GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Development and validation of single nucleotide polymorphism markers in *Odontobutis potamophila* from transcriptomic sequencing

H.W. Zhang, S.W. Yin, L.J. Zhang, X.Y. Hou, Y.Y. Wang and G.S. Zhang

College of Life Sciences, Nanjing Normal University,
Jiangsu Key Laboratory for Biodiversity and Biotechnology, Nanjing,
Jiangsu Co-Innovation Center for Marine Bio-Industry Technology,
Lian Yungang, China

Corresponding author: S. Yin
E-mail: yinshaowu@hotmail.com

Genet. Mol. Res. 14 (1): 2080-2085 (2015)
Received April 3, 2014
Accepted July 14, 2014
Published March 20, 2015
DOI http://dx.doi.org/10.4238/2015.March.20.18

ABSTRACT. Transcriptome sequencing technology has been applied in the development and discovery of single nucleotide polymorphism (SNP) markers in fish. In this study, a panel of 120 expressed sequence tag (EST)-derived SNPs was selected by several selection filters from the resultant EST library of *Odontobutis potamophila* using Illumina Sequencing. In total, 37 SNPs from 120 putative SNPs were considered as the true SNPs using Sanger sequencing. For each SNP locus of 30 individuals of one wild

population of *O. potamophila* that was successfully calculated, the number of alleles per locus was 2 with an observed heterozygosity of 0.0000-0.9000 and an expected heterozygosity of 0.1000-0.5263. A total of 33 loci conformed to Hardy-Weinberg equilibrium (HWE), and 4 loci deviated from HWE after Bonferroni correction. These 33 SNP markers will benefit the studies of population genetic structure, population evolution analysis, and construction of a high-density linkage map of *O. potamophila*.

Key words: *Odontobutis potamophila*; Expressed sequence tag; Illumina Sequencing; Single nucleotide polymorphism validation

INTRODUCTION

The freshwater sleeper, *Odontobutis potamophila* (Günther, 1861), is an important and commercially valuable fish that mainly relies on wild resources; it is widely distributed in the middle and the lower regions of the Yangtze River, Qiantang River, and Minjiang River systems in China. This fish was listed as an endangered animal by the International Union for the Conservation of Nature in 2012 (Huckstorf, 2012). Because of the rapid development of the aquaculture industry and the sharp expansion of the farming scale in recent years, improved selective breeding efforts need to be developed urgently. However, basic research related to the molecular-assisted breeding of the fish is still relatively lacking, and molecular markers, which are efficient tools for evaluating genetic resources and facilitating molecular maker-assisted breeding, are limited.

Transcriptome sequencing (RNA-seq) technology includes a series of processes such as the enrichment of single-stranded mRNA from total RNA, generation of double-stranded cDNA, and execution of high-throughput sequencing analysis. For species without a reference genome, *de novo* transcriptome sequencing can be performed, suggesting that *de novo* assembly of sequencing data can generate long segments to obtain a single gene sequence set (unigene) for the species. The developed transcriptome single nucleotide polymorphism (SNP) markers by 454 pyrosequencing technology have been applied in the breeding of blunt snout bream (*Megalobrama amblycephala*), and visible resources of fish transcriptome data have great potential in the field of breeding (Gao et al., 2012). The applications of developed and validated expressed sequence tag (EST)-derived SNP markers have been reported in aquatic species such as European hake (Milano et al., 2011), salmonids (Seeb et al., 2011; Lemay et al., 2013), Atlantic herring (*Clupea harengus*) (Helyar et al., 2012), rainbow trout (Boussaha et al., 2012; Salem et al., 2012), common carp (Xu et al., 2012), and turbot (*Scophthalmus maximus*) (Vera et al., 2013). However, very little is known about the development and validation of EST-derived SNP markers in *O. potamophila.*

In this study, we used Illumina Sequencing of cDNA pools of *O. potamophila* to develop EST-derived SNP markers. This study will lay the foundation for molecular marker-assisted selection of *O. potamophila*.

MATERIAL AND METHODS

RNA was extracted from the muscle, liver, and kidney of *O. potamophila*, a cDNA library was constructed, and an EST library was established by Illumina Sequencing. In total, 40,905 ESTs were obtained and assembled using the bowtie2 software. All heterozygous loci in the transcriptome data were determined by the samtools software. The quality of predicted SNPs was improved through a pipeline of several stringent filters, which was similar to the method implemented by Boussaha et al. (2012). Several selection filters were applied to select a panel of 120 EST-derived SNPs for validation.

Totally 30 *O. potamophila* individuals were genotyped for each of 120 SNPs using Sanger sequencing. These samples were collected from the Dangtu wild population in Anhui Province, China. Genomic DNA was extracted from fins using the Easy Pure Marine Animal Genomic DNA Kit (TRANS, Beijing TransGen Biotech Co., Ltd. Beijing, China) according to the manufacturer protocol. Briefly, primers were designed using the primer premier 5 software, polymerase chain reaction (PCR) was performed, and PCR products were subjected to Sanger sequencing on an ABI 3730 Genetic Analyzer to confirm their SNP genotype. Sequence comparisons were conducted using the Seqman software. The genetic analyses were carried out using POPGENE version 1.32 (Yeh et al., 1997).

RESULTS AND DISCUSSION

A total of 51,485 variation sites from transcriptome data of *O. potamophila* were obtained. After applying several selective filters, 431 SNP loci were scanned from 51,485 variable loci and 376 sequences. We selected 64 sequences among these 376 sequences, and each sequence contained 1-3 putative SNPs. One hundred twenty SNPs were used to genotype independent samples of *O. potamophila*. We successfully amplified 32 of 64 sequences for genotype validation, and the amplified fragment lengths varied from 150 to 550 bp. Finally, 37 SNPs of the successfully genotyped SNPs were considered as true SNPs. Seeb et al. (2011) demonstrated a detailed SNP discovery and validation pipeline that incorporates 454 pyrosequencing, high-resolution melt analysis, and 5' nuclease genotyping in duplicated salmonids; 5' nuclease genotyping was validated by 37 SNPs in 202 putative SNPs. The validation rate was 18.32% (37/202). Xu et al. (2012) performed Transcriptome sequencing of four strains of common carp with Solexa HiSeq2000 platform. Validation of selected SNPs with Sanger sequencing revealed that 48% percent of SNPs (12 of 25) were tested to be true SNPs. In our study, the 30.83% (37/120) validation rate is remarkably higher than that of the previous study of duplicated salmonids, but less than validation rate of common carp (48%). For each polymorphic locus of the fish population that was successfully calculated, the number of alleles per locus was 2, the observed heterozygosity was 0.0000-0.9000, and the expected heterozygosity was 0.1000-0.5263. A total of 33 loci conformed to Hardy-Weinberg equilibrium (HWE), and 4 loci (comp12768_c0_seq1_153, comp14820_c0_seq1_314, comp19282_c1_seq1_482, and comp776_c1_seq1_112) deviated from HWE after Bonferroni correction ($P < 0.00001$, Table 1).

Table 1. Single nucleotide polymorphism (SNP) markers in *Odontobutis potamophila* validated by Sanger sequencing.

Locus ID	Primers	Type of SNP	H_O	H_E	P
comp11244_c1_seq1_192	F: GCTCACCTAAAGGGATGACAT R: CTCACATTTAGCGGAAGAAGAT	G/T	0.4	0.5263	0.423711
comp11244_c1_seq1_402	F: GCTCACCTAAAGGGATGACAT R: CTCACATTTAGCGGAAGAAGAT	C/T	0.3	0.3947	0.40471
comp11552_c1_seq1_336	F: CGCATAAAGCAATGTAAAATAGC R: TAAACCCCAATGTAGCGTTGT	C/A	0.3	0.2684	0.65591
comp11552_c1_seq1_501	F: CGCATAAAGCAATGTAAAATAGC R: TAAACCCCAATGTAGCGTTGT	C/T	0.3	0.5211	0.156828
comp12768_c0_seq1_153	F: ACCAGTCCAAGGGAATCGTC R: GCTGAAGTTGCTCTTGGACG	T/C	0	0.1895	0.000013*
comp12768_c0_seq1_304	F: ACCAGTCCAAGGGAATCGTC R: GCTGAAGTTGCTCTTGGACG	A/G	0.1	0.3947	0.009534
comp14120_c0_seq1_245	F: CATTCAGGACAGTGTTTGCTTG R: ATTTGGGTAGGCATTTTAGGC	C/T	0.5	0.3947	0.354539
comp14120_c0_seq1_393	F: CATTCAGGACAGTGTTTGCTTG R: ATTTGGGTAGGCATTTTAGGC	G/T	0.7	0.4789	0.11956
comp14820_c0_seq1_240	F: CAGCACTCATGGTGAGCACTG R: AGCAGCAGGCTCGTTAGTCG	C/T	0.3	0.4789	0.207625
comp14820_c0_seq1_314	F: CAGCACTCATGGTGAGCACTG R: AGCAGCAGGCTCGTTAGTCG	C/T	0	0.1895	0.000013*
comp14820_c0_seq1_515	F: CAGCACTCATGGTGAGCACTG R: AGCAGCAGGCTCGTTAGTCG	T/A	0	0.3368	0.000347
comp14910_c0_seq1_232	F: AACTCCTGTGTGTTTGGATGG R: AAGTGAAATTAGTCAAAAGCAACAT	T/G	0.4	0.3368	0.502335
comp14910_c0_seq1_322	F: AACTCCTGTGTGTTTGGATGG R: AAGTGAAATTAGTCAAAAGCAACAT	A/G	0.4	0.3368	0.502335
comp16377_c0_seq1_364	F: CTTTCACAAACGCAGGCATC R: CGAAAATGGAGTGAAGCGAC	G/A	0.4	0.5053	0.485263
comp16377_c0_seq1_472	F: CTTTCACAAACGCAGGCATC R: CGAAAATGGAGTGAAGCGAC	G/A	0.3	0.3947	0.40471
comp16822_c0_seq1_412	F: AGACTGTGCCTTTTCTGATGG R: CGGAGGGATACAGTGCTACAG	G/A	0.6	0.5053	0.529954
comp18555_c0_seq2_186	F: GCTTTGAATTGAGCTAGGGC R: TAATGTCCATGATTTCTACCTCAA	T/C	0.9	0.5211	0.015219
comp18555_c0_seq2_458	F: GCTTTGAATTGAGCTAGGGC R: TAATGTCCATGATTTCTACCTCAA	T/C	0.4	0.4421	0.745333
comp19131_c2_seq1_323	F: AGACAGCACAGGTTTGTTGATG R: AAAGCTGAATGCTGGAGCAG	G/A	0.4	0.5053	0.485263
comp19131_c2_seq1_395	F: AGACAGCACAGGTTTGTTGATG R: AAAGCTGAATGCTGGAGCAG	A/G	0.1	0.1	1
comp19244_c1_seq1_352	F: CATTTCATCTGTTATGGGGTTC R: GTGGTCCTACAAGCTCATTATCA	A/G	0.2	0.1895	0.808365
comp19282_c1_seq1_482	F: CCGCTGATTATAGGCTGAGA R: TGACGGTGGTTCAACAAAAT	G/A	0	0.1895	0.000013*
comp19783_c0_seq1_501	F: AGTCTTTTCCAGGACGCTCT R: GAAAGACATTTGTGCTCCATAA	T/C	0.3	0.2684	0.65591
comp19978_c3_seq1_194	F: CACCAGCCATGATTAGCAGC R: GGCTTCAGTGTTATGGTTTGTCA	C/A	0.1	0.2684	0.017485
comp20141_c3_seq1_260	F: TTTACAACTCGGGAGCAGTG R: CATGTAAAGGTCCACAATCAATC	G/A	0.5	0.5211	0.892738
comp20141_c3_seq1_491	F: TTTACAACTCGGGAGCAGTG R: CATGTAAAGGTCCACAATCAATC	A/T	0.1	0.1	1
comp21284_c2_seq3_244	F: CTTGATAGCACCTGAAAATGTTG R: GGGATGGATTTGAAGTGATGTT	C/T	0.2	0.1895	0.808365
comp21284_c2_seq3_397	F: CTTGATAGCACCTGAAAATGTTG R: GGGATGGATTTGAAGTGATGTT	T/C	0.5	0.3947	0.354539
comp23148_c7_seq2_245	F: CATGAAATAAGCCTCAGTCCA R: AGTAACGGCTCAAAAATAACG	G/T	0.2	0.5053	0.04299
comp23307_c4_seq1_166	F: TTTTCAAATATGGAGCCACTTC R: GGTGACATTTTCTGAACTGCTT	G/C	0.4	0.5053	0.485263

Continued on next page

Table 1. Continued.

Locus ID	Primers	Type of SNP	H_O	H_E	P
comp23307_c4_seq1_255	F: TTTTCAAATATGGAGCCACTTC R: GGTGACATTTTCTGAACTGCTT	A/T	0.2	0.5263	0.038766
comp23827_c7_seq1_257	F: TCTTGTTGTAAACTGCTTTTCAA R: GCCATTACCACCGTTGTTAC	A/G	0.1	0.1	1
comp23827_c7_seq1_486	F: TCTTGTTGTAAACTGCTTTTCAA R: GCCATTACCACCGTTGTTAC	G/T	0.1	0.1	1
comp545_c1_seq1_134	F: CAGGTGTGCAAGCTAACAATCA R: AAAGTAATTCTGTCTCTGAAGCAGC	A/C	0.2	0.1895	0.808365
comp59_c0_seq1_126	F: ATGCTGTTTTGTCAGTTTTGCC R: GTTGCTCCAGGCTCAGTGCT	A/C	0.5	0.5211	0.892738
comp776_c1_seq1_112	F: CGAGTCCTGAAAAGTGGGTG R: GAGCAGTGAGGCTGTAAGGC	A/G	0	0.1895	0.000013*
comp917_c1_seq1_219	F: GCTGTCTGCCGTCTGTGAAG R: ACATTCGGTTCCCCACAAAG	G/C	0.2	0.1895	0.808365

H_O = observed heterozygosity; H_E = expected heterozygosity; P = exact P value for Hardy-Weinberg equilibrium test (*statistical significance after Bonferroni correction).

To our knowledge, this is the first report to validate 33 EST-derived SNPs in *O. potamophila*. The SNP markers that were developed in this study will be useful in population genetic studies and evolutionary studies, and they will also be used as important markers in the breeding and resource management of *O. potamophila*.

ACKNOWLEDGMENTS

Research supported by the National Spark Program Project (#2013GA690166), the Province R&D Program of Jiangsu (#BE2013441), Jiangsu Province Six Talent Peaks of High Level Talents Project (#2012-NY-032), and Science and Technology Achievement Transformation Fund Project of Nanjing Normal University (#2013-02).

REFERENCES

Boussaha M, Guyomard R, Cabau C, Esquerré D, et al. (2012). Development and characterisation of an expressed sequence tags (EST)-derived single nucleotide polymorphisms (SNPs) resource in rainbow trout. *BMC Genomics* 13: 238.

Gao ZX, Luo W, Liu H, Zeng C, et al. (2012). Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). *PLoS One* 7: e42637.

Helyar SJ, Limborg MT, Bekkevold D, Babbucci M, et al. (2012). SNP discovery using next generation transcriptomic sequencing in Atlantic herring (*Clupea harengus*). *PLoS One* 7: e42089.

Huckstorf V (2012). *Odontobutis potamophilus*. In: IUCN 2012. IUCN Red List of Threatened Species.

Lemay MA, Donnelly DJ and Russello MA (2013). Transcriptome-wide comparison of sequence variation in divergent ecotypes of kokanee salmon. *BMC Genomics* 14: 308.

Milano I, Babbucci M, Panitz F, Ogden R, et al. (2011). Novel tools for conservation genomics: comparing two high-throughput approaches for SNP discovery in the transcriptome of the European hake. *PLoS One* 6: e28008.

Salem M, Vallejo RL, Leeds TD, Palti Y, et al. (2012). RNA-seq identifies SNP markers for growth traits in rainbow trout. *PLoS One* 7: e36264.

Seeb JE, Pascal CE, Grau ED, Seeb LW, et al. (2011). Transcriptome sequencing and high-resolution melt analysis advance single nucleotide polymorphism discovery in duplicated salmonids. *Mol. Ecol. Resour*. 11: 335-348.

Vera M, Alvarez-Dios JA, Fernandez C, Bouza C, et al. (2013). Development and validation of single nucleotide polymorphisms (SNPs) markers from two transcriptome 454-runs of turbot (*Scophthalmus maximus*) using high-throughput genotyping. *Int. J. Mol. Sci*. 14: 5694-5711.

Xu J, Ji PF, Zhao ZX, Zhang Y, et al. (2012). Genome-wide SNP discovery from transcriptome of four common carp strains. *PLoS One* 7: e48140.

Yeh FC, Yang R and Boylet T (1997). POPGENE version 1.32: software Microsoft Window-based freeware for population genetic analysis. University of Alberta, Edmonton.

