Marine Genomics 11 (2013) 11–16

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Efficient isolation of polymorphic microsatellites from high-throughput sequence data based on number of repeats

Sara D. Cardoso [a,*], David Gonçalves [a,b,c], Joana I. Robalo [a], Vitor C. Almada [a], Adelino V.M. Canário [b], Rui F. Oliveira [a,d]

[a] *Unidade de Investigação em Eco-Etologia, Instituto Superior de Psicologia Aplicada — Instituto Universitário, Rua Jardim do Tabaco, 34, 1149-041 Lisboa, Portugal*
[b] *Centro de Ciências do Mar (CCMAR-CIMAR), Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal*
[c] *University of Saint Joseph, Rua de Londres, 16, Macau SAR, China*
[d] *Champalimaud Neuroscience Programme, Instituto Gulbenkian de Ciência, Rua da Quinta Grande, 6, 2780-901 Oeiras, Portugal*

ARTICLE INFO

Article history:
Received 4 February 2013
Received in revised form 17 April 2013
Accepted 18 April 2013

Keywords:
Microsatellite development
In silico polymorphism
Pyrosequencing
Number of repeats
Salaria pavo

ABSTRACT

Transcriptome data are a good resource to develop microsatellites due to their potential in targeting candidate genes. However, developing microsatellites can be a time-consuming enterprise due to the numerous primer pairs to be tested. Therefore, the use of methodologies that make it efficient to identify polymorphic microsatellites is desirable. Here we used a 62,038 contigs transcriptome assembly, obtained from pyrosequencing a peacock blenny (*Salaria pavo*) multi-tissue cDNA library, to mine for microsatellites and *in silico* evaluation of their polymorphism. A total of 4190 microsatellites were identified in 3670 unique unigenes, and from these microsatellites, *in silico* polymorphism was detected in 733. We selected microsatellites based either on their *in silico* polymorphism and annotation results or based only on their number of repeats. Using these two approaches, 28 microsatellites were successfully amplified in twenty-six individuals, and all but 2 were found to be polymorphic, being the first genetic markers for this species. Our results showed that the strategy of selection based on number of repeats is more efficient in obtaining polymorphic microsatellites than the strategy of *in silico* polymorphism (allelic richness was 8.2 ± 3.85 and 4.56 ± 2.45 respectively). This study demonstrates that combining the knowledge of number of repeats with other predictors of variability, for example *in silico* microsatellite polymorphism, improves the rates of polymorphism, yielding microsatellites with higher allelic richness, and decreases the number of monomorphic microsatellites obtained.

1. Introduction

Microsatellites, or simple sequence repeats (SSRs), are among the widely used genetic markers in biology. Because of their high mutation rates, Mendelian inheritance and high reproducibility they can be used for genome mapping and to answer a wide range of biological questions, from the level of the individual (identity, sex, parentage) to the level of the species (phylogenetics, conservation) (Chistiakov et al., 2006).

Until recently, the advantages of microsatellite markers were partially offset by the difficulties inherent in marker development which is required for each species. The most commonly used approaches rely on laborious procedures from preparation and screening of genomic libraries to sequencing of isolated clones and primer design and validation or testing microsatellite primers already developed for closely related species (cross-species microsatellites) (Selkoe and Toonen, 2006; Zane et al., 2002a,b). For species with genome sequences available, bioinformatic tools for *in silico* mining can be used to identify microsatellites and to design primers targeting these regions (Toth et al., 2000). And while sequencing entire genomes of non-model organisms is still out of reach for most researchers, sequencing smaller subsets of the genome or of the transcriptome, presents an attractive alternative. This can now be achieved at affordable prices through next-generation sequencing platforms, that offer the possibility of sequencing long reads (up to 1000 bp), and make possible *de novo* transcriptome assembly without a reference genome (Abdelkrim et al., 2009; Csencsics et al., 2010; Hoffman and Nichols, 2011; Vera et al., 2008; Vogiatzi et al., 2011). Microsatellites developed from expressed sequence tags (ESTs) represent a potential source of type I markers, which are loci situated in transcribed regions associated to genes of known functions (O'Brien, 1991), making them more useful for comparative genetic mapping, linkage and quantitative trait loci association studies (Scaglione et al., 2009). These microsatellites are less polymorphic, due to functional constraints (Serapion et al., 2004), compared to those derived from non-coding genomic sequences, but their flanking regions are expected to be more conserved across closely related species (Slate et al., 2007; Vogiatzi et al., 2011), decreasing the appearance of null alleles.

* Corresponding author at: Fax: +351 218860954.
E-mail address: sdcardoso@ispa.pt (S.D. Cardoso).

1874-7787/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.margen.2013.04.002

Sequence assemblies have been extensively used for finding single nucleotide polymorphisms (SNPs) (Grattapaglia et al., 2011; Louro et al., 2010; Seeb et al., 2011), but much less to find polymorphic microsatellites *in silico*. The first steps in this direction were given by developing PolySSR (Tang et al., 2008), a database that stores information about polymorphic SSRs using sequences from public EST databases (limited to seven organisms), and by Slate et al. (2007) in zebra finch and Shirasawa et al. (2012) in two cultivated peanut lines, which assembled sequences containing only microsatellites and inspected the alignments for contigs comprising sequences with different lengths of the same repeat motif. Recently, Hoffman and Nichols (2011) in Antarctic fur seal manually mined a transcriptome assembly for microsatellite polymorphism and obtained a positive relationship between the inferred number of alleles *in silico* and observed allele number. Furthermore, Neff and Gross (2001) by analyzing 592 AC microsatellite loci from 98 species obtained a positive relationship between microsatellite repeat length and the number of observed alleles across five vertebrate classes (fish, reptiles, amphibians, birds and mammals) and within each class.

We have therefore taken two different approaches for pre-screening microsatellites from next generation sequence data obtained from a normalized multi-tissue cDNA library in order to improve the level of polymorphism detected. In one approach microsatellites were mined for their polymorphism *in silico*, by screening the assembled contigs for variation in the number of repeats, and in the other approach microsatellites were selected based only on their number of repeats (repeat units comprising the microsatellite) which defines the alleles at each loci. Our species of choice was the peacock blenny (*Salaria* pavo) and its choice resulted from the lack of genetic markers for parentage assignment, an essential tool to understand the evolutionary advantage of the different reproductive tactics in this species (Goncalves et al., 2005, 1996). The microsatellites selected using the two approaches were evaluated on individuals from three peacock blenny populations and the efficiency of the two approaches compared.

2. Materials and methods

2.1. Fish samples

Fish used for collecting the tissue samples for the normalized library were euthanized by rapid severance of the spinal cord with a scalpel. The fin samples for the genotyping procedures from individuals at Culatra Island (36°59′N, 7°51′W, Algarve, Portugal) were collected by light anesthetizing the fish with MS222 (Sigma) followed by recovery in a container with abundant aeration. These fish were released into the same place where they had been captured. At Formentera (38°41′N, 1°27′E, Spain) and Borovac (43°9′N, 16°24′E, Croatia), samples were collected from fish killed for other research purposes by immersion in a lethal dosage of MS222. Animal protocols were performed in accordance with accepted veterinary practice under a "Group-1" license issued by the Directorate General for Veterinary of the Ministry for Agriculture, Rural Development and Fisheries of Portugal.

2.2. Sequence data and bioinformatic analysis

The peacock blenny transcriptome was sequenced on a GS-FLX System at Max Planck Institute (Berlin, Germany). Peacock blenny tissue samples were taken from 13 individuals (3 females, 3 Bourgeois males, 3 sneakers and 4 transitional males (transition from sneaker to Bourgeois male) sampled at Culatra Island. Total RNA was separately isolated with TRI Reagent® (Sigma-Aldrich) following standard procedures from 12 tissues (skin, muscle, bone, brain, olfactory epithelium, eyes, heart, kidneys, spleen, intestine, gonads and anal gland). Equal mass of total RNA from these tissues was pooled and used to construct one normalized multi-tissue cDNA library. Sample preparation and analytical processing such as base calling, were performed at Max Planck Institute using the manufacturer's protocol. After vector and quality trimming (≥q20), over 640,000 reads longer than 100 bp were assembled *de novo* using the MIRA3 assembler (Chevreux et al., 2004), in a total of 62,038 transcribed contigs with an average length of 452 bp. The mean coverage of these contigs was of 4.87 ± 17.3 reads (maxCoverage = 1054.5 and minCoverage = 1; mean ± standard deviation) and the mean nucleotide quality score was 35.71 ± 9.36. These contigs putatively correspond to different transcripts and henceforth were designated unigenes. The basic local alignment search tool (BLASTX) algorithm (Gish and States, 1993) was used to query for sequence similarities on all transcripts against the NCBI non-redundant (nr) protein sequence database (e-value <1e−5, release of May 2010) using Blast2GO suite (Gotz et al., 2008). Gene Ontology (GO) terms (Ashburner et al., 2000) were also obtained using Blast2GO with default parameters.

2.3. Microsatellite mining and selection

The identification and localization of perfect microsatellites in the assembled unigenes were accomplished using MSATCOMMANDER version 0.8.2 (Faircloth, 2008). The parameters were set for detection of di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of six repeats, and the option "Design Primers" was also chosen. Tab-delimited files were generated from the searches using this software, and converted to spread-sheet files for subsequent data manipulation as described in Santana et al. (2009). Unigenes harboring microsatellites were manually curated with the aid of Tablet (Milne et al., 2010), which allows graphical visualization of polymorphisms in unigene reads. Information collected with the software included number of reads covering completely the microsatellite (read coverage), completeness of the microsatellite at the 5′ end or 3′ end of the unigene and number of repeat unit variants found for the microsatellite (alleles). Microsatellites also received the annotation of its unigene.

In order to maximize the selection of polymorphic microsatellites for genotyping, two different strategies were pursued. The first strategy required the microsatellite to (i) display polymorphism in the reads forming the unigene; (ii) its unigene to have BLAST hits (e-value <1e−5), and to (iii) have at least a pair of primers. In the second strategy, the microsatellites were only selected based on the number of repeats and the existence of a pair of primers, irrespective of BLAST hits and *in silico* polymorphism.

2.4. PCR amplification and fragment analysis

A set of 63 microsatellites developed from *S. pavo* unigenes were selected for amplification test using one peacock blenny DNA sample. PCR amplifications were set up in 50 μl volume composed of ~100 ng DNA, 0.25 pmol of each primer (MWG), 1.5 mM $MgCl_2$, 120 μM of each dNTP, 5× Green GoTaq® Flexi Buffer 1×, and 1.5 u *Taq* DNA polymerase (Promega). PCRs were performed in a thermal cycler (Stratagene RoboCycler® Gradient 96) programmed as: 3 min at 94 °C for initial denaturation, followed by 35 cycles of 94 °C for 1 min, primer specific annealing temperature for 1 min, 72 °C for 45 s, and a final extension at 72 °C for 7 min. The success of the PCRs was determined by running 10 μL of each PCR product and co-running 6 μl of a mixture of DNA loading dye with a 50 bp DNA ladder (GeneRuler™ 50 bp DNA Ladder − 0.5 μg/μl; Fermentas) on a 1× Tris-acetate-EDTA buffer and 2% agarose gel stained with GelRed 3×, visualized under UV light and photographically documented.

For peacock blenny's loci that seemed to amplify well in the agarose gels, the respective forward primers were 5′ fluorescently labeled with 6-carboxyfluorescein (6-FAM) or with hexachloro-fluorescein (HEX) dyes. A total of 26 adult peacock blenny individuals sampled from Culatra (20 samples), Formentera (3 samples) and Borovac Islands (3 samples) were employed for polymorphism assessment. Individuals

Table 1
Locus primer sequences and microsatellite polymorphism characteristics. Microsatellites were identified *in silico* and developed for 28 loci from *Salaria pavo* unigenes, applied in twenty individuals from Culatra population. For each locus, the repeated motif and GenBank accession number are given.

Locus GenBank no.	Repeat motif	Primer sequence (5′-3′)	T_a (°C)	Culatra			
				Size (bp)	*k*	Ho	He
Spavo01[c] JQ619676	$(GT)_6$	F-CACCTCGAACAGTTGGCTTC GCTGCATTAGCCCAGATCC	58	387–397	3	0.30	0.27
Spavo02[c] JQ619677	$(GA)_8C(GA)_4$	F-CCCTGGCTGATGTGACTCC ACTCTCCAGGTGTAAGGCAC	61	250–258	5	0.25	0.28
Spavo03[c] JQ619678	$(AC)_6$-$(GT)_6$	F-GCACAAGTCGGCACTCAAG GCCAAGCCGAGTATGAAGC	60	229–237	4	0.50	0.58
Spavo04[c] JQ619679	$(AC)_6$	F-CCCACGTCTGTTCAGTTGAC GGAGTTGGCACATTCCGTG	58	259–266	3	0.40	0.45
Spavo05[c] JQ619680	$(AC)_9$	F-ATCAGCGCGAAACACATCG ACTGCACTCAAGTCAAAGCC	56	185–189	3	0.55	0.52
Spavo06[c] JQ619681	$(TG)_8$	F-GCTGGTCGATGGCAGAATG GCGTCGGAAATACCGTTCC	58	295–297	2	0.05	0.05
Spavo07[c] JQ619682	$(AC)_4G(AC)_{10}$	F-CACGACAGCTGGTCTCAAC GGGCTCACCAGTCCCATTC	58	331–337	3	0.35	0.42
Spavo08[c] JQ619683	$(CA)_9$	F-CGTGACTTCATGGCAAGGG TGTGTGGAAACGATATGTGC	58	221–235	7	0.75	0.79
Spavo09[c] JQ619684	$(AC)_8$	F-CGCTAAAAGGAGGCAACATC ACAGCGACGAGCTTCATCTT	61	196–200	3	0.10	0.10
Spavo10[c] JQ619685	$(AC)_9$	F-AGAGTAGGGGTCCGTCGATT TGGCAGTGAGAAAGTGCAAG	61	137–141	3	0.10	0.19
Spavo11[c] JQ619686	$(CT)_9$	F-GGTAGCGAGAGACGCAGAAG GGTAGACCAGCGGTCTGAAG	62	232–234	2	0.60	0.43
Spavo12[c] JQ619687	$(AC)_7G(AC)_{12}$	F-GCTGTAAAACTGCGTGGACA GGACGTGAACCTGGAGAAGA	61	179–204	5	0.60	0.56
Spavo13[c] JQ619688	$(AC)_{10}$	F-CCTCGCAGCAGTAACTCAGA TCCGTCTATGGAGGCTAACG	61[b]	136–146	3	0.60	0.59
Spavo14 JQ619689	$(AC)_{17}$	F-GGGGATCGAAATGTTTCACA CCACATGGAACCAACTTCCT	59	246–260	5	0.40	0.75**
Spavo15[c] JQ619690	$(AC)_6T(AC)_4$	F-CATGGCCTATCTGTTCCGC AGACCAACATCCCAGTCGC	58	240	1	–	–
Spavo16[c] JQ619691	$(AC)_5T(AC)_5$	F-GTTCAGGATGACCCGGTGG TGTGTATGAGTTCCTGCCC	56	168	1	–	–
Spavo17[c] JQ619692	$(TC)_7$	F-TGTCAAGCTCACAGCGAC ATGGCACCCATGCTTCAGG	56[a]	216	1	–	–
Spavo18[c] JQ619693	$(GA)_7$	F-CCATGACCAACTACGACGAG GGAGCTTAGGTCGCTCACC	62	175	1	–	–
Spavo19[c] JQ619694	$(CA)_3T(CA)_7$	F-ACCTTCCAGCCTACGAGAGC TGTGTCAGGAGTAGGCAGACC	62	170	1	–	–
Spavo20 JQ619695	$(AGC)_{10}$	F-TGCTCGGCTCTACGGTTC CCCTCACAGAGTTCACGGG	60	209–239	8	0.60	0.50
Spavo21 JQ619696	$(AATG)_{15}$	F-TGTGTTGGTTTGAGACGGC CCTCAAAGACATTGGATGCG	60	298–330	8	0.85	0.79
Spavo22 JQ619697	$(ATCC)_{14}$	H-GGCAGAAGGAAACCTGGAC GGCCCTTGAAACTCCACTCT	61	139–187	9	0.85	0.77
Spavo23 JQ619698	$(CATT)_8$	H-CGACCCATTTCGGTTACAAG GAACGAGTAACGTGATGCTGA	61	245–269	6	0.75	0.72
Spavo24 JQ619699	$(CTGT)_9$	F-GCTCCAACAGAGATAAAACGCTCT TCACTGTAGGAACACGGGAAT	62	170–182	4	0.30	0.27
Spavo25 JQ619700	$(CTGT)_{10}$	H-GAGTGAGCCGGAGTGTTCTG GGCTAAACTGTGGCTGCCTA	62	232–244	3	0.30	0.55*
Spavo26 JQ619701	$(GTTT)_9$	H-CACGTTGCCAATTCCAGTAG GAAGACGACAACCACTCTCAG	59	212–220	3	0.40	0.38
Spavo27 JQ619702	$(AAAC)_{13}$	F-GAGCTGGCGTTTCCCAAATA ACGGCGTAGTGAGCATGTTG	59	169–232	12	0.80	0.76
Spavo28 JQ619703	$(CTATT)_{10}$	H-GCAGAGTGACAATAAAGGACGA CCACAAGGCTCAGTTTGACA	59	292–328	7	0.75	0.68

Ta (°C) — annealing temperature; Ho — observed heterozygosity; He — expected heterozygosity; *k* — number of alleles; "F-" or "H-" at the 5′ end of the primer indicates FAM- or HEX-labeled primer; Hardy–Weinberg expectation deviations.

[a] Mg = 1.0 mM.

[b] Mg = 1.75 mM.

[c] Strategy 1.

* $P < 0.05$.

** $P < 0.001$.

from the populations of Formentera and Borovac were used in order to verify if the microsatellite primers worked in all DNA samples and not only on those of Culatra where the primers were designed. DNA was extracted from the dorsal fin using Extract-N-Amp™ Tissue PCR Kit (Sigma-Aldrich). Microsatellite amplification reactions were performed in 25 μl volume containing ~100 ng DNA, 0.25 pmol of each primer (MWG), 1.5 mM $MgCl_2$ (for exceptions see Table 1 and S2), 60 μM of each dNTP, 5× Green GoTaq® Flexi Buffer 1×, and 0.75 u *Taq* DNA polymerase (Promega). PCR thermal program was run as previously described (for annealing temperatures see Table 1).

DNA fragments were separated on a commercial ABI 3730XL DNA analyzer and sized by co-running a GeneScan HD400 (Applied Biosystems) size standard. DNA fragments were scored manually with the aid of GeneMarker® version 1.95 (SoftGenetics). For each working loci the type of microsatellite and the number of repeat variants were confirmed by commercial sequencing.

2.5. Microsatellite loci evaluation

Tests for Hardy–Weinberg equilibrium and genotypic linkage disequilibrium were performed using GENEPOP version 4.0.11 (Rousset, 2008) with the default setting (10,000 dememorization steps, 100 batches, and 5000 iterations per batch). Genetic diversity estimates, including expected (He) and observed (Ho) heterozygosities, were also calculated using GENEPOP. The test for the presence of null alleles was conducted using MICRO-CHECKER version 2.2.3 (Van Oosterhout et al., 2004).

In order to evaluate whether the two strategies used in this work potentially influence or not the polymorphism obtained, a Generalized Linear Model (GLM) was constructed within R (R. Development Core Team, 2011). The number of different alleles observed among all individuals genotyped from the three populations for each microsatellite locus was modeled as a response variable using a Poisson error structure. The microsatellite number of repeats (minimum number of repeat units observed *in silico*) and the number of alleles observed *in silico* were used as predictor variables and fitted as continuous variables. Each variable was dropped from the model and the change in deviance between full and reduced model was distributed as χ^2 with degrees of freedom equal to the difference in degrees of freedom between the models with or without the variable in question. The residual deviance (difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed) was used to perform a goodness of fit for the overall model. Allelic richness (mean number of observed alleles per locus) between strategies was examined by using a two-tailed Student's *t* test or a Welch Two Sample *t* test, after testing for variance homogeneity.

3. Results and discussion

3.1. Microsatellite mining and in silico assessment of polymorphism

A complete search of the peacock blenny assembly of 62,038 unigenes for 5 types of microsatellites with a minimum number of repeats of 6 units identified 4190 microsatellite loci in 3670 unique unigenes, representing 5.9% of the sequenced transcriptome. Dinucleotide repeats accounted for 79.0% of all microsatellite loci, followed by 14.5% for trinucleotide, 4.4% for tetranucleotide, 1.4% for pentanucleotide and 0.7% for hexanucleotide repeats, values in the range observed in other fish species (Ju et al., 2005). It was not possible to determine the *in silico* polymorphism in 1428 microsatellites either because they were incomplete (28.2%) or because of single read coverage for the microsatellite region (71.8%). Polymorphic microsatellites were 733, of which 727 were dinucleotides and only 6 were trinucleotides. Two dinucleotide microsatellites had the maximum of 4 alleles each, while the majority of the microsatellites had two alleles (91.5%).

3.2. Microsatellite application and evaluation

Applying the first strategy criteria, 108 microsatellites were available comprising only dinucleotide repeats. From these, 33 microsatellites were selected based on the quality of the microsatellite flanking regions for primer design (Table S1). These microsatellites had a mean read coverage of 21.72 ± 60.59 reads, of which two unigenes accounted for 98 and 346 reads, and a mean number of repeats of 7.24 ± 1.94 units. When following the second strategy, 1340 microsatellites were available, of which 30 microsatellites were selected, comprising ten dinucleotides, six trinucleotides, eleven tetranucleotides, one pentanucleotide and two hexanucleotides. They had a mean read coverage of 3.43 ± 2.75 reads and a mean number of repeats of 12.03 ± 3.26 units. With the exception of locus *Spavo14*, none of these microsatellites appeared to be polymorphic *in silico* or originated BLAST hits (Table S1).

When the 63 primer pairs selected by the two strategies were PCR checked on one peacock blenny DNA sample and reaction conditions optimized, 38.1% ($n = 24$) led to different or multiple PCR products and 61.9% ($n = 39$) resulted in PCR products of the expected size. Amplification of DNA samples from three different peacock blenny populations of the Islands of Culatra, Formentera and Borovac, showed that three microsatellites had mononucleotide variation (variation of only one nucleotide between different alleles) and six microsatellites had multiple peaks or lacked a clear peak in the target region and were discarded. Fragment variation in two other microsatellites was not in accordance to the corresponding type of microsatellite, possibly because of insertion–deletion (indel) polymorphisms (Vali et al., 2008) combined with the polymorphism of the microsatellite. Twenty-eight microsatellites, 18 from using the first strategy and 10 from using the second strategy, were successfully characterized in all individuals used from the three locations (Table 1 and S2), and their sequences submitted to GenBank with the accession numbers: JQ619676–JQ619703.

In 20 individuals genotyped from the population of Culatra, from which the cDNA library was originated, all but five dinucleotide microsatellite loci were found to be polymorphic (*Spavo15*–*Spavo19*; Table 1). The number of alleles ranged from 2 to 12 (4.83 ± 2.59) per locus and the observed and expected heterozygosities ranged from 0.05 to 0.85 and from 0.05 to 0.79, respectively. The mean number of alleles per locus and the expected heterozygosity were highest in microsatellite loci isolated using the second strategy (6.5 ± 2.88 and 0.62 ± 0.18) compared to the first strategy (3.54 ± 1.39 and 0.4 ± 0.22). Variation in allele number (Welch's unpaired *t* test, $t_{(12.233)} = 3.0$, $P = 0.01$) and expected heterozygosity (unpaired *t* test, $t_{(21)} = 2.52$, $P = 0.02$) were statistically significant. In this population, only *Spavo14* ($P < 0.001$) and *Spavo25* ($P < 0.05$) loci departed from Hardy–Weinberg (HW) equilibrium expectations, most probably because of heterozygote deficit (homozygote excess). The deviation of HW expectation in the first loci is significant possibly due to the presence of null alleles or stuttering leading to scoring errors. The presence of SNPs in *Spavo14* primer binding sites cannot be excluded considering the low depth read coverage (2 reads) of the unigene. No other loci were detected with null alleles. Two of the possible pairwise comparisons between loci were in linkage disequilibrium ($P < 0.01$: *Spavo05*–*Spavo08* and *Spavo08*–*Spavo25*). For the Formentera and Borovac samples, all but eight and twelve microsatellite loci, respectively, were found to be polymorphic (Table S2) in the 3 individuals genotyped from each population. The number of alleles ranged from 2 to 5 (mean 2.5 ± 0.83 and 2.69 ± 1.01 respectively) and expected heterozygosities from 0.33 to 0.93.

All but five microsatellites (*Spavo15*–*Spavo19*) were polymorphic in the Culatra population. Since the apparently monomorphic microsatellites were isolated using the first strategy and were therefore expected to be polymorphic, it is possible by increasing the number of individuals the polymorphism could be detected. In the other two populations, only two microsatellites (*Spavo15* and *Spavo18*) were not confirmed as polymorphic.

3.3. Relationship of in silico variability and microsatellite number of repeats with PCR polymorphism

The strategies described here were developed in order to achieve higher rates of polymorphism. The novelty in the approach relies on a pre-screening of microsatellites, based on their polymorphism *in silico* or based on their number of repeats. The relatively low success rate of nearly 45% of functional microsatellites obtained is in the range reported in other studies developing microsatellites from unigenes (mean = 65%; range 45%–76% (Csencsics et al., 2010; Hoffman and Nichols, 2011; Kim et al., 2008; Li et al., 2009; Vogiatzi et al., 2011)). The proportion of polymorphic microsatellites was also comparable to those reported by Li et al. (2009) in oyster (15/29 microsatellite loci), Hoffman and Nichols (2011) in Antarctic fur seal (23/38 microsatellite loci) and Csencsics et al. (2010) in dwarf bulrush (17/22 microsatellite loci).

The success of the rate of microsatellite PCR amplification was higher using the first strategy (54.5%) compared to the second (33.3%) (Table 2). The difference may have resulted from the lower read coverage of the flanking regions of the microsatellites isolated using the second strategy, affecting the base call confidence of these regions where the primers were designed. However, the second strategy was more effective in yielding more highly polymorphic microsatellites (8.2 ± 3.85 alleles per locus), considering all different alleles observed in the three populations, than the first strategy (4.56 ± 2.45 alleles per locus) (unpaired t test, $t_{(24)} = 2.96$, $P = 0.0069$). DNA slippage may increase in proportion to the number of repeats so that microsatellite loci with more repeats generally show higher mutation rates, which could explain these results (Petit et al., 2005). Furthermore, Li et al. (2004) reported that microsatellites present in protein-coding regions (strategy 1), could lead to gain or loss of gene function via frameshift mutations, which could explain the lower allele richness found in these loci. However, long stretches of repeats may also accumulate imperfections that persist because they favor slippage reduction and consequently improve microsatellite stability (Bhargava and Fuentes, 2010; Zhu et al., 2000), which is important for microsatellites harbored in genes. Examples of this are *Spavo02*, *Spavo07*, *Spavo12*, *Spavo15*, *Spavo16* and *Spavo19* loci, where one nucleotide was inserted or substituted interrupting the stretch of perfect repeats. Only in *Spavo12* locus the smaller stretch of repeats was confirmed as polymorphic.

To our knowledge only a recent study on Antarctic fur seal used an approach similar to our first strategy. Hoffman and Nichols (2011) selected microsatellites either on the basis of GO codes or high variability *in silico*, and obtained a positive relationship between the number of alleles *in silico* and the observed allele number. However, a lower number of polymorphic microsatellites were obtained (61% of the microsatellite loci compared to the 93% in the present study). A Generalized Linear Model (GLM) to evaluate the impact of the two strategies on the rates of polymorphism indicates that between the two predictor variables considered, number of alleles observed *in silico* and microsatellite number of repeats, only the latter was retained as a significant predictor variable (estimate = 0.75, $\chi^2_{(1)} = 6.67$, $P = 0.0098$) in the reduced model. One explanation could be that the microsatellites were not as polymorphic *in silico* (1 to 3 alleles; Table S1) as with Hoffman and Nichols (2011), where the microsatellites had between 1 and 6 alleles. Some variation may have been lost during normalization of the cDNA library, although to some extent this may have been compensated by a larger number of unigenes sequenced as a result. No conclusions can be drawn in relation to which type of microsatellite is more prone to be polymorphic. Although tetranucleotides appear to be candidates, this may be because they were the type of microsatellites successfully applied (7/11 microsatellite loci) in the second strategy.

Table 2
Summary of the microsatellite results obtained for each strategy. Microsatellite results are based on the 26 peacock blenny individuals genotyped from Culatra, Formentera and Borovac populations.

	Strategy 1	Strategy 2
Selected for application	33	30
Successfully genotyped	18	10
Polymorphic	16	10
Repeat length *in silico*[a]	7.17 ± 1.69	11.8 ± 4.16
Allelic richness[b]	4.56 ± 2.45	8.2 ± 3.85

Strategy 1 — *in silico* polymorphism with GO terms; strategy 2 — number of repeats.

[a] Welch's unpaired t test, $t_{(10.678)} = 3.37$, $P = 0.0065$.

[b] Unpaired t test, $t_{(24)} = 2.96$, $P = 0.0069$.

4. Conclusions

Using next-generation sequencing data offers a simple and relatively fast way to microsatellite screening and isolation for application. Combining the information of the microsatellite number of repeats with polymorphism *in silico* may help improve the number of polymorphic microsatellites and their allelic richness, important for species with low genetic variability, and at the same time, develop type I markers by using the annotation results.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2013.04.002.

Acknowledgments

The study was funded by the FCT (RG-LVT-331-2352 and PTDC/MAR/69749/2006) and by the FDCT (012/2012/A1).

References

Abdelkrim, J., Robertson, B.C., Stanton, J.A.L., Gemmell, N.J., 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46, 185–190.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29.

Bhargava, A., Fuentes, F.F., 2010. Mutational dynamics of microsatellites. Mol. Biotechnol. 44, 250–266.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Muller, W.E.G., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14, 1147–1159.

Chistiakov, D.A., Hellemans, B., Volckaert, F.A.M., 2006. Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture 255, 1–29.

Csencsics, D., Brodbeck, S., Holderegger, R., 2010. Cost-effective, species-specific microsatellite development for the endangered Dwarf Bulrush (*Typha minima*) using next-generation sequencing technology. J. Hered. 101, 789–793.

Faircloth, B.C., 2008. MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol. Ecol. Resour. 8, 92–94.

Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat. Genet. 3, 266–272.

Goncalves, E.J., Almada, V.C., Oliveira, R.F., Santos, A.J., 1996. Female mimicry as a mating tactic in males of the blenniid fish *Salaria pavo*. J. Mar. Biol. Assoc. U. K. 76, 529–538.

Goncalves, D., Matos, R., Fagundes, T., Oliveira, R., 2005. Bourgeois males of the peacock blenny, *Salaria pavo*, discriminate female mimics from females? Ethology 111, 559–572.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36, 3420–3435.

Grattapaglia, D., Silva, O.B., Kirst, M., de Lima, B.M., Faria, D.A., Pappas, G.J., 2011. High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biol. 11, 65.

Hoffman, J.I., Nichols, H.J., 2011. A novel approach for mining polymorphic microsatellite markers in silico. PLoS One 6, e23283.

Ju, Z., Wells, M.C., Martinez, A., Hazlewood, L., Walter, R.B., 2005. An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, *Fundulus*, and *Xiphophorus*. In Silico Biol. 5, 439–463.

Kim, K.S., Ratcliffe, S.T., French, B.W., Liu, L., Sappington, T.W., 2008. Utility of EST-Derived SSRs as population genetics markers in a beetle. J. Hered. 99, 112–124.

Li, Y.C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.

Li, Q., Liu, S.K., Kong, L.F., 2009. Microsatellites within genes and ESTs of the Pacific oyster *Crassostrea gigas* and their transferability in five other Crassostrea species. Electron. J. Biotechnol. 12, 3.

Louro, B., Passos, A.L.S., Souche, E.L., Tsigenopoulos, C., Beck, A., Lagnel, J., Bonhomme, F., Cancela, L., Cerda, J., Clark, M.S., et al., 2010. Gilthead sea bream (*Sparus auratus*) and European sea bass (*Dicentrarchus labrax*) expressed sequence tags: characterization, tissue-specific expression and gene markers. Mar. Genomics 3, 179–191.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., Marshall, D., 2010. Tablet-next generation sequence assembly visualization. Bioinformatics 26, 401–402.

Neff, B.D., Gross, M.R., 2001. Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. Evolution 55, 1717–1733.

O'Brien, S.J., 1991. Molecular genome mapping lessons and prospects. Curr. Opin. Genet. Dev. 1, 105–111.

Petit, R.J., Deguilloux, M.F., Chat, J., Grivet, D., Garnier-Gere, P., Vendramin, G.G., 2005. Standardizing for microsatellite length in comparisons of genetic diversity. Mol. Ecol. 14, 885–890.

R. Development Core Team, 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rousset, F., 2008. GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. Mol. Ecol. Resour. 8, 103–106.

Santana, Q.C., Coetzee, M.P.A., Steenkamp, E.T., Mlonyeni, O.X., Hammond, G.N.A., Wingfield, M.J., Wingfield, B.D., 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46, 217–223.

Scaglione, D., Acquadro, A., Portis, E., Taylor, C.A., Lanteri, S., Knapp, S.J., 2009. Ontology and diversity of transcript-associated microsatellites mined from a globe artichoke EST database. BMC Genomics 10, 454.

Seeb, J.E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., Seeb, L.W., 2011. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. Mol. Ecol. Resour. 11, 1–8.

Selkoe, K.A., Toonen, R.J., 2006. Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecol. Lett. 9, 615–629.

Serapion, J., Kucuktas, H., Feng, J.A., Liu, Z.J., 2004. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar. Biotechnol. 6, 364–377.

Shirasawa, K., Koilkonda, P., Aoki, K., Hirakawa, H., Tabata, S., Watanabe, M., Hasegawa, M., Kiyoshima, H., Suzuki, S., Kuwata, C., et al., 2012. In silico polymorphism analysis for the development of simple sequence repeat and transposon markers and construction of linkage map in cultivated peanut. BMC Plant Biol. 12, 80.

Slate, J., Hale, M.C., Birkhead, T.R., 2007. Simple sequence repeats in zebra finch (*Taeniopygia guttata*) expressed sequence tags: a new resource for evolutionary genetic studies of passerines. BMC Genomics 8, 52.

Tang, J.F., Baldwin, S.J., Jacobs, J.M.E., van der Linden, C.G., Voorrips, R.E., Leunissen, J.A.M., van Eck, H., Vosman, B., 2008. Large-scale identification of polymorphic microsatellites using an in silico approach. BMC Bioinformatics 9, 374.

Toth, G., Gaspari, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10, 967–981.

Vali, U., Brandstrom, M., Johansson, M., Ellegren, H., 2008. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. BMC Genet. 9, 8.

van Oosterhout, C., Hutchinson, W.F., Wills, D.P.M., Shipley, P., 2004. MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Mol. Ecol. Notes 4, 535–538.

Vera, J.C., Wheat, C.W., Fescemyer, H.W., Frilander, M.J., Crawford, D.L., Hanski, I., Marden, J.H., 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol. Ecol. 17, 1636–1647.

Vogiatzi, E., Lagnel, J., Pakaki, V., Louro, B., Canario, A.V.M., Reinhardt, R., Kotoulas, G., Magoulas, A., Tsigenopoulos, C.S., 2011. In silico mining and characterization of simple sequence repeats from gilthead sea bream (*Sparus aurata*) expressed sequence tags (EST-SSRs); PCR amplification, polymorphism evaluation and multiplexing and cross-species assays. Mar. Genomics 4, 83–91.

Zane, L., Bargelloni, L., Patarnello, T., 2002a. Strategies for microsatellite isolation: a review. Mol. Ecol. 11, 1–16.

Zane, L., Patarnello, T., Ludwig, A., Fontana, F., Congiu, L., 2002b. Isolation and characterization of microsatellites in the Adriatic sturgeon (*Acipenser naccarii*). Mol. Ecol. Notes 2, 586–588.

Zhu, Y., Strassmann, J.E., Queller, D.C., 2000. Insertions, substitutions, and the origin of microsatellites. Genet. Res. 76, 227–236.

Data accessibility

DNA sequences — GenBank accessions JQ619676–JQ619703.