CHAPTER 3

METHODOLOGY

In this study, sampling of *C. striata* aimed to collect individuals from selected locations so as to assess and discern the level of genetic variation and population structure between wild and cultured *C. striata* in peninsular Malaysia.

3.1. Sample collection sites and storage

Samples of adult wild *C. striata* were collected from sites across three states in peninsular Malaysia. The three collection sites (Johore, Kedah and Pahang) represent three natural populations of *C. striata*. All fish were caught by local government fisheries officers and the samples were used in a previous study that investigated the status of wild populations of *C. striata* in Malaysia (Salah, 2010). Fishes collected at each site were assumed to be representative of local population variation at and around that site, upon time of sampling.

In addition, *C. striata* reared in three local commercial farms; Kajang [Figure 3.1 (a) & (b)], Malacca [Figure 3.2 (a) & (b)], Rawang (Figure 3.3) were surveyed, representing three cultured populations in this study. The farm located in Kajang used broodstocks which originated from Klang while the Rawang farm employs breeding techniques using local broodstocks and fingerlings imported from Thailand. The exact origin of broodstock used at the Malacca farm could not be verified, however they are believed to have been obtained throughout Malaysia.

Geographical co-ordinates and information for each collection sites are presented in Table 3.0. Figure 3.4 illustrates location of collection sites. The muscle tissues were harvested and individually stored in labelled Falcon tubes containing absolute (100 %) ethanol at -80 °C until DNA extraction.

Table 3.0. Detail of samples, population names, collection sites, geographic co-ordinates, population type and sample size (*N*) of *C. striata* around peninsular Malaysia.

Population Name	Collection Site	Latitude and Longitude	Nature of Population	*N*
Kajang (KJ1-KJ49)	Rotomas Technology (M) Sdn. Bhd. (Farm)	2°57'26.90" N 101°50'38.10" E	Cultured	49
Malacca (M1-M50)	Sempadan Gemilang Enterprise (Farm)	2°24'18.24" N 102°13'13.86" E	Cultured	50
Rawang (R1-R50)	Batu Arang Farm (Farm)	3°19'21.68" N 101°31'00.74" E	Cultured	50
Johore (J1-J30)	Batu Pahat (Dam)	1°51'30.71" N 102°56'16.10" E	Wild	30
Kedah (K1-K30)	Kubur Panjang (Paddy fields)	6°06'51.10" N 100°32'57.53" E	Wild	30
Pahang (P1-P30)	Sungai Pahang (River)	3°41'30.12" N 102°53'42.56" E	Wild	30

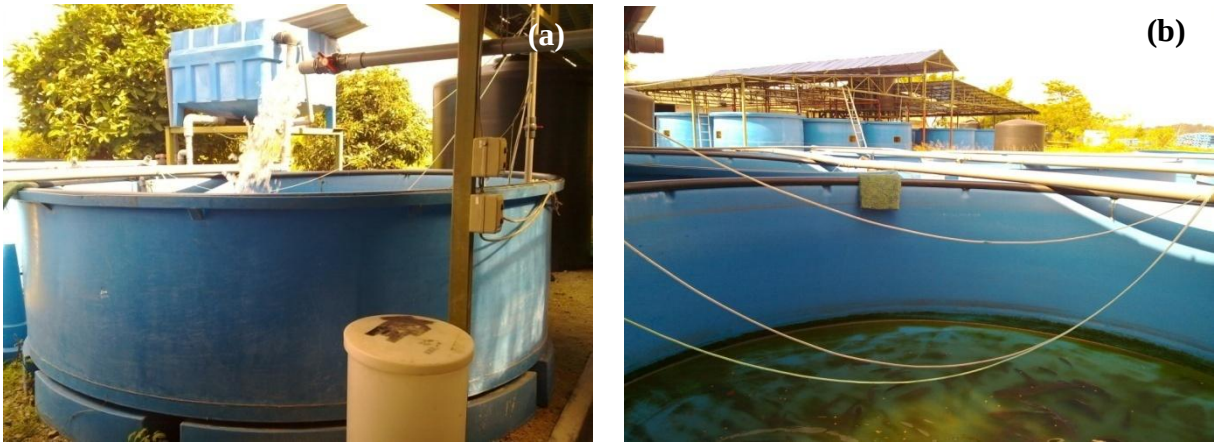
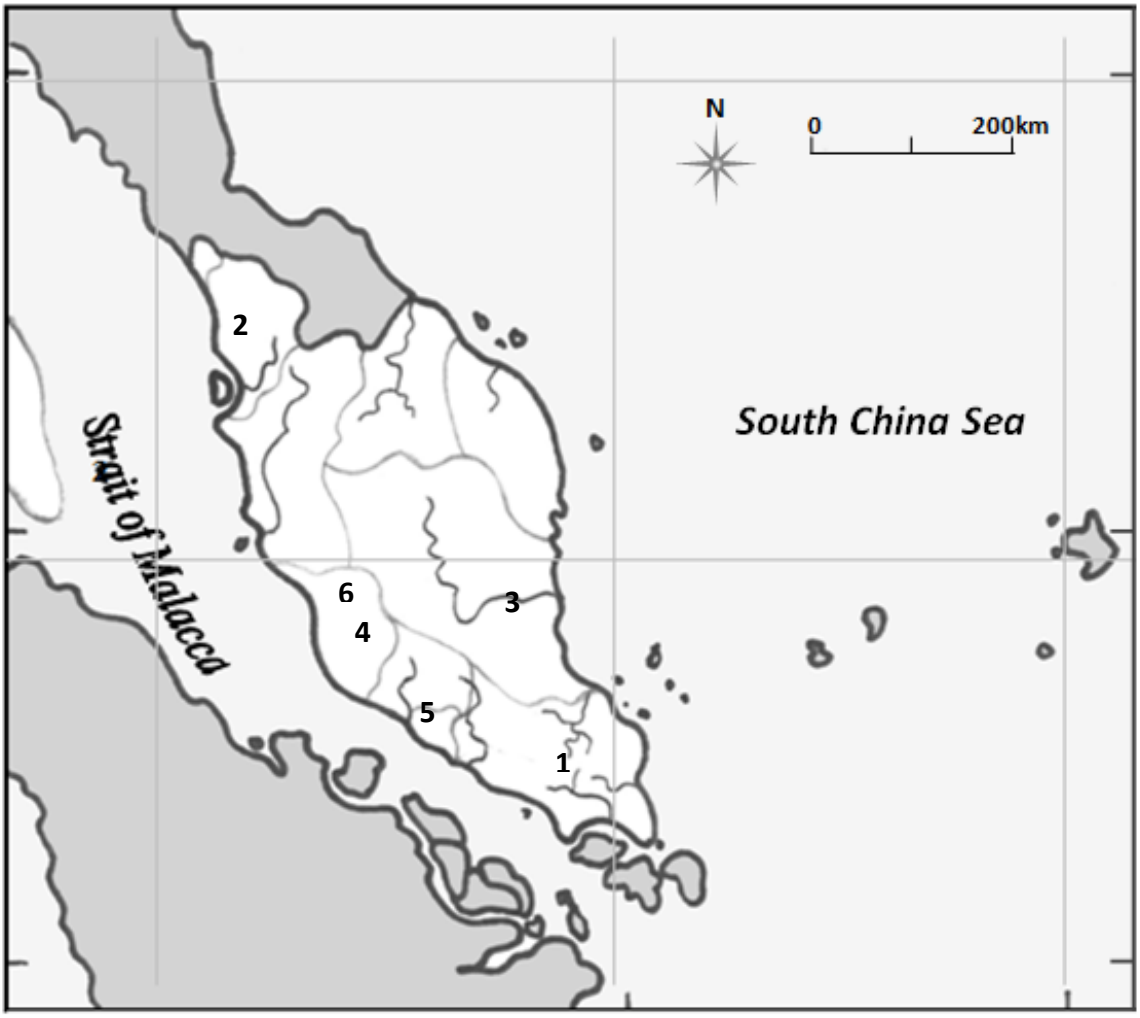


Figure 3.1. (a) & (b) Images from the *C. striata* farm in Kajang that employs holding tanks for the cultures.

Figure 3.2. (a) & (b) Images from the *C. striata* farm in Malacca.

Figure 3.3. View of an aquaculture pond on the *C. striata* farm in Rawang.



Legend :

1. Johore (Batu Pahat Dam)
2. Kedah (Kubur Panjang, Paddy Fields)
3. Pahang (Sungai Pahang)
4. Kajang (Commercial Farm)
5. Malacca (Commercial Farm)
6. Rawang (Commercial Farm)

Figure 3.4. Map of Malaysia showing the sampling sites for *C. striata*.

3.2. DNA Extraction

Total genomic DNA was extracted from all 239 samples using GF-1 Tissue Extraction Kit (Vivantis, Malaysia) prior to screening for genetic variation. Instructions provided by the manufacturer were followed with minor modifications (Appendix 1). The extracted DNA was subsequently quantified by NanoVue^{TM} (GE Healthcare, NJ, USA).

3.3. Molecular Methodology

3.3.1. Microsatellite Primer Optimisation, PCR Amplification, and Gel Electrophoresis

A total of fifty-five pair of microsatellite primers were tested in this study (Appendix 2). Fifty-one pair of microsatellite primers designed for *C. striata* were developed by Salah (2010). Four loci with primer sequences as described in Adamson (2010) were selected to further supplement the screening of genetic variation in *C. striata*.

PCR reaction master mixes for each primer set were optimised by trialling amplification of multiple samples under a range of $MgCl_2$ concentrations and annealing temperatures (T_a) to obtain maximum specificity and efficiency. The PCR reactions were performed in 10 µL reaction volumes using a C1000 Thermal Cycler (Bio-Rad). Each PCR reaction comprised of 20 ng of DNA template, 1× PCR Buffer (Promega), 2.0 – 3.75 mM $MgCl_2$ (Promega), 0.25 mM of each dNTPs component (dATP, dCTP, dGTP, dTTP) (Promega), 50.0 ρmole of each reverse and forward primer; and 7.5 unit of *Taq* DNA Polymerase (Promega). The temperature profile of the PCR was; 96 °C for 3 min

of an initial denaturing cycle followed by 40 cycles of 96 °C denaturation for 10 s, a 45-65 °C annealing for 10 s, a 72 °C extension cycle for 30 s; followed by a final extension step at 72 °C for 7 min.

In order to verify the presence of the desired PCR amplicons, 3 µL of each PCR product was electrophoresed on 1.0% w/v Agarose gels under the following conditions; 75 V, 150 mA using 1× TBE Running Buffer for 45 min. The gels were then stained with ethidium bromide (0.1 mg/mL) prior to visualisation under ultraviolet light (Alpha Innotech; CA, USA).

3.3.2. Preliminary Screening for Polymorphic Microsatellite Markers

Prior to polymorphic marker screening, primers that failed optimisation attempts were discarded. These included loci that failed to amplify, produced non-specific or undesired amplicons (based on the expected PCR product size), as well loci that generated products with low intensity.

The optimised primers were then screened for putative polymorphism using 16 randomly selected *C. striata* individuals from six populations. In this round of screening of the microsatellite markers, primer sets that display distinct allelic "signatures" consistent with null allele occurrence or monomorphic banding patterns were excluded from subsequent evaluation. The degree of polymorphism of each marker was assessed based the level of variation observed from the banding patterns. Those loci that were identified as putatively polymorphic were further validated by an additional round of testing with 16 different arbitrarily chosen individuals. High-resolution agarose,

Metaphor^{TM} agarose (FMC BioProducts, Rockland, ME) was used to segregate PCR products to facilitate the preliminary assessment of polymorphic markers. The electrophoretic conditions for the runs were similar to those stated in 3.3.1 with the exception of the run time, which was extended to 2 hours.

3.3.3. Fragment Analysis

Loci containing potentially polymorphic microsatellites were selected from the initial screening of 55 primer pairs. These loci were then subjected to a secondary screening whereby only the most highly informative markers were chosen for subsequent analysis. The markers were selected based on their polymorphism information content (PIC) (Botstein et al., 1980). PIC refers to the informative value of a molecular marker to detect polymorphism in a population. The number and frequencies of all alleles identified are indirectly related to the PIC value. The greater the number of alleles detected at a locus, the higher the PIC value for that locus. Likewise, for a known number of alleles, as allelic frequencies become more equal among alleles, the greater the PIC value (Liu and Cordes, 2004). PIC values of the microsatellite loci employed in this study were calculated on the basis of observed allelic frequencies in 32 randomly chosen genotypes from all populations using Cervus version 3.03 software (Kalinowski et al., 2007; Marshall et al., 1998).

The PIC values for each of the primer set were estimated by determining its allelic variation per locus according to the formula described by Varshney and colleagues (2007).

$$\text{PIC} = 1 - \sum_{i=1}^{k} (P_{i^2})$$

where *k* is the total number of alleles detected for a given marker locus and P_i is the frequency of the *i*th allele in the set of genotypes studied. In this study, a "good" polymorphic microsatellite primer would have a PIC value of 0.6 and above, an important criteria for the second selection process.

For the second selection, fragment analysis was the method chosen to assess variation in the microsatellite allele sizes at each locus. This approach utilises a capillary electrophoresis system to generate a size estimate for DNA fragments of interest relative to a size standard which contains fragments of known lengths (Applied Biosystems). The forward primers from the sets that have been shortlisted were fluorescently tagged with 6-FAM dye (First BASE). Note that the fluorescently labelled primers are photo-sensitive, hence exposure to any light source was kept at a minimum.

Freshly prepared PCR amplification products were diluted 10-fold with sterile ddH_2O. This was achieved by adding 1 µL of the PCR product with 9 µL of ddH_2O. One µL of this diluted mixture was then transferred into a sterile PCR tube whereby 10 µL of Hi-Di™ Formamide reagent (Applied Biosystems) and 0.2 µL of GeneScan™-500 LIZ™ Size Standard (Applied Biosystems) were subsequently added. The tube was vortexed briefly to allow even distribution of the mixture and then, spun down. The tube was heated for 5 min at 95 °C to denature double stranded DNA and immediately kept on

ice for precisely 5 min, before loading into a 96-well PCR analysis plate (Thermo Scientific) and covering it with a rubber septa. The plate was submitted for fragment analysis using ABI 3130 Genetic Analyzer (Applied Biosystems). The resulting electropherograms were analysed and scored using the software package GeneMapper® version 4.0 (Applied Biosystems, USA) for applications which include genotyping and allele size determination for microsatellite analysis.

The primer sites that yielded the highest PIC values from the second screening were further evaluated on all the samples to address the issue of the genetic variability among the wild and cultured populations. Information pertaining to these 8 microsatellite loci was presented in Table 3.1.

Table 3.1. Characteristics of microsatellite primers used for genetic diversity analyses in this study. Repeat motif refers to the tandem repeats as observed in sequence. Size range of PCR products is stated under alleles size. Information regarding optimisation of the individual set of primers for fragment analysis which include annealing temperature in degree Celsius (T_a) and concentration of $MgCl_2$ per 10 µL reaction was reported along with the corresponding PIC value calculated.

Locus	Primer Sequence	Repeat Motif	T_a (°C)	Alleles size (bp)	[$MgCl^2$] (mM)	PIC value	Reference
T113-11	5'-CCC TGT ATT TCA TTT CTC CA-3' 5'-ACC AAC ACT GCA ATC TCT CT-3'	(CTTT)3	56.9	267-303	3.75	0.481	(Salah, 2010)
BP6-2	5'-AGA AGA AGA AGA AGC CGA GT-3' 5'-GAA AAA CAG AGC AGG AAC AC-3'	(AGAGG)2	52.9	158-233	3.75	0.529	(Salah, 2010)
BP 6-4	5'-TCG AGC TGT GTT TAA GTG TG-3' 5'-GTT CGT GTT GTT TTC CAT CT-3'	(GT)16	58.7	254-298	3.75	0.876	(Salah, 2010)
PCT6-6	5'-CAC CTT TCC TTT GAG TCT TG-3' 5'-GTT CGT GTT GTT TTC CAT CT-3'	(CAGGTA)2	53.9	232-328	3.75	0.621	Unpublished.
BP13-6	5'-CTC TCT CTA ACA CAC ACA CAC C-3' 5'-ATT CAC TTC CTG TTC ACA CC-3'	(CA)10	60.0	215-227	3.75	0.569	Unpublished.
BP13-14	5'- TTT GAA AGA GCG AGA TAA GG-3' 5'-TAG AAA CAA AAT GGG GAC AG-3'	(CT)13	60.0	214-254	3.75	0.869	Unpublished.
CS-4	5'-TCG CAG TTT ATG TAC CGA CA-3' 5'- CTC CAG GGG AAT TTA CAG CA-3'	(CA)15	60.0	140-174	2.5	0.760	(Adamson, 2010)
CS-5	5'- AAA CCC AAA AGC CAC ACT TC-3' 5'- TGA AAT AGA GCC TGT GAC TGA TG-3'	(CA)14	50.0	134-174	2.5	0.807	(Adamson, 2010)

3.4. Data analysis

The raw microsatellite data from an Excel file was converted using the software CONVERT version 1.31 (Glaubitz, 2004) into input file formats for statistical software programmes used in subsequent analyses. The first step of the analyses was to screen the microsatellite data for any genotyping errors that may be caused by non-amplified alleles (null alleles), large allele dropout, stutter bands or even typographic errors using the programme Microchecker (Van Oosterhout et al., 2004). This application was used to detect the presence of null alleles, indicated by an overall significant excess of homozygotes. The corresponding null allele frequencies (*r*) were subsequently calculated by Microchecker (Brookfield 1996; Chakraborty et al., 1992).

Conformation to Hardy-Weinberg Equilibrium (HWE) at each locus for each population was carried out via exact tests (Guo and Thompson, 1992) using Arlequin version 3.11 (Excoffier et al., 2005) with the following parameters: No. of steps in Markov chain = 100 000, Dememorization = 1 000. The HWE test compares observed genotype frequencies with the frequencies expected for an ideal population (random mating, no mutation, no drift, no migration, no selection) (Selkoe and Toonen, 2006). In a population that achieves HWE, the frequency of the genotypes is dependent on the frequency of the genes and both will remain constant over generations. Departures from HWE may indicate that one or more of the HWE assumptions is violated in a given population. The significant criteria were adjusted for the number of simultaneous tests using False Discovery Rate (FDR) procedure (Benjamini and Hochberg 1995; Verhoeven et al., 2005). Arlequin was also used to compute the observed heterozygosity (H_O) and expected heterozygosity (H_e) values per locus and per population whereas the observed number of alleles (A), effective number of alleles (N_e) and percentage of

polymorphic loci were calculated by the software POPGENE version 1.31 (Yeh et al., 1997).

The log likelihood ratio statistic (G-test) method was employed in genotypic linkage disequilibrium (LD) analysis to indicate presence of significant associations between alleles across microsatellite loci using GENEPOP version 4.0 (Raymond and Rousset, 1995). This test was performed for pair-wise locus combinations for all populations, with the Markov chain parameters set at 10 000 dememorizations, 100 batches and 5 000 iterations per batch. Allelic association or LD, is measured as the departure of gametic frequencies from expectation under allelic independence among the loci (Hoelzel, 1998). Statistical significance levels used to detect genotypic equilibrium ratios were adjusted for the number of multiple tests with the FDR method. Both genotypic and allelic frequencies for each loci for each population were also estimated using GENEPOP.

To enable comparison of intrapopulation genetic diversity, allelic richness (A_R) was measured for each locus at each population in the software programme FSTAT Version 2.9.3.2 (Goudet, 2002), standardized to the smallest sample following rarefaction method (Leberg, 2002; Petit et al., 1998) (in this case, wild populations, n=30). Additionally, A_R may be relevant in establishing a decrease in population size or occurrence of a past bottleneck (Nei et al., 1975).

Arlequin was used to compute the Garza-Williamson ($G-W$) Index, denoted as M which refer to the mean ratio of the number of alleles to the range in allele size. The revised $G-W$ statistic (Excoffier et al., 2005; Garza and Williamson, 2001) is defined as:

$$G - W = \frac{k}{R + 1}$$

Where:

k = number of alleles at a given loci in a population sample

R = allelic range

The $G - W$ statistic evaluates if a population bottleneck is likely to have occurred by discriminating between populations that have recently reduced in size and those that have been small for an extensive period (Garza and Williamson, 2001). A significant bottleneck will result in a greater loss of alleles than a proportional loss of range of allele sizes, and as a consequence, the ratio of those parameters (M) will be lower for bottlenecked populations (Hundertmark and Daele, 2010). M varies from 0 (population bottleneck) to 1 (stationary population), with the critical range of $M < 0.68$ inferring a recent population reduction in size (given a data set with seven or more loci) (Garza and Williamson, 2001).

In the statistical analyses of population differentiation, Wright's fixation index or *F*-statistics (1951) was used to examine the variation in gene frequency among subpopulations, implemented in Arlequin. It is well established that population subdivision will result in a loss of genetic variation. Therefore, *F*-statistics were employed to quantify the level of genetic differentiation between subpopulations by measuring the extent of reduction in observed heterozygosity compared to the expected heterozygosity. This quantification has been formalised in a series of hierarchical *F*-statistics (Nguyen et al., 2006).

The series is primarily composed of three *F*-statistics measures; namely inbreeding coefficient (F_{IS}), fixation index (F_{ST}) and overall fixation index (F_{IT}). F_{IS} is the mean reduction in heterozygosity of an individual due to non-random mating within a subpopulation. Thus, F_{IS} is a measure of the extent of genetic inbreeding with subpopulations and ranges from -1.0 (all individuals are heterozygous) to +1.0 (no observed heterozygotes). On the other hand, F_{ST} represents the mean reduction in heterozygosity of a subpopulation relative to a total population, which resulted from genetic drift among subpopulations. In short, F_{ST} measures the extent of genetic differentiation among subpopulations, which ranges from 0.0 (no differentiation) to 1.0 (complete differentiation) where subpopulations are fixed for different alleles. The final coefficient F_{IT}, albeit infrequently utilised is the mean reduction of heterozygosity of an individual relative to total population.

The three measures of *F*-statistics can be expressed as:

$$F_{IS} = \frac{H_S - H_I}{H_S}$$

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

$$F_{IT} = \frac{H_T - H_I}{H_T}$$

Where:

H_I is the observed heterozygosity averaged across all subpopulations.

H_S is the expected heterozygosity across all subpopulations.

H_T is the expected heterozygosity for the total population.

Analysis of Molecular Variance (AMOVA) (Excoffier et al., 1992), implemented in Arlequin was used to examine population structure via hierarchical partitioning of genetic diversity within and among populations. In this study, the 6 populations studied were organised into 2 groups; wild (Johore, Kedah and Pahang) and cultured (Kajang, Malacca and Rawang). AMOVA performed a hierarchical analysis of variation that apportioned the total variance into covariance components in which tests included permutation of inferred haplotypes (allelic data) among populations and among groups (F_{IT}); inferred haplotypes among individuals within populations (F_{IS}); individual genotypes among populations but within groups (F_{SC}); and populations among groups (F_{CT}). The significance of the resultant statistics was tested with 1 000 random permutations.

Three pair-wise estimators of differentiation were computed to measure the different indices of dissimilarities (genetic distances) between pairs of populations. Firstly, F_{ST} and R_{ST} analogues (Slatkin 1991, 1995) were calculated using Arlequin (Excoffier et al., 2005), with 10 000 permutations. FDR adjustment of significance values was applied to correct simultaneous tests.

F_{ST} serves as the foundation for quantifying genetic distance where drift is the cause of divergence while values of R_{ST}, an analogue assuming the stepwise mutation model of microsatellite mutation (SMM), were measured by distance method as sum of squared size differences based on number of repeats (Balloux and Lugon-Moulin, 2002; Slatkin, 1995). As an estimator of population subdivision, F_{ST} is better suited to describing differences among on weakly structured populations. In contrast, R_{ST} performs best when the populations are highly structured, especially when presented with a small number of samples. However, investigation of population differentiation by means of

F-statistics and their analogues has been impeded by limitations when working with data containing highly polymorphic loci. Among the constraints faced by F_{ST} is the severe underestimation of differentiation in highly structured populations. R_{ST} is inundated by high variance and its performance is sensitive to deviations from SMM as R_{ST} gains independence from mutation only under strict SMM (Balloux and Lugon-Moulin, 2002).

A third measure of differentiation, D_{est} (estimator of actual differentiation) was developed by Jost (2008) to avoid the innate bias of fixation indices. D_{est} operates on the principle that the effective number of alleles scales linearly with an increase in equally frequent alleles, therefore, resulting in a more intuitive "true diversity" estimate (Jost, 2008; Meirmans and Hedrick, 2011). By eliminating heterozygosity, D_{est} is an explicit measure of relative differentiation between populations and is based on the effective number of alleles. D_{est} was calculated for each loci using a web-based application, SMOGD (Crawford, 2010) and can defined as:

$$D_{est} = \left[\frac{H_{\mathrm{t\ est}} - H_{\mathrm{s\ est}}}{1 - H_{\mathrm{s\ est}}}\right] \cdot \left[\frac{n}{(n-1)}\right]$$

Where:

$H_{\mathrm{t\ est}}$ = nearly unbiased estimator of total-subpopulation heterozygosity (Nei & Chesser, 1983)

$H_{\mathrm{s\ est}}$ = nearly unbiased estimator of within-subpopulation heterozygosity (Nei & Chesser, 1983)

n = number of populations.

Mantel tests were employed to examine the statistical significance of matrix correlations, in this case, among the three different pair-wise estimators of genetic differentiation (F_{ST}, R_{ST} and D_{est}) (Mantel, 1967). Arlequin was used (1 000 permutations) to conduct multiple bi-matrices tests of regression against F_{ST}, R_{ST} and D_{est} to determine whether correlation between these differentiation estimators would concur to a single analogous interpretation of the given dataset (Excoffier et al., 2005). Significant levels were adjusted with the FDR procedure.

To infer population genetic structure, the wild and cultured populations under study were clustered (disregarding their spatial distribution) using STRUCTURE version 2.2 (Pritchard et al., 2000). STRUCTURE utilises a Bayesian model-based clustering approach that assigns individuals to *K* populations based on their genotypes while the Markov chain Monte Carlo (MCMC) algorithm classifies individuals into populations and estimates the probability of membership in each population for each individual (Falush et al., 2003; Pritchard et al., 2000). The model simulated an admixture of individual ancestry and was run with correlated allele frequencies among populations with the following parameters: constant λ= 1.0, MCMC chain lengths = 50 000 iterations, with a burn-in period of 37 500 cycles. Each simulation of *K*, from 1 to 10 was performed in duplicate to assess the consistency of the analysis. The optimal number of clusters, *K* was determined by the highest mean LnP(D) value generated for each number of clusters, *K* across all replicate runs (Evanno et al., 2005).

Two approaches were employed in this study to illustrate the magnitude of differentiation among populations. The Neighbour-Joining (NJ) (Saitou and Nei, 1987) method implemented in MEGA version 4 (Tamura et. al., 2007) was used to construct a population tree based on pair-wise F_{ST} distances. In addition, Unweighted Pair-Group Method of Arithmetic Averages (UPGMA) clustering was performed based on Nei's 1978 unbiased genetic distances using genetic data analysis (GDA) version 1.1 (Lewis and Zaykin, 2001) to describe the relationship between populations.