CHAPTER 4

RESULTS & DISCUSSION

4.1. DNA extraction

The concentrations of the extracted DNA was measured using the NanoPhotometer™,UV/Vis spectrophotometer (Implen, munich, Germany) and the average concentration for 12 samples was 0.045µg/µl. Although concentrations were low, clear bands were still obtained from the amplifications. A probable reason for this would be because a kit was used instead of the conventional extraction method.

4.2. RAMs primers used to construct microsatellite library

A few RAMs primers were tested, however the ones that successfully amplified and gave the highest number of bands are the ones shown in Table 3.

Table 3 List of RAMs primers optimized. Sequence of the primers, annealing temperatures and $MgCl_2$ concentratiosn are shown in the table. RAMs Primers T79113 was designed by (See Leng Min, 2008), BP6, BP13, BP14 were designed by (Hoh, B. P. 2004). PCT6 on the other hand was from this paper (Rodrigues & Kumar, 2009). Repeat regions are in bold. K = G/T, V = G/C/A and R = G/A

RAMs primer name	Sequence 5' to 3'	Annealing temperature (^oC)	$MgCl_2$ concentration (mM)
T79113-7	KDMBMBDMRY$(CCG)_6$	45.8	3.75
BP14	(K_2)BNVSS$(GATA)_5$	49.5	3.75
BP13	$(K_2)(BS)_3B(CT)_6$	53.0	2.5
PCT6	(K_2)BNVSSG$(GATA)_6$	45.8	3.75
BP6	$(K_3)(YBH)_3(AAG)_4$	45.8	2.5

Note: R - A/G, M - A/C, W - A/T, H - A/T/C, V - G/A/C, D - G/A/T, Y - C/T, K - G/T, S - G/C, B - G/T/C, N - A/G/C/T (IUB code).

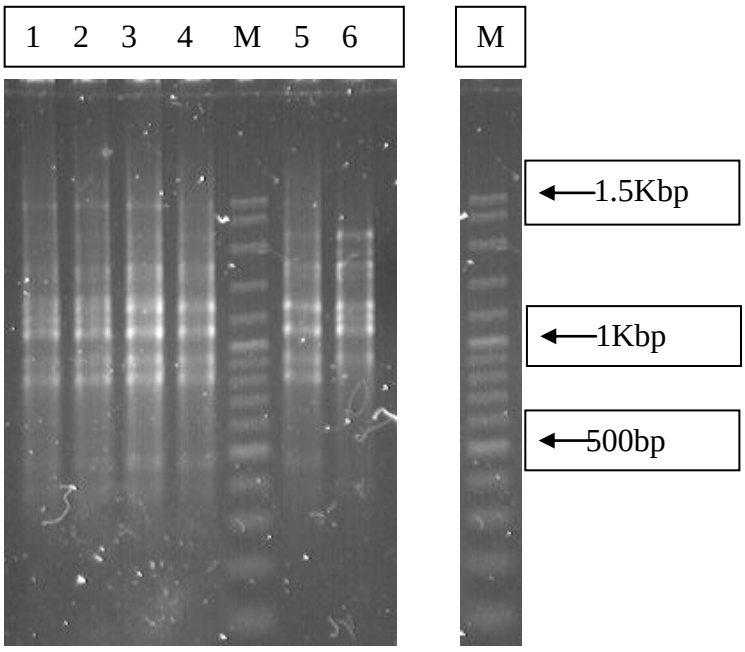

Figure 12 Optimization of RAMs primer BP14. All PCR products were with an $MgCl_2$ concentration of 25mM. Lanes 1-6 are identically prepared PCR products with different annealing temperatures. Annealing temperature for lane 1 is 40^oC and lane 6 is 54.1^oC. The 1% agarose gel ran at 70 volts for about an hour. M: is a 100bp ladder.

4.3. Cloning, sequencing and BLAST upload

RAMs primer products were all cloned into *E. coli* vectors. Several obstacles were encountered while conducting this procedure, such as culture plate contamination. This was resolved by taking extra precaution and using a Bunsen burner in the laminar flow. The issue did not reoccur.

A major problem faced was that after plating the cells, no growth was visible the following day. Even when there was growth, it was very minimal, such as 1 or 2 colonies only. After several attempts to trouble shoot this matter, it was finally concluded that the problem was due to non proper thawing of the competent cells and a probable delay in the transformation step; which should have been done as swiftly as possible.

There are many ways in which the number of transformed positive colonies per cloning procedure could have been increased and in some instances exponentially. This is done by enrichment. The utilization of 5' anchored primers (Fisher, et al., 1996) is just one of the several ways in which a researcher can construct a microsatellite library, in a cheap and highly efficient way. Not to mention that all inserts are practically guaranteed to contain repeats, even though a lot of them could not have primers designed for them; due to the proximity of these repeats to either side of the fragment containing it. Other factors include: low repeat number, or even inability to have microsatellite primers designed that meet the desired criteria. This PCR based method is less time-consuming than other tedious methods that involve library screening (Zane, Bargelloni, & Patarnello, 2002).

Other enrichment methods that are more or less expensive include: Biotin-labeled oligonucleotide attached to streptavidin-coated magnetic particles. The cloned

fragments are hybridized to the oligonucleotide probes (containing repeat sequences) that are attached to magnetic beads. The beads are then subjected to a magnetic field and after a series of washing steps the fragments containing the microsatellites are disassociated from the bead itself. This helps in increasing numbers of microsatellite libraries (Kandpal, Kandpal, & Weissman, 1994; Kijas, et. al., 1994).

Another method is to use different or mutated *E.coli* strains, such as the one used by Ostrander and others (Ostrander, et al., 1992). The team first spliced the genomic DNA of a dog and inserted then into "phagemid vectors". These were then transformed into the main key player of this procedure, a mutant *dut ung* (deficient in **dUT**Pase and **u**racil-**N**-**g**lycosylase enzyme) *E.coli* strain. This strain allows dUTPs to be incorporated in future replications of the vector instead of dTTPs. The absence of **u**racil-**N**-**g**lycosylase enzyme prevents dUTPs from being degraded. Then came the role of M13 helper phage. Once the phage infects bacteria, the DNA is isolated in a single stranded circular manner and subjected to a PCR amplification using primers that contain the desired repeat. Ostrander for instance chose either $(CA)_n$ or $(TG)_n$ (n being the number of repeats). Later the resulting fragments were transformed into a wild type *E.coli* strain. The results were a natural selection of fragments containing the wanted repeats. This is not only because transformation favors double stranded DNA, but also because the wild type *E. coli* strain contains the uracil-N-glycosylase enzyme, that degrades all uracil molecules, making the fragment a lot more vulnerable and subjected to degradation by nucleases (Ostrander, et al.,1992). This method supposedly increases efficiency of transformation to 50 times the normal bacterial strain (Ostrander, et al., 1992).

As a result of plasmid the extraction procedure, a total of 19 plasmids from T79113-7 RAM primer was sent for sequencing. Similarly, 15 were sent from BP6, 15 from BP14, 13 from PCT6 and 14 from BP13. An image of one of the plasmid extractions is shown in Figure 13.

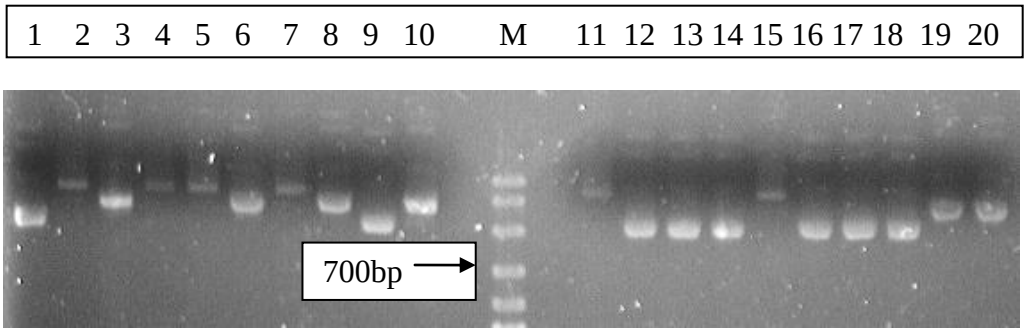

Figure 13 Gel image of plasmid extraction of RAM primer product T79113-7. The 1% agarose gel was run for about an hour or so at 70v. Lane 1-20: are bands resulting from plasmid DNA extraction. Lane M: is 100bp ladder.

When the sequences were received, they were cleared from vector sequences and compared to other mRNA sequences using BLAST and SynaTATE. Results are shown in Figure 14 and Table 4.

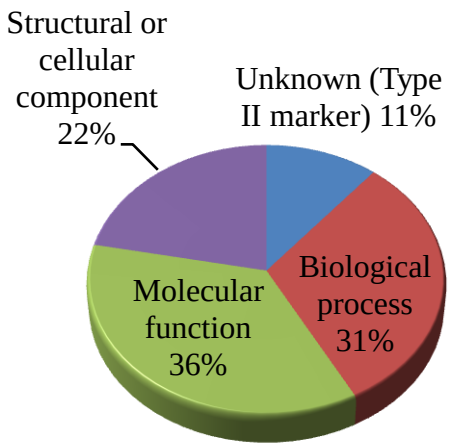

Figure 14 Pie chart showing a statistical overview of the functional genes. Percentages were obtained from the BLAST results' comparison of cloned *C. striata* sequences attained from RAM primers.

Table 4 Functions (or predicted functions) of some of the RAMs primer products after running them through BLAST. These results are only a few of those shown in Appendix A. The ones that actually had lower e-values and predicted functions were of proteins with somewhat known functions.

RAM primer	Species it was compared to	Function according to BLAST	Function according to UniProtKB	Biological process	Cellular component	Ligand	Molecular function	E-value
T79113-B6	Salmo salar (Atlantic salmon)	PREDICTED: Taeniopygia guttata similar to homeobox D4 (LOC100221788), mRNA	Unknown	Transcription regulation	Nucleus	DNA-binding	Developmental protein	1.00E-13
T79113-B3	Pan troglodytes (chimpanzee)	Pan troglodytes transforming growth factor beta regulator 1, transcript variant 3 (TBRG1), mRNA	Acts as a growth inhibitor. Can activate TP53/p53, causes G1 arrest and collaborates with CDKN2A to restrict proliferation, but does not require either protein to inhibit DNA synthesis. Redistributes CDKN2A into the nucleoplasm. Involved in maintaining chromosomal stability. Tumor supressor	Cell cycle	Nucleus			6.00E-10
BP6-C6	Danio rerio	Danio rerio tRNA aspartic acid methyltransferase 1 (trdmt1)	Methyltransferase Transferase	DNA methylation			DNA binding methyltransferase activity	1.00E-08
T79113-B1	Rattus norvegicus (Norway rat	Rattus norvegicus forkhead box E1 (thyroid transcription factor 2) (Foxe1), mRNA	Transcription and regulation of transcription	Negative regulation of transcription from RNA polymerase II promoter transcription from RNA	Nucleus	DNA-binding		6.00E-05

				polymerase II promoter				
BP6-C6	Xenopus laevis (African clawed frog)	Xenopus laevis similar to DNA (cytosine-5-)-methyltransferase 2 (MGC53207), mRNA	Unknown	DNA methylation			DNA binding	0.001
BP14-A4	Danio rerio (Zebra fish)	>ref|NM_194393.1| Danio rerio guanylate cyclase activator 1C (guca1c), mRNA				Calcium	• calcium ion binding • calcium sensitive guanylate cyclase activator activity	0.002
BP6-B3	Danio rerio (Zebra fish)	Danio rerio similar to zinc finger protein 513 (LOC100149174),	May be involved in transcriptional regulation.	• Transcription • Transcription regulation	Nucleus	• DNA-binding • Metal-binding • Zinc	• cellular transcription • regulation of cellular transcription	0.002
BP13-A4	Oncorhynchus mykiss (rainbow trout)	ref|NM_001124218.1| Oncorhynchus mykiss 11-beta-hydroxysteroid dehydrogenase (LOC100135822), mRNA	Oxidoreductase	oxidation reduction			• binding • oxidoreductase activity	0.002
BP14-B3	Danio rerio (Zebra fish)	>ref|NM_212827.2| Danio rerio monoamine oxidase (mao), mRNA	Oxidoreductase	oxidation reduction			oxidoreductase activity	0.01

4.4. Microsatellite primers

All microsatellites screened were perfect microsatellites. The primers designed can be seen in appendix B, along with their repeat motifs, ranging from di- to hexa-. Quite a number of the primers were not successfully optimized, even after multiple tries with a wide range of optimizing temperatures. Other primers did optimize, however they did not show signs of polymorphism.

As seen in the table some primers amplify segments that contain two or more sets of microsatellites. At the time of designing, this was thought of as a good thing, however in the long run and after the locus proved to be polymorphic, slight complications arose. The major issue was that most software that analyzes molecular data, requests the entry of a single repeat, the specific one that the microsatellite primer design was based on, therefore it was hard to distinguish the exact repeat that resulted in polymorphism and caused the basepair change in allele size.

In conclusion it would have been wise to choose microsatellite primers that amplify singular microsatellite repeats; to avoid confusion in future steps of the study. An overall look at the repeats isolated from *C.striata* is shown in Figure 15.

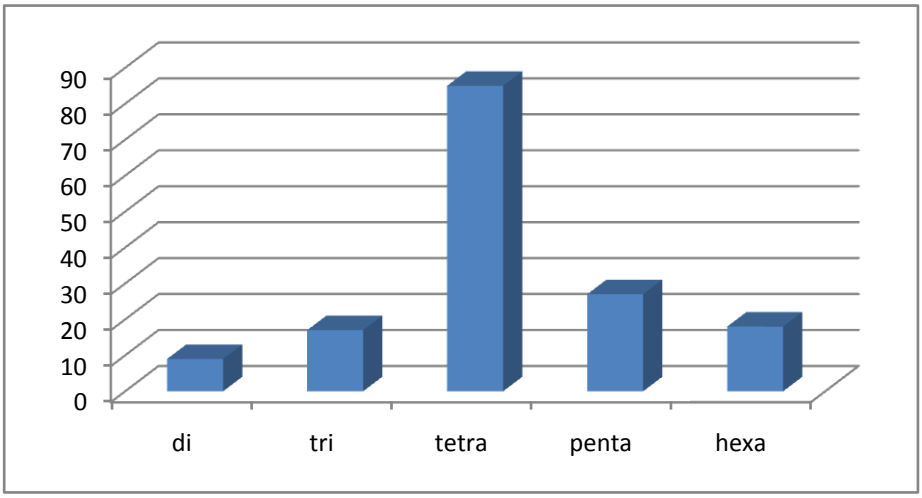

Figure 15 Range of repeats found in *C. striata*. The repeats are shown in numbers and not percentages.

4.5. Polymorphism (PCR products and fragment analyzer)

After testing the primers on a random 32 *Channa striata* individuals, only seven proved to be polymorphic out of the 60 pairs of primers that were optimized and tested. These are primers: T79113-7, T79113-11, BP6-2, BP6-4, BP14-1, BP14-5 and BP14-14. The reason for not testing out all of the 108 sets of primers was mainly time restrictions and the desire to conduct a population study on the microsatellite primers that were already available. A model example of a polymorphic microsatellite primer appears in Figure 16 as gel image and Figure 17 & Figure 18 using fragment analyzer.

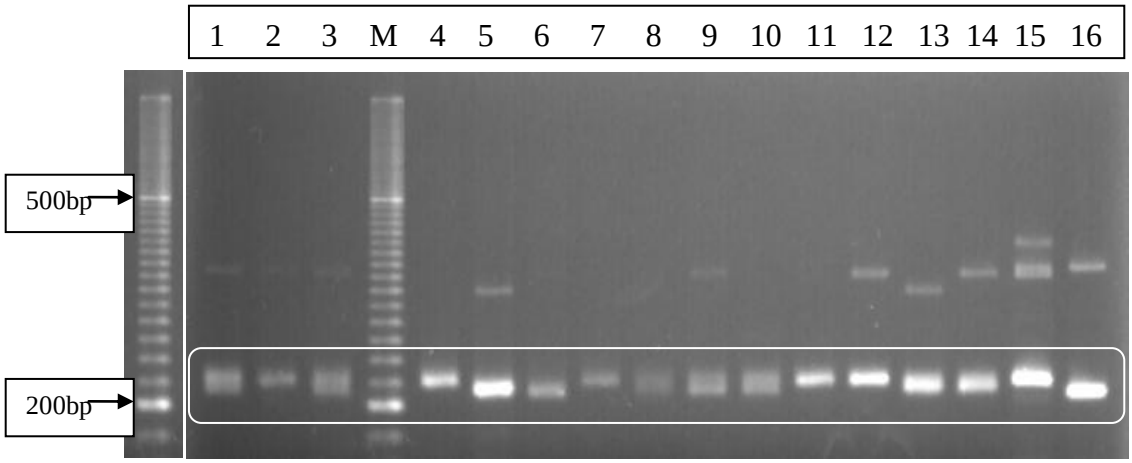

Figure 16 Polymorphism test of *C. striata* DNA using microsatellite primer BP6-2. The 4% (w/v) MetaPhore agarose gel (Cambrex, USA), shows individuals of the fish from Johore (lane 1-14) and Kedah (lane 15-16). As can be seen in the white box, the bands are clearly polymorphic (the difference in size is indicated by the uneven banding pattern). Lane M: is a 20bp ladder.

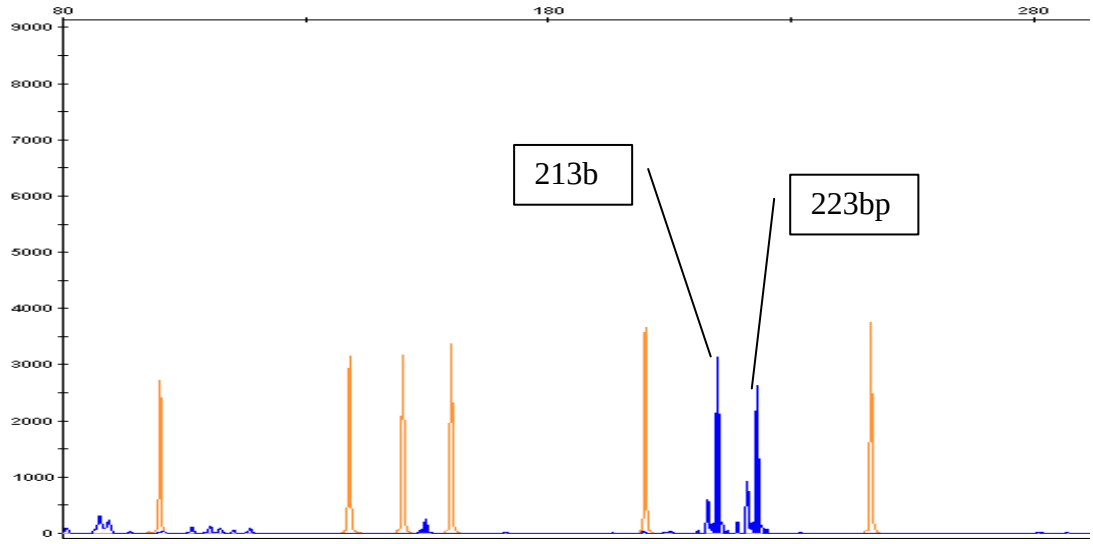

Figure 17 Electropherogram of an individual (ind. 1) from Johore using primer BP6-2. The image shows a graph with two peaks (alleles) of an individual of *C. striata* from Johore. The peak on the left shows the first allele that is 213bps and on the right is a 223bps. The yellow peaks is the side standard. The x-axis indicates the size of the fragment in bps, while y-axis the fluorescence intensity of each fragment**.**

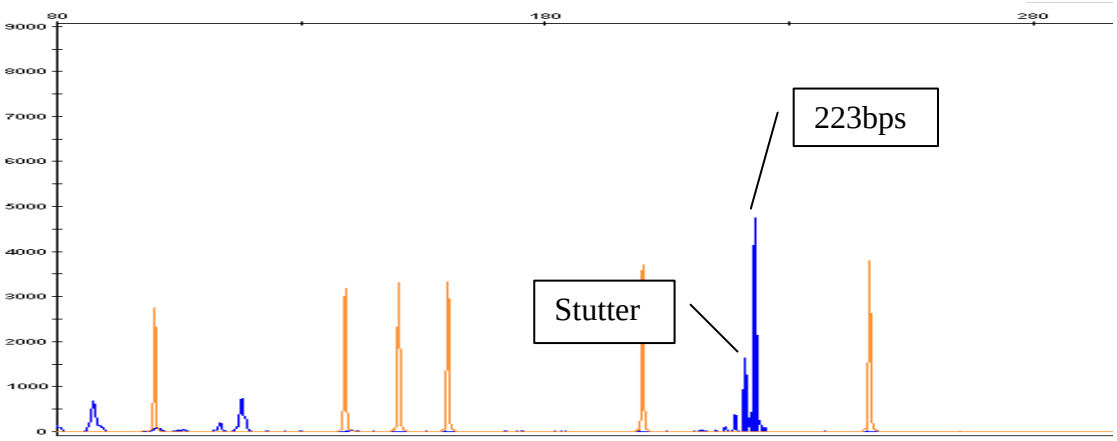

Figure 18 Electropherogram of different a individual (ind. 2) from Johore using primer BP6-2. The electropherogram shows another individual (ind. 2) from the same state that is homozygote for an allele that is 223bps. Clearly when comparing the two, the locus is polymorphic. The yellow peaks is the side standard. The x-axis indicates the size of the fragment in bps, while y-axis the fluorescence intensity of each fragment**.**

Even after testing 60 primers only seven (11.66%) were concluded to be polymorphic. The other 53 were discarded as polymorphic makers for these reasons:

1. Many primers would not anneal no matter what the annealing temperatures and PCR components' concentrations are. These primers were almost impossible to optimize and they were simply ignored. An example can be seen in Figure 19. There are many reasons as to why this happens. A reasonable explanation would be a mutation in the primer annealing site (Chapuis & Estoup, 2007).

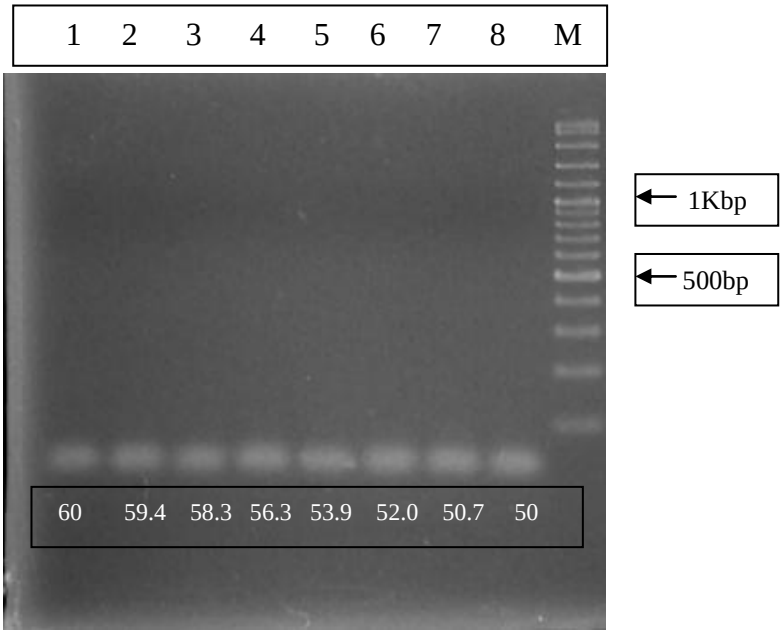

Figure 19. Primer T79113-4 optimization attempt. The gel shows a 4% (w/v) MetaPhore agarose gel (Cambrex, USA), with lanes 1-8 having the same PCR product however with different annealing temperatures, as seen at the bottom, are the annealing temperatures in Degrees Celsius. The bands at the bottom of the gel are primer dimers and not PCR products. Lane M: is a 100bp ladder.

2. Other primers would successfully optimize producing bright and clear bands, however rather a large number of individuals of *C. striata* would not amplify, producing what is called null alleles. Again the reasons for such findings are endless, nevertheless they could be summed up under these: Mutation in the primer annealing site (as mentioned before), could be a probable cause of null alleles (Kwok, et al., 1990). Also poor template quality can be a likely cause, however this is not the case here; since the exact same templates that are showing null alleles, amplified using other primers perfectly, not to mention that the same PCR reaction was conducted a few times with no change in the resulting product. An example of this can be seen in Figure 20.

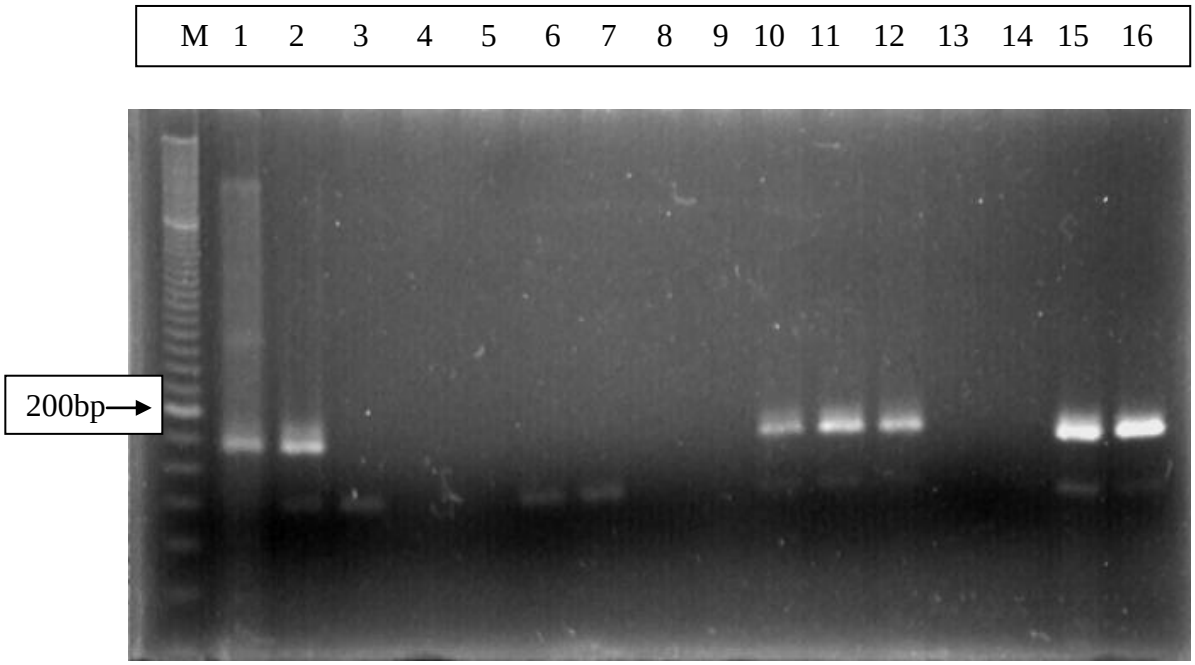

Figure 20 Gel image showing amplification of microsatellite primer T79113-17. This 4% (w/v) MetaPhore agarose gel (Cambrex, USA) was run on 75 volts for at least two hours and it shows the amplification of T79113-17 on individuals from Negeri Sembilan. As can be seen in the figure, lanes 3-9 and lanes 13-14, do not show any signs of amplification (faint bands at the bottom are null alleles). Lane M: is a 20bp ladder.

It has to be mentioned that these PCR products were double checked using the fragment analyzer and true enough there were no peaks what so ever, confirming that there was not even the slightest chance of an amplification which could have been missed when viewing the gel.

3. Another realization is that a few of the amplified microsatellite segments did show banding patterns that had polymorphic characteristics. These were two bands, initially thought to be two different alleles of the same locus; however they would turn out to be actually two different loci that consist of two different homozygote alleles**.** This recognition came through after running the same PCR product on a MetaPhore agarose gel (Cambrex, USA) and then comparing results on an ABI Prism® 3130*xl* Genetic Analyzer (Applied Biosystems, USA). An example of this appears in Figure 21 & Figure 22. In this particular PCR product of microsatellite primer T79113-7, the bands appeared to be polymorphic, this is indicated by the double banding pattern that appears in some lanes in Figure 21 and not in others. So true enough they were labelled and analyzed using the fragment analyzer and as it turns out the top band did not appear at all in the electrpherogram. Only the lower band appeared as two peaks in the graph that have a difference of 8bps, which is large enough to eliminating the possibility of the second peak being a stutter band.

   Luckily enough the second band itself (which is the actual locus that was to be amplified) turned out to be polymorphic, yet it had only a couple of alleles to show for. Therefore the microsatellite primer T79113-7 was still considered polymorphic.

   The band at the top of the gel in Figure 21 appears to be a product of reverse primer amplification. The reason behind this explanation is that the fragment did not show in the electropherogram, meaning that it did not fluoresce; therefore it could not

have been a product of the forward primer (which is fluorescently labeled), or a gene duplication; since the top band's size is only about 40bps larger than the lower fragment.

In conclusion, care must be taken before confirming a primer to be polymorphic. Cutting and sequencing the fragment was considered however this would not be of much use to us, since the band does not appear in all the PCR products, meaning it most likely would give us high numbers of null alleles.

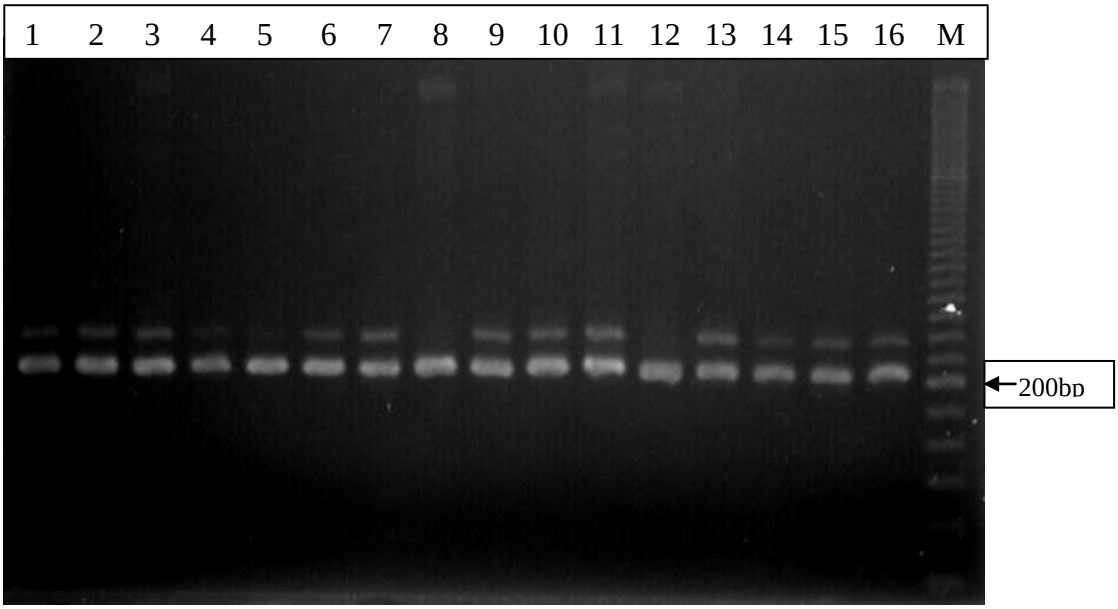

Figure 21 Amplification of T79113-7, showing false signs of polymorphism. The figure shows amplification of Kedah *C. striata* samples using primer T79113-7 on a 4% (w/v) MetaPhore agarose gel (Cambrex, USA). The image shows a double banding pattern from lane 1-16 except for lane 3, 4 and 12. As it turned out the lower band that is bright and rather thick, consists of actually two different alleles and not one. The other top band that appears as a single thin band, turned out to be an amplification that was not a result of the forward and reverse primer. This was concluded after viewing the graph resulting from the ABI Prism® 3130*xl* Genetic Analyzer (Applied Biosystems, USA) shown in Figure 22.

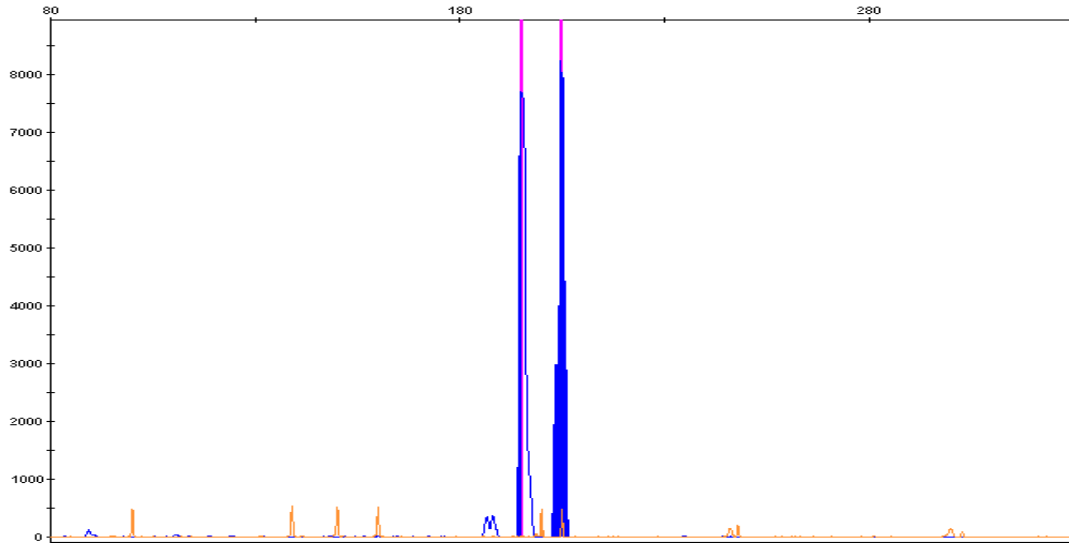

Figure 22 Fragment analyzer electropherogram of a *C. striata* individual from Kedah PCR product of microsatellite primer T79113-7. The graph shows two high peaks that are over scale in blue, which indicate the two alleles of the desired locus that was amplified (the peaks were over scale due to high concentration of PCR product). The peak on the left hand showed a size of 196bps, however the one on the right was 204bps. It is notable that the larger band that was shown in Figure 21, did not appear at all in the fragment analyzer, since it is about 40bps longer that the shorter band, indicating that it was not amplified by the same pair of primers. The yellow peaks are the size standards. On the horizontal line is the fragment size and the vertical axis indicates fluorescence intensity of each fragment.

Even after a primer has been concluded to be polymorphic and testing on individuals of *C. striata* species has commenced, other minor issues arose:

Stutter bands for instance were one of the major issues. This phenomenon occurs due to slippage in the polymerase during the PCR reaction. The shadow band that is produced differs from the actual allele by multiples of the exact number of repeat units (Walsh, Fildes, & Reynolds, 1996). It could be very confusing when interpreting results, since these faint bands that are thought to be stutter bands could actually be alleles that have not amplified intensely in the PCR; this is especially seen in multiplex PCR products

(Walsh, Fildes, & Reynolds, 1996). Stutter happens mostly in di-repeats (Figure 23). Wrongly indicting that a stutter band is in fact what it is, even though it could be an allele, could have serious consequences. It can cause the wrongful sizing of bands, not to mention the affect it has on data analysis, increasing homozygotes. It is thought that the cause of this could be slipped strand mispairing (Levinson & Gutman, 1987), or misannealing of segregated strands during replication (Hancock, 1999). It also can be due to the complexity of the DNA secondary structure (Miller & Yuan, 1997). Studies have proved that the more stringent the repeat the more likely slippage is to occur; this is because the polymerase does not encounter a change in energy profile that is needed for its flexibility (Hummel, 2003), in other word it can cause the polymerase to incorrectly read the nucleotides (Miller & Yuan, 1997).

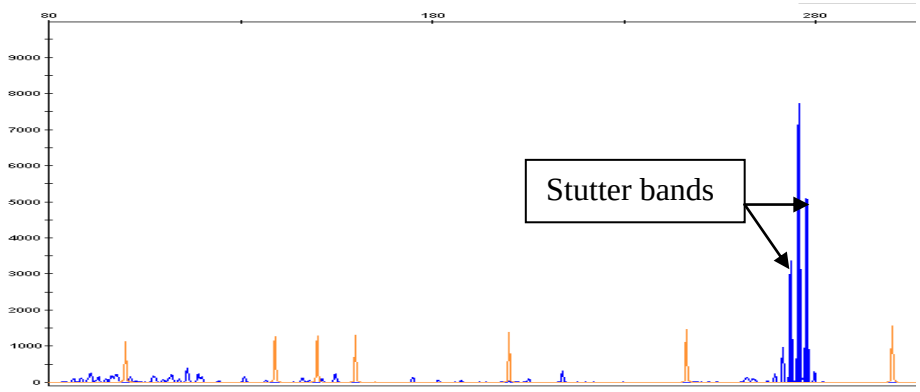

Figure 23 Stutter bands shown in electropherogram of *C. striata* sample from Penang using microsatellite BP6-4. This primer amplifies di- repeats and the peak is actually monomorphic for an allele the size of 276, however a peak with an extra 2 bps is seen on the left and another that is shorter by 2 bps can be seen on the left. Since all peaks are sharing the same base and not standing alone they are concluded to be stutter peaks.

To avoid this, many have resulted to choosing larger repeat units instead; because the numbers of stutter bands are reduced drastically. Others even suggest isolation microsatellites that are interrupted (Walsh, Fildes, & Reynolds, 1996). Some have even resorted to quantitative analysis of the PCR product to determine whether that fragment is true enough an allele instead of a stutter band. This can be done on sequencers, just like in our case where we used ABI Prism® 3130*xl* Genetic Analyzer (Applied Biosystems, USA).

It also said the optimizing PCR conditions and opting for "hot start" methods may reduce or eliminate the occurrence of stutter bands or even eradicate it. The idea is to raise the temperature before the actual PCR reaction initiates in order to get rid of any unspecific binding of primers to the templates that might occur in normal or mild temperatures (Kellogg, et.al., 1994). The protocol is simpler than what it is thought to be. *Taq* DNA Polymerase is withheld, using a heat sensitive material such as wax or a gel temperature sensitive bead from working until temperature is high enough to melt the material withholding the polymerase, then only is the reaction started (Clontechniques, 2008). Articles also have suggested designing microsatellite primers 10-20bps away from the actual repeat, however this was not known in our case when designing the primers, therefore some of the primers were designed to anneal at locations close to the repeat (Scribner & Pearce, 2000).

Overall the polymorphic loci at each population is shown in Table 5. Almost all the populations were polymorphic, except for loci BP14-1, which was monomophic in the Terengganu population.

The ability to use a Genetic Analyzer helped elevate the resolution and precisely estimate the sizing patterns of the bands. Which increased the numbered of polymorphic loci; as high resolution gels would not have given us similar results. Bands that would have had minor size differences (just a few bps), could have been binned together as one allele; decreasing our chances of finding polymorphic loci.

Table 5 Polymorphic loci for each population. It takes just one pair of alleles that is identical in all individuals of a population to consider it monomorphic.

Locus	Johore	Kedah	Negeri Sembilan	Pahang	Penang	Selangor	Terengganu
T79113-7	P	P	P	P	P	P	P
T79113-11	P	P	P	P	P	P	P
BP6-2	P	P	P	P	P	P	P
BP6-4	P	P	P	P	P	P	P
BP14-1	P	P	P	P	P	P	M
BP14-5	P	P	P	P	P	P	P
BP14-14	P	P	P	P	P	P	P

4.6. Data Analysis

4.6.1. Microsatellite Primers

The microsatellite data was run through Micro-Checker (van Oosterhout, et. al., 2004) after sizing them using the GeneMapper® Software Version 4.0 (Applied Biosystems, USA). There were no large allele drop outs, null alleles (Table 6) or stutter bands reported. Stutter was mainly eliminated because of the use of the Genetic Analyzer. Next, the data sheet was converted via Convert software (Glaubitz, 2004), to all the needed formats for further analysis.

Table 6 Null allele estimations obtained from Micro-Checker. No evidence of null allele was found.

Locus	Null Present	Oosterhout	Chakraborty	Brookfield 1	Brookfield 2
T113-7	no	-0.4836	-0.2245	-0.1836	0
T113-11	no	-0.0348	-0.0146	-0.0078	0
BP6-2	no	-0.0353	-0.0221	-0.0176	0
BP6-4	no	-0.0136	-0.0066	-0.0057	0
BP14-1	no	-0.0489	-0.0177	-0.01	0
BP14-5	no	-0.1125	-0.0834	-0.0528	0.1092
BP14-14	no	-0.1094	-0.0545	-0.0181	0.1515

4.6.2. Genetic Variation

The microsatellite primers were characterized as shown in Table 7 and allele numbers for each sampling location is shown in Table 8. The number of alleles ranged from 2 to 17 (with a di- repeat having the most numbers of alleles). Microsatellites markers are usually the method of choice in the last two decades when studying variability and population structure. They usually account for high allelic numbers but this is not in our case. The numbers of alleles were average (with the exception of loci BP6-4); indicating mild variation. This is because fish usually tend to have higher allele numbers than other organisms (Neff & Gross, 2001), whereby loci isolated did not show numbers comparable to that. Johore seemed to be having the highest variability, with a mean number alleles = 4.2857 (Std. dev= 3.9036).

Appendix C shows all the allele frequencies for each locus and their respective PIC values. As would have expected, the primer with the highest number of alleles (BP6-4), had the highest PIC value = 0.929815, making most likely to detect polymorphism in a population. The genotypic frequencies on the other hand are shown in appendix D.

The effective number of alleles in *C. striata* species, do not seem to be very high in most cases, however they do not deviate from the observed number of alleles in an abnormal manner. This tells us that the frequencies of the actual alleles for each loci that we found are not equal, meaning that one allele is usually dominating in most samples tested in a locus. The only way the effective and the observed number of alleles can be equal, is to get the frequencies of each of the alleles to be identical, which is extremely rare, if not impossible (McDonald, 2008).

Table 7 Characteristics of microsatellite primers isolated from *C. striata.*

Locus Gene Bank No.	Primer name	Primer Sequence (5'-3')	Repeat Motif	Annealing temperature (^{o}C)	n_a	n_e	Overall Allelic Richness*	Size range in bp	H_o	H_e
GU323273	T79113-7	F:GCAAAATGTAACTGCTACGA R:GAAAAGGCTTACATCACCTG	$[GCCA]_2$	62	4	1.9617	2.476	196-208	0.7931	0.4915
GU323274	T79113-11	F:CCCTGTATTTCATTTCTCCA R:ACCAACACTGCAATCTCTCT	[CTTT]3	56.9	3	2.6619	3.000	291-299	0.2850	0.6259
GU323263	BP6-2	F:AGAAGAAGAAGAAGCCGAGT R:GAAAAACAGAGCAGGAACAC	$[AGAGG]_2$	52.9	3	2.5439	2.999	213-223	0.6262	0.6084
GU323263	BP6-4	F:TCGAGCTGTGTTTAAGTGTG R:GTTCGTGTTGTTTTCCATCT	$[GT]_{16}$	58.7	17	7.2146	10.907	256-296	0.5825	0.8635
GQ853134	BP14-1	F:ATTACAGCTGTGTGGGTTTC R:CAATGTCGAAGGTTGTAGGT	$[CTGT]_2$	61.4	5	2.1790	3.334	204-220	0.7255	0.5424
GQ853136	BP14-5	F:GAGTTGGTGAATGAGGAAAA R:CCCCATCACACTGAACTATT	$[GTACCT]_2$	47.6	3	1.7610	2.103	274-286	0.6176	0.4332
GQ853139	BP14-14	F:GTCAGAGGGATGTTTAATGG R:AGCAGTTTTTGTTCCACATC	[ATGTAC]2	54	2	1.5403	2.000	254-266	0.4341	0.3516
					Mean=5.2857	Mean=2.8375			Mean= 0.5806	Mean= 0.5595
						St. Dev.= 2.8375				

H_o: Observed Heterozygosity
H_e: Expected Heterozygosity
HWE *(P)*: Overall Hardy Weinberg Equilibrium P-value
n_a: observed number of alleles
n_e: Effective number of alleles
St. Dev: Standard deviation
*: Allelic richness per locus overall the populations, based on minimal sample size of 21 individuals.

Table 8 Numbers of alleles observed in each sample location and loci.

	T79113-7	T79113-11	BP6-2	BP6-4	BP14-1	BP14-5	BP14-14	mean	Std. Dev.
Johore	2	3	3	13	4	3	2	4.28571429	3.9036
Kedah	2	3	3	8	2	2	2	3.14285714	2.193063
Negeri Sembilan	2	3	3	8	2	2	2	3.14285714	2.193063
Pahang	3	3	3	12	4	2	2	4.14285714	3.532165
Penang	2	3	2	9	2	2	2	3.14285714	2.609506
Selangor	2	3	2	9	2	2	2	3.14285714	2.609506
Terengganu	4	3	3	7	2	2	2	3.28571429	1.799471
Mean	2.429	3.000	2.714	9.429	2.571	2.143	2.000		
Std. Dev.	0.787	0.000	0.488	2.225	0.976	0.378	0.000		

St. Dev: Standard deviation

4.6.3. Heterozygosity and Inbreeding

Looking at heterozygosity values in Table 7, it seems as though the observed heterozygosity was rather average, ranging from 0.2850 to 0.7931, with T79113-11 having the lowest heterozygosity and T79113-7 the highest. The overall mean of the observed and expected heterozygosities is 0.5806 and 0.5595 respectively. We would not be expecting extremely high values of heterozygosity because of the low number of alleles.

It is clear that the difference between observed and expected heterozygosities for all loci is very much obvious, except for loci BP6-2, which is approximately the same. This is not so much a concern to us; as the average heterozygosity exceeds 50%, which means that according to the data from the studied loci, most individuals are experiencing heterozygosity excess.

The Inbreeding coefficient (F_{is}) is always associated with heterozygosity analysis; since it is the official withstanding parameter of any non-random mating between relatives, that result in having alleles that are identical by ancestral origin (Hartl, 1980).

In our case (as seen in Table 9), most loci show that our studied populations are experiencing heterozygous access. On the other hand, loci T79113-11 and BP6-4 do in fact show signs of inbreeding. T79113-11 primer's observed heterozygosity was in fact the lowest among all the loci, meaning that the high homozygosities observed have contributed to the inbreeding coefficient results shown in the table.

Table 9 Absolute values of inbreeding coefficient F_{IS}. Values are shown for each locus in every location under study. These values are absolute.

	Johore	Kedah	Negeri Sembilan	Pahang	Penang	Selangor	Terengganu	Average F_{IS}
T79113-7	-0.56757	-0.52632	-0.52632	-0.56941	-0.70588	-0.84615	-0.63636	-0.62269
T79113-11	-0.01270	0.28321	0.47273	0.40761	0.79109	0.42280	0.30000	0.38927
BP6-2	-0.02837	-0.14439	-0.72216	-0.01754	-0.18367	-0.16279	-0.52632	-0.26822
BP6-4	0.00373	0.21903	-0.00520	0.37340	0.20487	-0.02686	-0.05608	0.12258
BP14-1	-0.01903	-0.17391	-0.45000	-0.59146	-0.87097	-0.76380	-1.00000	-0.59966
BP14-5	-0.16493	-0.41463	-0.52632	-0.87097	-0.60000	-0.38889	-0.13725	-0.48155
BP14-14	-0.09804	-0.26087	-0.29396	-0.45000	-0.17681	-0.19048	-0.20833	-0.26100

Nevertheless when looking at overall F_{is} (AMOVA) values according to population Table 10, the populations do not seem to be undergoing inbreeding. Therefore results obtained earlier were probably due to sampling error (low sample sizes).

Table 10 Overall population F_{IS}. Values are calculated here using AMOVA and it shows population specific F_{IS} indices (1023 permutations).

Loci	F_{is}
Johore	-0.13818
Kedah	-0.18987
Negri Sembilan	-0.45063
Pahang	-0.28305
Penang	-0.36238
Selangor	-0.44674
Terengganu	-0.45487

4.6.4. Hardy-Weinberg Equilibrium (HWE)

When random sexual mating occurs in a diploid organism within a relatively large population, given that no overlapping of generations or natural selection occurs in the studied loci, then allele frequencies of the studied loci in a population are said to abide to the Hardy-Weinberg Law (Hardy, 1908; Weinberg, 1908; Hartl, 1980). In other words if allele frequencies are the same in male and females, their frequencies are to remain constant throughout the generations.

In Table 11, the P-values of HWE for each loci and location under study, is shown with significant values after basic Bonferroni adjustment (Rice, 1989) in bold; to make the significance more stringent. As indicated in Table 11, values in bold were all significantly deviating from HWE. Most of the loci (57.1%) accept the null hypothesis, indicating that gametes are randomly associating most of the time.

Alternatively, when looking at individual loci, T79113-7 seems to be falling out of equilibrium. This is a logical finding, given that it shows twice as much heterozygosity as would have been expected, not to mention the constant outbreeding values it had demonstrated in the inbreeding test. Clearly it has to assured that this by no means draws a conclusion as to what might be going on with species, however it does indicate that this locus is at a high recombination location, limiting its ability to pin point other possible happenings to the species, such as inbreeding and genetic drift.

On the other hand we have BP14-14, is the opposite T79113-7. It has a low number of observed alleles which is the minimalist of all (only two). Yet, its heterozygosity is about 43% which is mediocre. In addition to that there was no inbreeding observed, in fact mild outbreeding was detected. This questions the integrity of the primer whether it

is actually powerful enough to apprehend even the most obvious of changes to the genetic content as a whole.

When looking at the overall results this far, the high heterozygosities and the dismissal of inbreeding as an effect on the mating habits of the individuals, could mean that the Haruan individuals in almost all populations are not at a critical genetic standpoint. Nevertheless more statistical tests will be needed to confirm that argument.

Table 11 Probability values of HWE for each location under study. Values in bold are the ones deviating from HWE after basic Bonferroni adjustment.

	T79113-7	T79113-11	BP6-2	BP6-4	BP14-1	BP14-5	BP14-14
Johore	**0.0018**	0.0554	0.5204	0.2493	0.0633	0.7288	1.0000
Kedah	**0.0044**	0.1891	**0.0007**	0.0115	1.0000	0.0304	0.2909
Negeri Sembilan	**0.0044**	**0.0000**	**0.0000**	0.4012	0.0280	**0.0044**	0.2007
Pahang	**0.0011**	0.0159	**0.0042**	**0.0000**	**0.0000**	**0.0000**	0.0280
Penang	**0.0001**	**0.0000**	0.5628	0.0139	**0.0000**	**0.0013**	0.6377
Selangor	**0.0000**	0.0284	1.0000	0.4686	**0.0002**	0.0665	1.0000
Terengganu	**0.0000**	0.0774	**0.0001**	**0.0059**	**0.0000**	1.0000	0.5511

All values in bold are significant after Bonferroni correction ($P<0.007143$), whereby the α value= 0.05 and k=7.

4.6.5. Linkage Disequilibrium (LD)

On the other hand linkage disequilibrium has to be taken into consideration, because allele frequencies can also be affected if genes were to combine in a nonrandom fashion. It is safe to say that all loci do not necessarily exist independently; in fact some can be closely associated to others in a non random manner. It is necessary in all case studies dealing with multiple loci to evaluate this matter, before proceeding with assumptions and conclusions about a population.

In our case only two pair of loci showed signs of linkage disequilibrium, with a P-value that was significantly below 0.05, those loci were: T113-11 & BP6-2, with P-value= 0.013529 (shown in Appendix E).

The other pair that was almost statistically definite to be associated was BP6-2 and BP6-4, with a chi square value of infinity. This does not come as a surprise since the loci were actually amplified by the same RAM primer and the microsatellite primers were designed for both loci from the same amplified fragment that was cloned and sequenced. Therefore these two so loci are definitely linked. Therefore the question is: how much does this affect our statistical data? And could T113-11 be associated to the other BP6 loci?

When a pair of loci are suspected to be at LD (just as in T79113-11 & BP6-2), it means that there is a chance that there might be non random mating or association between these two loci, however our study involved a maximum of 30 individuals per population, therefore random genetic drift would probably play a key role in this finding and LD would extremely hard to certify. Another hypothesis would be that there could be fractionation in the population, which is decreasing the number of individuals in the sub-populations. Mutation and other factors are also possible causes, but this is highly

unlikely; since recombination occurs more often than mutation, jumbling the alleles and eliminating most non-random associations (Bio. Dept. UC Santa Cruz, 2005).

In conclusion it will be hard to infer that the two loci T79113-11 and BP6-2 are in fact in LD, since we have no proof that they exist on the same chromosome or even linked throughout the generations in many individuals, without the need to be on the same chromosome. Therefore more evidence will be needed; since the result obtained may be merely a statistical figure, with no truth to its holdings whatsoever. So until that could be proven, the concept will have to be put on hold (Durand, et al., 2003).

4.6.6. Genetic Distances

As mentioned earlier in the literature review, microsatellites arose from mutations and they continue to evolve and change in repeat number in the same manner. Regardless of the type of mutation model, these changes cause genetic similarities to occur, or divergence between individuals in different populations. Now these differences can be measured as distances for genetic studies. One of the first to attempt this was Nei (Nei, 1972; Nei, 1978). Who designed an algorithm and modified it later on to give an estimate of the standard genetic distance between any two populations or subpopulations.

The genetic distances were calculated using both Nei's 1972 and 1978 unbiased genetic distance. This is because some tree implementing software use either one of two methods of calculation. For instance PHYLIP uses Nei's 1972 (Nei M. , Genetic distance between populations, 1972) only, however GDA is able to implement tree for both Nei's 1972 and 1978.

Table 12 shows the genetic distance calculated using Nei's 1972 genetic distance (Nei M. , Genetic distance between populations, 1972); this was done using software GENDIST in PHYLIP. Another table was also constructed using GDA software (Table 13); in order to generate a UPGMA tree with a genetic distance scale, this table was also based on Nei's 1978 unbiased genetic distance and identity.

According to GENDIST (using Nei, 1972), the lowest genetic distance calculated was 0.058350 between Penang and Selangor populations. The highest genetic distance however was recorded as 1.070289 between Negeri Sembilan and Selangor populations.

On the other hand using GDA (using Nei's 1978) demonstrated that the lowest genetic distance was recorded between the Penang and Pahang populations at D = 0.0145069.

The highest was between the Negeri Sembilan and Selangor populations (D = 0.508099).

The differences in the distance values between both tables are far from negligible, yet they draw almost to the same conclusion, these will be much more obvious once the trees are constructed.

Although the distances in reality and geographical allocations do not relate to the genetic distances that are seen in the table, it is not unusual in our case, since the number of loci and individuals studied are not sufficient enough to give us expected results. According to Nei in order to avoid sampling error and unreliability of heterozygosity, it is ideal to test at least 50 loci in a species genome. If impossible, then increasing the number of sampled individuals would be wise (Nei M. , Estimation of average heterozygosity and genetic distance from a small number of individuals, 1978).

Table 12 Nei's biased Measures of genetic distance using GENDIST from the software PHYLIP (Nei M. , Genetic distance between populations, 1972).

Population	Johore	Kedah	Negeri Sembilan	Pahang	Penang	Selangor	Terengganu
Johore	****						
Kedah	0.256075	****					
Negeri Sembilan	0.776544	0.816527	****				
Pahang	0.153371	0.236339	0.608045	****			
Penang	0.259296	0.148705	0.958368	0.067476	****		
Selangor	0.351979	0.156334	1.070289	0.175975	0.058350	****	
Terengganu	0.155700	0.192792	0.811106	0.167045	0.186507	0.275179	****

Table 13 Nei's Unbiased Measures of Genetic Identity and Genetic distance using GDA v. 1.0 (Nei M. , Estimation of average heterozygosity and genetic distance from a small number of individuals, 1978).

Population	Johore	Kedah	Negeri Sembilan	Pahang	Penang	Selangor	Terengganu
Johore	****	0.765246	0.719968	0.895758	0.853914	0.823719	0.918432
Kedah	0.267558	****	0.623110	0.850332	0.869091	0.868906	0.878089
Negeri Sembilan	0.328549	0.473032	****	0.731619	0.634351	0.601638	0.700091
Pahang	0.110085	0.162128	0.312495	****	0.985598	0.930537	0.913083
Penang	0.157925	0.140307	0.455153	0.0145069	****	0.985332	0.903943
Selangor	0.193926	0.14052	0.508099	0.0719939	0.0147771	****	0.875528
Terengganu	0.085087	0.130007	0.356545	0.0909289	0.100989	0.132928	****

Note: Nei's genetic identity (above diagonal) and genetic distance (below diagonal).

A UPGMA consensus tree was constructed using Nei's unbiased genetic distance through PHYLIP (v.3.67), however the software based the tree on Nei's 1972 standard genetic distance and not on 1978. The figure shows that almost all the clusters shown (except for Kedah population, which clusters in the shown manner in about 46% of the 1000 UPGMA trees constructed) appear in at least 50% of the 1000 trees that were constructed to deduce the consensus. This makes the tree highly reliable. As shown in Figure 24, the population pairs that clustered together were Selangor and Penang, as well as Johore and Terengganu populations. The rest of the locations cluster as shown in Figure 24.

However what stands out strikingly is that the Negeri Sembilan population does not cluster at all with any other location. This is rather peculiar; because its geographical location falls right in the middle of both Selangor (which lies on the northern side of N. Sembilan) and Terengganu (which is allocated on the region's southern part). There are no geographical barriers or barricades that might hinder the mixing and mating of the freshwater fish, especially since it is an actively moving fish unlike anything of its kind.

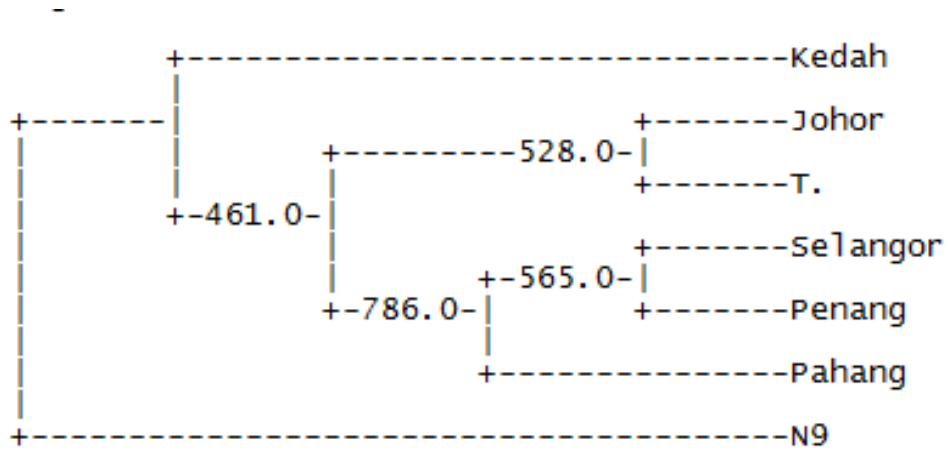

Figure 24 Majority rule consensus tree (UPGMA) resulting from a summerization of 1000 trees. Numbers indicate the times in which the particular branch reoccured in the process of creating 1000 trees (similar to the bootstrapping method). N9: Negeri Sembilan, T.: Terengganu

Another tree was constructed using GDA with a scale of Nei's 1978 genetic distance calculated using the same software (Figure 25).

The constructed GDA tree is slightly different than the consensus in Figure 24, probably because the consensus was based on Nei's 1972 genetic distance. Nevertheless it still shows almost the same clustering, but this time Penang and Pahang populations seem to cluster first and then the Selangor population. This nevertheless does not change the overall view of the tree.

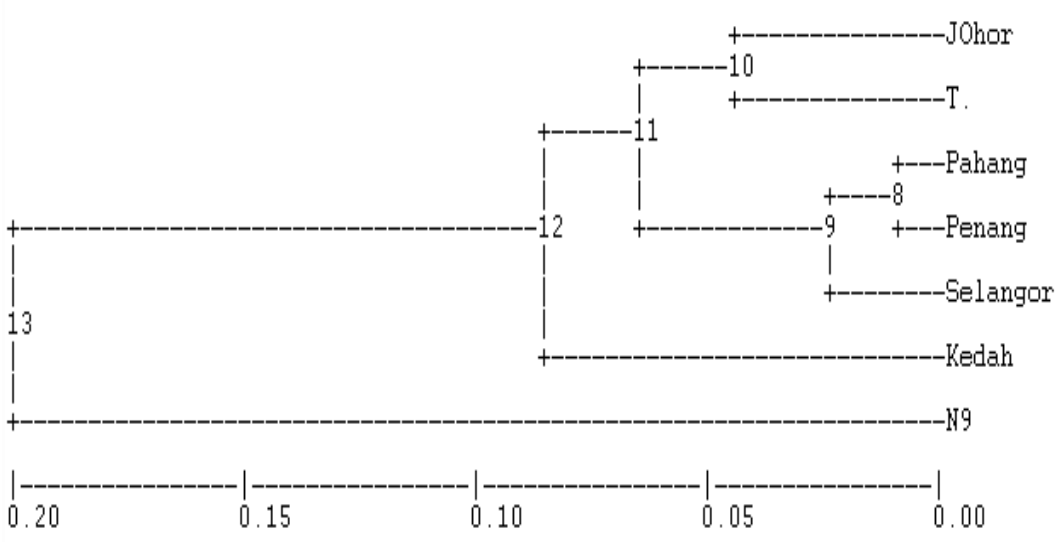

Figure 25 UPGMA tree constructed using GDA software based on Nei's unbiased Genetic distance. The numbers shown on the figures represent the step in which the node was created. The seven locations were added first and then the nodes were calculated for every two locations and placed on the tree, connecting each of the locations.

4.6.7. Genetic differentiation

The measurement of divergence (or in other words differentiation) between individuals of sampled populations is measured using Fixation index (F_{st}) (Hartl, 1980). Diversity is a factor that should be maintained constantly in a species, and throughout the timeline of its existence. This relives the organism from diseases and accumulation of mutations that might directly affect the individual itself or its progeny. A more drastic measure, would be the extinction of the species (Corti, *et. al.*, 2010).

Pairwise F_{st} values were all calculated and shown in Table 14. The values seen are somewhat high on the most part. The highest was found to be between Negeri Sembilan and Selangor populations F_{st} = 0.28838 (P-value of= 0.00000). The lowest F_{st} on the other hand was between Pahang and Penang populations, with a value of 0.00332 **(**P-value of= 0.02703±0.0139).

Overall, results with high F_{st} values indicate that there is a reduction in heterozygosity between the populations paired, compared to the overall total heterozygosity of the *C. striata* population in this region (Hamilton, 2009).This is reflected mostly when looking at the F_{st} values of populations such as Negeri Sembilan. When the Negeri Sembilan population is paired with other populations, the F_{st} values tend to be unusually high.

Other individuals from states such as Pahang population have had somewhat low F_{st} values when compared to Haruan from other states of the region (except for the Negeri Sembilan population). This indicates that the individuals from Pahang have probably maintained their variability across the generations and probably have heterozygosity levels that are very similar to the total heterozygosity of all locations.

F_{st} values are very closely related to genetic drift, the act of randomly choosing gametes in particularly small populations. This effect causes changes in the frequencies of alleles

under study. Genetic drift is almost always associated with minor inbreeding habits that are usually very mild. In our case that was excluded earlier on because F_{is} values indicated there were no signs of that (most values were below zero). However, according to Hartl, F_{st} values can vary according to the organism. For instance Humans have an F_{st} of 0.069, on the other hand jumping rodents (*Dipodomys ordii)* have an F_{st} value of 0.676 (Hartl, 1980).

Finally other elements might also contribute to the obtained results such us: effective population size, migration (this fish is not migratory, however it is known to maneuver and make its way through muddy lands) and natural selection (Hartl, 1980).

Table 14 Pairwaise F_{st} values according to ARLEQUIN. In between brackets are the P-values (Number of permutations: 110)

Population	Johore	Kedah	Negeri Sembilan	Pahang	Penang	Selangor
Kedah	0.20749(0)					
Negri Sembilan	0.14448(0)	0.24941(0)				
Pahang	0.09283(0)	0.12529(0)	0.18303(0)			
Penang	0.13970(0)	0.13963(0)	0.23572(0)	0.00332 (0.02703±0.0139)		
Selangor	0.18124(0)	0.15859(0)	0.28838(0)	0.06601(0)	0.02556 (0)	
Terengganu	0.09473(0)	0.11633(0)	0.17563(0)	0.08058(0)	0.09869(0)	0.13677(0)

4.6.8. AMOVA

The results obtained from AMOVA (Table 15) do in fact show that there is high variation among individuals of the same species, reflected in the access of heterozygosity reflected in F_{is} values obtained earlier. As a matter of fact the variation among the individuals of a population account for 85.64% of total variation.

As for among population variation, this source was only accountable for 15% of the overall variation. In other words, even though we observed high variability between populations in the region, they do not match up to the variation seen within individuals of a population.

Table 15 Molecular variance results obtained from AMOVA.

Source of variation	**d.f.**	**Sum of squares**	**Variance components**	**% of variation**
Among populations	6	92.718	0.23848 Va	14.36
Within populations	405	575.981	1.42217 Vb	85.64
Total	411	668.699	1.66065	
Fixation Index	**FST : 0.14361**			

No. of Permutations = 1000

4.6.9. Population structure

A schematic representation of the genetic structure of all the individuals under study is represented in Figure 26. The graphical image is not based on geographical allocations. The software uses allele frequencies from what is assumed to be potentially unlinked loci, and uses these figures to identify which set of genes are decedents of which population. Based on the allele frequencies the population is set (K), but K does not necessarily have to be the number of sampled locations (it can be anywhere in between one and the number of sampling locations); since it is not based on the actual, physical location. These k values or designated number of populations in the software, indicate the number of theoretical locations that represent our data in the best manner. Thereafter the individuals too are assigned to their significant K populations (Pritchard, Stephens, & Peter, 2000).

The K-value (number of populations) in our data set was chosen to be 5; since its ln(P(X|K) or Estimated Ln Prob of Data was the closest to 0 at K-value= 5 (ln(P(X|K)= -2434.4). This value consideration was recommended by the software manual.

The admixture model that was chosen is recommended for first time usage of the software and it assumes that individuals have mixed lineages. According to the bar plot intense mixture is found in the Pahang population. Earlier this population was found to have the lowest F_{st} values, which means that it has the least genetic differentiation when compared to all populations of *C. striata.* Now after viewing the bar plot it is clear why such low values persist in this location. The severe hybridizations that are seen in Pahang, illustrated with colors that are found in all the other five locations, are the reason why the population is heterozygous rich.

The second most hybridized according to the STRUCTURE image is the Penang population, with almost the same mixture of colors as in the Pahang population, which explains the clustering of populations Pahang and Penang together in the dendograms shown earlier. The Selangor population takes third place after Pahang and Penang populations when it comes to hybridizations in the bar plot, which again is very agreeable with the trees drawn earlier; as it clusters with the Penang population in the consensus tree.

On the other hand populations Johore, Kedah, Negeri Sembilan and Terengganu, very much stand alone with minimal mixing. This is magnified in the Negeri Sembilan population, as it has the minimalist number of colored lines penetrating its blue bar. Once again this is more or less a concurrence as to what has been viewed on the dendograms earlier. In other words the bar plot confirmed the UPGMA trees in the previous section.

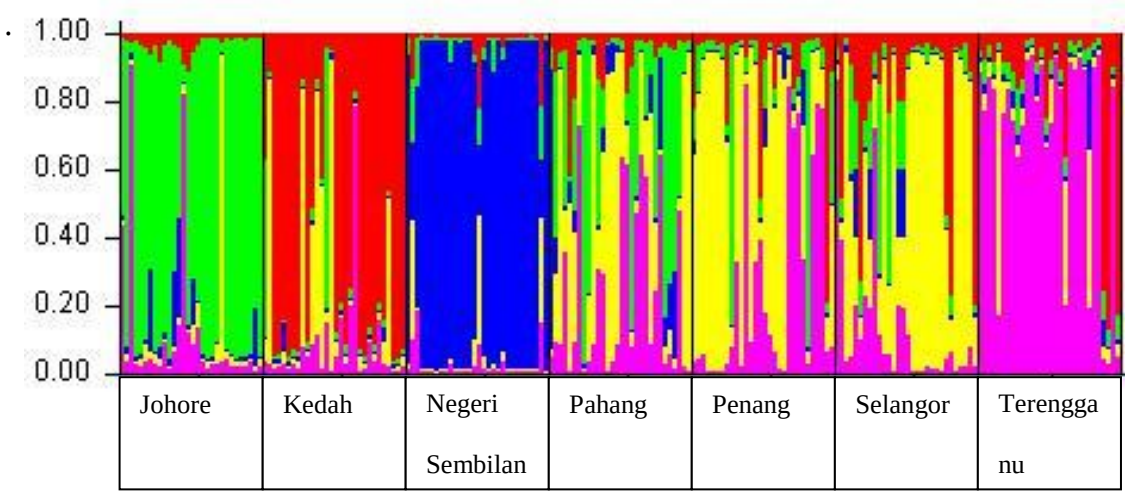

Figure 26 STRUCTURE image of the 7 microsatellites under study. K=5 was chosen, since it had the lowest Ln P(D) value making it the best K-value to represent the data under study. Each rectangular bar represents a different region out of the seven. Mixture of colours represents the probability of hybrids belonging to other clusters in the population.

An odd finding would be that the fish from three different geographical locations (Pahang, Penang and Selangor), have almost similar hybridizations and also cluster together in the UPGMA consensus tree. What makes this unusual is that these three regions are separated geographically, making extremely hard for individuals to mix. For instance states Pahang and Selangor are separated by a mountain range that divides the peninsular in two (Pahang being on the east and Selangor on the west). Penang on the other hand is very much secluded from either state, since the main land is located on the North-East.

A logical explanation would be human exploitation. Farmers and industrial fishermen, tend to illegally move the fish around for their own convenience. It also is extremely hard to monitor or restrict this, since the tropical nature of the country, is covered by jungles and sea coasts; making it almost impossible to inhibit smugglings of terrestrial species to different locations.

Another theory would be that most loci used in this study were tetra repeats or higher (six of the loci out of seven, to be exact). According to Chakraborty di- repeats tend to have 1.5-2 times more mutation rates than tetra. Therefore it is very much probable that the three populations that have clustered together (Pahang, Penang and Selangor), would have the same ancestral origins according to these loci that are under study. In other words they would have probably shared those alleles for generations, with minimal mutations (indels) occurring in those microsatellite regions (Chakraborty, *et. al.*, 1997; Zhivotovsky, Feldman, & Grishechkin, 1997; Lee, *et. al.*, 1999). However there is a very good probability that if other microsatellite loci with shorter motifs (such as di- repeats) could tell a different story, depending on their susceptibility to mutations across the generations.

In addition to repeat motif, repeat length can influence the variation obtained in the data analysis. Studies have suggested a correlation between repeat number and mutation rate (Goldstein & Clark, 1995; Neff & Gross, 2001). Since most of our alleles contain low repeat number, we could say again that they have had less mutations occurring to them, making them a little more stringent and somewhat conserved. Therefore these findings are probably an ascertainment to the originality of the species from these three locations. Meaning that they could have been once a single existing population that has fragmented throughout the years.

The clustering of Johore with Terengganu is probably due to the fact that samples in this study were originally taken from a damn in Johore (Table 2). Therefore mating habits of the species there could have been altered, because of the restriction of water flow. This leads to formation of lakes with high volumes of deoxygenated water that are high in toxic gases such as (H_2S) (Kechik, 1992). The change in habitat ultimately may lead to the clustering of the species from Johore with Terengganu. This is by no means a justification to the result, but just an explanation to why an odd clustering of these two populations on either side of the country could occur.

Perhaps a good suggestion would be to look into the stress affects the ecological changes have left on the species. Sometimes biological responses to environmental changes manifest first before population and genetic impacts occur. Cellular and biochemical responses should be able to indicate toxin accumulation and pathogen affects that are due to abnormal activity around the area (Subasinghe, Shariff, Fatimah, & Hassan, 1992).