Development and characterization of 21 novel microsatellite markers for the Australian smelt *Retropinna semoni* (Weber)

Md. Rakeb-Ul Islam; Daniel J. Schmidt; Jane M. Hughes

Australian Rivers Institute, Griffith University, Nathan, 4111, QLD, Australia.

Corresponding author contact information:

Email: r.islam@griffithuni.edu.au

Phone: +61 737357361

Keywords:

Freshwater fish; Retropinnidae; microsatellites; cryptic species; Next-Generation Sequencing

Introduction

The Australian smelt is a common freshwater fish species which is widely distributed throughout coastal and inland drainages of southeastern Australia. They form large shoals in the mid to upper water column and inhabit deep slow flowing pools as well as shallow fast flowing riffle-runs. Adults live in a diverse array of environments including slow flowing low land rivers, small coastal streams, river impoundments, upland river and streams, dune lakes and brackish water estuaries (Pusey et al., 2004). Australian smelt are formally recognised as two described taxa *R. semoni* (Weber), and *R. tasmanica* McCulloch, but recent genetic analyses have recognized a complex of five or more cryptic species throughout their geographic range based on allozymes, microsatellites and mitochondrial DNA data (Hammer et al., 2007; Hughes et al., 2014; Schmidt et al., 2016). Migratory life history patterns vary greatly among these genetic groups, ranging from freshwater residents to facultative amphidromy and estuarine-resident populations (Crook et al., 2008; Hughes et al., 2014; Woods et al., 2010). Information on population structure and life history variation in two northern lineages geographically designated 'CEQ' and 'SEQ' has received relatively little attention despite speculation these lineages may represent cryptic species (Hammer et al., 2007; Page and Hughes, 2010; Schmidt et al., 2016). Microsatellites are powerful DNA markers to measure genetic variation within and between populations (Muneer et al., 2009). They have been used extensively for studies of stock identification and population differentiation (Schmidt et al., 2014; Wright and Bentzen, 1994). Microsatellite variation has been examined within and between *R. semoni* lineages in southeastern Australia, demonstrating a range of population structures consistent with the diversity in migratory life histories (Hughes et al., 2014; Woods et al., 2010). In this study, twenty one new polymorphic microsatellite markers were developed and characterized for the study of northern lineages of Australian smelt. These markers will be used for analysis of population structure, life history variation and species delimitation.

Materials and methods

Randomly-sheared genomic DNA libraries were prepared for Illumina paired-end sequencing using two smelt specimens representing informal lineages designated 'SEQ' and 'CEQ' (Hammer et al., 2007; Schmidt et al., 2016). Genomic DNA was isolated from tissue voucher GUB_433 (= lineage SEQ, Twin Bridges Reserve, Brisbane River, -27.430457 152.639357) and voucher GUM_433 (= lineage CEQ, Conondale Bridge, Mary River, -26.727511 152.713604). Genomic DNA was extracted from fin tissue using the DNeasy Blood and

Tissue kit (Qiagen) following the manufacturer's directions. DNA was sheared to an approximate mean length of 400 bp using the M220 Focused-ultrasonicator (Covaris) and an Illumina MiSeq-compatible sequencing library was prepared. A double index two-step library preparation was used, with all steps performed in the presence of solid phase reversible immobilization (SPRI) beads (iTru protocol: Travis Glenn, pers. Comm.; Faircloth and Glenn, 2012). Paired-end sequencing was performed on the Illumina MiSeq Benchtop Sequencer at Griffith University DNA Sequencing Facility, using a 600 cycle MiSeq reagent kit v3. The two libraries each generated $\sim 1.2 \times 10^7$ paired-end reads. Overlapping paired reads were merged with Geneious v9.1.5 software (Kearse et al., 2012), using the Flash v1.2.9 plugin (min overlap 20, max overlap 200; Magoc and Salzberg, 2011). Then 1.6×10^6 merged reads in size range 200 – 580 bp were sampled for each library. Reads containing microsatellites from the genomic libraries were selected using the QDD pipeline version 3.1 (Meglécz et al., 2014). Selected loci contained ≥ 10 uninterrupted repeats and had minimal target region complexity indicated by the QDD design category 'A'. Complete mitochondrial genomes of both voucher specimens were previously assembled from a sample of the raw paired-end data used here (Schmidt et al., 2016).

A total of 48 primer pairs (24 pairs from each library) were screened using a sample of eight smelt specimens to check for successful amplification and polymorphism. Four unique 20-mer oligonucleotide tails were added to 5' end of the forward primers as described by Real et al. (2009). All subsequent microsatellite screening was carried out in 10 μ l PCR reaction mixture consisting of 0.5 μ l of genomic DNA, 0.2 mM reverse primer, 0.05 mM tailed forward primer, 0.2 mM tailed fluorescent tag (either FAM, VIC, NED or PET, Applied Biosystems), 1 $\times$ PCR buffer (Astral Scientific) and 0.02 units of *taq* polymerase (Astral Scientific). The following basic thermocycler settings for the polymerase chain reaction (PCR) were performed: initial denaturation at 94 $^{\circ}$ C for 4 min, followed by 35 cycles at 94 $^{\circ}$ C for 1 min, 57 $^{\circ}$ C for 30 s, 72 $^{\circ}$ C for 1 min and a final extension at 72 $^{\circ}$ C for 7 min.

Fluorescently labelled amplified PCR products were pooled and added to 10 μ l of Hi-DiTM formamide with 0.1 μ l of GeneScanTM 500 LIZ size standard and run on to an ABI PRISM 3130 Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions. Genemapper version 3.1 software (Applied Biosystems) was used to edit and score genotypes.

The best loci were selected on the basis of amplification success rate, the presence or absence of stutter peaks, peak intensity, the polymorphism of the loci, the number of alleles and

heterozygosity. Selected loci were further characterised using a population sample ($n=32$) of *R. semoni* collected from the Brisbane river ($27^{\circ}30'16.048''\text{S}$, $152^{\circ}55'49.991''\text{E}$), Queensland, Australia. The number of alleles (N_A), allelic range, observed heterozygosity (H_O) and expected heterozygosity (H_E), Hardy-Weinberg equilibrium (HWE) test were calculated using Arlequin version 3.5 (Excoffier and Lischer, 2010). Linkage disequilibrium (LD) between the loci was evaluated using GENETIX version 4.05 (Belkhir et al., 1996). The potential for null alleles, large allele dropout and stuttering to interfere with scoring accuracy was tested for each microsatellite locus in each sample using MICRO-CHECKER version 2.2.3 (Van Oosterhout et al., 2004). The polymorphic information content (PIC) for each locus was calculated using CERVUS version 3.0.3 (Kalinowski et al., 2007).

Results

Among 48 pairs of primers designed, 21 primer pairs were found polymorphic and selected for characterization in 32 Australian smelt individuals. Sequences of the selected loci were deposited in GenBank under accession numbers KY124334 - KY124354 (Table 1). Locus names beginning with letters “BS” were isolated from voucher GUB_433 (= lineage SEQ; Table 1). Locus names beginning with letters “MS” were isolated from voucher GUM_433 (= lineage CEQ; Table 1). Of the 21 primer pairs only 3 loci derived from the CEQ lineage were polymorphic (MS6, MS19 and MS24; Table 1). The remaining 18 polymorphic loci were all derived from the SEQ lineage (Table 1). Characteristics of these loci are described in Table 1. The number of alleles (N_A) per locus ranged from 3 to 20 (mean = 9.29). The observed (H_O) and expected (H_E) heterozygosity ranged from 0.111 to 0.969 (mean = 0.756) and 0.358 to 0.945 (mean = 0.734) respectively. The polymorphic information content (PIC) of all loci ranged from 0.334 to 0.926 (mean = 0.691) indicating that these markers are highly informative (Botstein et al., 1980). No locus pair showed significant linkage disequilibrium. However, deviation from Hardy-Weinberg equilibrium (HWE) was found in three loci (BS1, BS15 and MS19) after Bonferroni correction, and two loci (BS1 and BS2) showed potential presence of null alleles.

Discussion

Genetic diversity of the 21 polymorphic loci in the current study was similar to five loci previously developed for *R. semoni* by Hillyer et al. (2006), showing a moderate genetic diversity (Hillyer loci: mean $N_A = 10.8$, $H_O = 0.65$ and $H_E = 0.71$ vs new loci: mean $N_A = 9.29$, $H_O = 0.756$ and $H_E = 0.734$). Microsatellite loci of Hillyer et al. (2006) were developed

for *R. semoni* lineages from the southern extent of the species' range and have proven useful for assessing population structure at various spatial scales in southern populations (Hughes et al., 2014; Woods et al., 2010). However only two loci (sm-26, sm-80) described by Hillyer et al. (2006) amplified consistently in northern *R. semoni* lineages 'SEQ' and 'SEC' and these showed limited polymorphism (R. Islam pers. obs.). Allelic diversity of the 21 new loci presented here indicates that they will be valuable for population genetics research on northern lineages. Three loci that showed significant departure from HWE should be used with caution in further population genetic analysis. Potential presence of null alleles was found in two loci which might be the result of mutations in the primer annealing sites (Callen et al., 1993). In conclusion, most of the microsatellite loci which were developed here could be used to investigate the genetic diversity and population structure in Australian smelt fish.

Acknowledgements

We would like to thank Leo Lee for help in collecting samples and Kathryn Real and Jemma Somerville for assistance in the laboratory. Griffith University Animal Ethics Committee provided approval for this study in accordance with Australian Animal Ethics Approval No. ENV/23/14/AEC.

References

- Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N., & Bonhomme, F. (1996). GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire génome, populations, interactions, CNRS UMR. 1996-2004.
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics*, 32(3), 314-331.
- Callen, D. F., Thompson, A. D., Shen, Y., Phillips, H. A., Richards, R. I., Mulley, J. C., & Sutherland, G. R. (1993). Incidence and origin of null alleles in the (AC)_n microsatellite markers. *American Journal of Human Genetics*, 52(5), 922-927.
- Crook, D. A., Macdonald, J. I., & Raadik, T. A. (2008). Evidence of diadromous movements in a coastal population of southern smelts (Retropinninae : Retropinna) from Victoria, Australia. *Marine and Freshwater Research*, 59(7), 638-646.
- Excoffier, L., & Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3), 564-567.
- Faircloth, B. C., & Glenn, T. C. (2012). Not All Sequence Tags Are Created Equal: Designing and Validating Sequence Identification Tags Robust to Indels. *Plos One*, 7(8).
- Hammer, M. P., Adams, M., Unmack, P. J., & Walker, K. F. (2007). A rethink on *Retropinna*: conservation implications of new taxa and significant genetic sub-structure in Australian smelts (Pisces: Retropinnidae). *Marine and Freshwater Research*, 58(4), 327-341.

- Hillyer, M. J., Woods, R. J., & Hughes, J. M. (2006). Isolation and characterization of polymorphic microsatellite loci in the Australian smelt, *Retropinna semoni*. *Molecular Ecology Notes*, 6(1), 224-226.
- Hughes, J. M., Schmidt, D. J., Macdonald, J. I., Huey, J. A., & Crook, D. A. (2014). Low interbasin connectivity in a facultatively diadromous fish: evidence from genetics and otolith chemistry. *Molecular Ecology*, 23(5), 1000-1013.
- Kalinowski, S. T., Taper, M. L., & Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Molecular Ecology*, 16(5), 1099-1106.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12), 1647-1649.
- Magoc, T., & Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21), 2957-2963.
- Meglecz, E., Pech, N., Gilles, A., Dubut, V., Hingamp, P., Trilles, A., Grenier, R., & Martin, J. F. (2014). QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. *Molecular Ecology Resources*, 14(6), 1302-1313.
- Muneer, P. M. A., Gopalakrishnan, A., Musammilu, K. K., Mohindra, V., Lal, K. K., Basheer, V. S., & Lakra, W. S. (2009). Genetic variation and population structure of endemic yellow catfish, *Horabagrus brachysoma* (Bagridae) among three populations of Western Ghat region using RAPD and microsatellite markers. *Molecular Biology Reports*, 36(7), 1779-1791.
- Page, T. J., & Hughes, J. M. (2010). Comparing the performance of multiple mitochondrial genes in the analysis of Australian freshwater fishes. *Journal of Fish Biology*, 77(9), 2093-2122.
- Pusey, B. J., Kennard, M., & Arthington, A. (2004). *Freshwater fishes of north-eastern Australia*. Collingwood, VIC: CSIRO Publishing.
- Real, K. M., Schmidt, D. J., & Hughes, J. M. (2009). *Mogurnda adspersa* microsatellite markers: multiplexing and multi-tailed primer tagging. *Conservation Genetics Resources*, 1(1), 411-414.
- Schmidt, D. J., Crook, D. A., Macdonald, J. I., Huey, J. A., Zampatti, B. P., Chilcott, S., Raadik, T. A., & Hughes, J. M. (2014). Migration history and stock structure of two putatively diadromous teleost fishes, as determined by genetic and otolith chemistry analyses. *Freshwater Science*, 33(1), 193-206.
- Schmidt, D. J., Islam, M. R.-U., & Hughes, J. M. (2016). Complete mitogenomes for two lineages of the Australian smelt, *Retropinna semoni* (Osmeriformes: Retropinnidae). *Mitochondrial DNA Part B*, 1(1), 615-616.
- 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. *Molecular Ecology Notes*, 4(3), 535-538.
- Woods, R. J., Macdonald, J. I., Crook, D. A., Schmidt, D. J., & Hughes, J. M. (2010). Contemporary and historical patterns of connectivity among populations of an inland river fish species inferred from genetics and otolith chemistry. *Canadian Journal of Fisheries and Aquatic Sciences*, 67(7), 1098-1115.
- Wright, J. M., & Bentzen, P. (1994). Microsatellites - genetic-markers for the future. *Reviews in Fish Biology and Fisheries*, 4(3), 384-388.

Table 1

Characterization of 21 microsatellite loci in Australian smelt (*R. semoni*), including GenBank accession number, primer sequence, repeat motif, annealing temperature (T_m), number of alleles observed (N_A), allele size range, observed (H_O) and expected heterozygosity (H_E), polymorphic information content (PIC) and probability of Hardy-Weinberg equilibrium (P_{HWE})

Locus	GenBank Accession No.	Primer Sequence (5'-3') Forward (F) and Reverse (R)	Repeat motif	T _m (°C)	N _A	Alleles range (bp)	H _O	H _E	PIC	P _{HWE}
BS1	KY124334	F: TTCCAGCTTTGGGAAATGTT R: TCTGGTGAGAACAAGGGTCC	(AATG) ₁₅	57	4	118-150	0.111	0.358	0.334	0.000*
BS2	KY124335	F: CTGTCTCCTGCTTGGGTGAT R: CAATCCACAGTTTGGTGGTG	(ACGC) ₁₃	57	19	118-207	0.688	0.917	0.895	0.003
BS3	KY124336	F: TCTTTAAAGCAGCTGAAGGAGTT R: AAGGATGAATGAAATGATCTGACA	(ACAG) ₁₃	57	14	140-204	0.935	0.901	0.876	0.163
BS4	KY124337	F: CCAGTAAAGTTAAATACCAAAGACG R: TGAGCCTTAATTTTCATTGTCTGAG	(ACAG) ₁₄	57	12	122-173	0.906	0.861	0.831	0.985
BS5	KY124338	F: TGGGTAGAAGGATGAGACATCA R: GAGCATCAGGGTAGGTTGGA	(ACTG) ₁₆	57	14	109-168	0.931	0.901	0.876	0.910
BS6	KY124339	F: AACTCTGACGTGGGCTCTGT R: TGTCAGAACAGCTTGGATGG	(AGCC) ₁₁	57	8	126-157	0.906	0.845	0.812	0.445
BS8	KY124340	F: ATGTGGTTCTATGGGCGTTC R: AACCGTTCAGTCCAACCAAC	(AGCC) ₁₀	57	8	220-248	0.355	0.421	0.401	0.155
BS9	KY124341	F: TGCTAATCGCTCAGGCTAAA R: ATCCTCCACACCCAGAGTCA	(AATC) ₁₁	57	4	159-174	0.625	0.476	0.412	0.250
BS11	KY124342	F: AATTTCCCAGATTTGGATGAA R: TCAGTTGATTGACTGTGACTTTGA	(AGAT) ₁₄	57	12	126-179	0.875	0.872	0.843	0.079
BS14	KY124343	F: CGGAGATAAGGAAGCTGTGG R: AGGTTTGCTGCCTGTGAAAC	(AC) ₁₆	57	20	125-186	0.969	0.945	0.926	0.689
BS15	KY124344	F: TTCTGTGTAACCTAAGGCTATGAGAAT R: GCTACGAGTCCCCTCTACGCA	(AC) ₁₅	57	5	128-148	0.938	0.685	0.617	0.000*
BS16	KY124345	F: GTGGAAGTGCTGAAGAGGCT R: GGGTCAATTTGTCTATCCCA	(AC) ₁₈	57	12	148-176	0.969	0.863	0.833	0.393
BS18	KY124346	F: TCGAAAGACAGACAGAAGAACC R: GATGGATGGGTGGAATTCAG	(AC) ₁₅	57	3	151-159	0.655	0.671	0.585	0.822
BS20	KY124347	F: CTACGAAAGGGTCGTCCTGA R: CCTGGTGACTCCAGAACCCTT	(AC) ₁₇	57	11	122-160	0.867	0.859	0.827	0.663
BS21	KY124348	F: CCTTCCCTAATGAGACATAGTCA R: TATCATATGAAACCCAGCAAGC	(AC) ₁₉	57	8	157-175	0.704	0.778	0.73	0.075
BS22	KY124349	F: ACACATGTAAGCATTGAGAACTTTG R: GGAAGAGCCTGACAGGACAA	(AC) ₂₁	57	11	142-175	0.778	0.869	0.837	0.111
BS23	KY124350	F: TCTAATCGAAGTTGGCGACC R: GCGGTTGGCTAATCTCTGTG	(AC) ₁₅	57	7	190-206	0.625	0.735	0.674	0.783
BS24	KY124351	F: CAGCCCGCACTATGAAGAAT R: GCTTCACCTTGAGCCTGTCT	(AC) ₁₈	57	8	153-167	0.793	0.738	0.698	0.662
MS19	KY124352	F: CCAGTTGCCTGTGAGAAGTG R: GCCCTCATCCAGAGACTCAC	(AC) ₂₃	57	4	149-163	0.821	0.541	0.438	0.000*
MS24	KY124353	F: GAAGGGAAATGGGTGAGGAT R: GCTTCACCTTGAGCCTGTCT	(AC) ₂₆	57	6	248-279	0.615	0.553	0.516	0.902
MS6	KY124354	F: TGAACGAGTGAATAAGTGAATGA R: TGCACACTACCACACTTCTGC	(ACTC) ₂₀	57	5	78-104	0.813	0.632	0.548	0.175

*Loci deviating from Hardy-Weinberg equilibrium (HWE) at P < 0.001 after Bonferroni correction.