Sains Malaysiana 40(5)(2011): 497–502

Characterisation of Simple Sequence Repeats in the Asian Seabass, *Lates calcarifer* by Random Sequencing

(Pencirian Ulangan Jujukan Ringkas dalam Ikan Siakap, *Lates calcarifer* melalui Penjujukan Rawak)

PAN-PAN CHONG, ADURA MOHD. ADNAN & KIEW-LIAN WAN*

ABSTRACT

In recent years, there has been considerable interest in simple sequence repeats (SSRs) particularly as molecular markers with applications in many different fields. We have carried out an effort to identify and analyse SSRs in the genome of the Asian seabass, Lates calcarifer by random sequencing. Genomic DNA was isolated from the muscle tissue of L. calcarifer*, sheared by nebulisation and ligated into plasmid vector. Recombinant clones were selected randomly from the genomic libraries constructed. Subsequently, plasmid DNA was extracted and subjected to one-pass sequencing. A total of 4175 random sequences, also known as genome survey sequences (GSSs), with a total length of 1.7 Mb was generated. Screening of the whole* L. calcarifer *GSS data set allowed for the identification of a total of 151 perfect (100% similarity) SSRs. These SSR consensus patterns spread over a wide range of size (1 to 226 bp). The most frequent consensus pattern is dinucleotide, which represents 60% of all SSRs identified. The dinucleotides* $(AC)_n$, $(AT)_n$ *and* $(AG)_n$ *were also found to occur frequently in the* L. calcarifer *genome. Sequence comparison between* L. calcarifer *and other fish species showed variation in repeat content, indicating the different ways in which repeats may evolve in the genome of these species. Data generated from this random sequencing of the* L. calcarifer *genome should serve as a valuable resource for further studies of this organism.*

Keywords: Genome survey sequence; GSS; molecular marker; SSR

ABSTRAK

Sejak kebelakangan ini, terdapat minat yang mendalam mengenai ulangan jujukan ringkas (SSR), terutamanya sebagai penanda molekul dengan kegunaan dalam pelbagai bidang. Kami telah melaksanakan usaha untuk mengenal pasti dan menganalisis SSR dalam genom ikan siakap, Lates calcarifer *melalui penjujukan rawak. DNA genom telah dipencilkan daripada tisu otot* L. calcarifer*, diserpihkan dengan nebulisasi dan diligasikan ke dalam vektor plasmid. Klon rekombinan telah dipilih secara rawak daripada perpustakaan genom yang telah dibina. Seterusnya, DNA plasmid telah diekstrak dan diperlakukan penjujukan sekali lalu. Sejumlah 4175 jujukan rawak, yang juga dikenali sebagai jujukan tinjauan genom (GSS), dengan jumlah panjang 1.7 Mb telah dijana. Penabiran keseluruhan set data GSS* L. calcarifer *telah membolehkan pengenalpastian sejumlah 151 SSR sempurna (persamaan 100%). Corak konsensus SSR ini tersebar merentasi julat saiz yang luas (1 hingga 226 pb). Corak konsensus yang paling sering ditemui adalah dinukleotida, yang mewakili 60% daripada kesemua SSR yang dikenal pasti. Dinukleotida* $(AC)_n$, $(AT)_n$ *dan* $(AG)_n$ *juga dijumpai hadir dengan banyak dalam genom* L. calcarifer. *Perbandingan jujukan di antara* L. calcarifer *dengan spesies ikan lain mempamerkan variasi dalam kandungan ulangan, dan ini menunjukkan cara berbeza bagaimana ulangan berupaya berevolusi dalam genom spesies ini. Data yang terjana daripada penjujukan rawak genom* L. calcarifer *ini merupakan sumber yang berharga untuk kajian lanjut tentang organisma ini.*

Kata kunci: GSS; jujukan tinjauan genom; penanda molekul; SSR

INTRODUCTION

The Asian seabass, *Lates calcarifer* is one of the most economically important aquaculture species in South East Asia (Chou & Lee 1997). However, the molecular biology of this fish species remains largely unexplored and the number of genes identified at present is still limited (Mohd-Yusof et al. 2009; Tan et al. 2008). Genetic studies using appropriate markers will be useful to ascertain the diversity of this fish species and facilitate its production, including the management of brood stock for breeding purposes.

Simple sequence repeats (SSRs) or microsatellites are tandemly repeated tracts of DNA composed of 1 to 6 bp long unit. They show high levels of polymorphism, and are ubiquitous in prokaryotes and eukaryotes, present even in the smallest bacterial genome (Field & Wills 1996; Hancock 1996; Tautz 1989). SSRs can be found anywhere in the genome, both in protein-coding and non-coding regions. Thus, they are ideal markers for genetic variation studies (Watanabe et al. 2004), particularly among closely related populations or populations sampled over a reduced

geographical scale (Estoup et al. 1998; Wright & Bentzen 1994), and also genetic structure studies of cultivated and wild fish stocks (Jackson et al. 2003).

Realising the importance of SSR to facilitate further research, we have identified SSRs from the genome of *L. calcarifer* by random sequencing. To obtain a representative sampling of the genome, we have utilised libraries prepared from randomly sheared genomic DNA, and have sequenced inserts of clones randomly isolated from the libraries. These sequences are also termed genome survey sequences (GSSs). We found that this random sequencing approach is a productive and efficient method of SSR discovery.

Materials and Methods

GENOMIC LIBRARY CONSTRUCTION

Genomic DNA was isolated and purified from the muscle tissue of *L. calcarifer* obtained from the Fisheries Research Institute, Batu Maung, Penang, Malaysia using the Wizard® Genomic DNA Purification Kit (Promega, USA). Random genomic libraries from *L. calcarifer* were constructed using the TOPO® Shotgun Subcloning Kit (Invitrogen, USA). Briefly, the genomic DNA was sheared by nebulisation and purified to obtain a size range of 0.5 to 3.0 kb. Genomic DNA fragments were then blunt-end repaired with T4 DNA polymerase (Invitrogen, USA) and Klenow DNA polymerase (Invitrogen, USA), before they were dephosphorylated by calf intestinal phosphatase (Invitrogen, USA). The blunt-end and dephosphorylated genomic DNA fragments were subsequently ligated into the plasmid vector pCR®4Blunt-TOPO (Invitrogen, USA). Recombinant plasmids were transformed into TOP10 One Shot Electrocomp™ *Escherichia coli* (Invitrogen, USA) and plated on Luria-Bertani (LB) agar containing kanamycin. White colonies were picked into 96-well culture plates containing LB broth and kanamycin.

PLASMID PREPARATION AND DNA SEQUENCING

Plasmid DNAs were extracted and purified from randomly selected recombinant clones using the Montage Plasmid Miniprep$_{96}$ Kit (Millipore, USA). Analysis of insert size was carried out by agarose gel electrophoresis. These plasmid DNAs were then used as templates in cycle sequencing reactions with the T3 primer and the ABI PRISM® BigDye™ Terminator v3.0 Ready Reaction Cycle Sequencing Kit. Purified extension products were analysed on the ABI PRISM 3100 Genetic Analyzer (Applied Biosystems Inc., USA).

SIMPLE SEQUENCE REPEAT ANALYSIS

The generated sequences were base-called using Phred (Ewing et al. 1998; Ewing & Green 1998) and quality-trimmed at a cut-off value of 20. Vector sequences were masked by CrossMatch (Green 1996). Analysis of SSRs was done using Tandem Repeat Finder (version 3.21) (Benson 1999), with the following parameters: alignment match = 2, alignment mismatch = 7, alignment indel = 7, minimum alignment score to report repeat = 50 and maximun period size = 2,000. For the SSR comparative study, zebra fish (*Danio rerio*), medaka fish (*Oryzias latipes*) and fugu fish (*Tetraodon nigroviridis*) genomic sequences were downloaded from the GenBank database and scanned for various SSRs using Tandem Repeat Finder.

Results and Discussion

CHARACTERISATION OF GENOMIC LIBRARIES

Several small-insert libraries were prepared from genomic DNA that had been purified from the muscle tissue of a single fish to minimise potential complications due to allelic polymorphisms. *L. calcarifer* genomic DNA was mechanically fragmented by nebulisation to reduce bias, producing fragments that are randomly spread over the entire genome (Chong et al. 2005). By this method, a total of 4,175 random sequences were generated. These GSSs, with an average length of 410 bp, represent 1,711,858 bp of the *L. calcarifer* genome. Sequence assembly did not reveal any overlap among the sequences. This is expected since the amount of sequences that were generated represents only a small portion of the whole genome, and thus, showed that there is no bias towards certain regions of the *L. calcarifer* genome.

ANALYSIS OF *L. Calcarifer* SSRS

Analysis of the *L. calcarifer* sequence data using Tandem Repeats Finder revealed that based on different percentages of matches, between 151 to 435 SSRs were identified (Table 1). These results showed that the number of SSRs detected decreases with the increase of the percentage of matches. SSRs obtained at lower percentages of matches consist mainly of compound or imperfect repeats that are due to insertions and deletions between the copies of the repeats. Thus, only SSRs identified at 100% match were used for further analysis.

In this study, a total of 151 perfect SSRs, with a total length of 6,217 bp, were identified, accounting for 0.36% of the total length of the *L. calcarifer* GSSs. All of these perfect SSRs had a minimum copy number of 2, and a consensus pattern with a size of between 1 to 226 bp. As the sequences were generated from random libraries, consensus repeat patterns from both complimentary strands and in different reading frames were considered as in the same group. Thus, a polyA repeat is equivalent to a polyT repeat. Similarly, $(AC)_n$ is equivalent to $(CA)_n$, $(TG)_n$ and $(GT)_n$, while $(AAT)_n$ is equivalent to $(ATA)_n$, $(TAA)_n$, $(ATT)_n$, $(TTA)_n$ and $(TAT)_n$. As a result, two unique classes are possible for mononucleotide, whereas four for dinucleotide, 10 for trinucleotide and 33 for tetranucleotide repeats. In *L. calcarifer*, mono-, di-, tri- and tetranucleotides are the most frequent consensus patterns of SSR (Table 2).

In the *L. calcarifer* genome, dinucleotide constitutes the largest repeat type in the compound SSRs. All dinucleotide

TABLE 1. SSR analysis based on Tandem Repeats Finder with different percentage matches

SSR	X > 100					
	X ≥ 60	X ≥ 70	X ≥ 80	X ≥ 90	X ≥ 95	X ≥ 100
Mononucleotide	6	6	6	6	6	6
Dinucleotide	149	149	148	130	98	90
Trinucleotide	18	18	18	16	11	9
Tetranucleotide	17	17	16	12	9	9
Others	245	238	218	123	65	37
Total	435	428	406	287	189	151

TABLE 2. Different types of SSRs in the genome of *L. calcarifer*

SSR	SSR type	Repeat number	Copy number (n)
Mononucleotide	A	6	169
	Subtotal	6	169
Dinucleotide	AC	73	1310.5
	AG	8	110
	AT	9	164
	Subtotal	90	1584.5
Trinucleotide	AAC	1	9.7
	AAG	1	10.3
	AAT	5	78.1
	CAT	2	18.7
	Subtotal	9	116.8
Tetranucleotide	AATA	1	7
	ACAA	1	6.8
	AGAT	1	12.8
	ATAG	2	28.6
	ATCC	1	6.3
	CAGT	1	7.3
	TCTT	1	6.3
	TTTG	1	8.8
	Subtotal	9	83.9
Others	- - - -	37	101.3
	Total	151	2055.5

SSR combinations can be further grouped into four unique classes, namely $(AC)_n$, $(AG)_n$, $(AT)_n$ and $(GC)_n$. A total of 73 $(AC)_n$ repeats, which is the most frequent dinucleotide repeat type, were found. Analysis of the *L. calcarifer* GSSs also revealed nine $(AT)_n$ and eight $(AG)_n$ repeats. Interestingly, no $(GC)_n$ repeat was found in the genome of *L. calcarifer*. In most vertebrates and arthropods, $(AC)_n$ repeats are the most frequent dinucleotide repeat motif, while $(GC)_n$ repeats are very rare (Toth et al. 2000). Lower frequencies of $(GC)_n$ repeats in vertebrate genomes have been attributed to methylation of cytosine, which in turn, increases its chance of mutation to thymine by deamination (Schoreeret & Gartlar 1992).

Trinucleotide repeats were also found in the *L. calcarifer* genome with a significant frequency, with AAT as the most frequent repeat type. Trinucleotide repeats are of great interest because of the role they play in many human neurodegenerative disorders and in several human cancers (Arzimanoglou et al. 1998; Reddy & Housman 1997). The alteration responsible for this genetic disease is the expansion of trinucleotide repeats, where the rate of mutation depends on the number of tandem units within the repeat. Hence, analysis of trinucloetide repeats in *L. calcarifer* may be helpful in predicting genes that are predisposed to disease phenotypes.

For mononucleotides, six $(A)_n/(T)_n$ repeats were identified, while no $(G)_n/(C)_n$ repeat was found. The mononucleotides repeats identified tend to be longer than other repeats, with copy numbers from $(T)_{25}$ to $(A)_{34}$. Generally, introns contain more poly(A/T) than poly(G/C) in mamalia, vertebrata and human (Toth et al. 2000).

COMPARATIVE SSR ANALYSIS

Because the complete genome sequence of the *L. calcarifer* was not available, our comparative study was based on random genomic sequences of a similar size comprising of 4,175 *L. calcarifer* GSSs (1,711,858 bp), 2,870 *D. rerio* GSSs (1,711,796 bp), 2,265 *O. latipes* GSSs (1,711,465 bp) and 1,707 *T. nigroviridis* GSSs (1,711,503 bp). A total of 233, 47 and 154 SSRs, accounting for 0.70% (12,029 bp out of 1,711,796 bp), 0.13% (2,181 bp out of 1,711,465 bp) and 0.33% (5,597 bp out of 1,711,503 bp) of the total GSSs were identified in *D. rerio*, *O. latipes* and *T. nigroviridis*, respectively (Table 3). Among the different repeat types, mononucleotide has a minimum copy number of 25 which tend to be longer than other repeats. As in *L. calcarifer*, $(G)_n/(C)_n$ repeats were also not found in *D. rerio* and *O. latipes* but a $(C)_{26}$ repeat was identified in *T. nigroviridis*.

In this study, *D. rerio* was shown to have the most repeats (total repeat length of 12,029 bp) among the fish species compared. This may be attributed to the bigger genome size of *D. rerio*, which is about 1,700 Mb (Postlethwait et al. 1994). The compact genome of *T. nigroviridis* is about 380 Mb (Crollius et al. 2000), and contains only a total of 5,597 bp of repeats, which is much less abundant than *D. rerio*. The estimated size of the *O. latipes* genome is 680-850 Mb (Tanaka 1995), which corresponds to half the size of the *O. latipes* genome and twice the size of the *T. nigroviridis* genome. Surprisingly, only a total repeat length of 2,181 bp was identified in *O. latipes*, which is entirely different from the estimated half of the total repeat length of *D. rerio* (6,015 bp) and

TABLE 3. SSR frequencies in different fish genomes

Organism	Total number of random sequences	Total sequence length (bp)	Total SSR length (bp)	SSRs											
				Mononucleotide		Dinucleotide		Trinucleotide		Tetranucleotide		Others		Total	
				Repeat number	Copy number	Repeat number	Copy number	Repeat number	Copy number	Repeat number	Copy number	Repeat number	Copy number	Repeat number	Copy number
L. calcarifer	4,175	1,711,858	6,217	6	169	90	1584.5	9	116.8	9	83.9	37	101.3	151	2055.5
D. rerio	2,870	1,711,796	12,029	7	205	113	3271.5	27	367.7	40	536.9	46	216	233	4597.1
O. latipes	2,265	1,711,465	2,181	4	122	11	200.5	5	54.7	5	103.9	22	72.1	47	553.2
T. nigroviridis	1,707	1,711,503	5,597	8	219	84	1526.5	14	192.3	8	82.8	40	127.2	154	2147.8

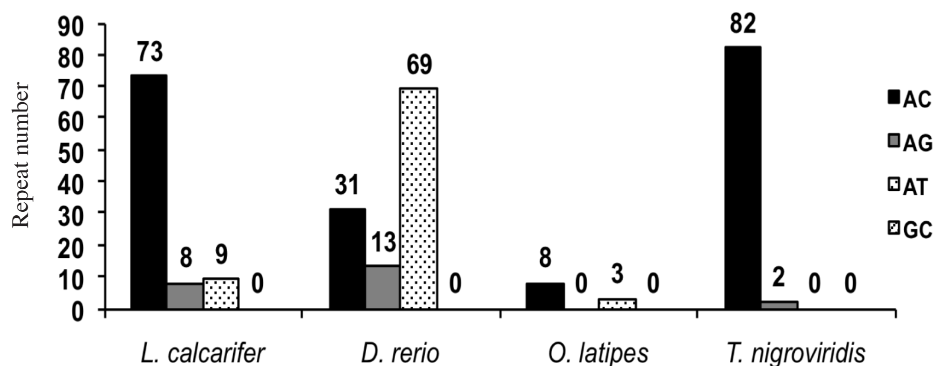

FIGURE 1. Comparison of selected dinucleotide repeats in different fish genomes

twice of the total repeat length of *T. nigroviridis* (11,194 bp). Neff & Gross (2001) found that larger genomes do not possess more microsatellite loci and therefore rejected the hypothesis that microsatellites function primarily to package DNA into chromosomes. Thus, the abundance of SSRs in *L. calcarifer* may not correspond to its genome size because SSR length varies among species.

Our study showed that *L. calcarifer*, *D. rerio*, *O. latipes* and *T. nigroviridis* have a significantly higher frequency of dinucletide repeats compared to other SSR types (Table 3). Figure 1 shows the repeat number of dinucleotide repeats in the different fish genomes. It is evident that *L. calcarifer*, *O. latipes* and *T. nigroviridis* contain more $(AC)_n$ repeats, followed by $(AT)_n$ repeats for *L. calcarifer* and *O. latipes* but $(AG)_n$ repeats for *T. nigroviridis*. In contrast, *D. rerio* contains more $(AT)_n$ repeats, followed by $(AC)_n$ repeats. Interestingly, $(GC)_n$ repeats are not found in all of the fish genomes studied. This is unusual as GC-rich are commonly found in coding sequences in vertebrates and prokaryotes (Oliver & Marin 1996).

It has been suggested that if the nucleotides on the single strand are self-complementary, they can base pair to form loops or hairpins and stabilise strand slippage (Gacy et al. 1995; Moore et al. 1999). If these mechanisms favour repeat expansions or deletions, repeats with higher hairpin propensities such as $(AAT)_n$ and $(CAG)_n$ or self-complementary repeats such as $(AT)_n$ and $(GC)_n$ are likely to be more abundant. However, the various di- and trinucleotide repeat classes within and between different fish genomes do not seem to support such an association. For example, *D. rerio* contains more $(AAT)_n$ trinucleotide repeats (21 repeats) but they are less abundant in *L. calcarifer* (4 repeats), *O. latipes* (1 repeat), and *T. nigroviridis* (none). This suggests that in addition to alternative DNA structures formed by repeat motifs, species-specific cellular factors interacting with them are likely to play an important role in the genesis of repeats (Toth et al. 2000).

CONCLUSIONS

A total of 151 perfect SSRs were identified from GSSs of *L. calcarifer*. These SSRs comprised of consensus patterns ranging from 1 to 226 bp, with dinucleotides as the most frequent consensus pattern present. Mononucleotide copy numbers tend to be longer than other repeats. These 151 SSRs were described for the first time in this species, and analysis of SSRs in different fish genomes provides a snapshot of *in vivo* accumulated repeats. These preliminary data indicate that this SSR collection is a useful resource in the study of the *L. calcarifer* genome organisation, recombination, quantitative genetic variation, gene regulation and evolution.

ACKNOWLEDGEMENTS

The authors thank the Ministry of Science, Technology and Innovation, Malaysia for the financial support of this project under the Top-Down program (Project no. 05-02-02-T002). We also thank the Fisheries Research Institute, Batu Maung, Penang, Malaysia for fish supply.

REFERENCES

Arzimanoglou, I.I., Gilbert, F. & Barber, H.R.K. 1998. Microsatellite instability in human solid tumors. *Cancer* 82: 1808-1820.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.* 27: 573-580.

Chong, P.-P., Mohd-Adnan, A. & Wan, K.-L. 2005. Construction and characterisation of a genomic library for seabass *Lates calcarifer*. *Sains Malaysiana* 34: 26-32.

Chou, R. & Lee, H.B. 1997. Commercial marine fish farming in Singapore. *Aquacult. Res.* 28: 767-776.

Crollius, H.R., Jailon, O., Dasilva, C., Ozouf-Costaz, C., Fizames, C., Fischer, C., Bouneau, L., Billault, A., Quetier, F., Saurin, W., Bernot, A. & Weissenbach, J. 2000. Characterization and repeat analysis of the compact genome of the freshwater pufferfish *Tetraodon nigroviridis*. *Genome Res.* 10: 939-949.

Estoup, A., Rousset, F., Michalakis, Y., Cornuet, J.M., Adriamanga, M. & Guyomard, R. 1998. Comparative analysis of microsatellite and allozyme markers: A case study investigating microgeographic differentiation in brown trout *Salmo trutta* L. *Mol. Ecol.* 7: 339-340.

Ewing, B. & Green, P. 1998. Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Res.* 8: 186-194.

Ewing, B., Hillier, L., Wendy, M.C. & Green, P. 1998. Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment. *Genome Res*. 8: 176-185.

Field, D. & Wills, C. 1996. Long, polymorphic microsatellites in simple organisms. *Proc. R. Soc. Lond*. 263: 209-215.

Gacy, A.M., Goellner, G., Juranic, N., Macura, S. & Mcmurray, C.T. 1995. Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell* 81: 533-540.

Green, P. 1996. *Crossmatch*. [http://bozeman.mbt.washington.edu/phrap.docs/general.html]

Hancock, J.M. 1996. Simple sequences in a 'minimal' genome. *Nat. Genet*. 14: 14-15.

Jackson, T.R., Martin-Robichaud, D.J. & Reith, M.E. 2003. Application of DNA markers to the management of Atlantic halibut (*Hippoglossus hippoglossus*) broodstock. *Aquaculture* 220: 245-259.

Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N. & Petes, T.D. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. *Proc. Natl. Acad. Sci. USA* 96: 1504-1509.

Neff, B.D. & Gross, M.R. 2001. Microsatellite evolution in vertebrates: Inference from AC dinucleotide repeats. *Evolution* 55: 1717-1733.

Mohd-Yusof, N.-Y., Hoh, C.-C., Mohd-Adnan, A. & Wan, K.-L. 2009. Identification of immune-related genes by analysis of spleen expressed sequence tags from the Asian seabass, *Lates calcarifer*. *Sains Malaysiana* 38(6): 939-945.

Oliver, J.L. & Marin, A. 1996. A relationship between GC content and coding-sequence length. *J. Mol. Evol*. 43: 216-223.

Postlethwait, J.H., Johnson, S.L., Midson, C.N., Talbot, W.S., Gates, M., Ballinger, E.W., Africa, D., Andrews, R., Carl, T., Eisen, J.S., Horne, S., Kimmel, C.B., Hutchinson, M., Johnson, M. & Rodrigues, A. 1994. A genetic linkage map of the zebrafish. *Science* 126: 699-703.

Reddy, P.S. & Housman, D.E. 1997. The complex pathology of trinucleotide repeats. *Curr. Opin. Cell Biol*. 9: 364-372.

Schoreeret, D.F. & Gartlar, S.M. 1992. Analysis of CpG suppression in methylated and nonmethylated species. *Proc. Natl. Acad. Sci. U.S.A*. 89: 957-961.

Tanaka, M. 1995. Characteristics of medaka genes and their promoter regions. *Fish Biol. J. Medaka* 7: 11-14.

Tan, S.-L., Mohd-Adnan, A., Mohd-Yusof, N.Y., Forstner, M.R.J. & Wan, K.-L. 2008. Identification and analysis of a prepro-chicken gonadotropin releasing hormone II (preprocGnRH-II) precursor in the Asian seabass, *Lates calcarifer*, based on an EST-based assessment of its brain transcriptome. *Gene* 411: 77-86.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucl. Acids Res*. 17: 6463-6471.

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

Watanabe, T., Fujita, H., Yamasaki, K., Seki, S. & Taniguchi, N. 2004. Preliminary study on linkage mapping based on microsatellite DNA and AFLP markers using homozygous clonal fish in ayu (*Plecoglossus altivelis*). *Mar. Biotechnol*. 6: 327-334.

Wright, J.M. & Bentzen, P. 1994. Microsatellites: genetic markers for the future. *Rev. Fish Biol. Fish*. 4: 384-388.

School of Biosciences and Biotechnology
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
Malaysia

Malaysia Genome Institute
Heliks Emas Block
UKM-MTDC Technology Centre
Unversiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor D.E.
Malaysia

*Corresponding author: email: klwan@ukm.my

Received: 1 June 2010
Accepted: 27 July 2010