PLOS ONE

RESEARCH ARTICLE

New diagnostic SNP molecular markers for the *Mytilus* species complex

Joanna Wilson[1,2], Iveta Matejusova[2], Rebecca E. McIntosh[2], Stefano Carboni[1], Michaël Bekaert[1]*

1 Institute of Aquaculture, Faculty of Natural Sciences, University of Stirling, Stirling, Scotland, United Kingdom, **2** Marine Scotland Science, Aberdeen, Scotland, United Kingdom

* michael.bekaert@stir.ac.uk

OPEN ACCESS

Citation: Wilson J, Matejusova I, McIntosh RE, Carboni S, Bekaert M (2018) New diagnostic SNP molecular markers for the *Mytilus* species complex. PLoS ONE 13(7): e0200654. https://doi.org/10.1371/journal.pone.0200654

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: March 19, 2018

Accepted: June 29, 2018

Published: July 12, 2018

Data Availability Statement: Reads were deposited at the European Bioinformatics Institute (EBI) Sequence Read Archive (SRA) study PRJEB7210.

Funding: The authors acknowledge the support of the MASTS pooling initiative (The Marine Alliance for Science and Technology for Scotland) in the completion of this study. MASTS is funded by the Scottish Funding Council (grant reference HR09011) to JW and MB and contributing institutions. JW was also supported by Marine

Abstract

The development of diagnostic markers has been a long-standing interest of population geneticists as it allows clarification of taxonomic uncertainties. Historically, there has been much debate on the taxonomic status of species belonging to the *Mytilus* species complex (*M. edulis*, *M. galloprovincialis* and *M. trossulus*), and whether they are discrete species. We analysed reference pure specimens of *M. edulis*, *M. galloprovincialis* and *M. trossulus*, using Restriction site associated DNA (RAD) sequencing and identified over 6,000 SNP markers separating the three species unambiguously. We developed a panel of diagnostic SNP markers for the genotyping of *Mytilus* species complex as well as the identification of hybrids and interspecies introgression events in *Mytilus* species. We validated a panel of twelve diagnostic SNP markers which can be used for species genotyping. Being able to accurately identify species and hybrids within the *Mytilus* species complex is important for the selective mussel stock management, the exclusion of invasive species, basic physiology and bio-diversity studies.

Introduction

Blue mussel (*Mytilus edulis*, Linneaus, 1758) has been integral part of humans' diet for millennia, their shells have been found in middens dated back to the late Mesolithic periods, 5,000 B. C. [1]. Today, Europe is a major contributor to mussel's production, supplying over a third of the total commercial outputs. Aquaculture is by far the main source of this commodity and it is responsible for over 90 percent of total landings. *M. edulis* and *Mytilus galloprovincialis* (Lamarck, 1819) are the two main species cultivated in Europe with an output of 550,000 tonnes and € 900 million per year [2]. These two species, together with the Baltic mussel (*Mytilus trossulus*, Gould, 1850), belong to the "*Mytilus* species complex". Hybridisation between these species has been observed across the world; *e.g.*, in the Pacific Ocean [3–6], in the Irish Sea [7–9] and Scotland where all combinations of species hybrids have been identified [10–13]. *M. trossulus* has been often associated with lower meat yield, thinner shell and reduced shelf life compared with *M. edulis* [10,14] and it is therefore considered undesirable within the European aquaculture context. Being able to accurately identify species and hybrids within the

Scotland Science. Publication fees were coved by the University of Stirling to MB.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: SNP, Single-Nucleotide Polymorphism; RAD, Restriction site Associated DNA; PCA, principal component analysis; KASP, Kompetitive Allele Specific PCR.

Mytilus species complex is therefore important for the management of a potentially economically damaging species.

Hybridisation and introgression between species are common evolutionary phenomena [15–17]. 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 [18]. Accurate identification of species (including cryptic or complex species) is important from commercial, conservation and research viewpoints and can be significantly impaired when hybridisation or introgression occur.

There are some distinguishing morphological features that could be employed for marine mussel species (*Mytilus* spp.) identification, such as shell colour, shape, texture and size [19–21]. However, a range of biotic and abiotic factors, including hydrodynamic conditions [22], water temperature and salinity [23], can affect these features making individuals often morphologically similar, especially in sympatric populations [24–26].

The development of diagnostic markers has been a long-standing interest of population geneticists as it allows clarification of taxonomic uncertainties. The first tools used to distinguish between mussel species were allozymes. Numerous allozyme markers have been developed and used for *Mytilus* species identification [10,14,23,27]. Nonetheless, low as well as high variation between individual of the *Mytilus* complex render the technique only marginally more useful than the identification by morphological features [28]. Over the past three decades, a range of species diagnostic markers have been developed for single locus genotyping of *Mytilus* species of which the most routinely used is the nuclear DNA marker Me15/16 [29]. Genotyping with the Me15/16 is favoured due to its simple methodology of PCR amplification and identification of a size-specific gene fragments unique for each of the *Mytilus* species [30]. Single locus genotyping delivers more accurate identification of *Mytilus* species. than studying morphology or allozymes, but has limited potential for analysing patterns of hybridisation or genome introgression [17]. Multilocus genotyping, using genome wide panels of single nucleotide polymorphism (SNP) markers by comparison, allows for a far better understanding of introgression [31–35]. There are only few studies on multilocus genotyping in Mytilus species complex [12,36], which are however limited to small number of markers used to resolve population structures based on allele frequencies.

In this study, we have employed an easy and rapid *de novo* SNP discovery method to develop genome-wide species-specific markers to genotype the *Mytilus* species complex, which also identify hybrids and introgressed individuals in field populations. More accurate characterisation of *Mytilus* specie populations' structure will aid improved conservation and aquaculture management strategies.

Materials and methods

Ethics statement

Animal (mussels) handling and collection was done under Marine Science Scotland (Scottish Government) authority and following both Marine Science Scotland and University of Stirling Ethical recommendations and guidance.

Sample collection

Adult mussels (at least 40 mm in length) were collected from regions where pure *Mytilus* species were reported to occur, based on historical, genetic analysis or morphological evidence [3,10,11,37,38]. Specimens of *M. edulis* were collected from two shoreline locations in south-west Scotland [Loch Ryan (LR) and Rascarrel Bay (RB)] and one shoreline location in east Scotland [Montrose (MON)], the three sites have shellfish farming activities in the vicinities;

Table 1. Details of sampling sites and the number *Mytilus* specimens used for marker discovery and marker validation. * One of the *M. trossulus* from Penn Cove was re-assigned as *M. edulis* during the marker development/validation stage (see Results). † Additional samples used only for marker validation.

Site location	Site coordinates	Species reported	Me15/16 screening	SNP discovery	KASP validation
LR—Loch Ryan	54˚56'06.8"N 5˚03'38.7"W	*M. edulis*	50	10	50
RB—Rascarrel Bay	54˚48'53.1"N 3˚51'22.7"W	*M. edulis*	50	10	50
BP—Bay of Piran	-	*M. galloprovincialis*	50	15	50
PC—Penn Cove	-	*M. trossulus*	8	4+1*	8
MON—Montrose	56˚42'16.3"N 2˚28'13.7"W	*M. edulis*	50	-	50
LET—Loch Etive†	56˚27'05.5"N 5˚19'13.3"W	*M. trossulus* (juvenal)	20	-	20
BDL—Bras d'Or Lake†	45˚59'55.4"N 60˚43'31.0"W	*M. trossulus*	50	-	50

https://doi.org/10.1371/journal.pone.0200654.t001

M. galloprovincialis were sourced from Slovenia [Bay of Piran (BP)]; and *M. trossulus* were acquired from Penn Cove (PC), USA (Table 1). Additional adults *M. trossulus* from Bras d'Or Lake (BDL), Canada, and juvenile mussels (approximately 15-months old) from Loch Etive (LET), Scotland were also obtained and used for markers validation purposes (Table 1). Tissue samples (gill/mantle from adults; all body tissues from juveniles) were taken and stored in 99% ethanol at -20˚C.

Me15/16 PCR genotyping

DNA was extracted from tissue and treated with RNase. Each sample was quantified by spectrophotometry (Nanodrop), quality assessed by agarose gel electrophoresis, and stored in 5 mmol/L Tris, pH 8.5. Preliminary PCR were carried out at a single locus with the Me15/16 primer set: Me15: `CCAGTATACAAACCTGTGAAGA`; Me16: `GTTGTCTTAATAGGTTTGTAAGA` [29]. Each 6 μL PCR reaction comprised 3 μL 2× MyTaq mix (Bioline); 0.4 μL of 10 μM forward and reverse primer; 0.5 μL template DNA (5–50 ng/μL); and 1.7 μL ultrapure water. PCR conditions were 95˚C for 1 min, [95˚C for 15 s, 56˚C for 15 s, 72˚C for 30 s] × 35 cycles. PCR products (1 μL) were run at 60 V for 40 mins on a 2% agarose gel (0.5× TAE, stained with 100 ng/μL EtBr). Using a UV transilluminator, *Mytilus* species and hybrids were identified based on size differentiation of PCR products: 180 bp (*M. edulis);* 168 bp (*M. trossulus);* 126 bp (*M. galloprovincialis)* or a combination in the case of hybrid individuals [29].

RAD library preparation and sequencing

A total of 40 pure specimens (21 *M. edulis*, 15 *M. galloprovincialis* and 4 *M. trossulus*) were chosen for species reference library construction (S1 Table). The RAD library was prepared as originally described in Baird *et al.* [39] and comprehensively detailed in Etter et al. [40], with minor modifications [41]. Briefly, each sample (0.25 μg DNA) was digested at 37˚C for 40 min with the high-fidelity restriction enzyme *Pst*I that recognises the CTGCA|G motif (New England Biolabs; NEB) using 6 U *Pst*I per μg genomic DNA in 1× Reaction Buffer 4 (NEB) at a final concentration of about 1 μg DNA per 50 μL reaction volume. The samples (12 μL final volume) were then heat-inactivated at 65˚C for 20 min. Individual specific P1 adapters, each with a unique 5 or 7 bp barcode (S1 Table), were ligated to the *Pst*I digested DNA at 22˚C for 15 min by adding 0.6 μL (DNA samples) 100 nmol/L P1 adapter, 0.15 μL 100 mmol/L rATP (Promega), 0.25 μL 10× Reaction Buffer 2 (NEB), 0.125 μL T4 ligase (NEB, 2,000 U/μL) and reaction volumes made up to 15 μL with nuclease-free water for each sample. After heat-inactivation at 65˚C for 20 min, the ligation reactions were slowly cooled to room temperature (over 1 h), then combined in appropriate multiplex pools. Shearing (Covaris S2 sonication) and initial size selection (100 to 800 bp) by agarose gel electrophoresis [41] was followed by gel

purification, end repair, dA overhang addition, P2 paired-end adapter ligation and library amplification. 120 μL of each amplified library was size-selected (about 250 to 500 bp) by gel electrophoresis. Final libraries were sent to BMR Genomics (Padua, Italy), for quality control and high-throughput sequencing. Libraries were accurately quantified by fluorimetry and calibrated by sequencing on an Illumina MiSeq at the Institute of Aquaculture using 100 base paired-end reads (v3 chemistry). The Libraries were sequenced in four lanes of an Illumina HiSeq 2000, using 100 base paired-end reads (v3 chemistry). Reads were deposited at the European Bioinformatics Institute (EBI) Sequence Read Archive (SRA) study PRJEB7210.

Genotyping RAD alleles

Reads of low quality (*i.e.*, with an average quality score less than 20), that lacked the restriction site or had ambiguous barcodes were discarded. Retained reads were sorted into loci and genotypes using Stacks v1.13 [42]. Stacks assigns loci based on nucleotide positions in RAD tags using a likelihood-based algorithm [43] to separate actual SNPs from SNPs likely to have arisen from sequencing error. Using the default parameters for *de novo* assembly pipeline, 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. Informative RAD markers were kept only when presenting a maximum of three SNPs and one to three alleles present in all three species and at least 50% of the samples. Diagnostic species-specific markers were Informative RAD markers with a fixed allele within species but presenting different allele between at least two of the three species.

Phylogenetic analysis

Sequencing data from filtered Informative RAD markers 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), using the RAxML v8.0.0 [44]. Maximum-likelihood phylogenetic trees were inferred using the GTR+CAT nucleotide substitution model [45] and bootstrap support values estimated from 10,000 replicate searches of randomly generated trees.

Data analysis

Data analysis was carried out using R v3.3.2 [46] and an associated R/*adegenet* package v1.4–1 [47] for Principal Component Analysis (PCA) and Discriminant Analysis of Principal Components (DAPC). PCA creates simplified models of the total variation within the dataset and DAPC identifies clusters of genetically related individuals [48].

SNP-assay design

Each tested locus comprised 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 bp from the end of a given sequence. This allowed enough sequence for compatible primers to be designed. SNP assays were designed and manufactured for use with KASP genotyping technology by LGC Genomics Ltd. (S2 Table). Each sample was genotyped in 5 μL reactions each containing approximately 40 ng template DNA. Optimisation assay conditions were 94˚C for 15 min; [94˚C for 20 s, 61–55˚C for 120 s (0.6˚C drop per cycle)] × 10; and [94˚C for 20 s, 55˚C for 120 s] × 40. Each 5 μL reaction comprised 2.5 μL 2× 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. An addition of 0.25 100% DMSO was added for markers G4, T4 and T5. Individual genotype assignment was performed through reading the fluorescence emission of the FAM and HEX fluorophores for each sample, in comparison to no-template control reactions, using a Techne Quantica Real Time PCR Thermal Cycler and Quansoft end-point genotyping software (Bibby Scientific).

Population structure

Structure v2.3 [49] was used to identify distinct genetic populations from multilocus (SNP) data, assigning individuals to populations, and identifying admixed individuals. The "Admixture Model" was used assuming that each genotyped individual could have mixed ancestry, inheriting some fraction of its genome from ancestors in a different population. This would be an assumption made for reference populations of pure species, while the other could have migrants that have interbred with native individuals.

Results

Me15/16 pre-evaluation

Putative pure *Mytilus* species individuals collected in locations where only pure species were previously reported [*M*. *edulis* (Loch Ryan and Rascarrel Bay); *M*. *galloprovincialis* (Bay of Piran); and *M*. *trossulus* (Penn Cove)] were screened using the Me15/16 locus (Table 2 and S3 Table) to confirm their genotype. A total of 40 individuals, genotyped as pure (homozygous) with Me15/16 were chosen for RAD library construction: these included of 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* produced a small sample size because of the very limited DNA material available for this species.

RAD library sequencing

High throughput sequencing of these 40 individuals produced 574,728,488 raw reads in total (four HiSeq lanes and one MiSeq lane). MiSeq technology was used to adjust the libraries. After the removal of low-quality and incomplete reads, 71.9% of the total raw reads were retained (413,377,018 reads). As only *M*. *galloprovincialis* has a published draft genome (NCBI assembly GCA_001676915.1) of over 1 million contigs and that only 35% of the reads from *M*. *galloprovincialis* samples were aligned to it, a *de novo* approached was used to assemble the RAD tags. A total of 3,253,798 RAD tags were detected (S1 Table).

Table 2. Summary results the preliminary Me15/16 genotyping. Genotypes are as follows: *M*. *edulis* [Me]; *M*. *galloprovincialis* [Mg]; *M*. *trossulus* [Mt]. Reported numbers are number of individual presenting a given genotype. Hybrid are shown as composites. Site names are abbreviated as detailed in Table 1.

Site	Me	Mg	Mt	Me/Mg	Me/Mt	Mg/Mt
LR	49	0	0	1	0	0
RB	50	0	0	0	0	0
MON	48	0	0	2	0	0
BP	0	50	0	0	0	0
PC	1	0	7	0	0	0
BDL	0	0	40	0	10	0
LET	0	0	20	0	0	0

https://doi.org/10.1371/journal.pone.0200654.t002

Sequence analysis

The number of RAD tag detected per individual was relatively consistent, ranging from 59,000–313,000 RAD tag (Table 1). There were two exceptions among *M. edulis* individuals with significantly lower number of tags obtained (RB_01 and PC_01, which had 18,220 and 5,459 RAD tags respectively) which was most likely caused by low quality DNA resulting in effecting the library preparation efficiency. Between 14% and 15% of the RAD tags were polymorphic. To identify robust genetic markers and to minimise the proportion of erroneous data, all informative markers were filtered to show only those with one to three alleles and a maximum of three SNPs, and which were detected in all three species and at least 50% of the samples. A total of 14,212 SNPs spread across 6,220 informative RAD markers were identified (some loci had more than one SNP), and used in subsequent analyses. A reduced set of markers, 378 SNPs spread across 365 diagnostics species-specific markers, were filtered out as the informative RAD markers when exhibiting fixed allele within species, but presenting different allele between at least two of the three species (S4 Table).

Phylogenetic reconstruction

The phylogenetic tree constructed from the composite genotypes of 6,220 informative RAD markers shared alleles (14,212 SNPs) showed three distinct clusters, accurately delineating the three species (3 sites for *M. edulis*, one site for *M. galloprovincialis* and *M. trossulus*) that were used for library construction (Fig 1A). *M. edulis* and *M. galloprovincialis* were the closest (a genetic distance of 33 nucleotide substitutions between the base *M. edulis* and *M. galloprovincialis* branches) and *M. trossulus* was more distant (a genetic distance of 79 nucleotide substitutions). One *M. edulis* from Penn Cove was grouped with the *M. edulis* from Loch Ryan and Rascarrel Bay, confirming its identity as *M. edulis* as suggested by the Me15/16 genotyping.

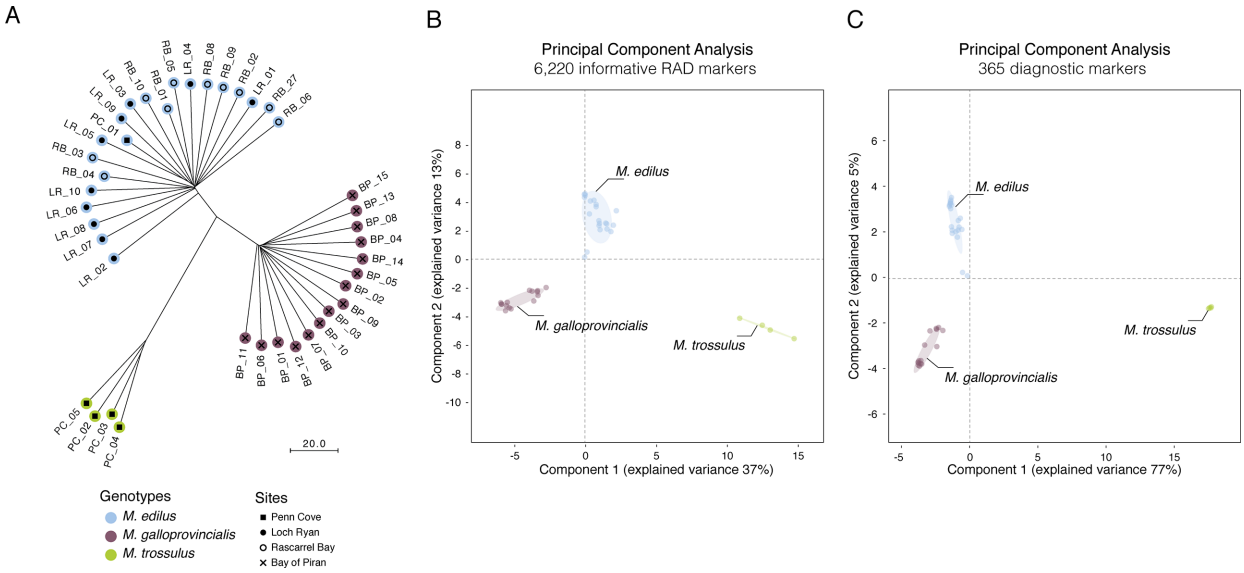

Fig 1. Capture the discriminant ability of RAD markers. (A) Phylogenetic reconstruction based on the SNP of the 6,220 informative RAD markers (RAxML). Genotype were established using preliminary Me15/16 PCR assay. The scale shows the number of nucleotide substitutions per site (B) Principal Component Analysis of the 6,220 informative RAD markers. (C) Principal Component Analysis of the 365 diagnostic markers. RB_01 and PC_01 while grouping with *M. edulis* exhibit lower polymorphism due to the highest number of missing data compared to all other samples.

https://doi.org/10.1371/journal.pone.0200654.g001

Marker selection

In order to better capture the *Mytilus* species complex structure the ability of each marker to discriminates each "pure" species, a principal component analysis (PCA) was conducted from 6,220 informative RAD markers using *R/adegenet* (Fig 1B). Three distinct clusters were separated using the first two components (88.7% of cumulative variance) despite the small number of samples examined.

A second PCA was applied on the reduced set of 365 diagnostics species-specific markers, to ensure that they kept their discrimination power (Fig 1C). Subsequently, the Discriminant Analysis of Principal Components (DAPC) sorted species diagnostic loci by their "loading values", values based on the population coverage at each locus [47]. Loci with the highest presence had the highest loading values and, thus, were assumed to be less likely to be false positive, improving their reliability as potential diagnostic markers. Loci with the highest "loading values" were preferred; as long as they matched the other selection criteria for KASP assay development.

SNP assay optimisation

All SNP assays were designed for use with KASP genotyping technology by LGC Genomics Ltd. Assay optimisation was carried out with the 40 samples used in RAD library construction. A total of 12 SNP assays, three assays per *Mytilus* species, were successfully optimised (S2 Table): 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 (KASP and RAD) were obtainable at all 12 loci for each of the 40 samples (S5 Table) revealing identical results regardless of the genotying technique.

SNP assay validation

238 samples, including 150 samples from 3 additional populations, were genotyped with the 12 SNP assays (Table 3 and S6 Table). Additional populations were sourced in order to validate the markers with individuals from different origins, thereby reducing the risk of developing markers that would be population- or location-specific rather than species-specific. Where all diagnostic alleles were attributed to only one species, individuals were identified as "pure" species (*M. edulis* [Me], *M. galloprovincialis* [Mg] or *M. trossulus* [Mt]). Individuals heterozygote at all diagnostic loci for two species would be identified as F1 hybrids; however, none were found in any of the population sampled. All other individuals were identified as introgressed individuals. Principal Component Analysis (PCA) discriminated the three "pure" species while the introgressed individuals were intermediates with respect to their genetic background

Table 3. Summary results the 12 SNP assay. Genotypes are as follows: *M. edulis* [Me]; *M. galloprovincialis* [Mg]; *M. trossulus* [Mt]; Hybrids are shown as composites. Site names are abbreviated as detailed in Table 1.

Site	Me	Mg	Mt	MeMg	MeMt	MtMg	MeMgMt
LR	48	0	0	2	0	0	0
RB	49	0	0	0	1	0	0
MON	44	0	0	3	2	0	1
BP	0	50	0	0	0	0	0
PC	1	0	7	0	0	0	0
BDL	0	0	12	0	5	10	23
LET	0	0	5	0	6	4	5

https://doi.org/10.1371/journal.pone.0200654.t003

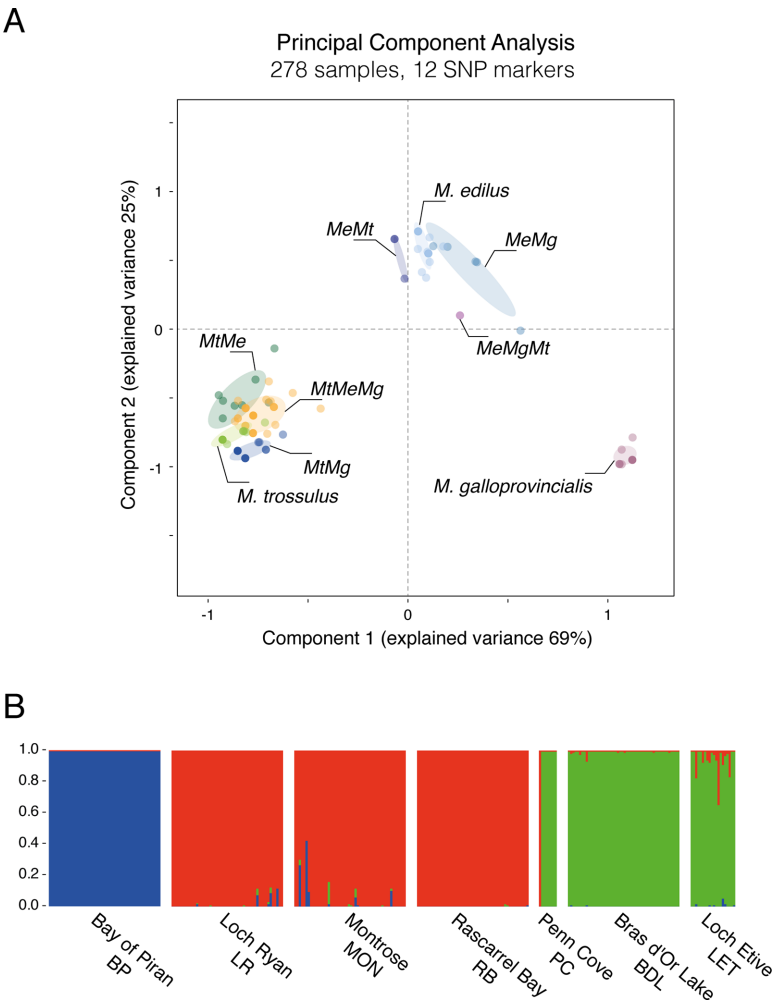

Fig 2. Multilocus genotyping across 12 SNP makers. (A) Principal Component Analysis of the 12 diagnostic markers across the 278 samples. The Genotype classes are clustered and annotated on the figure. (B) Structure plots constructed using the Admixture Ancestry Model with independent allele frequencies per population (K = 3, burnin = 10,000, reps = 100,000), showing the genetic composition of reference and validation population.

https://doi.org/10.1371/journal.pone.0200654.g002

(Fig 2A). Based on their position, the dominating background and one or more other species influencing the genetic make-up can be estimated.

Structure modelling showed three distinct clusters of genotypes; this model best corresponded to three distinct genotypes (*M. edulis*, *M. galloprovincialis* and *M. trossulus*) in the populations studied (Fig 2B). The four populations used for RADseq were suitable pure populations. These models suggest that, despite the introgression observable with the SNP genotyping, these populations are mostly pure with some mixing and thus were suitable for diagnostic marker design.

Bay of Piran (100%; 50/50), Penn Cove (100%; 8/8), Rascarrel Bay (98%; 49/50), and Loch Ryan (96%; 48/50) had the highest proportions of pure individuals, followed by Montrose (88%; 44/50). Bras d'Or Lake (24%; 12/50) and Loch Etive (25%; 5/20) showed a high proportion of introgressed individuals; Both location have been reported as having hybrids *M. edulis* × *M. trossulus* [10,37]. Loch Etive population was expected to have 100% *M. trossulus* individuals according to the Me15/16 locus genotyping (Fig 3A); however, the SNP analysis revealed

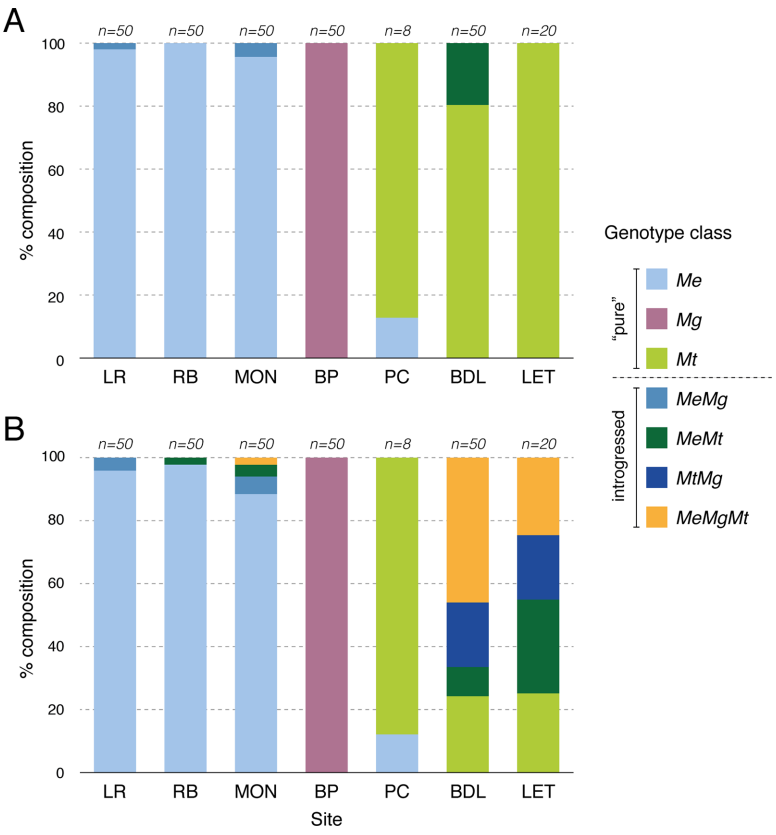

Fig 3. Genotype classes identified in seven populations. (A) Single locus Me15/16 (details in Table 2). (B) Multilocus genotyping across 12 SNP markers. No F1 hybrid was detected (details in Table 3).

https://doi.org/10.1371/journal.pone.0200654.g003

that only 25% were "pure" *M*. *trossulus*, and the other 75% of the individuals were hybrids with various degrees of introgression with *M*. *edulis* and/or *M*. *galloprovincialis* (Fig 3B).

Discussion

Historically, there has been much debate on the taxonomic status of species belonging to the *Mytilus* species complex (*M*. *edulis*, *M*. *galloprovincialis* and *M*. *trossulus*), and whether they are discrete species. Species diagnostic marker development could only take place if three discrete species were present; thus, assessing the phylogenetic relationships of the populations used for diagnostic marker development was crucial for this research. Indeed, Fig 1 shows a clear separation of individuals based on genotype and not population, multiple sites were used for *M*. *edulis*. As confirmed by the KASP assays on the same dataset (Fig 2), all diagnostic markers appeared fixed in the putative "pure" populations, despite the limitation caused by the small number of *M*. *trossulus* specimens (only 4 samples); However, both phylogenetic tree and PCA clearly distinct three groups, consistent with the three species. Furthermore, the identification of 6,220 informative RAD markers and the development of 12 diagnostics markers (chosen out of the 365-possible identified) is a unique opportunity to enhance understanding of the genotype for the application of basic physiological studies and to understand the biological differences between *M*. *trossulus*, *M*. *edulis* and *M*. *galloprovincialis*. The informative RAD markers (polymorphic marker present in at least 50% of the samples) and diagnostic markers (Informative RAD markers with fixed allele within species but presenting different allele between at least two of the three species) identified in this study can be used for many

purposed: bio-diversity evaluation, phylogenetic studies, species and population identification. This enhanced understanding of the genetic structure within and between populations, also provides opportunities to better clarify the relationship between genotype and phenotype, particularly in sympatric populations.

Within the aquaculture context, the availability of this new suit of diagnostic markers can allow the sourcing of seeds of known genotypes and the selection of seeds of desirable genotypes for broodstock for ongoing efforts into the hatchery production of seeds. This will significantly contribute to the future eradication of potentially economically damaging species. Improved stock management and potential selective breeding will result in superior resilience and increased productivity of the mussel's aquaculture industry. Indeed, with the application of multilocus genotyping on Scottish mussels, we have identified introgressed genotypes that were hitherto unrecognisable by using single locus (Me15/16) genotyping. By doing so we improved our understanding of the genetic diversity within and between populations currently present in farmed and wild populations along the Scottish coast. This quick and relatively in-expensive methodology can be extended to any species complex where the phenotype does not provide conclusive evidence for species assignment and where the genetic structure is unknown or poorly established.

Supporting information

S1 Table. Sample and RAD barcodes. Details each sample used: sample ID, library number, Me15/16 Genotype, RAD barcode (P1 adapter), RAD barcode (P1 adapter), number of raw reads (paired-ended) and number of RAD-tags.
(CSV)

S2 Table. KASP assay primer sequences. List of the allele-specific primers and common primer designed for the allele-specific PCR genotyping assay.
(CSV)

S3 Table. Details of the Me15/16 genotyping results. Genotypes results for Me15/16 preliminary assay for all 278 samples, summarised Table 2.
(CSV)

S4 Table. Details of the 378 selected SNP markers. N means no SNP reported for the species. To be selected a SNP needs to be present in at least 50% of all samples and to be reported in a species, the SNP needs to be in at least 50% of the samples of this species (see Materials and Methods).
(CSV)

S5 Table. Details of the RAD and KASP assay results. Genotypes of the 12 assays for 40 development individuals.
(CSV)

S6 Table. Details of the KASP genotyping results. Genotypes results for the 12 KASP assays for all 278 samples.
(CSV)

Acknowledgments

The authors thank 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.

Author Contributions

Conceptualization: Iveta Matejusova, Michaël Bekaert.

Data curation: Joanna Wilson, Michaël Bekaert.

Formal analysis: Joanna Wilson, Stefano Carboni, Michaël Bekaert.

Funding acquisition: Iveta Matejusova, Michaël Bekaert.

Investigation: Joanna Wilson, Michaël Bekaert.

Methodology: Rebecca E. McIntosh, Michaël Bekaert.

Project administration: Michaël Bekaert.

Resources: Joanna Wilson, Michaël Bekaert.

Supervision: Iveta Matejusova, Michaël Bekaert.

Validation: Joanna Wilson.

Writing – original draft: Michaël Bekaert.

Writing – review & editing: Joanna Wilson, Iveta Matejusova, Stefano Carboni, Michaël Bekaert.

References

1. Andersen SH. Shell Middens («Køkkenmøddinger»): The Danish Evidence. In: Andrzej A, Cipriani R, editors. Early Human Impact on Megamolluscs. 2008. pp. 135–156.
2. Monfort M-C. The European market for mussels. Rome; 2014.
3. Suchanek TH, Geller JB, Kreiser BR, Mitton JB. Zoogeographic Distributions of the Sibling Species *Mytilus galloprovincialis* and *M. trossulus* (Bivalvia: Mytilidae) and Their Hybrids in the North Pacific. Biol Bull. 1997; 193: 187–194. https://doi.org/10.2307/1542764 PMID: 28575597
4. Rawson PD, Joyner KL, Meetze K, Hilbish TJ. Evidence for intragenic recombination within a novel genetic marker that distinguishes mussels in the *Mytilus edulis* species complex. Heredity. 1996; 77: 599–607. PMID: 8972080
5. Rawson PD, Agrawal V, Hilbish TJ. Hybridization between the blue mussels *Mytilus galloprovincialis* and *M. trossulus* along the Pacific coast of North America: evidence for limited introgression. Mar Biol. 1999; 134: 201–211. https://doi.org/10.1007/s002270050538
6. Wonham MJ. Mini-review: Distribution of the Mediterranean mussel *Mytilus galloprovincialis* (Bivalvia: Mytilidae) and hybrids in the Northeast Pacific. J Shellfish Res. 2004; 23: 535–543.
7. Coghlan B, Gosling E. Genetic structure of hybrid mussel populations in the west of Ireland: two hypotheses revisited. Mar Biol. 2007; 150: 841–852. https://doi.org/10.1007/s00227-006-0408-z
8. Gosling E, Doherty S, Howley N. Genetic characterization of hybrid mussel (*Mytilus*) populations on Irish coasts. J Mar Biol Assoc UK. 2008; 88. https://doi.org/10.1017/S0025315408000957
9. Doherty SD, Brophy D, Gosling E. Synchronous reproduction may facilitate introgression in a hybrid mussel (*Mytilus*) population. J Exp Mar Bio Ecol. 2009; 378: 1–7. https://doi.org/10.1016/j.jembe.2009.04.022
10. Beaumont AR, Hawkins MP, Doig FL, Davies IM, Snow M. Three species of *Mytilus* and their hybrids identified in a Scottish Loch: natives, relicts and invaders? J Exp Mar Bio Ecol. 2008; 367: 100–110. https://doi.org/10.1016/j.jembe.2008.08.021
11. Dias PJ, Piertney SB, Snow M, Davies IM. Survey and management of mussel *Mytilus* species in Scotland. Hydrobiologia. 2011; 670: 127–140. https://doi.org/10.1007/s10750-011-0664-x
12. Zbawicka M, Drywa A, Śmietanka B, Wenne R. Identification and validation of novel SNP markers in European populations of marine *Mytilus* mussels. Mar Biol; 2012; 159: 1347–1362. https://doi.org/10.1007/s00227-012-1915-8
13. Michalek K, Ventura A, Sanders T. *Mytilus* hybridisation and impact on aquaculture: A minireview. Mar Genomics. 2016; 27: 3–7. https://doi.org/10.1016/j.margen.2016.04.008 PMID: 27157133
14. Penney RW, Hart MJ, Templeman ND. Shell Strength and Appearance in Cultured Blue Mussels *Mytilus edulis*, *M. trossulus*, and *M. edulis* × *M. trossulus* Hybrids. N Am J Aquac. 2007; 69: 281–295. https://doi.org/10.1577/A06-044.1

15. Mallet J. Hybridization as an invasion of the genome. Trends Ecol Evol. 2005; 20: 229–237. https://doi.org/10.1016/j.tree.2005.02.010 PMID: 16701374
16. Schwenk K, Brede N, Streit B. Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals. Philos Trans R Soc B Biol Sci. 2008; 363: 2805–2811. https://doi.org/10.1098/rstb.2008.0055 PMID: 18534946
17. Twyford AD, Ennos RA. Next-generation hybridization and introgression. Heredity. 2012; 108: 179–189. https://doi.org/10.1038/hdy.2011.68 PMID: 21897439
18. Rieseberg L, Wendel JF. Introgression and its consequences in plants. Hybrid Zo Evol Process. 1993; 70–10.
19. Hepper BT. Notes on *Mytilus galloprovincialis* Lamarck in Great Britain. J Mar Biol Assoc United Kingdom. 1957; 36: 33. https://doi.org/10.1017/S0025315400017045
20. Seed R. Factors Influencing Shell Shape in the Mussel *Mytilus Edulis*. J Mar Biol Assoc United Kingdom. 1968; 48: 561. https://doi.org/10.1017/S0025315400019159
21. Widdows J, Johnson D. Physiological energetics of *Mytilus edulis*: Scope for Growth. Mar Ecol Prog Ser. 1988; 46: 113–121. https://doi.org/10.3354/meps046113
22. Steffani CN, Branch GM. Growth rate, condition, and shell shape of *Mytilus galloprovincialis*: Responses to wave exposure. Mar Ecol Prog Ser. 2003; 246: 197–209. https://doi.org/10.3354/Meps246197
23. Riginos C, Cunningham CW. Local adaptation and species segregation in two mussel (*Mytilus edulis* × *Mytilus trossulus*) hybrid zones. Mol Ecol. 2004; 14: 381–400. https://doi.org/10.1111/j.1365-294X.2004.02379.x PMID: 15660932
24. Innes DJ, Bates JA. Morphological variation of *Mytilus edulis* and *Mytilus trossulus* in eastern Newfoundland. Mar Biol. Springer-Verlag; 1999; 133: 691–699. https://doi.org/10.1007/s002270050510
25. Hilbish TJ, Carson EW, Plante JR, Weaver LA, Gilg MR. Distribution of *Mytilus edulis*, *M. galloprovincialis*, and their hybrids in open-coast populations of mussels in southwestern England. Mar Biol. 2002; 140: 137–142. https://doi.org/10.1007/s002270100631
26. Koehn RK. The genetics and taxonomy of species in the genus *Mytilus*. Aquaculture. 1991; 94: 125–145. https://doi.org/10.1016/0044-8486(91)90114-M
27. McDonald JH, Seed R, Koehn RK. Allozymes and morphometric characters of three species of *Mytilus* in the Northern and Southern Hemispheres. Mar Biol. 1991; 111: 323–333. https://doi.org/10.1007/BF01319403
28. Schlötterer C. Opinion: The evolution of molecular markers—just a matter of fashion? Nat Rev Genet. 2004; 5: 63–69. https://doi.org/10.1038/nrg1249 PMID: 14666112
29. Inoue K, Waite JH, Matsuoka M, Odo S, Harayama S. Interspecific Variations in Adhesive Protein Sequences of *Mytilus edulis*, *M. galloprovincialis*, and *M. trossulus*. Biol Bull. 1995; 189: 370–375. https://doi.org/10.2307/1542155 PMID: 8555320
30. Dias PJ, Sollelis L, Cook EJ, Piertney SB, Davies IM, Snow M. 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. 2008; 367: 253–258. https://doi.org/10.1016/j.jembe.2008.10.011
31. Storey JD, Akey JM, Kruglyak L. Multiple Locus Linkage Analysis of Genomewide Expression in Yeast. PLoS Biol. 2005; 3: e267. https://doi.org/10.1371/journal.pbio.0030267 PMID: 16035920
32. Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ. Multiplex-Ready PCR: A new method for multiplexed SSR and SNP genotyping. BMC Genomics. 2008; 9: 80. https://doi.org/10.1186/1471-2164-9-80 PMID: 18282271
33. Linnen CR, Hoekstra HE. Measuring Natural Selection on Genotypes and Phenotypes in the Wild. Cold Spring Harb Symp Quant Biol. 2009; 74: 155–168. https://doi.org/10.1101/sqb.2009.74.045 PMID: 20413707
34. Davey JWL, Blaxter MLW. RADseq: Next-generation population genetics. Brief Funct Genomics. 2010; 9: 416–423. https://doi.org/10.1093/bfgp/elq031 PMID: 21266344
35. Zuo L, Wang K, Luo X. Use of diplotypes—matched haplotype pairs from homologous chromosomes—in gene-disease association studies. Shanghai Arch psychiatry. 2014; 26: 165–170. https://doi.org/10.3969/j.issn.1002-0829.2014.03.009 PMID: 25114493
36. Wenne R, Bach L, Zbawicka M, Strand J, McDonald JH. A first report on coexistence and hybridization of *Mytilus trossulus* and *M. edulis* mussels in Greenland. Polar Biol. 2016; 39: 343–355. https://doi.org/10.1007/s00300-015-1785-x
37. Tremblay MJ. Large epibenthic invertebrates in the Bras d'Or Lakes. Proc Nov Scotian Inst Sci. 2002; 21: 101–126.

38. Žižek S, Gombač M, Pogačnik M. Occurrence and effects of the bivalve-inhabiting hydroid *Eugymnanthea inquilina* in cultured mediterranean mussels (*Mytilus galloprovincialis*) in slovenia. Slov Vet Res. 2012; 49: 149–154.

39. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008; 3: e3376. https://doi.org/10.1371/journal.pone.0003376 PMID: 18852878

40. Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA. SNP Discovery and Genotyping for Evolutionary Genetics Using RAD Sequencing. In: Orgogoz V, Rockman M V, editors. Molecular Methods for Evolutionary Genetics. Institute of Molecular Biology, University of Oregon, Eugene, OR, USA.: Humana Press; 2012. pp. 157–178. https://doi.org/10.1007/978-1-61779-228-1_9

41. Houston RD, Davey JW, Bishop SC, Lowe NR, Mota-Velasco JC, Hamilton A, et al. Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics. 2012; 13: 244. https://doi.org/10.1186/1471-2164-13-244 PMID: 22702806

42. Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. Stacks: Building and Genotyping Loci *De Novo* From Short-Read Sequences. G3—Genes|Genomes|Genetics. 2011; 1: 171–182. https://doi.org/10.1534/g3.111.000240 PMID: 22384329

43. Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G. Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. Mol Ecol Resour. 2011; 11 Suppl 1: 117–122.

44. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014; 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

45. Lartillot N, Philippe HH. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 2004; 21: 1095–1109. https://doi.org/10.1093/molbev/msh112 PMID: 15014145

46. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austri: R Foundation for Statistical Computing; 2016.

47. Jombart T. Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; 24: 1403–1405. https://doi.org/10.1093/bioinformatics/btn129 PMID: 18397895

48. Jombart T, Devillard S, Balloux F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010; 11: 94. https://doi.org/10.1186/1471-2156-11-94 PMID: 20950446

49. Pritchard V, Jones K, Cowley D. Estimation of introgression in cutthroat trout populations using microsatellites. Conserv Genet. 2007; 8: 1311–1329. https://doi.org/10.1007/s10592-006-9280-0