Front. Agric. China 2010, 4(1): 121–127
DOI 10.1007/s11703-010-0101-3

RESEARCH ARTICLE

Jiaqing WANG, Lin HOU

Unigene derived SSR analysis for the *Fugu rubripes* and insights into the characteristics of EST-SSR distribution in tissues/organs

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2010

Abstract A total of 11348 Unigene derived ESTs from five cDNA libraries corresponding to five different tissues/organs were screened for EST-SSR, and 1097 EST–SSRs (9.67%) were detected. There were considerable variations in the amount of EST-SSRs among the five tissues/organs. In all the tissues/organs, trinucleotide repeats (67.74%) were found to be most abundant. The trinucleotide repeats GCG were the most abundant in three tissues/organs that ranged from 80.2% in skin, 83.3% in gut, to 90.2% in fin. Dinucleotide was the second abundant repeat type (13.58%), followed by tetra-(12.38%), hexa-(5.34%), and penta-(0.96%) nucleotide repeats. Furthermore, the results of the present study also indicated that the EST-SSR motifs were not randomly distributed in tissues/organs of fugu, such as trinucleotide repeats were found to be strikingly abundant in fin, followed by skin and gut, tetra-nucleotide repeats were especially rich in muscle, while pentanucleotide repeats were less common in any tissue or organ. Most of the EST-SSRs in five tissues/organs that were assigned functions during the present study represent housekeeping genes.

Keywords Unigene, *Fugu rubripes*, EST-SSRs, tissues/organs

Received September 16, 2009; accepted November 9, 2009

Jiaqing WANG (✉)
College of Ocean, Agricultural University of Hebei, Qinhuangdao 066003, China
E-mail: jiaqing5212@163.com

Lin HOU (✉)
College of Life Sciences, Liaoning Normal University, Dalian 116029, China
E-mail: houlin01@126.com

1 Introduction

Microsatellites or simple sequence repeats (SSRs) are the stretches of DNA consisting of exact simple tandemly repeated short motifs of 1–6 bp in length that frequently exhibit variation in the number of repeats at a locus. SSRs are ideal DNA markers because they are highly polymorphic between individuals and highly abundant dispersing evenly throughout the genomes (Klintschar and Wiegand, 2003). Moreover, microsatellite markers are reliable, codominant, multiallelic, and highly informative genetic markers well established for genetic analysis in many plant and animal species (Yu et al., 2006; Tu et al., 2007). The common procedure for developing SSRs involves the construction of a genomic library, the subsequent hybridization with tandemly repeated oligonucleotides, and the sequencing of candidate clones, thus, making the process time-consuming and labor-intensive. An alternative strategy for the development of SSRs arises from increasing information available in genomic DNA and EST databases (Ince et al., 2007).

Recently, the use of EST databases dramatically increased, and new strategies for the efficient use of ESTs have been developed. Although many researchers use EST data bases in gene discovery, a label to map a specific location on a chromosome, microsatellite mining, and various other studies, limited research is available for mining tissues, organs specific microsatellites (Bilgen et al., 2004; Ince et al., 2007). Microsatellites within genomic DNA sequences have been shown to be important in the formation of hairpin structures that may provide some structural or replication mechanisms (McMurray, 1999). It is also known that a growing number of neurological disorders associated with the repeated DNA

(Reddy and Housman 1997; Timchenko and Caskey, 1999). Therefore, studies on microsatellites located in coding regions of important genes would be very valuable to understand the repeat function in gene.

Fugu was introduced as a "genomic model" organism specifically due to its compact genome is only one-ninth the size of the human genome (Blair and Sudhir, 2002); yet, it contains approximately the same number of genes, permitting efficient comparison with other organism genomes. Studies on distribution of EST-SSR markers in tissues/organs would provide sufficient information for genetic studies on transcribed part of the genome and genome mapping of the fugu and its closely related fishes.

2 Methods

2.1 Retrieval of Unigene sequences

Unigenes comprising approximately 5.6 Mb in fugu were acquired in bulk through a FASTA sequence retrieval system from GenBank (http://www.ncbi.nlm.nih.gov). These sequences were downloaded from five different cDNA libraries corresponding to five different tissues/organs (adult gut, adult skin, adult fin, adult muscle, and adult ovary, respectively).

2.2 Mining microsatellites

Fugu unigene databases were used to identify and characterize SSRs using the Tandem Repeats Analyzer 1.5 program (TRA1.5), which was downloaded from ftp://ftp.akdeniz.edu.tr/Araclar/TRA/ (Mehmet et al., 2004). In this study, we searched for EST-SSRs using the exact module of the program. This computer program was run under windows, using a FASTA-formatted sequence file containing multiple sequences. TRA1.5 was set with the criteria of minimum number of seven repeats for dinucleotide, six for trinucleotide, five for tetranucleotide, four for pentanucleotide, and three for hexanucleotide. Single nucleotide repeats were not selected, because they were generally not considered as useful as polymorphic markers. The complementary sequences, for instance, AC/TG, CGA/GCT, were considered as the same repeat motif type.

2.3 Annotation of SSR-containing sequences

To obtain an idea about putative functions of SSR containing genes, these sequences were compared to the nonredundant (nr) protein database of the NCBI database (http://www.ncbi. nlm.nih.gov/blast) by BLASTX program using 1e-05 as a cutoff expected value. The proteins obtained during similarity search were classified into their respective classes.

3 Results

3.1 Searching for EST-SSRs and frequencies of fugu SSRs with different repeat motifs

A total of 11348 ESTs were screened for EST-SSR, and 1097 EST-SSR (9.67%) were detected (Table 1). There were considerable variations in the amount of EST-SSRs among the five tissues/organs. The highest number of EST-SSRs was in the adult fin of fugu (2894 ESTs and 21.49% of which were EST-SSRs). The adult muscle consisted of 1492 ESTs, and the EST-SSRs ratio was 8.51%. Adult gut consisted of 2447 ESTs and the EST-SSRs ratio was 6.05%. Adult skin contained 1970 ESTs, and the EST-SSRs ratio was 5.63%. Adult ovary, however, had the relatively higher ESTs (2545) than the other tissues/organs (except for adult fin) but the lowest EST-SSRs (3.5%) (Table 2). In general, the occurrence of the individual SSR motif was not evenly distributed within the whole sequences (Fig. 1): 155 (13.58%) were dinucleotide, with

Table 1 Overall statistics

parameter	No.
sequence processed	11348
sequence with repeat strings	1097
total repeat strings	1145
repeat index	0.101
average number of repeat strings per sequence	1.044
repeat percentage	9.67%
simple repeat strings	1141
compound-simple repeat strings	4

Note: Repeat index is a numerical value showing how many repeat strings were found among the all sequences processed, and it can be calculated by dividing the value of repeat string numbers to the number of ESTs processed. Repeat percentage (%) shows how many sequences were processed and how many of them contained repeats.

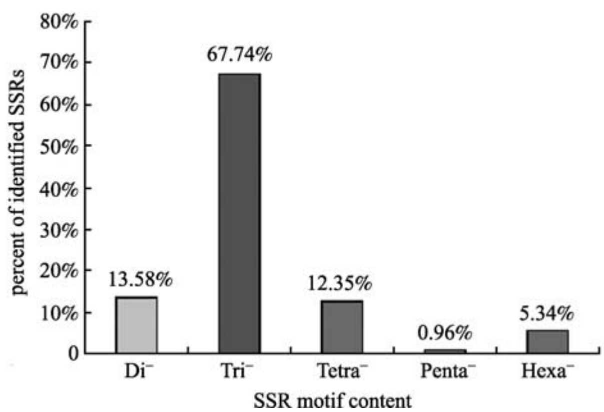

Fig. 1 Frequency distribution of different repeat types (2–6 motif units) microsatellite identified in Unigene sequences of fugu
Note: The numbers over the bars indicate the percent of each repeat type microsatellites in five tissues/organs in total.

Table 2 Distribution of fugu EST-SSRs among organs/tissues

tissues/organs	No. of seq. (processed)	No. of seq. (with repeat strings)	percentage	No. of repeat strings	repeat index
gut	2447	148	6.05	158	0.065
skin	1970	111	5.63	115	0.058
fin	2894	622	21.49	636	0.22
muscle	1492	127	8.51	132	0.088
ovary	2545	89	3.5	104	0.041
total	11348	1097	9.67	1145	0.101

Note: Repeat index is a numerical value showing how many repeat strings were found among the all sequences processed, and it can be calculated by dividing the value of repeat string number to the number of ESTs processed.

773 (67.74%) trinucleotide, 141 (12.35%) tetranucleotide, 11 penta-nucleotide (0.96%), and 61 (5.34%) hexanucleotide microsatellites.

Dinucleotide repeats were the most abundant in ovary and gut, while they were the least abundant in muscle among the five fugu organs/tissues. The most common dinucleotide repeat type was AC/TG and CA/GT in all the tissues/organs. Trinucleotide repeats were prominent in adult fin, followed by skin, gut, ovary, and muscle. The most abundant trinucleotide repeat type was GCG, which possibly codes the amino acid alanine. Tetranucleotide repeats were mostly in adult muscle. The most abundant tetranucleotide repeat type was CAAC in all the tissues/organs except for skin. Pentanucleotide repeats were less common in fugu ESTs. Hexanucleotide repeats were rich in ovary and gut, followed by fin, skin, and muscle (Fig. 2) (Table 3).

3.2 Repeat polymorphisms analysis

Repeat polymorphisms in EST-SSRs (different ESTs containing the same SSR) generally decreased as the

Table 3 Distribution of EST–SSRs motifs among fugu tissues/organs

tissues/organs	Di- (%)	Tri- (%)	Tetra- (%)	Penta- (%)	Hexa- (%)
gut	AC/TG (41.3)	AAC (1.3)	AAAT (7.1)	AAAAG (33.3)	AAGAAA (12.5)
	AG/TC (10.9)	AAG (1.3)	CAAC (64.3)	ATTTT (33.3)	AATAAA (6.3)
	AT/TA (8.7)	ATC (1.3)	CATC (7.1)	TTTAA (33.3)	AATCTG (6.3)
	CA/GT (32.6)	CAG (2.5)	GATG (7.1)		AGAACC (6.3)
	CT/GA (6.5)	GCC (1.3)	GGAT (7.1)		AGCCTG (6.3)
		GCG (83.3)	TCCA (7.1)		CTCGGT (6.3)
		GCT (5.1)			GCACGC (6.3)
		TCC (3.8)			GCCTAA (6.3)
					GTCCTC (6.3)
					GTTCCA (12.5)
					TCTGGT (6.3)
					TCTTTT (12.5)
					TTGTTC (6.3)
skin	AC/TG (23.5)	ATC (1.2)	GACA (100)	GAGCC (50.0)	AACAGG (16.7)
	AT/TA (17.6)	ATG (1.2)		TCTAT (50.0)	ACCAGA (16.7)
	CA/GT (41.2)	CAG (4.7)			AGAACC (33.3)
	CT/GA (11.8)	CCT (1.2)			GGTGTG (16.7)
	TC (5.9)	CGA/GCT (3.5)			GTGTGC (16.7)
		CTG (1.2)			
		GAG (2.3)			
		GCG (80.2)			
		TCA (1.2)			
		TCC (1.2)			
		TGA (2.3)			

(*Continued*)

tissues/organs	Di- (%)	Tri- (%)	Tetra- (%)	Penta- (%)	Hexa- (%)
fin	AC (18.5)	AAG (0.2)	AAAC (6.7)	TCTAT (50.0)	AACAGG (6.7)
	AG (3.7)	AGA (0.2)	ATAC (6.7)	TTAGC (50.0)	AAGAAA (46.7)
	AT/TA (18.5)	AGG (0.2)	CAAC (66.6)		AGATTC (6.7)
	CA/GT (48.1)	ATC (0.2)	CAGA (13.3)		AGCCTG (6.7)
	CT/GA (11.1)	ATT (0.2)	TGTC (6.7)		GACTCG (6.7)
		CAG (0.5)			GCCTAA (6.7)
		CGG (2.6)			TGCCGC (6.7)
		CTG (0.2)			TGGAAA (6.7)
		GAG (0.2)			TTCCTT (6.7)
		GCG (91.2)			
		GCT (0.2)			
		GGC (3.3)			
		TCA (0.2)			
		TGA (0.3)			
		TGG (0.5)			
muscle	AC/TG (30.8)	CAA/GTT (27.3)	CAAC (99.0)	AGAGG (33.3)	AAGAAA (25.0)
	AG/TC (23.1)	CAT (18.2)	CCAT (1.0)	GTCCT (66.7)	AGGACG (25.0)
	CA/GT (38.5)	GAG (9.1)			CGCCGG (25.0)
	GA (7.7)	GCG (27.3)			TGCCGC (25.0)
		GCT (9.1)			
		TAC (9.1)			
ovary	AC/TG (36.5)	AGC (4.8)	CAAC (33.3)	GGTTC (100)	AAGAAA (10.0)
	AT/TA (3.8)	CAG/GTC (19.0)	CGGA (11.1)		AGAACC (5.0)
	CA/GT (55.8)	CAT (4.8)	GATG (22.2)		AGCCTG (5.0)
	CT/GA (3.8)	CCT (4.8)	GCGT (22.2)		AGCGAC (15.0)
		CTG (4.8)	TCTT (11.1)		AGTGAC (5.0)
		GAG (9.5)			CCGCTG (5.0)
		GCA (9.5)			CCTGGT (5.0)
		GCG/CGC (9.5)			CTGAAT (5.0)
		GCT (4.8)			CTGGTT (5.0)
		TCC (9.5)			GAAGGG (5.0)
		TCT (9.5)			GACAGC (10.0)
		TGA (9.5)			GAGAAG (5.0)
					GAGGAA (5.0)
					GCAGCG (5.0)
					GCCTAA (5.0)
					TTAGCT (5.0)

motif lengths increased. For instance, motif length of 6 varied from 3 to 9 repetitions, whereas motif length of 2 varied from 7 to 39 repetitions. Variations in the EST–SSRs in most cases were due to the repeat number variations, whereas variations in the whole genome excluding EST-SSRs were mainly due to mismatches or base substitution in motifs. There were only four compound-simple repeat strings were found in all the tissues/organs.

3.3 Annotation of fugu EST-SSRs in tissues/organs

To determine the function of SSR containing sequences, the 1141 sequences from which SSRs were mined and annotated against the nonredundant (nr) protein database. Out of 148 SSR containing sequences in gut, 108 (72.97%) sequences were available for annotation. These proteins belong to repetitive elements (100, 67.57%), selenide, water dikinase (1, 0.68%), endothelin precursor (1,

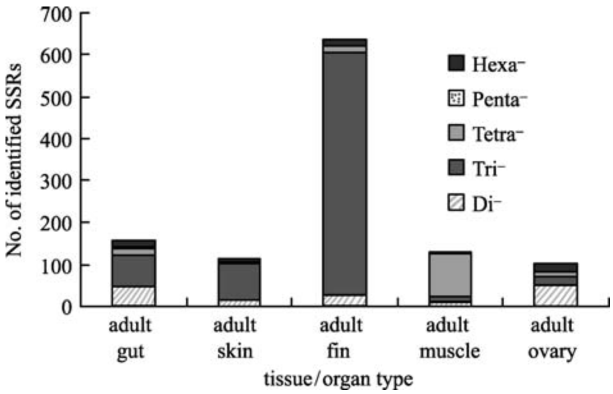

Fig. 2 Frequency distribution of different repeat types (2–6 motif units) microsatellite identified in Unigene sequences of fugu in five different tissues/organs, respectively

0.68%), acyl coenzyme (1, 0.68%), sterol o-acyltransferase (2, 1.35%), natural resistance-association macrophage protein (1, 0.68%), ras-related protein (1, 0.68%), ferritin (1, 0.68%), metastasis-associated protein (1, 0.68%), alkaline phosphatase (2, 1.35%), glococoricoid-induce leucine zipper gilz protein (1, 0.68%), T-complex protein (1, 0.68%), reductase (2, 1.35%), phosphoglycerate kinase (1, 0.68%), hypothetical 6.1 kd protein (2, 1.35%), cytochrome p450 (1, 0.68%), translation initiation factor (1, 0.68%), major vault protein (1, 0.68%), mps1 protein precursor (1, 0.68%), and others (12, 8.11%), which were not assigned to any specific class of proteins. Moreover, a number of sequences (15, 49.1%) remained unannotated due to the absence of a homolog in the protein sequence database. Out of 111 SSR containing sequences in skin, 98 (88.29%) sequences were available for annotation. These proteins belonged to repetitive elements (86, 77.48%), polo like kinase (1, 0.90%), tudor repeat associator with pctaire (1, 0.90%), ferritin (1, 0.90%), activin receptor-like kinase (1, 0.90%), type II basic cytokeratin (1, 0.90%), RNA binding motif protein (1, 0.90%), and others (6, 5.41%), which were not assigned to any specific class of proteins. In addition, a number of sequences (13, 11.71%) remained unannotated. Out of the 622 SSR containing sequences in fin, 618 (99.36%) sequences were available for annotations. These proteins belonged to repetitive elements (594, 95.50%), translation initiation factor (5, 0.80%), synaptobrevin (3, 0.48%), smt3a protein (1, 0.16%), glucocorticoid-induced leucine zipper gilz protein (1, 0.16%), calcium-independent phospholipase (1, 0.16%), ras-related protein (1, 0.16%), paired box protein (1, 0.16%), and others (11, 1.77%), which were not assigned to any specific class of proteins. Moreover, a number of sequences (4, 0.64%) remain unannotated. Out of 127 SSR containing sequences in muscle, (122, 96.06%) sequences were available for annotation. These proteins belonged to repetitive elements (13, 2.09%), translation initiation factor (1, 0.79%), ganglioside expression factor (1, 0.79%), sh3 domain-containing protein (1, 0.79%), myopodin protein (1, 0.79%), glycine max gmsti mRNA (1, 0.79%), amphiphysin-like protein (1, 0.79%), and others (103, 81.10%), which were not assigned to any specific class of proteins. Moreover, a number of sequences (5, 3.94%) remained unannotated. Out of 89 SSR containing sequences in the ovary, 60 (67.42%) sequences were available for annotation. These proteins belonged to repetitive element (33, 37.08%), selenide water dikinase (1, 1.12%), serine/threonine-protein kinase (1, 1.12%), ras-related protein (1, 1.12%), ferritin (1, 1.12%), Nf-kappa b essential modulator (1, 1.12%), cell division cycle control protein (1, 1.12%), hypothetical 55.4 KD protein (1, 1.12%), type II basic cytokeratin (1, 1.12%), epstein-barr virus complement receptor type II (CR2) precursor (1, 1.12%), T-complex protein (1, 0.68%), probable protein disulfide isomerase er-60 precursor (2, 2.24%), translation initiation factor (2, 2.24%), 26S proteasome regulatory subunit S2 (2, 2.24%), and others (12, 13.48%), which were not assigned to any specific class of proteins. In addition, a number of sequences (29, 32.58%) remained unannotated due to the absence of a homolog in the protein sequence database.

4 Discussion

In recent years, research efforts have yielded a considerable number of genomic and transcribed ESTs from many species, and researchers are now able to access the important information contained within these sequences. Earlier studies indicated that a significant portion of ESTs consisted of short tandem repeats (Kantety et al., 2002; Saha et al., 2003; Bilgen et al., 2004). Our preliminary study on fugu organs/tissues indicated that it might be possible to identify the genes contributed to a tissue or organ's unique characteristics. Furthermore, our analysis of EST-SSRs on fugu may be very useful to obtain advanced knowledge about the functions of the repeated DNA sequences in transcribed genes of tissues or organs of fugu.

The results of the present study indicated that SSR motifs were not randomly distributed in fugu coding genes, such as trinucleotide repeats were found to be strikingly abundant in fin, followed by skin and gut, tetranucleotide repeats were especially rich in muscle, while pentanucleotide repeats were less common in any tissue or organ. Similar results about unevenly distributed individual SSR motifs occurring in ESTs have also been observed in plants (Bilgen et al., 2004).

Our results also showed that all the EST-SSRs were the least in the ovary among the five tissues/organs, which can be explained by the fact that ovary is in charge of the function of inheritance from parents to offsprings, and any mutation in gene or alteration in gene expression will result in the instability to inheritance. However, some researchers considered microsatellites to be selectively neutral sequences and randomly or almost randomly distributed over the genome (Jeffreys et al., 1985; Schlotterer and Wiehe, 1999; Heslop-Harrison, 2003). Many aspects in the distribution of microsatellites across coding and noncoding regions, including presence or absence in a specific organ or tissue, functions in gene expression and evolutionary significance, genetic disorder and chromatin organization, contribution of replication slippage, recombination, and mutation still remain controversial (Timchenko and Caskey, 1999; Schlotterer and Wiehe, 1999).

In this study, trinucleotide repeats (67.74%) were found to be most abundant in all the tissues/organs, which was in agreement with previous reports on several plant and animal species (Cardle et al., 2002; Varshney et al., 2002; Franklin et al., 2005) but was different from some animal species, where dinucleotide repeats were abundant (Rohrer et al., 2002; Serapion et al., 2004; Ju et al., 2005; Perez et al., 2005). Dinucleotide was the second abundant repeat type (13.58%), followed by tetra-(12.38%), hexa-(5.34%), and penta-(0.96%) nucleotide repeats. The results indicated that the SSRs with different repeat motifs were unevenly distributed. These results were consistent with previous findings, which showed that the abundance of different repeats varied extensively depending on the species examined (Toth et al., 2000). The shorter repeat motifs were predominant among SSRs identified. As the length of the repeat unit increases, their occurrence decreases. This can be explained by the fact that longer repeats have higher mutation rates but are less stable (Wierdl et al., 1997).

Dimeric SSRs, the motifs CA/GT (32.6, 41.2, 48.1, 38.5, and 55.8% in five tissues/organs, respectively), and AC/TG (41.3, 23.5, 18.5, 30.8, and 36.5% in five tissues/ organs, respectively) were the most common ones. This finding of CA/GT motif rich is in accordance with recent report about common carp (Wang et al., 2007). Riley and Krieger (2004) reported that AC/GT was eight times more common in membrane-function mRNAs. This indicated that AC/GT repeats preferentially express in membrane proteins. The most common tetrameric microsatellite motifs were CAAC (64.3% in gut, 66.6% in fin, 99.0% in muscle, and 33.3% in ovary) in all the tissues/organs except for skin, which only contain one tetranucleotide repeat type GACA. The strikingly dominance of trimeric EST-SSRs (67.74%) in fugu unigene database can be explained by the suppression of nontrimeric SSRs in coding regions (ESTs) probably due to the risk of frameshift mutations (Metzgar et al., 2000). Among the trinucleotide repeats, the motifs GCG coding for amino acid alanine were the most abundant in three tissues/organs ranging from 80.2% in skin, 83.3% in gut, to 90.2% in fin (Table 3). One disease of synpolydactyly caused by duplication of the motif GCG had been reported (Muragaki et al., 1996). Trimeric motifs, TAA and TAG, were not found and TGA (2.3% in skin and 9.5% in ovary) also rarely appeared, probably because they code for stop codons that have a direct stop effect on protein synthesis (Chin, 1996).Trimeric SSRs sequences have attracted many researchers since a growing number of neurological disorders associated with this repeated sequence (Reddy and Housman, 1997; Zhuchenko, 1997; Timchenko and Caskey, 1999).

Out of 1141 SSR containing sequences, annotations were available only for 1006 (88.17%) sequences, which were categorized into different functional classes of proteins (repetitive element, kinases, acyltransferase, ras-related protein, ferritin, alkaline phosphatase, transcription factor, reductase, etc.), and for the remaining 135 (11.83%) SSR containing sequences, no homology could be found, or no obvious function has by far been assigned. Most of the EST-SSRs that are assigned functions during the present study represent housekeeping genes. In fugu, 88.17% functions for EST-SSRs sequences were assigned as a good source for annotating sequences.

Although ESTs in the sequence databases are annotated and provide information, in many cases, they are very large, unorganized, and redundant and relatively in low quality, while the unigene database has an advantage that is the elimination of redundant sequences. Our results based on unigene sequences, therefore, more accurately reflected the density of microsatellites in the expressed component of the genomes. Microsatellites or SSRs located in coding regions of important genes expressed in organ or tissue would lead to the development of tissues/organs specific SSRs, which would be very valuable to understand the repeat function in gene mapping for breeding and evolutionary studies.

In our study, a total of 11348 ESTs from five cDNA libraries corresponding to five different tissues/organs were screened for EST-SSR. Overall, EST amounts of each tissues/organs were basically evenly distributed: 2447 (21.5%) in gut, 1970 (17.4%) in skin, 2894 (25.5%) in fin, 1492 (13.2%) in muscle, and 2545 (22.4%) in ovary. The result, however, indicated that EST-SSR motifs were not randomly distributed in tissues/organs of fugu. Though a total of 11348 ESTs data used in the study were not enough abundant to reflect the normal tissues/organs level, as more and more EST data submitting to unigene database, investigating the distribution of EST-SSR motifs in different tissues/organs will become increasingly credible and meaningful.

Acknowledgements Our work is financially supported by the Foundation Project of Science & Technology of Qinhuangdao (No. 200901A068).

References

Bilgen M, Karaca M, Onus A N, Ince A G (2004). A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. Bioinformatics, 20: 3379–3386

Blair S H, Sudhir K, (2002). Vertebrate genomes compared. Science, 297: 1283–1285

Cardle L, Ramsay L, Milbourne, D, Macaulay M, Marshall D, Waugh R (2002). Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics, 156: 847–854

Chin E C L (1996). Maize simple repetitive DNA sequences: abundance and allele variation. Genome, 39: 866–873

Heslop-Harrison J S (2003). Tandemly repeated DNA sequences and centromeric chromosomal regions of *Arabidopsis* species. Chromosome Res, 11: 241–253

Ince A G, Onus A N, Elmasulu S Y, Bilgen M, Karaca M (2007). In silico data mining for development of *Capsicum* microsatellites. Acta Horticulturae, 729: 123–127

Jeffreys A J, Wilson V, Thein S L (1985). Hypervariable minisatellite regions in human DNA. Nature, 314: 67–73

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

Kantety R V, La R M, Matthews D E, Sorrells M E (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol, 48: 501–510

Klintschar M, Wiegand P (2003). Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. Forensic Sci Int, 135: 163–166

McMurray C T (1999). DNA secondary structure: a common and causative factor for expansion in human disease. Proc Natl Acad Sci USA, 96: 1823–1825

Mehmet B, Mehmet K, Naci A O, Ays G I (2004). A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences. Bioinformatics, 20: 3379–3386

Metzgar D, Bytof J, Wills C (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res, 10: 72–80

Muragaki Y, Mundlos S, Upton J, Olsen B R (1996). Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science, 272: 548–551

Perez F, Ortiz J, Zhinaula M, Gonzabay C, Calderon J, Volckaert F A (2005). Development of EST-SSR markers by data mining in three species of shrimp, *Litopenaeus vannamei, Litopenaeus stylirostris* and *Trachypenaeus birdy*. Marine Biotechnol, 7: 554–569

Reddy P S, Housman D E (1997). The complex pathology of trinucleotide repeats. Curr Opin Cell Biol, 9: 364–372

Riley D E, Krieger J N (2004). Short tandem repeats are associated with diverse mRNAs encoding membrane-targeted proteins. Bioassays, 26: 434–444

Rohrer G A, Fahrenkrug S C, Nonneman D, Tao N, Warren W C (2002). Mapping microsatellite markers identified in porcine EST sequences. Anim Genet, 33: 372–376

Saha S, Karaca M, Jenkins J N, Zipf A E, Reddy O U K, Pepper A E, Kantety R (2003). Simple sequence repeats as useful resources to study transcribed genes of cotton. Euphytica, 130: 355–364

Schlotterer C, Wiehe T (1999). Microsatellites, a neutral marker to infer selective sweeps. In: Goldstein D B, Schlstterer C, eds. Microsatellites: Evolution and Applications. Oxford: Oxford University Press, 238–247

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

Timchenko L T, Caskey C T (1999). Triplet repeat disorders: discussion of molecular mechanisms. Cell Mol Life Sci, 55: 1432–1447

Toth G, Gaspari Z, Jurka J (2000). Microsatellites in different eukaryotic genome, survey and analysis. Genome Res, 10: 1967–1981

Tu M, Lu B R, Zhu Y Y, Wang Y Y (2007). Abundant within-varietal genetic diversity in rice germplasm from Yunnan Province of China revealed by SSR fingerprints. Biochemical Genetics, 45: 789–801

Varshney R K, Thiel T, Stein N, Langridge P, Graner A (2002). In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett, 7: 537–546

Wang D, Liao X, Cheng L, Yu X, Tong J (2007). Development of novel EST-SSR markers in common carp by data mining from public EST sequences. Aquaculture, 271: 558–574

Wierdl M, Dominska M, Petes T D (1997). Microsatellite instability in yeast, dependence on the length of the microsatellite. Genetics, 146: 769–779

Yu X M, Zhou Q, Qian Z Q, Li S, Zhao G F (2006). Analysis of genetic diversity and population differentiation of *Larix potaninii* var. *chinensis* using microsatellite DNA. Biochemical Genetics, 44: 483–493

Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton D W, Amos C, Dobyns W B, Subramony S H, Zoghbi H Y, Lee C C (1997). Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the α1A-voltage-dependent calcium channel. Nature Genet, 15: 62–69