Null alleles are ubiquitous at microsatellite loci in the Wedge Clam (*Donax trunculus*)

Ciro Rico[1,2,*], Jose Antonio Cuesta[3], Pilar Drake[3], Enrique Macpherson[4], Louis Bernatchez[5] and Amandine D. Marie[1,*]

[1] School of Marine Studies, Molecular Analytics Laboratory (MOANA), Faculty of Science Technology and Environment, The University of the South Pacific, Suva, Fiji
[2] Estación Biológica de Doñana, (EBD, CSIC), Sevilla, Spain
[3] Instituto de Ciencias Marinas de Andalucía (ICMAN, CSIC), Puerto Real (Cádiz), Spain
[4] Centre d'Estudis Avançats de Blanes, (CEAB, CSIC), Blanes, Spain
[5] Institut de Biologie Intégrative et des Systèmes (IBIS), Département de Biologie, Pavillon Charles-Eugène-Marchand, Laval University, Quebec, Canada
[*] These authors contributed equally to this work.

ABSTRACT

Recent studies have reported an unusually high frequency of nonamplifying alleles at microsatellite loci in bivalves. Null alleles have been associated with heterozygous deficits in many studies. While several studies have tested for its presence using different analytical tools, few have empirically tested for its consequences in estimating population structure and differentiation. We characterised 16 newly developed microsatellite loci and show that null alleles are ubiquitous in the wedge clam, *Donax trunculus*. We carried out several tests to demonstrate that the large heterozygous deficits observed in the newly characterised loci were most likely due to null alleles. We tested the robustness of microsatellite genotyping for population assignment by showing that well-recognised biogeographic regions of the south Atlantic and south Mediterranean coast of Spain harbour genetically different populations.

INTRODUCTION

Recent studies have reported an unusually high frequency of non-amplifying alleles, also known as null alleles, at microsatellite loci in bivalves (*Hargrove et al., 2015*; *Chiesa et al., 2016*) including the wedge clam (*Nantón et al., 2014*). Null alleles occur when mutations in the binding site of the targeted DNA sequence prevent the efficient annealing of at least one primer resulting in failure of amplification during the PCR reaction. Null alleles can also occur via segmental aneuploidy, where one chromosome has a deletion containing the primer binding site. For instance, in a study of allelic BAC sequences in the Pacific oyster obtained by the Oyster Genome Consortium, 42 of 101 microsatellite loci were found to occur in the hemizygous state, owing to indels of various sizes (P. Gaffney, pers. com., 2016). Null alleles can occur in homozygous state when the samples produce no

amplification at all, and in heterozygous state, when the sample appears as a homozygous individual for a particular locus. *Nantón et al. (2014)*, characterised the first nineteen polymorphic microsatellite markers in *D. trunculus* to assist in the management of the species and allow the delineation of conservation units essential in fisheries. Of the 19 loci characterised in this study, at least 10 showed the presence of null alleles with frequencies ranging from 0.109 to 0.277. Another study aiming to examine the genetic variability and relationships between two Manila clam (*Ruditapes philippinarum*) populations in Korea also suggested the occurrence of null alleles in 7 loci out of 10 (*Kim et al., 2014*). More recently *Chiesa et al. (2016)* demonstrated the occurrence of null alleles through rigorous analytical tests in this species. The presence of null alleles has also been reported in a study that assessed the genetic diversity of four wild and six hatchery stocks of the hard clam (*Mercenaria mercenaria)* in Florida. Null alleles were present in all loci with frequencies ranging from 0.015 to 0.296, but for four loci out of seven, the frequency of null alleles was particularly high (>0.146) (*Hargrove et al., 2015*).

Simulation studies suggest that null alleles with frequencies between 5% and 8% should have only minor effects on classical estimates of population differentiation, but that higher frequencies would bias such parameters (*Chapuis & Estoup, 2007*). Null alleles have been associated with heterozygous deficits in many studies and several have tested their presence and effects in population structure estimates using various analytical and simulation tools (e.g., *Dąbrowski et al., 2015* and references therein). However, to the best of our knowledge no study has empirically tested for its consequences in estimating differentiation using population samples separated by well characterised biogeographic barriers.

The wedge clam constitutes an important fishing resource in southern Spain and Portugal due to its high economic value (*Gaspar, Ferreira & Monteiro, 1999*). In Europe alone, the recorded landings in the last 12 years total 11,202 tons with a maximum yield of 1,355 tons in 2005 (*FAO-FIGIS, 2016*). Overall, the species has experienced a steady decline in recent years reaching only 757 tons in 2014 (*FAO-FIGIS, 2016*). Despite its overexploitation threat, *D. trunculus* has only recently been studied from a population genetics (*Marie et al., 2016*). Understanding connectivity among populations through genetic structure provides tools to determinate the appropriate units and spatial scale for fisheries conservation and management (*Waples & Gaggiotti, 2006*; *Funk et al., 2012*). *D. trunculus* is an Atlantic–Mediterranean warm-temperate species found in the Black Sea, in the Mediterranean Sea (*Bayed & Guillou, 1985*) and from Senegal to the northern Atlantic coast of France (*Tebble, 1966*). It inhabits sandy beaches exposed to tidal rhythms, characterised by intense wave action and sediment instability (*Brown & McLachlan, 1990*). In these environments, populations are capable of reaching sufficiently high densities to support large commercial fleets (*Gaspar, Ferreira & Monteiro, 1999*). Because, wedge clams are filter feeders, they play important roles in the trophic structure of beaches, but can also accumulate xenobiotic compounds making them, ideal model organisms for environmental monitoring (*Saavedra & Bachere, 2006*). For this reason, they may also constitute potential risks for human health when they are eaten and thus have been extensively studied from an ecotoxicology perspective (*Tlili et al., 2010*; *Yawetz et al., 2010*;

Bouzas et al., 2011; *Company et al., 2011*; *Hamdani & Soltani-Mazouni, 2011*; *Tlili et al., 2011*).

Motivated by our interest in the conservation genetics of this important shellfish resource, we used sequences generated by one NGS platform (Roche-454) for microsatellite characterisation. Evidence of successful PCR amplification and allelic size variation across a subset of population samples allowed us to optimise 16 loci into four multiplex panels, permitting the PCR amplification of several markers simultaneously. We then genotyped two adjacent locations in the southern Atlantic Coast of Spain. One of them is situated within the boundaries of the Doñana National Park, exploited, due to historic rights, only by artisanal hand-operated dredges with restricted quotas. The second is approximately 85 km apart and is exploited by a commercial boat-operated dredging fleet. We also genotyped a sample from the south Mediterranean Coast of Spain. These samples, located on both sides of the Gibraltar strait, which is a well-known major barrier to gene flow (*Perez-Losada et al., 2007*; *Sá-Pinto et al., 2012*), allowed us to test the strength of the analyses using loci that have high frequencies of null alleles in this species. Finally, we also show through empirical and analytical tests that null alleles are ubiquitous in this species as previously suggested (*Nantón et al., 2014*).

MATERIALS & METHODS

Sampling

Samples of *D. trunculus* were collected from three sites, two located in the Atlantic coast and one in the Mediterranean Sea. In the southwest Atlantic Coast of Spain, we collected samples from two adjacent localities, Isla Canela ($n = 66$; 37°10′11.334″N, 7°22′14.25″W) and Doñana National Park ($n = 77$; 36°55′6″N, 6°28′20″W), in the province of Huelva. The third sample was obtained from the Mediterranean Coast of Spain from Caleta de Vélez, Malaga ($n = 52$; 36°44′49″N, 4°4′32″W). The Euclidean geographic distance between the first 2 sites over coast line is 85 km; while between the second and the third site is 287 km (128 km from Doñana National Park to the Strait of Gibraltar and 158 km from there to the third site). The first two samples are separated from the third one by the Strait of Gibraltar which is well-recognised as a major geographical barrier to dispersal among marine taxa (*Perez-Losada et al., 2007*; *Sá-Pinto et al., 2012*). Permit of sampling was approved by the Consejeria de Medio Ambiente (permit number 2011107300001463).

Library preparation and 454 GS-FLX pyrosequencing

DNA was extracted from different tissues using a slightly modified version of *Aljanabi & Martinez (1997)* salting out method. Namely, after addition of the saline solution, the mixture was centrifuged for 30 min at 10,000 g. Also, DNA precipitation was achieved by incubation at −20 °C with 600 μL isopropanol for 30 min. Then, after washing out the pellet with 70% ethanol, we centrifuged at 10,000 g for 10 min.

For the preparation of the SSR enriched library we followed the protocol of *Santana et al. (2009)*. In brief, we started with 20 μg of high molecular weight genomic DNA (gDNA) to enrich for SSR with ISSR-PCR method (*Zietkiewicz, Rafalski & Labuda, 1994*), but without using the final stage of cloning in a bacterial vector. The primers used were:

ISSR1 (5′-DDB (GTC) 5–3′), ISSR2 (5′-DHB (ATC) 5-3′), ISSR3 (5′YHY (GT) 5G-3′), ISSR4 (5′-HVH (CAT) 5-3′), ISSR5 (5′-NDB (TGT) 7C-3′), ISSR6 (5′-NDV (CT) 8–3′), and ISSR7 (5′-HBDB (AAC) 4–3′). For each trial microsatellite-enriched DNA we analysed in 5 μg in the Pyrosequencing platform Roche 454 GS-FLX. For each sample, a single-lane sequencing run using portioned sections of the PicoTiterPlate was performed according to the manufacturer's protocol preparation. Sample preparation and analytical processing such as base calling, were also performed on site according to the manufacturer's protocol.

Contig assemblies and microsatellite detection

To uncover sequences containing microsatellites in the pyrosequence reads, we used the mreps software which has been designed to isolate and characterise highly polymorphic microsatellite loci (*Kolpakov, Bana & Kucherov, 2003*). For the purpose of evaluating the number of microsatellite-containing reads from the dataset, we defined two filtering steps. In the first step, we included loci containing di- and tri-, nucleotides SSR with 20 to 60 base pairs of repeated units (i.e., at least 10 tandem repeats for di-, and seven for tri-nucleotides), and a maximum proportion of non-perfect repeats of 0.333.

Since hundreds of potential loci were detected in this filtering step and only loci with sufficiently large flanking sequences around the microsatellite are potentially useful for primer design for successful PCR amplification, we selected loci with a minimum of 40 flanking nucleotides on each side of the tandem repeated units. These loci are referred as "Potentially Amplifiable Loci", or PAL (*Castoe et al., 2012*). From the 85 PAL identified, we selected 52 sequences for subsequent analyses. Primer design was performed using the on-line version of Primer 3 (*Rozen & Skaletsky, 2000*). A total of 52 primer pairs were thus designed and synthesised with the universal M13 primer in the 5′ end of the forward primer for their later analysis using a fluorescently labelled M13 primer according to *Schuelke (2000)*.

Multiplex optimisation

In the initial development phase, eight individuals were used to verify the amplification of the expected products for the 52 markers. PCR amplifications were performed in a reaction volume of 15 μl, containing 15 ng of gDNA, 0.5 pmol of each primer, 75 μM of each dNTP, 1.5 mM of $MgCl_2$, 1× PCR Green Master Mix buffer (Promega, Madison, WI USA) and 0.5 units of *Taq* polymerase (Promega). The PCR amplifications were done on an Applied Biosystems 9700 DNA thermal cycler using the following conditions: an initial denaturation step at 95 °C for 10 min followed by 35 cycles of 30 s at 94 °C, 1min at either 58 or 50 °C, depending on the primer's melting point, and 1min 30 s at 72 °C, and a last cycle of extension at 72 °C of 10 min. The PCR conditions of successful amplifications were recorded for each locus in order to make the subsequent multiplexing easier. The amplification products were visualized by agarose gel electrophoresis using ethidium bromide staining. A total of 28 primer pairs amplified a clear PCR product and were selected for assessing their allelic size ranges using 48 individuals from the Doñana National Park. The rationale for using 48 samples is that by genotyping many, we aimed to detect a wide spectrum of the genetic diversity present in the population and thus

help in the subsequent design of the multiplex panels. For this, we directly incorporated the fluorescently labelled M13 primer with 6FAM^{TM} (*Schuelke, 2000*), using the same PCR conditions described above. Polymorphism and profile quality were first verified in simplex reactions. The PCR products were then separated on a capillary sequencer (ABI 3130× Genetic Analyzer; Applied Biosystems, Waltham, MA, USA) using GeneScan^{TM} 500 LIZ® Size Standard. Allele sizes were determined with the Gene Mapper® V4.1 program (Applied Biosystems, Waltham, MA, USA). From this analysis, we identified 21 loci that produced clear PCR products with relatively low levels of stuttering that were deemed of sufficient quality for multiplexing. The forward primer of each of these loci was then synthesised again, but this time incorporating in the 5′ end one of the four fluorescent markers 6FAM^{TM}, VIC®, NED^{TM} and PET® for multiplex PCR analysis.

We carried out a high number of tests to find optimal multiplexes, grouping several markers of different allelic sizes in a single reaction and adjusting the concentration of the primers according to the strength of the signal. For the multiplex reactions, we used the Qiagen® Multiplex PCR Kit (Qiagen, Toronto, CAN). Final volumes and concentrations of the master mixes were optimised to reduce the total genotyping cost (*Guichoux et al., 2011*). PCR reactions were done in a final volume of 10 µL for all multiplexes, with 3 µL of primer mix, 5 µL of Qiagen Multiplex Mix, and 2 µL of gDNA template (10–15 ng/µL). The cycling conditions for the four multiplexes differed only in the annealing temperature (Table 1): an initial step at 95 °C for 15 min; then 35 cycles at 94 °C for 30 s, 54 to 61 °C for 3 min and 72 °C for 1 min; and a final elongation step at 60 °C for 30 min. Only 16 loci produced codominant fragments (i.e., one allele of maternal and one of paternal origin) in the 48 samples analysed. Table 1 summarises the information about the primers sequences, concentrations in the primer premix (µM), GenBank accession numbers, and multiplex panel annealing temperatures.

Genotype scoring and analyses

Once multiplex conditions were optimised, we screened a total of 143 individuals from the south Atlantic coast and 52 from the south Mediterranean coast of Spain for variation at 16 polymorphic microsatellite loci (Table 1: 16 loci). Several steps were taken to ensure the consistency and accuracy of the genotype analysis and scoring. An initial allele size spread table was developed from the genotypes of the 195 samples used in the multiplex PCR amplifications for each locus. This table consisted of the observed size ranges for a given allele in each locus, as determined by the GeneMapper® v4.1 (Applied Biosystems). The table was further refined with the same set of samples genotyped in multiplex reactions and then used to manually size each sample according to the scoring range. Afterwards, any sample with a size value beyond the established range was reanalysed. In order to ensure that the allele spread calibration held for each set of samples analysed, we always included the same 6 individuals in each plate to be genotyped as reference standards. All genotype size scores were checked twice to ensure size scoring accuracy. Furthermore, for each marker, clear reading rules were defined and illustrated using screen shots of the chromatograms and provided to a second, experienced scorer to ensure scoring consistency. A first estimate of error rate was obtained by counting mismatches on the

Table 1 **Details of the 16 microsatellite markers developed and optimized for multiplex.** One hundred ninety-five individuals were used for this analysis. Locus identity, type of florescence dye in the forward primer, panel numbers (MP) indicate groupings of loci sharing a multiplex in the PCR reaction, primer sequences, repeat motif SSR, primer concentrations ([C] μM) in the primer mix, number of alleles (N_A) found, range of allele sizes (Range, bp), observed (H_O) and expected (H_E) heterozygosity, F_{IS}, departure from Hardy–Weinberg equilibrium (HWE), polymorphism information content (PIC), and GenBank accession numbers. Annealing temperatures for each panel were as follows: Kit 3 = 60° C, Kits 1,2,4 = 58 °C.

Locus ID	Fluoro	MP	Forward primer	Reverse primer	SSR	[C] μM	N_A	Range	H_O	H_E	F_{IS}	PIC	GenBank Acc. N°
D.tru2	6-FAM	4	ATTCTCCTACGGAGGGGCTA	GCGATGATTTCCTCCGTAAA	(ACA)	3	17	97–148	0.63	0.84	0.247	0.83	HG792255
D.tru4	6–FAM	3	TGCACTTATAATCAACCGGAAG	CTTCCAGCAACACCACGTC	(TG)	3.5	12	151–177	0.69	0.75	−0.069	0.71	HG792256
D.tru6	PET	3	GTTTTCTCACAGGCGTTCG	GCAGTGATAGGGTTAACGTATTTG	(CA)	3.5	23	69–131	0.59	0.85	0.316	0.84	HG792257
D.tru8	PET	4	AATATATTGCAGGCTGGTAGGG	TAAAATTGCCATGCGTGCAG	(ATC)	3	11	127–157	0.63	0.81	0.206	0.78	HG792258
D.tru11	6-FAM	2	AGAACCTGATGTGCTGTGGA	CACGTTAGTACAAAGACCCTTTCC	(GT)	1	21	101–145	0.57	0.79	0.268	0.78	HG792259
D.tru14	NED	4	TTTTTGTTCTTCTGAATAGTGCAA	TCGCCATCTTTTGTTGTTGT	CAA)	3	14	75–129	0.15	0.54	0.738	0.51	HG792260
D.tru15	6-FAM	1	TGTCACTAATACAGGATTTCTCACG	AATAGCATCTCTCACACAGACACA	(AC)	5	27	219–279	0.62	0.93	0.332	0.92	HG792261
D.tru16	VIC	3	TGCTCCTTATCATTTCAATTGTG	TGCAAACCATCTTCTGGTTG	(AC)	4.5	20	70–114	0.82	0.90	0.107	0.89	HG792262
D.tru19	VIC	1	AACACCCATAGCGACGAAAA	GATGACCTGTGAATACATGAAGGA	(AC)	1.5	23	126–180	0.73	0.94	0.214	0.93	HG792263
D.tru22	PET	1	TGAAGACATGGCAAAATCCA	TGAGCATATTTCTCTTTCGTAGG	(TTG)	3	8	218–245	0.27	0.69	0.601	0.64	HG792264
D.tru23	NED	1	CAAGCACGTTAGACAAAGTCC	ACCTGATGTGTTGTGGACGA	(AC)	1	28	105–167	0.73	0.84	0.123	0.83	HG792265
D.tru26	NED	2	TGGAGGTAATTAGATGGTCCAG	ACGCTGGCATCGTTCTCTAT	(AG)	0.8	22	57–119	0.65	0.92	0.297	0.91	HG792266
D.tru29	VIC	2	TGAATTTAGTGATTGGCAAAGCTA	ACGGGTGGCATACAACTTGA	(TGT)	4	20	262–322	0.31	0.88	0.638	0.87	HG792268
D.tru32	VIC	4	CCGAATGTCCCTTTTGTTGT	TGGGTCCTGGAGGGTAAAAT	(TTG)	3	18	196–281	0.28	0.66	0.572	0.64	HG792271
D.tru40	PET	1	GACATTAAGGAGTGGTTGCGTA	CATCAACCGAAAACTCTATAAACTG	(TG)	1	17	119–155	0.60	0.88	0.304	0.87	HG792273
D.tru49	NED	4	GAGTATTTCTAACGGTCTTCAAGTTAT	GCATTTATCTTATGTGGTGTTTGC	(CA)	3	11	148–182	0.24	0.78	0.663	0.74	HG792275

basis of these six positive controls repeated two times in each multiplex panel. Positive controls were essential to include as they allowed the verification of differences across genotyping runs. A second estimate of error rate was obtained by comparing scoring across scorers. Two types of errors were distinguished. Type A corresponds to cases where reader 1 called a genotype as heterozygous and reader 2 typed it as homozygous, or vice versa. Type B corresponds to cases where a different allele was called by either of the scorers (*Castoe et al., 2012*). For each marker, we calculated the total number of alleles (N_A), the observed (H_O) and expected (H_E) heterozygosities, the polymorphism information content (PIC) and F_{IS} using the program CERVUS 3.0 (*Kalinowski, Taper & Marshall, 2007*) while the null-allele frequencies (NAF) were determined using the program ML-NULLFREQ (*Kalinowski & Taper, 2006*). Private alleles, defined as alleles for a given locus present in only one of the two populations (South Atlantic Coast of Spain and South Mediterranean Coast of Spain) were estimated using Convert version 1.31 (*Glaubitz, 2004*). Samples were also tested for linkage disequilibrium and departure from Hardy-Weinberg equilibrium by the Markov chain method using 100,000 dememorization steps with the program Arlequin V 3.5 (*Schneider, Roessli & Excoffier, 2000*; *Excoffier, Laval & Schneider, 2005*). Finally, we investigated whether there was any genetic differentiation between the two Atlantic adjacent localities and subsequently between these and the Mediterranean sample using weighted F_{ST} (θ) statistics by estimating the pairwise fixation index based on allele frequency variation over all loci (*Weir & Cockerham, 1984*). The significance of genetic subdivision was assessed using 1,000 permutations in Arlequin.

We also estimated null allele frequencies and F_{ST} using the software FreeNa (*Chapuis & Estoup, 2007*) with a number of replicates fixed to 25,000. We ran this analysis using the ENA correction method to efficiently correct for the positive bias induced by the presence of null alleles on F_{ST} estimation and so, to provide an accurate estimation of F_{ST}. We ran both software programs in order to see if the presence of high frequencies of null alleles can affect the F_{ST} estimates knowing that Arlequin does not take into account the presence of null alleles compared to FreeNa.

Inbreeding, null allele estimations and genotype artefacts

The possibility of genotyping artefacts was examined in several ways. First, a subset of samples ($N = 94$) were re-amplified for each of the 16 loci under relaxed annealing temperatures (4 °C lower) to verify that additional alleles had not been missed. Second, we used the software MicroChecker (*Van Oosterhout et al., 2004*) to test for stuttering and large allele drop-out and to obtain various estimates of the expected frequency of null alleles (r). These, in turn, were used to calculate the expected number of null homozygotes (i.e., when both alleles remain undetected) using the $4(Nr^2)$ (*Brookfield, 1996*) so that we could compare these estimates to the actual number of individuals that fail to amplify systematically in repeated PCR reactions. The analysis yielded a close match between the expected and observed number of null homozygous. Third, we estimated the frequency of null alleles using ML-NULLFREQ (*Kalinowski & Taper, 2006*) which is a maximum likelihood estimator of the frequency of null alleles in a sample with or without missing data which has been shown to be the single best-performing method (*Dąbrowski et al., 2015*).

Finally, to test the scoring's robustness and power of resolution to depict population genetic structure of these microsatellite loci, we took advantage of the fact that it has extensively been shown, in many marine taxa, that Atlantic populations are often genetically distinct from Mediterranean populations (*Patarnello, Volckaert & Castilho, 2007*). Therefore it is reasonable to hypothesise that the samples from the Atlantic coast separated by 85 km of continuous shoreline would not be genetically distinct from each other, but would be differentiated from the Mediterranean population. To test this hypothesis even when the presence of null alleles in our loci was already confirmed, we took advantage of the algorithm introduced by *Falush, Stephens & Pritchard (2007)* in the program STRUCTURE V2.3.4 (*Hubisz et al., 2009*), which employs a Bayesian clustering method to infer the most likely number of populations (K) assuming no a priori structure and allows for the presence of null alleles in the dataset. To test the effects of null alleles in the population structure depicted by these loci, we first estimated the most likely number of populations (K) assuming no null alleles and using the LOCPRIOR option and subsequently, we repeated the analysis but incorporating the RECESSIVEALLELES as well as the LOCPRIOR options in STRUCTURE. In this case, the program assumes that the recessive allele is never observed in homozygous state but it might be present. This analysis, partitions multilocus genotypes into clusters, while minimising departure from Hardy–Weinberg expectations and linkage disequilibrium among loci, and it estimates individual ancestry proportions to each putative cluster. We therefore investigated the most likely K running ten independent simulations. All simulations were run using default parameters in the admixture model and with correlated allele frequencies. Each run included 100,000 iterations of burnin, followed by 500,000 MCMC iterations used for parameter estimation. The most likely value of K was chosen using the delta K (ΔK) statistic (*Evanno, Regnaut & Goudet, 2005*), based on the rate of change between successive K values using the STRUCTURE HARVESTER software (*Earl & Von Holdt, 2012*). The algorithm employed by this online program to determine ΔK by the Evanno method requires that at least three values of sequential K were analysed. Thus, we ran ten independent simulations of $K = 1$–3, despite knowing that the maximum number of populations between Atlantic and Mediterranean samples is likely 2. Then, potential admix individuals in each of the proposed clusters were identified using posterior probabilities calculated for each individual in STRUCTURE.

Because of the high frequencies of null alleles, we also ran STRUCTURE with the complete data but removing a different locus at a time for each new run using the same parameters as described above (and including the RECESSIVEALLELES option). This analysis was performed in order to identify the level of information provided by each locus on the level of assignment. Then, we selected the most informative loci and ran again STRUCTURE with and without the RECESSIVEALLELES options still using the same parameters. We selected the most informative loci from the run using the RECESSIVEALLELES option. For that, we created a reference assignment's level from the average level of individuals' assignment of the South-Atlantic Coast of Spain and the South-Mediterranean Coast of Spain respectively. If the average level of individuals' assignment of a run (run when removing a locus) was under the reference one, then, we considered that the locus we removed for this run was very informative.

RESULTS AND DISCUSSION

A total of 18,795 reads with an average length of 232 bp were obtained from 454 pyrosequencing of a microsatellite enriched library and assembled into 595 contigs. We identified 267 contigs (44.9%) that contained microsatellite loci with at least 10 di- or seven tri-nucleotide repetitions. Of these, there were a total of 85 contigs (31.8%) we identified as PAL of di- or tri-nucleotide microsatellites and 182 contigs (68.2%) as nonPAL (i.e., contigs containing di- or trinucleotide microsatellites without suitable flanking PCR-primer sites). Comparing across the two classes of repeats in PAL sequences, di-nucleotide repeats represented 66.8% and tri-nucleotides 33.2%.

From the microsatellite containing sequences, we selected 52 loci for testing their usefulness in multiplex genotyping and 37 produced PCR amplification when visualised on agarose gels and ethidium bromide staining. Of these 37 loci, nine gave multiple bands, faint bands, or fragments clearly defined but of a different size than the one expected from the original sequence. These loci were discarded from subsequent analyses. Of the remaining 28, 21 could be genotyped and displayed polymorphism in 48 individuals. These 21 loci were deposited in GenBank under accession numbers HG792255.1–HG792275.1. Five loci, D.tru28, 30, 31, 39 and 46, amplified two or three loci simultaneously and were discarded from further analyses. For the remaining loci, the number of alleles per locus ranged from eight to 28 ($\bar{x} = 18.25$, $\sigma = 5.86$). Expected and observed heterozygosities, and polymorphic information content ranged from 0.15 to 0.82 ($\bar{x} = 0.53$, $\sigma = 0.21$), 0.54 to 0.94 ($\bar{x} = 0.81$, $\sigma = 0.11$), and 0.51 to 0.93 ($\bar{x} = 0.79$, $\sigma = 0.12$), respectively. F_{IS} estimates ranged from -0.07 to 0.74 ($\bar{x} = 0.35$, $\sigma = 0.04$) (Table 1). Private alleles were present in all loci for the south Atlantic Coast of Spain and in 11 loci for the South Mediterranean Coast of Spain (Table S1). There was no evidence of large allele dropout or genotyping errors due to stutter peaks for the 16 loci. However, all loci but two (D.tru4 and D.tru16), showed a significant excess of homozygotes (deviation from Hardy–Weinberg equilibrium, $P < 0.001$). There was no evidence of linkage disequilibrium between any pair of loci. Genotyping was consistent across the 48 samples, which were each tested twice yielding the same genotypes. The multiplex optimisation yielded a total of four panels for the 16 primer pairs. The numbers of possible multiplex combinations was limited due to the allele size ranges of the markers and differences in the annealing temperatures of the primers. We were thus able to design two 3-plex, and two 5-plex multiplex reactions after excluding the duplicated loci.

Before the advent of next-generation sequencing technologies, no genetic marker has found such widespread use as microsatellites in the last two decades. This study joins an ever increasing number of others that have successfully employed NGS to detect a virtually unlimited number of polymorphic microsatellites across a wide variety of taxa (e.g., (Wang et al., 2012b; Wasimuddin et al., 2012; Whitney & Karl, 2012; Miller et al., 2013). Our study takes advantage of Roche-454 titanium chemistry. Microsatellite-containing sequences from shotgun and enriched libraries have also been reported across a broad range of taxa in recent years (Angeloni et al., 2011; Santure et al., 2011; Wang et al., 2012a; Wang et al., 2012b; Zhang et al., 2012). Together, these studies demonstrate that the main

obstacle of costly *de novo* isolation for using microsatellite markers for population studies and characterisation has been rapidly overcome by the advent of NGS technologies. More importantly, however, is the fact that a growing proportion of research projects, working on modest financial resources, are now capable of developing a suitable number of microsatellite markers at minimal laboratory costs for virtually any taxa (*Schoebel et al., 2013*). The number of microsatellite loci detected in this study illustrates how abundant these sequences are in eukaryotic genomes. Comparing the number of loci detected across studies and taxa is inherently difficult because genomes vary substantially in their frequency of microsatellites and genome size (*Schoebel et al., 2013*). However, the number of loci detected here favourably compares with other studies (e.g., *Wang et al., 2012a*; *Wang et al., 2012b*; *Zhang et al., 2012*). For example, in a recent study that used this technology, *Schoebel et al. (2013)* carried an extensive survey of microsatellite abundance in 17 non-model species including plants, fungi, invertebrates, birds and a mammals and determined if flanking regions were suitable for primer development. They concluded that that depending on the species, a different amount of 454 pyrosequencing data might be required for successful identification of a sufficient number of microsatellite markers for ecological genetic studies. Irrespective of the difficulty to compare microsatellite abundance in different taxa, an important message of our study is that a virtually unlimited number of microsatellite loci can be identified from the large amount of sequence data generated with NGS technologies. It is worth noticing as well that 454 pyrosequencing has been replaced by other technologies which generally yield many more sequences, but with shorter reads. The implications of these shorter reads may have an important impact on marker development as microsatellites require the availability of long flanking sequences for primer design. However, we have to consider that, despite the many advantages of microsatellite markers, admittedly, the development of SNP markers by NGS or SNP genotyping by sequencing have increasingly become attractive alternatives for many reasons such as availability of high numbers of annotated markers, low-scoring error rates, improved genotyping results for poor quality samples, highly informative in terms of segregation among populations, and the ability to examine neutral variation and regions under selection, among others (*Brumfield et al., 2003*; *Rosenberg et al., 2003*; *Morin et al,, 2004* ; *Liu et al., 2005*; *Grewe et al., 2015*).

Once the multiplex panels were optimised, we genotyped a total of 195 individuals from the three sampled locations. Table 1 summarises the genetic diversity found in each locus for all the samples analysed. The initial uncorrected typing error rate was 0.0023 per reaction or 0.0015 per allele. The frequency of single-locus genotypes missing in the data set was <0.05. Adjacent allele scoring error analysis of the final corrected data set yielded no significant deviations ($X^2 < 2$, $P > 0.15$), hinting that there were very few remaining scoring errors. From individuals genotyped as controls in each plate, we calculated a mistyping rate of 0.0019 per allele genotyped. Type A and type B error rates were 0.015 and 0.021 for multiplex 1, 0.013 and 0.008 for multiplex 2, 0.026 and 0.051 for multiplex 3, and 0.009 and 0.014 for multiplex 4, and 0.0152 and 0.023 across all loci and were mainly caused by excessive stuttering of some long alleles. Once again, all loci but D.tru4 and D.tru16 showed a significant excess of homozygotes (deviation from Hardy–Weinberg equilibrium,

Table 2 Characteristics of the 10 microsatellite markers with original (OA) and new (NA) annealing temperature for 94 individuals (Isla Canela: 27; Doñana: 37; Caleta de Vélez: 30) including the number of alleles (N_A), range of allele sizes (Range, bp), observed (H_O) and expected (H_E) heterozygosity, polymorphism information content (PIC) and null-allele frequency (NAF).

Locus	N_A		Range		H_O		H_E		PIC		NAF	
	OA	NA	OA	NA	OA	NA	OA	NA	OA	NA	OA	NA
D.tru4	11	34	151–179	99–209	0.641	0.809	0.759	0.930	0.718	0.920	0.093	0.065
D.tru6	18	26	71–131	69–161	0.527	0.923	0.842	0.882	0.821	0.868	0.239	−0.030
D.tru11	20	19	101–145	111–149	0.581	0.723	0.807	0.820	0.792	0.801	0.173	0.056
D.tru15	25	23	219–279	219–281	0.670	0.759	0.928	0.936	0.918	0.926	0.158	0.101
D.tru19	22	20	126–176	128–170	0.734	0.841	0.937	0.937	0.928	0.927	0.120	0.052
D.tru22	7	8	221–245	215–245	0.233	0.333	0.658	0.692	0.594	0.636	0.480	0.342
D.tru23	22	22	107–159	107–151	0.702	0.886	0.838	0.864	0.823	0.850	0.093	−0.020
D.tru26	19	21	57–103	67–111	0.696	0.783	0.924	0.934	0.913	0.925	0.140	0.086
D.tru29	17	13	268–322	265–310	0.344	0.203	0.872	0.851	0.855	0.828	0.434	0.611
D.tru40	17	18	119–163	121–161	0.596	0.830	0.866	0.885	0.849	0.869	0.182	0.030

$P < 0.001$) (Table 1). Exact tests for genotypic linkage disequilibrium confirmed the absence of physical linkage among most loci ($P < 0.05$ after Bonferroni correction).

The results of the re-amplification of a subsample of individuals under lower stringency mostly confirmed the above findings. However, only 10 loci produced codominant fragments (i.e., one allele of maternal and one of paternal origin) under relaxed annealing temperatures (Table S1 : Original Annealing Temperature, New Annealing Temperature). Locus D.tru16, which initially had not deviated from HWE, amplified multiple loci and thus could not be used for this comparison. Furthermore, two additional loci showed null allele frequencies not different from zero (D.tru6 and D.tru23). This result suggests that non-amplifying alleles can be detected by reducing the stringency of the annealing temperature. For the remaining eight loci, no differences were observed in the heterozygous deficit or the null allele frequencies supporting the presence of null alleles in the wedge clam genome (Table 2).

The fact that most loci had large heterozygous deficits was an intriguing result which deserves a thorough discussion. The first consideration that needs to be evaluated in addition of the occurrence of null alleles, is the presence of large allele dropout. Another consideration is the level of stuttering as it may hamper the reliable scoring of alleles differing within one or two repeated units in some loci. Another possible explanation could be the differences in allele size as reported by previous studies that have found that loci with longer alleles tend to have higher dropout rates than those with shorter alleles (*Sefc et al., 2003*; *Buchan et al., 2005*; *Broquet, Ménard & Petit, 2007*). In our study, allele size ranges ranged from 60 to 24 ($\bar{x} = 46.2$, $\sigma = 12.12$) and from 110 to 30 ($\bar{x} = 54.7$, $\sigma = 26$) for the original and relaxed annealing temperatures respectively (Table 2). It is thus expected that the largest allele at a locus would amplify with less efficiency than the shortest one as the latter would outcompete the former during amplifications (*Sefc et al., 2003*; *Buchan et al., 2005*; *Broquet, Ménard & Petit, 2007*). This will artificially increase the frequency of null alleles even in the absence of true null alleles.

The analysis of null allele frequencies per population using the maximum likelihood estimates implemented in ML-NullFreq confirmed the presence of null alleles in all loci but D.tru4 and D.tru16. For these loci, null-allele frequencies estimates ranged between 0% and 31.2% ($\bar{\chi} = 0.160$, $\sigma = 0.098$) and all loci but D.tru4, D.tru16 and D.tru23 had a null-allele frequencies above 5% (Table 3). Results from FreeNa showed lower null allele frequencies for each locus (except for the locus D.tru19) than results obtained using ML-NullFreq. Null allele frequencies estimates ranged between 0% and 29.3% ($\bar{\chi} = 0.149$, $\sigma = 0.090$; Table 3). This high frequency of null alleles found in this study is consistent with many reports in other bivalve species (*Launey et al., 2002*; *Nantón et al., 2014*; *Chiesa et al., 2016*). Clams are filter feeders and thus, are prone to accumulate xenobiotic compounds (*Tlili et al., 2010*; *Yawetz et al., 2010*; *Bouzas et al., 2011*; *Company et al., 2011*; *Hamdani & Soltani-Mazouni, 2011*; *Tlili et al., 2011*). It is plausible that the mutagenic action of some pollutants such as heavy metals (e.g., mercury or cadmium) (*Wong, 1988*) can be responsible for an increased mutation rate in clams. This could be further investigated by comparing samples representing genetically distinct populations where organisms are either heavily exposed to pollutants in beaches adjacent to major urban areas or those occupying habitats in remote locations where the level of pollution is likely to be low and far from industrial pollution.

As expected due to the absence of biogeographic barriers and the proximity of the sampling sites in the south Atlantic coast of Spain, the fixation index between samples from Isla Canela and Doñana National Park was not significantly different from zero (Arlequin: $F_{ST} = 0.0028$–$P > 0.05$; FreeNa: F_{ST} including ENA correction $= 0.0025$) indicating uninterrupted migration of larvae across the area. Furthermore, as predicted, the fixation index between these two localities and the Mediterranean sample was significantly different from zero (Arlequin: $F_{ST} = 0.033$–$P < 0.001$; FreeNa: F_{ST} including ENA correction $= 0.032$). The STRUCTURE analysis found that the two populations from the south Atlantic coast of Spain were genetically homogeneous while they were significantly different from the Mediterranean population with the highest ΔK value equal to the number of biogeographic regions ($K = 2$) and the programme showed that the 10 independent runs from $K = 1$ to $K = 3$ produced consistent results whether or not the RECESSIVEALLELES option was used in the analysis (Fig. 1). Calculation of the statistic ΔK from the STRUCTURE runs indicated that the two biogeographic regions analysed have clearly differentiated populations.

When running STRUCTURE with the complete data set without a different locus at a time for each run, our results also showed that three microsatellite loci were very informative (D.tru4, D.tru29 and D.tru32; Fig. S1, Table S2). Indeed, compared to the reference average level of individuals' assignment ($84.7\% \pm 2.5$ and $90.6\% \pm 1.7$ for the South Atlantic Coast of Spain and the South Mediterranean Coast of Spain, respectively), without these three loci, the average level of individuals' assignment dropped to $65.5\% \pm 6.5$, $68.2\% \pm 3.7$ and $77.9\% \pm 2.6$ in the South Atlantic Coast of Spain and to $38.2\% \pm 5.4$, $89.5\% \pm 2.2$ and $89.9\% \pm 2.8$ in South Mediterranean Coast of Spain, respectively, when excluding either locus D.tru4, D.tru29 or D.tru32. Surprisingly, two of these loci (D.tru29 and D.tru32) have large frequencies of null alleles (0.311 ± 0.030 and 0.228 ± 0.027 for the locus D.tru29 and D.tru32 respectively; Table 3) confirming that even when microsatellite loci display

Table 3 Estimate of null allele frequencies per locus and populations. The maximum likelihood and the ENA correction methods were used, using ML-NullFreq and FreeNa respectively.

| Locus ID	Isla Canela		Doñana		Caleta de Vélez		Mean		SD	
	ML-NullFreq	FreeNa	ML-NullFreq	FreeNa	ML-NullFreq	FreeNa	ML-NullFreq	FreeNa	ML-NullFreq	FreeNa
D.tru2	0.168	0.168	0.106	0.106	0.107	0.09	0.127	0.121	0.036	0.041
D.tru4	0	0	0	0	0	0	0	0	0	0
D.tru6	0.106	0.106	0.112	0.112	0.222	0.173	0.147	0.131	0.065	0.037
D.tru8	0.144	0.114	0.125	0.125	0.05	0.044	0.106	0.095	0.05	0.044
D.tru11	0.113	0.113	0.139	0.117	0.084	0.084	0.112	0.105	0.028	0.018
D.tru14	0.307	0.264	0.356	0.238	0.273	0.273	0.312	0.259	0.042	0.018
D.tru15	0.199	0.16	0.189	0.164	0.164	0.142	0.184	0.155	0.018	0.011
D.tru16	0.014	0.014	0.033	0.033	0.083	0.083	0.043	0.043	0.036	0.036
D.tru19	0.088	0.088	0.111	0.111	0.102	0.102	0.1	0.101	0.012	0.012
D.tru22	0.3	0.282	0.263	0.248	0.23	0.204	0.264	0.245	0.035	0.039
D.tru23	0.038	0.038	0.064	0.064	0.008	0.008	0.037	0.037	0.028	0.028
D.tru26	0.09	0.09	0.165	0.165	0.204	0.165	0.153	0.14	0.058	0.043
D.tru29	0.277	0.277	0.322	0.322	0.333	0.28	0.311	0.293	0.03	0.025
D.tru32	0.258	0.258	0.219	0.219	0.206	0.206	0.228	0.228	0.027	0.027
D.tru40	0.103	0.103	0.223	0.223	0.09	0.09	0.139	0.139	0.073	0.073
D.tru49	0.299	0.299	0.359	0.335	0.222	0.222	0.293	0.285	0.069	0.058

Notes.

SD, standard deviation.

A

B
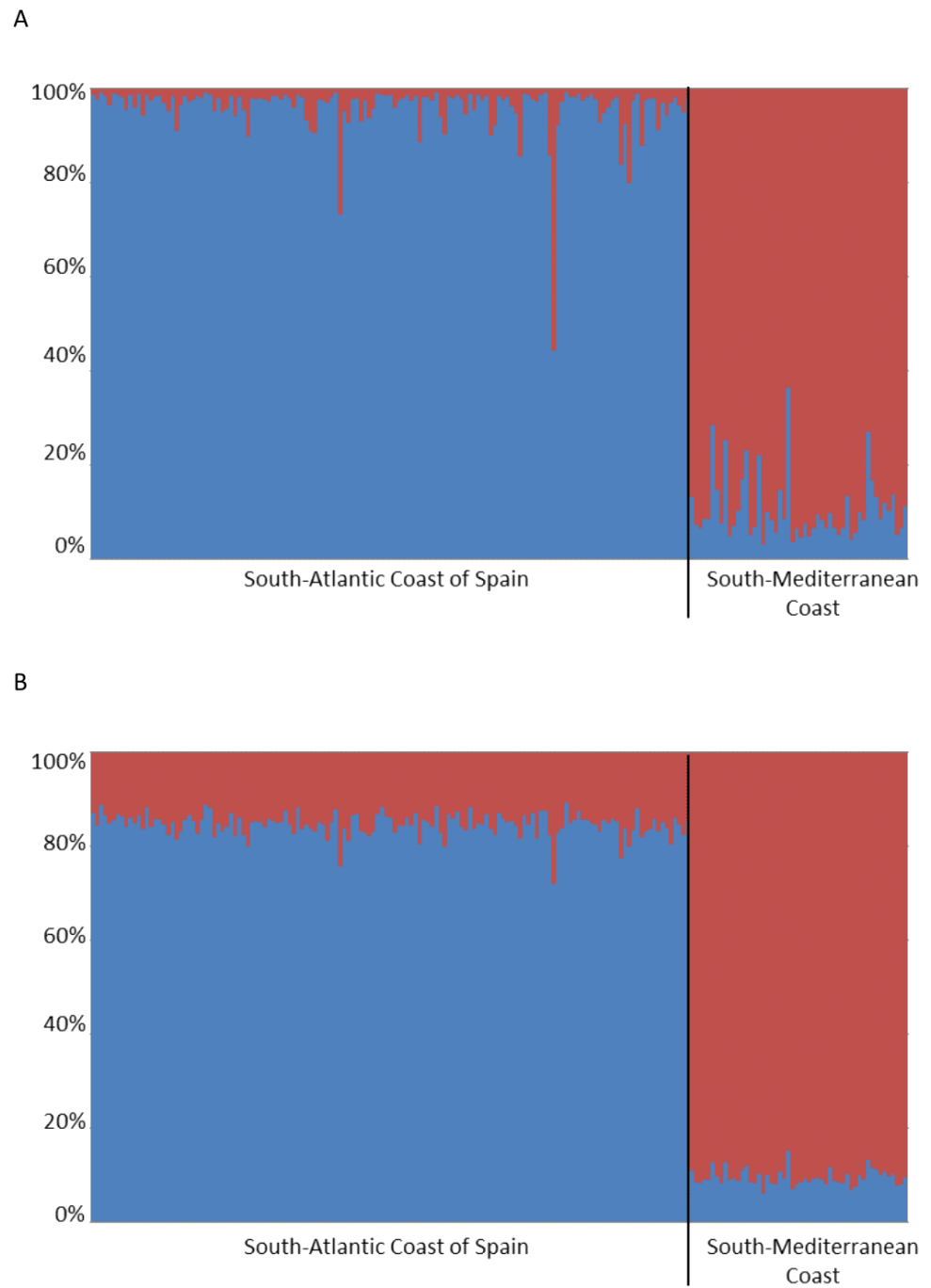
Figure 1 **Clustering analysis in STRUCTURE considering information about population of origin**
($k = 2$). Individuals are represented as vertical bars, where the amount of each colour indicates the proportion of each inferred cluster. A represents the clusters obtained in the simulations considering the absence of null alleles while B represents those assuming them. Isla Canela and Doñana are represented by the label "South-Atlantic Coast of Spain", and Caleta de Vélez by the label "South-Mediterranean Coast".

a large frequencies of null alleles, they can, nevertheless, be useful to identify the genetic structure of populations. This conclusion was also corroborated with the F_{ST} results using FreeNa and Arlequin, which showed that the use of the null allele's correction option had no effect in the F_{ST} estimates.

To the best of our knowledge, it is unusual that 87.5% of newly developed loci are affected by any form of genotyping errors or allele dropout during PCR due to the presence of long alleles. Although, simulation studies suggest that null alleles with frequencies between 5% and 8% should have only minor effects on classical estimates of population differentiation (*Chapuis & Estoup, 2007*), in our case null allele frequencies will render all these loci, but three, completely useless in population genetic analyses if the heterozygous deficits were due to null alleles. The fact that we found panmixia between samples from adjacent localities with no apparent barriers to gene flow and a well-defined structure for populations separated by a well-characterised biogeographic barrier suggests that the large heterozygous deficits we found, cannot be entirely due to null alleles or genotyping artefacts. If this was the case, one would have expected significant, albeit artificial, genetic differences between the two adjacent localities. Alternatively, it suggests that despite the presence of null alleles, clearly differentiated populations can be detected, although levels of genetic diversity are most likely underestimated. Admittedly, however, results pertaining to genetic structure presented here should be interpreted cautiously until they are corroborated by other approaches, namely SNP genotyping.

CONCLUSIONS

In conclusion, we have shown that null alleles at microsatellite loci are unusually abundant in the wedge clam and that genotyping artefacts are unlikely to completely explain the large heterozygous deficits observed in all population samples. We have also shown that the high frequencies of null alleles observed at these loci do not appear to have a significant effect in the population genetic parameters commonly assessed for microsatellite loci. The expected population structure between Atlantic and Mediterranean samples was confirmed for the tests carried out. Furthermore, the unusually high frequency of null alleles reported in the wedge clam and in other bivalve species is worth investigating further. As mentioned above, it is reasonable to hypothesise that pollutants may increase mutation rates in bivalves, resulting in high frequencies of null allele. Considering the economic value of *D. trunculus*, additional investigations are essential to validate this assumption.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Ciro Rico conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.
- Jose Antonio Cuesta, Pilar Drake and Enrique Macpherson contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper.
- Louis Bernatchez contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper, laboratory and sequencing facilities.
- Amandine D. Marie analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Permit of sampling was approved by the Consejeria de Medio Ambiente (permit number 2011107300001463).

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The locus described here are accessible via GenBank accession numbers: HG792255, HG792256, HG792257, HG792258, HG792259, HG792260, HG792261, HG792262, HG792263, HG792264, HG792265, HG792266, HG792268, HG792271, HG792273, HG792275.

Data Availability

The following information was supplied regarding data availability:

The raw data has been supplied as a Supplementary File.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3188#supplemental-information.

REFERENCES

Aljanabi SM, Martinez I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* **25**:4692–4693 DOI 10.1093/nar/25.22.4692.

Angeloni F, Wagemaker CAM, Jetten MSM, Den Camp HJMO, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Ouborg NJ. 2011. De novo transcriptome characterization and development of genomic tools for Scabiosa columbaria L. using next-generation sequencing techniques. *Molecular Ecology Resources* **11**:662–674 DOI 10.1111/j.1755-0998.2011.02990.x.

Bayed A, Guillou J. 1985. A contribution to the study of populations of the donax genus—the population of *donax-trunculus* l (mollusca, bivalvia). *Annales De L Institut Oceanographique* **61**:139–147.

Bouzas A, Aguado D, Marti N, Manuel Pastor J, Herraez R, Campins P, Seco A. 2011. Alkylphenols and polycyclic aromatic hydrocarbons in eastern Mediterranean Spanish coastal marine bivalves. *Environmental Monitoring and Assessment* **176**:169–181 DOI 10.1007/s10661-010-1574-5.

Brookfield JFY. 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology* **5**:453–455 DOI 10.1111/j.1365-294X.1996.tb00336.x.

Broquet T, Ménard N, Petit E. 2007. Noninvasive population genetics: a review of sample source, diet, fragment length and microsatellite motif effects on amplification success and genotyping error rates. *Conservation Genetics* **8**:249–260 DOI 10.1007/s10592-006-9146-5.

Brown AC, McLachlan A. 1990. *Ecology of sandy shores.* Amsterdam: Elsevier.

Brumfield RT, Beerli P, Nickerson DA, Edwards SV. 2003. The utility of single nucleotide polymorphisms in inferences of population history. *Trends in Ecology & Evolution* **18**:249–256 DOI 10.1016/S0169-5347(03)00018-1.

Buchan JC, Archie EA, Van Horn RC, Moss CJ, Alberts SC. 2005. Locus effects and sources of error in noninvasive genotyping. *Molecular Ecology Notes* **5**:680–683 DOI 10.1111/j.1471-8286.2005.01002.x.

Castoe TA, Streicher JW, Meik JM, Ingrasci MJ, Poole AW, De Koning APJ, Campbell JA, Parkinson CL, Smith EN, Pollock DD. 2012. Thousands of microsatellite loci from the venomous coralsnake Micrurus fulvius and variability of select loci across populations and related species. *Molecular Ecology Resources* **12**:1105–1113 DOI 10.1111/1755-0998.12000.

Chapuis M-P, Estoup A. 2007. Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* **24**:621–631 DOI 10.1093/molbev/msl191.

Chiesa S, Lucentini L, Freitas R, Nonnis Marzano F, Ferrari C, Filonzi L, Breda S, Minello F, Figueira E, Argese E. 2016. Null alleles of microsatellites for Manila clam *Ruditapes philippinarum*. *Animal Genetics* **47**:135–136 DOI 10.1111/age.12382.

Company R, Serafim A, Lopes B, Cravo A, Kalman J, Riba I, DelValls TA, Blasco J, Delgado J, Sarmiento AM, Nieto JM, Shepherd TJ, Nowell G, Bebianno MJ. 2011. Source and impact of lead contamination on delta-aminolevulinic acid dehydratase activity in several marine bivalve species along the Gulf of Cadiz. *Aquatic Toxicology* **101**:146–154 DOI 10.1016/j.aquatox.2010.09.012.

Dąbrowski M, Bornelöv S, Kruczyk M, Baltzer N, Komorowski J. 2015. 'True'null allele detection in microsatellite loci: a comparison of methods, assessment of difficulties and survey of possible improvements. *Molecular Ecology Resources* **15**:477–488 DOI 10.1111/1755-0998.12326.

Earl DA, Von Holdt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**:359–361 DOI 10.1007/s12686-011-9548-7.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* **14**:2611–2620 DOI 10.1111/j.1365-294X.2005.02553.x.

Excoffier L, Laval G, Schneider S. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics* **1**:47–50.

Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes* **7**:574–578 DOI 10.1111/j.1471-8286.2007.01758.x.

FAO-FIGIS. 2016. *Fisheries global information system. Available at http://www.fao.org/fishery/figis/en.*

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. 2012. Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**:489–496 DOI 10.1016/j.tree.2012.05.012.

Gaspar MB, Ferreira R, Monteiro CC. 1999. Growth and reproductive cycle of *Donax trunculus* L, (Mollusca : Bivalvia) off Faro, southern Portugal. *Fisheries Research* **41**:309–316 DOI 10.1016/S0165-7836(99)00017-X.

Glaubitz JC. 2004. convert: a user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes* **4**:309–310 DOI 10.1111/j.1471-8286.2004.00597.x.

Grewe PM, Feutry P, Hill PL, Gunasekera RM, Schaefer KM, Itano DG, Fuller DW, Foster SD, Davies CR. 2015. Evidence of discrete yellowfin tuna (Thunnus albacares) populations demands rethink of management for this globally important resource. *Scientific Reports* **5**:16916 DOI 10.1038/srep16916.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Leger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit RJ. 2011. Current trends in microsatellite genotyping. *Molecular Ecology Resources* **11**:591–611 DOI 10.1111/j.1755-0998.2011.03014.x.

Hamdani A, Soltani-Mazouni N. 2011. Changes in biochemical composition of the gonads of *Donax trunculus* L. (Mollusca, Bivalvia) from the Gulf of Annaba (Algeria)

in relation to reproductive events and pollution. *Jordan Journal of Biological Sciences* **4**:149–156.

Hargrove JS, Sturmer L, Scarpa J, Austin JD. 2015. Assessment of genetic diversity in wild and aquaculture stocks of mercenaria mercenariain Florida. *Journal of Shellfish Research* **34**:355–365 DOI 10.2983/035.034.0218.

Hauser L, Waples RS, Carvalho GR. 2008. Special issue: advances in marine fish and fisheries genetics. *Fish and Fisheries* **9**:331–332 DOI 10.1111/j.1467-2979.2008.00307.x.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources* **9**:1322–1332 DOI 10.1111/j.1755-0998.2009.02591.x.

Kalinowski ST, Taper ML. 2006. Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics* **7**:991–995 DOI 10.1007/s10592-006-9134-9.

Kalinowski ST, Taper ML, Marshall TC. 2007. Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Molecular Ecology* **16**:1099–1106 DOI 10.1111/j.1365-294X.2007.03089.x.

Kim EM, An HS, Kang JH, An CM, Dong CM, Hong YK, Park JY. 2014. New polymorphic microsatellite markers for the Korean manila clam (*Ruditapes philippinarum*) and their application to wild populations. *Genetics and Molecular Research* **13**:8163–8173 DOI 10.4238/2014.October.7.11.

Kolpakov R, Bana G, Kucherov G. 2003. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**:3672–3678 DOI 10.1093/nar/gkg617.

Launey S, Ledu C, Boudry P, Bonhomme F, Naciri-Graven Y. 2002. Geographic structure in the European flat oyster (*Ostrea edulis* L.) as revealed by microsatellite polymorphism. *Journal of Heredity* **93**:331–351 DOI 10.1093/jhered/93.5.331.

Liu N, Chen L, Wang S, Oh C, Zhao H. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genetics* **6**:S26 DOI 10.1186/1471-2156-6-S1-S26.

Marie AD, Lejeusne C, Karapatsiou E, Cuesta JA, Drake P, Macpherson E, Bernatchez L, Rico C. 2016. Implications for management and conservation of the population genetic structure of the wedge clam *Donax trunculus* across two biogeographic boundaries. *Scientific Reports* **6**:39152 DOI 10.1038/srep39152.

Miller AD, Good RT, Coleman RA, Lancaster ML, Weeks AR. 2013. Microsatellite loci and the complete mitochondrial DNA sequence characterized through next generation sequencing and de novo genome assembly for the critically endangered orange-bellied parrot, Neophema chrysogaster. *Molecular Biology Reports* **40**:35–42 DOI 10.1007/s11033-012-1950-z.

Morin PA, Luikart G, Wayne RK, The SNP workshop group. 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution* **19**:208–216 DOI 10.1016/j.tree.2004.01.009.

Nantón A, Arias-Pérez A, Méndez J, Freire R. 2014. Characterization of nineteen microsatellite markers and development of multiplex PCRs for the wedge clam

Donax trunculus (Mollusca: Bivalvia). *Molecular Biology Reports* **41**:5351–5357 DOI 10.1007/s11033-014-3406-0.

Patarnello T, Volckaert FAMJ, Castilho R. 2007. Pillars of Hercules: is the Atlantic-Mediterranean transition a phylogeographical break? *Molecular Ecology* **16**:4426–4444 DOI 10.1111/j.1365-294X.2007.03333.x.

Perez-Losada M, Nolte MJ, Crandall KA, Shaw PW. 2007. Testing hypotheses of population structuring in the Northeast Atlantic Ocean and Mediterranean Sea using the common cuttlefish *Sepia officinalis*. *Molecular Ecology* **16**:2667–2679 DOI 10.1111/j.1365-294X.2007.03333.x.

Rosenberg NA, Li LM, Ward R, Pritchard JK. 2003. Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics* **73**:1402–1422.

Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**:365–386.

Sá-Pinto A, Branco M.S, Alexandrino PB, Fontaine MC, Baird SJE. 2012. Barriers to gene flow in the marine environment: insights from two common inter-tidal limpet species of the atlantic and mediterranean. *PLOS ONE* **7**:e50330 DOI 10.1371/journal.pone.0050330.

Saavedra C, Bachere E. 2006. Bivalve genomics. *Aquaculture* **256**:1–14 DOI 10.1016/j.aquaculture.2006.02.023.

Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, Wingfield MJ, Wingfield BD. 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. *Biotechniques* **46**:217–223 DOI 10.2144/000113085.

Santure AW, Gratten J, Mossman JA, Sheldon BC, Slate J. 2011. Characterisation of the transcriptome of a wild great tit Parus major population by next generation sequencing. *BMC Genomics* **12**:283 DOI 10.1186/1471-2164-12-283.

Schneider S, Roessli D, Excoffier L. 2000. Arlequin Ver. 2.0: a software for population genetics data analysis. In: *Genetics and biometry laboratory*. Geneva: University of Geneva.

Schoebel CN, Brodbeck S, Buehler D, Cornejo C, Gajurel J, Hartikainen H, Keller D, Leys M, Ricanova S, Segelbacher G, Werth S, Csencsics D. 2013. Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequenc-ing. *Journal of Evolutionary Biology* **26**:1–12 DOI 10.1111/jeb.12044.

Schuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**:233–234 DOI 10.1038/72708.

Sefc KM, Payne RB, Sorenson MD, Fleischer RC. 2003. Microsatellite amplification from museum feather samples: effects of fragment size and template concentration on genotyping errors. *The Auk* **120**:982–989 DOI 10.1642/0004-8038(2003)120[0982:MAFMFS]2.0.CO;2.

Tebble N. 1966. *British bivalve shells: a handbook for identification*. London: Trustees of the British Museum (Natural History).

Tlili S, Metais I, Ayache N, Boussetta H, Mouneyrac C. 2011. Is the reproduction of *Donax trunculus* affected by their sites of origin contrasted by their level of contami-nation?. *Chemosphere* **84**:1362–1370 DOI 10.1016/j.chemosphere.2011.05.009.

Tlili S, Metais I, Boussetta H, Mouneyrac C. 2010. Linking changes at sub-individual and population levels in *Donax trunculus*: assessment of marine stress. *Chemosphere* **81**:692–700 DOI 10.1016/j.chemosphere.2010.07.064.

Van Oosterhout C, Hutchinson WF, Wills PM, Shipley P. 2004. Micro-checker: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**:535–538 DOI 10.1111/j.1471-8286.2004.00684.x.

Wang B, Ekblom R, Castoe TA, Jones EP, Kozma R, Bongcam-Rudloff E, Pollock DD, Hoglund J. 2012a. Transcriptome sequencing of black grouse (Tetrao tetrix) for immune gene discovery and microsatellite development. *Open Biology* **2(4)**:120054 DOI 10.1098/rsob.120054.

Wang R, Li C, Stoeckel J, Moyer G, Liu Z, Peatman E. 2012b. Rapid development of molecular resources for a freshwater mussel, Villosa lienosa (Bivalvia: Unionidae), using an RNA-seq-based approach. *Freshwater Science* **31**:695–708 DOI 10.1899/11-149.1.

Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology* **15**:1419–1439 DOI 10.1111/j.1365-294X.2006.02890.x.

Wasimuddin, Cizkova D, Ribas A, Pialek J, De Bellocq JG, Bryja J. 2012. Development and characterization of multiplex panels of microsatellite markers for Syphacia obvelata, a parasite of the house mouse (Mus musculus), using a high throughput DNA sequencing approach. *Molecular and Biochemical Parasitology* **185**:154–156 DOI 10.1016/j.molbiopara.2012.07.002.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. *Evolution* **38**:1358–1370 DOI 10.2307/2408641.

Whitney JL, Karl SA. 2012. Development of 38 microsatellite loci from the Arceye hawkfish, Paracirrhites arcatus, using next-generation sequencing and cross-amplification in other Cirrhitid species. *Conservation Genetics Resources* **4**:549–553 DOI 10.1007/s12686-011-9589-y.

Wong P. 1988. Mutagenicity of heavy metals. *Bulletin of Environmental Contamination and Toxicology* **40**:597–603 DOI 10.1007/BF01688386.

Yawetz A, Fishelson L, Bresler V, Manelis R. 2010. Comparison of the effects of pollution on the marine bivalve *Donax trunculus* in the vicinity of polluted sites with specimens from a clean reference site (Mediterranean Sea). *Marine Pollution Bulletin* **60**:225–229 DOI 10.1016/j.marpolbul.2009.09.022.

Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, Zhang Q, Liang X, Li Y. 2012. De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). *BMC Genomics* **13**:90 DOI 10.1186/1471-2164-13-90.

Zietkiewicz E, Rafalski A, Labuda D. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain-reaction amplification. *Genomics* **20**:176–183 DOI 10.1006/geno.1994.1151.