MARINE BIOTECHNOLOGY

Development of EST-SSR Markers by Data Mining in Three Species of Shrimp: *Litopenaeus vannamei*, *Litopenaeus stylirostris*, and *Trachypenaeus birdy*

Franklin Pérez,[1] Juan Ortiz,[1] Mariuxi Zhinaula,[1] Cesar Gonzabay,[1] Jorge Calderón,[1] Filip A.M.J. Volckaert[2]

[1]Fundación CENAIM-ESPOL, Km. 30.5 Viá Perimetral, Campus Politécnico, Guayaquil, Ecuador
[2]Katholieke Universiteit Leuven, Laboratory of Aquatic Ecology, Ch. de Bériotstraat 32, B-3000, Leuven, Belgium

Received: 4 August 2004 / Accepted: 8 February 2005 / Online publication: 14 July 2005

Abstract

We report on the data mining of publicly available *Litopenaeus vannamei* expressed sequence tags (ESTs) to generate simple sequence repeat (SSRs) markers and on their transferability between related Penaeid shrimp species. Repeat motifs were found in 3.8% of the evaluated ESTs at a frequency of one repeat every 7.8 kb of sequence data. A total of 206 primer pairs were designed, and 112 loci were amplified with the highest success in *L. vannamei*. A high percentage (69%) of EST-SSRs were transferable within the genus *Litopenaeus*. More than half of the amplified products were polymorphic in a small testing panel of *L. vannamei*. Evaluation of those primers in a larger testing panel showed that 72% of the markers fit Hardy-Weinberg equilibrium, which shows their utility for population genetic analysis. Additionally, a set of 26 of the EST-SSRs were evaluated for Mendelian segregation. A high percentage of monomorphic markers (46%) proved to be polymorphic by singles-stranded conformational polymorphism analysis. Because of the high number of ESTs available in public databases, a data mining approach similar to the one outlined here might yield high numbers of SSR markers in many animal taxa.

Key words: Data mining — EST-SSR — linkage mapping — *Litopenaeus* — population genetics — type I markers

Correspondence to: Franklin Pérez; E-mail: franklin@cenaim.espol.edu.ec

Introduction

Microsatellites or simple sequence repeats (SSRs) are highly polymorphic sequences present in plant and animal species (Toth et al., 2000). By virtue of their codominant nature, SSRs have a wide range of applications including genetic mapping, quantitative trait loci (QTL) association, kinship analysis, population genetics, and evolutionary studies. Most of the markers developed by this approach correspond to type II markers that lack known functions (Weber, 1990). Although their usefulness for genetic analysis has been widely demonstrated, orthodox approaches to their development require considerable investment. Traditionally, SSR isolation has relied on the screening of genomic libraries using repetitive probes and sequencing of positive clones in order to develop locus-specific primers.

Expressed sequence tags (ESTs) are generated by single-pass sequencing of complementary DNA clones obtained by reverse transcription of messenger RNA (Putney et al., 1983). High throughput sequencing generates information on thousands of ESTs, which can be compared with other DNA or protein sequences available in public databases. At the same time the new sequences are made accessible in various databases, increasing the growing information on gene expression. As ESTs are the direct product of gene expression, their analysis leads directly to description of the transcriptome, which is not the case with whole genome sequencing projects.

The use of ESTs as genetic markers can extend their utility beyond gene expression studies. Mouse sequences corresponding to the 5′ untranslated regions have demonstrated the usefulness of EST sequences and single-stranded conformational poly-

 DOI: 10.1007/s10126-004-5099-1 • Volume 7, 554–569 (2005) •

morphism (SSCP) analysis for generating large numbers of polymorphic markers and their use in genetic mapping (Brady et al., 1997). The drawing on ESTs without polarity selection rendered a high numbers of polymorphic markers in *L. monodon* useful for linkage mapping and population genetics studies (Tong et al., 2002). Intron sequences are also highly polymorphic, and the design of primers flanking those areas, based on *in silico* comparisons of ESTs with complete gene sequences available for different species, is possible using ESTs in the target species. This approach has been termed exon-primed intron-crossing (EPIC) polymerase chain reaction (PCR) (Bierne, 2000). Similar to noncoding DNA, EST sequences also contain SSR sequences, which can be used to developed SSR markers (Liu et al., 1999; Whan et al., 2000; Eujayl et al., 2002; Karsi et al., 2002).

A high percentage of publicly available plant EST sequences, (between 1.1% and 4.8%) have SSRs (Saha et al., 2003). Although the percentage of positive clones containing SSRs in nonenriched genomic libraries might be higher, information on ESTs is readily accessible and can be immediately used for development of specific markers known as EST-SSRs. As EST-SSRs are based on exon sequences, which are highly conserved, they are theoretically transferable between taxa. Furthermore, BLAST comparison with protein databases leads to the rapid putative identification of gene function of the EST-derived markers.

The use of molecular markers in shrimp genetics can ensure the long-term sustainability of breeding programs, speed up the genetic gain rate, and lower the costs. Here we report on the development of EST-SSR markers in the shrimps *Litopeanaeus vannamei*, *L. stylirostris* and *Trachypenaeus birdy* (Penaeidae, Crustacea) by data mining. EST-SSRs proved to be an effective approach for the development of transferable molecular markers. We also demonstrate the usefulness of EST-SSRs for population genetics studies and linkage mapping.

Materials and Methods

Data Mining. We downloaded 5832 *L. vannamei* EST sequences from the Marine Genomics repository (*http://www.marinegenomics.org*). Redundant clones were removed using a local nucleotide BLAST search with Bioedit Sequence Alignment Editor Software Version 7.0.1 (Hall, 1999). Short tandem repeats were identified using Tandem Repeats Finder (TRF) software (Benson, 1999) set to report tandem areas with a minimum alignment score of 20 bp (equivalent to finding repeats of 10-bp minimum length) containing mono-, di-, tri-, tetra-, and pentanucleotide repeats. The minimum number of mismatches and indels were 3 and 7, respectively. The results for each positive hit were exported from the individual Web page to a data sheet. Sequences containing poly(A) tails or tandem repeats with less than 30 bases far from the start or end of their EST sequences were excluded from further analysis.

PCR Analysis. Primer design using Primer Premier Software 5.0 (Premier Biosoft International, Palo Alto, Calif.) was carried out for each suitable EST-positive hit. Primers were designed with the default software parameters for a minimum and maximum length of 16 and 18 bp, respectively. PCR amplification for each primer was carried out under the following conditions: $MgCl_2$ 2 mM, 1× PCR buffer (Promega), 200 μM of each dNTP, 0.008 μl of *taq* polymerase per microliter of reaction (5 U/μl, Promega), and 0.4 μM of forward and reverse primer. Each reaction was carried out in 6 μl of PCR mix with 0.75 μl of DNA sample. The PCR reaction was carried out using a touchdown protocol (Don et al., 1991), as follows: initial denaturation at 94°C for 3 minutes, 12 cycles with denaturation at 92°C for 30 seconds, annealing at 55°C for 30 seconds in the first cycle, diminishing 1°C each cycle, and extension at 72°C for 1 minute. An additional 18 PCR cycles were run using the same program with annealing at 43°C and the denaturation and extension conditions as previously indicated. The program was finished with a final extension at 72°C for 1 minute.

PCR products were separated in nondenaturing 6% polyacrylamide gels (29:1 acrylamide-bisacrylamide mix in 1× TBE buffer) in vertical sequencing chambers at room temperature. Band visualization for all products was accomplished by silver staining (Dinesh et al., 1995). Gel documentation was carried out by a digital camera (Olympus Camedia C-5000) in Tiff mode. The picture was transformed to a gray scale and 16 bit mode with Adobe Photoshop 6.0. The Gene Profiler software 4.05 (Scanalytics Inc., Fairfax, Va.) was used for image analysis.

Primary Primer Screening. Primer pairs were initially evaluated in a multi species test panel containing 6 *L. vannamei* (2 parentals of a linkage mapping panel and 4 wild individuals), 2 wild *L. stylirostris*, and 2 wild *T. birdy*. Wild samples were collected along the Ecuadorian coast. DNA was extracted following a CTAB-based protocol (Shahjahan et al., 1995).

Analysis of Genetic Diversity and Mendelian Segregation. Genetic diversity was tested using a set of 16 wild *L. vannamei* collected in Pedernales

Table 1. SSR Motifs Found by Data Mining of *Litopenaeus vannamei* EST Sequences[a]

Motif type	*Number of ESTs*	*Number of different motifs*	*Three most frequent motifs*	*Number of repeats*	
				Min	*Max*
Mononucleotides	69	3	T(66); A(1); C(2)	15	55
Dinucleotides	60	10	AT(14); GT(13);AG(12)	8	143
Trinucleotides	74	30	ATT(10);GCT(8);CTT(7)	5	25
Tetranucleotides	38	27	AAAG(4); ATTT(4); TACA(3)	4	30
Pentanucleotides	43	19	AAAAT(6); AGGTT(5); GTTTT(4)	3	14
Total	284	89			

[a]Data are reported including reverse and complementary SSR sequences without further elaboration.

(00° 05′ N; 80° 06′ W), Ecuador. Samples were DNA extracted with a fat protocol: 400 μl of 5% Chelex plus 2 μl of proteinase K (20 mg/μl), heating at 65°C for 2 hours, boiling for 3 minutes, centrifugation at 12,000 rpm for 10 minutes, and transfer of the supernatant to 96-well plates. DNA was stored at −20°C for 9 months. This set was amplified with a total of 59 primers that showed polymorphism in the initial screening. Expected and observed heterozygosities, and Hardy-Weinberg equilibrium (HWE) were tested statistically by an empirical test (Monte Carlo simulation with 10 batches and 1000 permutations per run) using TFPGA software (Miller 1997).

Mendelian segregation was tested in a small mapping panel comprising both parents and 14 progeny, with the primers showing segregation in the initial primer screening. These DNA samples were extracted by the CTAB method (Shahjahan et al., 1995). A χ^2 test was used to evaluate the segregation hypothesis suggested by the parental genotypes.

SSCP Analysis. Monomorphic markers from the initial screening were amplified in a test panel comprising 14 wild individuals and 2 parentals of a mapping panel. DNA was extracted by the CTAB method (Sbahjahan et al., 1995). PCR samples were loaded with 2 μl of formamide 37% and 3 μl of blue dye, heated at 94°C in a thermocyeler for 5 minutes, and ice cooled. Product separation was carried out in 8% nondenaturating gels (29:1 acrylamide-bisacrylamide mix in 1× TBE buffer) at 10° to 15°C in a refrigerator. Staining, documentation, and gel analysis were accomplished as previously explained.

BLAST Analysis of Amplified Markers. All amplified marker sequences were compared against the GenBank nonredundant protein database using the Web-based HT BLAST Service (Wang and Mu, 2003) (http://mammoth.bii.a-star.edu.sg/webservices/htblast/index.html). All positive hits with scores larger than 60 and e-values lower 1×10^{-10} were included in our report.

Results

Data Mining. Out of 5832 downloaded ESTs 2848 were nonredundant. A total of 475 EST sequences had microsatellite-type repeats, Of these sequences 138 displayed mononucleotide repeats that might correspond to the cDNA poly(A) tail close to the start or the end of the sequence. Fifty-three sequences were eliminated from the analysis because the vicinity of the repeats to the start or the end of the sequence precluded primer design. A total of 284 sequences containing 89 different repeat motifs were isolated (Table 1). The most frequent repeat motifs were trinucleotides, followed by mononucleotides and dinucleotides, respectively. The number of repeats ranged from a minimum of 3 for pentanucleotide repeats to a maximum of 143 for a dinucleotide sequence. A total of 1353 kb of *Litopeanaeus vannamei* EST data was screened for the presence of repeat motifs, giving a frequency of one SSR every 4.01 kb (this calculation includes the 53 ESTs that showed repeats too close to the start or end of the sequence).

Two hundred six primers pairs were designed from the 284 SSR-containing sequences. These results showed that 7.2% of the nonredundant EST sequences had repeats appropriate for primer design.

Primary Primer Screening. Of the 206 designed primers, 112 (54%) yielded PCR products (Table 2). The highest success rate of PCR amplification was observed for *L. vannamei* (105 primer pairs amplified; 2 showing multiple bands), followed by *L. stylirostris* (76 primers; 8 with multiple bands) and *Trachypenaeus birdy* (29 primers; 12 showing multiple bands).

The number of polymorphic markers in the primary screening was high despite the reduced number of individual samples per species. In *L. vannamei*, 56% of the amplified products (59 products) displayed between 2 and 9 alleles, whereas in *L. stylirostris* 32% (24 products) gave between 2 and 4 alleles. In *T. birdy* the percentage of polymorphic

Table 2. EST-SSR Markers and Polymorphism Information Developed from *Litopenaeus vannamei* ESTs in a Small Multispecies Testing panel of *L. vannamei*, *L. stylirostris*, and *T. birdy*[a]

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 332	>2403	ACTGGACTAAGCAAGG GATTTACAACAAGAAGAA	(T)5 C (T)8 G (T)5 A (T)7	196	204(1)	219–237(3)	
CNM-MG 334	>2578	GAGTTCCAATGTAAGTAG AAAATGTAGGTCGGTC	(A)7 T(A)3 G(A)T (A)4 T(A)4	124	129(1)	129(1)	
CNM-MG 335	>3955	AGCCAGGAAGAGGAGG CATCGCCAGAAAGACAG	(GAGC)	112	112(1)	112(1)	112(1)
CNM-MG 338	>4799	TGCTCAAGTCGTTACT GAGGTTTCTGTTCTATAA	(TTTG)4	116	119(1)	119(1)	
CNM-MG 339	>6023	AAACAACATATTGCAGTTC AAGCGTCAGATTCCAG	(ACAAA)4	162	159–191(8)		
CNM-MG 344	>7025	TTACGGGTGAAGTGTT TTTATGCTTCCCTACC	(AC)7	289	309(1)	304–309(2)	
CNM-MG 345	>8364	GAAGTGAGCTTGGCATCCA GTAGAGCAGCGAGCCAGC	(TC)4 CC (TC)5	109	(MB)	(MB)	(MB)
CNM-MG 347	2630	TGATCGCAACAATAAAG GTCGAAGCTGGAAACT	(TGA)6	287	309–316(3)		
CNM-MG 350	>5810	ACAGAAAACCAAGCAA ACGGGATCATAGACAGC	(GT)4 TT (GT)6 TT (GT)2 AT (GT)3 AT (GT)	245	256–276(3)		
CNM-MG 351	>6093	GCAAACAGGAGACAAT CGGACTCTAGCAATAA	(T)20	218	216–227(4)	233(1)	
CNM-MG 354	>7065	AAGACAGAAAGGGTGA CAAGAGGGAGAAAGTAG	(T)15	190	203–214(4)	214(1)	
CNM-MG 355	>7175	TGGCATTCATCTTTGG AAGAGGCACTTCATCC	(AAAT) ATAT (AAAT)2 AATT (AAAT)	262	274–230(2)	275–279(2)	277(1)
CNM-MG 356	>7188	TGCGTTCACATTTCCA AATTGAGTGTCCCTTGC	(GATA) GAGA (GATA)3 GACN (GATA)3	177	180–192(2)		
CNM-MG 357	>7190	GCTTGAATCGCTACTGC GTTGCTGCCACTCATT	(CTG)6 CTA (CTG)3	278	287–290(2)	288(1)	
CNM-MG 359	>7229	TGACAGTAACTCCCAAAT GAATGCAGGAAACATG	(GATT)3	195	204(1)	254(1)	
CNM-MG 362	>2630	TACTTGGACCTCAGTCA GCACGCTTAGTCTCAA	(AAAAC) AAA (AAAAC)2	199	192–224(7)		
CNM-MG 363	>5567	TGCCTAAACCCAAGTC CAGTGGAATATGAAATAAGA	(AT)2 AC (AT)3 GT AA (AT)7	113	121(1)		
CNM-MG 364	>5587	CGTCGTAGTCACAAGAT CAGTATCAATACCGTCCT	(TA)2 TC (TA)7	166	170–173(2)		
CNM-MG 365	>5998	CTTCATACCCATTCTTTCT GCAATAGGCTACAGTTCC	(CTTC)4	300	305(1)	300(1)	
CNM-MG 366	>6145	TCACTTTCCAAATCAAAAC CTAGCAATCTTATTATTACTACC	(AG)3 AA (AG)2 GG (AG)7	196	199(1)		
CNM-MG 367	>6328	AAACCACCCTGACCATC CTGTGCCAAATTACAAGC	(ATTTT)4	284	281–308(9)	256(1)	

(continued)

Table 2. Continued

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 369	>6676	AGCAAGCATTCCTCCTA TTGTGGTCGAACCTAAAC	(T)19	239	251–255(2)	249–251(2)	
CNM-MG 370	>7353	ATAGCGGACCACCTAG CTTCCGTAAATCTTGG	(ACAA)2 A ATAA (ACAA)3	228	239(1)		
CNM-MG 371	>7446	CCAAGAGGGAGTAGAAA GGATAAACACGAAACC	(TA)6 TG (TA) AA (TA)3 T (TA)2	268	292–297(2)		
CNM-MG 372	>7462	TGGATTTGCCGATTGA TCCCAGCACTTGTCATC	(TTA)5	252	265–292(2)	265(1)	
CNM-MG 373	>7527	GATGTCTTATTGGAAA CAGAGCAGATATGGAA	(AAGAA)3	170	177(1)		
CNM-MG 374	>7553	TTGAAAAGCAAAGAAC CTTGGCAGGAGTAGTA	(AT)7	200	209(1)	209(1)	
CNM-MG 378	>2496	AAGGGTGAAGCATAT GTGTTTGGGTTGGTAT	(CA)4 GA (CA)5	199	207(1)	207(1)	291(1)
CNM-MG 379	>2545	GCACGATGGTTCAGTA CCAATGCAAAATACAGA	(TG)3 (TT)2 (TG)2 (TT) (TG)5 TA (TG)4	248	257–260(2)		
CNM-MG 380	>5602	CGAGCGTTATCAAATG GAAATGATGGGGAAGA	(ATT)3 GTT (ATT)5	238	237–260(6)	257–261(2)	
CNM-MG 383	>6156	TTCCTCGTCATTTCAC TGCTTACACCGCCAGA	(TA)2 TG (TA)4 C (TA)4 CA (TA)2	268	247–283(3)		
CNM-MG 384	>6534	ATCGGGAATACAATCG AACCCTAACAAACAATAAG	(AAACA)5	227	227–247(5)		
CNM-MG 386	>6623	CGAGCACAGGAAGATA TCTGGGAGAAGGGATA	(AAAAC)3	257	271–274(2)	273(1)	339(1)
CNM-MG 387	>6636	CAGCTCATACGGAGAC CTTGCGTGAAATTGTT	(AACA)2 TATA (AACA)2 AGCA (AACA)	221	212–223(3)	223(1)	
CNM-MG 390	>7251	CGTAAGATGTGCCAGT CAGTTATAAAGTCAAAAGTA	(TGA)5	248	254–260(2)	254(1)	
CNM-MG 393	>2113	TTTGACGGAATGAGCA GGGGAAATTAGTTAGAGG	(TTTTC) (T)8 (TTTTC)5	267	293–299(3)		
CNM-MG 396	>2518	GTTCTCGAACATGGGA GGGTGATGCAACCTAT	(AAAC)3	295	319(1)	326–335(2)	
CNM-MG 397	>2809	GACTTGGAAGGGAACTG AGAAATAAAGGCTCTATGC	(AGAAAA) AA (AGAA AA)2 AA	100	105(1)	105(1)	101–105(2)
CNM-MG 398	>2880	GGGAAGAATATGTAATG TAACAAGTGCCTGAAA	(AGAA)2 (CATA)5	178	177–197(6)		
CNM-MG 401	>7364	GACATGAGGTATAGCCATTA TATGCACCCTGCTGAC	(TTGT)4	208	212(1)	268(1)	214(1)
CNM-MG 402	>7415	CTTTTGGCTGGCTTAC TTCCTTTTGATCTACATTG	(AGAAA)3	178	187–194(3)		
CNM-MG 403	>7540	TTTCTTGAGAAGGGAG GCAATCTTACATGGTGG	(TAA)5 T (TAA)4	285	299(1)		
CNM-MG 405	>7789	GTGACTGCCCTTTCTACC CTTCCTTGCACGATTTT	(GA)17	251	298–317(4)	320–346(2)	

(continued)

Table 2. Continued

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 406	>7797	GATAAAGAAGCGAGAACG CTATGGCTAGATCCGAGA	(GA)18	256	318–354(8)	333–363(2)	
CNM-MG 407	>2077	GTCTCCTTGCCCGTGTC CGAGTCCGTTGATCCTT	(TTTCT)4	286	293–296(2)		
CNM-MG 408	>2272	ATGTAGTCCTTAACCCATTC GGTCATCAGTCCTGCTCT	(T)16	263		(MB)	
CNM-MG 412	>5818	GCCATTTGATTGCTCT TGACTTGGTCTTTGTTAG	(GT)8	235	236–245(2)	235(1)	
CNM-MG 416	>6631	TGCCAGTGCCATTTGA CCTCCTCCTCCCAACT	(TAT)4 TTT (TAT)2	258	286–288(2)		
CNM-MG 417	>7337	TAAGTTTCCGTAGTCTCA CATCATTATCATTATCGTTG	(ATG)2 GTG (ATG)4 ATA (ATG)2 ATA	205	212(1)	294(1)	
CNM-MG 418	>7393	TAGCCAACGAACAAGC GATTAGTTGATTAGCAGGA	(TAA)6	280	291–295(3)		
CNM-MG 421	>7555B	TTTCTGCCACGGAGTT CTGTTGCCCAAATAGC	(AAT)5	144	148–163(3)	149(1)	
CNM-MG 422		GCAACTATTTATCATCTAAC TTCTGGAAGACTGTGG	(AT)9	153	156(1)	164(1)	
CNM-MG 423	>7572	TTTGATGGGCAAGGAG AGTGGAGTGGCTGGAA	(TAAA)4	257	270(1)	270(1)	
CNM-MG 425		TAACCCAAGCAGAATG TGATCAATGCAAGAAA	(T)15	249	288(1)	286(1)	
CNM-MG 426	>2278	AGGGAGGCTGAGGACG CAATTAGCAGTGTATTATTTCG	(TTC)	205	209(1)	211–217(2)	252–256(2)
CNM-MG 430	>5553	GGGAAGCCCAAATAAGA AAAGAAGAGGAAAGGGATAG	(CT)3 CATT (CT)6 CA (CT)5	199	187–221(9)		
CNM-MG 431	>5616	ATGAAAAGACGAAATG ACGAGCGTTATCAAAT	(TAA)5 TAG CAA (TAA)2	246	248–268(3)	267–271(2)	
CNM-MG 432	>7343	TAGAAGGCAAAGCAGT ATTCTATCACCACCGT	(AAAG)4	275	291–301(3)	284–302(3)	
CNM-MG 433	>7374	TAGATCCCTTCTAGTTTC CTTTAGACAGCCAATT	(AAT)3 ... (AAT)2 AGT (AAT)4	292	317(1)		(MB)
CNM-MG 434	>7390	ACAGGGCAGGACAATA GTTAACTGAGCCATACTTT	(ATTTT)3	237	247 (1)	247 (1)	247 (1)
CNM-MG 435	>7525	CACTGATTGGCTGTTC TACTGCTCCTACTGTTTC	(AAAG)3 AAAA (AAAG)	235	244(1)	240–251(4)	246(1)
CNM-MG 436	>7567	AGAAGTTGCGGCCTAT TACCGAGTTATTCTTGCTG	(TA)10	295	320–331(5)		
CNM-MG 437	>7568	CAACCAGGAAATAGAACAG GCAGCCTTACCACGAC	(CAA)6	135	133–136(2)	230–244(4)	
CNM-MG 439	>7830	TGGCTAGATCCGAGACT CAACATCCCTTCACAAAC	(TC)17	225	291–324(6)	288–335(3)	
CNM-MG 443	>2501A	GAGGCAAGTCAAAGGG TCTGGCGTATCAATGTG	(CCA)2 (CCTCCACC TCCA)2 (CCA)	226	225(1)	225(1)	225(1)

(continued)

Table 2. Continued

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 444	>2501B	CGTACAAGGCATTGGG GCATCTACTTTGACGCACT	(GTT)4	278	274–294(5)	243–265(4)	260–331(2)
CNM-MG 447	>2687	TGATGAGCACCTTGAC CACTAGAGGCTTATACCA	(TACA)2 TAA (TACA)2 ... (TACA)3	240	252(1)	252(1)	252(1)
CNM-MG 450	>6344	ACTGACACCTGCATTG CACAGGCACAGGAATA	(TAT) TGT (TAT)2 ... (TAT) CAT (TAT)3	213	222(1)	244(1)	
CNM-MG 451	>6655	TCCACCATAGCCTCCA CCGCTGCAATGAACCA	(CCA)3 CCT(CCA)3 (TCA)3 (CCA)2	169	313 (MB)	306(MB)	342(MB)
CNM-MG 452	>6739	AGCCCAGCCCCGTGTT TGACAATAAAGCCTGAA	(AAC)2 AGC (AAC)3	489	534–540(2)	524(1)	
CNM-MG 455	>7414	GAGCGTATCTAACCTCA TATGGCTATTGTAACTCTTC	(AAAAT)4	284	307–316(2)	314(1)	
CNM-MG 456	>2157	TTCTTCACATATTGCCCTAC GATTCCGTCGCCAACT	(TTC)6 ATC (TCC)2 GTC (TTC)2	238	252(1)	234251(3)	
CNM-MG 457	>2450	CAATCTTCTGGTGGTTC TATGGCTCGGGTGTAT	(TC)8	243	247(1)	222–236(3)	
CNM-MG 459	>2461	ATCATGTAAGGGTATTTGG CATTTATTCGGCGTTTT	(T)14	136	134–139(3)	133–136(2)	
CNM-MG 460	>4295	TTCCATAATGCTGAATC CTGAGCGAAAGACGAG	(TA)9	134	238(1)		
CNM-MG 462	>5766	AGATACGCTTCCTAATGAT GCTAGTTCCTGCTCCC	(ATG)6	157	192(1)	192(1)	
CNM-MG 463	>5798	AACGCAGCGCAGAAGA TGGAATTGTGAGCGGATA	(AGC)6	268	265–283(5)	273–284(3)	
CNM-MG 465	>6596	AAGTCCAGACAACGAG TAACCTTTAGCAACCT	(TTA)5	256	300(1)	308(1)	
CNM-MG 467	>7413	CTTATTACTACTGCTGCTAG AGGCTGGACTTCTTGT	(TTA)4 TTG (TTA)3 CTA (TTA)	226	229(1)	229(1)	
CNM-MG 470	>7997	AAGTAACTTGGGTGAAA TAGGGCATAGACCATC	(TTTA)6	252			296(1)
CNM-MG 471	>4532	AAGTGTTGCTGGGTATG GCGGACGACAAGGTTT	(GAT) 3 GAG (GAT)3 GAG (GAT) 2	256	270–284(5)	259–272(3)	
CNM-MG 472	>4550	CCCTTCCACCGTGTTG CAGCCTTGCCCTTCTT	(AAAG)2 A (AAAG) C (AAAG)2	196	243–245(2)	250–257(2)	242–253(2)
CNM-MG 474	>5631	CTGGCTTGTGGAATGG CAACGAAAGGCAGATGG	(T)15	195	193–200(2)		
CNM-MG 477	>6643	TGATGATGACGACGATG ATTCTTGGGAGATGTTG	(GAT)6 (GAC)3 ATT (GAT)3 GGT (GAT)	366	353(1)		258–269(3)
CNM-MG 479	>6674B	GTGAAGTTGGGATTATAG CTGCCAGTTTAGCGAC	(T)16	102	96–106(4)		
CNM-MG 483	>7908	ATTTCGCTACATATCATCAC AAGAGGCAATAAGGGT	(T)7 A (T)11	281	294–303(4)		
CNM-MG 484	>8015	TCACCATCGCCAGAAA CCAGGGAAGAGGAGGA	(GCTC)4	115	117(1)	117(1)	

(continued)

Table 2. Continued

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 487	>3977	GACAGACAGTGGTGGCG CGTTCTCCTTGCGTGATG	(GGC)6	297	288–309(7)	653–687(MB)	622–715(MB)
CNM-MG 488	>4104	GCTGAAACGCTCGTCA TGGGCATACTGGGAAA	(AAAAT)3	265	266–620(2 B)	266–620(2 B)	
CNM-MG 489	>4141	GACAGCCACCACGATAAG GCAAGGAGGGACAGGAT	(AACTG)3	233	236–245(2)	222–228(MB)	219–637(MB)
CNM-MG 494	>7935	ACCACTGACTCCCACG CAGGGTCAAAGCAAGA	(AC)5 AG (AC)2 AT GG (AC) (ACGC)3	289	293–315(8)	325–337(4)	
CNM-MG 496	>7993	TGTCACTGTTGAGCCCTACT CAGATTCCTCAGCCTCCT	(TGG)4 CGG (TGG)	203	379–387(5)		
CNM-MG 498	>8181	TTGCTGCTTACTGTCTTGC CATCATCCGACTCTTCCT	(TGG)5	297	719–741(4)	560–566(2)	453(1)
CNM-MG 507	>5040	GATCCCGATGCCGTAGC TGTTTACCAGTTGGGTCCAT	(GCT)4 ACT (GCT)3	228	365–375(3)		
CNM-MG 508	>5063	GCAGCACTACAGGTAA AATTGCACAAGACTTCAT	(TG)4 CA (TG)5 TA (TG) TA (TG)3 TA (TG)2	118	119(1)	120(1)	105(1)
CNM-MG 512	>5625	TGGAAACCTGGCTTGA TCCTAAATACACGGACACT	(ATGT) GTGT AAGT (ATGT)3	240	249–257(4)	421–688(3)	
CNM-MG 514	>8069	TGGAGAAGACTGCCTGAT GATTTAGCCATACCTTTCA	(T)15	298	292–304(4)		
CNM-MG 516	>8230	GTGCCTATGGTGGTTC TCAATTTACTTCTTGGATC	(GCT)6	148		301(1)	
CNM-MG 518	>4876	CACAGTGCGAGATGGC TACCGAACAAGGAATACAAT	(GCAA)4	169	169(1)	167(1)	
CNM-MG 521	>5012	GGGATACAGCAATAAC ATTGGAACAGACAAGTA	(TA)2 TGA (TA)3 TT (TA)5	299	319(1)	315(1)	
CNM-MG 522	>5027	GCCTTTGGTGGTTCTC AATTACTTCTTGGATCCTC	(GCT)6	143		270(1)	
CNM-MG 526	>5175	TCTATAACAACACGTCCACTAG CACGAACGACTTGCTCC	(TTTTG)3	183	186(1)	188(1)	
CNM-MG 527	>5200	TAGCATTGTAGGGTCA CCTTACATTGCCTTTG	(ATT)3 A(ATT) T (ATT)3	188	198–205(2)	202(1)	
CNM-MG 528	>5286	GGGGAGTTGAGCATTG GGTTTACGGCGGAGAA	(CAC)3 CTAC (CAC)2	158			383–468(MB)
CNM-MG 529	>5329	TTGCTGCTTACTGTCTTGC ATCATCCGACTCTTCCTCT	(TGG) 5	296	724–755(4)	555–559(MB)	555–620(MB)
CNM-MG 531	>4282	GTTCTGTTTACAATTGGTTC GAGGAGGACTGAGGGTG	(CAC)5	206	728(1)	483–679(MB)	409–788(MB)
CNM-MG 532	>5026	TTGCGGCAGCGGTAAAGG GACGGCCAAGCCAAGACA	(GGA)3 GGT (GGA)3	108	106(1)	(MB)	(MB)
CNM-MG 533	>5825	CGGGCGGGTACAAGCT GTCGCTGAAAATAATGC	(AGC) 5	137			417–414(2)
CNM-MG 535	>6792	ACACTACAAGCAACCA TCCCAAAATAAACTCA	(T)17	280	334–345(3)		602–751(MB)
CNM-MG 543	>8410	TGAAGCCATTGTCTGT	(ATT)5	263	294(1)	294(1)	

(continued)

Table 2. Continued

Locus	*Entry*	*Primers 5′- 3′*	*Repeat sequence*	*Expected*	*L. vanamei*	*L. stylirostris*	*T. byrdi*
CNM-MG 548	>5851	TTGTGGGAGATGATGC ACAACTTCAAAGCTACA TCAGTCTCATCTTCCAT	(AAAT)4	252	287–303(2)	285(1)	
CNM-MG 554	>7410	TTTAGCTGGGCGACTT GTATGCAGCCTTCCCT	(AAAAG)3	126		927–1466(MB)	1420–1442(MB)
TOTAL				112	106	77	29

[a]In each species the range of the band size and the number of alleles (in parenthese) is presented.
MB indicates multiple bands.

markers was lower, 21% (6 products) with a maximum of 3 alleles.

Two *L. vannamei* samples from the primary screening panel corresponded to the parents of a mapping panel developed in our lab. Twenty-six EST-SSR sequences were polymorphic between those individuals. All those markers were tested for Mendelian segregation as explained below.

PCR amplification of EST-based markers can lead to the amplification of products with sizes different from the expected values, relative to the position of the primers in the original sequence. Sizes larger than expected might occur due to the presence of an intron in the genomic DNA. In Table 3 we summarize the PCR products with markers showing a minimum difference of 50 bp from the expected size product. In *L. vannamei*, 10 of the 109 PCR products showed sizes with 50 or more extra bases than expected. In *L. stylirostris*, 17 of the 79 primers showed unexpected sizes. In *T. birdy* 11 of the 31 amplified products showed differences from the expected size.

Genetic Diversity and Mendelian Segregation. Forty-seven (80%) of the 59 primers evaluated for HWE amplified DNA of 7 or more individuals in the wild animal test panel. Fourteen primers were excluded from the analysis because they showed less than 7 amplifications. Table 4 shows the observed and expected heterozygosities and *P* value of HWE. Thirteen loci showed significant deviations from equilibrium ($P < 5\%$). Average number of alleles per primer was 6.8, with a minimum and a maximum of 2 and 24 alleles, respectively.

Twenty-six primers showing polymorphism between the mapping panel parentals were evaluated for Mendelian segregation (Table 5). Evidence for the presence of null alleles was found for 5 primers (CNM-MG-362, -371, -383, -416, and -487).

SSCP Analysis. Forty-five markers that were monomorphic in the primary primer screening were evaluated under SSCP conditions. A variable number of polymorphic products (2 to 8) were detected in 21 (47%) of the markers. Eight markers were polymorphic between the parents of the mapping panel.

Sequence Identification. Twelve percent of the developed markers ($n = 13$) showed significant similarities with known protein sequences (Table 6). Three of the positive hits corresponded to ribosomal proteins. Eight of the positive hits corresponded to arthropod genes, and 2 positive hits were shrimp antimicrobial peptides of the penaeidin precursor type.

Table 3. EST-SSR Markers Developed from *Litopeanaeus vannamei* EST Sequences Showing Products of Unexpected Size (50 bp or greater difference from expected size) in Three Shrimp Species

		Observed – Expected size difference (bp)		
Locus	*Expected size (bp)*	*L. vannamei*	*L. stylirostris*	*T. birdy*
CNM-MG 359	195		59	
CNM-MG 378	199			92
CNM-MG 386				82
CNM-MG 401	208		60	
CNM-MG 405	251		69	
CNM-MG 406	256	62	77	
CNM-MG 417	205		89	
CNM-MG 437	135		95	
CNM-MG 439	225	66	63	
CNM-MG 451	169	144	137	173
CNM-MG 460	134	104		
CNM-MG 465	256		52	
CNM-MG 472	196		54	
CNM-MG 477	366			−108
CNM-MG 487	297		356	325
CNM-MG 496	203	176		
CNM-MG 498	297	422	263	156
CNM-MG 507	228	137		
CNM-MG 512	240		181	
CNM-MG 516	148		153	
CNM-MG 522	143		127	
CNM-MG 528	158			225
CNM-MG 529	296	428	259	259
CNM-MG 531	206	522	277	203
CNM-MG 533	137			280
CNM-MG 535	280	54		322
Total of putative introns		10	17	11

Discussion

We report the development of EST-SSR markers derived from publicly available EST sequences by data mining: A similar approach has been used in various species of animals (Yue et al., 2001; Rohrer et al., 2002; Yue and Orban 2002; Yue et al., 2004) and plants (Kantety et al., 2002; Gupta et al., 2003; Woodhead et al., 2003; and others). In our initial *in silico* screening, we found a frequency of one repeat every 4.018 kb in the screening of 1353 kb of nonredundant *Litopenaeus vannamei* ESTs. Data mining of EST-SSRs in wheat and barley showed close values with one SSR every 9.2 and 6.3 kb, respectively (Gupta et al., 2003; Thiel et al., 2003). The frequency of SSRs in *L. vannamei* genomic libraries varied according to the motifs and their number between one for every 1.43 kb and one for every 206 kb (Meehan et al., 2003). In *Penaeus monodon* the repeat frequency in two genomic libraries varied from one for every 93 kb to one for every 164 kb (Tassanakajon et al., 1998). The higher frequency of microsatellite-type repeats in EST sequences in *L. vannamei* in comparison with shrimp genomic libraries demonstrates the viability of the approach for large-scale SSR development in shrimp.

The most frequent type of repeats in *L. vannamei* EST sequences corresponded to trinucleotide motifs, followed by mononucleotide motifs (Table 1). Our results are in contrast with reports from genomic libraries in other Penaeid shrimp species in which dinucleotide repeats dominated (Tassanakajon et al., 1998; Meehan et al., 2003; Wuthisuthimethavee et al., 2003). Data on perfect microsatellite motifs in a wide range of eukaryotic genomes demonstrated that the frequencies of mononucleotides and dinucleotides are very similar (around 42%) and outnumber the frequency of trinucleotides in intergenic and intron regions. However, the frequency of trinucleotides in exonic regions (95%) largely surpassed the frequency of mononucleotides and dinucleotides (Toth et al., 2000). In our work we did not find such predominance of trinucleotide motifs. Differences in the data mining methods such as stringency of terms for declaring a microsatellite and the level of tolerance for nonperfect repeats might explain this difference.

SSR isolation in shrimp species has been shown to render variable yields. Pongsomboona et al.

Table 4. *Litopenaeus vannamei* EST-SSR Primer Polymorphism and Hardy-Weinberg Equilibrium in a Testing Panel of Wild Samples[a]

Primer	*Indiv.*	*Alleles*	*Min*	*Max*	H_e	H_o	*P*	*SE*
CNM-MG-339	14	9	150	192	0.86	0.86	0.694	0.009
CNM-MG-347	11	8	300	344	0.67	0.55	0.204	0.012
CNM-MG-350	14	12	230	302	0.88	0.79	0.002	0.002
CNM-MG-351	16	15	212	238	0.92	0.88	0.167	0.013
CNM-MG-354	15	10	200	210	0.84	0.80	0.206	0.007
CNM-MG-355	15	4	274	280	0.62	0.47	0.066	0.007
CNM-MG-356	11	4	180	192	0.55	0.18	0.003	0.002
CNM-MG-357	16	4	308	319	0.41	0.13	0.000	0.000
CNM-MG-362	15	21	189	224	0.94	0.93	0.439	0.016
CNM-MG-364	13	7	166	186	0.75	0.85	0.374	0.017
CNM-MG-367	16	6	285	308	0.82	0.94	0.874	0.008
CNM-MG-369	15	7	251	260	0.78	0.80	0.025	0.005
CNM-MG-371	13	10	284	309	0.86	0.31	0.000	0.000
CNM-MG-372	14	7	261	307	0.66	0.57	0.263	0.014
CNM-MG-379	14	2	256	260	0.48	0.36	0.571	0.013
CNM-MG-380	11	7	236	266	0.76	0.55	0.161	0.007
CNM-MG-383	7	5	273	286	0.72	0.29	0.004	0.002
CNM-MG-384	13	9	226	257	0.87	0.77	0.219	0.012
CNM-MG-386	13	4	273	293	0.33	0.23	0.005	0.001
CNM-MG-387	14	4	217	230	0.70	0.29	0.004	0.002
CNM-MG-390	16	5	259	268	0.53	0.35	0.007	0.003
CNM-MG-402	12	2	188	194	0.41	0.25	0.196	0.008
CNM-MG-405	9	12	269	333	0.88	0.89	0.634	0.020
CNM-MG-406	16	24	286	403	0.94	0.88	0.189	0.012
CNM-MG-407	16	2	290	297	0.06	0.06	1.000	0.000
CNM-MG-412	16	5	243	256	0.50	0.31	0.008	0.002
CNM-MG-416	10	7	294	324	0.80	0.80	0.216	0.010
CNM-MG-418	13	2	287	292	0.39	0.23	0.161	0.012
CNM-MG-421	15	4	145	153	0.24	0.27	1.000	0.000
CNM-MG-430	16	13	194	227	0.89	0.82	0.264	0.014
CNM-MG-431	11	8	247	274	0.80	0.64	0.280	0.006
CNM-MG-436	14	11	309	335	0.88	1.00	0.201	0.015
CNM-MG-437	16	2	133	136	0.17	0.19	1.000	0.000
CNM-MG-444	16	3	278	284	0.55	0.31	0.064	0.008
CNM-MG-455	16	5	303	332	0.60	0.50	0.085	0.007
CNM-MG-474	16	7	189	201	0.58	0.38	0.095	0.008
CNM-MG-479	16	12	96	109	0.85	0.56	0.001	0.001
CNM-MG-483	16	3	296	299	0.17	0.13	0.094	0.010
CNM-MG-487	15	7	287	305	0.80	0.73	0.278	0.015
CNM-MG-489	16	2	237	247	0.22	0.25	1.000	0.000
CNM-MG-494	12	9	290	311	0.74	0.33	0.000	0.000
CNM-MG-496	15	5	380	392	0.68	0.60	0.228	0.015
CNM-MG-498	10	2	717	727	0.10	0.10	1.000	0.000
CNM-MG-507	15	4	360	370	0.54	0.33	0.033	0.006
CNM-MG-512	16	6	210	265	0.71	0.88	0.964	0.007
CNM-MG-527	13	3	199	205	0.42	0.54	1.000	0.000
CNM-MG-548	15	2	280	288	0.28	0.33	1.000	0.000

[a]Number individuals amplified, number of alleles, minimum and maximum allele size (bp), expected and observed heterozygosities, *P* value, and standard error of the exact test for Hardy-Weinberg equilibrium are shown.

(2000) screened a *P. monodon* nonenriched genomic library with trinucleotide and tetranucleotide probes obtaining 79 positive clones and developed 6 polymorphic markers. The success rate from sequencing to polymorphic microsatellites was 7.6%. In *L. vannamei*, 251 positive clones derived from a nonenriched library and screened with di-, tri-, and tetranucleotide probes allowed the development of 93 polymorphic markers. In this case the success rate between positive clones to polymorphic microsatellites was 36.7% (Mehan et al., 2003). Following a similar protocol, Cruz et al. (2002) developed 5 microsatellites out of 68 positive clones with a success rate of 7.4%. In *L. schmitti* Espinosa et al. (2001) report the development of 2 microsatellites from 30 positive sequenced clones,

Table 5. Mendelian Segregation Model and *P* Values for the χ^2 Test in a Set of EST-SSR Markers Evaluated in a *Litopenaeus vannamei* Segregating Panel

Primer	*Model*	*P Value*
CNM-MG-339	1:1:1:1	0.84
CNM-MG-347	1:1	1.00
CNM-MG-351	1:1	0.29
CNM-MG-355	1:2:1	0.30
CNM-MG-362	1:1:1:1	0.01
CNM-MG-379	1:2:1	0.28
CNM-MG-380	1:1:1:1	0.18
CNM-MG-384	1:1:1:1	0.84
CNM-MG-398	1:1:1:1	0.37
CNM-MG-402	1:1	0.29
CNM-MG-406	1:1:1:1	0.46
CNM-MG-418	1:1	0.11
CNM-MG-430	1:1:1:1	0.11
CNM-MG-431	1:1:1:1	0.09
CNM-MG-437	1:2:1	0.48
CNM-MG-439	1:1:1:1	0.46
CNM-MG-459	1:1	0.29
CNM-MG-479	1:1:1:1	0.02
CNM-MG-483	1:1	0.59
CNM-MG-494	1:1:1:1	0.46
CNM-MG-496	1:1	0.11

giving a success rate of 6.6%. Xu et al. (1999) obtained a 12.5% success rate when they developed 10 microsatellites out of 83 *P. monodon* positive sequenced clones. Wuthisuthimethavee et al., (2003) developed 102 microsatellites out of 253 sequenced clones derived from a *P. monodon* enriched library, giving a 40.3% success rate from sequencing to polymorphic markers.

In our work we designed 206 primer pairs out of 282 SSR-containing EST sequences and generated 112 PCR amplifications (Table 2). The percentage of polymorphic markers reached 56%, 32%, and 21% of the amplified products for *L. vannamei*, *L. stylirostris*, and *Trachypenaeus birdy*, respectively. The success rate from designed primers to polymorphic markers was 27% in *L. vannamei*, 11% in *L. stylirostris* and 2.4% in *T. birdy*. However our data on polymorphism from the primary screening should be judged cautiously because they are the product of a small screening panel consisting of 6 *L. vannamei*, 2 *L. stylirostris*, and 2 *T. birdy* individuals.

A theoretical advantage of SSR markers developed from EST sequences is the high transferability between related species. In our research of the EST-SSRs that amplified products in *L. vannamei* 69% gave products in *L. stylirostris* and 21% in *T. birdy* (Table 2). Xu et al. (1999) report that 3 SSRs from a set of 10 SSRs developed in *P. monodon* showed PCR products in *L. vannamei*. Pongsomboona et al., (2000) report weak products obtained in 3 of 6 primers developed in the same species. Ball et al., (1998) showed that 4 of 6 SSRs developed for *P. setiferus* amplified in *P. aztecus*, *P. duorarum*, *L. vannamei*, and *L. stylirostris*. Although transferability of genomic SSR markers in shrimp remains to be tested on a broader scale, we have demonstrated that EST-SSRs give a higher rate of transferability between two closely related species than the genomic SSRs reported to date.

Table 6. *Litopenaeus vannamei* EST Markers with Positive Homologies to Known Proteins Identified from a Sequence Homology Search (BLAST)

Primer	*Protein Accesion*	*Function*	*Probability*	*Score*	*Species*
CNM-MG 365	Q9VXKO	NipSnap protein	5 × 10–23	105	*Drosophila melanogaster*
CNM-MG 369	P29341	Polyadenylate-binding protein	5 × 10–25	113	*Mus musculus*
CNM-MG 390	CAB41634.1	Iron regulatory protein 1–like protein	9 × 10–23	106	*Pacifastacus leniusculus*
CNM-MG 412	NP_501503	Polynucleotide 5′-kinase 3′-phosphatase	6 × 10–30	132	*Caenorhabditis elegans*
CNM-MG 416	P18262	Ras-like protein	6 × 10–23	105	*Artemia salina*
CNM-MG 426/463	Q59296	Catalase	2 × 10–11	68	*Campylobacter jejuni*
CNM-MG 462	NPJ02777	Proteasome α1 subunit isoform 2	8 × 10–20	98	*Homo sapiens*
CNM-MG 474	P81058	Penaeidin-3a precursor	3 × 10–28	124	Litopenaeus vannamei
CNM-MG 496	P02402	60S acidic ribosomal protein	9 × 10–26	114	*Artemia salina*
CNM-MG 512	P81057	Penaeidin-2a precursor	1 × 10–20	99	*Litopenaeus vannamei*
CNM-MG516/522	Q9NB34	60S ribosomal protein L34	3 × 10–25	67	*Ochlerotatus triseriatus*
CNM-MG 528	AAO92284	Putative β thymosin	9 × 10–30	132	*Dermacentor variabilis*
CNM-MG 529	Q29315	60 S acidic ribosomal protein P2	1 × 10–15	80	*Sus scrofa*

In initial primer screening we found that although we had designed primers based on *L. vannamei* EST sequences, 10 SSR sequences did not amplify in our target species but showed PCR products in *L. stylirostris* and *T. birdy* (Table 2). A possible explanation might be the presence of introns that hinder PCR amplification. PCR products amplified in nontargeted species but not in *L. vannamei* show on average products much larger than expected from the original EST sequences. In fact, taking as cutoff values a difference of 50 bp from the expected size, we found evidence for putative introns in 10 *L. vannamei* SSR amplified products. Six PCR products with 50 bp or greater difference from the product expected size that amplified in *L. vannamei* also showed products in *L. stylirostris* (Table 3). Since we did not sequence any of the amplicons obtained in this work, we cannot rule out the possibility that some of the products with unexpected size correspond to different genomic regions than those targeted by the designed primers. However, where introns were amplified, such markers are equivalent to the EPIC markers developed by the design of primers flanking specific intron sequences (Bierne et al., 2000).

High-resolution fingerprinting for population genetic studies requires large numbers of moderately polymorphic microsatellites. Hence we tested the utility of our EST-SSRs, evaluating HWE with 59 primers in a testing panel of wild animals. Those samples were DNA extracted with a fast Chelex protocol and stored for 9 months at −20°C. We used Chelex to select markers suitable for large-scale testing with an easy extraction method that avoids the cost and labor associated with more elaborate extraction methods. From the 59 tested primers, we obtained satisfactory PCR amplifications for 47 primers. The interaction between DNA quality and primers influences PCR amplification (our own observations and Coombs et al., 1999), which might explain the failure in 14 of our markers.

A high percentage of the evaluated primers (72%) did not show significant departures from HWE at the 0.01 *P* value (Table 4). Ball and Chapman (2003) reported a survey in *L. setiferus* in which 5 of the 6 microsatellites showed significant deviation from HWE that might be explained by the presence of null alleles and the Wahlund effect. In a population study in *P. monodon* in the Philippines, 6 microsatellites showed significant deviations from HWE. In this case the presence of null alleles was invoked but also the presence of allele scoring errors and genetic changes in the cultured populations evaluated (Xu et al., 2001). In *L.* vannamei a heterozygosity deficit in 4 of 5 evaluated microsatellites was also explained by the presence of null alleles (Cruz et al., 2002). In contrast, 6 polymorphic loci evaluated in *L. schmittii* gave no deviation from HWE (Maggioni et al., 2003). Although we used a small testing panel, the conformation to HWE and the small standard error of the *P* value of most of our markers points toward their utility for wider use in population genetic surveys of *L. vannamei.*

The number of alleles in our HWE testing panel varied from 2 to 24 (Table 4). When compared with SSR developed from genomic libraries, the EST-SSR level of polymorphism is lower. In other shrimp species SSR allele number varies from one allele (Maggioni et al., 2003; Meehan et al., 2003) to a maximum of 76 alleles (Ball et al., 2003). Some of the evaluated loci corresponded to SSRs with mononucleotide repeats, which can hamper allele scoring in population genetic studies. However, they can be useful for linkage mapping where allele sizes are known from the parental genotypes.

Mendelian segregation of EST-SSRs developed in this research was evaluated for 26 primers. Five primers showed evidence of null alleles in the segregating individuals. All 5 null alleles corresponded to a homozygous parental (4 for the male and one for the female parent) that did not segregate according to the expected model (data not shown). However, assuming the presence of null alleles, all primers might be useful for linkage analysis. As more EST-SSRs are developed and the amplified region is sequenced, the cause of null alleles in shrimp might be clarified.

With EST-SSRs, as with other PCR-based markers, SSCP analysis can disclose polymorphism where conventional polyacrylamide gel electrophoresis (PAGE) fails. This variability corresponds to single nucleotide polymorphism, whereas PAGE unveils length polymorphism. In our work 46% of EST-SSR markers that were monomorphic in the primary screening were found to be polymorphic by SSCP analysis. The presence of 8 markers that showed differential bands between the parentals of the mapping panel points toward the utility of these EST markers for genetic mapping. In *P. monodon* 30% of the EST markers were polymorphic and useful for population genetics and linkage mapping studies (Tong et at., 2002). Our higher rate of polymorphic markers might be explained by the low temperature and the higher polyacrylamide gel percentages, which are known to affect SSCP sensitivity (Humphries et al., 1997).

Thirteen markers showed significant homology with known proteins by BLAST comparison. Tong et al., (2002) found that 23% of *P. monodon* EST sequences corresponded to known proteins, which is

twice the percentage we found for *L. vannamei* ESTs. Because we used close BLAST cutoff values, the reason for this difference is not clear. However, both cases demonstrate the feasibility of using EST sequences in shrimp genetics to produce type I markers.

In this work we have shown the utility of data mining for the development of molecular markers in 3 shrimp species in which type I markers have not been reported previously. EST-SSR and EST-SSCP markers have been developed from publicly available sequences. These markers are highly transferable, at least between the evaluated species, and might prove useful for different research tasks in shrimp genetics. Genetic mapping by AFLPs has demonstrated that the *L. vannamei* genome covers around 4000 cM (Pérez et al., 2004). QTL analysis will require around 300 codominant markers or, alternatively, around 100 codominant markers plus a set of dominant markers in order to cover the genome at a 20 cM average space. The availability of EST sequences in various shrimp species is high in public databases. A line of work similar to the one presented here might render a high volume, in the order of hundreds, of new markers useful for shrimp genetics.

Use of data mining in plant-derived ESTs has identified hundreds of SSR markers in different species (Thiel et al., 2003). Although a fair number of EST-SSRs generated by data mining of publicly available ESTs has been previously reported in swine (Rohrer et al., 2002), animal geneticists have yet to take full advantage of EST data mining where large numbers of molecular markers are in order. Availability of EST sequences for different animal species is high (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). With the use of a new Web-based service for finding repeat motifs and designing primers (http://hornbill.cspp.latrobe.edu.au/cgi-binpub/index.pl) (Robinson et al., 2004), SSR isolation can become a straightforward task. To illustrate this point we examined 1000 ESTs from each of 3 different species (chicken *Gallus gallus*, pig *Sus scrofa*, and Atlantic salmon *Salmon salar*) and generated EST-SSR primers for 6.8%, 8.5%, and 5.7% of the sequences analyzed, respectively. In the specific case of the anadromous *Salmo salar*, whose linkage map comprises 64 markers (Gilbey et al., 2004), by April 2004 there were 87,982 EST sequences deposited at NCBI. Assuming a 1% success rate in marker development, around 900 EST-SSR markers could be tested for polymorphism and linkage with the available EST information. Percentages of EST-SSRs in chicken, pig, salmon, and shrimp are in the same range as those in plant species (Saha et al., 2003), which points toward a rich source of useful information.

The abundance of EST information available gives EST-SSR development by data mining various advantages over conventional development of genomic microsatellites. First, the cost of data mining for EST-SSRs is very low because it avoids the expensive work associated with the initial steps of microsatellite development—namely, library construction and sequencing. Second, as EST-SSR markers are derived directly from gene expression, product identity and function can be identified by comparison with protein databases, generating type I markers. Third, as we and others (Gupta et al., 2003; Thiel et al., 2003) have demonstrated, EST-SSRs are highly transferable across species. Transferability means that the net cost per developed marker will be even lower if they are used for different species. Expression studies using cDNA libraries might be carried out on a main target species, and EST-SSR data mining might be used to generate markers on different species. This approach will integrate transcriptome studies and marker development in a single task and open avenues in linkage mapping, population genetics, and kinship analysis of species for which funding might be scarce. Fourth, although the level of EST-SSR polymorphism might be lower than for genomic microsatellites isolated with conventional methods, the use of SSCP analysis might disclose single nucleotide polymorphism, further increasing the percentage of useful EST-SSR markers.

We conclude that, depending on genome length and EST availability, data mining can generate enough EST-SSR markers for a variety of genetics tasks in many organisms. For new projects, a quick download of ESTs from the species of interest or closely related taxa, combined with the appropriate *in silico* analysis, might save money and months of bench work.

Acknowledgments

We are grateful to Dr. Paul Gross and the research team at Marine Genomics for their work on EST development on *Litopenaeus vannamei*. Without their publicly available sequences, this research could have not been accomplished. This work was carried out with financial support from the Ecuadorian Science and Technology Foundation (FUNDACYT-SENACYT) under Project PFN-084 and the Belgian Technical Cooperation (BTC). We thank S. Sonnenholzner for the facilities provided during sampling of wild animals.

References

1. Ball AO, Chapman RW (2003) Population genetic analysis of white shrimp, *Litopenaeus setiferus*, using

microsatellite genetic markers. Mol Ecol 12, 2319–2330
2. Ball AO, Leonard S, Chapman RW (1998) Charaterization of $(GT)_n$ microsatellites from native white shrimp (*Penaeus setiferus*). Mol Ecol 7, 1251–1253
3. Benson G (1999) Tandem Repeats Finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580
4. Bierne N, Lehnert SA, Bedier E, Bonhomme F, Moore SS (2000) Screening for nitron-length polymorphisms in penaeid shrimps using exon-primed intron-crossing (EPIC)–PCR. Mol Ecol 9, 233–235
5. Brady KP, Rowe LB, Her H, Stevens TJ, Eppig J, Sussman DJ, Sikela J, Beier DR (1997) Genetic mapping of 262 loci derived from expressed sequences in a murine interspecific cross using single-strand conformational polymorphism analysis. Genome Res 7, 1085–1093
6. Coombs NJ, Gough AC, Primrose JN (1999) Optimisation of DNA and RNA extraction from archival formalin-fixed tissue. Nucleic Acids Res 27, e12
7. Cruz P, Mejia-Ruiz CH, Perez-Enriquez R, Ibarra AM (2002) Isolation and characterization of microsatellites in Pacific white shrimp *Penaeus* (*Litopenaeus*) *vannamei*. Mol Ecol Notes 2, 239–241
8. Dinesh KR, Chan WK, Lim TM, Phang VP (1995) RAPD markers in fishes—an evaluation of resolution and reproducibility. Asia-Pacific J Mol Biol Biotech 3, 112–118
9. Don RH, Cox PT, Wainwright BJ, Baker K, Mattick JS (1991) 'Touchdown' PCR to circumvent spurious priming during gene amplification. Nucleic Acids Res 19, 4008
10. Espinosa G, Jager M, García-Machado E, Borell Y, Corona N, Robainas A, Deutsch J (2001) Microsatellites from the white shrimp *Litopenaeus schmitti* (Crustacea, Decapada). Biotec Aplicada 18, 232–234
11. Eujayl I, Sorrells ME, Baum M, Wolters P, Powell W (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104, 399–407
12. Gilbey J, Verspoor E, McLay A, Houlihan D (2004) A microsatellite linkage map for Atlantic salmon (Salmo solar). Anim Genet 35, 98–105
13. Gupta PK, Rustgi S, Sharma S, Singh R, Kumar N, Balyan HS (2003) Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomes 270, 315–323
14. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41, 95–98
15. Humphries SE, Gudnason V, Whittall R, Day IN (1997) Single-strand conformation polymorphism analysis with high throughput modifications, and its use in mutation detection in familial hypercholesterolemia. Clin Chem 43, 427–435
16. Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 48, 501–510
17. Karsi A, Cao D, Li P, Patterson A, Kocabas A, Feng J, Ju Z, Mickett KD, Liu Z (2002) Transcriptome analysis of channel catfish (*Ictalums punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. Gene 285, 157–168
18. Liu Z, Karsi A, Dunham RA (1999) Development of polymorphic EST markers suitable for genetic linkage mapping of catfish. Mar Biotechnol 1, 437–447
19. Maggioni R, Rogers AD, Maclean N (2003) Population structure of *Litopenaeus schmitti* (Decapoda: Penaeidae) from the Brazilian coast identified using six polymorphic microsatellite loci. Mol Ecol 12, 3213–3217
20. Meehan D, Xu Z, Zuniga G, Alcivar-Warren A (2003) High frequency and large number of polymorphic microsatellites in cultured shrimp, *Penaeus* (*Litopenaeus*) *vannamei*. Mar Biotechnol 5, 311–330
21. Miller MP (1997) Tools for Population Genetics Analyses (TFPGA) 1.3: a Windows program for analysis of allozyme and molecular population data. Computer software distributed by the author
22. Pérez F, Erazo C, Zhinaula M, Calderón J, Volckaert F (2004) A sex specific linkage map of the white shrimp, Penaeus vannamei. Aquaculture 242, 105–118
23. Pongsomboona S, Whanb V, Mooreb SS, Tassanakajon A (2000) Characterization of tri- and tetranucleotide microsatellites in the black tiger prawn, *Penaeus monodon*. Science Asia 26, 1–8
24. Putney SD, Herlihy WC, Schimmel P (1983) A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature 302, 718–721
25. Robinson AJ, Love CG, Batley J, Barker G, Edwards D (2004) Simple sequence repeat marker loci discovery using SSR primer. Bioinformatics 20, 1475–1476
26. Rohrer GA, Fahrenkrug SC, Nonneman D, Tao N, Warren WC (2002) Mapping microsatellite markers identified in porcine EST sequences. Anim Genet 33, 372–376
27. Saha S, Karaca M, Jenkins JN, Zipf AE, Reddy O, Umesh K, Kantety RV (2003) Simple sequence repeats as useful resources to study transcribed genes of cotton. Euphytica 130, 355–364
28. Shahjahan RM, Hughes KJ, Leopold RA, DeVault JD (1995) Lower incubation temperature increases yield of insect genomic DNA isolated by the CTAB method. Biotechniques 19, 332–334
29. Tassanakajon A, Tiptawonnukul A, Supungul P, Rimphanitchayakit V, Cook D, Jarayabhand P, Klinbunga S, Boonsaeng V (1998) Isolation and characterization of microsatellite markers in the black tiger prawn, Penaeus monodon. Mol Mar Biol Biotechnol 7, 55–61
30. Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in bar-

ley (*Hordeum vulgare* L.). Theor Appl Genet 106, 411–422
31. Tong J, Lehnert SA, Byrne K, Kwan HS, Chu KH (2002) Development of polymorphic EST markers in *Penaeus monodon*: applications in penaeid genetics. Aquaculture 208, 69–79
32. Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eulcaryotic genomes: survey and analysis. Genome Res 10, 967–981
33. Wang J, Mu Q (2003) Soap-HT-BLAST: high throughput BLAST based on Web services. Bioinformatics 19, 1863–1864
34. Weber JL (1990) Informativeness of human (dC-dA)$_n$·(dG–dT)$_n$ polymorphisms. Genomics 7, 524–530
35. Whan VA, Wilson KJ, Moore SS (2000) Two polymorphic microsatellite markers from novel *Penaeus monodon* ESTs. Anim Genet 31, 143–144
36. Woodhead M, Russell J, Squirrell J, Hollingsworth PM, Cardle L, Ramsay L, Gibby M, Powell W (2003) Development of EST-SSRs from the Alpine lady-fern, *Athyrium distentifolium*. Mol Ecol Notes 3, 287–290
37. Wuthisuthimethavee S, Lumubol P, Vanavichit A, Tragoonrung S (2003) Development of microsatellite markers in black tiger shrimp (*Penaeus monodon* Fabricius). Aquaculture 224, 39–50
38. Xu Z, Dhar AK, Wyrzykowski J, Alcivar-Warren A (1999) Identification of abundant and informative microsatellites from shrimp (*Penaeus monodon*) genome. Anim Genet 30, 150–156
39. Xu Z, Primavera JH, Pena LD, Pettit P, Belak J, Alcivar-Warren A (2001) Genetic diversity of wild and cultured black tiger shrimp (*Penaeus monodon*) in the Philippines using microsatellites. Aquaculture 199, 13–40
40. Yue GH, Orban L (2002) Microsatellites from genes show polymorphism in two related *Oreochromis* species. Mol Ecol Notes 2, 99–100
41. Yue GH, Li Y, Orban L (2001) Characterization of microsatellites in the IGF-2 and GH genes of Asian seabass (*Lates calcarifer*). Mar Biotechnol 3, 1–3
42. Yue GH, Ho MY, Orban L, Komen J (2004) Microsatellites within genes and ESTs of common carp and their applicability in silver crucian carp. Aquaculture 234, 85–98