture of homozygous and hybrid (heterozygous) individuals. Loch Ryan comprised 94% homozygous *M. edulis* individuals and 6% heterozygous hybrids of *M. edulis* and *M. galloprovincialis*. Penn Cove comprised 75% homozygous *M. trossulus*, 12.5% homozygous *M. edulis* and 12.5% heterozygous hybrids of *M. galloprovincialis* and *M. trossulus*. Bras d'Or Lake comprised 80% homozygous *M. trossulus* and 20% heterozygous hybrids of *M. edulis* and *M. trossulus* (FIGURE 3.8A).

Multilocus genotyping with 12 SNP assays identified introgression where single locus genotyping had been unable, revealing a more complex array of hybrid genotypes in all population samples (FIGURE 3.8B). Across the six population samples genotyped, a total of 13 different genotype classes were identified.

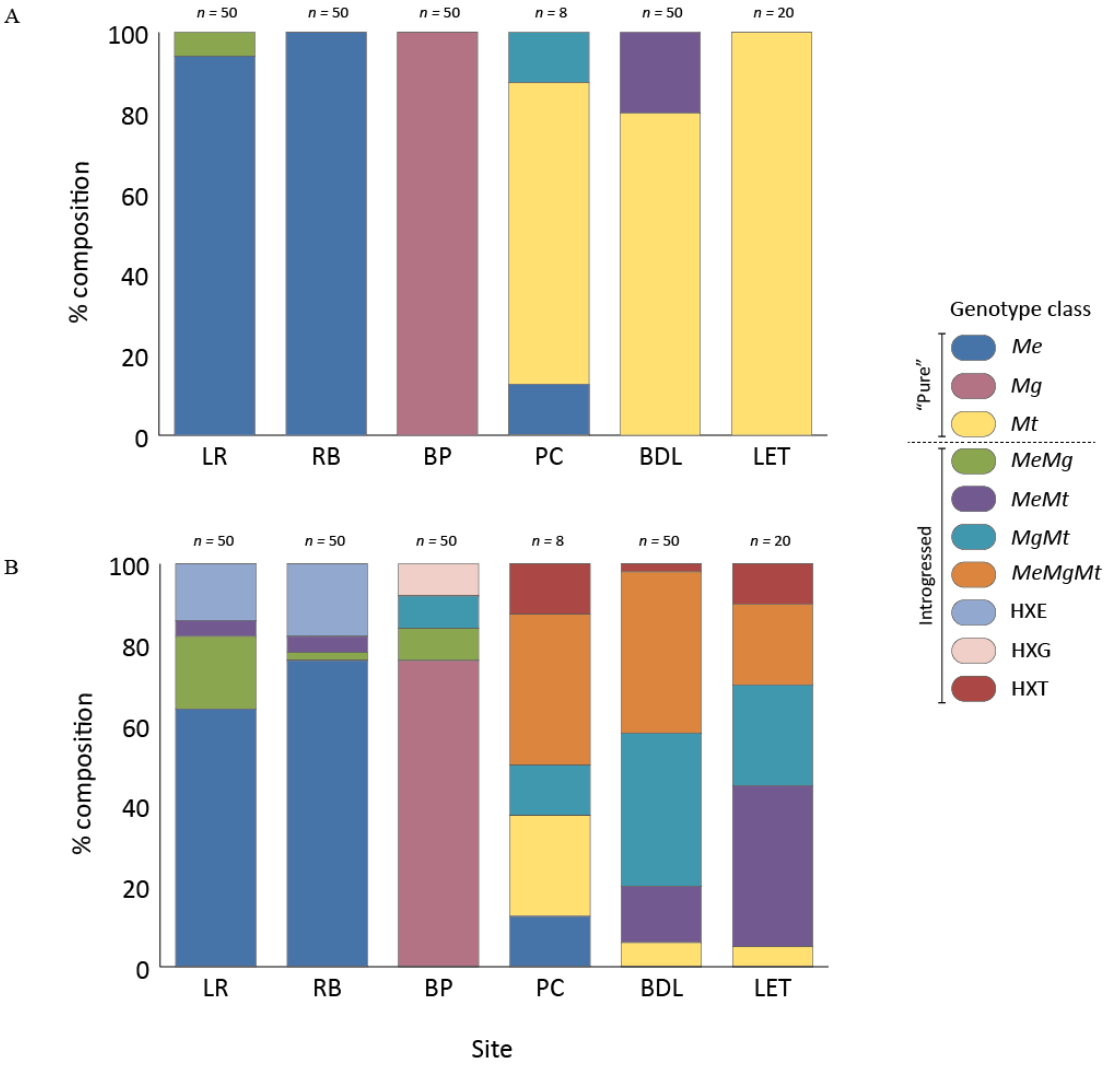

FIGURE 3.8 – Genotypes of *Mytilus* individuals generated from (A) single locus analysis with Me15/16 and (B) multilocus genotyping with 12 SNP markers. Genotype classes are as follows: *M. edulis* (*Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis* (*MeMg*); hybrid of *M. edulis* and *M. trossulus* (*MeMt*); hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt*); hybrid of all three species (*MeMgMt*); hybrid with confirmed allelic contribution from one species only, with heterozygous loci (HXE = *M. edulis;* HXG = *M. galloprovincialis;* HXT = *M. trossulus*). Site names are abbreviated as follows: Loch Ryan (LR); Rascarrel Bay (RB); Bay of Piran (BP); Penn Cove (PC); Bras d'Or Lake (BDL); Loch Etive (LET).

Pure genotype classes were identified in all population samples; these numbers were the same as in FIGURE 3.7. The specific genotype classes of introgressed (FX) hybrids per population is detailed in FIGURE 3.8B (see APPENDIX 6C for genotype classes per individual). Introgressed *MeMg* hybrids were identified in Loch Ryan (18%), Bay of Piran (8%) and Rascarrel Bay (2%;); *MeMt* hybrids were identified in Loch Etive (40%), Bras d'Or Lake (14%), Loch Ryan and Rascarrel Bay (4%); and

MgMt hybrids were identified in Bras d'Or Lake (38%), Penn Cove (37.5%), Loch Etive (25%) and Bay of Piran (8%;). *MeMgMt* hybrids were identified in Bras d'Or Lake (40%) and Loch Etive (20%) only. HXT hybrids were identified in Penn Cove (12.5%), Loch Etive (10%) and Bras d'Or Lake (2%); HXE hybrids were identified in Rascarrel Bay (18%) and Loch Ryan (14%;); and HXG hybrids were identified in Bay of Piran only (8%).

3.3.6. Analysis of genotyping data

3.3.6.1. PCA and DAPC analysis

On a broader scale, the suite of 12 SNP markers had consistently recognised species-diagnostic polymorphisms in population samples, and had successfully recognised introgression that had been unidentifiable with single locus genotyping. The ability of these markers to separate species and their hybrids was evaluated through PCA and DAPC analysis (FIGURE 3.9). The best-fitting PCA model

FIGURE 3.9- DAPC scatterplot of clusters generated by PCA of composite genotypes of 228 individuals with 12 SNP assays. PCA = 97.5% cumulative variance; DAPC = 95% cumulative variance. Genotypes are as follows: *M. edulis* (*Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis* (*MeMg*); hybrid of *M. edulis* and *M. trossulus* (*MeMt*); hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt*); hybrid of all three species (*MeMgMt*); hybrid with confirmed allelic contribution from one species only, with heterozygous loci (HXE = *M. edulis;* HXT = *M. trossulus*)

represented 97.5% of the cumulative variance within the SNP genotyping data from all 228 samples. DAPC analysis of PCA represented 95% of cumulative variance, and grouped this data into three clusters of presumed pure individuals interspersed with hybrid specimens.

3.3.6.2. Allele frequencies

The frequencies of alleles per locus per population sample were calculated using the GENALEX package for Microsoft Excel (version 6.5; Peakall and Smouse, 2006, 2012) (TABLE 3.8). Species reference samples for *M. edulis* (Loch Ryan and Rascarrel Bay) had diagnostic allele frequencies ≥ 0.92 at all *M. edulis* diagnostic loci; allele frequencies ≤ 0.05 at all *M. galloprovincialis* diagnostic loci; and allele frequencies ≤ 0.02 at all *M. trossulus* diagnostic loci. The species reference sample for *M. galloprovincialis* (Bay of Piran) had allele frequencies ≥ 0.96 at all *M. galloprovincialis* diagnostic loci; and allele frequencies ≤ 0.03 at *M. edulis* and *M. trossulus* diagnostic loci. The Penn Cove species reference sample for *M. trossulus* had a frequency of alleles ≥ 0.88 at all *M. trossulus* diagnostic loci, except at locus T4 which had an allele frequency of 0.563. *M. edulis* diagnostic loci had frequencies ≤ 0.188, and all *M. galloprovincialis* diagnostic loci had allele frequencies ≤ 0.250. The species reference sample for *M. trossulus* from Bras d'Or Lake did have some moderate input from *M. edulis* (0.32 at locus E1) and *M. galloprovincialis* (0.52 at locus G4) alleles, but all other allele frequencies at *M. edulis* and *M. galloprovincialis* diagnostic loci were ≤ 0.05. Overall, the input from *M. trossulus* alleles at *M. trossulus* diagnostic loci was much higher (≥ 0.95). The Loch Etive species reference sample for *M. trossulus* had a frequency of *M. trossulus* alleles ≥ 0.8 at all *M. trossulus* diagnostic loci. Frequencies of *M. edulis* alleles were very low at *M. edulis* diagnostic loci (≤ 0.2), and *M. galloprovincialis* alleles were only identified at the G3 locus (0.1). Data generated with GENALEX was exported to GENEPOP for Hardy Weinberg Exact tests (TABLE 3.9). It would be expected that a naturally outbreeding population was in Hardy Weinberg Equilibrium, but these tests revealed many loci to be out of equilibrium (i.e., $p < 0.05$). Here, where the values are statistically significant and out of equilibrium, it

TABLE 3.8 – Allelic frequencies per locus per population, calculated with GENALEX. For all loci, **D** refers to the diagnostic allele and n refers to the non-diagnostic allele. Sites are named as in TABLE 3.1

		Locus																					
		E1		E2		E3		G1		G2		G3		G4		T1		T2		T3		T4		T5	
	Allele	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n	**D**	n
Site	LR	0.93	0.07	0.95	0.05	0.97	0.03	0.05	0.95	0.05	0.95	0.03	0.97	0.04	0.96	0.02	0.98	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	RB	0.99	0.01	1.00	0.00	0.92	0.08	0.00	1.00	0.02	0.98	0.00	1.00	0.00	1.00	0.02	0.98	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	BP	0.02	0.98	0.00	1.00	0.03	0.97	1.00	0.00	1.00	0.00	0.99	0.01	0.96	0.04	0.03	0.97	0.01	0.99	0.00	1.00	0.00	1.00	0.00	1.00
	PC	0.19	0.81	0.19	0.81	0.13	0.88	0.00	1.00	0.13	0.88	0.00	1.00	0.25	0.75	0.88	0.13	0.88	0.13	0.88	0.13	0.56	0.44	0.88	0.13
	BDL	0.32	0.68	0.02	0.98	0.00	1.00	0.02	0.98	0.04	0.96	0.05	0.95	0.52	0.48	0.97	0.03	0.96	0.04	0.99	0.01	0.95	0.05	1.00	0.00
	LET	0.20	0.80	0.10	0.90	0.20	0.80	0.00	1.00	0.00	1.00	0.10	0.90	0.00	1.00	0.80	0.20	0.90	0.10	0.90	0.10	0.90	0.10	1.00	0.00

was due to a deficit in heterozygosity, likely arising from incomplete admixture of two (or more) populations.

TABLE 3.9 – p values generated per locus per population with Hardy Weinberg exact tests with GENEPOP: values <0.05 indicate a locus out of equilibrium; – denotes a locus for which no data was available. Site names are abbreviated as follows: Loch Ryan (LR); Rascarrel Bay RB); Bay of Piran (BP); Penn Cove (PC); Bras d'Or Lake (BDL); Loch Etive (LET).

Locus	LR	RB	BP	PC	BDL	LET
E1	0.0115	-	1.0000	0.1981	0.5330	0.0040
E2	1.0000	-	-	0.1964	1.0000	0.0003
E3	1.0000	1.0000	1.0000	0.0649	-	0.0002
G1	0.0018	-	-	-	1.0000	-
G2	0.0013	0.0093	-	0.0699	1.0000	-
G3	1.0000	-	-	-	1.0000	1.0000
G4	1.0000	-	0.0609	0.0151	0.0505	0.2055
T1	1.0000	1.0000	1.0000	0.0666	0.0300	0.0720
T2	-	-	-	0.0654	1.0000	0.0000
T3	-	-	-	0.0652	-	0.0000
T4	-	-	-	0.0499	1.0000	0.0000
T5	-	-	-	0.0670	-	0.0000

3.3.6.3. Genotype compositions per population sample

Multilocus genotyping detected a total of 65 composite introgressed genotypes across six population samples. The introgressed genotype class with the most variations (i.e., the most composite genotype combinations) was *MgMt* (*n*= 17), followed by *MeMgMt* (*n*=14); *MeMg* (*n*=13); *MeMt* (*n*=11); HXE (*n*=4); HXG and HXT (*n*=3) (APPENDIX 6D). Of these 65 composite genotypes, 56 were unique for one of six population samples, and 9 were shared by at least two of six population samples (TABLE 3.10).

TABLE 3.10 – Numbers of unique and shared composite genotypes in populations used for marker development and validation. Site names are abbreviated as follows: Loch Ryan (LR); Rascarrel Bay (RB); Bay of Piran (BP); Penn Cove (PC); Bras d'Or Lake (BDL); Loch Etive (LET)

Population	Nº unique genotypes	Nº shared genotypes
LR	9	4
RB	0	4
BP	8	0
PC	2	2
BDL	22	4
LET	15	2

There was no obvious bias towards a single marker as polymorphisms were distributed across loci and each population sample (with the exception of Rascarrel Bay) had its own range of unique composite genotypes. Of the additional population samples used for *M. trossulus* marker validation, Bras d'Or Lake shared three genotypes with Penn Cove (*Mt, MeMt* and HXT) and two genotypes with Loch Etive (*Mt* and *MgMt*). Bay of Piran shared no other genotypes with any population sample. Most hybrid genotypes identified in Bras d'Or Lake and Loch Etive had homozygous *M. trossulus* loci and polymorphisms at *M. edulis* and *M. galloprovincialis* loci.

3.4. DISCUSSION

Hybridisation and introgression amongst *M. edulis, M. galloprovincialis* and *M. trossulus* has, historically, led to debate among taxonomists about their exact relationship, in particular the relationship between *M. edulis* and *M. galloprovincialis,* because of widespread morphological and physiological similarities throughout their ranges [e.g., Lewis and Powell (1961); Seed (1971); Gosling (1984); Varvio *et al* (1988)]. Despite taxonomic uncertainties, there is evidence of genetic divergence in allozymes and mitochondrial DNA that has allowed *M. edulis, M. galloprovincialis* and *M. trossulus* to be considered distinct from each other (Edwards and Skibinski, 1987; Koehn, 1991; McDonald *et al.*, 1991), although *M. edulis* and *M. galloprovincialis* are sometimes regarded as being more closely related than either species is related to *M. trossulus* (Seed, 1971; Gosling, 1984; Brooks, 2000). The phylogenetic analysis of RADseq data presented here was consistent with such evidence. It showed three distinct genetic clusters, corresponding to *M. edulis, M. galloprovincialis* and *M. trossulus,* but also showed a greater genetic similarity between *M. edulis* and *M. galloprovincialis* than either *M. edulis* or *M. galloprovincialis* to *M. trossulus*. The distinctions identified between these species supported the idea that unique SNPs identified with RADseq could be suitable species diagnostic markers. F_{ST} values of potentially informative loci selected for assay design suggested clear differentiation between species, except in three cases where these values were negative (E5, G1 and G2). Negative values are, however, likely to have arisen from small and variable sample sizes between groups used in RAD library construction, a problem sometimes observed with *Stacks*

(Catchen *et al.,* 2013), so these values should be considered a guideline only. The small sample sizes used in RAD library construction (particularly *M. trossulus*) did call the diagnostic capabilities of potential markers into question, pending validation through additional genotyping – an essential step in genetic marker development (observed in, e.g., Zbawicka *et al.,* 2012; Palaiokostas *et al.,* 2013a; Peñarrubia *et al.,* 2015).

It was possible to genotype all additional 178 individuals with all 12 optimised assays. Far greater levels of hybridisation were detected than had been with single locus (Me15/16) genotyping. This did suggest the markers may not have been totally diagnostic, although it is extremely unlikely for any genetic marker, even a SNP with a low mutation rate, to be 100% diagnostic (Anderson and Thompson, 2002). An examination of allele frequencies demonstrated that in each species reference sample, diagnostic allele frequencies were consistently higher than non-diagnostic allele frequencies. Polymorphisms were detected across loci rather than an individual locus consistently generating unexpected results, and each population sample (except Bay of Piran) had a range of unique and shared genotypes. Thus, despite the hybridisation observed, each assay overall performed consistently and accurately in detecting *M. edulis, M. galloprovincialis* and *M. trossulus* alleles, thereby improving marker robustness and verifying their diagnostic properties on a small scale. These results offer an encouraging start in beginning a detailed assessment of genetic structure in field samples of *Mytilus* mussels.

RADseq, on average, identified hundreds of thousands of loci in the *M. edulis, M. galloprovincialis* and *M. trossulus* genomes. The estimated genome size of *M. edulis* is 1.56 Gb (Tanguy *et al.,* 2013) but there is no estimated genome size available for *M. galloprovincialis* or *M. trossulus*. Genome size can still be highly variable even between closely related eukaryotic organisms (Biémont, 2008; Muñoz-Diez *et al.,* 2011), which could explain the overall lower number of *M. trossulus* diagnostic loci assembled here. However, the smaller number of assembled loci is more likely a reflection of the small sample size of *M. trossulus* sequenced. Due to the fact that these species can hybridise to produce fertile offspring, they probably have similar genome sizes and numbers of chromosomes (Stelzer *et al.,* 2011). Thus, the RADseq approach adopted here could be considered an effective tool for recognising

polymorphisms within, and for *de novo* assembly of, *M. edulis*, *M. galloprovincialis* and *M. trossulus* genomes.

Although itself a powerful genotyping tool, since its development and publication by Baird *et al* (2008) the original RADseq approach has been adapted to suit a range of genetic studies, each of which has its own pros and cons (Puritz *et al.*, 2014): for instance, double digest RAD (ddRAD, Petersen *et al.*, 2012); ezRAD (Toonen *et al.*, 2013); and 2bRAD (Wang *et al.*, 2012). Following the successful generation of data from RADseq, ddRAD was also applied to the present study with additional samples of *M. trossulus* (including *M. edulis* control samples), given the small sample size used in the RAD library. A large majority of loci identified were highly polymorphic, were identified in very few individuals, and failed to reliably distinguish one species from the other. ddRAD can be particularly vulnerable to allelic dropout, the likelihood of which increases with the cumulative length of restriction enzyme sites (Arnold *et al.*, 2013; Gautier *et al.*, 2013; Andrews *et al.*, 2016), and subsequently may not be suitable for more sensitive population genetic studies (Puritz *et al.*, 2014). This perhaps presented a problem in the case of closely related *Mytilus* spp. genomes. It may be possible to further modify the ddRAD protocol attempted here (based on Petersen *et al.*, 2014) for successful sequencing of *Mytilus* species. However, ddRAD was not considered any further for the purposes of this study.

Previous studies identifying SNPs in *Mytilus* spp. (Vera *et al.*, 2010; Zbawicka *et al.*, 2012; Wenne *et al.*, 2016) have used capillary based electrophoresis rather than High Throughput Sequencing (HTS). HTS was applied here with the aim of genotyping a greater number of loci, and subsequently identifying a greater number of diagnostic markers (Davey and Blaxter, 2010; Peterson *et al.*, 2012) than possible with older sequencing technologies (Kircher and Kelso, 2010), and for a reduced cost (Hert *et al.*, 2008). A total of 349 loci (362 SNPs) were identified with RADseq, greater than the number identified and optimised by Vera *et al* (2010) (n=10), Zbawicka *et al* (2012) (n=21) and Wenne *et al* (2016) (n=54) for genotyping. However, of the 18 loci chosen for assay design, only 12 were optimisable. Difficulties with optimisation of six assays could have arisen from errors in sequencing and primer design, leading to failed or incomplete PCR amplification. Structural errors in the *de novo* sequence (from insertion, inversion or deletion) and

random point mutations cannot be identified without a reference genome (Davey and Blaxter, 2010; Leggett and MacLean, 2014). Mistakenly identifying these errors as actual sequence can lead to improper primer design and, subsequently, failed PCR reactions, which could have accounted for assays E4 and E6 failing to amplify DNA even after extensive optimisation attempts. In the event that actual SNPs were detected [i.e., polymorphisms occurring at a frequency of >1 % (Perkel, 2008)] and assays produced unclear genotyping clusters or low levels of fluorescence, primers had possibly been designed incorrectly on either side of the SNP in question. This may have arisen from incomplete contig (short sequences) overlap during *de novo* assembly (Salzberg and Yorke, 2005; Willing *et al.*, 2011), and could explain why assays E5, G5, G6 and G7 did not fluoresce properly. Unclear genotyping clusters may also have been due to SNPs located in null alleles, where mutated restriction sites are not recognised by restriction enzymes and DNA is not cut. Loci at these cut sites will not be sequenced, and any SNPs corresponding to diagnostic alleles will not be genotyped correctly (Andrews *et al.*, 2016). Additional diagnostic assays could be optimised from the remaining 331 loci that were not used for assay design, thereby giving rise to a larger panel of SNPs for future genotyping studies. This remains to be investigated.

A comparison of single and multilocus genotyping demonstrated that single locus genotyping (with Me15/16) had overestimated the degree of homozygosity (i.e., pure individuals) in population samples from Scotland, Slovenia, North America and Canada. A similar phenomenon was observed in Greenland (Wenne *et al.*, 2016), where multilocus SNP genotyping identified *M. trossulus* and its hybrids in a lake that, until then, had been assumed to contain only *M. edulis*. Multilocus genotyping in the present study detected a series of introgressed genotypes (*MeMgMt*, HXE, HXG and HXT) that single locus genotyping with Me15/16 failed to identify, confirming a single marker was insufficient for resolving intogression in field samples of *Mytilus* mussels. DAPC analysis (Jombart *et al.*, 2010) has been widely applied to multilocus genotyping studies of shellfish populations (e.g., Coscia *et al.*, 2013; Zardi *et al.*, 2015; Gormley *et al.*, 2015; Lal *et al.*, 2016; Lal *et al.*, 2017), including studies of *Mytilus* spp. (Giantsis *et al.*, 2014; Araneda *et al.*, 2016). As with any multivariate analysis technique reliant on transformation of data, there is a

risk of losing important information when modelling with DAPC (Jombart *et al.*, 2009; Dufresne *et al.,* 2014). Nevertheless, each of the DAPC models applied to the present study represented at least 90% of variation within the dataset, and showed consistent separation of individuals according to their genotype: three clusters, each corresponding to a different species, interspersed with hybrid genotypes where applicable. A similar trend is observable in studies of *Mytilus* spp. by Zbawicka *et al* (2012) and Wenne *et al* (2016). Both studies used Correspondence Analysis (Benzecri, 1992), a technique analogous to DAPC that models genetic structure at individual and population levels, to analyse the results of multilocus SNP genotyping. Species and hybrids were clearly separated according to their genotype, with more overlap observable between *M. edulis* and *M. galloprovincialis* when compared to *M. trossulus.* This was noticeable in DAPC of RAD data (and similar to phylogenetic analysis). DAPC of SNP assay data differed slightly, showing a more equal differentiation between the three clusters which was likely a result of the smaller number of markers (Jombart *et al.,* 2010). Nevertheless, DAPC of SNP assay data still successfully separated species and hybrid individuals, lending further support to the presence of three distinct species and the diagnostic capabilities of the markers on a small scale. The usefulness of Me15/16 for preliminary genotyping cannot be discounted, but if introgression is to be investigated using multilocus SNP genotyping would be more appropriate in future studies.

Although widespread introgression was detected, no F1 hybrids were identified; thus, it was not possible to say from these results if the optimised assays could differentiate F1 and FX hybrids. The apparent absence of F1 hybrids in these population samples may be explainable by the fact that successful interspecies hybridisation is relatively infrequent and has not occurred recently enough for F1 hybrids to be present or detected (Pujolar *et al.,* 2014). This certainly seems probable given the high levels of introgression identified (Fogelqvist *et al.,* 2015), and because each population sample included individuals of one species only (i.e., either *M. edulis, M. galloprovincialis* or *M. trossulus*). Both *M. edulis* and *M. trossulus* were identified in the Penn Cove population sample: however, this population sample had an unknown origin and such a small size that it is impossible to say that F1 hybrids were likely among the individuals genotyped. A mussel crossing experiment

(following steps outlined in Helm *et al.,* 2004) was trialled in order to validate and genotype pure specimens and F1 hybrids for control samples, but these attempts proved unsuccessful and were not considered further.

It is well documented that *M. edulis, M. galloprovincialis* and *M. trossulus* are three interfertile species that are often grouped together in the "*M. edulis* species complex" (e.g., Gardner, 1996; Rawson *et al.,* 1996; Brooks, 2000; Gardeström *et al.,* 2008). As a naturally diploid genus (Kiyomoto *et al.,* 1996), the appearance of *Mytilus* hybrids with genetic material from three related species (*MeMgMt*) could be considered unlikely and an error from inefficient genotyping. Genetic mixing over multiple generations has the capacity to create an enormous array of highly complex genotypes (Harrison and Larson, 2014; Patel *et al.,* 2015); certainly, alleles of all three species have been reported from single locus genotyping (Beaumont *et al.,* 2008; Dias *et al.,* 2011a) and multilocus SNP genotyping (Zbawicka *et al.,* 2012) of mussels in Loch Etive, so in theory it is feasible for *MeMgMt* hybrids to exist. *M. edulis* and *M. trossulus* have historically been documented in Bras d'Or Lake (Qiu *et al.,* 2002), and occur naturally in hybridising populations in Atlantic Canada (Penney *et al.,* 2002; LeBlanc *et al.,* 2005). No evidence is available that recognises *M. galloprovincialis* in or around this area. It is possible that *M. galloprovincialis* and its hybrids could have been introduced to Bras d'Or Lake through commercial activity at nearby mussel farms [e.g., on Prince Edward Island (Brooks, 2000; LeBlanc *et al.,* 2005)], and subsequently hybridised with *M. edulis* and *M. trossulus* hybrids to result in *MeMgMt* hybrids. Planktonic *Mytilus* larvae are easily transported outside of their natural range in the ballast water of ships (Carlton and Geller, 1993). In the case of HXE, HXG and HXT hybrids, it is possible that the proportion of genetic material from a second (or third) species was so low that not enough markers were used to detect the full extent of introgression (Currat *et al.,* 2008). Although each individual was genotyped at multiple loci, the actual numbers of loci diagnostic to a particular species remained small (*M. edulis n*=3; *M. galloprovincialis n*=4; *M. trossulus n*=5). With the widespread introgression observable here, especially in Penn Cove, Bras d'Or Lake and Loch Etive, it is possible that the individuals genotyped as pure were actually themselves introgressed, just less heavily than the individuals genotyped as

FX hybrids. Genotyping with a larger panel of diagnostic markers per species could potentially resolve this.

3.5. CONCLUSIONS AND SUMMARY

1. Analysis of SNPs identified with RADseq confirmed three genetically distinct species in the *M. edulis* species complex, presenting the possibility for exploring introgression with species diagnostic markers;

2. Using RADseq and KASP technology, and DAPC analysis, the present study successfully identified and validated a suite of 12 new species specific diagnostic SNP markers for multilocus genotyping of *Mytilus* mussel populations, allowing levels of admixture to be assessed;

3. The diagnostic properties of these markers on a broader scale has yet to be ascertained, but for the purposes of this study, multilocus genotyping of *Mytilus* populations enabled more confident designation of pure species individuals of *M. edulis, M. galloprovincialis* or *M. trossulus* compared to single locus genotyping with Me15/16*;*

4. Multilocus genotyping provided an insight into possible patterns of hybridisation and introgression: no F1 hybrids were identified but FX hybrids were present in all population ssamples, with hybridisation taking place across the genome

Acknowledgements

Thank you to Heiko Stuckas (Senckenberg Natural History Museum), Andreja Ramšak (NIB, Ljubljana), Barry MacDonald and Ellen Kenchington (Bedford Institute of Oceanography) for donation of *M. trossulus, M. galloprovincialis* and *M. trossulus* tissue samples respectively. Thanks to Rebecca McIntosh (Marine Scotland Science) for additional DNA extraction and Me15/16 genotyping. The work was funded by Marine Alliance for Science and Technology for Scotland (MASTS) and Marine Scotland Science.

Chapter 4

Assessing levels of genetic admixture of *Mytilus edulis* with congeneric species *M. galloprovincialis* and *M. trossulus*

4.1. INTRODUCTION

Hybridisation and introgression contribute to increasing genetic diversity in a natural environment (Rieseberg *et al.*, 2003; Arnold and Martin, 2009; Abbott *et al.*, 2013; Seehausen 2013). Increased genetic diversity can be beneficial as it may lead to improved performance of hybrid offspring compared to their parents, a phenomenon termed "hybrid vigour" or "heterosis" (Barton, 2001; Birchler *et al.*, 2010; Baranwal *et al.*, 2012). Hybrid vigour occurs most prominently in first generation (F1) hybrids and is commonly exploited commercially because of its positive impact on production: e.g., increased meat yields in hybrids between the channel and blue catfish (*Ictalurus* spp.) (Argue *et al.*, 2003; Bosworth *et al.*, 2004); and improved disease resistance in hybrids from different lines of *Cyprinus carpio* (common carp) (Linneaus, 1758) (Kirpichnikov *et al.*, 1993). F1 hybrids showing hybrid vigour are often sterile or have reduced fertility compared to their parents (Naisbit *et al.*, 2001; Suzuki and Nachman, 2015). This prevents F1 hybrids from outcompeting their parents and reduces the chances of introgression taking place, preserving the genetic integrity of parental populations. Introgression becomes more of a possibility when F1 hybrids are fertile but do not display hybrid vigour, thereby coexisting and backcrossing with parental forms rather than outcompeting them. These hybrids either have no obvious phenotypic difference from their parents [termed "cryptic hybrids" (Gibson and Dworkin, 2004; Haynes *et al.*, 2012; Mckean *et al.*, 2016], or they show an obvious (and possibly disadvantageous) phenotype compared to their parents (Burke and Arnold, 2001).

Both cryptic and obvious F1 hybrids have the potential to negatively affect production and profitability on a commercial scale. Genetic contamination of broodstock can cause problems for sourcing and rearing of pure species [e.g., in cryptic hybrids of the silver and bighead species of carp in Bangladesh (Hussain and Mazid, 2001; Mia *et al.*, 2005)]. In the Scottish shellfish industry, *Mytilus trossulus*

has been associated with undesirable traits in farmed mussels (i.e., fragile shells and poor meat yields) (Dias *et al.,* 2008). There is some evidence that hybridisation between *M. trossulus* and the native *Mytilus edulis* also confers undesirable traits upon individuals, but this is not guaranteed in all hybrids (Beaumont *et al.,* 2008). *M. trossulus* is native to the Pacific and colonised parts of the north Atlantic, including the Baltic Sea, around 3.5 million years ago after the Bering Strait opened (Riginos and Cunnigham, 2005). *M. trossulus* has been identified at a small number of sites on the west coast of Scotland (Gubbins et al., 2012), including Loch Etive. Evidence from mtDNA shows that *M. trossulus* in Loch Etive has a Pacific origin (Zbawicka *et al.*, 2010) and is most likely part of a relict population rather than being a recent introduction (Beaumont *et al.,* 2008). Within the last decade, an increased presence of fragile-shelled mussels has contributed to a decline in shellfish production at mussel farms in Loch Etive (Dias *et al.,* 2011a). Given the documented link between *M. trossulus* and shell fragility, *M. trossulus* is now classed as a commercially damaging species under Scottish law (Aquaculture and Fisheries (Scotland) Act 2013) because its spread could be detrimental for the industry. This legislation does not, however, apply to *M. trossulus* hybrids because morphological identification is unreliable and existing genotyping methodologies are not always practical for recognising hybridisation and introgression. Depending on the degree of introgression with *M. trossulus*, populations may or may not be at risk of shell fragility – if, indeed, shell fragility is controlled by a genetic factor. Thus, for effective population management, extra information about the presence of *M. trossulus* and its hybrids in Scotland is needed to investigate its distribution, and any possible threats to production, in greater detail.

M. edulis and *M. trossulus* in Scotland exist alongside a third related species, *Mytilus galloprovincialis* [all of which are grouped in the *"Mytilus edulis* species complex" (Fly and Hilbish, 2013)]. Hybridisation between species of the *M. edulis* species complex involves both pre- and post-zygotic isolation mechanisms that prevent or reduce interspecific gene exchange (Toro *et al.,* 2002; Bierne *et al.,* 2006; Doherty *et al.,* 2009; Monteiro *et al.,* 2012). Wherever interbreeding between two overlapping populations occurs and offspring with mixed ancestry are produced, a hybrid zone is formed (Barton and Hewitt, 1985). Hybrid zone stability is influenced

by multiple environmental and genetic factors (e.g., spawning time, larval dispersal rate and habitat preferences), and selection on exchanged genes (Hilbish *et al.*, 2003). Selection in hybrid zones depends on the relative fitness of parent species and their hybrid offspring, and environmental conditions (Hatfield and Schluter, 1999). In a tension hybrid zone, hybrids will be selected against while parental genotypes are favoured (Key, 1968). In a clinal hybrid zone, hybrids persist where the ranges of parental species overlap and occupy their own niche separate from the parents (Endler, 1977). In a mosaic hybrid zone, common amongst *Mytilus* species, parents and their hybrid offspring co-exist because neither displays an advantage over the other, resulting in equal selection pressures for both parents and hybrids (Gilg and Hilbish, 2003). There is evidence of *Mytilus* mosaic hybrid zones along the Atlantic coast of Europe (Bierne *et al.*, 2003; Daguin and Borsa, 2001; Varela *et al.*, 2007) and the Pacific coast of North America (Rawson and Hilbish, 1995; Rawson *et al.*, 1999). In Scotland, widespread hybridisation between species of the *M. edulis* species complex has been documented (Dias *et al.*, 2011a), but additional genetic data is required to estimate the degree of introgression and thus to confirm any existence of active hybrid zones.

To date, the majority of studies of *Mytilus* spp. in Scotland have been carried out with single locus genotyping using Me15/16 (Inoue *et al.*, 1995) (Beaumont *et al.*, 2008; Zbawicka *et al.*, 2010; Dias *et al.*, 2011a). These studies have acknowledged the presence of the *M. edulis* species complex and its hybrids. However, use of a single marker is extremely limiting in resolving introgression in a population (Twyford and Ennos, 2012), which in turn imposes limits on sourcing quality broodstock and effective stock management. Theoretically, where hybrids are sterile and no introgression takes place, controlling an undesirable species should be relatively straightforward (Huxel, 1999). Efficient stock management does, however, become more challenging when hybrids are fertile because there is a greater chance of interspecies genetic mixing in existing populations and in mussel broodstock (Bondad-Reantaso, 2007; Prado *et al.*, 2012; Jasper *et al.*, 2013). The performance of introgressed stock depends on the level of introgression and environmental conditions, which can vary between locations (Gibson and Dworkin, 2004). In comparison to single locus genotyping, multilocus genotyping allows a more reliable

assessment of population structure and introgression patterns. In a commercial aquaculture setting, genetic mixing between farmed and wild stocks of fish poses problems for production through the spread of undesirable traits (Hindar *et al.*, 2006). Several studies have used multilocus genotyping to investigate introgression and its effects on production: e.g., Glover *et al* (2013) used 99 SNP markers to test for introgression from farmed individuals in wild Norwegian stocks of Atlantic salmon; and Varne *et al* (2015) used 11 microsatellite loci to test for introgression of farmed and wild cod in Norway. Multilocus genotyping has been applied to mussels from Loch Etive (Zbawicka *et al.*, 2012), but there is otherwise no evidence of multilocus genotyping being used in Scottish *Mytilus* spp. populations. Applying multilocus genotyping on a wider scale would allow the extent of introgression with *M. trossulus* to be assessed in greater detail, subsequently enabling the development of improved management strategies for efficient production within the Scottish shellfish industry.

The present study utilises multilocus genotyping with 12 diagnostic markers to identify alleles of either *M. edulis*, *M. galloprovincialis* or *M. trossulus* in mussels collected from 23 sites around the Scottish coastline. The aims of the study are as follows:

1. To use multilocus SNP genotyping to assess current levels of admixture in Scottish population samples, by identifying pure individuals and different types of hybrids [First generation (F1) or second generation and beyond (FX)];

2. To determine potential links between species distribution and site type (rope or shoreline) and how this could relate to broodstock sourcing and site selection

3. To investigate the possible relationship between shell fragility and the *M. trossulus* allele

4.2. METHODS

4.2.1. Sample collection

Between 2012 and 2014, a total of 22 Scottish sites were sampled. Adult mussels (min. 40 mm in length) were collected from 21 sites: Dornoch Firth; Ferryness; Flotta; Kyelsku; Loch Ailort; Loch Eireasort; Loch Fyne; Loch Laxford; Loch Linnhe; Loch Long; Loch Roag; Loch Ryan; Loch Spelve; Lunderston Bay; Montrose; Northside; Rascarrel Bay; Scapa Beach; Shetland BR; Shetland BX; and St. Andrews. Juvenile mussels (approx. 15 months old) were collected from one site at Loch Etive (TABLE 4.1). Sites comprised a mixture of rope grown aquaculture

TABLE 4.1 – Sampling site details. Excluding Loch Etive, the coordinates of aquaculture sites are not provided. R = rope grown aquaculture; S = shoreline (bottom grown aquaculture or wild)

Site number	Site location	GPS coordinates	Source	*n*	Date sampled
1	Loch Eireasort	-	R	49	June 2014
2	Loch Roag	-	R	50	June 2014
3	Loch Spelve	-	R	50	Aug 2014
4	Scapa Beach	58°56'47.00"N | 2°59'13.27"W	S	10	Oct 2013
5	Northside	59°09'25.37"N | 3°12'50.53"W	S	10	Nov 2013
6	Flotta	-	R	45	Dec 2012
7	Shetland BX	-	R	45	Nov 2012
8	Shetland BR	-	R	45	Nov 2012
9	Montrose	56°42'16.31"N | 2°28'13.71"W	S	49	Feb 2014
10	St Andrews	56°20'07.67"N | 2°48'23.28"W	S	50	Feb 2014
11	Loch Ryan	54°56'06.83"N | 5°03'38.69"W	S	50	Feb 2013
12	Rascarrel Bay	54°48'53.11"N | 3°51'22.74"W	S	50	Feb 2013
13	Loch Laxford	-	R	30	Sept 2014
14	Kylesku	-	R	28	Sept 2014
15	Dornoch Firth	-	S	40	Oct 2014
16	Loch Ailort	-	R	50	Sept 2014
17	Loch Linnhe	-	R	28	Nov 2012
18	Ferryness	55°58'56.78"N | 2°54'40.65"W	S	50	Mar 2014
19	Lunderston Bay	55°55'31.49"N | 4°52'51.19"W	S	45	Feb 2013
20	Loch Long	56°02'09.51"N | 4°53'14.41"W	S	47	Feb 2013
21	Loch Fyne	-	R	92	Nov 2012
22	Loch Etive	56°27'05.53"N | 5°19'13.32"W	R	80	April 2013
23	Site X	-	R	39	May 2016

(n=12) and shoreline (wild and bottom grown aquaculture, n=10) locations on the Scottish mainland and islands (FIGURE 4.1). Sampling size varied between sites. A total of 991 individuals were collected. It was preferred that a minimum of 50 individuals were collected from each shoreline site, but this was not always possible due to variations in mussel availability and ease of sampling (i.e., weather

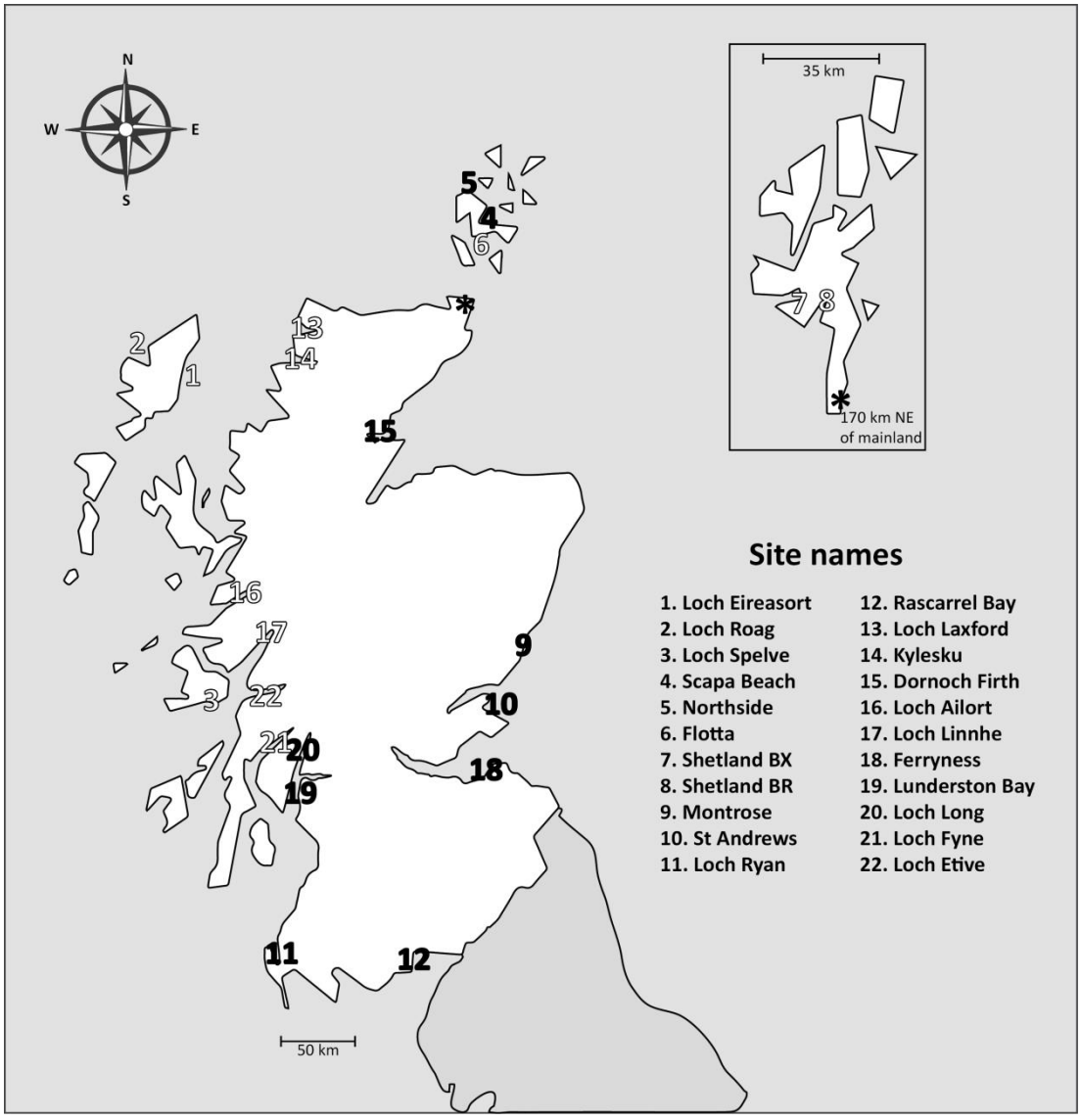

FIGURE 4.1 – Map of sampling site locations in Scotland. Numbers of wild shoreline and bottom grown aquaculture sites (15) are in black and numbers of rope grown aquaculture sites are in white.

conditions, tide height and number of people collecting samples). The distances from the shoreline and the depth of sampling were not recorded. A piece of gill/mantle (approximately 10 mg for manual extraction; 50 mg for automated extraction) was taken from all adult mussels and stored in 99% ethanol at -20°C prior to DNA extraction. All body tissues were taken from juvenile mussels and stored in 99% ethanol at -20°C prior to DNA extraction. DNA was extracted from both adults and juveniles using the automated and manual methods described in SECTION 2.2. In June 2016, an additional 39 adult mussels with both strong (n=20) and fragile (n=19) shells were obtained from "Site X", a site of commercial importance on the west coast of Scotland that wished to remain anonymised (TABLE 4.1). Shells from

Site X were classed as “strong” if they remained intact during mechanical harvesting, and classed as “fragile” if they did not. DNA was extracted from these mussels using manual SSTNE/SDS extraction (see SECTION 2.2.2.1), and their shells were retained for a comparison of phenotype with genotype.

4.2.2. Genotyping

4.2.2.1. Me15/16 PCR

Me15/16 was used to genotype a total of 532 individuals from 11 sites (Flotta, Loch Etive, Loch Linnhe, Loch Long, Loch Ryan, Lunderston Bay, Montrose, Rascarrel Bay, Shetland BR, Shetland BX and St Andrews) for some preliminary taxonomic data to compare with SNP genotyping. PCR with Me15/16 was carried out as per the conditions in SECTION 2.3.1.

4.2.2.2. SNP assays

Twelve diagnostic SNP markers were used to genotype a total of 1030 individuals (991 individuals from named sites, plus 39 individuals from Site X). All SNP assays were designed and manufactured by LGC Genomics Limited for use with KASP genotyping technology (see TABLE 3.4). SNP assay conditions and details of the SNP genotype calling process are detailed in SECTION 3.3.4.

4.2.3. Analysis of genotyping data

4.2.3.1. Inferring population structure

Population structure was determined using STRUCTURE (version 2.3) (Pritchard *et al.,* 2000) and NEWHYBRIDS (version 1.0) (Anderson and Thompson, 2002). STRUCTURE was used for a general overview of population admixture and grouping of population samples according to their degree of introgression, whereas NEWHYBRIDS was used to try and classify individuals into specific “genotype frequency classes” (i.e., pure or hybrid).

For STRUCTURE, most parameters were set to their default values as advised in the STRUCTURE 2.3 user manual (Pritchard *et al.,* 2010). Specifically, the admixture model with correlated allele frequencies between populations was chosen: this configuration is recommended by Falush *et al* (2003) as the most suitable for

resolving cryptic population structure [i.e., population structure that is difficult to detect phenotypically but which may be significant in genetic terms (Pritchard *et al.*, 2000), as in *Mytilus* spp. mussels]. The lengths of MCMC and burn-in were varied from 100 to 100,000. A value of 10,000 each proved to be sufficient; longer values did not obviously alter the results. The range of possible *K* values tested ranged from 1 to the total number of sampled populations (26). The optimal *K* was determined from 100 iterations of each *K* value, according to the method outlined by Evanno *et al* (2005) and tested using CLUMPAK software (Kopelman *et al.,* 2015). The optimal *K* by Evanno's method was 4 ($\Delta K = 64.706$).

For NEWHYBRIDS, only pure individuals and hybrids with a clear genetic input from two discrete species [i.e., *M. edulis* x *M. galloprovincialis* hybrid (*MeMg*); *M. edulis* x *M. trossulus* hybrid (*MeMt*); *M. galloprovincialis* x *M. trossulus* hybrid (*MgMt*)] could be included in simulations, as this software is specifically designed for use with crosses of two diploid species (Anderson and Thompson, 2002; Anderson, 2008). Any individuals that did not meet these criteria were excluded. To reduce simulation runtime, ten reference individuals for each species were included in initial simulations as relevant: *M. edulis* from a single Scottish site (Loch Ryan); *M. galloprovincialis* from a single site in Slovenia (Bay of Piran); and *M. trossulus* from one site in Scotland (Loch Etive, n=5), one site in North America (Penn Cove, n=2), and one site in Canada (Bras d'Or Lake, n=3). All hybrid individuals identified were included in simulations as relevant: *MeMg* n=274; *MeMt* n=128; *MgMt* n=6. Ten simulated hybrid individuals of each genotype frequency class were included as relevant to verify the reliability of category assignment.

Due to each hybrid cross having different numbers of loci (*MeMg*=7*; MeMt*=8; *MgMt*=9), separate simulations with a range of genotype frequency classes were trialled for each to find which simulation most accurately assigned individuals to a specific category, as determined prior to the simulation. A threshold of ≥0.55 (55%) was used to assign an individual to a specific class. Individuals that were either *M. edulis, M. galloprovincialis* or *MeMg* hybrids were segregated most clearly by the simulation using a total of four predefined genotype frequency classes (PureMg, PureMe, F1 and FX), the frequencies of which are detailed in TABLE 4.2.

TABLE 4.2 - Genotype frequency classes assumed for a NEWHYBRIDS model for *M. edulis, M. galloprovincialis* and their hybrids. PureMg and PureMe refer to pure *M. galloprovincialis* and pure *M. edulis* respectively; F1 refers to first generation hybrid; FX refers to a hybrid of second generation or beyond

Genotype frequency class	Frequency of AA	Frequency of AB	Frequency of BA	Frequency of BB
PureMg	1.00	0.00	0.00	0.00
PureMe	0.00	0.00	0.00	1.00
F1	0.00	0.5	0.5	0.00
FX	0.25	0.25	0.25	0.25

Individuals that were either *M. edulis, M. trossulus* or *MeMt* hybrids were segregated most clearly by the simulation using a total of five predefined genotype frequency classes (PureMt, PureMe, F1, 1FX and 2FX), the frequencies of which are detailed in TABLE 4.3.

TABLE 4.3 - Genotype frequency classes assumed for a NEWHYBRIDS model for *M. edulis, M. trossulus* and their hybrids. PureMt and PureMe refer to pure *M. trossulus* and pure *M. edulis* respectively; F1 refers to first generation hybrid; 1FX and 2FX all refer to hybrids of second generation or beyond, with different allele frequency combinations

Genotype frequency class	Frequency of AA	Frequency of AB	Frequency of BA	Frequency of BB
PureMt	1.00	0.00	0.00	0.00
PureMe	0.00	0.00	0.00	1.00
F1	0.00	0.5	0.5	0.00
1FX	0.25	0.25	0.25	0.25
2FX	0.5	0.00	0.00	0.5

All NEWHYBRIDS simulations had a burnin and MCMC length of 100,000 for a total of five chains, as per the simulation parameters used in Anderson and Thompson (2002). Once optimal simulation conditions had been determined, simulations were re-run with the same parameters using all *M. edulis* (n=373) and *M galloprovincialis* (n=50) individuals where relevant to verify the accuracy of category assignment across a broader scale. Additional *M. trossulus* was unavailable. Crosses of *MgMt* hybrids could not be adequately modelled due to small sample numbers and lack of genetic variation among hybrids, so these were excluded from further analysis.

4.2.3.2. PCA, DAPC and allele frequencies

Following multilocus SNP genotyping, Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were carried out using the *adegenet* package (version 1.4-1; Jombart, 2008) for R (version 3.1.0) (R Core

Team, 2014). PCA creates simplified models of the total variation within the dataset (Jackson, 1991), and DAPC identifies clusters of genetically related individuals (Jombart *et al.,* 2010) within the most statistically likely PCA model, determined using the Bayesian Information Criterion (Schwarz, 1978) (see APPENDIX 4B for PCA and DAPC script). DAPC was used to group individuals into clusters of like individuals, showing the possible relationships between pure and admixed genotypes.

In order to infer genetic contributions from *M. edulis, M. galloprovincialis* and *M. trossulus* in population samples, an equation was designed based on the equations for calculating genotype and allele frequencies (EQUATION 3 in SECTION 2.6.3.1, detailed below; see SECTION 2.6.3.1 for full details of genotype and allele frequency calculations).

$$\text{R(A) per individual} = \frac{[2(n\text{AA}) + n\text{AB}]}{2M}$$

$$\text{R(A) per population} = \frac{\sum_{.}^{T} \left(\frac{[2(n\text{AA}) + n\text{AB}]}{2M} \right)}{T} \qquad (3)$$

This estimated only the proportion of diagnostic alleles in a population sample relative to the number of diagnostic markers used for genotyping. In the equation, R(A) is the relative proportion of diagnostic alleles; *n* represents the number of individuals; and *T* represents the total number of individuals in a population sample. "A" and "B" refer to two alleles at a biallelic locus: "AA" is homozygous diagnostic and "AB" is heterozygous; homozygous non-diagnostic "BB" individuals were excluded. *M* represents the number of diagnostic markers, which differed according to species: *M*=3 for *M. edulis; M*=4 for *M. galloprovincialis; M*=5 for *M. trossulus.* Values for *n*(AA) and *M* are multiplied by 2 to represent diploid organisms with two alleles at each locus.

4.3. RESULTS

4.3.1. Genotypes per site

SNP assay genotyping divided each population into a series of Types (pure, hybrid and introgressed) and genotype classes (genotype based on alleles identified with 12 SNP assays) (TABLE 4.4; see also SECTION 3.3.4.2 for further details). SNP assays successfully genotyped all 1030 sampled individuals at all 12 loci.

TABLE 4.4 – Types and Genotype classes (Pure and hybrid) according to genotyping results with 12 SNP assays. Genotype classes are as follows: *M. edulis* (*Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis* (*MeMg*); hybrid of *M. edulis* and *M. trossulus* (*MeMt*); hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt*); hybrid of all three species (*MeMgMt*); hybrid with confirmed allelic contribution from one species only (HXE = *M. edulis;* HXT = *M. trossulus).*

Type	Genotype class	Notes
Pure	*Me*	Only *Me* diagnostic alleles
	Mg	Only *Mg* diagnostic alleles
	Mt	Only *Mt* diagnostic alleles
F1 hybrid	F1 *MeMg*	100% heterozygous *Me* and *Mg*
	F1 *MeMt*	100% heterozygous *Me* and *Mt*
	F1 *MgMt*	100% heterozygous *Mg* and *Mt*
Introgressed (FX) hybrid	*MeMg*	*Me* and *Mg* diagnostic alleles
	MeMt	*Me* and *Mt* diagnostic alleles
	MgMt	*Mg* and *Mt* diagnostic alleles
	MeMgMt	*Me*, *Mg* and *Mt* diagnostic alleles
	HXE	*Me* contribution confirmed
	HXT	*Mt* contribution confirmed

4.3.1.1. Individual Type: pure species or introgressed

In total, 375 pure, 7 F1 hybrids and 609 FX hybrids were identified. FX hybrids were present at all 22 named sites. The highest percentage was in the wild Orkney populations, Scapa Beach and Northside (both 100%), and the lowest percentage was at Rascarrel Bay on the southwest coast (24%). With the exception of Scapa Beach and Northside, pure individuals were present at all sites. The highest percentage was in Rascarrel Bay (76%) and the lowest percentage was in Loch Laxford in the Highlands (3.3%). F1 hybrids were identified on the east coast in St Andrews (4.1%, F1 *MeMg*), and on the west coast in Loch Fyne (3.2%, F1 *MeMt*) and Loch Etive (2.5% F1 *MeMt*) (FIGURE 4.2).

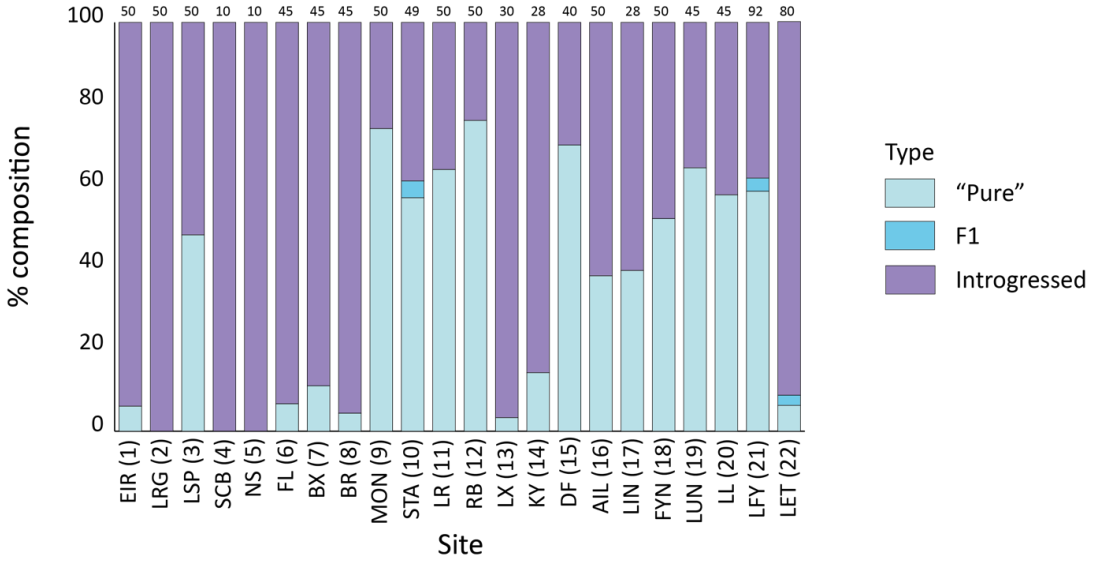

FIGURE 4.2 – The proportions of presumed pure, F1 hybrids and introgressed individuals at 22 Scottish sites, detected with multilocus genotyping using 12 SNP assays. The numbers of individuals genotyped from each site are at the top of each bar. Sites numbers and names are as follows: 1. Loch Eireasort (EIR); 2. Loch Roag (LRG); 3. Loch Spelve (LSP); 4. Scapa Beach (SCB); 5. Northside (NS); 6. Flotta (FL); 7. Shetland BX (BX); 8. Shetland BR (BR); 9. Montrose (MON); 10. St. Andrews (STA); 11. Loch Ryan (LR); 12. Rascarrel Bay (RB); 13. Loch Laxford (LX); 14. Kylesku (KY); 15. Dornoch Firth; 16. Loch Ailort (AIL); 17. Loch Linnhe (LIN); 18. Ferryness (FYN); 19. Lunderston Bay (LUN); 20. Loch Long (LL); 21. Loch Fyne (LFY); 22. Loch Etive (LET)

4.3.1.2. Genotypes with Me15/16 and SNP assays

Genotyping of 11 sites (rope grown aquaculture n=5; shoreline sites n=6) with Me15/16 identified six different genotype classes: *M. edulis* (*Me*), *M. galloprovincialis* (*Mg*) and *M. trossulus* (*Mt*), plus hybrids between *M. edulis* and *M. galloprovincialis* (*MeMg*), *M. edulis* and *M. trossulus* (*MeMt*), and *M. galloprovincialis* and *M. trossulus* (*MgMt*) (FIGURE 4.3A). SNP assays detected introgression and a greater number of genotype classes than Me15/16 had. Genotype classes were one of eight possibilities. *Me* and *Mt* referred to presumed pure *M. edulis* and *M. trossulus* respectively. Hybrids were named according to the combination of alleles identified in each, with *Mg* to represent *M. galloprovincialis* diagnostic alleles: *MeMg*, *MeMt* and *MgMt* hybrids had allele contributions from two species; *MeMgMt* hybrids had confirmed diagnostic allele contributions from three species; and HXE and HXT hybrids had a confirmed allele contribution from one species only (HXE = *M. edulis;* HXT = *M. trossulus*), but were heterozygous at one or more diagnostic loci (FIGURE 4.3B). The introgressed *MeMg* genotype class was

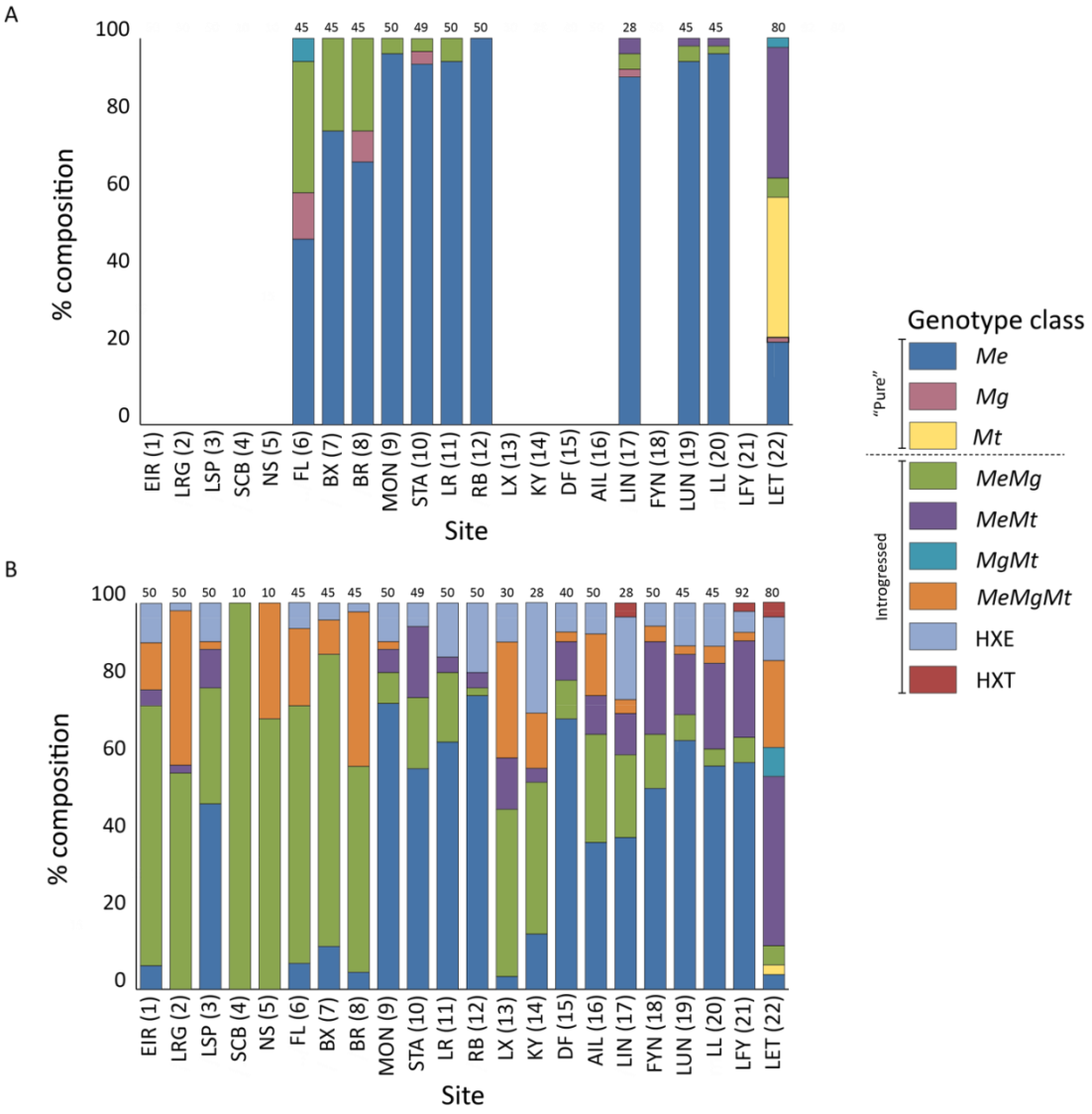

FIGURE 4.3 – Genotype classes per site, with data from (A) single locus genotyping with Me15/16 and (B) multilocus genotyping with 12 SNP assays. Sites are in the same order in both cases for ease of comparison, even though Me15/16 genotyping data was not available for all sites. The numbers of individuals genotyped from each site are at the top of each bar. Genotype classes are as follows: *M. edulis* (*Me); M. galloprovincialis (Mg); M. trossulus (Mt);* hybrid of *M. edulis* and *M. galloprovincialis* (*MeMg*); hybrid of *M. edulis* and *M. trossulus* (*MeMt*); hybrid of *M. galloprovincialis* and *M. trossulus* (*MgMt*); hybrid of all three species (*MeMgMt*); hybrid with confirmed allelic contribution from one species only (HXE = *M. edulis;* HXT = *M. trossulus).* Site numbers and names are as follows: 1. Loch Eireasort (EIR); 2. Loch Roag (LRG); 3. Loch Spelve (LSP); 4. Scapa Beach (SCB); 5. Northside (NS); 6. Flotta (FL); 7. Shetland BX (BX); 8. Shetland BR (BR); 9. Montrose (MON); 10. St. Andrews (STA); 11. Loch Ryan (LR); 12. Rascarrel Bay (RB); 13. Loch Laxford (LX); 14. Kylesku (KY); 15. Dornoch Firth; 16. Loch Ailort (AIL); 17. Loch Linnhe (LIN); 18. Ferryness (FYN); 19. Lunderston Bay (LUN); 20. Loch Long (LL); 21. Loch Fyne (LFY); 22. Loch Etive (LET)

identified at all 22 sites. The highest proportion was found at Scapa Beach (100%) and the lowest at Rascarrel Bay (2%). HXE was the next most widespread genotype class, being absent from Scapa Beach and Northside only, with the highest

proportion in the Highlands at Kylesku (28.6%), and the lowest in the Isle of Lewis at Loch Roag (2%). The *Me* genotype class was most abundant at Rascarrel Bay (76%) and the least abundant at Loch Laxford (3.3%), and was absent from Loch Roag, Scapa Beach and Northside. *MeMt* hybrids were most abundant at Loch Etive (40%), least abundant at Loch Spelve on the west coast (2%), and were absent from five sites (Scapa Beach, Northside and Flotta in Orkney, plus Shetland BR and Shetland BX). HXT hybrids were present at three sites on the west coast: Loch Etive (3.8%), Loch Linnhe (3.6%) and Loch Fyne (2.2%). Loch Etive was the only site where the *MgMt* (7.5%) and *Mt* genotype classes were identified (2.5%).

4.3.2. All genotypes

A total of 991 individuals (911 adults and 80 juveniles) were genotyped with 12 SNP assays. Overall, *Me* was the most abundant genotype class (*n*=373). *MeMg* was the next most abundant genotype class (*n*=274), followed by *MeMt* (*n*=128), *MeMgMt* (*n*=110), HXE (*n*=92), and *MgMt* and HXT (*n*=6). *Mt* was the least abundant genotype class (*n*=2) (FIGURE 4.4). Across all 22 named sites, a total of

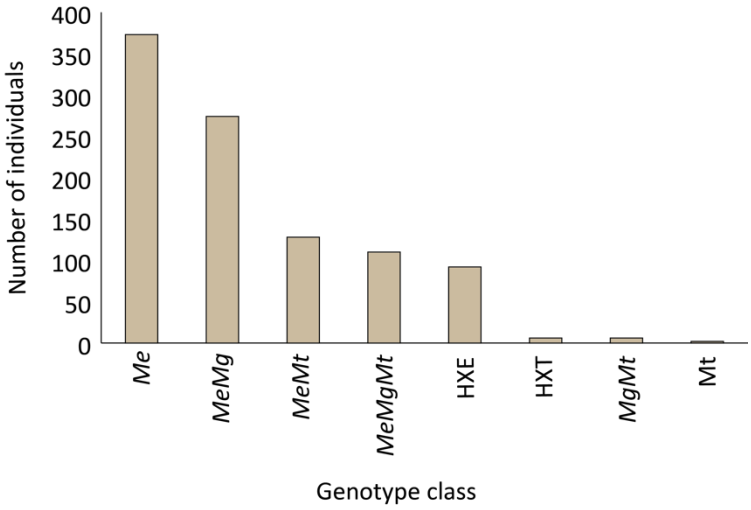

FIGURE 4.4 – Total numbers of all genotype classes identified among 22 Scottish population samples (*n*=991) after multilocus genotyping with 12 SNP assays

339 different composite genotypes (i.e., combined genotypes from 24 alleles, identifiable by SNPs spread across 12 loci) were identified in 991 individuals. The genotype class with the most variations was *MeMg* (*n*=144), followed by *MeMgMt*

(*n*=101); *MeMt* (*n*=75); HXE (*n*=10); HXT (*n*=5) and *MgMt* (*n*=4). Of these 339 composite genotypes, 283 were identified in one of 22 populations only, and the remaining 56 were shared by at least two populations (APPENDIX 7). Several PCA models were trialled to identify that which best represented variation among composite genotypes. Accordingly, the PCA model representing 90% of the cumulative variance within the dataset, and with the smallest Bayesian Information Criterion (BIC) value, was selected for DAPC (FIGURE 4.5). DAPC of SNP data

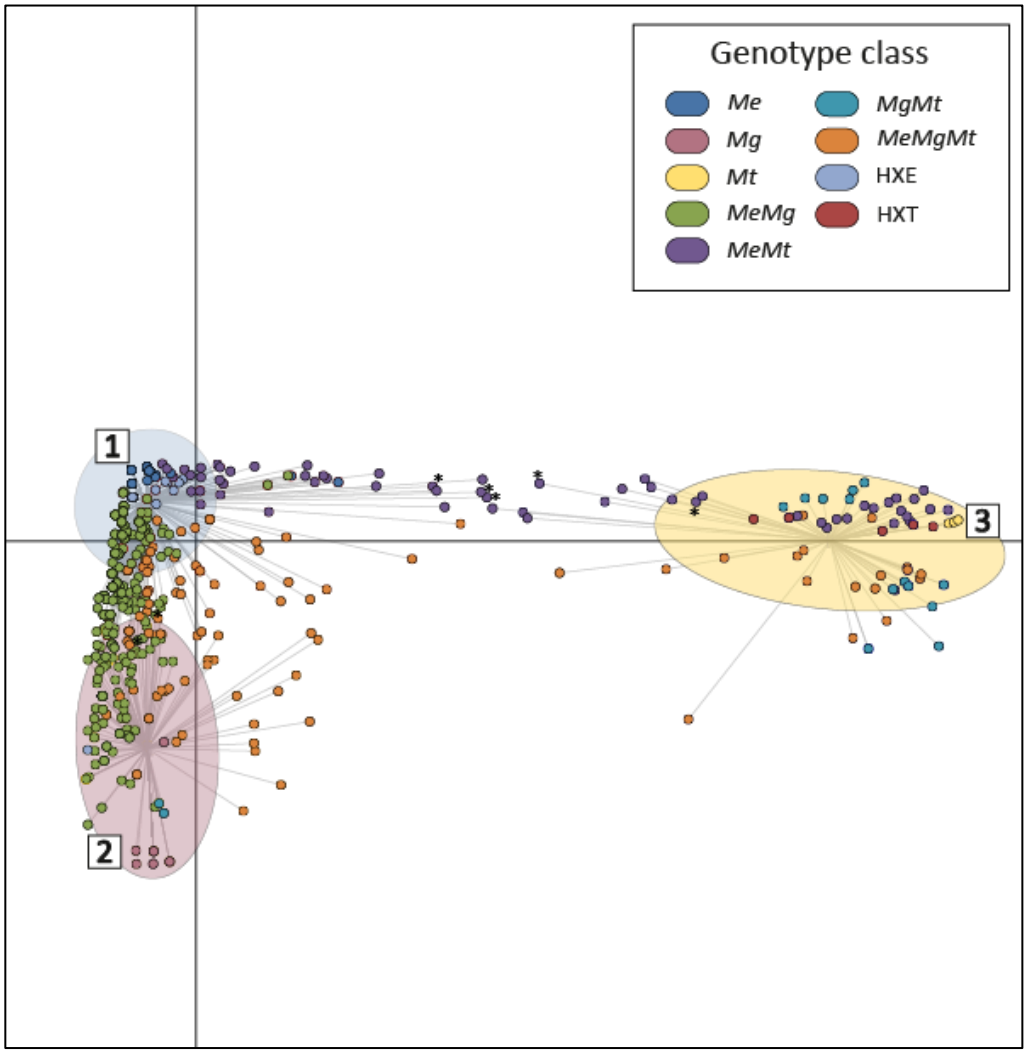

FIGURE 4.5 – DAPC scatterplot of clusters generated by PCA of 339 composite genotypes of 991 individuals at 12 biallelic loci. PCA = 90% cumulative variance; DAPC = 100% cumulative variance. F1 hybrids (individuals with 100% heterozygous loci) have been marked with *.

retained 100% of cumulative variance of PCA data, clearly differentiating *M. edulis*, *M. galloprovincialis* and *M. trossulus* while also showing widespread admixture

between Scottish population samples. *M. edulis* from all population samples was indistinguishable and formed a very tight group (Group 1). Loch Etive *M. trossulus* and additional *M. trossulus* from Bras d'Or Lake also formed a tight cluster in Group 3. *M. galloprovincialis* from Bay of Piran (used for reference) formed a less tight grouping but all individuals were in the same group (Group 2) and were distinct from *M. edulis* and *M. trossulus. MeMt* hybrids grouped either with *M. edulis* or *M. trossulus* or were placed between Group 1 and Group 3. *MeMg* hybrids grouped with either *M. edulis* or *M. galloprovincialis* or were placed between Group 1 and Group 2, but were clustered more tightly and showed more overlap than *MeMt* hybrids. HXE hybrids were mostly clustered with *M. edulis* in Group 1 and all HXT were clustered with *M. trossulus* in Group 3. *MeMgMt* hybrids mostly clustered outside of Groups 1-3, but some overlap was observable between *MeMgMt* from Group 1 and *MeMgMt* from Group 2. F1 *MeMt* hybrids were placed approximately halfway between Group 1 and Group 3 and were mostly distinguishable from FX hybrids. This distinction was not obvious between F1 *MeMg* and FX *MeMg* hybrids, which were clustered more closely together.

4.3.3. Admixture analysis

4.3.3.1. STRUCTURE

STRUCTURE (*K*=4, burnin = 10,000, reps = 10,000) identified a total of six different groups among the Scottish sites sampled (Group A, Group B, Group C, Group D, Group E and Group F), each of which comprised sites of similar genetic composition as denoted by membership proportion (q) values (TABLE 4.5; FIGURE 4.6). A seventh group (Group G) comprised three sites from outside of Scotland that were used as species reference samples for pure *M. galloprovincialis* (Bay of Piran) and pure *M.trossulus* (Penn Cove, Bras d'Or Lake). Sites in Group G were kept together for ease of reference and were not grouped by genotypic structure. Species reference samples for pure *M. edulis* were in Group F (Loch Ryan and Rascarrel Bay, both in Scotland). FIGURE 4.7 shows the location of all groups in Scotland (excluding Site X), determined by genetic structure. All sites in all groups exhibited

TABLE 4.5 – Average membership proportion (q) of each pre-defined population in each of the four clusters assigned by STRUCTURE. Groups A-F are arranged by genetic composition; Group G is for reference to *M. galloprovincialis* and *M. trossulus* and is not arranged by similar genetic composition. Species reference samples for *M. edulis* (Loch Ryan and Rascarrel Bay), *M. galloprovincialis* (Bay of Piran) and *M. trossulus* (Penn Cove and Bras d'Or Lake) are highlighted in grey. Nº refers to site number.

Group	Site	Nº	Type	*Mg*	*Mt*	*MeMg*	*Me*
A	Loch Roag	2	R	0.062	0.016	0.845	0.077
	Scapa Beach	4	S	0.22	0.005	0.612	0.163
	Northside	5	S	0.335	0.011	0.511	0.143
	Shetland BR	8	R	0.046	0.03	0.82	0.105
B	Loch Eireasort	1	R	0.067	0.013	0.667	0.253
	Flotta	6	R	0.089	0.042	0.629	0.24
	Shetland BX	7	R	0.047	0.005	0.663	0.286
	Loch Laxford	13	R	0.021	0.054	0.704	0.221
C	Kylesku	14	R	0.016	0.019	0.567	0.398
	Loch Ailort	16	R	0.028	0.019	0.408	0.545
D	Loch Linnhe	17	R	0.015	0.064	0.272	0.649
	Loch Fyne	21	R	0.005	0.079	0.143	0.773
E	Loch Etive	22	R	0.01	0.672	0.107	0.211
F	Loch Spelve	3	R	0.007	0.006	0.308	0.679
	Montrose	9	S	0.01	0.006	0.166	0.818
	St Andrews	10	S	0.021	0.02	0.182	0.777
	Loch Ryan	11	S	0.019	0.007	0.207	0.767
	Rascarrel Bay	12	S	0.004	0.004	0.116	0.876
	Dornoch Firth	15	S	0.005	0.013	0.178	0.805
	Ferryness	18	S	0.006	0.022	0.236	0.737
	Lunderston Bay	19	S	0.005	0.013	0.156	0.826
	Loch Long	20	S	0.004	0.024	0.183	0.788
	Site X	23	-	0.009	0.019	0.254	0.718
G	Bay of Piran	24	S	0.979	0.005	0.008	0.008
	Penn Cove	25	-	0.009	0.858	0.011	0.121
	Bras d'Or Lake	26	S	0.008	0.983	0.005	0.004

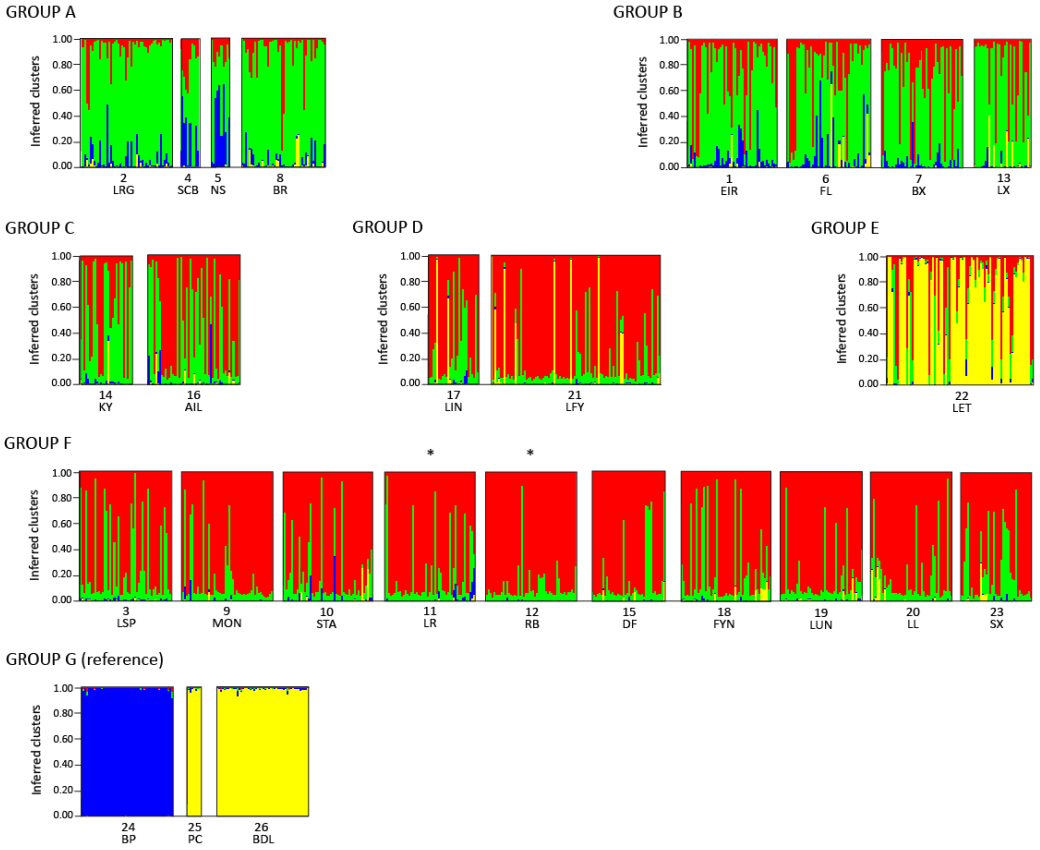

FIGURE 4.6 – Structure plots constructed using the Admixture Ancestry Model with independent allele frequencies per population [*K*=4 (Δ*K* = 64.706, determined from 100 iterations using Evanno's method (2005)), burnin = 10,000, reps = 10,000], showing the genetic composition of each site sampled. Each vertical line represents the genetic composition of an individual broken into four coloured segments, with lengths proportional to each of the inferred clusters: *M. edulis* = red*; M. galloprovincialis* = blue*; M. trossulus* = yellow*;* introgressed *MeMg.* = green. Sites in Groups A-F have been clustered according to their structure, and are listed along the x-axis. Groups A-F are all Scottish populations, with site names and numbers as follows: 1. Loch Eireasort (EIR); 2. Loch Roag (LRG); 3. Loch Spelve (LSP); 4. Scapa Beach (SCB); 5. Northside (NS); 6. Flotta (FL); 7. Shetland BX (BX); 8. Shetland BR (BR); 9. Montrose (MON); 10. St. Andrews (STA); 11. Loch Ryan (LR); 12. Rascarrel Bay (RB); 13. Loch Laxford (LX); 14. Kylesku (KY); 15. Dornoch Firth; 16. Loch Ailort (AIL); 17. Loch Linnhe (LIN); 18. Ferryness (FYN); 19. Lunderston Bay (LUN); 20. Loch Long (LL); 21. Loch Fyne (LFY); 22. Loch Etive (LET); 23. Site X (SX). Group G is a reference for pure *M. galloprovincialis* [Bay of Piran (BP)] and *M. trossulus* [Penn Cove (PC) and Bras d'Or Lake (BDL)]. Species reference samples for *M. edulis* are marked with a *.

some degree of admixture, which was generally higher in rope grown aquaculture sites compared to shoreline sites (TABLE 4.5). In Groups A-D, there was gradually less introgression between *M. edulis* and *M. galloprovincialis* (*MeMg*), and a greater presence of pure *M. edulis* (*Me*). Sites in Group A, Group B and Group C had the highest average q values in the *MeMg* cluster, ranging from q=0.845 (Loch Roag;

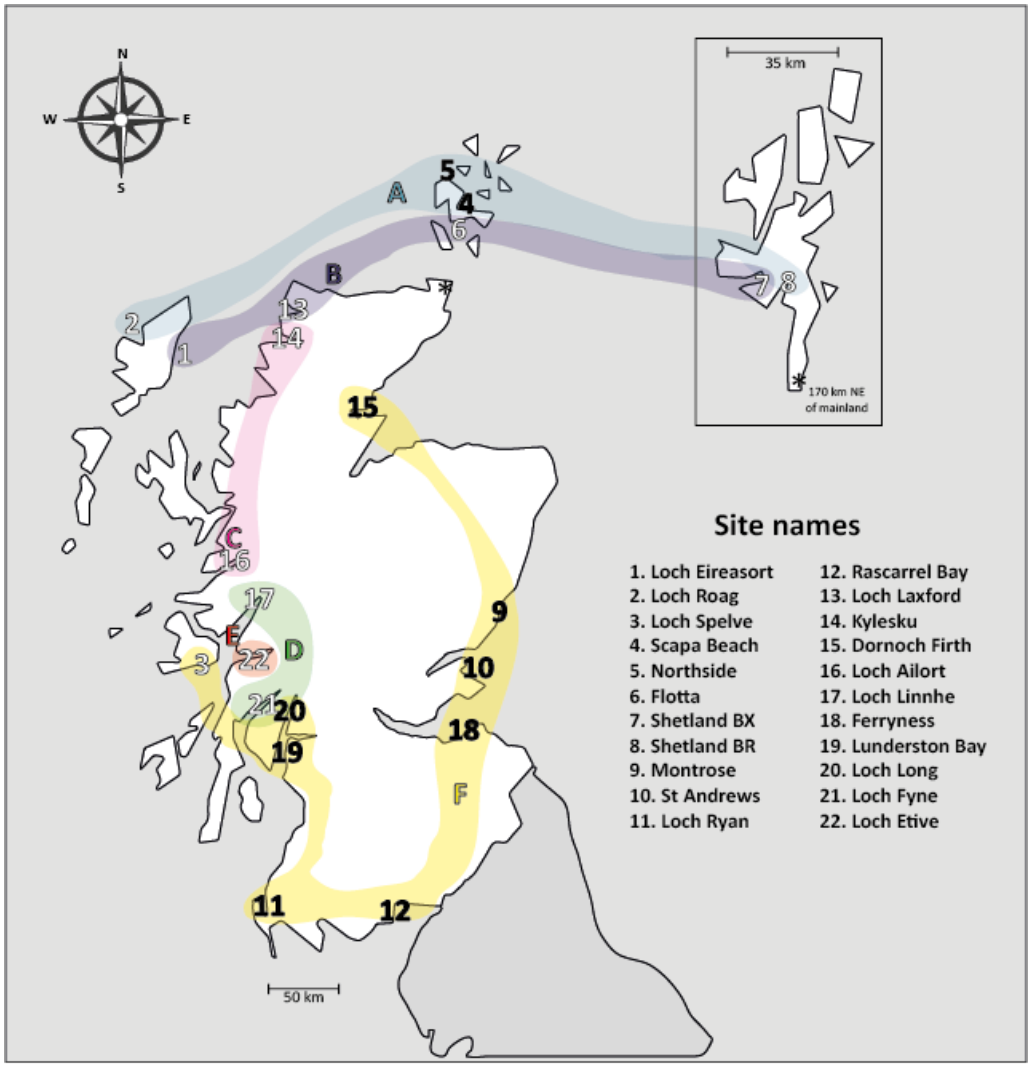

FIGURE 4.7 – Map of named sampling sites colour-coded according to group. Numbers of wild shoreline and bottom grown aquaculture sites (15) are in black and numbers of rope grown aquaculture sites are in white

Group A) to Loch Ailort (q=0.408; Group C). Sites in Group A, Group B and Group C all had membership proportions in the *Me* cluster: these values were lowest at sites in Group A (ranging from 0.077–0.163), higher at sites in Group B (ranging from 0.221–0.286); and highest in Group C [q=0.398 (Kylesku); q=0.545 (Loch Ailort)]. There was less *MeMg* input at sites in Group D [q=0.272 (Loch Linnhe); q=0.143 (Loch Fyne)], and a higher input from *Me* [q=0.649 (Loch Linnhe); q=0.773 (Loch Fyne)]. Sites in Group D had a higher input in the *Mt* cluster [q=0.064 (Loch Linnhe); q=0.079 (Loch Fyne)] than any of the sites in Groups A-C. Sites in Group F had the highest membership proportions in the *Me* cluster: Loch Ryan (q=0.767) and Rascarrel Bay (q=0.876) were species reference samples for pure *M. edulis:* all sites

in Group F (except Loch Spelve, Ferryness and Site X) had q values ≥0.767, and could thus be considered to comprise mostly pure *M. edulis.* Loch Spelve, Ferryness and Site X were included in Group F because they had higher *Me* q values than sites in Groups A-E, and also had very low input from *Mt.* Group E comprised a single site at Loch Etive with a notably different genetic composition from any other Scottish site genotyped. Excluding the species reference samples for pure *M. trossulus* in Group G, Loch Etive had the highest membership proportion in the *Mt* cluster (q=0.672). The input from *Me* (q=0.211) was similar to sites in Group B, and the *MeMg* input (q=0.107) was less than all sites in Group A, Group B, Group C, Group D and Group F. These values indicate levels of admixture between *M. edulis* and *M. trossulus* that was not observable at any other site genotyped.

STRUCTURE allele frequency divergence (F_{ST}) values indicated that, despite widespread admixture, a clear separation between the three *Mytilus* species (*M. edulis, M. galloprovincialis* and *M. trossulus*) remained (TABLE 4.6). All F_{ST} values

TABLE 4.6 – Allele frequency divergence (net nucleotide distance; F_{ST}) among *K* clusters, calculated using STRUCTURE. Clusters 1-4 correspond to the following genotype classes, as randomly assigned by STRUCTURE software: 1. *Mg* (*M. galloprovincialis*); 2. *Mt* (*M. trossulus*); 3. *MeMg* (hybrid of *M. edulis* and *M. galloprovincialis*); 4. *Me* (*M. edulis*)

Cluster	1 (*Mg*)	2 (*Mt*)	3 (*MeMg*)	4 (*Me*)
1 (*Mg*)	-	0.6521	0.3949	0.5634
2 (*Mt*)	0.6521	-	0.4981	0.5368
3 (*MeMg*)	0.3959	0.4891	-	0.0282
4 (*Me*)	0.5634	0.5368	0.0282	-

between clusters (*Me. Mg, Mt*) exceeded 0.25 which, according to the threshold values specified by Hartl and Clark (1997), indicated very great genetic differentiation. F_{ST} values between the *MeMg* cluster and the *Me*, *Mg* and *Mt* clusters were also >0.25, indicating very great genetic differentiation between this introgressed genotype, its parental species and *M. trossulus.*

At the individual level, STRUCTURE distinguished between pure and hybrid individuals but had limited capacity to discriminate between F1 hybrids and some FX hybrids, depending on the degree of genetic admixture. All (100%) of pure individuals were correctly assigned and had, on average, the following q values: *M. edulis* (q=0.93); *M. galloprovincialis* (q=0.99); *M. trossulus* (q=0.99). Lower q values indicated admixture. Of the hybrid Types identified in SECTION 4.3.1.1,

85.7% of F1 hybrids were correctly assigned (50:50 split between clusters), and 47% of FX hybrids were correctly assigned (split between clusters in variable proportions). The remaining 53% of FX hybrids were indiscriminate from F1 hybrids.

4.3.3.2. NEWHYBRIDS

The NEWHYBRIDS simulations that best represented genetic variation within datasets had the least number of ambiguous assignments among hybrids and consistent, high assignment probabilities (≥0.55) of pure individuals to specific genotype frequency classes (FIGURE 4.8). The overall efficiency of NEWHYBRIDS in correctly assigning high probabilities to the relevant genotype frequency class varied with simulation. Both simulations for *MeMg* and *MeMt* crosses consistently identified pure species, assigning 100% of pure individuals with high probability (0.99) to either PureMe, PureMg or PureMt. All simulated hybrid individuals were assigned with 100% accuracy to the relevant categories in selected simulation parameters (data not shown). The simulation for *MeMt* crosses correctly identified 63.7% FX hybrids, assigning individuals with high probability (≥0.55) to an introgressed genotype frequency class (1FX or 2FX). The remaining 36.3% of FX hybrids were misclassified as PureMe or PureMt. The capacity of NEWHYBRIDS to distinguish pure and hybrid individuals was greatly diminished when applied to the *MeMg* dataset: only 4.8% of FX *MeMg* were assigned with high probability (≥0.55) to the introgressed genotype frequency class (1FX), while the remaining 95.2% of FX hybrids were assigned with higher probability as PureMe. In both simulations for *MeMg* and *MeMt* crosses, NEWHYBRIDS failed to assign any F1 hybrids with high probability to the F1 genotype group. F1 assignment values were as follows: *MeMg* hybrids from St Andrews (STA_16=0.00 and STA_29=0.02); F1 *MeMt* hybrids from Loch Fyne (LFY_14=0.43; LFY_71=0.28; LFY_72=0.31); and F1 *MeMt* hybrids from Loch Etive (LET_33=0.22; LET_40=0.12). Instead, both F1 *MeMg* hybrids from St Andrews were assigned with highest probability to the PureMe genotype group (STA_16 =0.55 and STA_29 =0.89). All F1 *MeMt* hybrids from Loch Fyne and Loch Etive were assigned with highest probability to the 1FX genotype group (LFY_14=0.57; LFY_71=0.71; LFY_72=0.68; LET_33=0.77; LET_40=0.86).

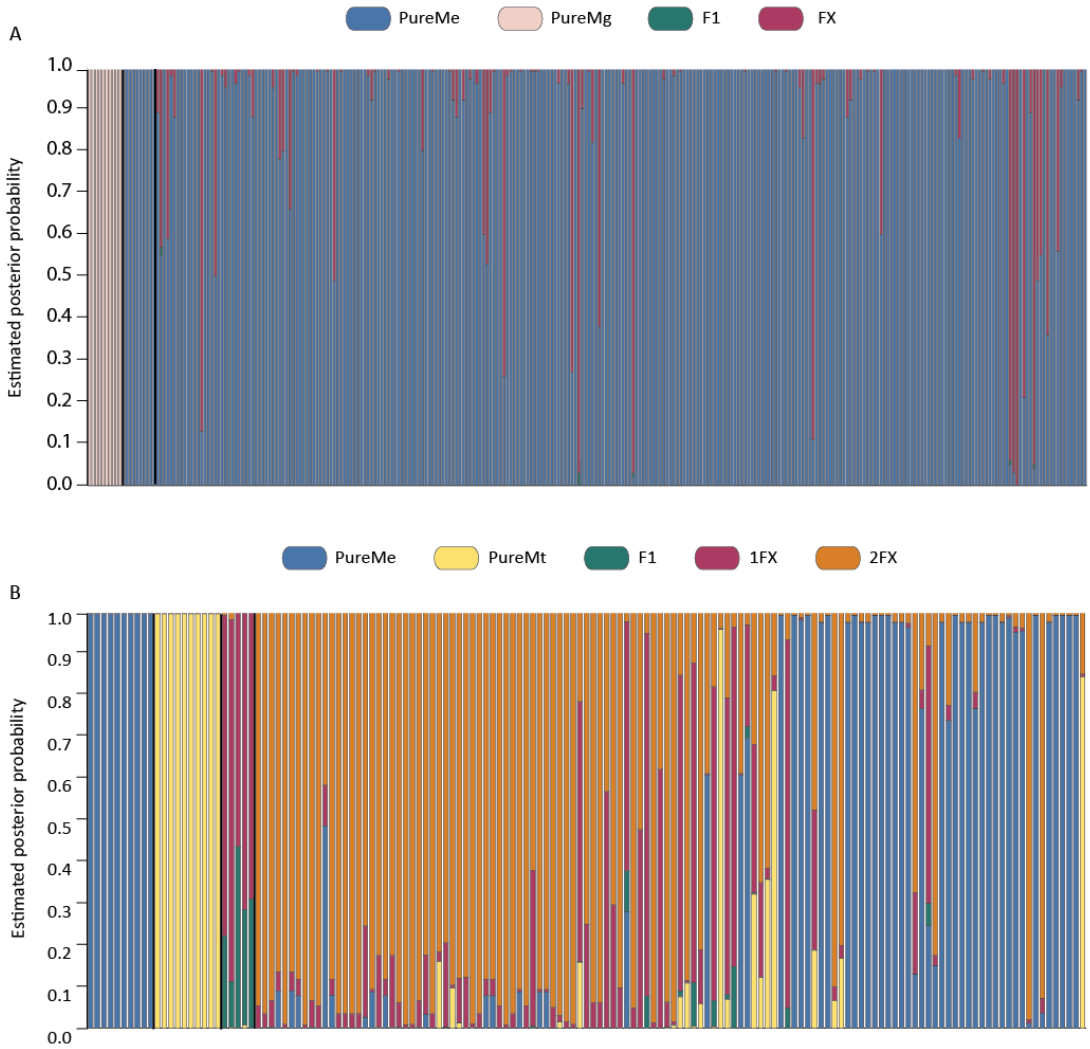

FIGURE 4.8 – NEWHYBRIDS classifications for (A) *M. edulis, M. galloprovincialis* and their hybrids, and (B) *M. edulis, M. trossulus* and their hybrids, with posterior probabilities for each given genotype frequency class. "PureMe" = *M. edulis;* "PureMg" = *M. galloprovincialis;* "PureMt" = *M. trossulus;* F1 = First generation hybrid; FX = introgressed hybrid (numbered 1-2 to denote different genetic combinations). Each vertical line represents the genetic composition of an individual broken into coloured segments, with lengths proportional to each genotype class. Only 10 pure individuals are shown in each graph for the sake of clarity: all pure individuals had the same assignment. Groups of Pure, F1 and FX individuals (genotyped with SNP assays) are separated by black lines.

4.3.4. M. galloprovincialis *and* M. trossulus *introgression in Scotland*

4.3.4.1. Relative proportion of diagnostic alleles

Across all 22 named sites, the relative proportion of *M. edulis* alleles was highest (0.874), followed by *M. galloprovincialis* (0.121) and *M. trossulus* (0.086), indicating widespread hybridisation and introgression of *M. galloprovincialis* and *M. trossulus* with native *M. edulis.* The relative proportion of introgression from *M.*

galloprovincialis and *M. trossulus* [R(AA)] at all sampling sites was calculated and compared to the q values generated by STRUCTURE (TABLE 4.7), as a means of assessing the effectiveness of this calculation for estimating the proportion of introgression in a population sample. Given that NEWHYBRIDS was looking at

TABLE 4.7 – Comparison of two methods measuring the levels of *M. galloprovincialis* and *M. trossulus* introgression at each sampling site: the relative proportion of diagnostic alleles [R(AA)], an equation designed specifically for use in this study based on standard allele frequency calculations; and the q values assigned by STRUCTURE which denote membership to the *M. trossulus* and *M. galloprovincialis* clusters (also detailed in TABLE 4.5)

Group	Site	R(AA) *Mg*	STRUCTURE *Mg*	R(AA) *Mt*	STRUCTURE *Mt*
A	Loch Roag	0.405	0.062	0.050	0.016
	Scapa Beach	0.400	0.220	0.000	0.005
	Northside	0.588	0.335	0.030	0.011
	Shetland BR	0.278	0.046	0.089	0.030
B	Loch Eireasort	0.298	0.067	0.022	0.013
	Flotta	0.311	0.089	0.056	0.042
	Shetland BX	0.194	0.047	0.009	0.005
	Loch Laxford	0.213	0.021	0.097	0.054
C	Kylesku	0.125	0.016	0.021	0.019
	Loch Ailort	0.098	0.028	0.036	0.019
D	Loch Linnhe	0.041	0.015	0.054	0.064
	Loch Fyne	0.004	0.005	0.085	0.079
E	Loch Etive	0.022	0.010	0.638	0.672
F	Loch Spelve	0.048	0.007	0.018	0.006
	Montrose	0.020	0.010	0.010	0.006
	St Andrews	0.041	0.021	0.029	0.020
	Loch Ryan	0.038	0.019	0.004	0.007
	Rascarrel Bay	0.005	0.004	0.004	0.004
	Dornoch Firth	0.013	0.005	0.020	0.013
	Ferryness	0.030	0.006	0.046	0.022
	Lunderston Bay	0.006	0.005	0.020	0.013
	Loch Long	0.008	0.004	0.042	0.024
	Site X	0.008	0.009	0.068	0.019

individual genotype classes and not the overall proportion of genetic admixture in a population sample, it was not used for comparison here. The relative proportion of diagnostic *M. galloprovincialis* alleles was higher than STRUCTURE q values at every site except Loch Fyne and Site X. Overall, R(AA) values and STRUCTURE q values varied little and were in the same order of magnitude; most variation was observable

at sites in Group A, Group B and Group C, particularly Loch Roag [R(AA)=0.405; q=0.062]; Shetland BR [R(AA)=0.278; q=0.046]; Loch Eireasort [R(AA)=0.298; q=0.067]; Flotta [R(AA)=0.311; q=0.089]; Shetland BX [R(AA)=0.194; q=0.047]; Loch Laxford [R(AA)=0.213; q=0.021]; and Kylesku [R(AA)=0.125; q=0.016]. Nevertheless, the R(AA) values remained consistent with STRUCTURE q values in showing a general decreasing trend of *M. galloprovincialis* introgression from Groups A-F. The relative proportion of diagnostic *M. trossulus* alleles was slightly higher than STRUCTURE q values at the majority of sites, except for Scapa Beach, Loch Linnhe, Loch Etive and Loch Ryan which were slightly lower. Overall, however, R(AA) values were not highly variable and remained within the same order of magnitude as STRUCTURE q values. Consistent with STRUCTURE results, Loch Etive had by far the highest proportion of *M. trossulus* alleles [R(AA)=0.638]. Next highest were Loch Laxford [R(AA)=0.097], Shetland BR [R(AA)=0.089] and Loch Fyne [R(AA)=0.085], compared to Loch Laxford (q=0.054), Loch Fyne (q=0.079) and Loch Linnhe (q=0.064) with STRUCTURE. Loch Etive, With both STRUCTURE (q=0.005) and the relative proportion of diagnostic alleles calculation [R(AA)=0.000], Scapa Beach had the lowest proportion of admixture from *M. trossulus.*

Sampling sites were divided into rope grown aquaculture sites (n=12) or shoreline [wild and bottom grown aquaculture (n=10)] sites. Of the 991 individuals genotyped, 592 were collected from ropes and 399 were collected from the shoreline. Higher average proportions of *M. galloprovincialis* and *M. trossulus* alleles were observable at rope grown aquaculture sites compared to the shoreline. The average proportion of *M. trossulus* alleles on ropes [R(AA)=0.130] was over six times higher than on the shoreline [R(AA)=0.021]. There was less variation in the average proportion of *M. galloprovincialis* alleles on ropes [R(AA)=0.170] and the shoreline [R(AA)=0.115].

4.3.4.2. Shell fragility and M. trossulus

To assess the possible relationship between shell fragility and *M. trossulus* introgression, a total of 39 mussels (20 strong shelled and 19 fragile shelled) from Site X were genotyped (FIGURE 4.9). Amongst 20 strong shelled mussels, *Me* was most abundant (60%), followed by *MeMg* (25%) and HXE (15%). Amongst 19

fragile shelled mussels, pure *Me* and *MeMt* were the two most abundant genotype classes (36.8%), followed by *MeMg* and HXE (10.5% each) and *MeMgMt* (5.4%) (FIGURE 4.9A).

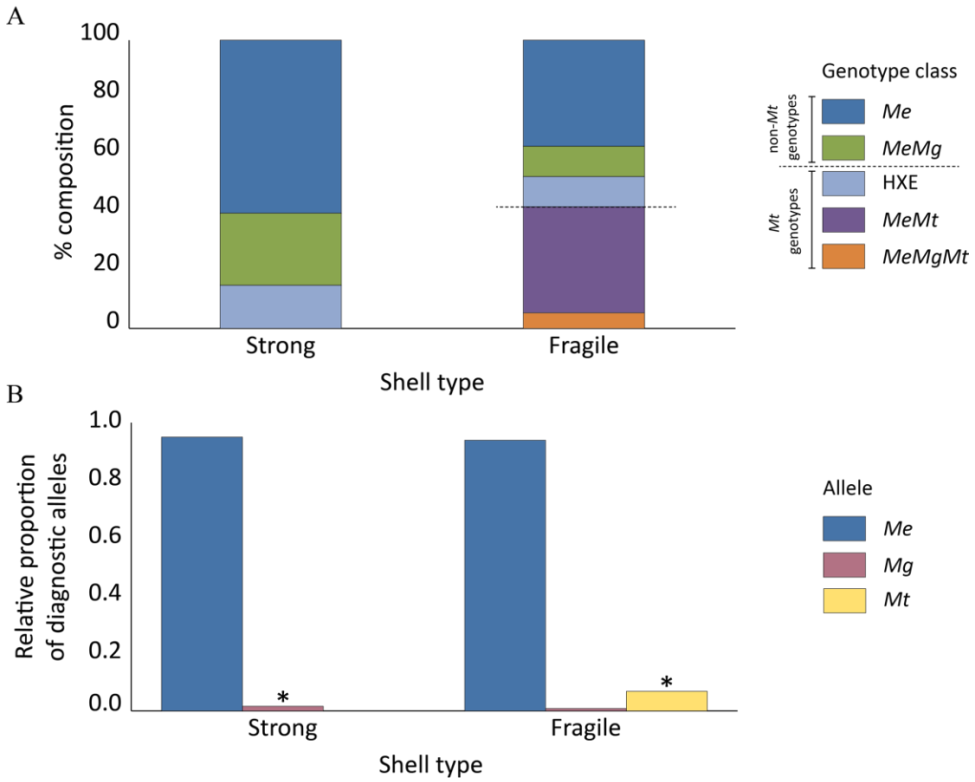

FIGURE 4.9 – Bar graphs showing (A) species composition ["*M. trossulus* genotype classes" comprise the lower part of each bar and "non-*M. trossulus* genotype classes" comprise the upper part, separated by a dotted line], **and (B) The relative proportions of *M. edulis, M. galloprovincialis* and *M. trossulus* alleles amongst strong and fragile shelled mussels from Site X in Scotland**. Alleles with a statistically significant effect (p = 0.002) on shell strength are marked with *.

To estimate levels of introgression in "strong" and "fragile" shelled mussels, the relative proportion of diagnostic alleles [R(AA), for *M. edulis*, *M. galloprovincialis* and *M. trossulus*] was calculated. In strong shelled mussels, the relative proportions of diagnostic alleles were *M. edulis* [R(AA)=0.950], *M. galloprovincialis* [R(AA)=0.016], and *M. trossulus* [R(AA)=0.00]. In fragile shelled mussels, the relative proportions of alleles were *M. edulis* [R(AA)=0.939], *M. galloprovincialis* [R(AA)=0.008], and *M. trossulus* [R(AA)=0.068] (FIGURE 4.9B). Two-way ANOVA analysis was carried out using R (version 3.1.0) (R Core Team, 2014), to determine whether a relationship between the type of species diagnostic allele and shell strength was statistically likely. A significant relationship between allele type and shell strength (p=0.002) was detected. However, these results should be regarded

with caution because of the limited data available for analysis. Additionally, no data on shell fragility was collected at other sites so it was not possible to establish any possible link between the proportion of *M. trossulus* alleles and shell fragility outside of Site X.

4.4. DISCUSSION

Historically, the presence of *M. trossulus* and its hybrids has been detected with single (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a) and multilocus (Zbawicka *et al.*, 2012) genotyping in Loch Etive, and *M. trossulus* has also been acknowledged in Loch Fyne and Loch Eil (Gubbins *et al*., 2012). There is, however, no evidence of multilocus genotyping having taken place in Scotland outside of Loch Etive, and to date single locus genotyping with Me15/16 has been the only tool utilised for wide scale species composition surveys (Dias *et al.,* 2009c; Dias *et al.,* 2011a). Multilocus genotyping in the present study was carried out with a suite of 12 SNP assays which, previously, had only been applied on a small scale as part of this PhD thesis (see CHAPTER 3). It was possible to successfully genotype all 1030 sampled individuals at all 12 loci, thereby verifying the diagnostic properties of the SNP assays on a larger scale and emphasising their robustness for future genotyping studies. Multilocus genotyping identified the presence of pure *M. trossulus* in Loch Etive (consistent with multilocus genotyping by Zbawicka *et al.,* 2012), but did not identify pure *M. trossulus* at any other site. Introgressed hybrids of *M. trossulus* were, however, widespread, generally in higher proportions at rope grown aquaculture sites and lower proportions at wild or bottom grown aquaculture sites. It is not possible to compare these results directly with previous species composition surveys (Dias *et al.,* 2009c; Dias *et al.,* 2011a) because, excluding Loch Etive and some sites unused in this study, specific sampling sites were not named. Nevertheless, it is clear that single locus genotyping has underestimated the overall extent of hybridisation and the distribution of *M. trossulus* alleles in Scotland: this could be highly significant for the shellfish industry when considering, for instance, broodstock sourcing and spat importation, and raises questions about the effectiveness of single locus genotyping (with Me15/16) as a universal management tool applicable across different environments (Araneda *et al.,* 2016).

Widespread distribution of introgressed genotypes, but an apparent restriction of pure *M. trossulus* to Loch Etive and absence of *M. galloprovincialis*, does call into question how *M. trossulus* and *M. galloprovincialis* alleles came to be so widespread around Scotland. Scottish *M. trossulus* (in Loch Etive) is believed to be a non-native invader from the Pacific which represents a relict population (Beaumont *et al.*, 2008, Zbawicka *et al.*, 2010). *M. galloprovincialis* and its hybrids are thought to have moved into Scottish waters from Cornwall and Devon [where they were first identified in the UK by Ahmad and Beardmore (1976) and Skibinski *et al* (1978)], with increasing surface temperatures of the North Atlantic favouring a natural northward migration (Jones *et al.*, 2001; Gosling *et al.*, 2008). Extensive hybridisation and introgression with a native species can greatly decrease the size of a relict population, or reduce the number of immigrants (Pryor, 1951; Lopez *et al.*, 2000): perhaps, introgression of Pacific *M. trossulus* and Mediterranean *M. galloprovincialis* with native *M. edulis* (Varvio *et al.*, 1988; Anderson *et al.*, 2002) has, over time, reduced the proportion of pure *M. trossulus* and *M. galloprovincialis* to such small numbers that nowadays both are very rare in Scottish population samples, but their hybrids remain widespread. Artificial spat movement between locations [i.e., translocation of mussel ropes between farms (Hickman, 1992; Goulletquer and Le Moine, 2002) and discharge of ballast water from shipping activity (Carlton and Geller, 1993; Tamelander *et al.*, 2010)] could have aided the spread of hybrids outside of their natural dispersal range and encouraged introgression (Daguin *et al.*, 2001; Bierne *et al*., 2003). The natural dispersal range of *M. edulis* is typically estimated at around 30 km per generation (Gilg and HIlbish, 2003), far less than the distances between the Isle of Lewis and Orkney (c. 200 km) and Shetland and Orkney (c. 140 km) which would suggest anthropogenic factors are facilitating similar genetic compositions at geographically distant sites. Certainly, the proportions of *M. galloprovincialis* and, particularly, *M. trossulus* hybrids at rope grown aquaculture sites were higher than at wild or bottom grown aquaculture sites, and there is evidence that aquaculture ropes were moved from Loch Etive to surrounding locations before the genetic link between *M. trossulus* and shell fragility was suggested (Gubbins *et al.*, 2012). Rope grown aquaculture sites would offer fragile mussels greater protection from predation (as demonstrated in studies of

Scottish mussels by Beaumont *et al.*, 2008; Dias *et al.*, 2009a; Dias *et al.*, 2011a; Gubbins *et al.*, 2012), thereby creating an opportunity unavailable for exploitation on the exposed shoreline (Riginos and Cunningham, 2005; Jensen and Patursson, 2011; Dias *et al.*, 2011b). However, not all farmers import spat as a means of reducing any potential risks associated with genetic contamination from undesirable species (Douglas Wilson, Personal Communication, June 2016). Spat translocation would therefore not have an equal impact across rope grown aquaculture sites, nor explain the presence of *M. trossulus* and *M. galloprovincialis* alleles at non-commercial sites, as observed here. Perhaps long term natural dispersal, influenced by hydrodynamic variation across coastal areas (Gilg and Hilbish, 2003), has allowed hybrids to reach non-commercial sites over multiple generations, albeit more slowly and in smaller proportions than artificially translocated spat. Species composition among adult mussels could also be influenced by variation in larval settlement and larval fitness (Bierne *et al.*, 2002; Bierne *et al.*, 2006). Perhaps, an improved fitness of hybrid *M. trossulus* and *M. galloprovincialis* larvae could explain the prevalence of hybrid genotypes but very low abundance (or seeming absence) of pure *M. trossulus* and *M. galloprovincialis.* Studies into the mortality rates of hybrid *Mytilus* larvae, in comparison to that of pure species, have been carried out *in vitro* by several authors (e.g., Bierne *et al.*, 2002; Miranda *et al.*, 2010; Toro *et al.*, 2012). In the present study, bivalve "D-larvae" were sorted from plankton samples that had been collected with two different mesh sizes (200 μM and 68 μM) from a site at Loch Ewe, where Marine Scotland Science collects temporal samples for use in monitoring species diversity and composition (Cook, 2013). DNA was extracted from pooled samples comprising at least 100 bivalve larvae (following the protocol outlined by Zhan *et al.*, 2008), and amplified with universal eukaryotic primers E528F/UI429R (Edgcomb *et al.*, 2002) and LCO1490/HCO2198 (Folmer *et al.*, 1994) to confirm the presence of DNA. However, these results yielded limited success and, within time constraints of the study, could not be pursued further. The application of similar techniques to future studies could nevertheless be beneficial in comparing the genetic composition of larvae and settled adults and in determining possible genetic effects on fitness.

The detection of F1 hybrids, albeit in very low numbers, suggests that ongoing hybridisation is possible at three of the sites surveyed (Loch Etive, Loch Fyne and St Andrews). "True" F1 hybrids (i.e., those with an equal genetic contribution from two pure parent species) are less likely to exist in populations with high levels of introgression (Fogelqvist *et al.,* 2015), so the low numbers of F1 hybrids identified here was not unexpected. F1 hybrids must have been present at some point if such an abundance of FX hybrids was observed (Jiggins and Mallet, 2000), but their extremely low proportion relative to the high proportion of FX hybrids is consistent with widespread introgression which would make hybridisation between pure species less likely. No pure *M. galloprovincialis* was detected at any of the Scottish sites sampled, consistent with previous studies (Beaumont *et al.,* 2008; Dias *et al.,* 2009c; Dias *et al.,* 2011a) that have identified little or no *M. galloprovincialis* in Scotland. The appearance of "true" F1 *MeMg* hybrids in St Andrews is, based on new and historical data, very unlikely, and may simply have arisen from genotyping with a small number of markers (*n*=7). However, as no other studies of species composition have been conducted specifically in St Andrews, the likelihood of such hybrids existing cannot be determined without a wider scale study of this area. The appearance of F1 *MeMt* hybrids in Group D (Loch Fyne) and Group E (Loch Etive), alongside introgressed *MeMt,* implies ongoing hybridisation and, perhaps, that this is an active mosaic hybrid zone (Gilg and Hilbish, 2003; Smietanka *et al.,* 2004), something that has not been suggested by previous studies of species composition in Scotland (Dias *et al.,* 2011a). Although not identified in the present study, pure *M. trossulus* has previously been detected in Loch Fyne (Gubbins *et al.,* 2012) so the appearance of "true" F1 *MeMt* hybrids is perhaps feasible. However, without further research confirming the presence of *M. trossulus* in Loch Fyne, the existence of an active hybrid zone between *M. edulis* and *M. trossulus* remains unconfirmed.

DAPC analysis (Jombart *et al.,* 2010) has been applied to multilocus genotyping studies of *Mytilus* spp. populations in the Mediterranean (Giantsis *et al.,* 2014) and in South America (Araneda *et al.,* 2016). Correspondence Analysis (CA) (Benzecri, 1992), a technique analogous to DAPC, has also been used to model multilocus SNP data collected from mixed species *Mytilus* populations (Zbawicka *et al.,* 2012; Wenne *et al.,* 2016) from Scotland (Loch Etive) and Greenland respectively. DAPC

of data from the present study displayed a trend similar to previous CA analyses: clear differentiation of pure species (*M. edulis*, *M. galloprovincialis* and *M. trossulus*) interspersed with admixed individuals, highlighting the usefulness of the 12 SNP markers for large scale genotyping. Although the three are considered as distinct species (e.g., Edwards and Skibinski, 1987; Koehn, 1991), various studies have demonstrated *M. edulis* and *M. galloprovincialis* to be less diverged than either species is from *M. trossulus* (e.g., Gérard *et al.*, 2008; Zbawicka *et al.*, 2012; Astorga *et al.*, 2015; Fraisse *et al.*, 2015; Mathiesen *et al.*, 2016). This trend was also observable in our DAPC data and CA analysis by Zbawicka *et al* (2012) and Wenne *et al* (2016). This may have explained the greater degree in overlap between *MeMg* hybrids when compared to *MeMt* hybrids, although it is possible that the larger sample size of *MeMg* (n=274) compared to *MeMt* (n=128) simply increased the likelihood of individuals having similar genotypes. Wenne *et al* (2016) found three clusters of pure species and a distinct hybrid cluster that did not overlap with the main groups, but this was less obvious in our data which showed more overlap between hybrids and pure species. The histories of natural populations are complex and tend not to follow simple patterns of genetic exchange, which gives rise to a wide array of admixed genotypes (Harrison and Larson, 2014; Patel *et al.*, 2014) and could explain the patterns of overlap observable in our data. Alternatively, the greater degree of overlap may have arisen from the smaller number of markers used (n=12) compared to Wenne *et al* (2016) (n=54), which would perhaps offer less clear differentiation between introgressed genotypes.

STRUCTURE and NEWHYBRIDS are two Bayesian clustering methods that have slightly different approaches to modelling population genetic data: STRUCTURE focuses on detecting admixture at the population level (Pritchard *et al.*, 2000; Vähä and Primmer, 2006), whereas NEWHYBRIDS focuses on the individual level to identify specific types of hybrid (Anderson and Thompson, 2008; Marie *et al.*, 2011). Combined use of STRUCTURE and NEWHYBRIDS can provide a more detailed assessment of genetic structure, and potentially allows more reliable assignment of individuals to specific genetic clusters than use of a single program does (Frantz *et al.*, 2009). However, the choice of program used ultimately depends on the aims of the study (Marie *et al.*, 2011) and how effectively empirical data fits into chosen

model parameters (Burgarella *et al.*, 2009). Although both approaches were applied to data in the present study STRUCTURE produced more consistent results than NEWHYBRIDS, possibly due to our data more suitably fitting the STRUCTURE model. Anderson (2002) and Anderson and Thompson (2008) state specifically that NEWHYBRIDS is suitable for use in a scenario where only two diploid species are hybridising, and where recent hybridisation is expected to have taken place. Conversely, the STRUCTURE model outlined by Pritchard *et al* (2000) makes no such stipulations about the number of different species, and is appropriate for use in a scenario with multiple generations of admixture (Anderson, 2002). Our data recognised the presence of three discrete *Mytilus* species and their hybrids and a far greater proportion of introgressed (FX) hybrids than F1 hybrids, which indicated widespread, long standing genetic admixture and less recent hybridisation between two pure species. Additionally, *MeMgMt* hybrids with genetic material from three species were detected but, due to their unsuitability for use with the NEWHYBRIDS model, had to be excluded from this analysis. On the other hand, the model employed by STRUCTURE allowed inclusion of *MeMgMt* hybrids because it was possible for individuals to have genetic input from more than two inferred clusters. STRUCTURE also allowed inclusion of hybrids with an indeterminate genotype (HXE and HXT), whereas these could not be included in the NEWHYBRIDS model. The main aim of the study was to assess levels of genetic admixture in Scottish population samples, and furthermore to identify which types of hybrid were present. Analysis of the entire dataset with STRUCTURE, rather than selected parts of the dataset with NEWHYBRIDS, offered a more complete picture of genetic diversity and enabled classification of population samples into discrete groups, information that could be of great relevance to the Scottish shellfish industry.

STRUCTURE has been considered less accurate than NEWHYBRIDS in assigning "hybrid type" (e.g., Vähä and Primmer, 2006; Burgarella *et al.*, 2009; Marie *et al.*, 2011); our results, however, demonstrated the opposite. STRUCTURE correctly assigned six of seven F1 hybrids (i.e., those with an equal genetic contribution from parental species), whereas NEWHYBRIDS failed to assign any F1 hybrids with high probability to the specified F1 category. In datasets with low proportions of (F1) hybrids, STRUCTURE can be more effective in quantifying admixture (Marie *et al.*,

2011) while NEWHYBRIDS can be less efficient in reliable hybrid identification and classification, particularly when low numbers of markers are used (Vähä and Primmer, 2006). Additionally, genetic variation in data from real population samples can produce inconsistent results with different software if it differs too much from "model" data (Sanz *et al.*, 2009; Marie *et al.*, 2011). This could explain why NEWHYBRIDS could not adequately assign hybrids to their relevant categories (particularly in the *MeMg* dataset) and also why it could not be used with data from *MgMt* crosses.

The abundance of FX hybrids points towards widespread introgression amongst Scottish mussel populations. Both STRUCTURE and NEWHYBRIDS correctly assigned 100% of pure individuals. STRUCTURE recognised admixture in all hybrid individuals; although its capacity to distinguish between F1 and FX hybrids was limited (as acknowledged by, for instance, Vähä and Primmer, 2006; Corander *et al.*, 2008; Marie *et al.*, 2011), it remained a useful tool in investigating overall admixture levels. Measuring admixture with STRUCTURE has been applied to multiple studies of *Mytilus* spp. mussel populations: e.g., Zbawicka *et al.*, 2012; Larrain *et al.*, 2015; Katolikova *et al.*, 2016; Larsson *et al.*, 2016; and Mathiesen *et al.*, 2016. Our STRUCTURE results (K=4) bore striking resemblance to the STRUCTURE results (K=4) in Mathiesen *et al* (2016): both studies allow clear discrimination between *M. edulis*, *M. galloprovincialis*, and *M. trossulus*, and also have a cluster corresponding to hybrids of *M. edulis*. Rather than hybrid *M. edulis* individuals having input from *M. edulis* and *M. galloprovincialis* clusters, they have been assigned their own, unique cluster; thus, it is possible that multiple generations of backcrossing has altered the structure of these populations so that, genetically, they are quite distinct from either of their parental species (Gross and Rieseberg, 2005). In our data, in the case of rope grown aquaculture sites with high assignment in the *MeMg* cluster, anthropogenic spat importation could have affected the genetic composition (Hickman, 1992; Goulletquer and Le Moine, 2002); in the case of shoreline sites with high assignments in the *MeMg* cluster, this could have been an effect of farming activities affecting natural species composition, a common occurrence in aquaculture systems (Naylor *et al.*, 2001). However, this cannot be determined without a more detailed study of the areas in question.

NEWHYBRIDS recognised admixture but overestimated the proportion of pure individuals. It may be difficult for NEWHYBRIDS to distinguish genotype frequency classes with low genetic differentiation – as, for instance, in cryptic hybrids (Anderson and Thompson, 2002). This was observable in our data: if too few genotype frequency classes were used in simulations pure individuals were wrongly classified as hybrids, whereas if too many genotype frequency classes were used hybrids were wrongly classified as pure individuals. The array of genetic variation in SNP data made reliable category assignment difficult, and although the best models were chosen for discussion here, overall they were not an ideal fit for the datasets in question. A study of introgression between *M. galloprovincialis* and *M. trossulus* in central California (Saarman and Pogson, 2015) applied NEWHYBRIDS to genetic data from 1337 SNPs, which correctly assigned 100% of F1 and backcross hybrids into the relevant genotype frequency class. Perhaps in our data, the poor distinction of backcrosses from parental individuals was due to the small number of markers used (Vähä and Primmer, 2006) which subsequently increased the chance of error in correct group assignment (O'Hara *et al.*, 2008). However, given such a huge difference between genotyping effort in our study and the approach used by Saarman and Pogson (2015), this is perhaps an unreliable comparison and further studies should be conducted before more reliable conclusions can be drawn about the effectiveness of NEWHYBRIDS for analysis of *Mytilus* spp. datasets. Nevertheless, because the STRUCTURE model could include all types of hybrid and produced more consistent results than NEWHYBRIDS it was considered, for the purposes of this study, a better approach for investigating admixture levels in population samples, and analysis with NEWHYBRIDS was not considered further.

The equation devised to calculate the relative proportion of diagnostic alleles [R(AA)] generated results consistent with trends in levels of introgression demonstrated by STRUCTURE, despite most R(AA) values being higher. This equation could be a useful tool when estimating the overall levels of admixture in a population sample, either as a supplement to STRUCTURE data or by itself to reduce the time invested in simulations: however, application of this equation outside of this study would be required to verify its usefulness in data analysis. Although both STRUCTURE and R(AA) equation demonstrated widespread introgression of *M.*

galloprovincialis and *M. trossulus* with the native *M. edulis,* the proportions of *M. trossulus* and *M. galloprovincialis* alleles were lower than those of *M. edulis* (excluding *M. trossulus* in Loch Etive). The historical low abundance of pure *M. galloprovincialis* (Beaumont *et al.*, 2008; Dias *et al.*, 2009c; Dias *et al.*, 2011a) and its cultivation in other countries does suggest that *M. galloprovincialis* is unlikely to impact significantly on a commercial scale in Scotland (Dias *et al.*, 2009c). The possible commercial impact of *M. trossulus* and its hybrids remains unevaluated. However, due to recent increases in production values throughout Scotland; prolific mussel production in the Shetland Isles (Munro and Wallace, 2015) despite an abundance of *M. trossulus* hybrids; and no other documented cases of shell fragility (outside of Site X), it is probably unlikely that current levels of *M. trossulus* introgression pose an immediate threat to Scottish mussel aquaculture. The effects of *M. trossulus* on shell characteristics is unknown. A tentative link between the *M. trossulus* allele and shell fragility was demonstrated at Site X, but this data was not reliable for comparison or application elsewhere. It is likely that shell characteristics are influenced by a combination of genetic and environmental factors. Other alleles could interact with *M. trossulus* alleles to mitigate their effects on phenotype. For example, Beaumont *et al* (2008) tested the strength of *Mytilus* species shells and found that those with higher proportions of *M. edulis* and *M. galloprovincialis* alleles had more robust shells than individuals with *M. trossulus* alleles. In our data, Shetland BR (and Loch Laxford) had high proportions of the *MeMg* genotype class, which may have counteracted the effects of *M. trossulus* alleles on shell fragility. Additionally, environmental conditions could affect gene expression (Jaenisch and Bird, 2003) or could themselves directly influence shell characteristics. Perhaps the environmental conditions of Site X and Loch Etive favour the persistence and expression of a *M. trossulus* allele that confers shell fragility more than the environmental conditions in Shetland BR, Loch Laxford or Loch Fyne do. Further study is needed to establish the exact cause of shell fragility in Scottish mussels, but care should still be taken to mitigate the spread of *M. trossulus* alleles until this has been more fully evaluated. Perhaps farming at wild sites or sites with low existing levels of *M. trossulus* introgression could be an option, or these areas could be used to source less introgressed broodstock. Additional studies will be required before any

management decisions can be made: nevertheless, multilocus SNP genotyping will undoubtedly be a very useful tool in more completely understanding the genetic structure of Scottish mussel populations, and holds great potential for improving sustainability and profitability in the Scottish shellfish industry.

4.5. CONCLUSIONS AND SUMMARY

1. Multilocus genotyping with a suite of 12 SNP markers has revealed widespread admixture in Scotland that previous, single locus genotyping studies were unable to identify, marking an important step forward in assessing genetic structure of field samples of *Mytilus* spp. mussels;

2. DAPC, STRUCTURE, the R(AA) calculation and, to a lesser extent, NEWHYBRIDS, demonstrated genetic diversity within and between population samples; STRUCTURE was deemed a more appropriate tool than NEWHYBRIDS for measuring overall proportions of admixture;

3. Introgressed (FX) hybrids appear far more abundant than F1 hybrids, suggesting long-standing admixture as opposed to more recent hybridisation events;

4. Generally, there was more hybridisation at rope grown aquaculture sites than at shoreline sites, suggesting anthropogenic influences on spat movement and possible areas for selection of less introgressed broodstock;

5. A tentative link between shell fragility and *M. trossulus* introgression was demonstrated in a single farmed population (Site X), but this has yet to be confirmed or investigated elsewhere

Acknowledgements

A special thanks to Douglas Wilson for his wealth of information on the world of mussel aquaculture, and to Andrew Mayes (Marine Scotland Science) for his advice and guidance with site selection. Thanks to all farmers and fellow researchers who

helped with sample acquisition, collection and processing. Thanks to Rebecca McIntosh (Marine Scotland Science) for additional DNA extraction and Me15/16 genotyping. The work was funded by Marine Alliance for Science and Technology for Scotland (MASTS) and Marine Scotland Science.

Chapter 5
Temporal distribution of *Mytilus* spp. mussels in a Scottish Loch

5.1. INTRODUCTION

Loch Etive is a historically important site for mussel aquaculture, situated 6 km north of Oban on the west coast of Scotland. The sea loch measures over 30 km in length and extends over two large basins shaped by glaciation: the lower basin has a maximum depth of approximately 60 m, and the upper basin has a maximum depth of approximately 150 m (Gage, 1972; Howe *et al.,* 2002). The catchment area measures 1400 km^2 (the largest of any Scottish sea loch) (Edwards and Edelsten, 1997), which causes extensive freshwater and sediment input into the loch after heavy rainfall (Howe *et al.*, 2002). Loch Etive is connected to the sea by a sill 300 m wide, 4 km long and 10 m deep. Its tidal range is small (2 m) compared to the outside sea (4 m), resulting in slow exchange of surface water and stagnation of deeper water in the loch (Edwards and Edelsten, 1997). Tidal flow influences multiple abiotic factors, such as temperature, salinity, pH and oxygen concentration (Gage, 1972, Austin and Inall, 2002), which subsequently affects species composition in the sea loch.

Loch Etive mussel farming was at its peak of productivity in 2002, employing more than 30 people and contributing to at least half of Scottish production by farming 1000 tonnes of *Mytilus edulis* (blue mussel). However, by 2004 it had become noticeable that the amount of marketable product was declining due to the presence of a thin-shelled mussel with poor quality meat (Gubbins et al., 2012), which over subsequent years continued to cause extensive problems. Production dropped by 50% between 2008 and 2009 (Dias *et al.,* 2011a/b). By 2011, farming was rendered inviable: thousands of mussel ropes were stripped and their contents deposited on the seabed, with the intention that natural predators would eradicate the more fragile mussels and, over time, enable mussel farming in Loch Etive to continue (Gubbins et al., 2012). Several genetic studies of mussels in Loch Etive have suggested that these fragile specimens largely comprise *Mytilus trossulus* and,

often, its hybrids with the native *M. edulis* (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a). *M. trossulus* is an unmarketable species with a meat yield some 75% lower than the minimum specification (20% by weight) set in Scotland (Gubbins et al., 2012). *M. trossulus* is a species native to the Pacific Ocean, which colonised parts of the north Atlantic, including the Canadian Maritimes and the Baltic Sea, around 3.5 million years ago after the Bering Strait opened (Riginos and Cunnigham, 2005). When *M. trossulus* was first identified in Scotland, it was thought to be a recent invader from the Baltic Sea. However, mtDNA studies revealed that *M. trossulus* in Loch Etive has a Pacific origin, which instead suggests that it is part of a long-established, relict population (Zbawicka *et al.,* 2010). The surface waters of Loch Etive can have a very low salinity after heavy rainfall, an environment in which *M. trossulus* thrives in comparison to *M. edulis* and *M. galloprovincialis* (Qiu *et al.,* 2002). Perhaps this environment aided the initial settlement of *M. trossulus* in Loch Etive and contributed to its persistence over the years. As of 2013, *M. trossulus* is recognised as a commercially damaging species under Scottish law [The Aquaculture and Fisheries (Scotland) Act (2013)] because of the potential negative impact it could have on meat yields and shell strength among Scottish populations.

Environmental conditions in Loch Etive could have a direct influence on shell characteristics, contributing to the abundance of fragile individuals present as well as (or instead of) genetic factors. Freshwater input from heavy rainfall lowers salinity and also affects pH. Salinity and pH both impact upon the carbonate chemistry of the water column and, subsequently, affect the growth, physiology and survival of calcareous, shell-forming organisms like *Mytilus* spp. (Doney *et al.,* 2009; Byrne, 2011). In addition to genetic influences on shell fragility in Loch Etive, current research has found some evidence to suggest *M. trossulus* could be linked to shell fragility at a site of current commercial importance on the west coast of Scotland (see SECTION 4.3.3.). However, despite the widespread occurrence of introgressed *M. trossulus* around the Scottish coast, there are no other reports of problems with shell fragility and this tentative link cannot be confirmed. This does suggest that environmental factors may be more strongly affecting shell characteristics than genetics are, but the impact of genetics cannot be ruled out and remains important in

assessing the suitability of Loch Etive – and other potential sites – for mussel farming.

Fallowing of Loch Etive was performed in attempt to eradicate *M. trossulus* from the area, stripping ropes to deposit on the seabed with the intention that fragile *M. trossulus* would be predated. Fallowing first took place in August 2010 and then in July 2011, when all ropes were stripped from the loch (Gubbins et al., 2012); since 2011, there has been no mussel aquaculture. Farmers have agreed not to resume production until the outcome of fallowing has been ascertained, which is determinable by genotyping of individuals that have settled in the area post-fallowing. A single site in Loch Etive (at Achnacloich) has retained ropes for the settlement of natural spat which is used for monitoring purposes and is genotyped to identify *M. trossulus* and its hybrids. Although the future of mussel farming in Loch Etive remains uncertain at present, ongoing temporal monitoring of species composition will be useful in assessing the effectiveness of fallowing as a method for controlling *M. trossulus* and its hybrids, which will be important for long term management of the area.

Previous studies of Loch Etive have used a single genetic marker [Me15/16 (Inoue *et al.,* 1995)] to check for the presence of *M. trossulus* and its hybrids. This study aims to further improve such monitoring by detecting introgression with multilocus SNP genotyping. Understanding the extent of hybridisation and introgression in Loch Etive will assist with management decisions and allow the future commercial potential of this site to be established. The aims of this study are as follows:

1. To more accurately ascertain the *Mytilus* species composition in Loch Etive and any potential changes over a short time period after fallowing (19 months);

2. To monitor the levels of *M. trossulus* in Loch Etive and assess whether hybridisation and introgression are ongoing;

3. To determine whether Loch Etive is a suitable site for commercial mussel farming.

5.2. METHODS

5.2.1. Sample collection and genotyping

Juvenile mussels were collected from monitoring ropes at Achnacloich, Loch Etive (FIGURE 5.1), and sent to Marine Scotland Science. The dates of collection were approximately 15 months after settlement: estimated settlement dates were January 2012, July 2012 and August 2013. These dates are subsequently used throughout the chapter to refer to each temporal sample. Sample sizes were as follows: January 2012 (n=80); July 2012 (n=80); and August 2013 (n = 150).

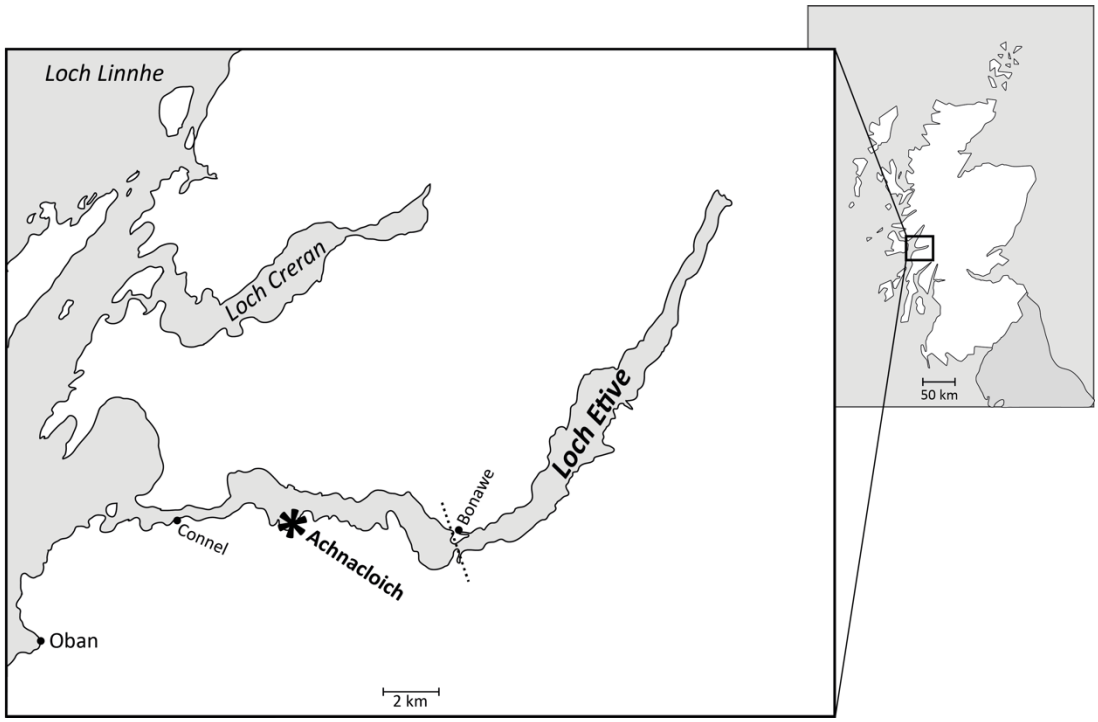

FIGURE 5.1 – Map showing the location of Loch Etive in Scotland. The sampling site at Achnacloich is labelled and marked with a "*" symbol. The lower basin starts at Connel and the upper basin starts at Bonawe. The division between the lower and upper basins is marked with a dotted line

No record of sampling depth was provided. DNA was extracted from juvenile mussels using the automated system described in SECTION 2.2.1. Each individual was genotyped by Marine Scotland Science at a single locus with Me15/16 (see

SECTION 2.3.1. for PCR conditions); DNA samples were then transported to The University of Stirling for genotyping with 12 diagnostic SNP markers under the KASP conditions in detailed in SECTION 3.3.4. Scoring of SNP assays was carried out as detailed in SECTION 3.3.4.

5.2.2. Inferring population structure with STRUCTURE

Population structure was determined using STRUCTURE (version 2.3) (Pritchard *et al.,* 2000), which provided a general overview of levels of introgression per temporal sample. Most parameters were set to their default values as advised in the STRUCTURE 2.0 user manual (Pritchard and Wen, 2003). Specifically, the admixture model with correlated allele frequencies between populations was chosen: this configuration is recommended by Falush *et al* (2003) as the most suitable for resolving cryptic population structure. The lengths of MCMC and burn-in were varied from 100 to 100,000. A value of 10,000 each proved to be sufficient; longer values did not obviously alter the results. The range of possible *K* values tested ranged from 1 to the total number of temporal samples (3), assuming each temporal sample could have a unique genetic composition. The optimal *K* was determined from 100 iterations of each *K* value, according to the method outlined by Evanno *et al* (2005) and tested using CLUMPAK software (Kopelman *et al.,* 2015). The optimal *K* by Evanno's method was 2 ($\Delta K = 37.974$).

5.2.3. PCA and DAPC analysis

Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC) were carried out using the *adegenet* package (version 1.4-1; Jombart, 2008) for R (version 3.1.0) (R Core Team, 2014). PCA creates simplified models of the total variation within the dataset (Jackson, 1991), and DAPC identifies clusters of genetically related individuals (Jombart *et al.,* 2010) within the most statistically likely PCA model, determined using the Bayesian Information Criterion (Schwarz, 1978) (see APPENDIX 4C for PCA and DAPC script). DAPC was used to group individuals into clusters of like individuals, showing the possible relationships between pure and admixed genotypes.

5.2.4. M. trossulus *genetic contribution*

In order to infer the genetic contribution from *M. trossulus* in each temporal sample, an equation was designed based on the equations for calculating genotype and allele frequencies (EQUATION 3 in SECTION 2.6.3.1, detailed below; see SECTION 2.6.3.1 for details of genotype and allele frequency calculations).

$$\text{R(A) per individual} = \frac{[2(n\text{AA}) + n\text{AB}]}{2M}$$

$$\text{R(A) per population} = \frac{\sum^{T} \left(\frac{[2(n\text{AA}) + n\text{AB}]}{2M} \right)}{T} \quad (3)$$

This estimated only the proportion of diagnostic *M. trossulus* alleles relative to the number of diagnostic markers (M=5) used for genotyping. In the equation, R(A) is the relative proportion of diagnostic alleles; n represents the number of individuals; and T represents the total number of individuals in a population sample. "A" and "B" refer to two alleles at a biallelic locus: "AA" is homozygous diagnostic and "AB" is heterozygous; homozygous non-diagnostic "BB" individuals were excluded. Values for n(AA) and M are multiplied by 2 to represent diploid organisms with two alleles at each locus.

5.3. RESULTS

5.3.1. Single and multilocus genotyping

5.3.1.1. Me15/16 PCR

Genotyping with Me15/16 identified a total of six different genotypes, which were as follows: *M. edulis* (*Me*), *M. galloprovincialis* (*Mg*) and *M. trossulus* (*Mt*), plus hybrids between *M. edulis* and *M. galloprovincialis* (*MeMg*), *M. edulis* and *M. trossulus* (*MeMt*), and *M. galloprovincialis* and *M. trossulus* (*MgMt*) (FIGURE 5.2A). All six genotypes were identified in the January 2012 and August 2013 samples, but *MeMt* and *MgMt* were absent from July 2012. *Mt* was the most abundant genotype in January 2012 (36.3%) and August 2013 (46%), and *Me* was the most abundant genotype in July 2012 (78.8%). *Mg* was least abundant in January

2012 (1.25%) and August 2013 (0.67%), and *Mt* was the least abundant in July 2012 (1.25%).

5.3.1.2. SNP assays

SNP assays detected overall eight genotype classes overall (FIGURE 5.2B). *Me* and *Mt* referred to pure *M. edulis* and *M. trossulus* respectively. Hybrids

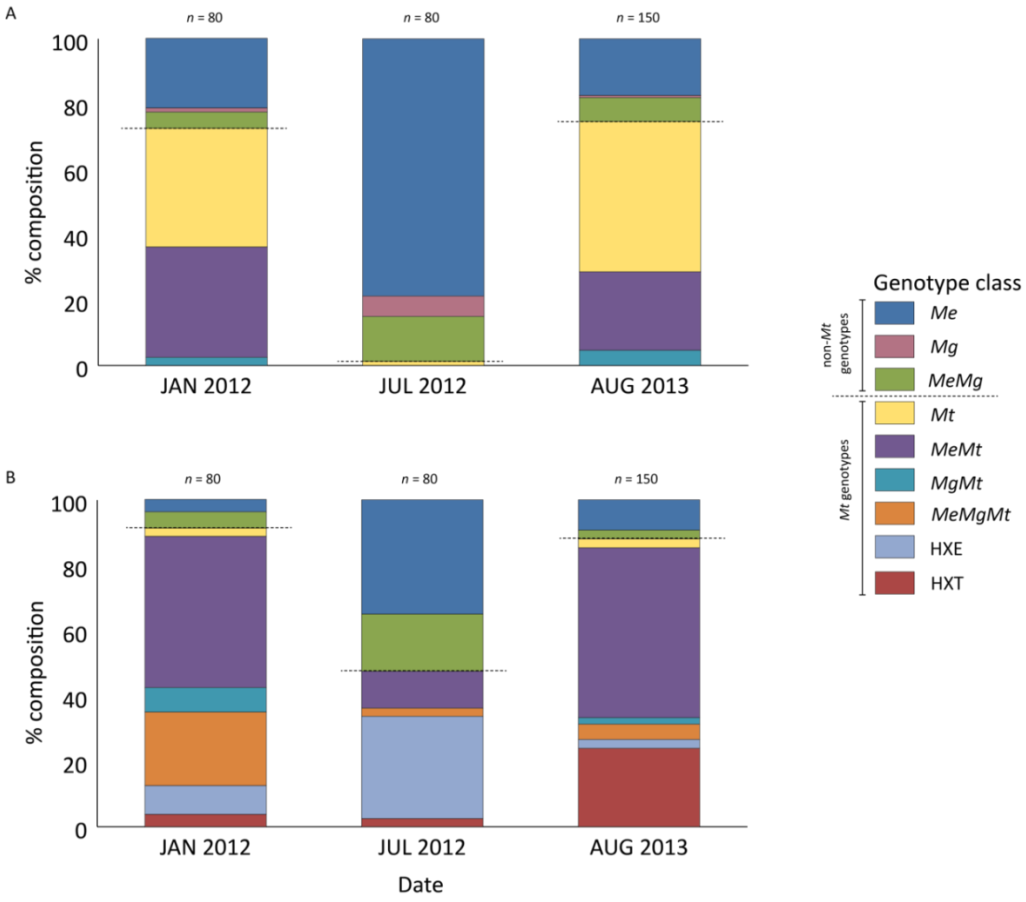

FIGURE 5.2 – Bar graph showing species composition by sampling date after genotyping with (A) Me15/16 and (B) SNP assays. "*M. trossulus* genotypes" comprise the lower part of each bar and "non-*M. trossulus* genotypes" comprise the upper part, separated by a dotted line.

were named according to the combination of alleles identified in each, with *Mg* to represent *M. galloprovincialis* diagnostic alleles: *MeMg*, *MeMt* and *MgMt* hybrids had allelic contributions from two species; *MeMgMt* hybrids had confirmed allelic contributions from three species; and HXE and HXT hybrids had a confirmed allelic contribution from one species only (HXE = *M. edulis;* HXT = *M. trossulus*), but were heterozygous at one or more diagnostic loci. All eight genotype classes were

identified in the samples from January 2012 and August 2013, but the *Mt* and *MgMt* genotype classes were absent from the July 2012 sample. *MeMt* hybrids were the most abundant in January 2012 (46.3%) and July 2013 (52%), and the *Me* genotype class was the most abundant in August 2013 (35%). *Mt* was the least abundant in January 2012 (2.5%), *MgMt* hybrids were least abundant in July 2012 (2%), and both *MeMgMt* and HXT hybrids were least abundant in August 2013 (2.5%).

5.3.1.3. Individual Type: pure species or introgressed

Using data from SNP assay genotyping, the highest proportion of pure *M. edulis* was identified in July 2012 (35%), followed by August 2013 (9.3%) and January 2012 (3.8%). Pure *M. trossulus* proportions were very low, and identified in January 2012 (2.5%) and August 2013 (2.7%) only. Introgressed (FX) hybrids were most abundant in January 2012 (92.5%), followed by August 2013 (85.3%) and July 2012 (65.1%). F1 hybrids were identified in August 2013 (2.7%) and January 2012 (1.3%) only (FIGURE 5.3).

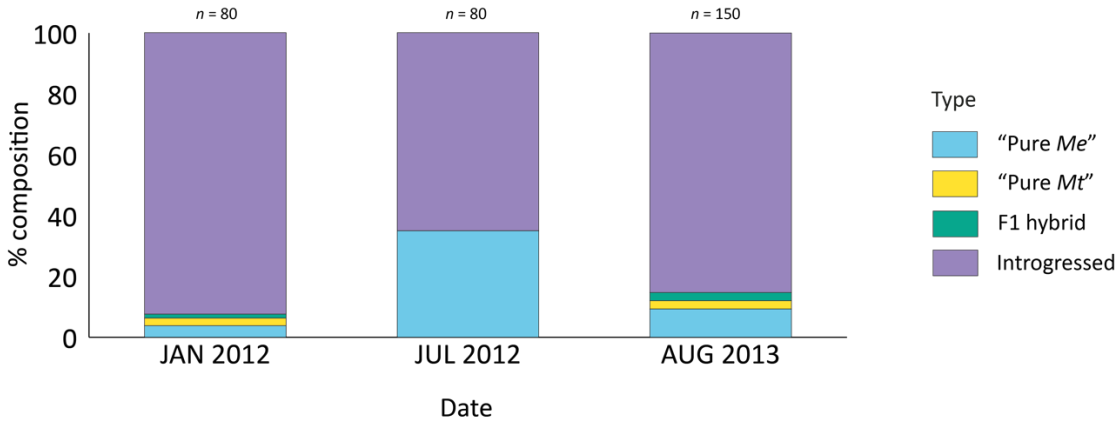

FIGURE 5.3 – Bar graph showing proportions of pure individuals [*M. edulis* (*Me*) and *M. trossulus* (*Mt*)] and different types of hybrid [F1 and introgressed (FX)] at each sampling date

5.3.2. Inferring population structure with STRUCTURE

Structure (*K*=2, burnin = 10,000, reps = 10,000) identified two different groups within the data. The genetic composition of groups was denoted by membership proportion (q) values in one of two clusters, corresponding to *M. edulis* and *M. trossulus* genotypes (FIGURE 5.4; TABLE 5.1). January 2012 and August 2013 had

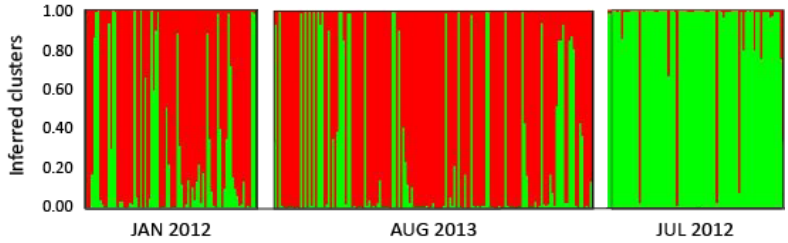

FIGURE 5.4 – Structure plots constructed using the Admixture Ancestry Model with independent allele frequencies per population [*K*=2 (Δ*K* = 37.974, determined from 100 iterations using Evanno's method (2005)), burnin = 10,000, reps = 10,000], showing the genetic composition of temporal samples from Loch Etive. Samples are grouped according to their genetic composition.

the same genetic structure, while July 2012 was distinct. All samples had membership in both clusters: Jan 2012 (q=0.697) and Aug 2013 (q=0.749) had much higher membership values in the *Mt* cluster than Jul 2012 (q=0.074). Jul 2012 had a higher membership value in the *Me* cluster (q=0.926) than Jan 2012 (q=0.303) or Aug 2013 (q=0.251).

TABLE 5.1 – Average membership proportion (q) of temporal samples in each of the two clusters assigned by STRUCTURE software

Group	Sample	*Mt*	*Me*
1	Jan 2012	0.697	0.303
	Aug 2013	0.749	0.251
2	Jul 2012	0.074	0.926

At the individual level, STRUCTURE recognised pure and hybrid individuals but did not clearly distinguish between individual Type (as designated in FIGURE 5.3). A total of 73.3% of pure *Me* individuals had q values ≥0.82, and 26.7% of pure *Me* had q values ≤0.81 in the *Me* cluster; 80% of pure *Mt* had a q value of 0.99, and the remaining 20% of pure *Mt* had a q value of 0.65. This was indistinguishable from FX hybrids: 90.9% had q values ≥0.82, and 9.1% had q values ≤0.81 in either the *Me* or *Mt* cluster. Of five F1 hybrids, only one was assigned a q value of 0.5 in both the *Me* and *Mt* clusters; the remaining four were indistinguishable from pure or FX hybrids. Although poorly distinguishing individual types within groups, STRUCTURE nevertheless recognised a clear separation between groups: $F_{ST}1=0.484$; $F_{ST}2=0.741$. Both F_{ST} values exceeded 0.25 which, according to the threshold values specified by Hartl and Clark (1997), indicated very great genetic differentiation.

5.3.3. DAPC analysis

DAPC analysis (retaining 12 Principal Components representing 100% of cumulative variance in the dataset) grouped genetic data into two clusters which had a small degree of genetic overlap (FIGURE 5.5). Consistent with Structure data,

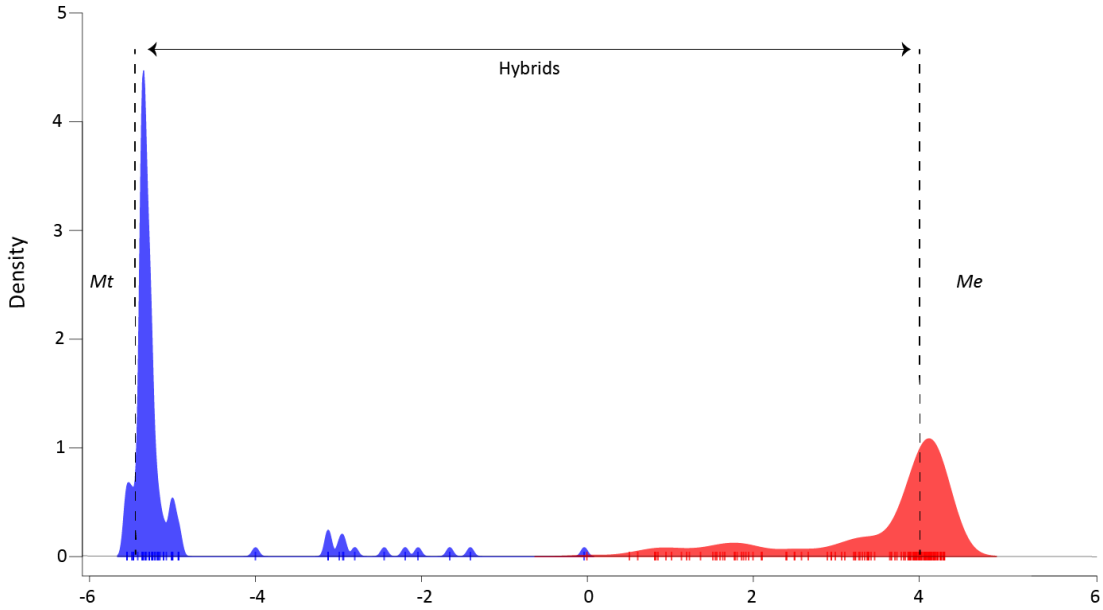

FIGURE 5.5 – DAPC scatterplot of temporal genetic data from Loch Etive (retaining 12 Principal Components and representing 100% of cumulative variance). The blue group corresponds to the Jan 2012 and Aug 2013 samples (comprising mostly *M. trossulus* and its hybrids); the red group corresponds to the Jul 2012 sample (comprising mostly *M. edulis* and its hybrids). Individuals are represented by vertical lines on the x-axis; the height of the peak corresponds to the number of individuals with a given genotype. The approximate positions of *M. edulis*, *M. trossulus* and hybrid individuals is indicated.

individuals from Jan 2012 and Aug 2013 were grouped together, and individuals from Jul 2012 were in their own group. The peak of each cluster represented pure species and introgressed hybrids with few heterozygous loci, while overlap was observable between hybrid individuals with greater proportions of heterozygous loci that, subsequently, displayed a greater degree of genetic variation.

5.3.4. M. trossulus *genetic contribution*

To estimate levels of *M. trossulus* introgression, the relative proportion of diagnostic *M. trossulus* alleles [R(AA)] was calculated for each temporal sample and compared to STRUCTURE q values (TABLE 5.2).

TABLE 5.2 - Comparison of two methods measuring the levels of *M. trossulus* introgression in temporal samples from Loch Etive: the relative proportion of diagnostic alleles [R(AA)], an equation designed specifically for use in this study based on standard allele frequency calculations; and the q values assigned by STRUCTURE which denote membership to the *M. trossulus* clusters (also detailed in TABLE 5.1)

Sample	R(AA) *Mt*	STRUCTURE *Mt*
Jan 2012	0.635	0.697
Jul 2012	0.073	0.074
Aug 2013	0.654	0.749

Both approaches were consistent in showing high proportions of *M. trossulus* alleles in Jan 2012 and Aug 2013, and very small proportions of *M. trossulus* alleles in Jul 2012. R(AA) and STRUCTURE q values differed slightly for Jan 2012 [R(AA)=0.635; q=0.697] and Aug 2013 [R(AA)=0.654; q=0.749], and were almost identical in Jul 2012 [R(AA)=0.073; q=0.074].

5.4. DISCUSSION

Controlling the genetics of broadcast spawners in a natural environment is very challenging because it requires knowledge of current species distribution, likely routes of gamete and larval transport, and any possible effects of undesirable species on an ecosystem (Palumbi, 1994; Byres *et al.,* 2002; Molnar *et al.,* 2008). When hybridisation and introgression take place between related species, as in *M. edulis* and *M. trossulus,* this challenge becomes more pronounced because it is extremely unlikely that any natural populations will be free of undesirable alleles. Even after an area is fallowed there is no way to prevent genetic flow from surrounding waters, and restocking of the area with introgressed broodstock would only continue to cause problems for production and productivity on a commercial scale (Hussain and Mazid, 2001; Mia *et al.*, 2005). The aim of stripping ropes in Loch Etive was to reduce the presence of fragile-shelled *M. trossulus* mussels and, ultimately, to allow mussel farming to resume in the area once it had become commercially viable. However, if a natural source of wild (or introgressed) *M. trossulus* exists in or near Loch Etive, such fallowing would most likely be unsuccessful.

Consistent with our Me15/16 data, previous studies applying single locus genotyping to Loch Etive identified the presence of *M. trossulus* and its hybrids with *M. edulis* and *M. galloprovincialis* (Beaumont *et al.,* 2008; Dias *et al.,* 2011a). Dias *et al* (2011a) identified greater proportions of pure *M. edulis* and pure *M. trossulus*

than *MeMt* hybrids. This contrasted with our multilocus genotyping data, and that from Zbawicka *et al* (2012) which also recognised alleles of all three *Mytilus* species in Loch Etive, but acknowledged there was a greater proportion of hybridisation between *M. edulis* and *M. trossulus* than single locus genotyping had indicated. Our data also recognised a distinct difference in species composition at different sampling dates, something not recognised by previous temporal studies in Loch Etive (Dias *et al.,* 2009b). Mussels estimated to have settled in January 2012 and August 2013 had an overall similar genetic composition (comprised mostly of *MeMt* hybrids) compared to mussels estimated to have settled in July 2012 (comprising a larger input from *M. edulis* and a smaller proportion of hybridisation). The reasons for such distribution is unclear. Bayesian clustering with STRUCTURE suggested that two pure species (and thus, two subpopulations) exist in Loch Etive, rather than each temporal sample belonging to a larger, interbreeding population. This would be consistent with native *M. edulis* existing alongside relict *M. trossulus* (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010). STRUCTURE did not reliably distinguish between types of hybrid and pure species [most likely a result of the small sample size (n=310) and small number of loci (n=12) (Vähä and Primmer, 2006)], so such an interpretation should perhaps be regarded with caution; however, STRUCTURE did recognise admixture and the presence of two pure species, which was consistent with STRUCTURE analysis of Loch Etive mussels (Zbawicka *et al.,* 2012) that also recognised different species (and possible subpopulations) within a single area. Additionally, the trends recognised in our data by the R(AA) equation (*M. trossulus* admixture) and DAPC analysis (two distinct, pure species) were equivalent to STRUCTURE data. If two subpopulations with little genetic overlap exist in Loch Etive, one with a low proportion of *M. trossulus* introgression and higher proportions of pure *M. edulis*, this could have potential for seasonal mussel aquaculture (Gosling, 1992). However, further genetic data, plus additional data on mussel shell characteristics and environmental conditions, would be required before any such conclusions could be drawn or management strategies adapted accordingly.

It is also a possibility that the temporal samples analysed represent a single, interbreeding population, rather than a relict population of *M. trossulus* and a population of native *M. edulis*. Some farmers do consider this more likely than a

relict population of *M. trossulus* in Loch Etive, because shell characteristics have not always been affected by the presence of *M. trossulus* (Gubbins *et al.,* 2012). In this case, differences in observed species composition may have arisen from environmental conditions affecting the spawning times of introgressed *M. trossulus* compared to *M. edulis.* Spawning of *Mytilus* spp. takes place throughout the year and peak spawning times can vary between species [i.e., April-May for *M. edulis,* and July-October for *M. trossulus* (Chipperfield, 1953; Toro *et al.,* 2002)], influenced by abiotic factors including tidal flow, salinity, pH and temperature (Bierne *et al.*, 2003; Fly and Hilbish, 2013). Perhaps in July 2012 environmental conditions more strongly favoured *M. edulis* than they did introgressed *M. trossulus*, which affected spawning time and subsequent spat settlement. This is, however, unlikely based on a previous study of spawning times in Loch Etive (Dias *et al.,* 2009b), which demonstrated little monthly difference between *M. edulis, M. trossulus* and their hybrids in a large, hybridising population. This study subsequently concluded that the timing of rope deployment was unlikely to affect the species composition of settling larvae. Fallowing, which took place in Loch Etive in August 2010 and in July 2011, could also have affected species composition in temporal samples. After initial fallowing in 2010, divers observed that over 99% of mussels deposited on the seabed were predated by crabs, whelks and starfish within a three-month period (Gubbins *et al*., 2012). However, there is no record of such observations being made after the second fallowing in July 2011, so it is not known if similar proportions of mussels were eradicated from the loch; if their larvae were already present in the water column prior to fallowing; or if they persisted and had a chance to spawn. If spawning did take place prior to or after the 2011 fallowing event, it is possible that larvae estimated to have settled in January 2012 came from these fragile individuals, thereby explaining the high proportion of introgressed *M. trossulus* identified. Perhaps larvae estimated to have settled in July 2012 (i.e., 12 months after fallowing) were those of less fragile mussels that had avoided predation, accounting for the lower proportion of introgressed *M. trossulus* detected in this sample.

The proportion of introgressed *M. trossulus* in the August 2013 sample, believed to comprise individuals estimated to have settled some 24 months after fallowing, is interesting because it suggests that *M. trossulus* retains a significant presence in Loch

Etive or its close environments and, potentially, that fallowing was ineffective after some initial success in reducing the presence of *M. trossulus*. Although this study did not genotype any samples taken after 2013, there is anecdotal evidence of high proportions of *M. trossulus* remaining in Loch Etive in 2015 (Marine Scotland Science, unpublished genotyping data); this subsequently raises doubt about the effectiveness of fallowing and, furthermore, raises the question of whether other strategies should be considered for the long term management of *M. trossulus* in Loch Etive. With such extensive levels of introgression observed here it will be very difficult to eliminate *M. trossulus* genetic contribution from a marine environment, presenting challenges when devising management plans for its control (Thresher and Kuris, 2004). However, it must be acknowledged that, without knowledge of the proportion of introgressed *M. trossulus* before fallowing (in 2011), it is impossible to say whether or not these results reflect "successful" fallowing measures. It may also be too early to draw conclusions about the impacts of fallowing from such a short term study, because a 25 month period only covers a single generation in the *Mytilus* life cycle (i.e., from spawning to settlement). Thus, the collection and analysis of further data is needed before a more complete picture of species composition in Loch Etive can be made available. Management approaches in Loch Etive cannot be revised until the effects of *M. trossulus* hybridisation on shell fragility have been more completely evaluated. This is a problem that needs to be addressed by the Scottish shellfish industry and is an issue that single locus genotyping (with Me15/16) will be insufficient to tackle in any great detail. Multilocus genotyping will be essential in future studies of Loch Etive if species composition and its potential suitability as an aquaculture site are to be assessed, and if more productive and profitable shellfish aquaculture is ever to become a possibility in this once thriving area.

5.5. CONCLUSIONS AND SUMMARY

1. Despite fluctuations in abundance, the relative proportion of *M. trossulus* alleles in Loch Etive was high in August 2013, 25 months after a major fallowing event in July 2011;

2. Without a longer term study to determine the levels of *M. trossulus* introgression and the subsequent effect on shell fragility, Loch Etive cannot yet be considered a suitable site for mussel farming;

3. Single locus genotyping with Me15/16 is insufficient to detect the extent of hybridisation in Loch Etive; multilocus SNP genotyping should be applied to future studies for more effective site management.

Acknowledgements

Thanks to Walter Speirs for the Loch Etive backstory, and thanks to Rebecca McIntosh (Marine Scotland Science) for DNA extraction and Me15/16 genotyping. The work was funded by Marine Alliance for Science and Technology for Scotland (MASTS) and Marine Scotland Science.

Chapter 6
General discussion and conclusions

The present research investigated the phylogenetic relationships between field samples of Scottish mussels; these were taken from both farmed and wild sites on the mainland and in Orkney, and farmed sites in The Inner and Outer Hebrides and Shetland. There were four main outcomes:

1. Multilocus SNP analysis based on RADseq data confirmed that the three species in the *M. edulis* species complex are genetically distinct. This presented the possibility of exploring introgression through surveying species diagnostic markers;
2. Twelve novel SNP markers were successfully established as robust, effective genotyping tools for genotyping all surveyed samples, which were capable of reliably distinguishing pure species (*M. edulis*, *M. galloprovincialis* and *M. trossulus*) from introgressed (FX) hybrids and recognising F1 hybrids, thereby revealing complex genetic structure where previous studies had been unable to;
3. A clear difference was observable between the genetics of most farmed stock and wild populations, indicating an anthropogenic influence on introgression levels and, subsequently, population genetic structure;
4. Despite temporal variation in population genetic structure over a short term study (19 months), the relative proportion of *M. trossulus* alleles in Loch Etive was high when measured 25 months after fallowing had taken place.

The following chapter reviews these main outcomes in more detail, along with their possible limitations, and discusses how the findings could be applied to future studies.

6.1. OVERALL CONCLUSIONS

Examining the genetics of the *M. edulis* species complex is challenging because it comprises three closely-related sympatric species, among which hybridisation

frequently occurs (e.g., Gardner, 1996; Rawson *et al.,* 1996; Brooks, 2000; Gardeström *et al.,* 2008). Although useful for preliminary studies of species identification, single locus genotyping is not suited to monitoring introgression and, if contributing to a decrease in fitness, introgression could have negative consequences for aquaculture (Bekkevold, 2006). Hence, utilising multilocus genotyping in *Mytilus* spp. populations is beneficial for investigating introgression, and therefore population structure, in greater detail. There has historically been debate about the actual taxonomic status of *M. edulis, M. galloprovincialis* and *M. trossulus,* and whether or not they are discrete species (Seed, 1971; Gosling, 1984). Previous studies of Scottish mussels (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a; Zbawicka *et al.,* 2012) have considered *M. edulis, M. galloprovincialis* and *M. trossulus* to be separate species, a convention retained in the present study for ease of comparison. Thus, assessing the phylogenetic relationships of the populations used for species diagnostic marker development was crucial for the commencement of this research. Diagnostic marker development could only take place if three discrete species were, as assumed, present in the *Mytilus edulis* species complex. Our data demonstrated a clear separation of individuals based on genotype and not geographic distance: *M. edulis* from three different population samples (two Scottish and one North American) was grouped together (n=21), separate from the single population sample of *M. galloprovincialis* from Slovenia (n=15), and the single population sample of *M. trossulus* (n=4) from North America (see FIGURE 3.3, p68 for reference). All diagnostic markers appeared fixed in presumed pure population samples. The main limitation of this research has arisen from the particularly small sample size and single reference site used to develop *M. trossulus* specific markers, which could affect the robustness of these markers on a larger scale. That said, the diagnostic power of the *M. trossulus* markers held during validation with *M. trossulus* samples from additional sites (Scotland and Canada; CHAPTER 3) and when applied to a larger scale study of Scottish population samples (CHAPTER 4), thereby strengthening their diagnostic usefulness for future studies. The *M. edulis* and *M. galloprovincialis* markers can be considered more reliable because they were developed from slightly larger sample sizes. Overall, the small number of species-specific differences recognised with RADseq (i.e., 349

species-diagnostic loci from over 38,000 shared markers), which separated the members of the *M. edulis* species complex, would perhaps indicate the "species" classification for *M. edulis, M. galloprovincialis* and *M. trossulus* was not wholly reliable. However, consistent and clear discrimination of all surveyed individuals with SNP genotyping, plus DAPC and STRUCTURE analysis, demonstrated robustness of the optimised assays in recognising unique polymo1rphisms. Subsequently, for the purposes of this study, *M. edulis, M. galloprovincialis* and *M. trossulus* were still considered as three discrete species despite widespread genetic overlap in Scottish population samples, consistent with previous studies in the country.

Multilocus genotyping of Scottish mussels has improved our knowledge of genetic admixture within and between populations, identifying introgressed genotypes that were hitherto unrecognisable with single locus (Me15/16) genotyping (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a). Such results challenge the effectiveness of Me15/16 as a monitoring tool, and indicate that more in depth genotyping will be of far greater benefit in making profitable management decisions. This is particularly relevant in the context of a new mussel hatchery being developed in Shetland, which aims to cultivate and sell native *M. edulis.* However, contrary to Me15/16 genotyping which showed high proportions of *M. edulis,* our data instead revealed a deficit of pure *M. edulis* and widespread admixture between *M. edulis* and *M. galloprovincialis* at the Shetland sites genotyped. With such widespread introgression from *M. galloprovincialis,* it will not only be difficult to source pure species, but it will furthermore be impossible to eliminate *M. galloprovincialis* alleles from sites in Shetland and in the rest of Scotland. Although in far lower proportions than *M. galloprovincialis* alleles, *M. trossulus* alleles were also present at the Shetland sites genotyped. While alleles from either species may not necessarily affect the quality of seed, given previous evidence of the negative effects of *M. trossulus* (e.g., Dias *et al.,* 2009b; Dias *et al.,* 2011a) it is unlikely to have a high economic value, which could result in economic losses for the hatchery itself. This also raises issues for product marketing and product branding, which in turn could affect customer satisfaction and, subsequently, overall production at key shellfish sites. Such outcomes are not guaranteed but certainly raise questions about the feasibility of cultivating and marketing pure *M. edulis* in Shetland, and

furthermore emphasise that management decisions based solely on evidence from Me15/16 genotyping could be ill-advised and unsustainable in the long term.

When population samples were grouped according to genotype [see FIGURE 4.6 (p108) and FIGURE 4.7 (p109)], genetic structure did not segregate according to geographic proximity. In most cases, a clear difference was observable between the genetic composition of rope grown aquaculture sites (Groups A-E), and shoreline sites (Group F). Exceptions to this were Scapa Beach and Northside (Group A), two wild shoreline sites with a genetic composition similar to rope grown aquaculture sites in Shetland and the Isle of Lewis; and Loch Spelve, a rope grown aquaculture site with a genetic composition similar to shoreline population samples in Group F. Group F comprised much higher levels of pure *M. edulis* than Groups A-E, which instead comprised higher proportions of hybrid genotypes. It is not known whether all rope grown aquaculture sites in Groups A-E import spat from outside sources; however, such a notable difference from the wild stock did indicate anthropogenic influences on current species composition (as suggested by, e.g., Dias *et al.,* 2009a and Gubbins *et al.,* 2012). Possible genetic contamination of wild stock by anthropogenic spat translocation was notable in Orkney: two shoreline sites (Scapa Beach and Northside) which had genetic compositions similar to farms hundreds of miles away in Shetland and the Isle of Lewis – a pattern that would be unlikely to arise with natural spat migration [i.e., an estimated dispersal distance of 30 km per generation (Bierne *et al.,* 2003)]. Genetic contamination of wild stock from farms could present challenges for farmers wishing to source pure broodstock or set up new farms (Hussain and Mazid, 2001; Mia *et al.,* 2005). In the context of Scottish shellfish aquaculture this raised questions about existing regulations and their effects on population genetics. It is impossible to stop hybridisation in a natural environment where there is free exchange of gametes in the water column (Palumbi, 1992), but reducing anthropogenic interference could mitigate hybrid spread (Dias *et al.,* 2008; Dias *et al.,* 2009c). Legally, shellfish farms must keep records of spat imports and exports, but there is generally no restriction on spat movement in Scotland if it is known to come from disease free areas in Britain and Ireland (Andrew Mayes, personal communication, 23rd June 2015). From a commercial perspective, if introgression is not affecting production then there may be no need to revise such

regulations (Dias *et al.*, 2009b). By all accounts the widespread introgression and co-existence of hybrid and parental forms observed in our data would indicate that, overall, hybridisation does not impact on individual survival or fitness (Spaak and Hoekstra, 1995) and that current levels of introgression in Scottish populations are not an immediate threat to production in the absence of contradictory evidence. Nevertheless, more control on the translocation of spat could still benefit farming by reducing continued genetic contamination and dilution of the native gene pool, and by making taxonomic classifications and "species" management less complicated in the long term (Allendorf *et al.*, 2001; Stronen and Paquet, 2013). Until any changes to these regulations are implemented, however, the effects of anthropogenic interference on species composition cannot be evaluated.

With such genetic variability observable between FX hybrids, hybridisation and introgression in *Mytilus* spp. mussels must be complex processes in which no single genetic or environmental factor determines the degree of interspecies gene exchange (Marques *et al.*, 2007). The suite of 12 diagnostic SNP markers confirmed the presence of three distinct *Mytilus* species and recognised widespread introgression in Scottish population samples. While this fulfilled the overall aim of the thesis, it is possible that the small number of genetic markers led to incorrect assignment of hybrid type (Vähä and Primmer, 2006) in some cases, particularly with regards to F1 hybrids. Our data identified F1 *MeMg* hybrids in St Andrews, even though no pure *M. galloprovincialis* was recognised at this site or any other site genotyped, and detected F1 *MeMt* hybrids in Loch Fyne despite no pure *M. trossulus* being detected in this area. Possible errors in individual category assignment was also demonstrated by some discrepancies in Bayesian analysis: STRUCTURE identified pure and hybrid individuals but was limited in its ability to distinguish between F1 and FX hybrids; and NEWHYBRIDS identified pure individuals but was limited in its ability to distinguish pure from FX, and was unable to distinguish F1 from FX. Nevertheless, the recognition of F1 *MeMt* hybrids in Loch Fyne and the nearby Loch Etive is an interesting discovery that could indicate an active hybrid zone, and furthermore suggests a capacity of *M. trossulus* to spread and to become a production issue at sites outside of Loch Etive (Gubbins *et al.*, 2012). However, this cannot be confirmed without further evidence on the distribution of *M. trossulus*, or evidence to

detail the actual effects of *M. trossulus* on shell fragility. Although confirmed as discrete species capable of hybridisation, the taxonomic relationships of *M. edulis, M. galloprovincialis* and *M. trossulus* could still affect the readiness with which they hybridise; for instance if *M. edulis* and *M. galloprovincialis* are less diverged than either species is from *M. trossulus* (as suggested by e.g., Gérard *et al.,* 2008; Zbawicka *et al.,* 2012 and Astorga *et al.,* 2015; and observed in our DAPC and phylogenetic analyses). Although hybridisation does occur between *M. galloprovincialis* and *M. trossulus,* greater genetic divergence between these two species in comparison to less genetic divergence between *M. edulis* and *M. galloprovincialis* could account for the historical rarity of *MgMt* hybrids in Scotland (Beaumont *et al.,* 2008; Dias *et al.,* 2009c; Dias *et al.,* 2011a) and the low abundances in our data, in comparison to much higher proportions of *MeMg* hybrids.

The temporal genetic data from Loch Etive indicated either the presence of two genetically distinct subpopulations in the loch [consistent with native *M. edulis* and relict *M. trossulus* (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010)]; or of one large sympatric population of hybridising *M. edulis* and *M. trossulus. M. trossulus* alleles were present in at least 50% of individuals at each sampling date, and pure *M. trossulus* itself was identified in two of the three temporal samples. Recognition of a greater degree of hybridisation than in previous studies (Beaumont *et al.,* 2008; Zbawicka *et al.,* 2010; Dias *et al.,* 2011a**)** emphasised the usefulness of multilocus SNP genotyping in understanding population dynamics, a recurring theme throughout the thesis. Previous studies suggest that the low salinity environment in the upper layers of Loch Etive favours the settlement of *M. trossulus* over *M. edulis* and *M. galloprovincialis,* which influences shell fragility (Beaumont *et al.,* 2008; Dias *et al.,* 2009b; Dias *et al.,* 2011a). However, the actual factors causing shell fragility remain unevaluated. A tentative relationship between the *M. trossulus* allele and shell fragility was demonstrated at a site of current commercial importance outside Loch Etive, but this link was not established elsewhere, even in areas where levels of *M. trossulus* introgression appeared higher. It is possible that a "weak" *M. trossulus* allele conferring shell fragility exists and persists on aquaculture ropes because they are more sheltered than the shoreline (Dias *et al.,* 2009c). The actual effects of this allele on shell strength could be mediated by environmental variation

and/or epistatic interactions [i.e., the effects of other genes (Philips, 2008)], explaining why there was no evidence to indicate that almost all rope grown sites with high proportions of *M. trossulus* were not known to exhibit fragility-related production issues. Since fallowing in Loch Etive took place, species composition has only been monitored at a single site in Achnacloich, and only juvenile mussels have been examined. Achnacloich may not have conditions representing the whole of Loch Etive because it is in the lower basin: this is shallower and more frequently exchanges water with the outside sea than the upper basin, which has deeper water prone to stagnation (Edwards and Edelsten, 1997). Additionally, juvenile mussels have thin, fragile shells compared to the more robust shells of adults (Helm *et al.*, 2004), so only examining juveniles will not necessarily reflect the characteristics of adults. The upper and lower basins do have slightly different environmental conditions [i.e., temperature, salinity and pH (Gage, 1972)] which could have different effects on the relative proportion of introgressed *M. trossulus* present; on shell fragility; or on both. It could be worthwhile having an additional monitoring site in the upper basin and comparing the proportion of introgressed *M. trossulus* to that in Achnacloich, and for both of these sites to genotype and test the shell strength of adult and juvenile mussels. Beaumont *et al* (2008) and Dias *et al* (2008) looked at multiple sites throughout Loch Etive and revealed that *M. trossulus* and its hybrids were widespread, so restricting focus to a single site in the region is not particularly reliable for comparison. There is no evidence to indicate that the other sites where *M. trossulus* has been identified have reported shell fragility or been studied in any detail (Gubbins *et al*., 2012). In the present study, results from Site X were only observed in a very small sample size (n=19); thus, the genetic link with shell fragility is somewhat tenuous and until further study is carried out, it remains no more than a hypothesis. In depth multilocus genotyping will be crucial in carrying such research forward, and it would be prudent to favour its use over that of Me15/16 if more informed and effective management decisions are to be made by the Scottish shellfish industry.

6.2. DIRECTIONS FOR FUTURE RESEARCH

Genotyping of commercially important shellfish species has been carried out worldwide for both commercial and conservational purposes. For instance, genotyping commercially important *Crassostrea* spp. oysters in China (Wang *et al.*, 2014) and Unionoid mussels in Thailand (Vannarattanarat *et al.,* 2013) resolved taxonomic uncertainties among morphologically similar individuals. Both studies provided tools for genetic screening that could aid the development of breeding programmes. Population genetic studies of Dreissenid mussels in the Caspian and Black Seas revealed the presence of distinct species whose distribution is affected by salinity gradients. Dreissenid mussels are biofouling organisms which can be problematic for ecosystems, so knowledge of their distribution is beneficial from a population management perspective (Therriault *et al.,* 2004). The present research provides an overview of the levels of admixture in Scottish *Mytilus* spp. Mussel populations. This holds a great deal of potential for future research aiming to evaluate the possible impacts of hybridisation and introgression on the shellfish industry. It also has potential in helping to establish management strategies for mitigating such effects, and emphasises the benefits of multilocus (SNP) genotyping compared with the limited genetic data available from single locus (Me15/16) genotyping.

Although a clear difference between farmed and wild stock was observable, there were still areas of the coast that were not sampled, thereby providing an incomplete picture of species composition on both the mainland and the islands. Continuing to use the panel of 12 SNP markers for genotyping, sampling could be extended and sample sizes increased for a better representation of sites that were not included in this study. If further extended to cover a number of years, temporal species composition for the whole of Scotland could be established. There is also the need for research focused on specific sites for more targeted management: for instance, studying wild and farmed populations in close proximity could establish any possible genetic effects of farmed stock on wild stock. Such data would have useful implications for broodstock sourcing and new site selection, which could be of particular benefit to development and expansion of the proposed Scottish mussel hatchery. Utilising SNP markers on a broader scale would enable genotyping of

Mytilus spp. populations worldwide, facilitating a more detailed look at the structure of "hybrid zones" which have, until now, only been identified with single locus genotyping. A more detailed study of the suggested "hybrid zone" in Loch Etive and Loch Fyne, using SNP markers and mitochondrial markers to investigate individual origins, would also be useful in understanding the capacity of *M. trossulus* to spread around Scotland.

A larger scale, temporal study of Site X would benefit in helping to establish the cause of shell fragility. In addition to affecting gene persistence and expression, environmental variation could exert a direct influence on shell characteristics. Abiotic factors (e.g., pH, temperature and salinity) can affect calcium carbonate absorption in *Mytilus* spp. mussels and thus the robustness of the outer shell (Doney *et al.*, 2009). Inner shell strength can be affected by the chemical composition of body fluids, which may become increasingly acidic and corrosive during periods of stress (such as poor food availability and low oxygen levels) (Melzner *et al.*, 2011). SNP markers could be used as a starting point in identifying possible Quantitative Trait Loci associated with shell robustness under a range of environmental parameters (as in, e.g., a study of the bivalve *Hyriopsis cumingii* (triangle sail mussel) (Lea, 1852) (Bai *et al.*, 2016), which used SNPs as a starting point to identify QTL associated with shell characteristics**)**. Investigating the genetic composition of species under different environmental conditions would allow the possible threat from *M. trossulus* to be investigated in greater detail. It needs to be established whether this species and its hybrids are dangerous for mussel production, or if they can instead be considered a harmless component of natural populations which do not pose any real threat to Scottish aquaculture.

Management techniques promoting productivity and profitability could be designed if the effects of introgression on production, and the causes of shell fragility were understood. This research demonstrated differences in temporal species composition in Loch Etive, potentially related to differential spawning times of *M. edulis* and *M. trossulus.* Investigating species composition and differences in settlement time at other sites could possibly establish the optimal time for casting spat collection ropes, thereby maximising the collection of spat with "desirable" alleles. There is evidence that positioning ropes deeper in the water of Loch Etive

favours the settlement of *M. edulis* over *M. trossulus* (Dias *et al.*, 2008), another potential management strategy that could be exploited by farmers if genetic composition with spatial variation was studied. However, in terms of a "model site" for population management, Loch Etive may not be entirely suitable: our data indicates Loch Etive has a genetic composition unlike any other site in Scotland, and a hyposaline upper layer (Gubbins *et al.*, 2012). Thus, for the purposes of modelling population management, it might be better if studies were focussed towards sites with genetic compositions and environments more like those found in "normal" farming environments. Identifying and monitoring such sites, using genotypic and environmental data, could prove more beneficial for population management in Scottish aquaculture than previous case studies have permitted.

To conclude, this research demonstrated that the *M. edulis* species complex comprises three genetically distinct but interfertile mussel species: *M. edulis, M. galloprovincialis* and *M. trossulus.* To date, single locus genotyping has overestimated the proportion of pure *Mytilus* species in Scotland. Utilising 12 new SNP markers facilitated higher resolution genotyping than had been permissible by single locus genotyping in previous research, thereby presenting a new genotyping tool that will be invaluable in future studies aimed at improving productivity and profitability in the Scottish shellfish industry. Although the small number of markers used may have led to some errors in assignment of hybrid type, multilocus genotyping still revealed widespread introgression within the *M. edulis* species complex and indicated anthropogenic influences could have a strong effect on genetic composition at farmed sites. However, despite this more complex genetic structure, introgression does not appear disadvantageous, and it is likely that environmental factors play a role alongside genetics in influencing the appearance of commercially undesirable characteristics among mussels.

BIBLIOGRAPHY

Abbott, R., Albach, D., Ansell, S., Arntzen, J.W., Baird, S.J.E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C.A., Buggs, R., Butlin, R.K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S.H., Hermansen, J.S., Hewitt, G., Hudson, A.G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-Rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A.W., Parisod, C., Pfennig, K., Rice, A.M., Ritchie, M.G., Seifert, B., Smadja, C.M., Stelkens, R., Szymura, J.M., Väinölä, R., Wolf, J.B.W., Zinner, D., 2013. Hybridization and speciation. J. Evol. Biol. 26, 229–246.

Abdi, H., Williams, L. J. (2010). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433–470.

Ahmad, M., Beardmore, J.A., 1976. Genetic Evidence that the "Padstow Mussel" is *Mytilus galloprovincialis*. Mar. Biol. 147, 139–147.

Aljanabi, S.M., Martinez, I., 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Res. 25, 4692–4693.

Allendorf, F., Leary, R., Spruell, P., Wenburg, J., 2001. The problems with hybrids: Setting conservation guidelines. Trends Ecol. Evol. 16, 613–622.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. Lipman, D.J., 1990. "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

Amish, S., Hohenlohe, P., Painter, S., Leary, R., Muhlfeld, C., Allendorf, F., Luikart, G., 2012. RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. Mol. Ecol. Resour. 12, 653–660.

Anderson, E.C., Thompson, E.A., 2002. A model-based method for identifying species hybrids using multilocus data. Genetics 160, 1217 –1229.

Anderson, E. C., 2008. Bayesian inference of species hybrids using multilocus dominant genetic markers. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 363(1505), 2841–50.

Anderson, A., Bilodeau, A., Gilg, M., Hilbish, T., 2002. Routes of introduction of the Mediterranean mussel (*Mytilus galloprovincialis*) to Puget Sound and Hood Canal. J. Shellfish Res. 21, 75–79.

Anderson, E. C., Thompson, E. A., 2002. A model-based method for identifying species hybrids using multilocus data. Genetics, 160(3), 1217–1229.

Anderson, E., Hubricht, L., 1938. Hybridization in Tradescantia. III. The Evidence for Introgressive Hybridization. Am. J. Bot. 25, 396–402.

Andrews S., 2010. FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., Hohenlohe, P. A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet (2), 81–92.

Araneda, C., Larraín, M. A., Hecht, B., Narum, S., 2016. Adaptive genetic variation distinguishes Chilean blue mussels (*Mytilus chilensis*) from different marine environments. Ecology and Evolution, 6(11), 3632–3644.

Argue, B.J., Liu, Z., Dunham, R.A., 2003. Dress-out and fillet yields of channel catfish, Ictalurus punctatus, blue catfish, *Ictalurus furcatus*, and their F1, F2 and backcross hybrids. Aquaculture 228, 81–90.

Arnold, M., 1997. Natural Hybridization and Evolution. Oxford University Press.

Arnold, M., Martin, N., 2009. Adaptation by introgression. J. Biol. 8, 82.

Astorga, M. P., Cardenas, L., Vargas, J., Ambientales, C., 2015. Phylogenetic Approaches To Delimit Genetic Lineages of the *Mytilus* Complex of South America : How Many Species Are There ? Journal of Shellfish Research, 34(3), 1–12

Austin, W., Inall, M., 2002. Deep-water renewal in a Scottish fjord: Temperature, salinity and oxygen isotopes. Polar Res. 21, 251–258

Back, B., Laitinen, Teija; Sere, K. M. V. W., 1996. Choosing Bankruptcy Predictors Using Discriminant Analysis , Logit Analysis , and Genetic Algorithms. Proceedings of the Ist International Meeting on Artificial Intelligence in Accounting, Finance and Tax, (40), 337--356.

Bai, Z.-Y., Han, X.-K., Liu, X.-J., Li, Q.-Q., Li, J.-L., 2016. Construction of a high-density genetic map and QTL mapping for pearl quality-related traits in Hyriopsis cumingii. Scientific Reports, 6(August), 32608.

Baird, N., Etter, P., Atwood, T., Currey, M., Shiver, A., Lewis, Z., Selker, E., Cresko, W., Johnson, E., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3, e3376.

Baker, A. M., Bartlett, C., Bunn, S. E., Goudkamp, K., Sheldon, F., Hughes, J. M.., 2003. Cryptic species and morphological plasticity in long-lived bivalves (Unionoida: Hyriidae) from inland Australia. Molecular Ecology, 12(10), 2707–2717

Baranwal, V., Mikkilineni, V., Zehr, U., Tyagi, A., Kapoor, S., 2012. Heterosis: emerging ideas about hybrid vigour. J. Exp. Bot. 63, 695–709.

Bartley, D.M., Rana, K., Immink, A.J., 2000. The use of inter-specific hybrids in aquaculture and fisheries. Rev. Fish Biol. Fish. 10, 325–337

Barton, N.H., 2001. The role of hybridization in evolution. Mol. Ecol. 10, 551–568

Barton, N.H., Hewitt, G., 1985. Analysis of Hybrid Zones. Annu. Rev. Ecol. Evol. Syst. 16, 113–148

Battonyai, I., Specziár, A., Vitál, Z., Mozsár, A., Görgényi, J., Borics, G., Tóth, L.G., Boros, G., 2015. Relationship between gill raker morphology and feeding habits of hybrid bigheaded carps (*Hypophthalmichthys* spp.). Knowl. Manag. Aquat. Ecosyst. 416, 1–11

Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D., Blaxter, M.L., 2011. Linkage Mapping and Comparative Genomics Using Next-Generation RAD Sequencing of a Non-Model Organism. PLoS One 6, 11.

Bayne, B.L., 1965. Growth and the delay of metamorphosis of the larvae of *Mytilus edulis* (L.). Ophelia 2, 1–47.

Beaumont, A., Gjedrem, T., Moran, P., 2007. Blue mussel - *Mytilus edulis* Mediterranean mussel - *M. galloprovincialis*, in: Genimpact - Evaluation of Genetic Impact of Aquaculture Activities on Native Populations (Final Scientific Report). pp. 62–69.

Beaumont, A., Hawkins, M.P., Doig, F.L., Davies, I.M., Snow, M., 2008. Three species of *Mytilus* and their hybrids identified in a Scottish Loch: natives, relicts and invaders? J. Exp. Mar. Bio. Ecol. 367, 100–110.

Beaumont, M. A, Rannala, B., 2004. The Bayesian revolution in genetics. Nature Reviews. Genetics, 5(4), 251–261.

Bekkevold, D., Hansen, M., Nielsen, E., 2006. Genetic impact of gadoid culture on wild fish populations: Predictions, lessons from salmonids, and possibilities for minimizing adverse effects. ICES J. Mar. Sci. 63, 198–208.

Benzecri J. P., 1992. Correspondence analysis handbook. In: Statistics: a series of textbooks and monographs, vol 125. Marcel Dekker Inc, New York

Berry, P.Y., Low, M.P., 1970. Comparative Studies on Some Aspects of the Morphology and Histology of *Ctenopharyngodon idellus*, *Aristichthys nobilis*, and Their Hybrids (Cyprinidae). Am. Soc. Ichthyol. Herpetol. 708–726.

Bierne, N., David, P., Boudry, P., Bonhomme, F., 2002. Assortative Fertilization and selection at larval stage in the mussels *Mytilus edulis* and *M. galloprovincialis*. Evolution (N. Y). 56, 292–298.

Bierne, N., Borsa, P., Daguin, C., Jollivet, D., Viard, F., Bonhomme, F., David, P., 2003. Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. Mol. Ecol. 12, 447–461.

Bierne, N., Bonhomme, F., Boudry, P., Szulkin, M., David, P., 2006. Fitness landscapes support the dominance theory of post-zygotic isolation in the mussels *Mytilus edulis* and *M. galloprovincialis*. Proc. R. Soc. B Biol. Sci. 273, 1253–1260.

Birchler, J., Yao, H., Chudalayandi, S., Vaiman, D., Veitia, R., 2010. Heterosis. Plant Cell 22, 2105–2112.

Bondad-Reantaso, M., 2007. Assessment of freshwater fish seed resources for sustainable aquaculture, FAO Fisheries Technical Paper.

Bostock, J., McAndrew, B., Richards, R., Jauncey, K., Telfer, T., Lorenzen, K., Little, D., Ross, L., Handisyde, N., Gatward, I., Corner, R., 2010. Aquaculture: global status and trends. Philos. Trans. R. Soc. London - Ser. B Biol. Sci. 365, 2897–2912.

Bosworth, B.G., Wolters, W.R., Silva, J.L., Chamul, R.S., Park, S., 2004. Comparison of production, meat yield, and meat quality traits of NWAC103 line Channel catfish, Norris line Channel catfish, and female Channel catfish × male Blue catfish F1 Hybrids. N. Am. J. Aquac. 66, 177–183.

Brooks, K., 2000. Literature review and model evaluation describing the environmental effects and carrying capacity associated with the intensive culture of mussels (*Mytilus edulis galloprovincialis*).

Burgarella, C., Lorenzo, Z., Jabbour-Zahab, R., Lumaret, R., Guichoux, E., Petit, R. J., Gil, L. (2009). Detection of hybrids in nature: application to oaks (Quercus suber and Q. ilex). Heredity, 102(5), 442–452.

Burke, J.M., Arnold, M.L., 2001. Genetics and the fitness of hybrids. Annu. Rev. Genet. 35, 31–52.

Byrne, M., 2011. Impact of ocean warming and ocean acidification on marine invertebrate life history stages: vulnerabilities and potential for persistence in a changing ocean. Oceanogr. Mar. Biol. An Annu. Rev. 49, 1–42.

Capote, N., Pastrana, A.M., Torres-tomejil, L., Andalucía, J. De, Río, A., 2012. Molecular Tools for Detection of Plant Pathogenic Fungi and Fungicide Resistance, in: Plant Pathology. InTech, p. 374

Carlton, J.T., Geller, J.B., 1993. Ecological roulette: the global transport of nonindigenous marine organisms. Science 261, 78–82.

Catchen, J. M., 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology, 22(11), 3124–3140.

Catchen, J., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J., 2011. Stacks: building and genotyping Loci de novo from short-read sequences. G3 Genes|Genomes|Genetics 1, 171–82.

Chen, X., Sullivan, P.F., 2003. Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. Pharmacogenomics J. 3, 77–96.

Chipperfield, P.N.J., 1953. Observations on the Breeding and Settlement of *Mytilus edulis* (L) in British Waters. J. Mar. Biol. Assoc. United Kingdom 32, 449–476.

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., O'Toole, P. W., 2010. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Research, 38(22), e200.

Coghlan, B., Gosling, E., 2007. Genetic structure of hybrid mussel populations in the west of Ireland: two hypotheses revisited. Mar. Biol. 150, 841–852.

Collin, R., 2000. Phylogeny of the *Crepidula plana* (Gastropoda : Calyptraeidae) cryptic species complex in North America. Can. J. Zool. 78, 1500–1514.

Cook, K. (2013). ICES Zooplankton Status Report 2010/2011. In ICES Cooperative Research Report (pp. 123–126).

Corander, J., Marttinen, P., 2006. Bayesian identification of admixture events using multilocus molecular markers. Molecular Ecology, 15(10), 2833–2843.

Coscia, I., Robins, P. E., Porter, J. S., Malham, S. K., Ironside, J. E., 2013. Modelled larval dispersal and measured gene flow: Seascape genetics of the common cockle *Cerastoderma edule* in the southern Irish Sea. Conservation Genetics, 14(2), 451–466.

Costedoat, C., Pech, N., Chappaz, R., Gilles, A., 2007. Novelties in hybrid zones: Crossroads between population genomic and ecological approaches. PLoS One 2.

Cowles, D., 2005. *Mytilus trossulus* fact sheet [WWW Document]. URL http://www.wallawalla.edu/academics/departments/biology/rosario/inverts/Mollusca/Bivalvia/Mytiloida/Mytilidae/*Mytilus_trossulus*.html (accessed 12.18.12).

Cullingham, C. I., James, P. M. A., Cooke, J. E. K., Coltman, D. W., 2012. Characterizing the physical and genetic structure of the lodgepole pine × jack pine hybrid zone: mosaic structure and differential introgression. Evolutionary Applications, 5(8), 879–891.

Daguin, C., Bonhomme, F., Borsa, P., 2001. The zone of sympatry and hybridization of *Mytilus edulis* and *M. galloprovincialis*, as described by intron length polymorphism at locus mac-1. Heredity (Edinb).

Dare, P., Edwards, D., Davies, G., 1983. Experimental collection and handling of spat mussels (Mytiluls *edulis* L.) on ropes for intertidal cultivation.

Dare, P.J., 1980. Mussel Cultivation in England and Wales. Crown Copyright.

Davey, J.L., Blaxter, M.W., 2010. RADseq: Next-generation population genetics. Brief. Funct. Genomics 9, 416–423.

Davey, J.W., Hohenlohe, P., Etter, P., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499–510.

Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., Reid, G., 2013. High throughput sequencing methods and analysis for microbiome research. J. Microbiol. Methods 95, 401–414.

Dias, P., Sollelis, L., Cook, E.J., Piertney, S.B., Davies, I.M., Snow, M., 2008. Development of a real-time PCR assay for detection of *Mytilus* species specific alleles: Application to a sampling survey in Scotland. J. Exp. Mar. Bio. Ecol. 367, 253–258.

Dias, P., Bland, M., Shanks, A., Beaumont, A. R., Piertney, S., Davies, I., Snow, M., 2009a. *Mytilus* species under rope culture in Scotland: implications for management. Aquaculture International, 17, 437–448.

Dias, P., Batista, F.M., Shanks, A., Beaumont, A.R., Davies, I., Snow, M., 2009b. Gametogenic asynchrony of mussels *Mytilus* in a mixed-species area: Implications for management. Aquaculture 295, 175–182.

Dias, P. J., Dordor, A., Tulett, D., Piertney, S., Davies, I. M., Snow, M., 2009c. Survey of mussel (*Mytilus*) species at Scottish shellfish farms. Aquaculture Research, 40(15), 1715–1722.

Dias, P., Piertney, S., Snow, M., Davies, I., 2011a. Survey and management of mussel *Mytilus* species in Scotland. Hydrobiologia 670, 127–140.

Dias, P., Malgrange, B., Snow, M., & Davies, I., 2011b. Performance of Mussels, *Mytilus edulis*, *Mytilus trossulus*, and Their Hybrids in Cultivation at Three Scottish Lochs. Journal of the World Aquaculture Society, 42(1), 111–121.

Dobzhansky T., 1935. A critique of the species concept in biology. Philos Sci 2:344–355.

Doherty, S.D., Brophy, D., Gosling, E., 2009. Synchronous reproduction may facilitate introgression in a hybrid mussel (*Mytilus*) population. J. Exp. Mar. Bio. Ecol. 378, 1–7.

Drummond, A. J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7(1), 214.

Dudu, A., Suciu, R., Paraschiv, M., Georgescu, S. E., Costache, M., Berrebi, P., 2011. Nuclear markers of Danube sturgeons hybridization. International Journal of Molecular Sciences, 12, 6796–6809.

Dufresne, F., Stift, M., Vergilino, R., Mable, B. K., 2014. Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. Molecular Ecology, 23(1), 40–69.

Edgcomb, V., Kysela, D., Teske, A., de Vera Gomez, A., Sogin, M., 2002. Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. Proceedings of the National Academy of Sciences of the United States of America, 99(11), 7658–7662.

Edwards, A., Edelsten, D., 1977. Deep water renewal of Loch Etive: A three basin Scottish fjord. Estuar. Coast. Mar. Sci. 5, 575–595.

Edwards, C.A., Skibinski, D.O.F., 1987. Genetic variation of mitochondrial DNA in mussel (*Mytilus edulis* and *M. galloprovincialis*) populations from South West England and South Wales. Mar. Biol. 94, 547–556.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and cross-validation. American Statistician, 37, 36–48.

Ekblom, R., Galindo, J., 2011. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity (Edinb). 107, 1–15.

Endler, J., 1977. Geographic variation, speciation and clines. Princeton University Press.

Etter, P., Bassham, S., Hohenlohe, P., Johnson, E., Cresko, W., 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol. Biol., Methods in Molecular Biology 772.

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.

Falush, D., Stephens, M., Pritchard, J. K., 2003. Inference of population structure using multilocus genotype data: Linked loc and correlated allele frequencies. Genetics, 164(4), 1567–1587.

Falush, D., Stephens, M., Pritchard, J. K., 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Molecular Ecology Notes, 7(4), 574–578.

FAO. 2012a. FAO Fisheries and Aquaculture Department information on *Mytilus edulis*. http://www.fao.org/fishery/species/2688/en (Accessed December 18, 2012).

FAO. 2012b. FAO Fisheries and Aquaculture Department information on *Mytilus galloprovincialis*. http://www.fao.org/fishery/species/3529/en (Accessed December 18, 2012).

Felsenstein, J., 1985. Confidence Limits on Phylogenies : An Approach Using the Bootstrap. Evolution, 39(4), 783–791.

Fisher RA: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 1936, 7:179-188.

Fly, E., Hilbish, T., 2013. Physiological energetics and biogeographic range limits of three congeneric mussel species. Oecologia 172, 35–46.

Fogelqvist, J., Verkhozina, A., Katyshev, A., Pucholt, P., Dixelius, C., Rönnberg-Wästljung, A., Lascoux, M., Berlin, S., 2015. Genetic and morphological evidence for introgression between three species of willows. BMC Evol. Biol. 15, 193.

Folmer, O., Black, M., Hoeh, W., Lutz, R., Vrijenhoek, R., 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology, 3(5), 294–299.

Fraïsse, C., Belkhir, K., Welch, J. J., Bierne, N., 2015. Local interspecies introgression is the main cause of extreme levels of intraspecific differentiation in mussels. Molecular Ecology (2015), 1-18.

Fraley, C., Raftery, A. E., 1998. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal, 41(8), 578–588.

Frantz, A. C., Cellina, S., Krier, A., Schley, L., Burke, T., 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: Clusters or isolation by distance? Journal of Applied Ecology, 46(2), 493–505.

Gabriel, S., Ziaugra, L., Tabbaa, D., 2009. SNP genotyping using the sequenom massARRAY iPLEX Platform. Current Protocols in Human Genetics, (SUPPL. 60), 1–18.

Gage, J., 1972. A preliminary survey of the benthic macrofauna and sediments in Lochs Etive and Creran, sea-lochs along the west coast of Scotland. J. Mar. Biol. Assoc. UK 237–276.

Gardeström, J., Pereyra, R., André, C., 2007. Characterization of six microsatellite loci in the Baltic blue mussel *Mytilus trossulus* and cross-species amplification in North Sea *Mytilus edulis*. Conserv. Genet. 9, 1003–1005.

Gérard, K., Bierne, N., Borsa, P., Chenuil, A., Féral, J. P., 2008. Pleistocene separation of mitochondrial lineages of *Mytilus* spp. mussels from Northern and Southern Hemispheres and strong genetic differentiation among southern populations. Molecular Phylogenetics and Evolution, 49(1), 84–91.

Giantsis, I. A., Mucci, N., Randi, E., Abatzopoulos, T. J., Apostolidis, A. P., 2014. Microsatellite variation of mussels (*Mytilus galloprovincialis*) in central and eastern Mediterranean: genetic panmixia in the Aegean and the Ionian Seas. Journal of the Marine Biological Association of the United Kingdom, 94(4), 797–809.

Gibson, G., Dworkin, I., 2004. Uncovering cryptic genetic variation. Nat. Rev. Genet. 5, 681–690.

Gilg, M., Hilbish, T., 2003. Patterns of larval dispersal and their effect on the maintenance of a blue mussel hybrid zone in southwestern England. Evolution 57(5), 1061–1077.

Glover, K., Pertoldi, C., Besnier, F., Wennevik, V., Kent, M., Skaala, Ø., 2013. Atlantic salmon populations invaded by farmed escapees: quantifying genetic introgression with a Bayesian approach and SNPs. BMC Genet. 14, 74.

Goffredi, S.K., Hurtado, L. A., Hallam, S., Vrijenhoek, R.C., 2003. Evolutionary relationships of deep-sea vent and cold seep clams (Mollusca: Vesicomyidae) of the "'Pacifica/Lepta'" species complex. Mar. Biol. 142, 311–320.

Gonen, S., Lowe, N.R., Cezard, T., Gharbi, K., Bishop, S.C., Houston, R.D., 2014. Linkage maps of the Atlantic salmon (Salmo salar) genome derived from RAD sequencing. BMC Genomics 15, 166.

González, J.R., Armengol, L., Solé, X., Guinó, E., Mercader, J.M., Estivill, X., Moreno, V., 2007. SNPassoc: An R package to perform whole genome association studies. Bioinformatics 23, 644–645.

Goodwin, S., McPherson, J.D., McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 17, 333–351.

Gormley, K., Mackenzie, C., Robins, P., Coscia, I., Cassidy, A., James, J., Porter, J., 2015. Connectivity and dispersal patterns of protected biogenic reefs: Implications for the conservation of *Modiolus modiolus* (L.) in the Irish Sea. PLoS ONE, 10(12), 1–17.

Gosling, E., 1992. The mussel *Mytilus*: ecology, physiology, genetics and culture. Developments in Aquaculture and Fisheries Science, 25.

Gosling, E., 1984. The systematic status of *Mytilus galloprovincialis* in western Europe: a review. Malacologia 25, 551–568.

Gosling, E., Doherty, S., Howley, N., 2008. Genetic characterization of hybrid mussel (*Mytilus*) populations on Irish coasts. J. Mar. Biol. Assoc. UK 88, 341–346.

Goulletquer, P., Moine, O. Le., 2002. Shellfish farming and Coastal Zone Management (CZM) development in the Marennes-Oléron Bay and Charentais Sounds (Charente Maritime, France): A review of recent developments. Aquaculture International, 10(6), 507–525.

Gross, B., Rieseberg, L., 2005. The ecological genetics of homoploid hybrid speciation. J. Hered. 96, 241–252.

Gubbins, M., 2012. *Mytilus trossulus* : Managing Impact On Sustainable Mussel Production In Scotland.

Gut, I.G., 2001. Automation in genotyping of single nucleotide polymorphisms. Hum. Mutat. 17, 475–492.

Hansen, M., Nielsen, E., Bekkevold, D., Mensberg, K., 2001. Admixture analysis and stocking impact assessment in brown trout (*Salmo trutta*), estimated with incomplete baseline data. Can. J. Fish. Aquat. Sci. 58, 1853–1860.

Harismendy, O., Ng, P., Strausberg, R., Wang, X., Stockwell, T., Beeson, K., Schork, N., Murray, S., Topol, E., Levy, S., Frazer, K., 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 10:R32.

Harrison RG. 1990. Hybrid zones: windows on evolutionary process. In: Futuyma D, Antonovics J, editors. Oxford surveys in evolutionary biology. Vol. 7. New York: Oxford University Press. p. 69–128

Harrison, R.G., Larson, E.L., 2014. Hybridization, introgression, and the nature of species boundaries. J. Hered. 105, 795–809.

Hartl, D.L, Clark, A.G., 1997. Principles of population genetics, 3rd edn. Sunderland, MA: Sinauer Associates, Inc.

Hartl, D. L., Grant, A., 2007. Principles of Population Genetics. 4th Edition. Sinauer Associates.

Hatfield, T., Schluter, D., 1999. Ecological Speciation in Sticklebacks : Environment-Dependent Hybrid Fitness. Evolution (N. Y). 53, 866–873.

Hayden, M.J., Nguyen, T.M., Waterman, A., Chalmers, K.J., 2008. Multiplex-Ready PCR: A new method for multiplexed SSR and SNP genotyping. BMC Genomics 9, 80.

Haynes, G., Gongora, J., Gilligan, D., Grewe, P., Moran, C., Nicholas, F., 2012. Cryptic hybridization and introgression between invasive Cyprinid species Cyprinus carpio and Carassius auratus in Australia: Implications for invasive species management. Anim. Conserv. 15, 83–94.

Heath, D.D., Rawson, P.D., Hilbish, T.J., 1995. PCR-based nuclear markers identify alien blue mussel *Mytilus* spp. genotypes on the west coast of Canada. Can. J. Fish. Aquat. Sci. 52, 2621–2627.

Heled, J., Drummond, A. J., 2010. Bayesian Inference of Species Trees from Multilocus Data. Molecular Biology and Evolution, 27(3), 570–580.

Helm, M., Bourne, N., Lovatelli, A., 2004. Hatchery Culture of Bivalves. A Practical Manual. FAO.

Hepper, B., 1957. Notes on *Mytilus galloprovincialis* Lamarck in Great Britain. J. Mar. Biol. Assoc. UK 36, 33–40.

Hickman, R.W., 1992. Mussel cultivation, in: Gosling, E. (Ed.) The mussel *Mytilus*: ecology, physiology, genetics and culture. pp. 465-510.

Hilbish, T. J., 1985. Demographic and temporal structure of an allele frequency cline in the mussel *Mytilus edulis*. Marine Biology, 86(2), 163–171.

Hilbish, T. J., Carson, E. W., Plante, J. R., Weaver, L. A., Gilg, M. R., 2002. Distribution of *Mytilus edulis*, *M. galloprovincialis*, and their hybrids in open-coast populations of mussels in southwestern England. Marine Biology, 140(1), 137–142.

Hilbish, T., Timmons, J., Agrawal, V., Schneider, K.R., Gilg, M.R., 2003. Estuarine habitats protect hybrid mussels from selection. J. Exp. Mar. Bio. Ecol. 292, 177–186.

Hindar, K., Fleming, I.A., McGinnity, P., Diserud, O., 2006. Genetic and ecological effects of salmon farming on wild salmon: modelling from experimental results. ICES J. Mar. Sci. 63, 1234–1247.

Hohenlohe, P., Amish, S., Catchen, J., Allendorf, F., Luikart, G., 2011. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol. Ecol. Resour. 11 Suppl 1, 117–22.

Hohenlohe, P., Bassham, S., Etter, P., Stiffler, N., Johnson, E., Cresko, W., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet. 6, e1000862.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441.

Houston, R., Davey, J., Bishop, S., Lowe, N., Mota-Velasco, J., Hamilton, A., Guy, D., Tinch, A., Thomson, M., Blaxter, M., Gharbi, K., Bron, J., Taggart, J., 2012. Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics 13, 244.

Howe, J., Shimmield, T., Austin, W., Longva, O., 2002. Post-glacial depositional environments in a mid-high latitude glacially-overdeepened sea loch, inner Loch Etive, western Scotland. Mar. Geol. 185, 417–433.

Hubisz, M. J., Falush, D., Stephens, M., Pritchard, J. K.., 2009. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources, 9(5), 1322–1332.

Huff, S. W., Campbell, D. C., Gustafson, D. L., Lydeard, C., Altaba, C. R., 2002. Investigations into the phylogenetic relationships of the threatened freshwater pearl-mussels (Bivalvia, Unionoidea, Margaritiferidae) based on molecular data: implications for their taxonomy and biogeography. Journal of Molluscan Studies, 70(1), 379–388.

Hussain, M.G. Mazid, M.A., 2001. Genetic improvement and conservation of carp species in Bangladesh. Bangladesh Fisheries Research Institute, Mymensingh, Bangladesh and ICLARM, Penang Malaysia, 74p.

Huxel, G.R., 1999. Rapid displacement of native species by invasive species: Effects of hybridization. Biol. Conserv. 89, 143–152.

Hvilsom, M.M., Theisen, B.F., 1984. Inheritance of allozyme variations through crossing experiments with the blue mussel, *Mytilus edulis* L., 7, 1–7.

Innes, D.J., Bates, J.A., 1999. Morphological variation of *Mytilus edulis* and *Mytilus trossulus* in eastern Newfoundland. Mar. Biol. 133, 691–699.

Inoue, K., Odo, S., Noda, T., Nakao, S., Takeyama, S., Yamaha, E., Yamazaki, F., Harayama, S., 1997. A possible hybrid zone in the *Mytilus edulis* complex in Japan revealed by PCR markers. Mar. Biol. 128, 91–95.

Inoue, K., Waite, J., Matsuoka, M., Odo, S., Harayama, S., 1995. Interspecific Variations in Adhesive Protein Sequences of *Mytilus edulis*, *M. galloprovincialis*, and *M. trossulus*. Biol. Bull. 189, 370–375.

Jackson, J.E., 1991. A User's Guide To Principal Components. Wiley-Interscience.

Jaenisch, R., Bird, A., 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. 33 Suppl, 245–254.

Jasper, J.R., Habicht, C., Moffitt, S., Brenner, R., Marsh, J., Lewis, B., Fox, E.C., Grauvogel, Z., Olive, S.D.R., Grant, W.S., 2013. Source-sink estimates of genetic introgression show influence of hatchery strays on wild chum salmon populations in Prince William Sound, Alaska. PLoS One 8.

Jensen, F., Patursson, E., 2011. Blue Mussel (*Mytilus edulis*) in Faroese Fjords : Biology and Farming Potential. University of the Faroe Islands.

Jiggins, C.D., Mallet, J., 2000. Bimodal hybrid zones and speciation. Trends Ecol. Evol. 15, 250–255.

Jombart, T., 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–5.

Jombart, T., Pontier, D., Dufour, A.-B., 2009. Genetic markers in the playground of multivariate analysis. Heredity, 102, 330–41.

Jombart, T., Devillard, S., Balloux, F., 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11, 94.

Jones, P., Osborn, T., Briffa, K.R., 2001. The evolution of climate over the last millennium. Science 292, 662–7.

Jorgenson, J., Lukacs, K., 1983. Capillary zone electrophoresis. Science, 222(4621), 266–272.

Kai, W., Nomura, K., Fujiwara, A., Nakamura, Y., Yasuike, M., Ojima, N., Masaoka, T., Ozaki, A., Kazeto, Y., Gen, K., Nagao, J., Tanaka, H., Kobayashi, T., Ototake, M., 2014. A ddRAD-based genetic map and its integration with the genome assembly of Japanese eel (*Anguilla japonica*) provides insights into genome evolution after the teleost-specific genome duplication. BMC Genomics 15, 233.

Karayücel, S., Karayücel, I., 1999. Growth and Mortality of Mussels (*Mytilus edulis* L.) Reared in Lantern Nets in Loch Kishorn, Scotland. Turkish J. Vet. Anim. Sci. 23, 397–402.

Karger, B. L., Guttman, A., 2009. DNA Sequencing by Capillary Electrophoresis. National Institute of Health, 30 (Suppl 1), 1–11.

Katolikova, M., Khaitov, V., Vainola, R., Gantsevich, M., Strelkov, P., 2016. Genetic, ecological and morphological distinctness of the blue mussels *Mytilus trossulus* Gould and *M. edulis* l. in the White Sea. PLoS ONE, 11(4), 1–25.

Kenchington, E., Hamilton, L., Cogswell, A., Zouros, E., 2009. Paternal mtDNA and Maleness Are Co-Inherited but Not Causally Linked in Mytilid Mussels. PLoS One 4, 13.

Key, K., 1968. The Concept of Stasipatric Speciation. Syst. Zool. 17, 14–22.

Kijewski, T., Zbawicka, M., Väinölä, R., Wenne, R., 2006. Introgression and mitochondrial DNA heteroplasmy in the Baltic populations of mussels *Mytilus trossulus* and *M. edulis*. Mar. Biol. 149, 1371–1385.

Kim, S.C., Rieseberg, L., 1999. Genetic architecture of species differences in annual sunflowers: implications for adaptive trait introgression. Genetics 153, 965–77.

Kircher, M., Kelso, J., 2010. High-throughput DNA sequencing - Concepts and limitations. BioEssays 32, 524–536.

Kirpichnikov, V.S., J.I. Ilyasov, L.A. Shart, A.A. Vikhman, M.V. Ganchenko,A.L. Ostashevsky,V.M. Simonov,G.F. Tikhonovand V.V.Tjurin. 1993. Selection of Krasnodar common carp (Cyprinus carpio L.) for resistance to dropsy: principal results and prospects. Aquaculture 111:7-20.

Knowlton, N., 2000. Molecular genetic analyses of species boundaries in the sea. Hydrobiologia 420, 73–90.

Koehn, R., 1991. The genetics and taxonomy of species in the genus *Mytilus*. Aquaculture 94, 125–145.

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, (March), 1179–1191.

Kothera, L., Nelms, B. M., Reisen, W. K., Savage, H. M., 2013. Population genetic and admixture analyses of *Culex pipiens* complex (Diptera: Culicidae) populations in California, United States. American Journal of Tropical Medicine and Hygiene, 89(6), 1154–1167.

Kovach, A. I., Walsh, J., Ramsdell, J., Thomas, W. K., 2015. Development of diagnostic microsatellite markers from whole-genome sequences of *Ammodramus* sparrows for assessing admixture in a hybrid zone. Ecology and Evolution, 5(11), 2267–2283.

Kruse, C., Hubert, W., Rahel, F., 2000. Status of Yellowstone Cutthroat Trout in Wyoming Waters. North Am. J. Fish. Manag. 20, 693–705.

Lachenbruch, P., Goldstein, M., 1979. Discriminant Analysis. Biometrics, 35(1), 69–85.

Lal, M. M., Southgate, P. C., Jerry, D. R., Zenger, K. R., 2016. Fishing for divergence in a sea of connectivity: The utility of ddRADseq genotyping in a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. Marine Genomics, 25, 57–68.

Lal, M. M., Southgate, P. C., Jerry, D. R., Bosserelle, C., Zenger, K. R., 2017. Swept away: ocean currents and seascape features influence genetic structure across the 18,000 Km Indo-Pacific distribution of a marine invertebrate, the black-lip pearl oyster *Pinctada margaritifera*. BMC Genomics, 18(1), 66.

Lallias, D., Stockdale, R., Boudry, P., Lape, S., Beaumont, A., 2009. Characterization of ten microsatellite Loci in the blue mussel *Mytilus edulis*. J. Shellfish Res. 28, 547–551.

Larraín, M. A., Díaz, N. F., Lamas, C., Uribe, C., Jilberto, F., Araneda, C., 2015. Heterologous microsatellite-based genetic diversity in blue mussel (*Mytilus chilensis*) and differentiation among localities in southern Chile. Latin American Journal of Aquatic Research, 43(5), 998–1010.

Larson, E.L., Guilherme Becker, C., Bondra, E.R., Harrison, R.G., 2013. Structure of a mosaic hybrid zone between the field crickets *Gryllus firmus* and *G. pennsylvanicus*. Ecol. Evol. 3, 985–1002.

Lartillot, N., Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution, 21(6), 1095–1109.

Latch, E. K., Dharmarajan, G., Glaubitz, J. C., Rhodes, O. E., 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics, 7(2), 295–302.

LeBlanc, N., Landry, T., Stryhn, H., Tremblay, R., McNiven, M., Davidson, J., 2005. The effect of high air and water temperature on juvenile *Mytilus edulis* in Prince Edward Island, Canada. Aquaculture, 243(1–4), 185–194.

Lee, C., Abdool, A., Huang, C.-H., 2009. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics, 10(Suppl 1):S73.

Lee, T., Foighil, D.Ó., 2004. Hidden Floridian biodiversity: Mitochondrial and nuclear gene trees reveal four cryptic species within the scorched mussel, *Brachidontes exustus*, species complex. Mol. Ecol. 13, 3527–3542.

Leggett, R., MacLean, D., 2014. Reference-free SNP detection: dealing with the data deluge. BMC Genomics 15, S10.

Lemey, P., Salemi, M., Vamdamme, A.-M., 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing Second Edition, Book. Cambridge University Press.

Lessios, H.A., 1992. Testing eletrophoretic data for agreement with Hardy-Weinberg expectations. Mar. Biol. 112, 517–523.

Lewis, J.R., Powell, H.T., 1961. The occurrence of curved and ungulate forms of the mussel *Mytilus edulis* L. in the British Isles and their relationship to *M. galloprovincialis* Lamarck. J. Zool. 137, 583–598.

Li, H., Liang, Y., Sui, L., Gao, X., He, C., 2011. Characterization of 10 polymorphic microsatellite markers for Mediterranean blue mussel *Mytilus galloprovincialis* by EST database mining and cross-species amplification. J. Genet. 90, 30–33.

Linnen, C.R., Hoekstra, H.E., 2009. Measuring natural selection on genotypes and phenotypes in the wild. Cold Spring Harb. Symp. Quant. Biol. 74, 155–68.

Liu, Z., Cordes, J., 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture 238, 1–37.

Lopez, G. A., Potts, B. M., Tilyard, P. A., 2000. F1 hybrid inviability in Eucalyptus: the case of *E. ovata*, *x E. globulus*. Heredity, 85(3), 242–250.

Low, Y.L., Wedren, S., Liu, J., 2006. High-throughput genomic technology in research and clinical management of breast cancer. Evolving landscape of genetic epidemiological studies. Breast Cancer Res. 8, 209.

Lydeard, C., Mulvey, M., Davis, G. M., 1996. Molecular systematics and evolution of reproductive traits of North American freshwater unionacean mussels (Mollusca: Bivalvia) as inferred from 16S rRNA gene sequences. Philosophical Transactions of the Royal Society of London B Biological Sciences, 351(1347), 1593–1603.

Mallet, J., 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20, 229–237.

Maloy, A., 2001. Gametogenic Cycles of Marine Mussels, *Mytilus edulis* and *Mytilus trossulus* in Cobscook Bay, Maine.

Maloy, A.P., Barber, B.J., Rawson, P.D., 2003. Gamentogenesis in a sympatric population of blue mussels, *Mytilus edulis* and *Mytilus trossulus*, from Cobscook Bay (USA). J. Shellfish Res. 22, 119–123.

Manel, S., Gaggiotti, O. E., Waples, R. S., 2005. Assignment methods: Matching biological questions with appropriate techniques. Trends in Ecology and Evolution, 20(3), 136–142.

Marie, A. D., Bernatchez, L., Garant, D., 2011. Empirical assessment of software efficiency and accuracy to detect introgression under variable stocking scenarios in brook charr (*Salvelinus fontinalis*). Conservation Genetics, 12(5), 1215–1227.

Marie, A., Bernatchez, L., Garant, D., 2012. Environmental factors correlate with hybridization in stocked brook charr (*Salvelinus fontinalis*). Canadian Journal of Fisheries and Aquatic Sciences, 69, 884–893.

Marine Life Information Network (MarLIN), 2006. BIOTIC - Biological Traits Information Catalogue. United, Marine Life Information Network. Plymouth: Marine Biological Association of the United Kingdom [WWW Document], n.d. URL http://www.marlin.ac.uk/biotic/browse.php?sp=4250 (accessed 9.16.16).

Marques, I., Rosselló-Graell, A., Draper, D., Iriondo, J., 2007. Pollination patterns limit hybridization between two sympatric species of Narcissus (Amaryllidaceae). Am. J. Bot. 94, 1352–1359.

Mathiesen, S. S., Thyrring, J., Hemmer-Hansen, J., Berge, J., Sukhotin, A., Leopold, P., Nielsen, E. E., 2016. Genetic diversity and connectivity within *Mytilus* spp. in the subarctic and Arctic. Evolutionary Applications 10 (1), 39–55.

Matsuoka, Y., Vigouroux, Y., Goodman, M. M., Sanchez G, J., Buckler, E., Doebley, J., 2002. A single domestication for maize shown by multilocus microsatellite genotyping. Proceedings of the National Academy of Sciences of the United States of America, 99(9), 6080–6084.

Mayes, A., Fraser, D., 2012. Scottish Shellfish Farm Production Survey 2011 report.

Mayr E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Cambridge (MA): Harvard University Press.

McCartney, M., Lima, T., 2011. Evolutionary consequences of introgression at m7 lysin, a gamete recognition locus, following secondary contact between blue mussel species. Integr. Comp. Biol. 51, 474–484.

McDonald, J., Seed, R., Koehn, R., 1991. Allozymes and morphometric characters of three species of *Mytilus* in the Northern and Southern Hemispheres. Mar. Biol. 111, 323–333.

Mckean, N., Trewick, S., Morgan-Richards, M., 2016. Little or no gene flow despite F1 hybrids at two interspecific contact zones. Ecol. Evol. 6, 2390 – 2404.

McKindsey, C., Archambault, P., Callier, M., Olivier, F., 2011. Influence of suspended and off-bottom mussel culture on the sea bottom and benthic habitats: a review. Can. J. Zool. 89, 622–646.

Melzner, F., Stange, P., Trubenbach, K., Thomsen, J., Casties, I., Panknin, U., Gorb, S., Gutowska, M., 2011. Food supply and seawater pCO 2 impact calcification and internal shell dissolution in the blue mussel *Mytilus edulis*. PLoS One 6.

Mia, M., Taggart, J., Gilmour, A., Gheyas, A., Das, T., Kohinoor, A., Rahman, M., Sattar, M., Hussain, M., Mazid, M., Penman, D., McAndrew, B., 2005. Detection of hybridization between Chinese carp species (*Hypophthalmichthys molitrix* and *Aristichthys nobilis*) in hatchery broodstock in Bangladesh, using DNA microsatellite loci. Aquaculture 247, 267–273.

Miles, C., Wayne, M., 2008. Quantitative Trait Locus (QTL) Analysis. Nat. Educ. 1, 208.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W. A, Johnson, E., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17, 240–248.

Minoche, A., Dohm, J., Himmelbauer, H., 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biol. 12, R112.

Miranda, M.B.B., Innes, D.J., Thompson, R.J., 2010. Incomplete reproductive isolation in the blue mussel (*Mytilus edulis* and *M. trossulus*) hybrid zone in the Northwest Atlantic: role of gamete interactions and larval viability. Biol. Bull. 218, 266–281.

Molnar, J.L., Gamboa, R.L., Revenga, C., Spalding, M.D., 2008. Assessing the global threat of invasive species to marine biodiversity. Front. Ecol. Environ. 6, 485–492.

Monteiro, C., Serrão, E., Pearson, G., 2012. Prezygotic barriers to hybridization in marine broadcast spawners: Reproductive timing and mating system variation. PLoS One 7.

Morin, P.A., Luikart, G., Wayne, R.K., 2004. SNPs in ecology, evolution and conservation. Trends Ecol. Evol. 19, 208–216.

Morozova, O., Marra, M., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92, 255–264

Muñoz-Diez, C., Vitte, C., Ross-Ibarra, J., Gaut, B., Tenaillon, M., 2012. Using Nextgen Sequencing to Investigate Genome Size Variation and Transposable Element Content, in: Plant Transposable Elements. Springer Berlin / Heidelberg, pp. 41–58.

Munro, L., Wallace, I., 2015. Marine Scotland Science Scottish Shellfish Farm Production Survey 2014.

Munro, L., Wallace, I., 2016. Marine Scotland Science Scottish Shellfish Farm Production Survey 2015.

Naisbit, R.E., Jiggins, C.D., Mallet, J., 2001. Disruptive sexual selection against hybrids contributes to speciation between *Heliconius cydno* and *Heliconius melpomene*. Proc. R. Soc. London B Biol. Sci. 268, 1849–1854.

Naylor, R., Williams, S. L., Strong, D. R., 2001. Aquaculture - A gateway for exotic species. Science, 294 (November), 1655–1666.

Newell, R., 1989. Species Profiles : Life Histories and Environmental Requirements of Coastal Fishes and Invertebrates (North and Mid-Atlantic): BLUE MUSSEL. U.S. Fish and Wildlife Service.

Nielsen, E. E., Hansen, M. M., Ruzzante, D. E., Meldrup, D., Grønkjær, P., 2003. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. Molecular Ecology, 12(6), 1497–1508.

O'Hara, R. B., Cano, J. M., Ovaskainen, O., Teplitsky, C., Alho, J. S., 2008. Bayesian approaches in evolutionary quantitative genetics. Journal of Evolutionary Biology, 21(4), 949–957.

Palaiokostas, C., Bekaert, M., Khan, M.G.Q., Taggart, J.B., Gharbi, K., McAndrew, B.J., Penman, D.J., 2013a. Mapping and validation of the major sex-determining region in Nile tilapia (*Oreochromis niloticus* L.) Using RAD sequencing. PLoS One 8, e68389.

Palaiokostas, C., Bekaert, M., Davie, A., Cowan, M.E., Oral, M., Taggart, J.B., Gharbi, K., McAndrew, B.J., Penman, D.J., Migaud, H., 2013b. Mapping the sex determination locus in the Atlantic halibut (*Hippoglossus hippoglossus*) using RAD sequencing. BMC Genomics 14, 566.

Palumbi, S., 1992. Marine speciation on a small planet. Trends in Ecology and Evolution 7(4), 114–118.

Palumbi, S., 1994. Genetic divergence, reproductive isolation, and marine speciation. Annu. Rev. Ecol. Syst. 25, 547–572.

Patel, S., Schell, T., Eifert, C., Feldmeyer, B., Pfenninger, M., 2015. Characterizing a hybrid zone between a cryptic species pair of freshwater snails. Mol. Ecol. 24, 643–655.

Pawiro, S., 2010. Bivalves : Global production and trade trends, in: Rees, G., Pond, K., Kay, D., Bartram, J., Santo Domingo, J. (Eds.), Safe Managemnet of Shellfish and Harvest Waters. IWA Publishing, London, p. 360.

Peakall, R. and Smouse P. 2006. GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6, 288-295.

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

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(1), 559–572.

Peñaloza, C., Bishop, S., Toro, J., Houston, R., 2006. RAD Sequencing reveals genome-wide heterozygote deficiency in pair crosses of the Chilean mussel *Mytilus* spp. 10th World Congress of Genetics Applied to Livestock Production, 2–4.

Peñarrubia, L., Sanz, N., Pla, C., Vidal, O., Viñas, J., 2015. Using Massive Parallel Sequencing for the Development, Validation, and Application of Population Genetics Markers in the Invasive Bivalve Zebra Mussel (*Dreissena polymorpha*). Plos One, 10(3), e0120732.

Penney, R. W., Hart, M. J., Templeman, N., 2002. Comparative growth of cultured blue mussels, *Mytilus edulis*, M-*trossulus* and their hybrids, in naturally occurring mixed-species stocks. Aquaculture Research, 33, 693–702.

Perkel, J., 2008. SNP genotyping: six technologies that keyed a revolution. Nat. Methods 5, 575–575.

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.

Philipp, E., Kraemer, L., Melzner, F., Poustka, A., Thieme, S., Findeisen, U., Schreiber, S., Rosenstiel, P., 2012. Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan *Mytilus edulis*. PLoS One 7.

Phillips, P., 2008. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. 9, 855–867.

Poly, W.J., 1997. Characteristics of an Intergeneric Cyprinid Hybrid, *Campostoma anomalum x Luxilus* sp. indet. (Pisces: Cyprinidae), from the Portage River, Ohio. Ohio J. Sci. 97, 40–43.

Prado, F., Hashimoto, D., Senhorini, J., Foresti, F., Porto-Foresti, F., 2012. Detection of hybrids and genetic introgression in wild stocks of two catfish species (Siluriformes: Pimelodidae): The impact of hatcheries in Brazil. Fish. Res. 125-126, 300–305.

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

Pritchard, J., Wen, W., 2003. Documentation for STRUCTURE software: Version 2. Available from http://www.pritch.bsd. uchicago.edu.

Pritchard, V., Jones, K., Cowley, D., 2007. Estimation of introgression in cutthroat trout populations using microsatellites. Conserv. Genet. 8, 1311–1329.

Pryor, L., 1951. A genetic analysis of some Eucalyptus species. Proc. Linn. Soc. New South Wales 76, 140–148.

Pujolar, J.M., Jacobsen, M.W., Als, T.D., Frydenberg, J., Magnussen, E., Jónsson, B., Jiang, X., Cheng, L., Bekkevold, D., Maes, G.E., Bernatchez, L., Hansen, M.M., 2014. Assessing patterns of hybridization between North Atlantic eels using diagnostic single-nucleotide polymorphisms. Heredity (Edinb). 112, 627–637.

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., Bird, C. E., 2014. Demystifying the RAD fad. Molecular Ecology, 23(24), 5937–5942.

Qiu, J., Tremblay, R., Bourget, E., 2002. Ontogenetic changes in hyposaline tolerance in the mussels *Mytilus edulis* and *M. trossulus*: Implications for distribution. Mar. Ecol. Prog. Ser. 228, 143–152.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rafalski, A., 2002. Applications of single nucleotide polymorphism in crop genetics. Curr. Opin. Plant Biol. 5, 94–100.

Rambaut, A., 2007. FigTree, phylogenetic tree viewer, software downloadable from http://tree.bio.ed.ac.uk/software/figtree/.

Rawson, P., Agrawal, V., Hilbish, T., 1999. Hybridization between the blue mussels *Mytilus galloprovincialis* and *M. trossulus* along the Pacific coast of North America: evidence for limited introgression. Mar. Biol. 134, 201–211.

Rawson, P., Hilbish, T., 1995. Distribution of male and female mtDNA lineages in populations of blue mussels, *Mytilus trossulus* and *M. galloprovincialis*, along the Pacific coast of North America. Mol. Biol. Evol. 124, 245–250.

Rawson, P., Joyner, K.L., Meetze, K., Hilbish, T., 1996. Evidence for intragenic recombination within a novel genetic marker that distinguishes mussels in the *Mytilus edulis* species complex. Heredity (Edinb). 77, 599–607.

Rawson, P., Slaughter, C., Yund, P., 2003. Patterns of gamete incompatibility between the blue mussels *Mytilus edulis* and *M. trossulus*. Mar. Biol. 143, 317–325.

Raymond M., Rousset F. 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Heredity, 86:248-249

Rieseberg, L., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A., Lexer, C., 2003. Major Ecological Transitions in Wild Sunflowers Facilitated by Hybridization. Science (80) 301, 1211–1216.

Rieseberg, L., Wendel, J.F., 1993. Introgression and its consequences in plants. Hybrid Zo. Evol. Process 70–10.

Riginos, C., Cunningham, C.W., 2005. Local adaptation and species segregation in two mussel (*Mytilus edulis* x *Mytilus trossulus*) hybrid zones. Mol. Ecol. 14, 381–400.

Riginos, C., Hickerson, M., Henzler, C., Cunnigham, C., 2004. Differential Patterns of Male and Female MtDNA Exchange Across the Atlantic Ocean in the Blue Mussel, *Mytilus edulis* 58, 2438–2451.

Rodríguez-Ramilo, S. T., Toro, M. A., Fernández, J., 2009. Assessing population genetic structure via the maximisation of genetic distance. Genetics, Selection, Evolution : GSE, 41, 49.

Roe, K. J., Lydeard C., 1998. Molecular systematics of the freshwater mussel genus *Potamilus* (Bivalvia, Unionidae). Malacologia, 39(1–2): 195–205.

Rosenberg, N. A., 2011. Genetic Structure of Human Populations. Science, 2381(2002).

Saarman, N. P., Pogson, G. H., 2015. Introgression between invasive and native blue mussels (genus *Mytilus*) in the central California hybrid zone. Molecular Ecology, 24(18), 4723–4738.

Salzberg, S.L., Yorke, J.A., 2005. Beware of mis-assembled genomes. Bioinformatics 21, 4320–4321.

Sanz, N., Araguas, R. M., Fernández, R., Vera, M., García-Marín, J. L., 2009. Efficiency of markers and methods for detecting hybrids and introgression in stocked populations. Conservation Genetics, 10(1), 225–236.

Schlötterer, C., 2004. The evolution of molecular markers--just a matter of fashion? Nat. Rev. | Genet. 5, 63–69.

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464.

Schwenk, K., Brede, N., Streit, B., 2008. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 363, 2805–2811.

SEAFISH, 2012. The Seafish Guide To Aquaculture.

Seed R, 1971. A physiological and biochemical approach to the taxonomy of *Mytilus edulis* L. and *M. galloprovincialis* LMK from SW England. Cah. Biol. Mar. 12, 291–322.

Seed, R., 1968. Factors influencing shell shape in the mussel *Mytilus edulis*. J. Mar. Biol. Assoc. UK 48, 561–584.

Seed, R., 1969. The ecology of *Mytilus edulis* L. (Lamellibranchiata) on exposed rocky shores - I. Breeding and settlement. Oecologia 3, 277–316.

Seehausen, O., 2013. Conditions when hybridization might predispose populations for adaptive radiation. J. Evol. Biol. 26, 279–281.

Semagn, K., Babu, R., Hearne, S., Olsen, M., 2013. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Mol. Breed. 33, 1–14.

Sharma, R., Goossens, B., Kun-Rodrigues, C., Teixeira, T., Othman, N., Boone, J.Q., Jue, N.K., Obergfell, C., O'Neill, R.J., Chikhi, L., 2012. Two different high throughput sequencing approaches identify thousands of de novo genomic markers for the genetically depleted Bornean elephant. PLoS One 7, e49533.

Skibinski, D., Ahmad, M., Beardmore, J., 1978. Genetic Evidence for Naturally Occurring Hybrids Between *Mytilus edulis* and *Mytilus galloprovincialis*. Evolution (N. Y). 32, 354–364.

Śmietanka, B., Zbawicka, M., Wołowicz, M., Wenne, R., 2004. Mitochondrial DNA lineages in the European populations of mussels (*Mytilus* spp.). Mar. Biol. 146, 79–92.

Śmietanka, B., Zbawicka, M., Sańko, T., Wenne, R., Burzyński, A., 2013. Molecular population genetics of male and female mitochondrial genomes in subarctic *Mytilus trossulus*. Mar. Biol. 160, 1709–1721.

Sobrino, B., Brión, M., Carracedo, A., 2005. SNPs in forensic genetics: A review on SNP typing methodologies. Forensic Sci. Int. 154, 181–194.
Sousa, V.C., Carneiro, M., Ferrand, N., Hey, J., 2013. Identifying Loci Under Selection Against Gene Flow in Isolation-with-Migration Models. Genetics 194, 211–233.

Spaak, P., Hoekstra, J., 1995. Life History Variation and the Coexistence of a Daphnia Hybrid With Its Parental Species. Ecology 76, 553–564.

Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics, 22(21), 2688–2690.

Stamatakis, A., 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Systematic Biology, 57(5), 758–71.

Stelzer, C.-P., Riss, S., Stadler, P., 2011. Genome size evolution at the speciation level: the cryptic species complex *Brachionus plicatilis* (Rotifera). BMC Evol. Biol. 11, 90.

Storey, J.D., Akey, J.M., Kruglyak, L., 2005. Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol. 3, e267.

Strimmer, K., von Haeseler, A., 2009. Section III: Phylogenetic inference. In The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing Second Edition (pp. 111–140). Cambridge University Press.

Stronen, A. V., Paquet, P.C., 2013. Perspectives on the conservation of wild hybrids. Biol. Conserv. 167, 390–395.

Suchanek, T.H., Geller, J.B., Kreiser, B.R., Mitton, J.B., 1997. Zoogeographic distributions of the sibling species *Mytilus galloprovincialis* and *M. trossulus* (Bivalvia: Mytilidae) and their hybrids in the North Pacific. Biol. Bull. 193, 187–194.

Suzuki, T., Nachman, M., 2015. Speciation and reduced hybrid female fertility in house mice. Evolution (N. Y). 69, 2468–2481.

Syvanen A-C., 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930– 942.

Tamelander, J., Riddering, L., Haag, F., Matheickal, J., 2010. Guidelines for development of a national ballast water management strategy. GEF-UNDP-IMO Glob. IUCN 45.

Tanguy, M., McKenna, P., Gauthier-Clerc, S., Pellerin, J., Danger, J.-M., Siah, A., 2013. Sequence analysis of a normalized cDNA library of *Mytilus edulis* hemocytes exposed to *Vibrio splendidus* LGP32 strain. Results Immunol. 3, 40–50.

The Aquaculture and Fisheries (Scotland) Act 2013 (Specification of Commercially Damaging Species) Order 2014.

Therriault, T., Docker, M., Orlova, M., Heath, D., MacIsaac, H., 2004. Molecular resolution of the family Dreissenidae (Mollusca: Bivalvia) with emphasis on Ponto-Caspian species, including first report of *Mytilopsis leucophaeata* in the Black Sea basin. Mol. Phylogenet. Evol. 30, 479–489.

Thresher, R.E., Kuris, A.M., 2004. Options for managing invasive marine species. Biol. Invasions 6, 295–300.

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., Bird, C. E., 2013. ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ* 1: e203.

Toro, J., Thompson, R.J., Innes, D.J., 2002. Reproductive isolation and reproductive output in two sympatric mussel species (*Mytilus edulis*, *M. trossulus*) and their hybrids from Newfoundland. Mar. Biol. 141, 897–909.

Toro, J., Innes, D.J., Thompson, R.J., 2003. Genetic variation among life-history stages of mussels in a *Mytilus edulis-M. trossulus* hybrid zone. Mar. Biol. 713–725.

Toro, J., Oyarzun, P., Penaloza, C., Alcapan, A., Videla, V., Tilleria, J., Astorga, M., Martinez, V., 2012. Production and performance of larvae and spat of pure and hybrid species of *Mytilus chilensis* and *M. galloprovincialis* from laboratory crosses. Lat. Am. J. Aquat. Res. 40, 243–247.

Tremblay, M.J., 2002. Large epibenthic invertebrates in the Bras d'Or Lakes. Proc. Nov. Scotian Inst. Sci. 42, 101–126.

Twyford, A.D., Ennos, R.A., 2012. Next-generation hybridization and introgression. Heredity (Edinb). 108, 179–189.

Utter, F., 2000. Patterns of subspecific anthropogenic introgression in two salmonid. Rev. Fish Biol. Fish. 10, 265–279.

Vähä, J. P., Primmer, C. R., 2006. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Molecular Ecology, 15(1), 63–72.

Van de Peer, Y., 2009. Section III: Phylogenetic inference. In The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing Second Edition (pp. 142–180). Cambridge University Press.

Vannarattanarat, S., Zieritz, A., Kanchanaketu, T., Kovitvadhi, U., Kovitvadhi, S., Hongtrakul, V., 2014. Molecular identification of the economically important freshwater mussels (Mollusca-Bivalvia-Unionoida) of Thailand: developing species-specific markers from AFLPs. Anim. Genet. 45, 235–239.

Varela, M.A., González-Tizón, A., Mariñas, L., Martínez-Lage, A., 2007. Genetic divergence detected by ISSR markers and characterization of microsatellite regions in *Mytilus* mussels. Biochem. Genet. 45, 565–578.

Varne, R., Kunz, K.L., Johansen, T., Westgaard, J.-I., Uglem, I., Mork, J., 2015. Farmed cod escapees and net-pen spawning left no clear genetic footprint in the local wild cod population. Aquac. Environ. Interact. 7, 253–266.

Varvio, S., Koehn, R., Vainola, R., 1988. Marine Biology Evolutionary genetics of the *Mytilus edulis* complex in the North Atlantic region. Mar. Biol. 98, 51–60.

Vera, M., Pardo, B., Pino-Querido, A., Alvarez-Dios, J., Fuentes, J., Martinez, P., 2010. Characterization of single-nucleotide polymorphism markers in the Mediterranean mussel, *Mytilus galloprovincialis*. Aquac. Res. 41, 568–575.

Vercaemer, B., Spence, K. R., Herbinger, C. M., Lapègue, S., Kenchington, E. L., 2006. Genetic Diversity of the European Oyster (*Ostrea edulis* L.) in Nova Scotia: Comparison With Other Parts of Canada, Maine and Europe and Implications for Broodstock Management. Journal of Shellfish Research, 25(2), 543–551.

Vermeij, G., 1991. Anatomy of an invasion: the trans-Arctic interchange. Paleobiology 17, 281–307.

Vignal, A., Milan, D., SanCristobal, M., Eggen, A., 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34, 275–305.

Walther, G.-R., Roques, A., Hulme, P., Sykes, M., Pys, P., Robinet, C., Semenchenko, V., 2009. Alien species in a warmer world : risks and opportunities 24, 686–693.

Wang S, Meyer E., McKay J. K., 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature methods, 9,808–810.

Warwick, T., Knight, A.J., Ward, R.D., 1990. Hybridisation in the *Littorina saxatilis* species complex (Prosobranchia: Mollusca). Hydrobiologia 193, 109–116.

Weising, K., Nybom, H., Pfenninger, M., Wolff, K., Kahl, G., 2005. DNA Fingerprinting in Plants: Principles, Methods, and Applications, Second Edition. CRC Press.

Wenne, R., Bach, L., Zbawicka, M., Strand, J., McDonald, J. H., 2016. A first report on coexistence and hybridization of *Mytilus trossulus* and *M. edulis* mussels in Greenland. Polar Biology, 39(2), 343–355.

Westfall, K.M., Wimberger, P.H., Gardner, J.P.A, 2010. An RFLP assay to determine if *Mytilus galloprovincialis* Lmk. (Mytilidae; Bivalvia) is of Northern or Southern hemisphere origin. Mol. Ecol. Resour. 10, 573–575.

Widdows, J., Johnson, D., 1988. Physiological energetics of *Mytilus edulis* : scope for growth. Mar. Ecol. - Prog. Ser. 46, 113–121.

Wilkinson, S., Haley, C., Alderson, L., Wiener, P., 2011. An empirical assessment of individual-based population genetic statistical techniques: application to British pig breeds. Heredity, 106(2), 261–9.

Willing, E.-M., Hoffmann, M., Klein, J.D., Weigel, D., Dreyer, C., 2011. Paired-end RAD-seq for de novo assembly and marker design without available reference. Bioinformatics 27, 2187–93.

Wilson, G. A., Rannala, B., 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics, 163(3), 1177–91.

Wonham, M.J., 2004. Mini-Review : Distribution of the Mediterranean mussel *Mytilus galloprovincialis* (Bivalvia: Mytiliade) and hybrids in the northeast Pacific. J. Shellfish Res. 23, 535–543.

Wood, A.R., Beaumont, A.R., Skibinski, D.O.F., Turner, G., 2003. Analysis of a nuclear DNA marker for species identification of adults and larvae in the *Mytilus edulis* complex. J. Mol. Stud. 69, 61–66.

Zardi, G. I., Nicastro, K. R., McQuaid, C. D., Castilho, R., Costa, J., Serrão, E. A., Pearson, G. A. 2015. Intraspecific genetic lineages of a marine mussel show behavioural divergence and spatial segregation over a tropical/subtropical biogeographic transition. BMC Evolutionary Biology, 15, 100.

Zbawicka, M., Burzyński, A., Wenne, R., 2007. Complete sequences of mitochondrial genomes from the Baltic mussel *Mytilus trossulus*. Gene 406, 191–198.

Zbawicka, M., Burzyński, A., Skibinski, D., Wenne, R., 2010. Scottish *Mytilus trossulus* mussels retain ancestral mitochondrial DNA: complete sequences of male and female mtDNA genomes. Gene 456, 45–53.

Zbawicka, M., Drywa, A., Śmietanka, B., Wenne, R., 2012. Identification and validation of novel SNP markers in European populations of marine *Mytilus* mussels. Mar. Biol. 159, 1347–1362.

Zhan, A., Bao, Z., Hu, X., Lu, W., Wang, S., Peng, W., Hu, J., 2008. Accurate methods of DNA extraction and PCR-based genotyping for single scallop embryos/larvae long preserved in ethanol. Molecular Ecology Resources, 8(4), 790–5.

Žižek, S., Gombač, M., Pogačnik, M., 2012. Occurrence and effects of the bivalve-inhabiting hydroid Eugymnanthea inquilina in cultured Mediterranean mussels (*Mytilus galloprovincialis*) in Slovenia 49, 149–154.

Zuo, L., Wang, K., Luo, X., 2014. Use of diplotypes - matched haplotype pairs from homologous chromosomes - in gene-disease association studies. Shanghai Arch. Psychiatry 26, 165–170.

APPENDIX 1

APPENDIX 1A – SSTNE extraction buffer recipe

1. For 1 L of extraction buffer, add the following reagents to approximately 500 mL ddH_2O

    - 17.5 g NaCl
    - 6.05 g Tris base (Fisher Scientific)
    - 1 mL 0.2 M EDTA (Fisher Scientific)
    - 76 mg EGTA (Sigma Aldrich, ref E3889)
    - 72 mg spermidine (Sigma Aldrich, ref SO266)
    - 52mg spermine (Sigma Alrdich, ref S1141)

2. Stir gently with a magnetic stirrer. DO NOT VORTEX as the buffer contains detergents that will foam if shaken
3. Top up to the total volume (1 L) with ddH_2O
4. Measure pH; should be between 9.5-10
5. Autoclave and store buffer at room temperature

APPENDIX 2

Genome assembly from Illumina sequencing data was performed with a series of shell scripts, which are presented in this appendix.

APPENDIX 2A – Barcode demultiplexing script "build_samples.sh". A total of four HiSeq and two MiSeq runs were performed. The following script was used to demultiplex barcodes from HiSeq run 1; the file name, which has been highlighted **bold** in this script, was changed accordingly to demultiplex the other data sets.

```
#!/bin/bash
mkdir samples
fqz_comp -d < 358W16403_NoIndex_L000_R1_001.fqz >
358W16403_NoIndex_L000_R1_001.fastq
rm -r 358W16403_NoIndex_L000_R1_001.fqz
fqz_comp -d < 358W16403_NoIndex_L000_R2_001.fqz >
358W16403_NoIndex_L000_R2_001.fastq
rm -r 358W16403_NoIndex_L000_R2_001.fqz
process_radtags -E phred33 --filter_illumina -i fastq -y fastq -1
358W16403_NoIndex_L000_R1_001.fastq -2 358W16403_NoIndex_L000_R2_001.fastq -o samples/
-e pstI -t 93 -c -q --inline_inline -b barcodes.5.5.txt
cd samples
mv process_radtags.log process_radtags.5.5.log
rm -rf *.rem.?.fq
cd ..
process_radtags -E phred33 --filter_illumina -i fastq -y fastq -1
358W16403_NoIndex_L000_R1_001.fastq -2 358W16403_NoIndex_L000_R2_001.fastq -o samples/
-e pstI -t 93 -c -q --inline_inline -b barcodes.5.6.txt
cd samples
mv process_radtags.log process_radtags.5.6.log
rm -rf *.rem.?.fq
cd ..
process_radtags -E phred33 --filter_illumina -i fastq -y fastq -1
358W16403_NoIndex_L000_R1_001.fastq -2 358W16403_NoIndex_L000_R2_001.fastq -o samples/
-e pstI -t 93 -c -q --inline_inline -b barcodes.7.5.txt
cd samples
mv process_radtags.log process_radtags.7.5.log
rm -rf *.rem.?.fq
cd ..
process_radtags -E phred33 --filter_illumina -i fastq -y fastq -1
358W16403_NoIndex_L000_R1_001.fastq -2 358W16403_NoIndex_L000_R2_001.fastq -o samples/
-e pstI -t 93 -c -q --inline_inline -b barcodes.7.6.txt
cd samples
mv process_radtags.log process_radtags.7.6.log
rm -rf *.rem.?.fq

for file in *.fq
do
  wc -l ${file} >>count.log
done
cd ..
rm -r 358W16403_NoIndex_L000_R1_001.fastq 358W16403_NoIndex_L000_R2_001.fastq
```

APPENDIX 2B – Script for merging samples "merge_samples.sh", which was used to associate all samples from four HiSeq and two MiSeq runs.

```
#!/bin/bash
mkdir samples
thepath=$(pwd)

#mkdir samples.miseq
#mkdir samples.hiseq
#cd /repository/queues/results/mussel.miseq/samples
#for A in *.1.fq.fqz;
#do
#  B=`echo "$A" | cut -f1 -d'.'`
#  fqz_comp -d < "$B.1.fq.fqz" > "$thepath/samples.miseq/$B.1.fq"
#  fqz_comp -d < "$B.2.fq.fqz" > "$thepath/samples.miseq/$B.2.fq"
#done
#cd /repository/queues/results/mussel.hiseq/samples
#for A in *.1.fq.fqz;
#do
#  B=`echo "$A" | cut -f1 -d'.'`
#  fqz_comp -d < "$B.1.fq.fqz" > "$thepath/samples.hiseq/$B.1.fq"
#  fqz_comp -d < "$B.2.fq.fqz" > "$thepath/samples.hiseq/$B.2.fq"
#done
#cd "$thepath"
#cd samples.miseq
#for A in *.fq;
#do
#  cat "../samples.miseq/$A" "../samples.hiseq/$A" > "../samples/$A"
#done
#cd ..
#rm -rf samples.miseq
#rm -rf samples.hiseq

cd /repository/queues/results/mussel.miseq/samples
for A in *.fq;
do
  cat "/repository/queues/results/mussel.miseq/samples/$A"
"/repository/queues/results/mussel.miseq.2/samples/$A"
"/repository/queues/results/mussel.hiseq.1/samples/$A"
"/repository/queues/results/mussel.hiseq.3/samples/$A"
"/repository/queues/results/mussel.hiseq.4/samples/$A" | fastx_renamer -n COUNT >
"$thepath/samples/$A"
done

cd "$thepath"
cd samples
for file in *.fq
do
  wc -l ${file} >>count.log
done
cd ..
```

APPENDIX 3

Appendix 3A – Complete PERL script "find.pattern.pl"

```
#!/usr/bin/perl
# $Revision: 0.4 $
# $Date: 2014/04/23 $
# $Id: find_pattern.pl $
# $Author: Michael Bekaert $
# $Desc: Find fix allele patterns $
use strict;
use warnings;
use Getopt::Long;
use List::MoreUtils qw/ uniq /;

#-------------------------------------------------------------
our $VERSION = 0.4;
my $shift = 2;    #exported.haplotypes.tsv samples start at the third position!

#-------------------------------------------------------------
my ($verbose, $fix, $min, $ploidy, $min_snp, $max_snp, $grouping, $population, $haplofile, $whitefile, $arff, $genepop, $ade,
$fasta, $mapfile, $snpfile) = (0, 0, 0, 2, 1, 2, 0);
GetOptions(
          'haplotypes=s' => \$haplofile,
          'population=s' => \$population,
                    'tag:s'        => \$mapfile,
          'snp:s'        => \$snpfile,
          'whitelist:s'  => \$whitefile,
          'min:i'        => \$min,
          'group:i'      => \$grouping,
          'arff:s'       => \$arff,
          'fasta:s'      => \$fasta,
          'genepop:s'    => \$genepop,
          'ade:s'        => \$ade,
          'minsnp:i'     => \$min_snp,
          'maxsnp:i'     => \$max_snp,
          'f|fix+'       => \$fix,
          'v|verbose!'   => \$verbose
         );
if (defined $haplofile && -r $haplofile && defined $population && -r $population && $min >= 0 && $grouping >= 0 &&
$grouping <= 2)
{
   my %whitelist;
   if (defined $whitefile && -r $whitefile && open my $IN, q{<}, $whitefile)
   {
      while (<$IN>)
      {
         chomp;
         my @tmp = split m/\t/x;
         $whitelist{$tmp[0]} = $tmp[0] if (exists $tmp[0]);
      }
      print {*STDERR} (scalar keys %whitelist), " markers in the whitelist!\n" if ($verbose);
      close $IN;
   }
   my (%physmap, %snpsmap);
   if (defined $mapfile && -r $mapfile && open my $IN, q{<}, $mapfile)
   {
      while (<$IN>)
      {
         chomp;
         my @tmp = split m/\t/x;
         if (scalar @tmp >= 10 && (!%whitelist || exists $whitelist{$tmp[2]}) && defined $tmp[9] && length $tmp[9] > 0)
         {
            if (defined $tmp[3] && length $tmp[3] > 0 && defined $tmp[4] && length $tmp[4] > 0 && defined $tmp[5] &&
length $tmp[5] > 0)
            {
               $tmp[3] = $1 if ($tmp[3] =~ m/^(.*):\d+\.\.\d+$/g);
               @{$physmap{$tmp[2]}} = ($tmp[9], $tmp[3], $tmp[4], $tmp[5]);
            }
```

```
            else { @{$physmap{$tmp[2]}} = $tmp[9]; }
         }
      }
      print {*STDERR} (scalar keys %physmap), " markers mapped!\n" if ($verbose);
      close $IN;
      if (defined $snpfile && -r $snpfile && open $IN, q{<}, $snpfile)
      {
         my ($nb, $last) = (0);
         while (<$IN>)
         {
            chomp;
            my @tmp = split m/\t/x;
            if (scalar @tmp >= 6 && (!%whitelist || exists $whitelist{$tmp[2]}))
            {
               my @seq;
               for my $i (5 .. (scalar @tmp - 1)) { push @seq, $tmp[$i] if (length($tmp[$i]) > 0); }
               if (!defined $last || $last != $tmp[2]) { $nb = 0; }
               else                                    { $nb++; }
               @{$snpsmap{$tmp[2] . q{_} . chr(65 + $nb)}} = ($tmp[3], join(q{}, sort @seq));
               $last = $tmp[2];
            }
         }
         print {*STDERR} (scalar keys %snpsmap), " SNP identified!\n" if ($verbose);
         close $IN;
      }
   }
   my (%pop, %group);
   if (open my $IN, q{<}, $population)
   {
      my %class;
      while (<$IN>)
      {
         chomp;
         my @tmp = split m/\t/x;
         if (scalar @tmp >= 2 && $tmp[1] ne q{-})
         {
            if (!exists $class{$tmp[1]})
            {
               $class{$tmp[1]} = scalar keys %class;
               $group{$class{$tmp[1]}} = 0;
            }
            $pop{$tmp[0]} = $class{$tmp[1]};
            $group{$class{$tmp[1]}}++;
         }
      }
      close $IN;
      if ($verbose)
      {
         print {*STDERR} (scalar keys %pop), " samples to be used!\n";
         foreach my $item (sort keys %class) { print {*STDERR} 'Group ', $item, ' [', $class{$item}, '] have ',
$group{$class{$item}}, ' member', ($group{$class{$item}} > 1 ? q{s} : q{}), "\n"; }
      }
   }
   $min = int((scalar keys %pop) * 0.75) if ($min == 0 || $min > scalar keys %pop);
   print {*STDERR} 'Threshold fixed at ', $min, "\n" if ($verbose);
   if (%pop && %group)
   {
      my @traits;
      if ($grouping == 1) { push @traits, {%pop}; }
      elsif ($grouping == 2)
      {
         my @groups = keys %group;
         my $size   = scalar @groups;
         for (my $i = 0 ; $i < 2**$size ; $i++)
         {
            my $str = sprintf("%*.*b", $size, $size, $i);
            my %combination;
            for (my $j = 0 ; $j < $size ; $j++)
            {
               #if (substr($str, $j, 1)) { $combination{$groups[$j]} = $groups[$j]; }
               if (substr($str, 0, 1) ne '1' && substr($str, $j, 1)) { $combination{$groups[$j]} = $groups[$j]; }
            }
            if (scalar keys %combination > 0 && scalar keys %combination < $size)
```

```
            {
               my %tmp;
               foreach my $item (keys %pop) { $tmp{$item} = (exists $combination{$pop{$item}} ? 1 : 0); }
               push @traits, {%tmp};
            }
         }
      }
      my @ind;
      if (open my $IN, q{<}, $haplofile)
      {
         my ($nb_all, $nb_selected, $nb_good, $nb_good_snp, $index) = (0, 0, 0, 0, 0);
         foreach my $item (split m/\t/x, <$IN>)
         {
            chomp $item;
            $index++;
            if ($index > $shift) { push @ind, $item; }
         }
         if (scalar @ind == scalar keys %pop)
         {
            my (@header, @good_markers);
            my (%line_arff, %line_genepop, %line_ade, %line_fasta, %header_arff, %header_good);
            while (<$IN>)
            {
               $nb_all++;
               chomp;
               my @data = split m/\t/x;
               if (scalar @data == (scalar @ind + $shift) && $data[1] >= $min && (!%whitelist || exists $whitelist{$data[0]}))
               {
                  $nb_selected++;
                  my ($num_allele, $num_snp) = (0, 0);
                  my %alleles;
                  $index = 0;
                  foreach my $item (@data)
                  {
                     last if ($item eq 'consensus' || $num_allele > $ploidy || $num_snp > $max_snp);
                     $index++;
                     if ($index > $shift && $item ne '-')
                     {
                        my @item2 = sort(split(m/\//x, $item));
                        $num_allele = scalar @item2 if ($num_allele < scalar @item2);
                        $num_snp = length($item2[0]);
                        foreach my $allele (@item2)
                        {
                           my $i = 0;
                           foreach my $subitem (split m//x, $allele) { push @{$alleles{$ind[$index - ($shift + 1)]}[$i++]},
$subitem; }
                        }
                     }
                  }
                  my %all_alleles;
                  if ($index == scalar @data)
                  {
                     foreach my $item (keys %pop)
                     {
                        if (exists $alleles{$item})
                        {
                           foreach my $subitem (@{$alleles{$item}}) { push @{$all_alleles{$item}}, join(q{}, uniq(sort
@{$subitem})); }
                        }
                        else
                        {
                           for my $i (1 .. $num_snp) { push @{$all_alleles{$item}}, 'N'; }
                        }
                     }
                     undef %alleles;
                  }
                  if (%all_alleles && $num_snp >= $min_snp)
                  {
                     my $lasti = -1;
                     for my $i (0 .. ($num_snp - 1))
                     {
                        my $flag_fix = 1;
                        if ($fix)
```

```
                {
                   my @thelist;
                   foreach my $item (keys %pop)
                   {
                      if (exists $all_alleles{$item}[$i])
                      {
                         push @thelist, $all_alleles{$item}[$i] if ($fix == 2 && $all_alleles{$item}[$i] ne 'N');
                         $flag_fix = 0 if (length($all_alleles{$item}[$i]) > 1);
                      }
                   }
                   $flag_fix = 0 if (@thelist && (scalar uniq sort @thelist) > 2);
                }
                if ($flag_fix)
                {
                   if ($grouping == 1 || $grouping == 2)
                   {
                      foreach my $trait (@traits)
                      {
                         my $true = 0;
                         foreach my $refs (keys %pop)
                         {
                            $true = 0;
                            my $ref = $all_alleles{$refs}[$i];
                            next if ($ref eq 'N');
                            $ref = $ref x $ploidy if (length($ref) == 1);
                            if (length($ref) == $ploidy)
                            {
                               my (%tmpline_arff, %tmpline_ade, %tmpline_genepop, %tmpline_fasta);
                               my $flag = $trait->{$refs};
                               foreach my $item (keys %pop)
                               {
                                  if (exists $all_alleles{$item}[$i])
                                  {
                                     my $tmp = $all_alleles{$item}[$i];
                                     $tmp = $tmp x $ploidy if (length($tmp) == 1);
                                     if (length($tmp) == $ploidy)
                                     {
                                        if ($tmp eq 'NN' || ($flag eq $trait->{$item} && $ref eq $tmp) || ($flag ne $trait-
>{$item} && $ref ne $tmp)) { $true++; }
                                        if (defined $arff)
                                        {
                                           my $tmp2 = $tmp;
                                           $tmp2 =~ s/N+/\?/g;
                                           push @{$tmpline_arff{$item}}, $tmp2;
                                           push @{$header_arff{$data[0] . q{_} . chr(65 + $i)}}, $tmp2 if ($tmp2 ne '?');
                                        }
                                        if (defined $ade)
                                        {
                                           my $tmp2 = $tmp;
                                           $tmp2 =~ tr/ATCGN/1234 /;
                                           push @{$tmpline_ade{$item}}, $tmp2;
                                        }
                                        if (defined $genepop)
                                        {
                                           my $tmp2 = $tmp;
                                           $tmp2 =~ s/A/01/g;
                                           $tmp2 =~ s/C/02/g;
                                           $tmp2 =~ s/G/03/g;
                                           $tmp2 =~ s/T/04/g;
                                           $tmp2 =~ s/N/00/g;
                                           push @{$tmpline_genepop{$item}}, $tmp2;
                                        }
                                        if (defined $fasta) { push @{$tmpline_fasta{$item}}, $tmp; }
                                        if (!(defined $arff || defined $genepop || defined $fasta || defined $ade))
                                        {
                                           my $tmp2 = $tmp;
                                           $tmp2 =~ s/N+/\?/g;
                                           push @{$header_good{$data[0] . q{_} . chr(65 + $i)}{$pop{$item}}}, $tmp2 if
($tmp2 ne '?');
                                        }
                                     }
                                  }
                               }
                            }
```

```
                                    if ($true != (scalar keys %pop))
                                    {
                                       #remove header
                                       if (defined $arff) { delete $header_arff{$data[0] . q{_} . chr(65 + $i)}; }
                                       if (!(defined $arff || defined $genepop || defined $fasta || defined $ade)) { delete
$header_good{$data[0] . q{_} . chr(65 + $i)}; }
                                    }
                                    else
                                    {
                                       if ((defined $arff || defined $genepop || defined $fasta || defined $ade))
                                       {
                                          foreach my $item (keys %pop)
                                          {
                                             if (defined $arff) { push @{$line_arff{$item}}, @{$tmpline_arff{$item}}; }
                                             if (defined $ade)  { push @{$line_ade{$item}},  @{$tmpline_ade{$item}}; }
                                             if (defined $genepop) { push @{$line_genepop{$pop{$item}}{$item}},
@{$tmpline_genepop{$item}}; }
                                             if (defined $fasta) { push @{$line_fasta{$item}}, @{$tmpline_fasta{$item}}; }
                                          }
                                       }
                                       push @header,       $data[0] . q{_} . chr(65 + $i);
                                       push @good_markers, $data[0];
                                       last;
                                    }
                                 }
                              }
                              last if ($true == (scalar keys %pop));
                           }
                        }
                        else
                        {
                           my ($ref, $flag);
                           foreach my $item (keys %pop)
                           {
                              if (exists $all_alleles{$item}[$i])
                              {
                                 my $tmp = $all_alleles{$item}[$i];
                                 $tmp = $tmp x $ploidy if (length($tmp) == 1);
                                 if (length($tmp) == $ploidy)
                                 {
                                    push @header, $data[0] . q{_} . chr(65 + $i) if ($i != $lasti);
                                    push @good_markers, $data[0] if ($i != $lasti);
                                    if (defined $arff)
                                    {
                                       my $tmp2 = $tmp;
                                       $tmp2 =~ s/N+/\?/g;
                                       push @{$line_arff{$item}}, $tmp2;
                                       push @{$header_arff{$data[0] . q{_} . chr(65 + $i)}}, $tmp2 if ($tmp2 ne '?');
                                    }
                                    if (defined $ade)
                                    {
                                       my $tmp2 = $tmp;
                                       $tmp2 =~ tr/ATCGN/1234 /;
                                       push @{$line_ade{$item}}, $tmp2;
                                    }
                                    if (defined $genepop)
                                    {
                                       my $tmp2 = $tmp;
                                       $tmp2 =~ s/A/01/g;
                                       $tmp2 =~ s/C/02/g;
                                       $tmp2 =~ s/G/03/g;
                                       $tmp2 =~ s/T/04/g;
                                       $tmp2 =~ s/N/00/g;
                                       push @{$line_genepop{$pop{$item}}{$item}}, $tmp2;
                                    }
                                    if (defined $fasta) { push @{$line_fasta{$item}}, $tmp; }
                                    if (!(defined $arff || defined $genepop || defined $fasta || defined $ade))
                                    {
                                       my $tmp2 = $tmp;
                                       $tmp2 =~ s/N+/\?/g;
                                       push @{$header_good{$data[0] . q{_} . chr(65 + $i)}{$pop{$item}}}, $tmp2 if ($tmp2 ne
'?');
                                    }
```

```
                                    }
                                    else
                                    {
                                       print {*STDERR} $data[0], ': Ambibuity detected: ploidy unresolved!', "\n" if ($verbose);
                                       undef @header;
                                       last;
                                    }
                                 }
                                 $lasti = $i;
                              }
                           }
                        }
                     }
                  }
               }
            }
            $nb_good = scalar uniq sort @good_markers;
            undef @good_markers;
            $nb_good_snp = scalar @header;
            if (defined $arff && @header && scalar @header > 0)
            {
               print {*STDOUT} "\@RELATION $haplofile\n\n\@ATTRIBUTE Sample STRING\n\@ATTRIBUTE Population
{", join(q{,}, uniq(sort(keys %group))), "}\n";
               foreach my $item (@header) { print {*STDOUT} '@ATTRIBUTE ', $item, ' {', join(q{,},
uniq(sort(@{$header_arff{$item}}))), "}\n"; }
               print {*STDOUT} "\n\@DATA\n";
               foreach my $item (keys %line_arff) { print {*STDOUT} $item, q{,}, $pop{$item}, q{,}, join(q{,},
@{$line_arff{$item}}), "\n"; }
            }
            if (defined $ade && @header && scalar @header > 0)
            {
               print {*STDOUT} "Samples\t", join("\t", @header), "\n";
               my $counter = 0;    ##
               foreach my $item (keys %line_ade)
               {
                  my $tmp = $item;
                  $tmp =~ s/[^a-zA-Z0-9]+//gx;

                  #print {*STDOUT} substr($tmp, 0, 8), "\t", join("\t", @{$line_ade{$item}}), "\n";
                  print {*STDOUT} substr($tmp, 0, 5), $counter++, "\t", join("\t", @{$line_ade{$item}}), "\n";    ##
               }
               my @tmp;
               foreach my $item (keys %line_ade) { push @tmp, $pop{$item}; }
               print {*STDERR} "R/adegenet population vector:\n pop <- c('", join('\',\'', @tmp), "');\n";
            }
            if (defined $genepop && @header && scalar @header > 0)
            {
               print {*STDOUT} "Find_pattern.pl version $VERSION; Genepop 4.+\n";
               print {*STDOUT} join(q{,}, @header), "\n";
               foreach my $pop (keys %line_genepop)
               {
                  print {*STDOUT} "pop\n";
                  foreach my $item (keys %{$line_genepop{$pop}}) { print {*STDOUT} $item, ",\t", join("\t",
@{$line_genepop{$pop}{$item}}), "\n"; }
               }
            }
            if (defined $fasta && @header && scalar @header > 0)
            {
               foreach my $item (keys %line_fasta) { print {*STDOUT} q{>}, $item, ' [', $pop{$item}, "]\n", join(q{},
@{$line_fasta{$item}}), "\n"; }
            }
            if (!(defined $arff || defined $genepop || defined $fasta || defined $ade) && @header && scalar @header > 0)
            {
               print {*STDOUT} 'Marker_SNP';
               foreach my $pop (sort keys %group) { print {*STDOUT} "\tAllele_", $pop; }
               print {*STDOUT} "\n";
               foreach my $item (@header)
               {
                  if (exists $header_good{$item})
                  {
                     my $id = substr $item, 0, -2;
                     print {*STDOUT} $item;
```

```
                    foreach my $pop (sort keys %group) { print {*STDOUT} "\t", (exists $header_good{$item}{$pop} ? q{\{} .
join(q{,}, uniq(sort(@{$header_good{$item}{$pop}}))) . q{\}} : '{NN}'); }
                    if (%snpsmap && exists $snpsmap{$item} && %physmap && exists $physmap{$id})
                    {
                        print {*STDOUT} "\t", substr($physmap{$id}[0], 0, $snpsmap{$item}[0]), q{[}, $snpsmap{$item}[1],
q{]}, substr($physmap{$id}[0], $snpsmap{$item}[0] + 1);
                        print {*STDOUT} "\t", $physmap{$id}[1], "\t", $physmap{$id}[2], "\t", $physmap{$id}[3] if (exists
$physmap{$id}[1]);
                    }
                    print {*STDOUT} "\n";
                }
            }
        }
    }
    close $IN;
    print {*STDERR} "Total markers read: $nb_all\nMarker analysed:    $nb_selected\nMarker selected:    $nb_good\nSNP
selected:      $nb_good_snp\n\n";
  }
 }
}
else
{
  print
   "Usage: $0 --haplotypes <batch_<num>.haplotypes.tsv> --population <popmap.txt>\nDescription: Test for diagnostic
alleles or patterns between populations\n\n--haplotypes <file>\n    Raw haplotype file, automaticaly generated by Stacks, and
called\n    batch_<num>.haplotypes.tsv.\n\n--population <file>\n    Population file used by Stacks.\n\n--tag <file>\n
batch_<num>.catalog.tags.tsv, required for physical mapping and marker sequences.\n\n--snp <file>\n
batch_<num>.catalog.snps.tsv, required for marker sequences.\n\n--whitelist <file>\n    Text file with the list of marker to only
consider.\n\n--min <integer>\n    Minimum number of sample sharing alleles. [default 75%]\n\n--group <file>\n    Grouping
[default 0].\n      0 between individuals [all];\n      1 between populations [species];\n      2 between groups of population
[group].\n\n--arff <file>\n    Output as an ARFF file format.\n\n--fasta <file>\n    Output as a FASTA file format (SNP
only).\n\n--genepop <file>\n    Output as an genepop file format.\n\n--minsnp <integer>\n    Minimum number of SNP. [default
1]\n\n--maxsnp <integer>\n    Maximum number of SNP. [default 2]\n\n--fix\n    Force fixed alleles only.\n\n--fix --fix\n
Force ONE fixed fallele only.\n\n--verbose\n    Becomes very chatty.\n\n";
}

#grouping
#old        new
#all        0       done!
#species    1       done!
#group      2       done!
#./find_pattern.pl --haplotypes batch_2.haplotypes.tsv --population farmed.txt -v --group 2 -d
```

APPENDIX 4

All Principal Component Analyses and Discriminant Analysis of Principal Components were performed with the *adegenet* package (version 1.4-1; Jombart, 2008) for R (version 3.1.0) (R Core Team, 2014). All scripts are presented in this appendix.

APPENDIX 4A – PCA and DAPC scripts for Chapter 3

#1. 349 RAD MARKERS 40 SAMPLES

```
#LOAD ADEGENET
library(adegenet)

#FIND CLUSTERS
#CREATE GENIND OBJECT
haplo <- read.csv("pattern.ade",header=TRUE,sep="\t")
pop<-
c('0','1','0','2','2','0','1','0','1','0','0','1','1','0','0','0','2','2','1','2','0','1','1','1','0','0','0','1','0','1','1','0','1','0','0','1','
0','1','0','0')
row.names(haplo) <- haplo[,1]
haplo <- haplo[,!(names(haplo) == "Samples")]
obj_pop <-df2genind(haplo, ploidy=2, pop=pop)

#FIND VARIANCE EXPRESSED BY PCA
grp<-find.clusters(obj_pop, npca=23,n.clust=3)
#PLOT GRAPH OF ORIGINAL GROUPS & INFERRED GROUPS
table.value(table(pop(obj_pop), grp$grp), col.lab=paste("inf", 1:3),row.lab=paste("ori", 1:3))
table(pop(obj_pop), grp$grp)

#DESCRIBING CLUSTERS USING DAPC
dapc1<-dapc(obj_pop,grp$grp)
30
1400

#MAKE GRAPH
scatter(dapc1, scree.da=FALSE, scree.pca=FALSE)
```

#2. DAPC RAD MARKERS 40 SAMPLES

```
#LOAD ADEGENET
library(adegenet)

#CREATE GENIND OBJECT
haplo <- read.csv("12radlociexcel.txt",header=TRUE,sep="\t")

pop<-
c('0','1','0','2','2','0','1','0','1','0','0','1','1','0','0','0','2','2','1','2','0','1','1','1','0','0','0','1','0','1','1','0','1','0','0','1','
0','1','0','0')
row.names(haplo) <- haplo[,1]
haplo <- haplo[,!(names(haplo) == "Samples")]
obj_pop <-df2genind(haplo, ploidy=2, pop=pop, ncode=0)
```

```
#FIND VARIANCE EXPRESSED BY PCA
grp<-find.clusters(obj_pop)
8
3

#PLOT GRAPH OF ORIGINAL GROUPS & INFERRED GROUPS
table.value(table(pop(obj_pop), grp$grp), col.lab=paste("inf", 1:3),row.lab=paste("ori", 1:3))
table(pop(obj_pop), grp$grp)

#DESCRIBING CLUSTERS USING DAPC
dapc1<-dapc(obj_pop,grp$grp)
12
150

#MAKE GRAPH
scatter(dapc1, scree.da=TRUE, scree.pca=TRUE,pch=19, cstar=FALSE)
```

#3. DAPC KASP ASSAYS 40 SAMPLES

```
#LOAD ADEGENET
library(adegenet)

#2. FIND CLUSTERS
#CREATE GENIND OBJECT
haplo <- read.csv("KASP 40.txt",header=TRUE,sep="\t")
pop<-
c('0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','3','3','3','3','3','3')
row.names(haplo) <- haplo[,1]
haplo <- haplo[,!(names(haplo) == "Samples")]
obj_pop <-df2genind(haplo, ploidy=2, pop=pop, ncode=0)

obj_pop


#FIND VARIANCE EXPRESSED BY PCA
grp<-find.clusters(obj_pop)
10
3

#PLOT A TABLE OF GROUPS
table(pop(obj_pop), grp$grp)


#PLOT GRAPH OF ORIGINAL GROUPS & INFERRED GROUPS
table.value(table(pop(obj_pop), grp$grp), col.lab=paste("inf", 1:3),row.lab=paste("ori", 1:3))
table(pop(obj_pop), grp$grp)

#DESCRIBING CLUSTERS USING DAPC
dapc1<-dapc(obj_pop,grp$grp)
10

10000

#MAKE GRAPH
```

```
scatter(dapc1, scree.da=FALSE, cstar=FALSE)


#4. 12 KASP ASSAYS 178 SAMPLES
#LOAD ADEGENET
library(adegenet)

#CREATE GENIND OBJECT
haplo <- read.csv("12locikasp.txt",header=TRUE,sep="\t")

pop<-
c('0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','
0','0','0','0','0','0','0','0','0','0','0','0','0','0','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1',
'1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','2','2','2','2','2','2','2','2','2','2'
,'2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2
','2','2','2','3','3','3','3','3','3','3','3','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','
4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','5','5','5','5','5','5','5','5','5','5','5','5',
'5','5','5','5','5','5','5','5')
row.names(haplo) <- haplo[,1]
haplo <- haplo[,!(names(haplo) == "Samples")]
obj_pop <-df2genind(haplo, ploidy=2, pop=pop, ncode=0)

#FIND VARIANCE EXPRESSED BY PCA
grp<-find.clusters(obj_pop)
15
3

#PLOT GRAPH OF ORIGINAL GROUPS & INFERRED GROUPS
table.value(table(pop(obj_pop), grp$grp), col.lab=paste("inf", 1:3),row.lab=paste("ori", 1:3))
table(pop(obj_pop), grp$grp)

#DESCRIBING CLUSTERS USING DAPC
dapc1<-dapc(obj_pop,grp$grp)
12
10200

#MAKE GRAPH
scatter(dapc1, scree.da=TRUE, scree.pca=TRUE,pch=19, cstar=FALSE)
```

APPENDIX 4B – PCA and DAPC script for Chapter 4

```
#LOAD ADEGENET
library(adegenet)
#CREATE GENIND OBJECT

hap <- read.csv("coastline.txt", header=TRUE,sep="\t")
pop<-
c('0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','
0','0','0','0','0','0','0','0','0','0','0','0','0','0','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1',
'1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2'
,'2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','3','3','3','3','3','3','3
','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','3','4','4','
4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4','4',
'4','4','4','4','4','4','4','4','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5'
,'5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','5','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6
','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','6','7','7','7','7','7','7','7','
7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','7','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8',
```

```
'8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8'
,'8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','8','9','9','9','9','9','9','9','9','9','9
','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','
9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9','9',
'9','9','9','9','9','9','9','9','9','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','10','
10','10','10','10','10','10','10','10','10','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','1
1','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11','11'
,'11','11','11','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','
12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','12','1
2','12','12','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13'
,'13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','13','
13','13','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','1
4','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14','14'
,'14','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','
15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','15','16','16','16','16','1
6','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16','16'
,'17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','
17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','17','1
8','18','18','18','18','18','18','18','18','18','19','19','19','19','19','19','19','24','24','24','24','24','24','24','24','24','24'
,'24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','
24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','24','21','21','21','21','21','21','21','21','2
1','21','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22'
,'22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','22','
22','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','2
3','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23','23'
)
row.names(hap) <- hap[,1]
hap <- hap[,!(names(hap) == "Samples")]
obj_pop <-df2genind(hap, ploidy=2, pop=pop, ncode=0)

#PCA
grp<-find.clusters(obj_pop)
8
3

#TABLE
table(pop(obj_pop), grp$grp)

#DAPC analysis
dapc1<-dapc(obj_pop,grp$grp)
12
5000
scatter(dapc1, scree.da=FALSE, pch=19)
```

APPENDIX 4C – PCA and DAPC script for Chapter 5

```
#LOAD ADEGENET
library(adegenet)

#2. FIND CLUSTERS
#CREATE GENIND OBJECT
haplo <- read.csv("LET.txt",header=TRUE,sep="\t")
pop<-
c('0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','
0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0','0',
'0','0','0','0','0','0','0','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1'
,'1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1','1
```

```
','1','1','1','1','1','1','1','1','1','1','1','1','1','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','
2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2',
'2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2'
,'2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2
','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2','2')
row.names(haplo) <- haplo[,1]
haplo <- haplo[,!(names(haplo) == "Samples")]
obj_pop <-df2genind(haplo, ploidy=2, pop=pop, ncode=0)


# PCA
grp<-find.clusters(obj_pop)
12
3

#PLOT A TABLE OF GROUPS
table(pop(obj_pop), grp$grp)


#3. DAPC
dapc1<-dapc(obj_pop,grp$grp)
12
5000

#MAKE GRAPH
dapc1
scatter(dapc1, scree.da=TRUE, scree.pca=TRUE, pch=19, cstar=FALSE)
```

APPENDIX 5

APPENDIX 5A – Table of the full sequences of loci chosen for SNP assay design, and their corresponding loading values. The SNP of interest for assay design is highlighted in **bold** and is only highlighted in optimised sequences. All sequences below the dotted line are SNP assays that could not be optimised. SNPs are represented by IUPAC codes (A/C = M; A/G = R; A/T = W; C/G = S; C/T = Y; G/T = K). In cases where contig assembly did not generate a complete sequence, missing nucleotides are represented by N.

Assay name	Complete RAD tag sequence	Loading value
E1	TGCAGTTTACCGATTTGGAAGCGGTGGGCGGCGCTT**[R]**TAATTTGTTGGTCGCGCAAAATGTTAAACAGGCAAAAGTTAATAAGTTTTAACTGG	0.00294
E2	TGCAGGGGACTACTTGTTCCACTTGACATGAGTAAGATGGGTAAAGTGKCTCAAGTGATATAT**[K]**ATGGAATAAATTTGTGGTAAATATCAGGT	0.00228
E3	TGCAGGCCAAAGTGTTTCTTCCTGAT**[W]**CAGTTTCATCTTATATGGTGACCAAT[M]TGTAATCTCTTGCTAGAGGGTGCGATTTCAAATTATACA	0.00269
G1	TGCAGAGTGAGAGCCCTAGCAGAAAGAGGAGAAAAACCTC**[M]**GGCAGTTATATTGATTTGACACATTCTCCAAGCCCCCCACCATCTTGGACCA	0.00033
G2	TGCAGATTTAAAGTTGATAAAACTCAACCTACCTTTATAGT**[K]**GTATCTTTATCT[W]TTATAAAATAAAATCCCTTGTTTATTTGCAATATTAAA	0.00042
G3	TGCAGT[Y]GTAGGGAATCTGTTAGTCATATTTACATTAGTACATAATAA**[R]**CGGATGAGAACTGTTACAAACGTATTTTTACTGAACCTTGCCGT	0.00144
G4	TGCAGC[W]GCAACAGCAGCAAACCTTTCATCCTTATCATC**[S]**AAGATGGCTGTCACAAAACAACATTTACAAGAACTAAGCAACAAGACTTATCC	0.00128
T1	TGCAGCTTTCAAAAAGGAATCTGGTTTATTCGATTCA[**R**]TGAATGTTACCCCTACTATATGACTGCTATGGTTTGCTCAATATTTTGTTATTTA	0.00133
T2	TGCAGATGCAATTACTTCTAAATGTGGATGCCACACAAAGATAATT**[K]**CACACAACATCCCTAATTAATTTTGTTTTCTCTTGTAGAACATGCT	0.00131
T3	TGCAGTAATGGACCTTGCTTCCTTTGCCGCYTCCATTGCAAA**[Y]**GATACTGCTAATTTAACGCAAATGACAATATCTCCTACAAATTTGGATGC	0.00126
T4	TGCAGAGAAACTTGATCCTTTCTTCTGTAAAGGTTGTAATAACCTTGTACA**[Y]**AAACCATCAGCATCCTTTGTATAATGCTGAAAGATACAATT	0.00114
T5	TGCAGATGCAAAAGATAGAGCAAAAAGAATATTGGCAGGTTGTAGAGGAGG**[W]**AGTATTGGGTCTTACACTGCTAGCCCTGGTGTTGAGGTTAT	0.00119
E4	TGCAGAGCTGGATTKACTGCTGGGGGCGA[Y]CGTGGGTCCTCCTTTTCCTAGCTGTTGGTTAATTCTTTGAGCTGCTTCTTTAGCAGCTTGAGCT	0.00287

E5	TGCAGTCATTTCCATTAAAACSAGTCAGTATCTGACGTATTCACT CAGCTGCACTATTTCAGAGAAACCAATTT[K]CTAATGAGAAGTT GTTGCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNTAGTTAGCATACTGCCAGGCCAACTCACTAA TAACACTAAAGTTGTGCTCAGCCGATAAACTCGCCGGTATTTTT GAAAAAACGGAGAAAACGATAAGACCCCATAACGTTAGCGGAT AGCGTCATTTTGTTAACGTAATCGTAAACGTGCACACCTTAGAT TGACAGACGTTGCGGCCACACTATTTGAGCCACATATAAAAAAT TGATATCGACTTGTAGCACAGGAACAGGGCAACAAATCCGTAA CAGTAGATGTATGC	0.00348
E6	TGCAGCCAACAAAAAGCTTCAAGCAAAACAGCATACCCAAACA TAAATGATGAG[R][R]TGGCTAAATAATGAGGGCAACTAATGAA AAACCATGC	0.00200
G5	TGCAGCGCCCTCGAACTCACCCGACACAAATACCTTCCCATTGT TCGTGTGCAATCCTG[W]TAGCATGAAAAACAAG[K]AATGTCTTA TTAATTT	0.00055
G6	TGCAGGAACCCAACAACCAAGAAACCAACCACCCCCACCACAA CAGT[W]TGCCCAACCACCACCACCACAAGAAGCAGATAGATTT TC[Y]GGAAA	0.00045
G7	TGCAGCATATCAAAATTCACAATCAACAACACATCTGCTACTGT GAGCAATGTGGCGA[Y]GGCTTTATGGAGTATGGGAAATTAAGG TTGCACA	0.00112

APPENDIX 6

APPENDIX 6A – Tables showing PCR results from genotyping of all individuals with Me15/16: i. Loch Ryan; ii. Rascarrel Bay; iii. Bay of Piran; iv. Penn Cove; v. Bras d'Or Lake; vi. Loch Etive. Individuals chosen for RAD library construction are highlighted in grey

i. Loch Ryan

Individual ID	Genotype
LR_01	*Me*
LR_02	*Me*
LR_03	*Me*
LR_04	*Me*
LR_05	*Me*
LR_06	*Me*
LR_07	*Me*
LR_08	*Me*
LR_09	*Me*
LR_10	*Me*
LR_11	*Me*
LR_12	*Me*
LR_13	*Me*
LR_14	*Me*
LR_15	*Me*
LR_16	*Me*
LR_17	*Me*
LR_18	*Me*
LR_19	*Me*
LR_20	*Me*
LR_21	*Me*
LR_22	*Me*
LR_23	*Me*
LR_24	*Me*
LR_25	*Me*
LR_26	*Me*
LR_27	*Me*
LR_28	*Me*
LR_29	*Me*
LR_30	*Me*
LR_31	*Me*
LR_32	*Me*
LR_33	*Me*
LR_34	*Me*
LR_35	*Me*
LR_36	*Me*
LR_37	*Me*
LR_38	*Me*
LR_39	*Me*
LR_40	*Me*
LR_41	*Me*
LR_42	*Me*
LR_43	*Me*
LR_44	*Me*
LR_45	*Me*
LR_46	*Me*
LR_47	*Me*
LR_48	*MeMg*
LR_49	*MeMg*
LR_50	*MeMg*

ii. Rascarrel Bay

Individual ID	Genotype
RB_01	*Me*
RB_02	*Me*
RB_03	*Me*
RB_04	*Me*
RB_05	*Me*
RB_06	*Me*
RB_07	*Me*
RB_08	*Me*
RB_09	*Me*
RB_10	*Me*
RB_11	*Me*
RB_12	*Me*
RB_13	*Me*
RB_14	*Me*
RB_15	*Me*
RB_16	*Me*
RB_17	*Me*
RB_18	*Me*
RB_19	*Me*
RB_20	*Me*
RB_21	*Me*
RB_22	*Me*
RB_23	*Me*
RB_24	*Me*
RB_25	*Me*
RB_26	*Me*
RB_27	*Me*
RB_28	*Me*
RB_29	*Me*
RB_30	*Me*
RB_31	*Me*
RB_32	*Me*
RB_33	*Me*
RB_34	*Me*
RB_35	*Me*
RB_36	*Me*
RB_37	*Me*
RB_38	*Me*
RB_39	*Me*
RB_40	*Me*
RB_41	*Me*
RB_42	*Me*
RB_43	*Me*
RB_44	*Me*
RB_45	*Me*
RB_46	*Me*
RB_47	*Me*
RB_48	*Me*
RB_49	*Me*
RB_50	*Me*

iii. Bay of Piran

Individual ID	Genotype
BP_01	*Mg*
BP_02	*Mg*
BP_03	*Mg*
BP_04	*Mg*
BP_05	*Mg*
BP_06	*Mg*
BP_07	*Mg*
BP_08	*Mg*
BP_09	*Mg*
BP_10	*Mg*
BP_11	*Mg*
BP_12	*Mg*
BP_13	*Mg*
BP_14	*Mg*
BP_15	*Mg*
BP_16	*Mg*
BP_17	*Mg*
BP_18	*Mg*
BP_19	*Mg*
BP_20	*Mg*
BP_21	*Mg*
BP_22	*Mg*
BP_23	*Mg*
BP_24	*Mg*
BP_25	*Mg*
BP_26	*Mg*
BP_27	*Mg*
BP_28	*Mg*
BP_29	*Mg*
BP_30	*Mg*
BP_31	*Mg*
BP_32	*Mg*
BP_33	*Mg*
BP_34	*Mg*
BP_35	*Mg*
BP_36	*Mg*
BP_37	*Mg*
BP_38	*Mg*
BP_39	*Mg*
BP_40	*Mg*
BP_41	*Mg*
BP_42	*Mg*
BP_43	*Mg*
BP_44	*Mg*
BP_45	*Mg*
BP_46	*Mg*
BP_47	*Mg*
BP_48	*Mg*
BP_49	*Mg*
BP_50	*Mg*

iv. Penn Cove

Individual ID	Genotype
PC_01	*Me*
PC_02	*Mt*
PC_03	*Mt*
PC_04	*Mt*
PC_05	*Mt*
PC_06	*Mt*
PC_07	*Mt*
PC_08	*MgMt*

v. Bras d'Or Lake

Individual ID	Genotype
BDL_01	*Mt*
BDL_02	*Mt*
BDL_03	*Mt*
BDL_04	*Mt*
BDL_05	*Mt*
BDL_06	*Mt*
BDL_07	*Mt*
BDL_08	*Mt*
BDL_09	*Mt*
BDL_10	*Mt*
BDL_11	*Mt*
BDL_12	*Mt*
BDL_13	*Mt*
BDL_14	*Mt*
BDL_15	*Mt*
BDL_16	*Mt*
BDL_17	*Mt*
BDL_18	*Mt*
BDL_19	*Mt*
BDL_20	*Mt*
BDL_21	*Mt*
BDL_22	*Mt*
BDL_23	*Mt*
BDL_24	*Mt*
BDL_25	*Mt*
BDL_26	*Mt*
BDL_27	*Mt*
BDL_28	*Mt*
BDL_29	*Mt*
BDL_30	*Mt*
BDL_31	*Mt*
BDL_32	*Mt*
BDL_33	*Mt*
BDL_34	*Mt*
BDL_35	*Mt*
BDL_36	*Mt*
BDL_37	*Mt*
BDL_38	*Mt*
BDL_39	*Mt*
BDL_40	*Mt*
BDL_41	*MgMt*
BDL_42	*MgMt*
BDL_43	*MgMt*
BDL_44	*MgMt*
BDL_45	*MgMt*
BDL_46	*MgMt*
BDL_47	*MgMt*
BDL_48	*MgMt*
BDL_49	*MgMt*
BDL_50	*MgMt*

vi. Loch Etive

Individual ID	Genotype
LET_01	*Mt*
LET_02	*Mt*
LET_03	*Mt*
LET_04	*Mt*
LET_05	*Mt*
LET_06	*Mt*
LET_07	*Mt*
LET_08	*Mt*
LET_09	*Mt*
LET_10	*Mt*
LET_11	*Mt*
LET_12	*Mt*
LET_13	*Mt*
LET_14	*Mt*
LET_15	*Mt*
LET_16	*Mt*
LET_17	*Mt*
LET_18	*Mt*
LET_19	*Mt*
LET_20	*Mt*

APPENDIX 6B – Numbers of assembled loci from RADseq data per individual, separated by population: LR = Loch Ryan; RB = Rascarrel Bay; BP =. Bay of Piran; PC = Penn Cove

Individual ID	Number of assembled loci
LR_01	250882
LR_02	307972
LR_03	244153
LR_04	251484
LR_05	202519
LR_06	258430
LR_07	296590
LR_08	297764
LR_09	212925
LR_10	208759
RB_01	18220
RB_02	292175
RB_03	313406
RB_04	138934
RB_05	271840
RB_06	250194
RB_07	294965
RB_08	255841
RB_09	201064
RB_10	279931
BP_01	294042
BP_02	234102
BP_03	260588
BP_04	307442
BP_05	301665
BP_06	292548
BP_07	301090
BP_08	258468
BP_09	257836
BP_10	229886
BP_11	253482
BP_12	294835
BP_13	303760
BP_14	268664
BP_15	300134
PC_01	5459
PC_02	187292
PC_03	59773
PC_04	268765
PC_05	96006

APPENDIX 6C –SNP assay genotypes per individual per population: i. Loch Ryan; ii. Rascarrel Bay; iii. Bay of Piran; iv. Penn Cove; v. Bras d'Or Lake; vi. Loch Etive

i. Loch Ryan

Individual ID	Genotype
LR_01	*Me*
LR_02	*Me*
LR_03	*Me*
LR_04	*Me*
LR_05	*Me*
LR_06	*Me*
LR_07	*Me*
LR_08	*Me*
LR_09	*Me*
LR_10	*Me*
LR_11	*Me*
LR_12	*MeMg*
LR_13	*MeMg*
LR_14	*Me*
LR_15	*Me*
LR_16	*Me*
LR_17	*Me*
LR_18	HXE
LR_19	*Me*
LR_20	*Me*
LR_21	*Me*
LR_22	*Me*
LR_23	*Me*
LR_24	*MeMg*
LR_25	*Me*
LR_26	*Me*
LR_27	*MeMt*
LR_28	*Me*
LR_29	*Me*
LR_30	*Me*
LR_31	*MeMg*
LR_32	HXE
LR_33	HXE
LR_34	*Me*
LR_35	*Me*
LR_36	*Me*
LR_37	*MeMg*
LR_38	*Me*
LR_39	HXE
LR_40	HXE
LR_41	HXE
LR_42	*Me*
LR_43	*Me*
LR_44	HXE
LR_45	*MeMg*
LR_46	*Me*
LR_47	*Me*
LR_48	*MeMg*
LR_49	*MeMg*
LR_50	*MeMg*

ii. Rascarrel Bay

Individual ID	Genotype
RB_01	*Me*
RB_02	*Me*
RB_03	HXE
RB_04	*Me*
RB_05	*Me*
RB_06	*MeMg*
RB_07	*Me*
RB_08	*Me*
RB_09	*Me*
RB_10	*Me*
RB_11	*Me*
RB_12	*Me*
RB_13	*Me*
RB_14	*Me*
RB_15	*Me*
RB_16	*Me*
RB_17	*Me*
RB_18	*Me*
RB_19	HXE
RB_20	*Me*
RB_21	*Me*
RB_22	*Me*
RB_23	*Me*
RB_24	*Me*
RB_25	*MeMt*
RB_26	*Me*
RB_27	HXE
RB_28	*Me*
RB_29	*Me*
RB_30	HXE
RB_31	HXE
RB_32	HXE
RB_33	HXE
RB_34	*Me*
RB_35	*Me*
RB_36	*Me*
RB_37	*Me*
RB_38	*Me*
RB_39	*Me*
RB_40	*MeMt*
RB_41	HXE
RB_42	*Me*
RB_43	*Me*
RB_44	*Me*
RB_45	*Me*
RB_46	*Me*
RB_47	*Me*
RB_48	*Me*
RB_49	*Me*
RB_50	HXE

iii. Bay of Piran

Individual ID	Genotype
BP_01	*Mg*
BP_02	*MeMg*
BP_03	HXG
BP_04	*Mg*
BP_05	*MeMg*
BP_06	*Mg*
BP_07	*MgMt*
BP_08	*Mg*
BP_09	*Mg*
BP_10	*Mg*
BP_11	*Mg*
BP_12	*Mg*
BP_13	*Mg*
BP_14	*Mg*
BP_15	*Mg*
BP_16	*Mg*
BP_17	*Mg*
BP_18	*Mg*
BP_19	*Mg*
BP_20	*Mg*
BP_21	*Mg*
BP_22	*Mg*
BP_23	*Mg*
BP_24	*Mg*
BP_25	*Mg*
BP_26	*MgMt*
BP_27	*Mg*
BP_28	*Mg*
BP_29	*Mg*
BP_30	*Mg*
BP_31	*Mg*
BP_32	*Mg*
BP_33	*MeMg*
BP_34	*MgMt*
BP_35	HXG
BP_36	*Mg*
BP_37	*Mg*
BP_38	*Mg*
BP_39	*Mg*
BP_40	*Mg*
BP_41	*Mg*
BP_42	*Mg*
BP_43	HXG
BP_44	*Mg*
BP_45	*Mg*
BP_46	*Mg*
BP_47	*MgMt*
BP_48	HXG
BP_49	*Mg*
BP_50	*MeMg*

iv. Penn Cove

Individual ID	Genotype
PC_01	*Me*
PC_02	HXT
PC_03	*Mt*
PC_04	*MgMt*
PC_05	*MeMt*
PC_06	*MgMt*
PC_07	*Mt*
PC_08	*MgMt*

v. Bras d'Or Lake

Individual ID	Genotype
BDL_01	*MgMt*
BDL_02	*MeMgMt*
BDL_03	HXT
BDL_04	*MgMt*
BDL_05	*MeMt*
BDL_06	*MeMt*
BDL_07	*Mt*
BDL_08	*MeMgMt*
BDL_09	*MeMt*
BDL_10	*MeMt*
BDL_11	*MgMt*
BDL_12	*MgMt*
BDL_13	*MeMt*
BDL_14	*MgMt*
BDL_15	*MeMt*
BDL_16	*MgMt*
BDL_17	*Mt*
BDL_18	*MeMgMt*
BDL_19	*MeMgMt*
BDL_20	*MeMgMt*
BDL_21	*MgMt*
BDL_22	*MgMt*
BDL_23	*Mt*
BDL_24	*MeMgMt*
BDL_25	*MeMgMt*
BDL_26	*MgMt*
BDL_27	*MeMgMt*
BDL_28	*MeMgMt*
BDL_29	*MgMt*
BDL_30	*MgMt*
BDL_31	*MgMt*
BDL_32	*MeMt*
BDL_33	*MgMt*
BDL_34	*MeMgMt*
BDL_35	*MeMgMt*
BDL_36	*MeMgMt*
BDL_37	*MgMt*
BDL_38	*MgMt*
BDL_39	*MeMgMt*
BDL_40	*MeMgMt*
BDL_41	*MgMt*
BDL_42	*MgMt*
BDL_43	*MeMgMt*
BDL_44	*MeMgMt*
BDL_45	*MeMgMt*
BDL_46	*MeMgMt*
BDL_47	*MeMgMt*
BDL_48	*MgMt*
BDL_49	*MgMt*
BDL_50	*MeMgMt*

vi. Loch Etive

Individual ID	Genotype
LET_01	*Mt*
LET_02	*MgMt*
LET_03	*MeMgMt*
LET_04	*MeMt*
LET_05	*MeMgMt*
LET_06	*MeMt*
LET_07	*MgMt*
LET_08	*MeMt*
LET_09	*MeMt*
LET_10	HXT
LET_11	HXT
LET_12	*MeMt*
LET_13	*MeMt*
LET_14	*MgMt*
LET_15	*MeMgMt*
LET_16	*MgMt*
LET_17	*MeMt*
LET_18	*MeMt*
LET_19	*MgMt*
LET_20	*MeMgMt*

APPENDIX 6D – Complete list of composite introgressed genotypes for each genotype class identified in individuals used for marker development and marker validation, *n*I is the total number of individuals with a given genotype class; *n*P refers to the number of populations each genotype class appeared in

CLASS	Composite genotype	BP	RB	LR	BDL	LET	PC	*n*I	*n*P
HXE	GGTTATCCTTAACCAATTCCCCAA		8	2				10	2
HXE	AAGTATCCTTAACCAATTCCCCAA			1				1	1
HXE	AGGTAACCTTAACCAATTCCCCAA			2				2	1
HXE	GGGTAACCTTAACCAATTCCCCAA			1				1	1
HXG	AAGGTTAAGGAGGGAATTCCCCAA	1						1	1
HXG	AAGGTTAAGGGGCCAATTCCCCAA	1						1	1
HXG	AAGGTTAAGGGGCGAATTCCCCAA	2						2	1
HXT	AAGGTTCCTTAACCAGGGTTTTTT					1		1	1
HXT	AAGGTTCCTTAACCGGGTTTCTTT					1		1	1
HXT	AAGGTTCCTTAACCGGGGTTCTTT				1		1	2	2
MeMg	GGTTAACCGGAACCAATTCCCCAA		1	1				2	2
MeMg	GGTTAACCTTAGCCAATTCCCCAA			2		2		4	2
MeMg	AAGGATAAGGGGGGAATTCCCCAA	2						2	1
MeMg	AAGTTTCCTTAACGGGGGTTTTTT				1			1	1
MeMg	AATTAACCTTAAGCAATTCCCCAA			1				1	1
MeMg	AGGGATAAGGGGGGAATTCCCCAA	1						1	1
MeMg	AGGGTTAAGGGGGGAATTCCCCAA	1						1	1
MeMg	GGGTAACCTTAAGCAATTCCCCAA			1				1	1
MeMg	GGTTAAAAGGAACCAATTCCCCAA			1				1	1
MeMg	GGTTAAAATTAACCAATTCCCCAA			1				1	1
MeMg	GGTTAAACGTAGCGAATTCCCCAA			1				1	1
MeMg	GGTTAACCTTAACGAATTCCCCAA			1				1	1
MeMg	GGTTAACCTTAAGGAATTCCCCAA					1		1	1
MeMgMt	AGGGAAACTTAACCAGGGCTCTTT					1		1	1
MeMgMt	AGGGTTACTTAACGGGGGTTTTTT				1			1	1
MeMgMt	AGGGTTCCGTAACCAGGGTTTTTT					1		1	1
MeMgMt	AGGGTTCCGTAACGGGGGTTTTTT				1			1	1
MeMgMt	AGGGTTCCTTAACGAGGGCTCTTT					1		1	1
MeMgMt	AGGGTTCCTTAACGGGGGTTTTTT				6			6	1
MeMgMt	AGGGTTCCTTAACGGGGTTTTTTT				1			1	1
MeMgMt	AGGGTTCCTTAAGGGGGGCTTTTT				1			1	1
MeMgMt	AGGGTTCCTTAAGGGGGGTTTTTT				3			3	1
MeMgMt	AGGGTTCCTTAGCCGGGGTTTTTT					1		1	1
MeMgMt	AGGTTTCCTTAACCGGGGTTCTTT					1		1	1
MeMgMt	GGGGTTCCGTAAGGGGGTTTTTTT				1			1	1
MeMgMt	GGGGTTCCTTAACCGGGGTTCTTT				2			2	1
MeMgMt	GGGGTTCCTTAAGGGGGGTTTTTT				1			1	1
MeMt	AAGGATCCTTAACCAGGGCTTTTT					1		1	1
MeMt	GGTTAACCTTAACCAGTTCCCCAA		2	2				4	2
MeMt	AGGTTTCCTTAACCGGGGTTTTTT				1		1	2	2
MeMt	AGTTAACCTTAACCAATTCCCCAA		1	1				2	2
MeMt	AAGGATCCTTAACCGGGGTTTTTT					2		2	1
MeMt	AAGTTTCCTTAACCAGGGTTTTTT					2		2	1
MeMt	AGGGATCCTTAACCGGGTCTTTTT					1		1	1
MeMt	AGGGTTCCTTAACCAGGGCTTTTT					1		1	1
MeMt	AGGGTTCCTTAACCGGGGTTTTTT				3			3	1
MeMt	AGGGTTCCTTAGGGGGGGTTTTTT				1			1	1
MeMt	AGGTATCCTTAACCAGGTTTCTTT					1		1	1
MgMt	AAGGTTAAGGGGGGAAGTCCCCAA	1						1	1
MgMt	AAGGTTCCTTAACGGGGGTTTTTT				4	2		6	2
MgMt	AAGGTTCCTTAAGGGGGGTTTTTT				7	1		8	2
MgMt	AAGGTTAAGGGGGGAGTTCCCCAA	3						3	1
MgMt	AAGGTTCCGGAACCGGGGTTTTTT						1	1	1
MgMt	AAGGTTCCGTAACCGGGGTTTTTT				1			1	1
MgMt	AAGGTTCCGTAACGGGGGTTTTTT				1			1	1
MgMt	AAGGTTCCTTAACGAGGGTTCTTT					1		1	1
MgMt	AAGGTTCCTTAACGGGGTTTTTTT				1			1	1
MgMt	AAGGTTCCTTAAGGAGGGTTTTTT				1			1	1

MgMt	AAGGTTCCTTAAGGGGGGTTCCTT		2	2	1
MgMt	AAGGTTCCTTAAGGGGGGTTCTTT	1		1	1
MgMt	AAGGTTCCTTAGCCGGGGTTTTTT	3		3	1
MgMt	AGGGTTCCTTAACGGGGGTTCTTT	1		1	1
MgMt	AGGGTTCCTTAGGGGGGTTTTTTT	1		1	1
MgMt	GGGGTTACTTAACGGGGGTTTTTT	1		1	1
MgMt	GGGGTTCCTTAACCAAGGTTTTTT	1		1	1

APPENDIX 7

APPENDIX 7A – Complete list of composite genotypes for each genotype class identified in 22 Scottish sites used for DAPC analysis, *n*I is the total number of individuals with a given genotype class; *n*P refers to the number of populations each genotype class appeared in

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	nI	nP
Me	GGTTAACCTTAACCAATTCCCCAA	19		5	28	3	3	26	4	3	54	11	26	32		23	29		36		38		27	386	18
Mt	AAGGTTCCTTAACCGGGGTTTTTT									2														7	1
HXE	AAGTATCCTTAACCAATTCCCCAA													1										1	1
HXE	AGGTAACCTTAACCAATTCCCCAA													2										2	1
HXE	AGGTATCCTTAACCAATTCCCCAA									1														1	1
HXE	AGGTNNCCTTAACCAATTCCCCAA									1														1	1
HXE	GGGTTTCCTTAACCAATTCCCCAA					2																		2	1
HXE	GGGGAACCTTAACCAATTCCCCAA									4									1					5	2
HXE	GGGTATCCTTAACCAATTCCCCAA											2												3	1
HXE	GGTTTTCCTTAACCAATTCCCCAA	2	2			1			4				1		1				1					12	7
HXE	GGGTAACCTTAACCAATTCCCCAA	2	1	1		2	2		2	2	3	1		1		1	4							24	12
HXE	GGTTATCCTTAACCAATTCCCCAA				2			3	2		2	3	4	2		4	1	4	2		8		3	42	13
HXT	AAGGTTCCTTAACCAAGGTTCTTT											1												1	1
HXT	AAGGTTCCTTAACCAGGGTTTTTT									1														1	1
HXT	AAGGTTCCTTAACCGGGGTTCTTT										1													3	1
HXT	AAGGTTCCTTAACCGGGTTTCTTT									1														1	1
HXT	AAGGTTCCTTAACCGGGTTTTTTT									1	1													2	2
MeMg	AAGTTTCCTTAACGGGGGTTTTTT									1														2	1
MeMg	AATTAACCGGGGGGAATTCCCCAA														1									1	1
MeMg	AATTAACCGTGGCCAATTCCCCAA														1									1	1
MeMg	AATTAACCTTAAGCAATTCCCCAA													1										1	1
MeMg	AGGGATCCTTGGCGAATTCCCCAA																			1				1	1
MeMg	AGGTAAAAGGAGCGAATTCCCCAA																					1		1	1
MeMg	AGGTAAAAGTAGGGAATTCCCCAA																			1				1	1
MeMg	AGGTAAACGGGGGGAATTCCCTAA						1																	1	1
MeMg	AGGTAAACGTAGCCAATTCCCCAA						1																	1	1
MeMg	AGGTAAACGTAGCGAATTCCCCAA																						1	1	1
MeMg	AGGTAACCGGAGCCAATTCCCCAA	1																						1	1
MeMg	AGGTAACCGTAAGGAATTCCCCAA					1																		1	1
MeMg	AGGTAACCGTGGCCAATTCCCCAA					1																		1	1
MeMg	AGGTAACCTTAAGGAATTCCCCAA		1																					1	1
MeMg	AGGTAACCTTAGCCAATTCCCCAA											1												1	1
MeMg	AGGTATACTTAAGGAATTCCCCAA			1																				1	1
MeMg	AGGTATCCTTAACGAATTCCCCAA														1									1	1
MeMg	AGGTATCCTTAGCGAATTCCCCAA																						1	1	1
MeMg	AGGTTTAAGGAGCCAATTCCCCAA																			1				1	1
MeMg	AGGTTTACGGAGGGAATTCCCCAA						1																	1	1
MeMg	AGGTTTCCGTAACGAATTCCCCAA																					1		1	1
MeMg	AGTTAAAAGGGGGGAATTCCCCAA																			1				1	1
MeMg	AGTTAAAAGTAAGGAATTCCCCAA					1																		1	1

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMg	AGTTAAACGTAACCAATTCCCCAA			1																				1	1
MeMg	AGTTAAACGTAAGGAATTCCCCAA														1									1	1
MeMg	AGTTAAACGTAGCCAATTCCCCAA			1																				1	1
MeMg	AGTTAAACGTGGGGAATTCCCCAA					1																		1	1
MeMg	AGTTAAACTTAACGAATTCCCCAA					1																		1	1
MeMg	AGTTAACCGGAGCGAATTCCCCAA																					1		1	1
MeMg	AGTTAACCGTAGCCAATTCCCCAA			1																				1	1
MeMg	AGTTAACCTTAGCCAATTCCCCAA						1																	1	1
MeMg	AGTTAACCTTGGGGAATTCCCCAA																	1						1	1
MeMg	AGTTATACGTAGCCAATTCCCCAA						1																	1	1
MeMg	AGTTATCCGTGGGGAATTCCCCAA														1									1	1
MeMg	AGTTTTCCGGAAGGAATTCCCCAA																					1		1	1
MeMg	AGTTTTCCGTGGCGAATTCCCCAA		1																					1	1
MeMg	AGTTTTCCTTAACGAATTCCCCAA		1																					1	1
MeMg	AGTTTTCCTTAGCCAATTCCCCAA	1																						1	1
MeMg	AGTTTTCCTTAGGGAATTCCCCAA		1																					1	1
MeMg	AGTTTTCCTTGGCGAATTCCCCAA																	1						1	1
MeMg	GGGGAAACGGAACGAATTCCCCAA																					1		1	1
MeMg	GGGGAAACTTAGCCAATTCCCCAA																		1					1	1
MeMg	GGGGAACCTTAGCCAATTCCCCAA														1									1	1
MeMg	GGGGATCCTTAACGAATTCCCCAA																	1						1	1
MeMg	GGGGTTACGGAACCAATTCCCCAA	1																						1	1
MeMg	GGGTAAAATTAACCAATTCCCCAA					1																		1	1
MeMg	GGGTAAACGGGGCGAATTCCCCAA																					1		1	1
MeMg	GGGTAAACGGGGGGAATTCCCCAA					1																		1	1
MeMg	GGGTAAACGTAACGAATTCCCCAA			1																				1	1
MeMg	GGGTAAACGTAGCCAATTCCCCAA			1																				1	1
MeMg	GGGTAAACTTGGCCAATTCCCCAA					1																		1	1
MeMg	GGGTAACCGGAAGGAATTCCCCAA						1																	1	1
MeMg	GGGTAACCGGAGCCAATTCCCCAA			2																				2	1
MeMg	GGGTAACCGGAGGGAATTCCCCAA																					1		1	1
MeMg	GGGTAACCGTAACCAATTCCCCAA			1																				1	1
MeMg	GGGTAACCGTAAGGAATTCCCCAA					1																		1	1
MeMg	GGGTAACCTTAACGAATTCCCCAA					1																		1	1
MeMg	GGGTAACCTTAAGCAATTCCCCAA													1										1	1
MeMg	GGGTAACCTTAAGGAATTCCCCAA	1																						1	1
MeMg	GGGTAACCTTAGGGAATTCCCCAA						1																	1	1
MeMg	GGGTATACGGAAGGAATTCCCCAA						1																	1	1
MeMg	GGGTATACTTAACCAATTCCCCAA						1																	1	1
MeMg	GGGTATCCGTAACCAATTCCCCAA					1																		1	1
MeMg	GGGTATCCGTAAGGAATTCCCCAA						1																	1	1
MeMg	GGGTATCCTTAGCCAATTCCCCAA							1																1	1
MeMg	GGGTATCCTTAGCGAATTCCCCAA																						1	1	1
MeMg	GGGTATCCTTGGCCAATTCCCCAA								1															1	1
MeMg	GGGTTTACTTAGCCAATTCCCCAA			1																				1	1
MeMg	GGGTTTCCGTGGCGAATTCCCCAA	1																						1	1
MeMg	GGGTTTCCTTAAGGAATTCCCCAA		2																					2	1
MeMg	GGGTTTCCTTGGCGAATTCCCCAA		1																					1	1
MeMg	GGGTTTCCTTGGGGAATTCCCCAA		1																					1	1
MeMg	GGTTAAAAGGAACCAATTCCCCAA													1										1	1

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMg	GGTTAAAAGTAAGGAATTCCCCAA						1																	1	1
MeMg	GGTTAAAAGTAGCCAATTCCCCAA			1																				1	1
MeMg	GGTTAAAAGTAGGGAATTCCCCAA					1																		1	1
MeMg	GGTTAAAAGTGGCCAATTCCCCAA														1									1	1
MeMg	GGTTAAACGGAGGGAATTCCCCAA																						1	1	1
MeMg	GGTTAAACGTAACCAATTCCCCAA			1																				1	1
MeMg	GGTTAAACGTAAGGAATTCCCCAA														1									1	1
MeMg	GGTTAAACGTAGCGAATTCCCCAA													1										1	1
MeMg	GGTTAAACGTAGGGAATTCCCCAA					1																		1	1
MeMg	GGTTAAACTTAGCCAATTCCCCAA						1																	1	1
MeMg	GGTTAAACTTAGGGAATTCCCCAA						1																	1	1
MeMg	GGTTAAACTTGGGGAATTCCCCAA					1																		1	1
MeMg	GGTTAACCGGAGGGAATTCCCCAA						1																	1	1
MeMg	GGTTAACCGGGGCCAATTCCCCAA					1																		1	1
MeMg	GGTTAACCGTAGCGAATTCCCCAA																			1				1	1
MeMg	GGTTATAATTGGCCAATTCCCCAA								1															1	1
MeMg	GGTTATACGTAACCAATTCCCCAA						1																	1	1
MeMg	GGTTATACGTAGCCAATTCCCCAA						1																	1	1
MeMg	GGTTATACGTAGCGAATTCCCCAA			1																				1	1
MeMg	GGTTATACTTAACGAATTCCCCAA											1												1	1
MeMg	GGTTATACTTAAGGAATTCCCCAA						1																	1	1
MeMg	GGTTATACTTAGCGAATTCCCCAA			1																				1	1
MeMg	GGTTATACTTGGCGAATTCCCCAA								1															1	1
MeMg	GGTTATCCGTAACGAATTCCCCAA			2																				2	1
MeMg	GGTTATCCGTAAGGAATTCCCCAA																	1						1	1
MeMg	GGTTATCCGTGGCCAATTCCCCAA			1																				1	1
MeMg	GGTTATCCGTGGGGAATTCCCCAA		1																					1	1
MeMg	GGTTATCCTTAAGGAATTCCCCAA																	2						2	1
MeMg	GGTTATCCTTAGCGAATTCCCCAA					1																		1	1
MeMg	GGTTATCCTTAGGGAATTCCCCAA						2																	2	1
MeMg	GGTTTTAATTAAGGAATTCCCCAA		1																					1	1
MeMg	GGTTTTAATTGGCCAATTCCCCAA														1									1	1
MeMg	GGTTTTACGGAGGGAATTCCCCAA						1																	1	1
MeMg	GGTTTTCCGGGGCCAATTCCCCAA					1																		1	1
MeMg	GGTTTTCCGTAACGAATTCCCCAA					1																		1	1
MeMg	GGTTTTCCGTGGGGAATTCCCCAA					1																		1	1
MeMg	GGTTTTCCTTGGCCAATTCCCCAA														2									2	1
MeMg	GGTTTTCCTTGGCGAATTCCCCAA		1																					1	1
MeMg	GGTTTTCCTTGGGGAATTCCCCAA														1									1	1
MeMg	AGTTAACCGTAACCAATTCCCCAA											1												2	1
MeMg	GGGTAAACTTAACCAATTCCCCAA			1																			1	2	2
MeMg	GGTTAAAATTAACCAATTCCCCAA						2							1										3	2
MeMg	GGTTAAACTTGGCCAATTCCCCAA			1		2																		3	2
MeMg	GGTTAACCGGAACGAATTCCCCAA																	1				1		2	2
MeMg	GGTTAACCGTGGGGAATTCCCCAA					1									1									2	2
MeMg	GGTTAACCTTAGCGAATTCCCCAA			3				1																4	2
MeMg	GGTTAACCTTGGCGAATTCCCCAA	1	2																					3	2
MeMg	GGTTATCCTTGGGGAATTCCCCAA		1															1						2	2
MeMg	GGTTTTCCGGAGCCAATTCCCCAA					1									1									2	2
MeMg	GGTTTTCCGTAACCAATTCCCCAA	1													1									2	2

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMg	GGTTTTCCTTAGCCAATTCCCCAA						1			1														2	2
MeMg	AGTTAACCTTAACGAATTCCCCAA					1		1																3	2
MeMg	GGGTAACCTTAGCCAATTCCCCAA		1	1			1																	3	3
MeMg	GGTTAAACTTAACCAATTCCCCAA					2										1							2	5	3
MeMg	GGTTAACCGGGGGGAATTCCCCAA					1									1					1				3	3
MeMg	GGTTAACCGTAAGGAATTCCCCAA		1			1									1									3	3
MeMg	GGTTAACCGTAGCCAATTCCCCAA	1		1											1									3	3
MeMg	GGTTAACCGTGGCCAATTCCCCAA														1			1				1		3	3
MeMg	GGTTAACCTTAGGGAATTCCCCAA	1	1				1																	3	3
MeMg	GGTTATACTTAACCAATTCCCCAA								1		1					1								3	3
MeMg	GGTTATCCGTAACCAATTCCCCAA			1	1													1						3	3
MeMg	GGTTATCCTTAGCCAATTCCCCAA						1									2		1						4	3
MeMg	GGTTAAACTTAACGAATTCCCCAA			2												1				1				5	3
MeMg	GGTTAACCTTAAGGAATTCCCCAA		3				2			1					2									8	4
MeMg	GGTTAACCTTGGGGAATTCCCCAA	1	2			1									1									5	4
MeMg	GGTTAACCGGAACCAATTCCCCAA					1		2						1	1						1			6	5
MeMg	GGTTATCCTTGGCCAATTCCCCAA		1				1		2						3			1						8	5
MeMg	GGTTAACCTTGGCCAATTCCCCAA					1			3						2	1		1	1					9	6
MeMg	GGTTAACCGTAACCAATTCCCCAA	1		4		1	1				1					3	1					1	1	16	9
MeMg	GGTTAACCTTAACGAATTCCCCAA		1	1	1				1		3		1	1		1	1							12	9
MeMg	GGTTAACCTTAGCCAATTCCCCAA	3		2	2	1		2	1	2	1	3	1	2		4	1		2				1	29	15
MeMgMt	AAGTTTCCTTAACGAGGGTTTTTT									1														1	1
MeMgMt	AATTTTCCTTAACGAGGGTTTTTT									1														1	1
MeMgMt	AATTTTCCTTAACGGGGTTTCCTT									1														1	1
MeMgMt	AGGGAAACTTAACCAGGGCTCTTT									1														1	1
MeMgMt	AGGGTTCCGTAACCAGGGTTTTTT									1														1	1
MeMgMt	AGGGTTCCTTAACGAGGGCTCTTT									1														1	1
MeMgMt	AGGGTTCCTTAACGAGGGTTCTTT									1														1	1
MeMgMt	AGGGTTCCTTAACGGGGGTTCCTT									1														1	1
MeMgMt	AGGGTTCCTTAAGGGGGGCTTTTT									1														2	1
MeMgMt	AGGGTTCCTTAGCCGGGGTTTTTT									1														1	1
MeMgMt	AGGTAAAAGGAGGGAAGTCCCCAA																			1				1	1
MeMgMt	AGGTATACTTAGCCAAGTCTCCTT						1																	1	1
MeMgMt	AGGTATCCGTAGGGAAGTTTTTTT						1																	1	1
MeMgMt	AGGTTTCCTTAACCGGGGTTCTTT									1														1	1
MeMgMt	AGGTTTCCTTAGCCAAGTCCCCAT								1															1	1
MeMgMt	AGTTAACCGTAACCAAGGCCCCAA						1																	1	1
MeMgMt	AGTTAACCGTGGCCAATTCCCCAT														1									1	1
MeMgMt	AGTTAACCGTGGGGAGTTCCCCAA														1									1	1
MeMgMt	AGTTAACCTTAACGAGTTCCCTAA		1																					1	1
MeMgMt	AGTTAACCTTAGGGGGTTCCCCAA		1																					1	1
MeMgMt	AGTTAACCTTGGCGAATTCCCTAA			1																				1	1
MeMgMt	AGTTATCCTTAGCGAGTTCCCCTT	1																						1	1
MeMgMt	AGTTTTCCTTGGCGAATTCCCCAT					1																		1	1
MeMgMt	GGGGAAACTTAGCCAGTTCCCCAA																		1					1	1
MeMgMt	GGGGAACCGTGGCCAAGTCCCCAA														1									1	1
MeMgMt	GGGGAACCTTAACGGGGGCTCTTT									1														1	1
MeMgMt	GGGGAACCTTAAGGAGGGTTTTTT									1														1	1
MeMgMt	GGGGATCCTTAACGGGGGTTTTTT									2														2	1
MeMgMt	GGGTAAACTTAGCCAGTTCCCCAA			1																				1	1

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMgMt	GGGTAAACTTAGCCGGTTCCCCAT						1																	1	1
MeMgMt	GGGTAAACTTGGGGGGTTCCCCAA		1																					1	1
MeMgMt	GGGTAACCGGAGGGAGTTCCCCAA			1																				1	1
MeMgMt	GGGTAACCGGGGCCAATTCCCCAT														1									1	1
MeMgMt	GGGTAACCGTGGCCAGTTCCCCAA					1																		1	1
MeMgMt	GGGTAACCTTAACGAGGGTTCCTT									1														1	1
MeMgMt	GGGTAACCTTAACGGGTTCCCCAA		1																					1	1
MeMgMt	GGGTAACCTTAAGGGGTTCCCCAA		1																					1	1
MeMgMt	GGGTAACCTTGGCGGGTTCCCCAA		1																					1	1
MeMgMt	GGGTAACCTTGGCGGGTTCTCCAT		1																					1	1
MeMgMt	GGGTAACCTTGGGGGGTTCCCCAA		1																					1	1
MeMgMt	GGGTATCCTTAACGGGGTTTTTTT									1														1	1
MeMgMt	GGGTTTCCGGAGCCAATTCCCCAT					1																		1	1
MeMgMt	GGGTTTCCTTGGCGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAAACTTGGGGAAGTCCCCAA						1																	1	1
MeMgMt	GGTTAACCGGAACCAATTCCCTAA							1																1	1
MeMgMt	GGTTAACCGGAAGGGGGTCCCCAA		1																					1	1
MeMgMt	GGTTAACCGGGGCCAATTCCCTAA														1									1	1
MeMgMt	GGTTAACCGGGGCGAGTTCCCCAA														1									1	1
MeMgMt	GGTTAACCGGGGGGAGTTCCCTAA														1									1	1
MeMgMt	GGTTAACCGTAACGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAACCGTAAGGAATTCCCCAT														1									1	1
MeMgMt	GGTTAACCGTAAGGAATTCCCTAA														1									1	1
MeMgMt	GGTTAACCGTGGCCGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAACCGTGGGGAAGTCCCCAA														1									1	1
MeMgMt	GGTTAACCGTGGGGAATTCCCCAT														1									1	1
MeMgMt	GGTTAACCGTGGGGAATTCCCTAA					1																		1	1
MeMgMt	GGTTAACCTTAACGAAGTCCCCAA										1													1	1
MeMgMt	GGTTAACCTTAAGGAGTTCCCCAA														1									1	1
MeMgMt	GGTTAACCTTAAGGAGTTCCCTAT							1																1	1
MeMgMt	GGTTAACCTTAGCCAATTCCCTAA						1																	1	1
MeMgMt	GGTTAACCTTAGGGAGTTCCCCAA											1												1	1
MeMgMt	GGTTAACCTTAGGGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAACCTTGGCCAATTCCCCAT								2															2	1
MeMgMt	GGTTAACCTTGGCCAATTCCCTAA																	1						1	1
MeMgMt	GGTTAACCTTGGCCAATTCTCCAA	1																						1	1
MeMgMt	GGTTAACCTTGGGGAATTCCCCTT															1								1	1
MeMgMt	GGTTAACCTTGGGGAGTTCCCCAA						1																	1	1
MeMgMt	GGTTAACCTTGGGGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAACCTTGGGGGGTTCCCTAA																	1						1	1
MeMgMt	GGTTAACCTTGGGGGGTTCTCTAA																	1						1	1
MeMgMt	GGTTATAAGTAGCGAATTCTCCAA																			1				1	1
MeMgMt	GGTTATAAGTAGGGAATTCCCCAT																			1				1	1
MeMgMt	GGTTATAAGTGGCCGGTTCCCCAT														1									1	1
MeMgMt	GGTTATAATTAACGAGTTCCCCAT						1																	1	1
MeMgMt	GGTTATCCGGAACCAGTTCCCCAA																	1						1	1
MeMgMt	GGTTATCCGGAACCGGTTCCCCAA																	1						1	1
MeMgMt	GGTTATCCGTAACCAGTTCCCCAA																	1						1	1
MeMgMt	GGTTATCCGTAACGAGTTCCCCAA					1																		1	1
MeMgMt	GGTTATCCGTAAGGAGTTCCCCAA														1									1	1

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMgMt	GGTTATCCGTAGCGAATTCCCCAT	1																						1	1
MeMgMt	GGTTATCCTTAAGGAATTCCTTAT						1																	1	1
MeMgMt	GGTTATCCTTAGCCAATTTTCCAA				1																			1	1
MeMgMt	GGTTATCCTTAGCGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTATCCTTGGCCAATTCCCCTT		1																					1	1
MeMgMt	GGTTATCCTTGGCCAATTCTCCAA																	1						1	1
MeMgMt	GGTTATCCTTGGGGAATTTTTTAA																	1						1	1
MeMgMt	GGTTTTACTTGGGGAAGTCCCCAA														1									1	1
MeMgMt	GGTTTTCCGGGGCCAATTCCCCAT														1									1	1
MeMgMt	GGTTTTCCGGGGCCAATTCTCCAA														1									1	1
MeMgMt	GGTTTTCCGGGGGGAATTCTCCAT														1									1	1
MeMgMt	GGTTTTCCGTAACCAGTTCCCCAA												1											1	1
MeMgMt	GGTTTTCCGTAACGGGGTCCCCAA		1																					1	1
MeMgMt	GGTTTTCCTTAACGAATTCTCCAA	1																						1	1
MeMgMt	GGTTTTCCTTAAGGAATTCCCTAA														1									1	1
MeMgMt	GGTTTTCCTTAGCGGGTTCCCCAA		1																					1	1
MeMgMt	GGTTAACCGTAAGGAGTTCCCCAA					1									1									2	2
MeMgMt	GGTTAACCTTAACGAGTTCCCCAA										1						1							2	2
MeMgMt	GGTTAACCTTGGCCGGTTCCCCAA	1																1						2	2
MeMgMt	GGTTAACCTTAGCCAATTCCCCAT	2											1			1								4	3
MeMgMt	GGTTAACCTTGGCCAGTTCCCCAA	1		1					1															3	3
MeMt	AGTTATCCTTAACCAGTTCCCCAA																		1					1	1
MeMt	AAGGATCCTTAACCAGGGCTTTTT									1														1	1
MeMt	AAGGATCCTTAACCGGGGTTTTTT									2														2	1
MeMt	AAGTAACCTTAACCGGGGTTTTTT									1														1	1
MeMt	AAGTATCCTTAACCAGGGTTCTTT									1														1	1
MeMt	AAGTTTCCTTAACCAAGGTTTTTT										1													1	1
MeMt	AAGTTTCCTTAACCAGGGTTTTAT										1													1	1
MeMt	AAGTTTCCTTAACCAGGGTTTTTT									2														2	1
MeMt	AAGTTTCCTTAACCAGGTCTCCAT										1													1	1
MeMt	AATTATCCTTAACCAGGTCTCTTT									1														1	1
MeMt	AATTTTCCTTAACCGGGGTTTTTT									1														1	1
MeMt	AGGGAACCTTAACCAAGGCTCCAT									1														1	1
MeMt	AGGGAACCTTAACCAGGTCTCTTT									1														1	1
MeMt	AGGGATCCTTAACCGGGGTTTTTT									1														1	1
MeMt	AGGGATCCTTAACCGGGTCTTTTT									1														1	1
MeMt	AGGGNNCCTTAACCAGGGTTTTTT									1														1	1
MeMt	AGGGTTCCTTAACCAGGGCTTTTT									1														1	1
MeMt	AGGGTTCCTTAACCAGTTCTCTTT									1														1	1
MeMt	AGGGTTCCTTAACGAAGGTTTTTT									1														1	1
MeMt	AGGTAACCTTAACCAGGTCTCTAT										1													1	1
MeMt	AGGTATCCTTAACCAGGGCTTTTT									1														1	1
MeMt	AGGTATCCTTAACCAGGTCCCCAT									1														1	1
MeMt	AGGTATCCTTAACCAGGTTTCTTT									1														1	1
MeMt	AGGTATCCTTAACCGGGGTTCTAT									1														1	1
MeMt	AGGTTTCCTTAACCGGGGTTCCTT										1													1	1
MeMt	AGTTAACCTTAACCAATTCCCTAA																1							1	1
MeMt	AGTTAACCTTAACCAATTCTCCAT																1							1	1
MeMt	AGTTAACCTTAACCAATTCTTTAA																						1	1	1
MeMt	AGTTATCCTTAACCAAGTCCCTAT									1														1	1

CLASS	Composite genotype	AIL	BR	BX	DF	EIR	FL	FYN	KY	LET	LFY	LIN	LL	LR	LRG	LSP	LUN	LX	MON	NS	RB	SCB	STA	*n*I	*n*P
MeMt	AGTTATCCTTAACCAGGTCTCCAT										1													1	1
MeMt	AGTTTTCCTTAACCAGGGTTCCTT									1														1	1
MeMt	AGTTTTCCTTAACCAGGTCTCTTT									1														1	1
MeMt	GGGGAACCTTAACCAGGGTTTTTT									1														1	1
MeMt	GGGGTTCCTTAACCAATTTTCTTT											1												1	1
MeMt	GGGGTTCCTTAACCAGGGTTCCTT									1														1	1
MeMt	GGGTAACCTTAACCAAGTCCCCAA												1											1	1
MeMt	GGGTAACCTTAACCAATTCCCTAA																1							1	1
MeMt	GGGTAACCTTAACCAATTCTCCAA										1													1	1
MeMt	GGGTAACCTTAACCAGGTCTTTTT									1														1	1
MeMt	GGGTAACCTTAACCAGTTCCCCAA																1							1	1
MeMt	GGGTAACCTTAACCAGTTCTCCAT									1														1	1
MeMt	GGGTAACCTTAACCGGGGTTCTTT									1														1	1
MeMt	GGGTATCCTTAACCAGGGCTTTTT									1														1	1
MeMt	GGGTATCCTTAACCAGGTCTCCAT										1													1	1
MeMt	GGGTATCCTTAACCGGGGTTTTTT									1														1	1
MeMt	GGTTAACCTTAACCAAGTCCCCAA																						2	2	1
MeMt	GGTTAACCTTAACCAAGTCCCCAT	1																						1	1
MeMt	GGTTAACCTTAACCAAGTCTCTAT												1											1	1
MeMt	GGTTAACCTTAACCAATTCCCTAT										1													1	1
MeMt	GGTTAACCTTAACCAGTTCCTTAA												1											1	1
MeMt	GGTTAACCTTAACCGGGGTTTTTT									1														1	1
MeMt	GGTTAACCTTAACCGGTTCCCCAA				1																			1	1
MeMt	GGTTAACCTTAACCGGTTCTCCAA							1																1	1
MeMt	GGTTATCCTTAACCAATTCCCTAA								1															1	1
MeMt	GGTTATCCTTAACCAATTCTCCTT												1											1	1
MeMt	GGTTATCCTTAACCAGGGTTTTTT									1														1	1
MeMt	GGTTATCCTTAACCAGTTCCCCTT																	1						1	1
MeMt	GGTTATCCTTAACCAGTTCCCTAA							1																1	1
MeMt	GGTTTTCCTTAACCAATTCCCCAT														1									1	1
MeMt	GGTTTTCCTTAACCAATTCTCCAA				1																			1	1
MeMt	GGTTTTCCTTAACCAATTTTCCAA																	1						1	1
MeMt	GGTTTTCCTTAACCAATTTTCCTT																	1						1	1
MeMt	GGTTAACCTTAACCAATTCCCCTT							1																4	1
MeMt	GGTTAACCTTAACCAATTCCTTAA	1						3																4	2
MeMt	GGTTAACCTTAACCAATTTTCCAA				1																		1	2	2
MeMt	GGTTAACCTTAACCAATTTTTTAA					1																	1	2	2
MeMt	GGTTATCCTTAACCAATTCCCCAT	2											1											3	2
MeMt	GGTTATCCTTAACCAATTCTCCAA											1	1											2	2
MeMt	GGTTATCCTTAACCAGTTCCCCAA																	1						2	1
MeMt	GGTTAACCTTAACCAATTCCCTAA	1						2			1					1							1	6	5
MeMt	GGTTAACCTTAACCAATTCCCCAT					1		2			2					5	1		1				1	13	7
MeMt	GGTTAACCTTAACCAATTCTCCAA				1						6		1				1		1				1	12	6
MeMt	AGTTAACCTTAACCAATTCCCAA			1	1		1			2				1					1		1		2	11	8
MeMt	GGTTAACCTTAACCAGTTCCCCAA							2		1	5	1	3	2			1		1		2			19	9
MgMt	AAGGTTCCTTAACGAGGGTTCTTT									1														1	1
MgMt	AAGGTTCCTTAACGGGGGTTCTTT									2														2	1
MgMt	AAGGTTCCTTAACGGGGGTTTTTT									2														6	1
MgMt	AAGGTTCCTTAAGGGGGGTTTTTT									1														8	1

APPENDIX 8

The following comprises a collection of articles that were distributed in publications accessible by both a general audience and potential stakeholders. All articles were written to raise awareness of the current research and the possible benefits it could bring to Scottish shellfish aquaculture. Some of the articles used the same pictures, the figure references for which have been altered so that this reads as a complete chapter.

APPENDIX 8A – ARTICLE 1

HYBRID MUSSELS IN SCOTLAND

Author: Joanna Wilson

The following article was published in the Scottish Consortium for Rural Research Newsletter in July 2014. The final copy is accessible under the following link: www.scrr.ac.uk/downloads/scrr-news-80.pdf

Plans are afoot to encourage the development of sustainable and productive shellfish farms throughout Scotland, where the shellfish sector of the aquaculture industry has grown strongly over the last ten years. There is only one problem with these plans: farmers and researchers are not sure what mussel species live in Scottish waters. Parts of the Scottish shellfish sector have suffered in recent years due to increased numbers of a mussel with a particularly fragile shell and a reduced meat yield. The latest studies of mussel populations in Scotland have revealed these more fragile specimens are often hybrids of the farmed blue mussel, *Mytilus edulis*, and the bay mussel, *Mytilus trossulus*. Hybrids could, therefore, negatively affect productivity and profitability if such characteristics make them less marketable, and could cause problems for mussel growers if their presence is left unchecked. Farmers need to know what species live in Scottish waters to effectively manage their stocks and to mitigate any possible risks associated with an expansion in hybrid range. Existing genetic tests are not sensitive enough to detect all mussel species in Scotland. To enable a better assessment of population genetics, new markers are currently being developed from large scale sequencing of the mussel genome. This collaborative research between The Institute of Aquaculture (University of Stirling) and Marine Scotland Science (Aberdeen) will look at locations throughout Scotland to give a more complete picture of mussel species existing in the country.

APPENDIX 8B – ARTICLE 2

MUSSELS: DO YOU KNOW WHAT SPECIES ARE GROWING ON YOUR FARM?

Author: Joanna Wilson

The following article was published in the Association of Scottish Shellfish Growers summer newsletter in July 2014. The final copy is accessible under the following link:

http://assg.org.uk/#/the-grower/4532754744

The blue mussel, *Mytilus edulis*, is a popular source of food throughout the world and is an important contributor to the aquaculture industries of many countries. The shellfish sector of Scottish aquaculture has seen a strong growth over the last ten years. There are measures in place to increase the development of sustainable and productive mussel farms within Scotland, but there is a problem behind this idea: neither mussel farmers or researchers are completely sure what *Mytilus* species are living in Scottish waters.

Mytilus species were previously classified by the shape and colour of their shell, although this is no longer considered a very accurate method of identification due to considerable morphological overlap between some species. *M. edulis* exists with two other closely related species, *M. galloprovincialis* and *M. trossulus*, in the "*M. edulis* species complex", so referred to because of similarities in their shell characteristics: for instance, *M. edulis* has a blue, brown or purple shell with a pearly-white interior; *M. galloprovincialis* looks very similar to *M. edulis* but can be slightly larger with a darker outer shell, while *M. trossulus* has a finely lined and more fragile shell. The taxonomy of mussels within the *Mytilus edulis* complex has been investigated in greater detail through examining mussel DNA. These studies of genetics have shown that shell morphology is not a reliable method for species identification: environmental conditions can affect shape, size and colour to the point where different species can look like the same species, while individuals of the same species can look completely different. Additionally, genetic analysis has revealed that hybrid individuals are produced wherever the ranges of these species overlap.

"Hybrid zones" have been identified and studied around the world. For example, on the Irish coastline where the Irish Sea meets the North Atlantic Ocean, there is a hybrid zone between *M. edulis* and *M. galloprovincialis; M. galloprovincialis* and *M. trossulus* hybridise along the Pacific coast of North America; and *M. edulis* has been found to hybridise with both *M. galloprovincialis* and *M. trossulus* on the West coast of Scotland.

In spite of the strong growth of the Scottish shellfish industry in the last ten years there has more recently been a slight decline in production, which is in part due to the increased presence of a mussel with a particularly fragile shell and lower meat yield – characteristics that tend to be associated with *M. trossulus* (and its hybrids). Farmers wishing to cultivate a single species could run into problems if they do not monitor the genetics of their populations: they need to know what species live in Scottish waters in order to effectively manage their stocks and mitigate risks associated with an expansion in hybrid range.

M. trossulus was first reported in wild populations of Scottish mussels in a study by Joana Dias in 2007, which was featured in the June and December issues of The Grower that year. Subsequent studies have used the same methods of genetic analysis for species identification, but a new and potentially more reliable genetic method has now become

available for a deeper investigation of the *Mytilus* genome and the species present in Scottish waters. A new project, which is a joint venture of the Institute of Aquaculture (University of Stirling) and Marine Scotland Science (Aberdeen), plans to utilise this new technique in order to get an updated picture of *Mytilus* species throughout Scotland, covering as much of the coastline as possible. To date, samples have been collected from 23 sites (9 farms and 14 shorelines) around the East and West coasts of the mainland, Orkney, Shetland and the Outer Hebrides, plus future collections are being planned at sites in the Highlands and the Inner Hebrides. Additionally, historical samples, from the year after *M. trossulus* was first reported in wild populations to the present day, will be analysed to give a more complete picture of species distribution and hybrid zones around Scotland.

APPENDIX 8C – ARTICLE 3

HYBRIDISATION AMONGST SCOTTISH MUSSEL POPULATIONS: IS MYTILUS TROSSULUS MORE WIDESPREAD THAN WE PREVIOUSLY THOUGHT?

Author: Joanna Wilson

The following article was published in the Association of Scottish Shellfish Growers winter newsletter in December 2015, as a follow up to the article in APPENDIX 8B.

The final copy is accessible under the following link:

http://assg.org.uk/#/the-grower/4532754744

In my previous article for the Grower (July 2014), I outlined the benefits of studying mussel genetics when identifying species and hybrids within the "*Mytilus edulis* species complex" (ie, *Mytilus edulis*, *Mytilus galloprovincialis* and *Mytilus trossulus*). Genetic studies are more reliable than simply looking at morphology because shell characteristics are so heavily influenced by environmental conditions, making it easier to misidentify individuals and, potentially, to underestimate hybridisation. Despite the most recent figures showing a high point for mussel production in Scotland, previous years have seen slight declines which are thought to have been exacerbated by mussels with particularly fragile shells and poor meat yields – traits characteristic of *M. trossulus* (and its hybrids). Indeed, *M. trossulus* is now recognised as a commercially damaging species under The Aquaculture and Fisheries (Scotland) Act 2013, but this legislation does not apply to the management and control of *M. trossulus* hybrids because such laws are more difficult to regulate. Hybrids may or may not threaten production depending on the levels of hybridisation present; thus, an awareness of the extent of hybridisation is important in assessing any potential threats from hybrid individuals, but without the appropriate tools, it may be difficult to detect all hybrids in a given population. Existing studies of Scottish *Mytilus* populations have used a single genetic marker to identify species and their hybrids, but if multiple genetic markers are used instead it could give a more in-depth picture of genetic composition – a benefit when trying to assess the capacity of *M. trossulus* to spread.

In my PhD project, a collaborative effort between The University of Stirling, Marine Scotland Science and MASTS, I have

designed 12 new genetic markers for identifying *Mytilus* species and their hybrids, which have subsequently been used to genotype mussels collected throughout 2012, 2013 and 2014 from a mix of farm and shoreline locations in Scotland. Within the data, there are three main points for discussion that will be covered here.

Firstly, there are notable differences when comparing the genetic composition on the mainland to that of the surrounding islands (see FIGURE 8C.1).

FIGURE 8C.1 - Comparison of the genetic composition of the islands and mainland

The mainland is dominated by *M. edulis* (*e*) while the islands are dominated by *M edulis/M. galloprovincialis* hybrids (*e/g*). A greater proportion of *M. trossulus* genes [either pure (*t*) or hybrid with *M. edulis* (*e/t*) or *M. galloprovincialis* (*g/t*), or hybrids between all three species (*egt*)] were detected on the mainland than on the islands. The presence of *egt* hybrids could be explained by, for instance, *M. edulis* and *M. galloprovincialis* individuals that hybridised over many generations so that, gradually, genes of one species became incorporated into the genome of the other. When these offspring expressing genes of two species went on to hybridise with a third species – in this case, *M. trossulus* – *egt* hybrids would be produced. It makes sense here that, proportionately, more *egt* hybrids were identified on the islands than the mainland because the islands had a higher proportion of *M. edulis/M. galloprovincialis* hybrids.

Secondly, comparing the genetic composition of mussels from ropes with those from the shoreline revealed a much higher incidence of hybridisation amongst rope grown individuals (FIGURE 8C.2). It is possible that movement of ropes between locations promotes more genetic mixing than movement of spat by natural ocean currents, artificially increasing the capacity for hybridisation. Additionally, if hybrids were more fragile-shelled than their pure counterparts, ropes would offer greater shelter from predators than the exposed shoreline.

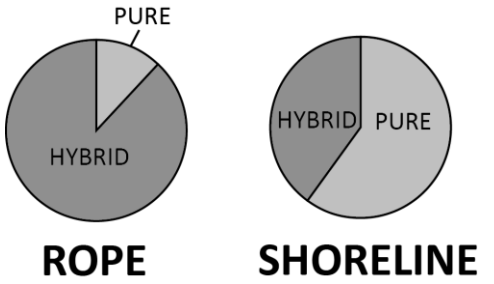

FIGURE 8C.2 - Comparison of the level of pure and hybrid individuals from rope and shoreline samples

Lastly, genotyping with the new markers revealed that *M. trossulus* genes were distributed more widely around the Scottish coast than existing studies had suggested. Consistent with previous research, the new markers identified pure *M. trossulus* in the Strathclyde region of the mainland and nowhere else. Contrary to previous research, however, the new markers not only identified *M. trossulus* hybrids in the Strathclyde region, but in all regions of the mainland (except the Grampian region, from which no samples were obtained), Orkney, Shetland, and the Inner and Outer Hebrides. The proportion of pure *M. trossulus* was very low: it equated to 0.8% of the total sampled in the Strathclyde region; 0.3% of the total sampled from the mainland; and 0.2% of the total sampled from all sites. The proportion of *M. trossulus* hybrids was notably greater: on the mainland, 26.6% of individuals were hybrids, compared to 22.6% on the islands which, overall, equated to 25.4% of the total number sampled. The proportion of hybrids varied considerably between regions, with Strathclyde having the highest (42.4%) and Dumfries and Galloway having the lowest (4%). The proportions of hybrids in all regions can be seen in FIGURE 8C.3.

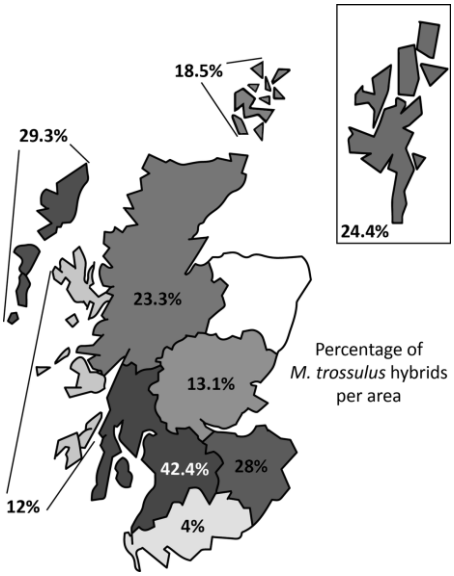

FIGURE 8C.3 - Percentage of *M. trossulus* hybrids in different regions of the mainland and islands

The new data presented here offers an updated picture of the genetic diversity of *Mytilus* species in Scotland, showing a far greater presence of *M. trossulus* hybrids on both the mainland and islands than previously documented. It cannot be ascertained without further study but, based on the results of this work and the most recent increasing trend in mussel production values, is possible that current levels of pure *M. trossulus* and its hybrids do not pose any immediate threat. Nevertheless, awareness of the distribution of *M. trossulus* and its hybrids remains important in Scottish aquaculture and, to maintain production and profitability at a high level, care should still be taken to minimise the spread of such individuals.

APPENDIX 8D – ARTICLE 4

MYTILUS TROSSULUS: IS THIS SPECIES STILL THREATENING SCOTTISH MUSSEL PRODUCTION?

Author: Joanna Wilson

The following article was published in Fish Farming Expert Magazine in February 2016.

Introduction

The blue mussel (*Mytilus edulis*) is an important source of food worldwide and dominates production within the shellfish sector of Scottish aquaculture. For the year 2014, production of *M. edulis* reached a high point of 7683 tonnes. This was a 14% increase from 2013, and the second year that mussel production was seen to increase following a production drop in 2011 and 2012. This drop in production is believed to have arisen from reduced spatfall rates, elevated toxin levels, and business closures affected by hybrids between *M. edulis* and a related *Mytilus* species, *Mytilus trossulus*. Understanding what caused this decline and subsequently introducing management practices to tackle these causes is essential in ensuring sustainable growth of the industry in future. However, without the proper tools, it may not be possible to fully assess any potential problems or deal with them accordingly, resulting in further issues for the shellfish industry. In my PhD project, a collaborative effort between The University of Stirling, Marine Scotland Science and MASTS, I have designed 12 new genetic markers for identifying *Mytilus* species and their hybrids, which have subsequently been used to genotype mussels collected throughout 2012, 2013 and 2014 from a mix of rope and shoreline locations in Scotland.

The MESC and hybridisation: how do we identify different species?

M. edulis and *M. trossulus*, alongside a third related species *Mytilus galloprovincialis*, are grouped together in the so-called "*Mytilus edulis* species complex" (*MESC*) and are capable of hybridisation wherever their ranges overlap. *M. edulis* and *M. galloprovincialis* both have robust shells and a high meat yield, whereas *M. trossulus* is considered far less marketable because of its thinner shell and reduced meat yield, so its presence on farms is not desirable. However, it is very difficult to identify species of the *MESC* simply by sight because they share a range of morphological similarities that are heavily influenced by environmental conditions. A more reliable method of distinguishing species from each other is by looking at their genetics. Indeed, the presence of the *MESC* in Scotland has been verified through studies with the genetic marker Me15/16, which looks at a single region (locus) in the genome to produce a band of a given size depending on which species – or hybrids – are present. The most recent study of *Mytilus* species distribution in Scotland (by Joana Dias in 2011), using Me15/16 to identify species and their hybrids, found *M. edulis* on the east coast and the southwest; *M. edulis* and *M. galloprovincialis* plus their hybrids on the west coast and in the Highlands; and *M. trossulus* plus hybrids with both *M. edulis* and *M. galloprovincialis* on the west coast, including in Loch Etive which has, over the years, been particularly affected by the presence of *M. trossulus.* As a result of these problems, *M. trossulus* is now recognised as a commercially damaging species under Scottish law, but this legislation does not apply to the management and control of *M. trossulus* hybrids because such laws are more difficult to regulate, given that hybrids may or may not threaten production depending on the levels of hybridisation present.

What different types of hybrids are there?

If hybridisation occurs between two pure species – ie, *M. edulis* and *M. trossulus* – all of the resulting first generation (F1) offspring will have the same genotype (50% from *M. edulis* and 50% from *M. trossulus*). In this situation, use of a single marker like Me15/16 is ideal because there is no backcrossing and no introgression (ie, incorporation of genetic material from one species into the genome of another through repeated backcrossing). However, F1 hybrids are known to be fertile and if they reproduce, with their parents or with their genetically identical siblings, things become more complex in subsequent generations of hybrids (F2, F3, etc), and use of a single marker may not be enough to tell us the full extent of hybridisation because it cannot separate F1 hybrids from F2, F3 or beyond (FIGURE 8D.1). Use of multiple genetic markers could, in comparison, give a more in-depth picture of genetic composition, which in this case would be a definite benefit when trying to assess the capacity of *M. trossulus* to spread and any threats that it may pose to production.

How are new markers developed?

To begin the development of new genetic markers, we need a way to look at the *Mytilus* genome in more depth. Next Generation Sequencing (NGS) allows us to look at multiple loci in a genome. Restriction Site Associated DNA sequencing (RADseq) uses NGS to create a reduced representation of a genome. DNA is digested with restriction enzymes and then each fragment is sequenced to identify Single Nucleotide Polymorphisms (SNPs). A SNP is a single base change in DNA that functions as a genetic "tag" to allow ID of different species. Finding SNPs in *Mytilus* species identifies multiple new markers for species ID and allows SNP assays to be developed for genotyping individuals. Much in the same way that Me15/16 produces a diagnostic band of a given size to identify a species, a SNP assay identifies species by detecting differently coloured fluorescent signals, enabling fast and reliable detection of *Mytilus* species and their hybrids.

What genotypes did the new markers find?

I have designed a total of 12 SNP assays that have been used to genotype mussels collected from 22 sites around Scotland between 2012 and 2014. This includes five regions of the

mainland (Central, Dumfries and Galloway, Highland, Lothian, and Strathclyde) and three island regions (The Hebrides, Orkney and Shetland), and comprises a mixture of rope and shoreline locations. Within the data, there are three main points for discussion that shall be covered here.

Firstly, there are notable differences when comparing the genetic composition on the mainland to that of the surrounding islands (see FIGURE 8C.1). The mainland is dominated by *M. edulis* while the islands are dominated by *M edulis/M. galloprovincialis* hybrids. A greater proportion of *M. trossulus* genes [either pure or hybrid with *M. edulis* or *M. galloprovincialis*, or hybrids between all three species (*egt*)] were detected on the mainland than on the islands. Assuming that the new markers are detecting genuine polymorphisms, and these results have not arisen from errors with efficiency, the presence of *egt* hybrids could be explained by, for instance, *M. edulis* and *M. galloprovincialis* individuals that hybridised over many generations so that, gradually, introgression occurred. If introgressed offspring expressing genes of two species went on to hybridise with a third species – in this case, *M. trossulus – egt* hybrids would be produced. It makes sense here that, proportionately, more *egt* hybrids were identified on the islands than the mainland because the islands had a higher proportion of *M. edulis/M. galloprovincialis* hybrids.

Secondly, comparing the genetic composition of mussels from ropes with those from the shoreline revealed a much higher incidence of hybridisation amongst rope grown individuals (77.9% on ropes; 38.8% on shoreline). It is possible that movement of ropes between locations promotes more genetic mixing than movement of spat by natural ocean currents, artificially increasing the capacity for hybridisation. Additionally, if hybrids were more fragile-shelled than their pure counterparts, ropes would offer greater shelter from predators than the exposed shoreline.

Lastly, genotyping with the new markers revealed that *M. trossulus* genes were distributed more widely around the Scottish coast than existing studies had suggested. Consistent with previous research, the new markers identified pure *M. trossulus* in the Strathclyde region of the mainland and nowhere else. Contrary to previous research, however, the new markers not only identified *M. trossulus* hybrids in the Strathclyde region, but in the other four regions of the mainland that were sampled (Central, Dumfries and Galloway, Lothian and Highland), plus the Hebrides, Orkney and Shetland. The proportion of pure *M. trossulus* was very low: it equated to 0.8% of the total sampled in the Strathclyde region; 0.3% of the total sampled from the mainland; and 0.2% of the total sampled from all sites. The proportion of *M. trossulus* hybrids was notably greater: overall, 25.4% of the total number sampled was made up of *M. trossulus* hybrids. The proportion of hybrids varied considerably between regions. On the mainland, Strathclyde had the highest (41.6%), followed by Lothian (28%), Highland (23.3%), Central (13.1%), and Dumfries and Galloway (4%). Of the island regions, Shetland had the highest proportion of *M. trossulus* hybrids (24.4%) followed by Orkney (18.5%) and the Hebrides (12%) (see FIGURE 8C.3).

Did the markers identify different generations of hybrids?

F1, F2 and F3 and beyond hybrids were all detected. Hybrids that were classed as F3 and beyond were the most abundant and F1 hybrids were the least abundant. Despite their seemingly low abundance, the identification of F1 hybrids does still suggest that hybridisation between pure individuals is an ongoing phenomenon that has occurred within recent generations. The identification of F2 and later generations of hybrids suggests that F1 hybrids are reproducing, which subsequently has implications for the capacity of *M. trossulus* alleles to spread across generations.

What conclusions can we draw about the possible threat of* M. trossulus *and its hybrids to the Scottish shellfish industry?

The new data presented here offers an updated picture of the genetic diversity of *Mytilus* species in Scotland, showing a far greater presence of *M. trossulus* hybrids on both the mainland and islands than previously documented. Additionally, the presence of F1, F2 and later generation hybrids suggests that *M. trossulus* does have the capacity to spread throughout generations, and that hybridisation is an ongoing phenomenon in Scottish populations. It cannot be ascertained without further study but, based on the results of this work and the most recent increasing trend in mussel production values, it is possible that current levels of pure *M. trossulus* and its hybrids do not pose any immediate threat, and that existing legislation for the management of *M. trossulus* is sufficient. Nevertheless, awareness of the distribution of *M. trossulus* and its hybrids remains important in Scottish aquaculture and, to maintain production and profitability at a high level, care should still be taken to minimise the spread of such individuals.

I would like to thank all individuals who have provided me with and helped me collect samples throughout the course of my project. This research was funded by Marine Scotland Science and MASTS, and has been supervised by Dr John Taggart, Dr Michaël Bekaert (Institute of Aquaculture, University of Stirling) and Dr Iveta Matejusová (Marine Scotland Science).

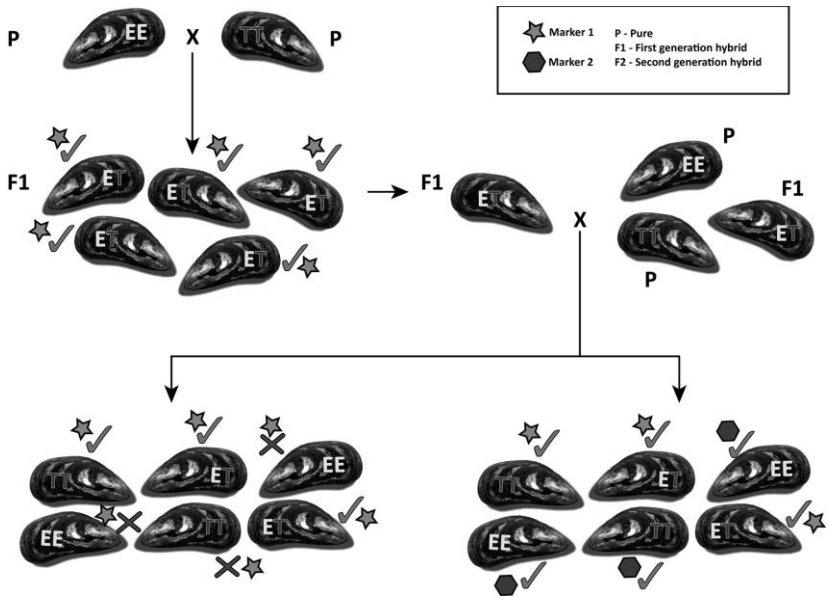

FIGURE 8D.1 – Diagram illustrating the benefits of using multiple markers when hybrid offspring have more than one genotype. Pure *M. edulis* and second generation hybrids with two *M. edulis* alleles are represented by EE; pure *M. trossulus* and second generation hybrids with two *M. trossulus* alleles are represented by TT; first generation hybrids with an allele of each species are represented by ET