动 物 学 研 究 2008，Aug. 29(4)：373-378 CN 53-1040/Q ISSN 0254-5853
Zoological Research doi:10.3724/SP.J.1141.2008.04373

Searching for Protein-coding Genes Using Microsatellites in Common Carp by Comparing to Zebrafish EST Database

CHANG Yu-mei[1, 2], KUANG You-yi[1], LIANG Li-qun[1],
LU Cui-yun[1], HE Jian-guo[2], SUN Xiao-wen[1,*]

(1. *Heilongjiang River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Harbin* 150070, *China*;
2. *College of Life Sciences, Zhongshan University, Guangzhou* 510275, *China*)

Abstract: In this study, an *in silico* approach was utilized to identify homologies existing between common carp microsatellite sequences and GenBank database using Blastn and Blastx searches. About 875 microsatellite sites with flanking sequences over 50bp of common carp were first compared to the zebrafish EST database. The results showed that 121 homologies were found using Blastn. Subsequent Blastx searches confirmed 94 sites recorded in the protein database. Except for 33 hypothetical proteins and three unknown proteins, seven out of 58 characterized proteins have been mapped to two linkage maps. In addition, two polymorphic STS markers were developed using matched zebrafish EST sequences by PCR-SSCP method, of which one marker HLJZe33 was mapped successfully. This study was a pilot for comparative studies between common carp and zebrafish, and the results demonstrated that more genetic and genomic resources of zebrafish can be used for the genome research of common carp.

Key words: *Cyprinus carpio*; *In silico*; Microsatellites; Blast

借助斑马鱼 EST 数据库从鲤鱼微卫星序列中寻找蛋白编码基因

常玉梅[1,2]，匡友谊[1]，梁利群[1]，鲁翠云[1]，何建国[2]，孙效文[1,*]

(1. 中国水产科学研究院黑龙江水产研究所, 黑龙江 哈尔滨 150070；2. 中山大学 生命科学学院, 广东 广州 510275)

摘要：应用 *in silico* 的方法，利用 Blastn 和 Blastx 搜索引擎，将鲤鱼微卫星序列与 GenBank 数据库进行同源序列比对。利用 Blastn，将侧翼序列长度>50bp 的 875 个鲤鱼微卫星序列与斑马鱼的 EST 数据库首先进行比对，结果找到了 121 个同源序列。随后采用 Blastx 搜索蛋白质数据库，有 94 个微卫星位点存在同源蛋白。除了 33 个假定和 3 个未知蛋白外，剩余的 58 个微卫星位点被成功地进行了功能注释，而且其中的 7 个位点已经定位在了鲤鱼连锁图谱上。另外，通过 PCR-SSCP 的方法，将两个与鲤鱼微卫星侧翼序列相匹配的斑马鱼 EST 序列开发成鲤鱼的 STS 标记，并将其中的一个标记 HLJZe33 定位到鲤鱼连锁图谱上。以上研究结果表明，通过比较基因组研究，模式生物斑马鱼的很多遗传和基因组资源都可以被利用到鲤鱼的基因组研究中。

关键词：鲤鱼；*in silico*；微卫星标记；Blast
中图分类号：Q959.468；Q75；S917 **文献标识码**：A **文章编号**：0254-5853-(2008)04-0373-06

The common carp (*Cyprinus carpio* L.) is an important freshwater aquaculture species, widely distributed from Southeast Asia to Europe and the Mediterranean region (David et al, 2001). Though it is important for aquaculture, its genomic information is limited and lags behind those of tilapia (*Oreochromis* spp.), rainbow trout (*Oncorhynchus mykiss*) and channel catfish (*Ictalarus punctatus*), whose genome projects were launched by the United States Department of Agriculture (USDA) in 1997. Presently, about 32103 ESTs and 1871 protein sequences are available in GenBank (http://www.ncbi.nlm.nih.gov). The first

Received date：2008-04-22；Accepted date：2008-06-25
Fundation item: Supported by National Basic Research Program of China (2004CB117405)
收稿日期：2008-04-22；接受日期：2008-06-25
*通讯作者(Corresponding author) , E-mail : sunxw2002 @163.com
第一作者简介：博士生，E-mail: ymchang2002@sohu.com

genetic linkage map of common carp published contained only 268 DNA markers (Sun & Liang, 2004). Additionally, some microsatellite markers have been reported for this species (Crooijmans et al, 1997; Aliah et al, 1999; Wei et al, 2001). Most microsatellite markers are generally considered as type Π markers due to their distribution in non-coding regions of the genome, some of these markers are in protein coding regions and have the connection with certain functional genes. The latter has been confirmed using two methods. One is that microsatellite markers were identified by screening cDNA libraries or EST sequences (Kantety et al, 2002; Yue & Orban, 2004; Ju et al, 2005). The other is that the flanking sequences of microsatellite repeats matched genes using a BLAST search engine. This *in silico* data mining approach has already been used to identify genes in mouse (*Mus musculus*), cattle (*Bos taurus*), pigs (*Sus scrofa*), chicken (*Gallus gallus*) and tilapia (Cnaani et al, 2002; Herron et al, 1998; Farber & Medrano, 2003). In this study, we utilized a similar approach to search for genes using microsatellite sequences of common carp. This study will help annotations of these microsatellites and anchor more genes to linkage map. More importantly, it will pave the way for comparative genomic research between common carp and zebrafish (*Zebra danio*) in the future.

1 Materials and methods

1.1 Data collections

Approximately 1000 microsatellite sequences for common carp were mostly developed in our laboratory, and a small part of them were downloaded from GenBank. All sequences were saved in FASTA format. Zebrafish db ESTs were retrieved in FASTA format from the GenBank database using the Entrez nucleotide query webpage (http://www.ncbi.nlm.nih.gov/sites/entrez?) at the National Center for Biotechnology Information (NCBI).

1.2 Batched blast

All repeats were masked and deleted from the target sequences using software programmed in the lab. Flanked sequences (>50bp) were queried against expressed sequence tags (EST) database of zebrafish using blastn searches locally. We used batched blast programmed by ourselves for individual searches of each locus. The detailed parameters for blastn were set as at least 11 (word size) consecutive nucleotide alignments and minimum sum expect value (E) of less than e-10, penalizing five scores for mismatches and three for gap open and extension. Matched zebrafish ESTs were selected manually and queried against protein NR database using Blastx searches. The Blastx search was limited to at least three consecutive amino acids alignments and E of less than 0.001.

1.3 Primers designed and PCR-SSCP

Sequences annotated above were chosen for primer designing using Primer3 (Rozen & Skaletsky 2000). A first set of 10 primer pairs were synthesized and used for PCR amplification. Six primer pairs amplified target bands in 36 F_2 hybrids (*Cyprinus carpio* var. *wuyuanensis* ×*C. pellegrini pellegrini*). Amplification reactions were performed on GeneAmp 9700 (Applied Biosystems, Inc.) in a 15μL volume containing 10-50ng of template, 10mmol/L Tris-HCl (pH 8.3), 50 mmol/L KCl, 200μmol/L each dNTP, 0.1% Triton X-100, 0.1% NP-40, 0.01% gelatin, 0.4mmol/L of each primer, and 1 U of *Taq* polymerase (Shanghai Sangon, China). PCR cycle parameters for all primers were: initial denaturation for 3min at 94℃ followed by 24 cycles of 30s at 93℃, 30s at 50-55℃, 30s at 72℃, and a final extension step of 10min at 72℃. The mixture of 1μL PCR products and 5μL loading buffer (95% Formamide，0.2% Bromophenol blue，0.2% Xylene cyanol FF) was denaturated for 10 min at 98℃ on a thermcycler, then chilled quickly on ice for at least 10 min. The 6 μL mixture was subjected to gel electrophoresis on an 8% non-denaturing polyacrylamide gel. After electrophoresis, gel was silver-stained and scanned.

2 Results

2.1 Genes found by Blastn and Blastx homologous searches

After eliminating short flanked sequences (<50bp), the rest of the 875 carp microsatellites with flanking sequences over 50bp were compared to zebrafish EST database using blastn searches locally, a total of 121 homologous sequences were found. The blastx search confirmed 94 sequences recorded in GenBank protein database on 5th December, 2007. Fifty eight are characterized proteins, 33 hypothetical proteins and three unknown proteins. Hypothetical proteins and uncharacterized proteins were excluded in the following analysis (Tab. 1). These results were compared to those of two different carp genetic linkage maps constructed by our research team (unpublished). One is a recombinant inbred line (RIL) map, the other is a self-crossed F_2 map. Tweleve microsatellite sites denoted by HLJ in Tab. 1 were used in genetic linkage analysis. Of three sites,

Tab. 1 Genes found by blastn and blastx homologous comparisons between common carp microsatellite sequences and the GenBank database

Carp clone No.	Zebra fish EST (Accession No.)	Identity Blastn (%)	Matched gene (Accession No.)	Similarity Blastx(%)	E-value	Species[a]
LIC_41	**DN857925**	61/66 (92)	SORCS receptor (**XP_421750**)	52/110 (47)	2e-18	GG,MM,TN
HLJZe1	**BM861022**	128/137 (93)	Serine/threonine-protein kinase (**Q9NYY3**)	134/216 (62)	4e-50	HS,RN,MM
LIC_77	**EB929248**	50/53 (94)	HnRNP methyltransferase (**XP_001366105**)	78/89 (87)	4e-57	TN,MM,HS
LIC_65	**AL928408**	243/263 (92)	Down syndrome cell adhesion molecule (**XP_693739**)	112/183 (61)	4e-44	DR,TN,MM
LID_53	**EB885308**	40/41 (97)	Capping protein (**NP_956229**)	110/113 (97)	7e-59	DR,MM
HLJ174	**EB777153**	93/102 (91)	General transcription factor (**XP_001345566**)	63/63 (100)	1e-30	DR,MM,HS
HLJ321	**CF997715**	114/122 (93)	Gamma-dystrobrevin (**ABF55376**)	233/235 (99)	3e-139	DR,MM,RN
HLJ721	**CN319735**	111/115 (96)	Na+/K+ ATPase (**AAK33032**)	204/204 (100)	1e-116	DR,MM,HS
LIF_46	**DN897423**	39/40 (97)	Pre-B cell enhancing factor (**XP_686052**)	85/91 (93)	2e-18	DR,TN
LIF_72	**EB905145**	60/65 (92)	Transposase (**XP_001338227**)	64/107 (59)	9e-29	DR, OM
LIH_63	**CN171307**	191/208 (91)	Transposase (**XP_001339494**)	44/90 (48)	1e-26	DR.OM
LII-78	**CT703965**	75/81 (92)	Transposase (**CAB51371**)	110/271 (40)	8e-42	PP,OM,DR
LIK_08	**CT664262**	35/35 (100)	Transposase (**CAB51372**)	57/86 (66)	9e-51	PP,OM,DR
LIF_90	**DN597649**	55/60 (91)	Solute carrier family 39 member (**NP_997748**)	198/207 (95)	9e-82	DR,TN,OM
LIF-30	**EB937013**	65/71 (91)	Talin 1 (**NP_001009560**)	191/191 (100)	4e-101	DR,RN
LIF-36	**CT695338**	49/52 (94)	Alpha-actinin (**XP_001345739**)	23/24 (95)	0.001	DR
LIF-57	**EE303946**	60/61 (98)	MGC83562 protein (**XP_421214**)	52/126 (41)	2e-24	GG,XL
LIF-81	**AL925325**	65/69 (94)	VAV-3 protein (**XP_001515846**)	61/108 (56)	3e-20	OA,MM,RN
LIF-20	**EB900829**	45/47 (95)	Protein tyrosine phosphatases epsilon (**XP_695831**)	25/25 (100)	8e-08	DR,TN,HS
HLJ354	**CT635692**	97/103 (94)	Activin receptor IIB (**XP_697649**)	134/134 (100)	2e-80	DR,HS,GG
HLJ356	**CN014012**	128/145 (88)	T-cell immune regulator (**NP_998234**)	248/260 (95)	1e-122	DR,TN,RN
HLJ870	**EE325778**	91/95 (95)	Phenylalanyl-tRNA synthetase (**AAI28806**)	222/222 (100)	8e-132	DR,HS,MM
LII_01	**CT609085**	49/53 (92)	Sorting nexin 13 (**AAO15004**)	23/38 (60)	3e-05	TR,GG,MM
HLJ376	**DY556876**	89/91 (97)	Actin-related protein (**NP_001003944**)	212/212 (100)	2e-121	DR,TN,MM
HLJ380	**EE716632**	37/38 (97)	Diaphorase (NADH) (**NP_956483**)	230/233 (98)	2e-123	DR,TN
LII_47	**EG580779**	62/65 (95)	136342 protein (**AAI46619**)	240/242 (99)	8e-149	DR,MD,GG

(to be continued)

(continued)

Carp clone No.	Zebra fish EST (Accession No.)	Identity Blastn (%)	Matched gene (Accession No.)	Similarity Blastx(%)	E-value	Species[a]
HLJ390	**EB852635**	85/90 (94)	DNA binding protein (**XP_696090**)	154/154 (100)	1e-72	DR,HS,MM
LII-82	**EB885308**	40/41 (97)	Capping protein (**NP_956229**)	110/113 (97)	7e-59	DR,TN,HS
LIK_06	**CT676279**	84/89 (94)	Alpha-actinin (**AAN77132**)	190/190(100)	9e-107	DR,TN
LIK_56	**EG571207**	77/80 (96)	VAMP -associated protein (**NP_001002546**)	172/173 (99)	1e-115	DR,HS,RN
LIK_70	**CT602809**	35/35 (100)	Oviductin (**XP_001377953**)	23/26 (88)	0.005	MD
LIK_80	**EG573763**	35/36 (97)	G protein pathway suppressor (**EDM06902**)	142/145 (97)	1e-72	RN,DR,MM
LIK_82	**CT609085**	49/53 (92)	Sorting nexin 13 (**AAO15004**)	23/38 (60)	3e-05	TR,GG,MM
HLJ418	**EB927888**	99/110 (90)	Damage-specific DNA binding protein (**EDM12828**)	182/216 (84)	8e-100	DR,RN,TN
LIL-14	**CN014012**	132/149 (88)	T-cell immune regulator (**NP_998234**)	248/260 (95)	1e-122	DR,TN,GG
LIL-19-2	**CK684754**	37/38 (97)	ORF2-encoded protein (**BAE46429**)	43/50 (86)	6e-16	DR, PO
HLJZe33	**CK691226**	71/76 (93)	Meningioma (**NP_001093672**)	117/183 (63)	2e-45	XT,TN,RN
LIL-63	**CK146298**	59/62 (95)	Nrp1b protein (**AAI33731**)	107/109 (98)	3e-59	DR,MAM
LIL-73	**EE699974**	91/102 (89)	Fibroblast growth factor (**NP_001093754**)	138/193 (71)	9e-70	DR,XT,HS
LIL-91	**EB952613**	61/66 (92)	DnaJ (Hsp40) homolog (**NP_955956**)	126/126 (100)	4e-68	DR,MD
LIL-96	**DV590107**	76/85 (89)	Dynactin 3 (**NP_001002220**)	186/187 (99)	4e-99	DR,TN,XL
LIM_18	**EB931517**	46/49 (93)	Triple functional domain (**NP_001097996**)	132/145 (91)	2e-60	DR,TN,HS
LIM_64	**DY554324**	122/135 (90)	Protein tyrosine phosphatase (**NP_956140**)	174/233 (74)	2e-72	DR,XL,HS
LIM-46	**EB961560**	95/103 (92)	Ubiquitin protein ligase E3 (**AAH81553**)	90/90 (100)	2e-29	DR,MM,HS
LIN-15	**DN597286**	94/102 (92)	Mitochondrial peptide chain release factor (**CAC24560**)	70/126 (55)	9e-37	PF,TN,RN
LIN-60	**CN505827**	50/52 (96)	Sema domain (**NP_998164**)	256/259 (98)	7e-130	DR,MD,HS
HLJ779	**EE718464**	35/36 (97)	ATP-binding domain (**NP_001014203**)	129/181 (71)	4e-89	RN,DR
LIN-72	**DT060464**	104/114 (91)	RNA splicing factor (**NP_596908**)	97/143 (67)	2e-38	DR,RN,MM
LIN-76	**EB946165**	137/139 (98)	Testican-2 (**XP_690238**)	94/94 (100)	8e-55	DR,TN,XL
LIN-91	**CK396800**	144/159 (90)	ROD1 protein (**XP_001335967**)	131/131 (100)	2e-68	DR,RN
LIO-11	**BM889554**	52/54 (96)	Osmotic stress transcription factor (**AAT84345**)	44/61 (72)	5e-12	TN,MD
LIP-C6	**EE699974**	91/102 (89)	Fibroblast growth factor (**NP_001093754**)	138/193 (71)	9e-70	XT,DR,PT

(to be continued)

(continued)

Carp clone No.	Zebra fish EST (Accession No.)	Identity Blastn (%)	Matched gene (Accession No.)	Similarity Blastx(%)	E-value	Species[a]
LIP-E07	**EE303946**	60/61 (98)	MGC83562 protein (**XP_421214**)	52/126 (41)	2e-24	GG,XL
LIP-H21	**AL911930**	118/128 (92)	66097 protein (**AAH74055**)	59/59 (100)	6e-27	DR, TN
LIP-H07	**DN597649**	55/60 (91)	Solute carrier family 39 (**NP_997748**)	198/207(95)	9e-82	DR,TR
LIP-H11	**DV585441**	44/46 (95)	Acetylglucosaminyltransferase (**XP_001114281**)	67/118 (56)	1e-32	MAM,DR,HS
LIQ_48	**CK704847**	36/37 (97)	Caveolin 3 (**NP_991301**)	150/150 (100)	7e-85	DR,TN,GG
HLJ526	**EB912091**	112/126 (88)	Vertebrate t-complex testis expressed 1 (**CAI12013**)	101/104 (97)	4e-52	DR, TN

[a]The names of species were abbreviated as follows: DR, *Danio rerio*; GG, *Gallus gallus*; HS, *Homo sapiens*; MM, *Mus musculus*; RN, *Rattus norvegicus*; MD, *Monodelphis domestica*; OA, *Ovis aries*; PO, *Paralichthys olivaceus*; XL, *Xenopus laevis*; TN, *Tetraodon nigroviridis*; OM, *Oncorhynchus mykiss*; PP, *Pleuronectes platessa*; XT, *Xenopus tropicalis*; MAM, *Macaca mulatta*; PF, *Platichthys flesus*; PT, *Pan troglodytes*.

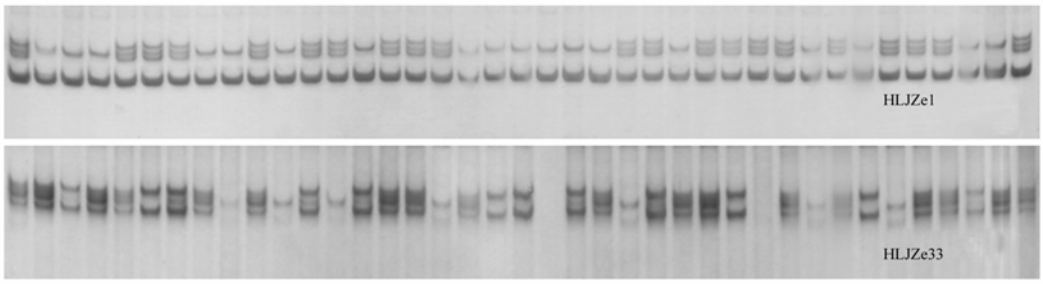

Fig. 1 Polymorphisms of two STS markers in 36 F_2 hybrids, each lane stands for one individual

HLJ380(LG40), HLJ418(LG19) and HLJ870(LG4) were linked to RIL map, four sites, HLJ321(LG6), HLJ376(LG8), HLJ526(LG14) and HLJ779(LG8) were anchored to self-crossed map.

2.2 Two new sequence tag sites (STS) developed by PCR-SSCP method

Ten primer pairs were redesigned from matched zebrafish EST sequences according to the primer designing criteria. As a result, six primer sets amplified target bands clearly, but only two primers, HLJZe1 and HLJZe33, had polymorphisms in 36 F_2 individuals (Fig. 1). Subsequently, two STS markers were used for linkage analysis, and only HLJZe33 were mapped successfully (Fig. 2).

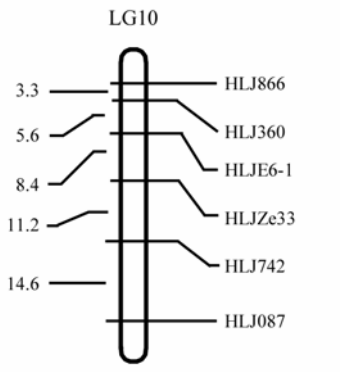

Fig. 2 One STS marker HLJZe33 was mapped to linkage 10 of the self-crossed F_2 map

3 Discussion

According to our findings in microsatellite development and application, only 10%-20% of microsatellites were useful in population genetics or other research fields. Most microsatellite sequences were abandoned due to failures in primer design or PCR amplification. However, flanking sequences of microsatellites usually are highly conserved in one species and its close relatives. Those conserved sequences might lie in coding regions. Sequences of coding regions are more conserved than non-coding regions and thus are better for relating genes

between different species (Ju et al, 2005). Common carp belongs to the same Cypriniae Family as zebrafish, which has been developed as a genetics model organism and accumulated abundant genetic data. It is greatly advantageous and significantly meaningful to apply zebrafish genomic and genetic data to the genome research of the common carp. In fact, the first genetic linkage map of common carp contained 65 microsatellites of zebrafish (Sun & Liang, 2004). Quan et al (2006) utilized 6072 zebrafish mircrosatellites to detect polymorphisms in *C. c haematopterus*, *C. c* var. *wuyuanensis* and *C. pellegrini pellegrini*. As a result, 646 amplified target bands (9%), of 563 were polymorphic and useful in three different carp. In our study, by comparing flanking sequences of common carp to zebrafish dbESTs, 121 homologous genes (7%) were found with high identities (>85%). These results demonstrated that the genetic and genomic resources of zebrafish can be used in the genetic and genomic research of the common carp.

Using Blastx search, 58 sites were annotated and established associations with some functional genes, and eight of them were anchored to linkage maps (seven microsatellite sites and one STS marker). Despite the fact that the remaining 50 sites have not been mapped yet, 48 sites were localized to the same genes in two or more vertebrates, thus supporting the identification of a true gene. Another two sites, LIF-36 and LIK_70, were detected in one vertebrate alone using Blastx searches with lower E-values. Thereafter, sequences of them were blasted against the database of the fugu genome project (http://fugu.hgmp.mrc.ac.uk/blast), however, none of them found homologous genes, indicating these two matched sequences might be false matches.

This study was a pilot for comparative studies between common carp and zebrafish, nevertheless, the results provided insights for proceeding with further studies into this area of research. In addition, with the genome sequencing completion of zebrafish, more and better genetic and genomic information will be available for comparative studies and more results can be expected from comparative genome research between these two fish species.

Acknowledgements: The authors thank Dr. ZHOU Yi for re-editing the English language of this manuscript

References:

Aliah RS, Takagi M, Dong SM, Teoh CT, Taniguchi N. 1999. Isolation and inheritance of microsatellite markers in the common carp *Cyprinus carpio* [J]. *Fish Sci,* **65**: 235-239.

Cnaani A, Ron M, Hulata G, Seroussi E. 2002. Fishing in silico: searching for tilapia genes using sequences of microsatellite DNA markers [J]. *Anim Genet,* **33**: 474-476.

Crooijmans RPMA, Bierbooms VAF, Koman J, Van der Poel JJ, Groenen MAM. 1997. Microsatellite markers in common carp [J]. *Animl Genet,* **28**: 129-134.

David L, Rajasekaran P, Fang J, Hillel J, Lavi U. 2001. Polymorphism in ornamental and common carp strains (*Cyprinus carpio* L.) as revealed by AFLP analysis and a new set of microsatellite markers [J]. *Mol Genet Genomics,* **266**: 353-362.

Farber CR, Medrano JF. 2003. Putative in silico mapping of DNA sequences to livestock genome maps using SSLP flanking sequences [J]. *Anim Genet,* **34**: 11-18.

Herron BC, Silva GH, Flaherty L. 1998. Putative assignment of EST to the genetic map by use of the SSLP database [J]. *Mamm Genome,* **9**: 1702-1704.

Ju ZL, Wells MC, Martinez A, Hazlewood L, Walter RB. 2005. An *in silico* mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, *Fundulus*, and *Xiphophorus* [J]. *In Silico Biology,* **5**: 0041 </isb/2005/05/0041/>.

Kantety RV, Rota ML, Matthews DE, Sorrells ME. 2002. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat [J]. *Plant Mol Biol,* **48**: 501-510.

QuanYC, Liang LQ, Sun XW, Yao K, Lei QQ. 2006. Identification of genetic polymorphism between three species of common carp using zebrafish Danio rerio microsatellite molecular markers [J]. *J Fish Sci China,* **13**: 300-304.

Rozen S, Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: S. Krawetz, S. Misener (eds), Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa (NJ): Humana Press, 365-386.

Sun XW, Liang LQ. 2004. A genetic linkage map of common carp (*Cyprinus carpio* L.) and mapping of locus associated with cold tolerance [J]. *Aquaculture*, **238**: 165-172.

Wei DW, Lou YD, Sun XW. 2001. Isolation of microsatellite markers in the common carp [J]. *Zool Res*, **22**: 238-241.

Yue GH, Orban L. 2004. Microsatellites within genes and ESTs of common carp and their applicability in silver crucian carp [J]. *Aquaculture*, **234**: 85-98.