Manual on Application of Molecular Tools in **Aquaculture and Inland Fisheries Management**

Part 2

Laboratory protocols and data analysis

Manual on Application of Molecular Tools in Aquaculture and Inland Fisheries Management

Part 2:

Laboratory protocols and data analysis

Contributors

Thuy Nguyen
Network of Aquaculture Centres in Asia-Pacific

David Hurwood, Peter Mather
School of Natural Resource Sciences, Queensland University of Technology

Uthairat Na-Nakorn
Kasetsart University, Thailand

Wongpathom Kamonrat
Department of Fisheries, Thailand

Devin Bartley
Food and Agriculture Organization of the United Nations

NACA MONOGRAPH SERIES
NACA is an intergovernmental organization that promotes rural development through sustainable aquaculture. NACA seeks to improve rural income, increase food production and foreign exchange earnings and to diversify farm production. The ultimate beneficiaries of NACA activities are farmers and rural communities.

Visit NACA online at **www.enaca.org** for hundreds of freely downloadable publications on aquaculture and aquatic resource management.

© Network of Aquaculture Centres in Asia-Pacific
PO Box 1040, Kasetsart University Post Office
Ladyao, Jatujak
Bangkok 10903
Thailand
Email: info@enaca.org

Nguyen, T.T.T., Hurwood, D., Mather, P., Na-Nakorn, N, Kamonrat, W. and Bartley, D. 2006. Manual on applications of molecular tools in aquaculture and inland fisheries management, Part 2: Laboratory protocols and data analysis. NACA Monograph No. 2, 134p.

ISBN 978-974-88246-2-8

Printed by Scandmedia, Bangkok.

Contents

List of tables

List of figures

Acknowledgements

The senior author would like to thanks Mr. Pedro Bueno, former Director General of NACA, without whose support the manual could not have become possible. Encouragement from Prof. Sena De Silva, Deakin University (current Director General of NACA) is very much appreciated. P. Mather and D. Hurwood would like to acknowledge the Australian Centre for International Agricultural Research (ACIAR) for funding support. We would like to thank Dr. Menchie Ablan from the WorldFish Centre for her constructive comments.

Aims, scope and format of the manual

This is the second part of the manual, "**Application of molecular genetic techniques in aquaculture and inland fisheries management**". The major aim of this part of the manual is to provide step-by-step laboratory protocols and methodologies for data analysis, and a guideline to design a population genetic study.

The scope covers most commonly used techniques for screening genetic variation, general background on the methodologies for estimation of important parameters in population genetic studies for different forms of molecular genetic markers.

Part 2 includes three sections:

- **Section I - Molecular markers - an overview:** will provide an overview of common molecular markers used in population genetic studies.

- **Section I - Laboratory protocols:** will provide step-by-step protocols of commonly used molecular genetic techniques.

- **Section III - Data analysis and project design:** will deal with aspects of data management such as data analysis, interpretation and presentation, and a guideline to design a population genetic studies.

Section 1

Molecular markers - an overview

There are many levels at which we can assess how different populations or individuals may be, genetically. The traditional method and the approach still employed in classical systematics has been to examine external morphological traits and infer divergence in the underlying genes that produce morphological phenotypes. There are a number of problems that can be associated with this approach however, including the fact that morphological traits can often be highly conserved, convergent evolution can confuse relationships and associated with this, environments can modify expression of underlying genes in diverse ways hiding true patterns of relationship. Thus biologists have sought more fundamental markers with which to assess genetic relationships and the advent of molecular genetics has provided potential markers of genetic relationship that do not face the same problems evident in comparisons of morphology. This is because the genotype is fixed at fertilisation and consequently cannot be influenced directly by the environment. Potential sources of direct genetic markers of differentiation include; chromosome morphology, protein variants, whole DNA fragments and DNA sequences. Basically the simple linear arrangement of four nucleotides (A, T, C, and G) contains a large amount of information on evolutionary rates, mutation rates, fixation rates and selection pressures. There are also now a diverse array of genetic markers to choose, both from the nuclear and cytoplasmic genomes.

1.1 Nuclear markers

The nuclear genome (nDNA) is very large in most eukaryote species and recombines. Most species (although not all) have a relatively large number of chromosomes with many DNA sequences available for analysis. Most eukaryote species are diploid (2N) although some groups of fishes (e.g. salmonids and catastomids) contain polyploid lineages that can make genetic differentiation analyses more complex. For most nuclear markers (unlike mitochondrial DNA (mtDNA) where offspring inherit from the female parent only in most species), male and female contribution to offspring is nearly equal (the exception being where sex determination of the individual is via sex chromosomes). For nDNA both coding and non-coding DNA sequences are present and individual sequences vary widely in their rates of evolution that are at least in part determined by functional constraints on some coding sequences. Thus there is a large amount of DNA available for analysis and many ways of targeting individual sequences for study.

To the uninitiated a diverse array of molecular techniques and methodologies are now available for analysing DNA sequences and choosing appropriate methods for the specific questions to be addressed can be a complex and confusing process. When deciding on what type of genetic marker to use to characterise DNA genetic diversity in a new study, a number of issues should be considered: (1) choice of marker to a large extent will depend on the research question(s) to be addressed - choose a marker that can adequately address the specific research questions, (2) relative costs, technical difficulty and necessary facilities/equipment may require compromise – but choose the most powerful marker available within resource limitations, with the capacity to address the questions and (3) some questions are better addressed by synthesising data from different types of markers (e.g. combining data from nDNA and mtDNA markers). Two broad classes of marker are available for analysis of nDNA, so-called Dominant and Co-dominant markers. Dominant markers (e.g. Random Amplified Polymorphic DNAs [RAPDs] and Amplified Fragment Length Polymorphisms [AFLPs]) are assessed by presence /absence of a band on a gel and so we are not able to directly assess if an individual with a band is homozygous (aa) or heterozygous (ab) for the fragment. So for genetic diversity studies with dominant markers we can only estimate expected heterozygosity assuming H/W equilibrium and have no way of determining if the population conforms or not. Thus, while dominant markers have their uses (especially for example AFLP's in gene mapping studies), co-dominant markers (e.g. allozymes, microsatellites, SSCPs etc.) provide better analytical power because they allow direct determination of heterozygosity in a sample. While many studies have used dominant RAPD markers to document genetic diversity, in recent times questions have been raised about the reliability and repeatability of RAPD's particularly in animal species because they can be very sensitive to the PCR conditions used to generate them. This means that there is now more pressure to demonstrate that RAPDs are repeatable and inherited in a Mendelian fashion before data are necessarily accepted.

The markers that led the revolution in modern genetic diversity analyses were Allozymes. First developed for analysis of some human metabolic disorders in the early 1960's, evolutionary biologists quickly realised that they could be used to investigate diversity in a wide array of organisms at a much more fundamental level than had been possible previously. While allozymes are not affected by environmental factors, unlike most DNA technologies, the variation observed is 'one step' away from variation at the most fundamental level because it is protein phenotypes that are compared rather than DNA sequences directly. Studies commenced in the late 1960's and early 1970's and now a huge data set is available. While new technologies have taken over from allozymes to a large extent in many labs, they still provide a relatively inexpensive, quick and sensitive technique for screening genetic diversity. In some labs, researchers are returning to use this method because often the results achieved are comparable with that achieved from more sophisticated but technically difficult and expensive methodologies (e.g. microsatellites). The main shortcomings of this technique is the fact that tissue needs to be stored at -80°C and that sampling tissue is almost always invasive (i.e. the animal is killed).

One of the first technologies to directly target diversity at the DNA level directly was restriction fragment length polymorphism (RFLP). This relatively simple approach uses commercially available restriction enzymes (REs) that are harvested from bacteria species to cut DNA sequences at specific sites (restriction sites). Each RE has a unique recognition site, so that when a particular RE (e.g. EcoR-1) recognises its own specific restriction sequence along a DNA molecule it cuts the DNA at the site. This is known as a restriction digest. The more restriction sites along the sequence the more places the RE will cut the DNA and hence the greater will be the number of fragments produced. By exposing the same piece of DNA from different individuals to the same RE's separately we can detect any mutations in the recognition sites in individuals by the pattern of fragments produced on a gel that separates DNA fragments by size. For each individual, the combination of all size fragments after 'cutting' with the RE should equal the total size of the original DNA fragment. Thus we can estimate genetic divergence among individuals or populations by comparing the number of DNA fragments (or RFLPs) they share after restriction digest. Individuals with identical genotypes will have exactly the same set of fragments for the same piece of DNA. While RFLP analysis is still used for genetic diversity studies, the approach is less popular now because it is limited by (a) the ability of RFLP analysis to only provide information about the sites at which REs cut and not the larger intervening sequence of DNA and (b) the relatively high cost of the restriction enzymes.

Development of the Polymerase Chain Reaction (PCR) in the late 1980's revolutionised the study of DNA sequences. This is because the method in theory allows any DNA sequence to be amplified and millions of copies made of just the targeted fragment that is then available for analysis. Essentially PCR is *in vitro* DNA replication. Providing we have some knowledge about the DNA sequence surrounding the sequence of interest we can design Primers (short DNA fragments usually (10 to 25bp in length - oligonucleotides) that bind to the DNA either side of a target sequence and act as initiators of replication for the target sequence. A similar approach is used to obtain the complete sequence of a DNA fragment, because in sequencing reactions the new nucleotides that are added to a growing fragment have been prior labelled with fluorescent dyes (a different colour each for A, T, G and C) and a laser reads the colour as they are incorporated individually in the growing strand and hence the complete sequence of bp's along a fragment can be deciphered. The sequence data can then be stored on large databases (e.g. GenBank, Australian National Genetic Information System – ANGIS) and accessed via the net for comparative purposes.

Microsatellites (or simple sequence repeats – SSRs; or variable number tandem repeats VNTRs) currently are the nDNA marker most favoured in genetic diversity studies as they usually show very high levels of individual and population variation. SSRs are either simple (e.g. TGTGTGTGTG) or complex (e.g GAA$(GA)_{17}$GAA) tandem repeats of short DNA sequences that are found at regular intervals right across the genome of most eukaryote species. While they are quite costly and time-consuming to develop for a new species they have great utility in intra-specific studies. Most SSRs are embedded in non-coding DNA sequences, are noncoding themselves and consequently accumulate mutations very rapidly. Thus individuals in outbred populations rarely share the same SSR genotype across several loci. While this characteristic is advantageous in many intra-specific genetic diversity applications it can be problematical in comparisons of more distantly related individuals or populations because both Homoplasy and Null alleles can be a problem. Where either or both occur they will tend to result in under-estimates of true genetic divergence. Another issue for use of microsatellite data in phylogenetic studies is that the mode of evolution of microsatellite alleles has been debated. Unless it is it has been determined for a particular species that mutations occur as accumulation or loss of single (stepwise mutation model) or multiple repeats then making the wrong assumption can bias phylogenetic reconstructions of relationship. due to the problems outlined here, it is unwise to use SSRs for phylogenetic reconstruction.

Once genetic data are available from individuals and or populations there are a number of different genetic diversity indices that we can calculate (depending on the data type). The ones most frequently used include; % polymorphic loci, relative allelic diversity, average heterozygosity, Perhaps % polymorphic loci is the least informative index (unless a relatively large number of independent loci are available) because the interpretation can be heavily biased by both the number of loci screened and their function in the organism (if any). Relative allelic diversity and average heterozygosity are considered better measures of genetic diversity than % polymorphic loci but cannot always be calculated e.g. heterozygosity cannot be calculated directly from markers that show dominant inheritance. Other complications for specific types of marker include the recognition that phylogenetic inference from most nDNA is compromised potentially by the fact that associations of alleles/sequences along individual chromosomes in the nDNA have the potential to be mixed up by chance genetic recombination events and confused by bi-parental inheritance (i.e. an offspring can receive various combinations of the parental chromosomes). For fast evolving nDNA sequences like SSR's we also need to consider that homoplasy and null alleles could affect our interpretations about genetic diversity. Demographic events can also influence the patterns of genetic diversity in some populations and not in others, for example when one population has been pushed through a population bottleneck while another has not. Population bottlenecks can result in loss of allelic diversity as population size declines. This loss is stochastic and essentially unrelated to individual fitness. This means that current levels of allelic diversity may be related to past demographic events (e.g. historical population bottlenecks) and hence may not accurately reflect current relationships.

1.2. Mitochondrial DNA markers

Mitochondrial DNA (mtDNA) differs significantly from nuclear DNA in structure and mode of inheritance. MtDNA is a circular molecule that undergoes no recombination and is maternally inherited. The molecule (generally 16-20kb in size) is made up of 13 protein coding genes, 2 ribosomal RNA genes (rRNA), 22 transfer RNA genes (tRNA) and section generally known as the D-loop or Control Region (which is non-coding but is involved in the replication of the molecule). Unlike the nuclear genome, the mitochondrial genome contains very little non-coding DNA.

Several characteristics of mtDNA make it a good choice of molecular marker for population studies. Firstly, its maternal inheritance and haploid nature dictate that populations will exist at ¼ the effective population size (N_e) as that seen with nuclear markers. This characteristic will amplify the effects of drift (causing populations to differentiate), so mtDNA is therefore sensitive for detecting population structure. Secondly, the lack of recombination means that mtDNA lineages will evolve without the history of descent becoming jumbled over time as on homologous chromosomes. This allows us to differentiate between historical and contemporary processes that may have influenced or determined the observed population genetic structure. Thirdly, mtDNA is generally considered selectively neutral (although this has been challenged recently). Therefore the effects of selection can largely be removed as a confounding factor when interpreting the data. In contrast, past or ongoing selection can be a significant problem for some nuclear markers.

The most powerful method for determining variation is by direct sequencing (PCR) of the same DNA fragment for all individuals. This method is expensive, particularly for population studies where sample sizes and number of populations may be high. The previous section has described several useful techniques but not all of these are appropriate for mtDNA analysis. Here we will discuss two methods that are particularly useful: Temperature Gradient Gel Electrophoresis (TGGE) and Single Strand Conformational Polymorphism (SSCP). Both methods have reasonable high throughput and a high power to detect single base pair mutations (but slightly less resolution than sequencing) therefore they can identify unique haplotypes. Once all individuals have been assessed, analysis of haplotypic frequency data is then possible. This method however, does not achieve the full potential of the data. Also, the power to isolate contemporary from historical gene flow is significantly reduced. For a little more financial outlay, one or two representatives of each unique haplotype detected in the screening process can be sequenced. This method can reduce the number of individuals requiring sequencing by an order of magnitude (e.g. from 100's to 10's).

As the name suggests, TGGE relies on a temperature gradient to denature double-stranded DNA fragments that are differentially separated based on their respective melting profiles. Another method, Denaturing Gradient Gel Electrophoresis (DGGE) is similar to TGGE but uses a chemical gradient of increasing urea and formamide concentrations to denature the DNA instead of a temperature gradient.

The procedure involves electrophoresing DNA through a polyacrylamide gel that is running parallel to a temperature gradient. The double-stranded duplexes of DNA migrate through the gel until they reach their respective melting points where progress is greatly reduced when the dsDNA begins to unwind. The melting point of a specific fragment of DNA is a function of both the effect of base sequence on the helix structure and the electrophoretic mobility of the strand as it starts to unwind. Therefore DNA fragments with different base pair sequences tend to display different melting points and hence stop at different points on the gel. More details on TGGE techniques will be illustrated in Section 2.10.

To improve the resolution of the technique, TGGE is often conducted in conjunction with Heteroduplex Analysis (TGGE/HA). Nuclear (diploid) DNA fragments can be heteroduplexed to themselves but because mtDNA is haploid, an extra reference DNA fragment (ideally from a moderately divergent conspecific individual) needs to be added. The heteroduplexing process involves heating both the reference and sample DNA together in order to reduce them to single strands. Upon cooling the strands reanneal. Apart from the original double strands from the reference and sample fragments recombining to themselves (homoduplexes), mismatch pairings occurs with one strand from the reference and one strand from the sample also recombining (heteroduplexes). Where heteroduplex fragments have nonperfect complementary matches, they tend to have lower and more variable melting points than homoduplexes resulting in additional bands on the gel. TGGE is a reliable method for DNA fragments up to ~700 base pairs in length.

SSCP relies on electrophoresing single stranded DNA through a nondenaturing polyacrylamide gel. Under appropriate conditions (i.e. nondenaturing conditions) single stranded DNA folds into a particular shape (tertiary structure), with different sequences generally resulting in different structure. The electrophoretic mobility through an acrylamide gel is largely dependent on the shape of the fragment that is determined by its unique DNA sequence, therefore different haplotypes will provide different banding patterns on the gel when visualised. This method does not require any special apparatus like TGGE does and can detect nucleotide differences in fragments up to ~700 base pairs in length. However, there tends to be an inverse relationship between the length of the fragment and the sensitivity of the technique (i.e. resolution is better with smaller fragments).

SECTION 2

Laboratory protocols

2.1. Allozyme electrophoresis

The term "allozyme" refers to products of different allelic forms of an enzyme coding gene. These products are strands of amino acids called polypeptides. Five of the 20 common amino acids that make up polypeptide chains have weak electric charges (lysine, arginine and histidine are positively charged with NH_3+; aspartic acid and glutamic acid are negatively charged with COO-). Thus, polypeptide chains comprising different numbers of these amino acids have different net electrical charges so does the allozyme molecule which is made of one or more than one type of polypeptide chains. As such if crude extracts of enzymes are placed in an electric field, allozyme molecules will be separated according to their net-charge.

Allozyme electrophoresis involves the separation of products from isozyme alleles on the basis of differential migration due to varying surface charge when subjected to an electric current. Thus different alleles are detected based on the mobility differences of their products at the end of the run, which are visualized via specific histochemical staining procedures (Richardson et al. 1986).

The advantages of allozymes are their co-dominant nature, relatively low cost and availability of protocols common for wide range of organisms. However the requirement for fresh tissues can be a major draw back. Moreover, allozymes commonly have low levels of polymorphism hence they may not be suitable for detecting genetic diversity of organisms showing weak population differentiation such as marine organisms. Lastly, the technique only detects a portion (usually <25%) of the actual genetic variation because not all nucleotide changes lead to amino acid substitutions. Additionally not all amino acid substitutions result in electrophoretically detectable mobility differences (Ryman & Utter 1987).

Allozyme data are used for detecting population structure, hybridisation, species boundaries and for estimating levels of gene flow (Hillis et al. 1996), and investigating systematic relationships (Richardson et al. 1986; Swofford et al. 1996).

There are four most commonly used methods of allozyme electrophoresis, depending on types of medium used: starch, polyacrylamide, agarose and cellulose acetate.

2.1.1. Gel preparation

Buffer for electrophoresis

Running buffer systems are chosen according to the types of tissue sampled and enzymes. Often stocks of buffers are made for subsequent dilution before use. The proper dilutions and running conditions are given in Annex 1.

The buffer used for electrophoresis is called running or electrode buffer, and that used for making gel is called gel buffer. There are two different running systems based on the combination of running and gel buffers - **continuous** and **discontinuous** electrophoresis systems. The running system that use the same buffer for electrode and gel is called continuous (e.g. the TC6, TC8, TEB, TG and TM buffers; and when the electrode buffer is different to gel buffer it is called discontinuous (e.g. the LiOH and Poulik buffers) (See Annex 1).

Making starch gel

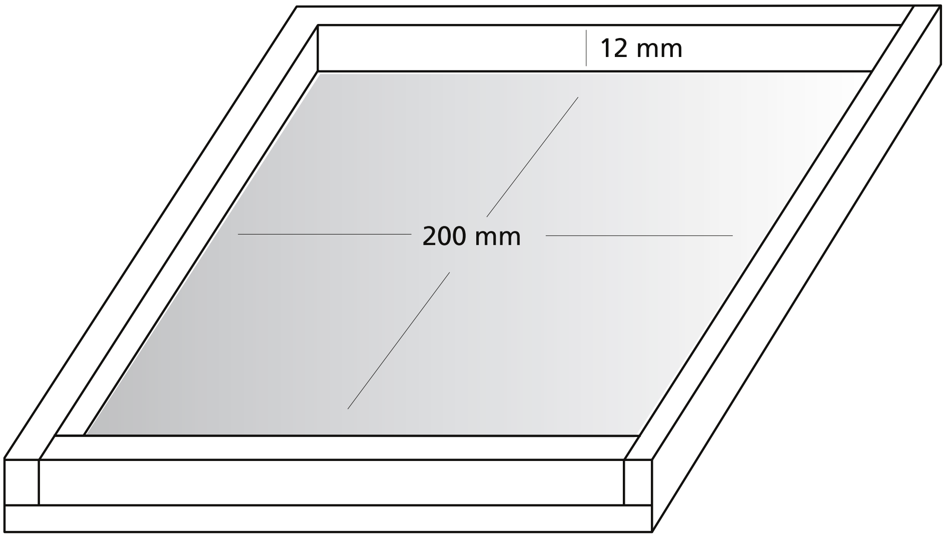
1. Clean gel mould(s) using tissue paper absorbed with ethanol (75-80%) to clean gel mould. Clean once again using distilled water. The reason to clean the gel mould properly is to avoid the gel sticking onto the plate and breaking. Set up the plates on a sheet of newspaper or paper towel, in case there is any spillage.

2. Gel mould(s) are often made from Perspex (5mm thick should be fine). There is no restriction on size of the mould, however we recommend 200 mm x 200 mm x 12 mm for easy handling and suitable number of samples that can be run (about 20 samples could be stained for four enzymes per gel) (see Figure 1).

3. Place the appropriate amount of starch into the boiling flask. The suggested amount is 55g of starch per gel (200mm x 200mm x 12mm). A one-litre boiling flask is sufficient to make one gel; a 2-litre boiling flask for making 2 gels.

4. Add 450 ml of buffer (exact amount varies with the buffer system; see Annex 1) to the starch, and swirl the mixture to suspend the starch. The specific instructions for each buffer are indicated on the stock bottles. Each stock requires dilution for use.

Figure 1. An example of a gel mould.

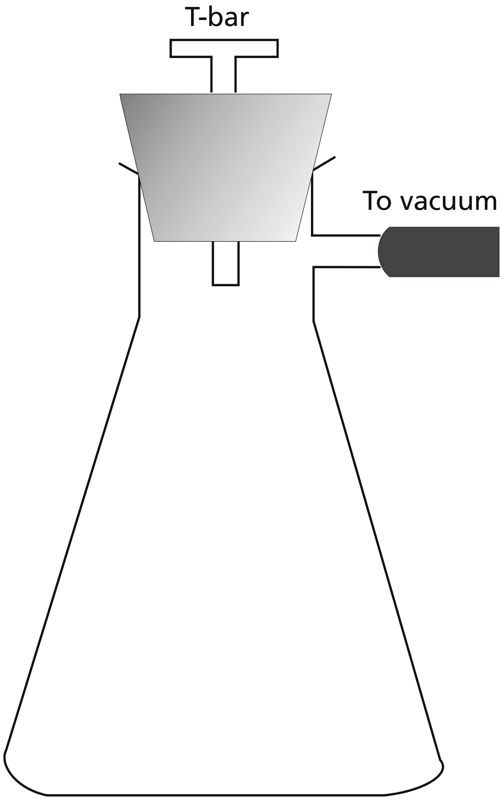


5. Place flask with fully suspended starch on a hot plate with stirring bar inside. Set heat on high and stirrer on approximately 3/4. Pick up and swirl flask by hand every five minutes or so and check bottom of flask for signs of boiling. If indications of boiling keep swirling vigorously until it abates. Return to hot plate and closely monitor. Localised boiling of starch while cooking is to be avoided at all costs as it causes uneven cooking and lumpy gels. Swirling need not be violent, but just sufficient to prevent over-cooking on the bottom of the flask. As the starch heats up, the suspension becomes more viscous and the stirring bar will stop; remove flask from hot plate every 30 seconds or so and swirl for a few seconds before returning it to the hotplate for a total of a further 3 - 5 minutes until the starch mixture begins to loosen up again and becomes semi-translucent. At that point, it is cooked.

6. De-gas the cooked starch with the aspirator pump. Place the flask on a towel or sponge, rather than the bench or sink, to minimise the chance of breakage. The neck of the flask becomes very hot - keep a glove on one hand. Use a thumb and finger on the holes of the "T-bar" on top of stopper as a safety valve, to prevent starch from being sucked into the pump. De-gas the starch until it has come to a full boil (Figure 2)

7. Pour the starch into the gel mould with a continuous action; avoid backtracking and zigzagging, as these can create discontinuities in the gel. Make certain that you use all of the starch, equally distributed between the plates.

Figure 2. Connection to vacuum to de-gas.



8. Immediately place the flask under running (preferably hot) water. If the starch sets on the glass, it will be difficult to clean.

9. Place the lid on the gel, taking care not to trap bubbles. Do not press the lid flush with the sides of the plate - if you do, large bubbles may be sucked in as the gel cools.

10. Leave the covered gel at room temperature for at least 1 hour. It is generally most convenient to pour the gels one night before use. Depending on the weather, they may be kept longer, but it is risky. It is best to keep a consistent pattern, as the gels gradually lose moisture.

2.1.2. Sample preparation

1. Tissues (or whole animals) should be stored as intact as possible, in the freezer. Depending on the tissue and the enzyme, activity remains reasonably good at -20°C for several months (even years in some cases), but for longer preservation, use the liquid nitrogen cylinder or a -80°C freezer

2. Label all items in the freezer, indicating contents, your name, and the date. Also, when you are certain that you will no longer need specimens, remove them from the freezer, as space is at a premium.

3. Small pieces of tissue are homogenised in depression wells on ceramic tiles (spotting plates) or in 0.5-2.0 ml plastic micro tubes. Add about 2 volumes of extractant.

4. **Extractant contains 10% sucrose, 0.1% mercaptoethanol, and 0.1% bromphenol blue. Variations exist for different applications and tissue types.**

5. It is essential to keep the extract cold to prevent deterioration of the enzymes (i.e. place grinding plate or tubes on ice).

6. Tough tissue may require powdered glass to improve homogenisation.

7. Soft, messy tissue works best if not over-homogenised.

8. Extracts should be used or frozen immediately. Frozen extracts should be good for several days, after which most enzymes will deteriorate considerably.

9. It is generally most convenient to prepare samples the day before running gels, as this allows an early start to the run.

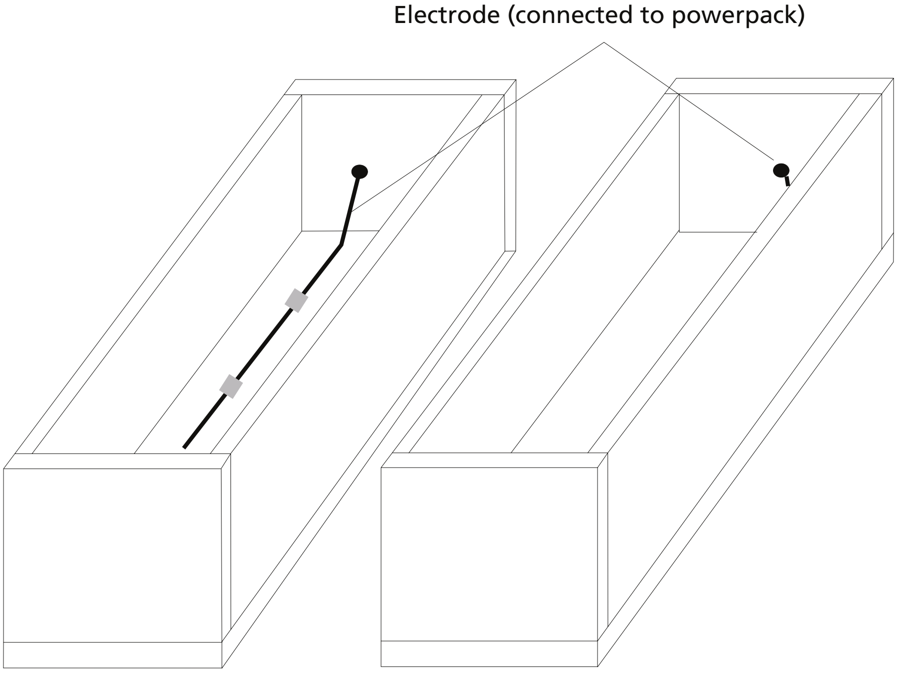
2.1.3. Electrophoresis

1. Add buffer to electrode boxes (Figure 3); the buffer is diluted from stock solutions. If possible, use 50% new buffer and 50% used buffer, in order to save money. Volumes: 500 ml per box if run two gels, 250 ml if run one gel.

2. Keep the sample extracts on ice, and soak up extract onto sample inserts (rectangles of filter paper, 5mm x 6mm - cut your own). Avoid getting bits of tissue on the inserts.

3. Remove the lids from the gel plates, and make a cut across the gel, parallel to the edge of the gel (if you use a scalpel, be careful not to cut into the Perspex mould). The

Figure 3. Example of a set of electrode boxes.



exact placement of this cut depends on the buffer and particular enzymes (Discontinuous buffers: 3 cm from end of gel; Continuous buffers: 3-6 cm, depending on whether any of the enzymes migrate cathodally (backwards, towards the black leads (–ve)).

4. Blot the excess liquid from the inserts using tissue paper, and line up along the open cut (Figure 4). Leave about 1.5mm between samples; there is room for 25-30 samples on each gel. It is a good idea to have replicates of some samples on different gels, to be certain of electrophoretic resolution on each gel.

5. Push the gel together, making certain that there are no gaps.

6. Place the gel with samples onto the buffer boxes, with the samples at the cathodal (black) end. Most enzymes will migrate towards the anodal (red) end.

7. Connect the gel to the buffer with sponges (well rinsed in distilled water).

8. Cover the gel with a sheet of plastic, to prevent dehydration and **accidental electrocution**.

9. Connect leads (red to red, black to black) to the buffer boxes, then to the power supply.

10. Turn on power to desired current or voltage. (Current varies with the number of gels; voltage does not).

11. Let it run until the bromphenol blue migrates about 10cm, or until the required separation has occurred (determined empirically for each system).

12. As the run approaches completion, prepare staining solutions. These will keep for hours in the refrigerator. Check to make certain that there is sufficient melted agar on hand.

13. Turn off the power supply and disconnect the leads at the power supply, before disconnecting the leads at the buffer boxes.

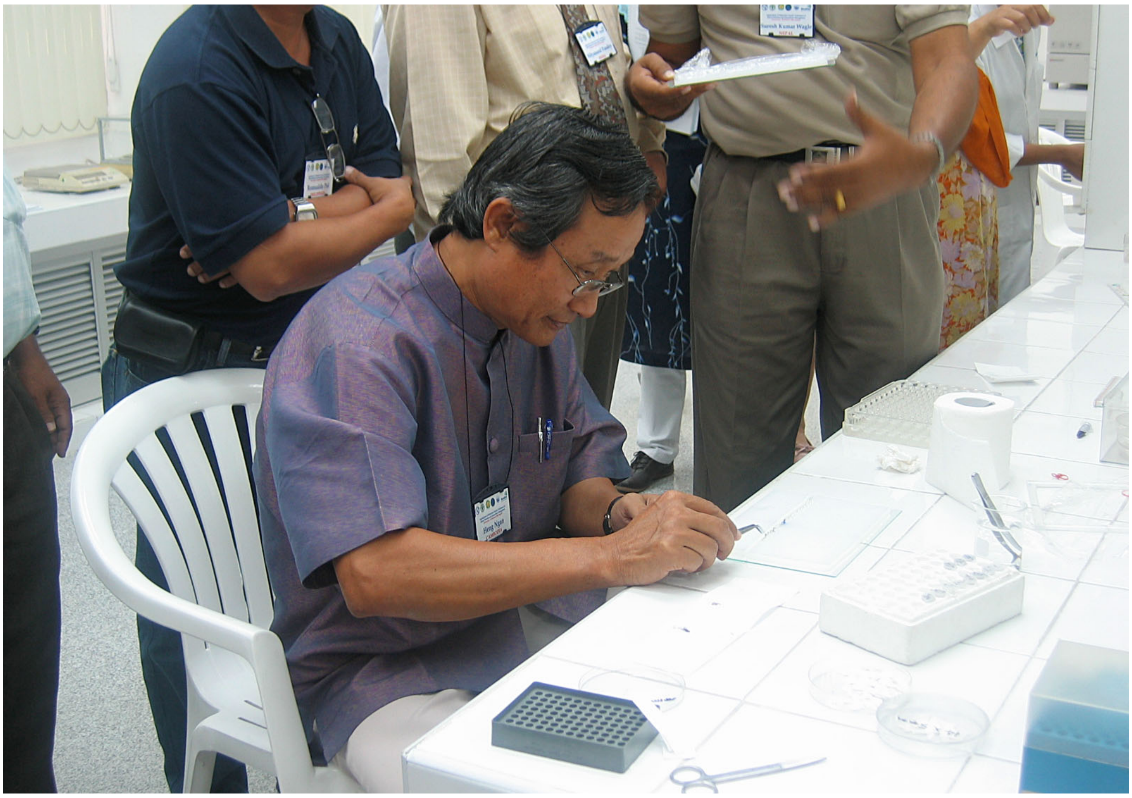
14. Remove and sponge off the plastic sheet.

15. Prepare the gel for slicing by cutting off excess sections, and marking one corner for orientation.

16. Slice gels with fishing line. Flip the top slice over for staining (remember to read it from right to left!)

17. Place slices on perspex sheets (no air bubbles underneath) for stains with an agar overlay; for liquid stains, use a glass dish.

Figure 4. Loading samples into a starch gel.



18. **Melted agar** is made by suspending the flask in a beaker of boiling water, until the agar is clear. Store in 65°C oven in flask. When using it, open the oven only as necessary, and return the flask to the oven immediately. Make certain that there are always two full 250 ml flasks in the oven during busy times.

2.1.4. Gel staining

See Annex 3 for a list of staining recipes for common enzymes. Staining can be conducted in two ways, one is to mix the staining chemicals, then add agar and poor over the gels, and another method involves dissolving staining ingredients into staining buffer, which is used to soak the gel. Some enzymes may stain well in one method but not the other.

Note that most of staining ingredients are relatively expensive, and either can cause allergy or are poisons. Information on effects of chemicals is given on the label or on a Materials Safety Data Sheet (MSDS) provided by the chemical supplier, so read carefully before dealing with them. Care must be taken during handling, gloves and a mask must be worn at all times during the staining process.

2.1.5. Gel scoring

If the staining is successful bands of enzyme/protein appear. The bands are then scored for presumed genotypes based on nature of each enzyme/protein.

A monomer protein/enzyme comprises of single polypeptide chain which is called a subunit. Therefore each homozygote would produce a single band of different mobility and the heterozygote would produce two bands of different mobility (Figure 5).

A dimer protein/enzyme comprises two polypeptide chains, either the same amino acid sequence or different. Therefore homozygote individuals produce only one type of polypeptide chain and would show one band while a heterozygote would produce 3 bands, with a thick one equally distant between the two outside bands. This is because there are twice as many ways of producing a band comprised of two different peptides as there is one comprising two identical peptides.

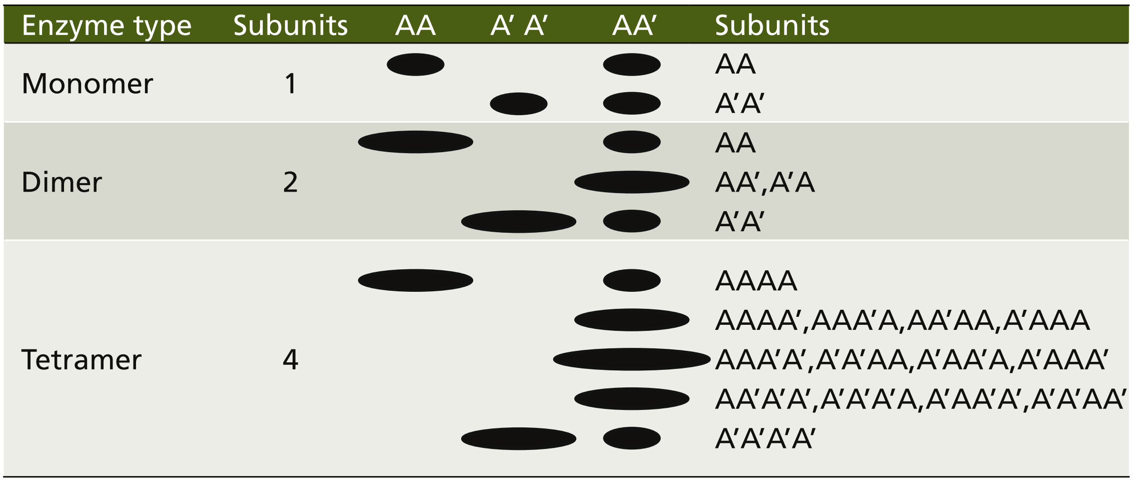
A tetramer comprises 4 polypeptide chains thus a heterozygote produces 5 bands. sometimes it is difficult to see the outer bands of a tetramer as they are approximately 1/6 of the intensity of the middle band.

Each staining may show products of more than a single locus. This means that multiple enzymes catalyse the same reaction but are products of different genes. They are termed “isozymes”. While the products of different alleles of a locus are termed “allozymes”.

2.1.6. Trouble shooting

Allozyme electrophoresis needs experience. It is quite difficult for beginners to get good results. However, the

Figure 5. Electrophoretic phenotypes when one locus is expressed.



Enzyme type	Subunits	AA	A’ A’	AA’	Subunits
Monomer	1	●		●	AA
			●	●	A’A’
Dimer	2	●		●	AA
				●	AA’,A’A
			●	●	A’A’
Tetramer	4	●		●	AAAA
				●	AAAA’,AAA’A,AA’AA,A’AAA
				●	AAA’A’,A’A’AA,A’AA’A,A’AAA’
				●	AA’A’A’,A’A’A’A,A’AA’A’,A’A’AA’
			●	●	A’A’A’A’

common problems are always minor and can be easily fixed. Table 1 shows some of the most common problems and solutions.

2.2. DNA extraction

There are many different and versatile methods for isolating genomic DNA, including a large variety of commercial kits. In this handbook, two methods that are most commonly used and inexpensive compared to commercial kits but result in reasonably good DNA quality are presented.

2.2.1. Salt precipitation method

This basic DNA extraction protocol is slightly modified from Crandall et al. (1999), and it works well for both fresh, frozen and ethanol preserved tissues. It is advisable to include a negative control (a control which does not contain any tissues sample but only extraction chemicals) when performing DNA extractions to ensure solutions are not contaminated.

Solutions required

- **Cell lysis solution**: 10mM Tris, 100mM EDTA, 2%SDS, pH 8.0
- **Protein precipitation solution**: 7.5M Ammonium Acetate
- **TE buffer**: 10mM Tris, 0.1 M EDTA, pH 8.0
- **70% ethanol**: 70ml of pure ethanol (molecular biological grade) with 30ml autoclaved deionised water.
- **Proteinase K**: 20mg/ml
- **Ribonuleanase-A (R-Nase)**: 10mg/ml

The TE buffer and Cell Lysis Solution should be autoclaved before use.

Table 1. Common problems in starch gel electrophoresis and solutions.

Problem	Possible cause(s)
Lumps in gels	Flask was not swirled well during cooking process
Gel breaks during moving after slicing	Gel can be under- or over-cooked Gel may have been run under high voltage, or gets heated up during the run
No staining	At least one staining ingredient is missing Staining chemicals may be degraded Is the enzyme needed to be viewed under UV light?
Enzymes migrate reversely	Gels were not placed in right direction Enzymes may migrate cathodally Electrodes may be connected in the reverse orientation
Enzyme stains weakly (bands are faint)	Chemicals may be not fresh, or may be degraded Amount of tissue is not sufficient, or too much homogenising buffer was added
Enzyme stains too intensely (bands are to thick/smeary)	Substrate solutions or linking enzymes are too concentrated Too much tissue, and not sufficient homogenising buffer
Air bubbles occurred in the gel	Improper pouring
Bands are curved	Gels are warp, may be caused by heating up during the run (system can be run in a fridge at 4°C), or gels are not evenly cooked Wicks are not in touch evenly with gels

Protocol

1. If you are using fresh tissue skip to step 2. If using ethanol preserved tissue you need to eliminate as much ethanol as possible. Dissect about 50 mg tissue and place on a clean paper towel and press out the sample a couple of times to remove ethanol. For fin clips or other tissue types not very absorbent this is adequate. For absorbent tissue (eg. liver) the sample is then placed in a 1.7 ml tube with 900 µl TE buffer. Mix well and centrifuge for 1 minute at high speed. Draw off the TE buffer and blot the tissue on a paper towel and proceed to step 2.

2. Pipette 720 µl of Cell Lysis Solution to a 1.5 ml Eppendoff tube and add tissue sample (about 50 mg).

3. Add 5 µl of Proteinase-K (20 mg/ml) to each tube. Mix by inversion 10-20 times and incubate at 55°C for several hours or overnight (with periodic mixing) until tissue is completely dissolved. Once digested cool to room temperature.

4. Add 4 µl Rnase-A (10 mg/ml) to each sample and mix by inverting tube 25 times. Incubate samples at 37°C for 1 hour then cool to room temperature.

5. Add 300 µl of Protein Precipitation Solution to samples and vortex vigorously for 20 seconds to mix samples. Incubate on ice (or in freezer) for 30 minutes. Centrifuge at high speed for 3-5 minutes. The precipitated protein should form a tight pellet in the bottom of the tubes. If not repeat the vortex and ice incubation steps.

6. Pour off the supernatant containing DNA (leaving behind the precipitated protein) into a 1.7 ml centrifuge tube containing 720 µl of isopropanol. Mix by inverting gently 25-50 times. Centrifuge at high speed for 5 minutes to pellet the DNA (placing tubes at -20°C for several hours or overnight aids in DNA precipitation). Pour off the supernatant and briefly drain tube on a clean paper towel. Add 750 µl 70% ethanol and invert tube several times to wash DNA. Centrifuge at high speed for 1 minute. Carefully pour off the ethanol. The DNA pellet may be loose so make certain that you don't discard it. Drain the tube on clean paper towel and allow DNA pellet to dry (room temp or 37°C oven). It is essential that the pellet is dry but if it is over dry it is difficult to rehydrate.

7. Depending on the size of the pellet, add 30-200 µl TE buffer to the dried DNA pellet. Some literature suggests that TE may inhibit subsequent PCR reactions. Although this does not seem to be a major problem it may be advisable to store DNA in 1/10 TE or deionised water. Allow the DNA to rehydrate overnight at room temperature. Store at 2-8°C short term or –20°C long term.

8. DNA extraction of some tissues can be difficult by this method and if this is the case there are several other methods available. They are generally more tedious and use some potentially nasty chemicals that need to be used in a fume hood.

2.2.2. Phenol chloroform method

Solutions required

- TNES-urea: 10mM Tris-HCl pH 7.5, 125mM NaCl, 10 mM EDTA pH 7.5, 0.5% SDS, 4M Urea)

- TE: 10 mM Tris-HCl pH 7.5, 1mM EDTA

- 2xCTAB: 100mM Tris pH 8.0, 1.4M NaCl, 20mM EDTA pH 8.0, 2% CTAB

- 70% ethanol: 70ml of pure ethanol (molecular biological grade) with 30ml autoclaved deionised water.

- Proteinase K: 20mg/ml

- Phenol chloroform

- Isoamyl alcohol

Protocol

1. Place tissue in 200µl TNES.

2. Grind up tissue using plastic mortar or tweezers. Add 500µl 2xCTAB to tube.

3. Add 5µl of proteinase-K (20mg/ml solution).

4. Vortex briefly, incubate at 65°C for 1 hour.

5. Add 600µl of chloroform-isoamyl (24:1); mix well and centrifuge for 15 minutes at room temperature.

6. Pipette off supernatant, add to new tube, extract with 600µl phenol-chloroform-isoamyl (25:24:1).

7. Extract one final time with chloroform-isoamyl (24:1).

8. Add 600µl cold (-20°C) isopropanol, mix gently but thoroughly. White stringy pellets are formed.

9. Let it sit for at least 1 hour, then spin in cold room for at least 30 minutes.

10. Pipette off supernatant, add 1 ml 70% cold (-20°C) ethanol. Mix gently, then spin in cold room for 5 minutes. Repeat once more to ensure that all salts are removed.

11. Dry pellet in vacuum centrifuge for 25 minutes or until ethanol is evaporated.

12. Add 100-200µl of TE and let sit for several hours at room temperature.

2.2.3. DNA extraction from fish blood

The following protocol is developed by Doug I. Cook and Danielle Paquet (Marine Gene Probe Laboratory, Dept. of Biology, Dalhousie University, Halifax, NS B3H 4J1, Canada), Wongpathom Kamonrat (National Aquaculture Genetics Research Institute, Dept. of Fisheries, Chatujak, Bangkok 10900, Thailand), and Elizabeth R. Pitman (Dept. Of Renewable Resources, University of Alberta, Edmonton).

Blood samples

Blood samples should be preserved in EtOH at about 1:1 (v/v) ratio and mixed well. Preserved samples are then stored in fridge.

Reagents

- High TE: 100 mM Tris-HCl, 40 mM EDTA
- MGPL lysis buffer: 10 mM Tris-HCl, 1 mM EDTA, 200 mM LiCl, 0.8% SDS
- Proteinase K
- TE: 10 mM Tris-HCl, 1 mM EDTA
- Isopropanol
- 70% ethanol
- NaCl

Procedures

1. An aliquot of 20-100 µl* of the blood/ethanol was transferred to a 1.9 ml microcentrifuge tube containing 1 ml of high TE (100 mM Tris-HCl, 40 mM EDTA, pH 8.0). The sample was vortexed and pulse spun for 10-15 seconds to pellet the cells. The supernatant was then removed.

2. The cell pellet was resuspended in 250 µl MGPL lysis buffer (10 mM Tris-HCl, 1 mMEDTA, 200 mM LiCl, pH 8.0 and 0.8% SDS).To this was added Proteinase K to a final concentration of 200 µg/ml. The sample was then incubated at 45°C. After 15 min the sample was vortexed briefly and further incubated until the cells were completely lysed (10-20 mm).

3. Following incubation, 500 µl of TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) and

4. 750 µl of cold isopropanol were added, and the sample mixed by vortexing.

5. The DNA was then pelleted by a 1 minute spin at 14,000 rpm. The supernatant was removed and the DNA pellet was washed with 70% ethanol, pulse spun and the ethanol removed. This was followed by a second pulse spin and removal of the residual ethanol.

6. The sample was air dried for 5 to 10 minutes and then resuspended in 100 µl of TE. The samples were then adjusted to the appropriate template concentration for use in PCR.

7. *The amount of blood/ethanol used in the extraction (step 1) must be determined for each set of samples and for each species since the initial dilution factor of the blood may vary between collections and the amount of DNA in the cells nay vary from species to species. Both of these factors will affect the yield from the extraction and must be taken into account.

8. With the following modifications this procedure has been used to extract DNA, of sufficient quality for PCR, from samples of skeletal muscle, tail fin (Ruzzante et al. 1996) and adipose fin (D. Cook, unpublished results):

9. The time of digestion with Proteinase K (step 2 above) was extended.

10. Prior to the addition of isopropanol (step 3 above) the samples were spun, for one minute, to pellet insoluble debris and the supernatant was transferred to a clean tube.

11. NaCl to 50 mM was added after the TE (step 3 above). This was necessary, to ensure efficient recovery of DNA, because of the lower concentration of DNA obtained from tissue samples compared to that obtained from blood.

2.3. DNA quality and quantification

2.3.1. Solutions required

- **50X TAE (generally diluted to 1X for use)**: 242 g Tris; 700 ml ddH$_2$O; 57.1 ml glacial acetic acid; 100 ml 0.5 M EDTA pH 8.0. Adjust volume to 1L using dd H$_2$O. Autoclave and store at room temperature.

- **1X TAE agarose gel (1%)**: 2 g agarose; 200 ml 1 X TAE; heat in microwave oven until agarose dissolved (do not boil). Store in 65°C oven.

- **Load dye**: 50 mM EDTA; 30% Glycerol; 0.25% bromophenol blue; 0.25% xylene cyanol

- **Marker**: 50 mM EDTA; 20 μl load dye; 175 μl 1 X TAE; 5 μl DNA marker (HindIII digest of λ phage DNA for example)

- **Ethidium bromide**: HAZARD!!! Wear double gloves when handling.

2.3.2. Methods

There are two methods to quantify DNA sample. The conventional method of electrophoresis of DNA sample of unknown concentration with a known standard is applied in most labs where a spectrophometer is not available. Another method is using a spectrophometer that directly quantifies the DNA concentration in the resuspended DNA extraction.

Electrophoresis of a DNA sample of unknown concentration with a known standard

1. Place the gel plate into gel mould, position the comb and ensure that the gel is horizontal – check with a spirit level is necessary (different supplier have different designs, do follow the instructions from manufacturers).

2. Prepare a 1% agarose gel: dissolve 1g agarose in 100 ml 0.5x TBE or 1x TAE. Heat the mixture in a microwave oven until completely dissolved. Cool to 60°C.

3. Pour agarose onto the gel tray and allow it to set for at least 30 min.

4. Remove the comb. Place the gel into the electrophoresis tank and pour 0.5xTBE or 1xTAE (same as the buffer that was used to make gel) until the gel is completely covered.

5. Mix 1 μl loading dye and 2 μl DNA and load into the well.

6. Load 2 of DNA marker (HindIII digested λ DNA for example) into one of the wells.

7. Run the gel at 70-100V until the dye is about 2.5 cm from the origin.

8. Move the gel to a tray with ethidium bromide (1µl ethidium bromide in 100 µl ddH$_2$0) (HAZARD!!!). Let the gel stain for 5-10 min, and then de-stain for about 2 min in another container with ddH$_2$O only.

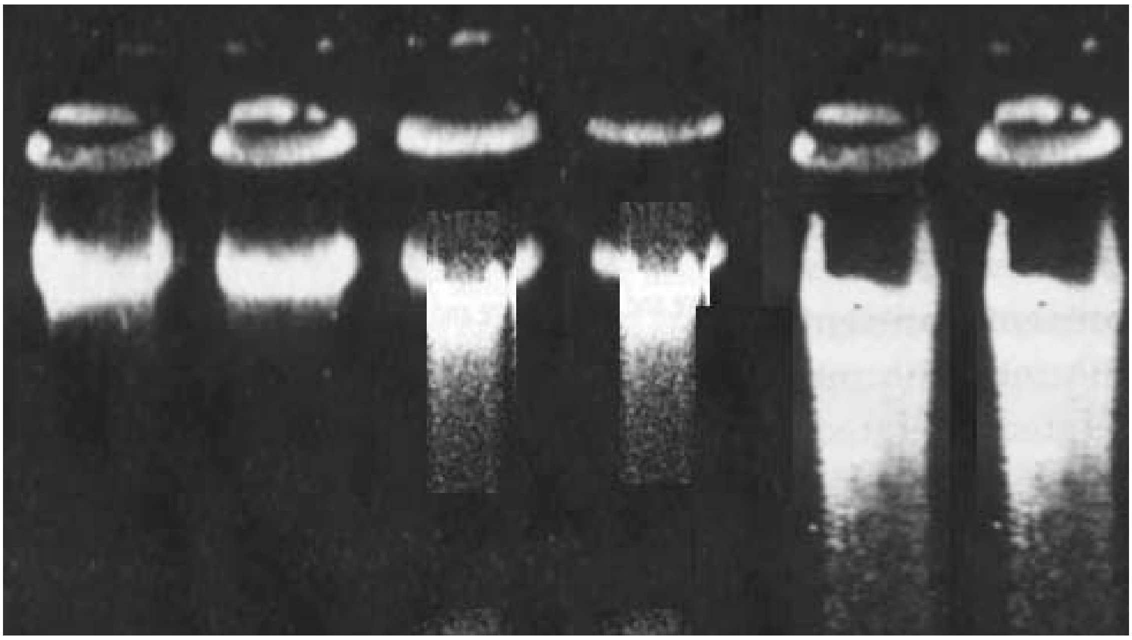
9. Illuminate the gel with UV light (CAUTION – UV LIGHT IS HAZARDOUS!!! – WEAR MASK OR UV PROTECTION GLASSES IF EXPOSED TO UV LIGHT).

10. Photograph the gel under the UV.

11. Compare the intensity of the DNA bands of the samples with the intensity of the λ bands. As the amount of DNA present in each λ band is known (information is often provided by the supplier), the amount of DNA of each sample can be estimated by comparing the fluorescent yield of the sample with those of the λ bands.

12. Quality of extracted DNA can also be assessed by looking at the gel. Good quality DNA will show as a sharp intense band. Degraded DNA extracts will show various degree of smearing. See Figure 7 for example.

Figure 6. Agarose gel analysis of genomic DNA isolated from *Tor tambroides*: 1, 2: good DNA quality; 3-4: DNA with RNA; 5-6: degraded DNA; 7-8; DNA with salts.



Spectrophotometric determination of DNA concentration

Dilute 1.5 µl of DNA to 1500 with deionised water and read at A_{230}, A_{260} and A_{280}. The A_{260}/A_{280} ratio provides an estimate of the purity of the DNA. In a pure sample, this ratio is approximately 1.8. Lower values indicate protein or phenol contamination. A_{230} should be less than A_{260} and may be the same as A_{280}. High A_{230} reading indicates that residual phenol remains in the preparation. An A_{260} of 1 corresponds to approximately 50 µl/ml of double-stranded DNA in a 1 cm quartz cuvette. Nucleic acid concentration is calculated as follows:

A_{260} * 50 mg/µl * 0.001 µl/ml * dilution factor (1500 µl/1.5 µl) (µg/µl)

2.3.3. Trouble Shooting

Common problems and solutions are summarised in Table 2.

2.4. Polymerase chain reaction (PCR)

The advent of PCR has greatly accelerated the progress of studies on the genomic structure and processes of various organisms. Any gene region, even in highly complex genomes, can be specifically amplified using this technique if the flanking nucleotide sequences are known (Saiki et al. 1988). The PCR procedure involves replicating target regions of DNA, which are flanked by regions of known sequences (Erlich 1989). Synthetic nucleotide primers (usually 18-30 bp long) that are complementary to each of the

Table 2. Common problems and appropriate solutions.

Problem	Possible causes / solutions
Lysis of the cell is incomplete	• Incubation time is not long enough • At least one of the ingredients is lacking in cell lysis solution, especially SDS. Check by shaking the tube, if no foam formation is seen, then add 100 µL of 10% SDS. If there is foam formation, add 5 µL of Proteinase K (20mg/ml) and continue to incubate for one hour
No DNA precipitation	• Cell lysis is not complete • Short centrifuge time and/or low speed • Wrong precipitation solution (ethanol/isopropanol)
No DNA on gel	• DNA pellet is lost during wash step • The gel was not stained with ethidium bromide
DNA extraction fail from preserved samples	• The two methods described in this manual are often used to produce reasonably good DNA quality. If these two methods fail, especially when dealing with preserved samples, the best solution is probably to use commercial DNA extraction kits.

flanking regions are needed. These are combined with a small amount of DNA (at nanogram levels), plus free deoxynucleotides (dNTPs), a reaction buffer, and *Taq* DNA polymerase (isolated from the hot spring bacteria *Thermus aquatica*). During a series of up to 30 heating and cooling cycles, the DNA is denatured into single-stranded molecules, the two primers anneal to their complementary sequences on either side of the target region, and the DNA polymerase replicates the region downstream from each primer. The amount of target DNA will double with each temperature cycle, so that even low starting copy numbers of the target sequence will generate trillions of copies by the end of the last cycle.

PCR can significantly decrease the amount of time required to isolate a desired segment of the genome. Also, PCR allows DNA analysis to be performed from small and sometimes minute tissue sample. However, for most uses of PCR, one must determine the sequences of regions flanking a given locus, and this can entail considerable effort when working with a new species. The use of arbitrary primers (in random amplified polymorphism DNA - RAPD), does allow one to identify genetic markers relatively quickly in a species for which extensive sequence information may not be available. Furthermore, a significant number of universal primers have been developed and used to amplify a number of gene regions in different species.

The following basic PCR protocol is from *The Simple Fools Guide to PCR* by Palumbi *et al.* (1991). The amounts listed below are for 25 µl reaction volumes so if you use larger or smaller volumes, adjust each component accordingly. Again, it is advisable to include a negative control in each round of PCR to ensure solutions are not contaminated. Similarly, the inclusion of a positive control (i.e. a sample the has been amplified successfully previously) can help isolate problems should amplification fail.

2.4.1. Chemicals required

- 10X PCR buffer (supplied with Taq)
- 50mM $MgCl_2$ (supplied with Taq)
- Taq DNA polymerase
- 2.5 mM of each dNTPs, mix together
- 10 µM each primer
- Extracted genomic DNA
- Autoclaved deionised H_2O

2.4.2. PCR mixture

Table 3 represent the volume and concentration of each component in one PCR reaction of 25 µL total volume.

To minimise pipetting errors, add the components together in a master mix except for sample DNA and *Taq*. Simply multiply the amounts per reaction by the number of reactions to be performed plus two, one for the blank and one as a spare (in case of pipetting error). The *Taq* is added at the last moment. Just before adding the DNA template, as it degrades very

Table 3. Basic PCR mixture.

Component	Volume (µL)	Concentration
10X Buffer	2.5	1X
2.5mM dNTPs	2.0	0.2µM
50mM $MgCl_2$	2.5	0.5µM
10 µM Primer 1	0.5	0.02µM
10 µM Primer 2	0.5	0.02µM
ddH_2O	16.95	
5U/µL *Taq*	0.05	0.1 U
Genomic DNA	0.5	0.5-1ng/µL

rapidly. Keep it out of the freezer for a minimal period and keep on ice whilst processing.

In order to save time during PCR preparation, it is advisable that a PCR mixture sheet should be attached to the notebook. This is because the number of samples used for PCR varies each time, sometimes it is 5 (therefore need to prepare for 7 – one extra for negative control, and another one for pipetting error), another time it is 10 (prepare enough for 12) for example. Table 4 below shows an example PCR mixture sheet:

2.4.3. Thermal cycling

For most primers, thermal condition for PCR will be recommended in previous publications. However, if you intend to use primers for which there is no information in the literature a good starting point is:

Initial denaturing at 94°C for 3 minutes to completely dissociate double stranded DNA

Perform 30-40 cycles as follows:

- Denature at 94°C for 15-30 seconds (depending on length of desired PCR product)

Table 4. An example of PCR mixture sheet.

Solution	1 sample	2 samples	3 samples	…	N samples
10X PCR buffer	2.50 µl	5.00	7.50	…	N x 2.50
dNTPs (2.5 mM each)	2.00 µl	4.00	6.00	…	N x 2.00
$MgCl_2$ (50 mM)	2.50 µl	5.00	7.50	…	N x 2.50
Primer 1 (10 µM)	0.50 µl	1.00	1.50	…	N x 0.50
Primer 2 (10 µM)	0.50 µl	1.00	1.50	…	N x 0.50
ddH_2O	16.95 µl	33.9	50.85	…	N x 16.95
Taq	0.05 µl	0.1	0.15	…	N x 0.05

- Annealing at (**Tm**-10°C) for 15-30 seconds. Tm is provided with primers. The standard estimate of annealing temperature is:

  Tm (°C)= [(G+C)x4]+[(A+T)x2]-5

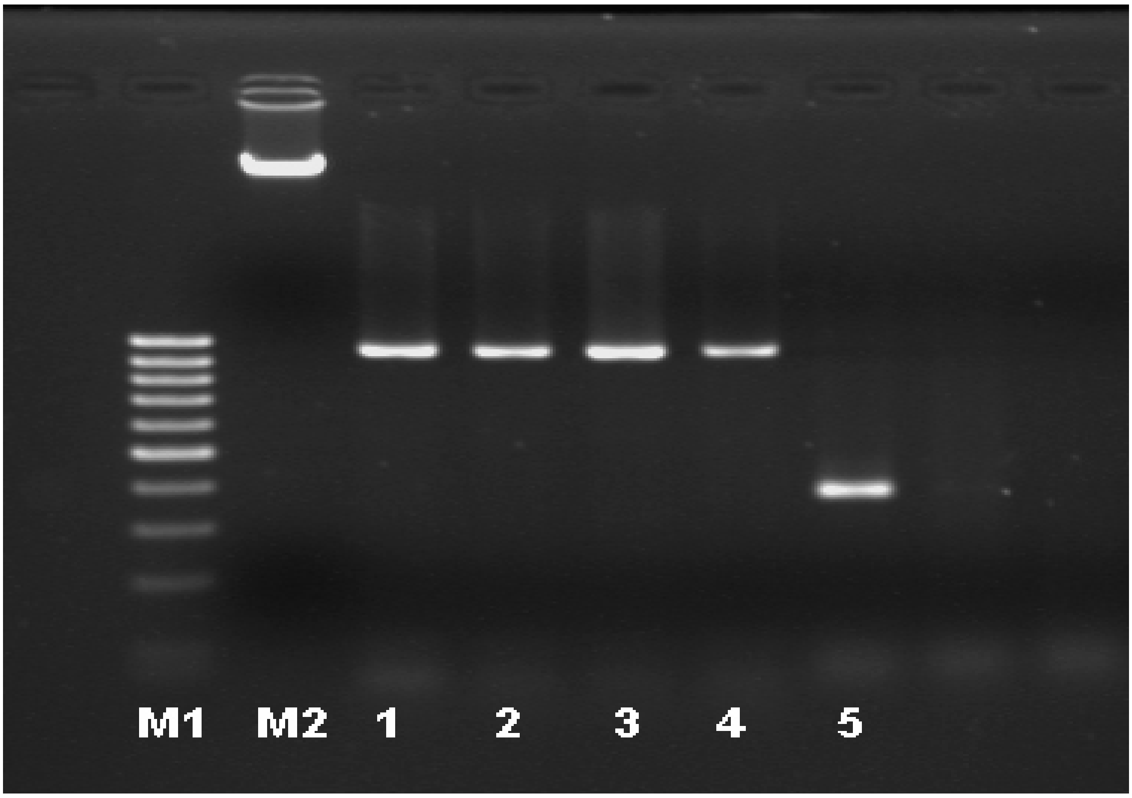
- Always use the lower Tm of the two primers

- Extension at 72°C for 15-90 seconds (depending on the length of the gene fragment to be PCR)

- A further extension step of 2 minutes at 72°C following the last cycle.

This basic protocol can be adjusted to maximise PCR efficiency if required. Generally annealing temp may be lowered if the primers were not "perfect" or increased for increased specificity. Also, concentration of $MgCl_2$ can also be changed; it is advised that a series of $MgCl_2$ concentration should be tried and a general observation is that high concentration of $MgCl_2$ provides higher yield, however it may also produce non-specific PCR products.

Figure 7. Verification of the PCR product on gel. (Lane M1: 100bp DNA ladder; lane 1-4: PCR products of ATPase6-8 fragment amplified from *Tor* species DNA, fragment is approximately 1000 bases long; lane 5: PCR products of cytochrome *b* fragment amplified from *Tor* species DNA, the fragment is approximately 450 bases long; the last two lanes are negative control.



2.4.4. Visualising PCR products

PCR products are assessed in the same way as DNA extracts (see Protocol 4). However, rather than a high molecular weight DNA band close to the sample origin the PCR product should migrate at the rate appropriate to its predicted size. This is verified by comparison with the DNA marker (which also assesses quantity). If the predicted size of the PCR product is small, be careful not to run the gel too long so as samples do not run off the gel.

Before the PCR product is used in further applications, it has to be checked if:

1. **There is product formed.**
   Not every PCR is successful. There is for example a possibility that the quality of the DNA is poor, that one of the primers doesn't fit, or that there is too much starting template, or even at least one ingredient was left out!

2. **The product is of the right size.**
   It is possible that there is a product, for example a band of 500 bases, but the expected gene should be 1800 bases long. In this case, one of the primers probably fits on a part of the gene closer to the other primer. It is also possible that both primers bind to a totally different gene (non-specific priming).

3. **Only one band is formed.**
   As in the description above, it is possible that the primers fit to the desired locations, and also at other locations. In this case, different bands may be present in one lane on a gel.

2.4.5. Trouble shooting

Common PCR problems are summarised in Table 5.

2.5. Restriction fragment length polymorphism (RFLP)

RFLPs were the first DNA markers to be used by population biologists (Parker et al. 1998). The technique involves cutting a DNA strand at specific nucleotide sequences using a restriction endonuclease and thereby producing a pool of different sized DNA fragments. RFLP variation can be visualised directly by staining with ethidium bromide following electrophoresis of the DNA in an agarose gel. This can be done for small molecules, such as the entire mitochondrial DNA, which produce a manageable number of fragments with many restriction enzymes (Landsmann et al. 1981; Tegelström 1992). The most appropriate method of analysis involves restriction sites, whereby actual sites are mapped to specific positions on the strand of DNA of interest. Scoring is based on the loss or gain of a site, this giving an accurate resolution of relationships. An alternative method is fragment analysis which scores the different fragments as either present or absent. However, this method assumes that fragments of similar length on a

Table 5. Common PCR problems.

Problem	Possible cause(s)
No PCR products	• Mg^{2+} concentration is not optimal • Amount of DNA template is not optimal • An enzyme inhibitor is present in the reaction (residue ethanol, phenol) • Primer annealing temperature is too high • Primers are degraded or not optimal • Incomplete template denaturation • No DNA template was added in the reaction
Low fidelity	• Mg^{2+} concentration is too high • dNTPs concentration is too high or unbalanced • Mispriming caused by secondary structure of template • Damaged template DNA
Non-specific bands	• Mg^{2+} concentration is too high • dNTPs concentration is too high or unbalanced • Mispriming caused by secondary structure of template • Primers are degraded or sequence is not optimal • Annealing temperature is too low • Primer concentration is too high • DNA contamination
Smeared bands	• Annealing temperature is too low • Mg^{2+} concentration is too high • dNTPs concentration is too high or unbalanced • Mispriming caused by secondary structure of template • DNase activity (smears visible on gel below expected band size)
Low yield	• Annealing temperature is too high • Template is not clean or degraded • An enzyme inhibitor is present in the reaction (residue ethanol, EDTA, phenol) • Extension temperature is too high • Too many cycles • Primer concentration is too low
No product but +ve control amplified	• DNA extraction unsuccessful

gel are homologous. Unfortunately, this assumption can be misleading since multiple fragments may make up a single band on a gel. In addition, different cleavage sites may produce similar banding patterns, thus giving erroneous relationship among samples (Fetzner & Crandall 2001).

A more efficient approach which is now used relatively commonly is to amplify the DNA region of interest then conduct restriction analysis on the amplified fragment. PCR products are treated with restriction enzymes and the fragments separated on an agarose gel and visualised by ethidium bromide staining to identify RFLP profiles (Karl & Avise 1992). An advantage of this approach is that the PCR fragment can first be sequenced from a number of individuals in the first instance, to allow detection of polymorphic restriction sites. Mitochondrial DNA is most commonly used for this method of genetic analysis as it has been proven to allow easy detection of genetic differences at population levels (Hillis et al. 1996).

2.5.1. Restriction digest mix

For one reaction:

- Restriction enzyme buffer 1.0 µl
- Restriction enzyme 0.2 µl
- PCR product 8.8 µl

Incubate for at least one hour at required temperature (enzyme manufacturer's instruction). Once incubation time is up, place the tube on ice to bring down the condensation, then freeze until required.

Similar to setting up PCR reaction, it is advisable to make a restriction digest mix sheet for multiple samples. Make a master mix and then aliquot the appropriate volume to each tube. PCR product is added last.

2.5.2. Running buffer

10x TBE (dilute to 1x for use):

- 10.8 g Tris
- 5.5 g Boric acid
- 0.744 g EDTA
- make up to 1 litre with dd H_2O

2.5.3. Running a gel

1. Gels used for RFLP analyses are often made of 2% Agarose in 1x TBE buffer.

2. Place TBE gel into running chamber, pour 1x TBE buffer over the gel.

3. Load all (10 µL) digested PCR product into a well (mix with loading dye). Load marker for justifying band size.

4. Run the gel at 150V, 60 mA.

Table 6. An example of restriction digest mix for multiple samples.

Chemicals	1 sample	2 samples	…	N samples
Amount of restriction enzyme buffer	1.0 µl	2.0	…	N x 1.0
Amount of restriction enzyme	0.2 µl	0.4	…	N x 0.2

5. Once the run is finished, place the gel in a water bath with 1% ethidium bromide (HAZARD!!!) for 10 min, then destain in distilled H_2O for 5 min.

6. View the gel under UV light.

2.5.4. Documentation of results

Restriction digest of PCR product results in banding patterns as shown in Figure 8 when running on an agarose gel. Each pattern is a haplotype and often assigned a letter (e.g. A, B, C…) or a number (1, 2, 3…). Example from the gel in Figure 8 shows that individuals 1 and 2 have the same haplotype A, individuals 3, 4, 7 have haplotype B and individuals 5, 6 and 8 have haplotype C.

When more than one restriction enzyme is used, the final haplotype is the combination of all haplotypes that resulted from single restriction digest (see Table 7).

2.5.5. Trouble shooting

Common problems with RFLP are summarised in Table 8.

Figure 8. An example of RFLP results.

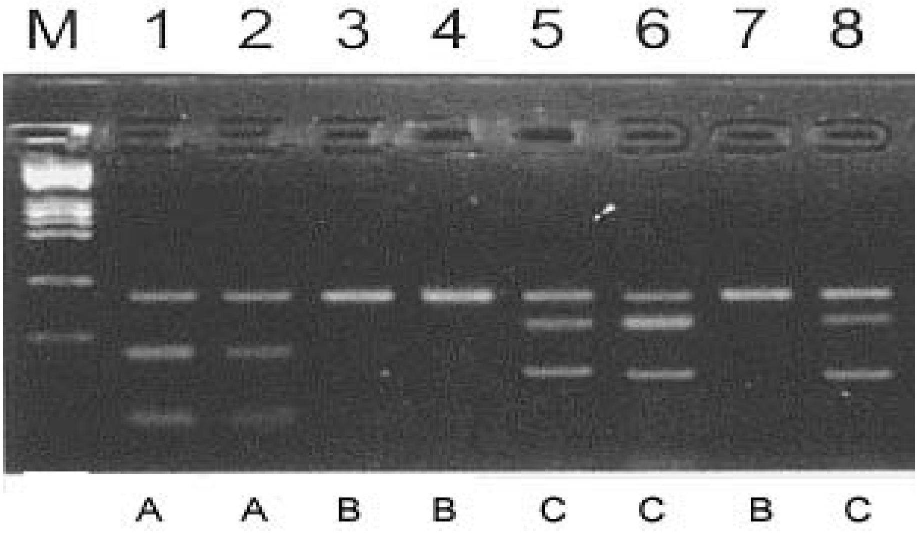


Table 7. An example of how to score and arrange RFLP results.

Individual	RsaI	EcoI	FokI	Combination
1	A	A	A	AAA
2	A	B	B	ABB
3	B	C	B	BCB
4	B	C	C	BCC
5	B	B	C	BBC
…	…	…	….	…

Table 8. Common problems with RFLP.

Problem	Possible cause(s)/ Solution(s)
Not complete digestion	• Not enough enzyme • Not following enzyme manufacturers' instruction about incubation temperature and time
No digestion	• Enzyme does not work, e. g. check sequences, enzyme age • Enzyme not added to master mix
Slow electrophoresis	• Excess digested products were loaded into wells
Gel warp or bands do not run	• Gel made with water, not buffer
Missing short bands	• Electrophoresis time is too long, short bands run out of the gel

2.6. Single strand conformation polymorphism (SSCP)

SSCP analysis is technically the simplest method available for rapid screening of DNA fragments for nucleotide sequence polymorphisms. The method relies on variable electrophoretic mobility of secondary structures formed by single-stranded DNA (ssDNA) fragments of different primary structure (i.e. nucleotide sequence). PCR products are denatured by heat (95°C) and immediately placed on ice. Under these conditions most of the single strands will not anneal to its compliment. Instead most of the fragments will fold upon themselves into stable conformations according to base pairing rules (G-C & A-T). Sequence differences cause different conformations that are detected by electrophoresis.

2.6.1. Chemical preparation

Formamide loading dye:

- 950 µl formamide, 50 µl load dye

SSCP polyacrylamide gel:

- 8% acrylamide (37.5:1 acrylamide: bis-acrylamide)
- 5% glycerol
- 0.5X TBE buffer (44.5 mM *Tris*, 44.5 mM boric acid, 1 mM EDTA) (TBE is usually made at 10X concentration and diluted for use)
- Start with 40% acrylamide and 0.625X TBE containing 6.25% glycerol
- For 50 ml total volume; 10 ml acrylamide + 40 ml TBE/glycerol
- Catalyze with 400 µl 10% ammonium persulfate (APS) and 100 ul TEMED (N,N,N',N'-tetramethylenediamine)
- Stir and pour gel immediately. Allow to set for 2 hours

2.6.2. Sample preparation

1. Mix 2 ul PCR product with 7 ul formamide loading dye.

2. Heat mixture to 95°C for 5 minutes and immediately place on ice (keep on ice until you load sample).

2.6.3. Gel running

1. Running Buffer is 0.5X TBE.

2. Cool gel in refrigerator for 1 hour and then pre-run for 15 minutes at 5W constant power prior to loading.

3. Flush wells with buffer and load samples.

4. Run at 5W for 2-10 hrs (you must determine this yourself but remember that single strand DNA migrates slower than double stranded DNA).

5. It is important power be kept constant throughout the run.

2.6.4. Silver staining

- **Wash solution**: 10% ethanol, 0.5% acetic acid

- **Silver nitrate solution**: 0.1% silver nitrate

- **Developing solution**: 1.5% sodium hydroxide, 0.15% formaldehyde (4ml/l 37% formaldehyde stock), 0.001% sodium borohydride

- **Fixing solution**: 0.74% sodium carbonate

- **All solutions should be made fresh just before use**. All staining steps are performed with gentle shaking of the gel at room temperature in a fume hood. Use enough of each solution to cover the gel.

1. Remove the gel from electrophoresis apparatus, separate the glass plates and remove the spacers. Transfer the gel on the bottom glass plate to a clean staining tray. The gel may float off the plate during staining and can be removed.

2. Wash gel twice for 3 minutes in wash solution. Take care to remove the last of the wash solution before adding the silver nitrate solution.

3. Incubate the gel in silver nitrate solution for 10 minutes (remove solution).

4. Rinse the gel twice for 10 seconds in distilled water (remove).

5. Add a little developing solution (~10 ml) to the tray. A black precipitate will form which should be removed before adding the rest of the solution. The length of time required for staining depends on the amount of DNA loaded, but is usually ~20 minutes.

6. Stop the development by adding fixing solution. Leave for 10 minutes. Any longer and the gel will swell.

7. The gel can now be photographed or dried onto filter paper.

2.7. Random amplified polymorphic DNA (RAPD)

RAPD markers are produced by PCR using short oligonucleotide primers of random sequences. Different RAPD patterns arise when genomic regions vary according to the presence/absence of complementary primer annealing sites. The primers are typically 10 bp long (Williams et al. 1990) and no specific knowledge of a particular DNA sequence is required. Allelic variation usually consists of the presence or absence of particular amplification products, which can be separated on agarose gels stained with ethidium bromide. The RAPD process typically reveals several polymorphic genetic segments per primer within populations; other segments may appear as monomorphic bands within or across populations (Hadrys et al. 1992). The degree of variability observed for many primers suggests that the technique will be useful for a variety of questions, including individual identification, pedigree analysis, strain identification, and phylogenetic analysis (Parker et al. 1998).

RAPD markers are rarely inherited as codominant alleles (Parker et al. 1998). Losses of a priming site results in complete absence of the enclosed amplified segment, not simply a shift in mobility on the gel. In heterozygotes, therefore, differences may appear only as differences in band intensity, which is not usually a reliable phenotype for PCR analysis. As a consequence, information on the parental origin of alleles may be inaccessible for RAPD markers, as compared to codominant nuclear markers revealed by RFLPs, allozymes or microsatellites (Lewis & Snow 1992) although it should be noted occasionally RAPD markers have been detected that show patterns consistent with codominant markers which contain microsatellite loci (Garcia & Benzie 1995). Because of their short length, RAPD markers may produce some artifactual amplification products, and careful control of DNA quality and amplification conditions is necessary to ensure reproducible banding patterns (Carlson et al. 1991; Riedy et al. 1992; Scott et al. 1993). Ideally and where possible, undertake crosses of known genotypes to demonstrate Mendelian inheritance of bands, and repeat experiments to confirm banding patters before scoring.

2.7.1. PCR preparation

Stock and final concentrations per 25 µl of reaction mixture is presented in Table 9.

2.7.2. DNA amplification

Place PCR tubes in a thermal cycler. Amplify using the most commonly used temperature profile as shown in Table 10.

After amplification remove the PCR tubes from the thermal cycler. Add 3 µl of 10x loading buffer to each tube. Mix by flicking the bottom of the tube and spin to collect the mixture. The mixture is now ready for loading in the agarose gel. It is may be advisable to run and

Table 9. Stock and final concentration of RAPD-PCR.

Components	Stock concentration	Final concentration	Vol/reaction
dNTPs	100 mM	0.2 mM each	0.2 µl
PCR buffer	10x	1x	2.5 µl
$MgCl_2$	50 mM	2.5 mM	1.25 µl
Taq	5 u/µl	0.5 u/rxn	0.1 µl
Primer	10 µM	0.4 µM	1.0 µl
ddH_2O			17.35 µl
DNA	2 ng/µl	5 ng/µl	2.5 µl

check for the presence of PCR products before running on bigger gels to check for RAPD bands.

2.7.3. Electrophoresis

Get a gel mould and seal both edges with 1" masking tape. Place in a level platform and attach combs. Some gel making design may not need this step.

1. Prepare 1.4% agarose by weighing 3.5 g agarose. Transfer this to a 500 ml flask and add 250 ml 0.5x TBE buffer.
2. Boil for 6 min in a microwave. Allow the solution to cool to 60°C.
3. Pour agarose unto the gel mould and allow solidifying.
4. Fill the electrode tank with 0.5x TBE buffer.
5. Remove masking tape from both ends of the gel mould. Mount the gel mould on to the electrode tank making certain that bubbles do not form beneath the mould.
6. Gently remove the comb.
7. Load 10 µl of 1 Kb DNA ladder on the first well and 10 µl of each reaction mixture in the succeeding wells making certain oil is not pipetted out with the mixture.
8. Close tank and attach electrode wires to the power supply. Run for 3 h at 150 V (running time depending on the buffer used and length of the gel/ length need to run).

2.7.4. Staining and documentation

1. After electrophoresis, switch off the power supply and remove the tank cover.

Table 10. Example temperature profile for RAPD-PCR.

Temperature	Time	No. cycles
94 °C	3 min	1
94 °C	30 sec	35
35 °C	1 min	
72 °C	2 min	
72 °C	5 min	1
Hold temperature: 25 °C		

2. Remove the gel from the moulder and transfer in a tray with ethidium bromide (5 µl) staining solution in a 500 ml H_2O. Stain for 10 min. EtBr staining solution can be reused. Please take care when working with ethidium bromide, it is a mutagen, hence may need to wear double gloves when handling gels with ethidium bromide.

3. After staining, destain by rinsing with ddH_2O in a different container.

4. Photograph/score the gel under UV light. Again, some laboratories are well equipped with gel imaging system, but some other laboratories may have to wear UV protection glasses while viewing gels under UV light.

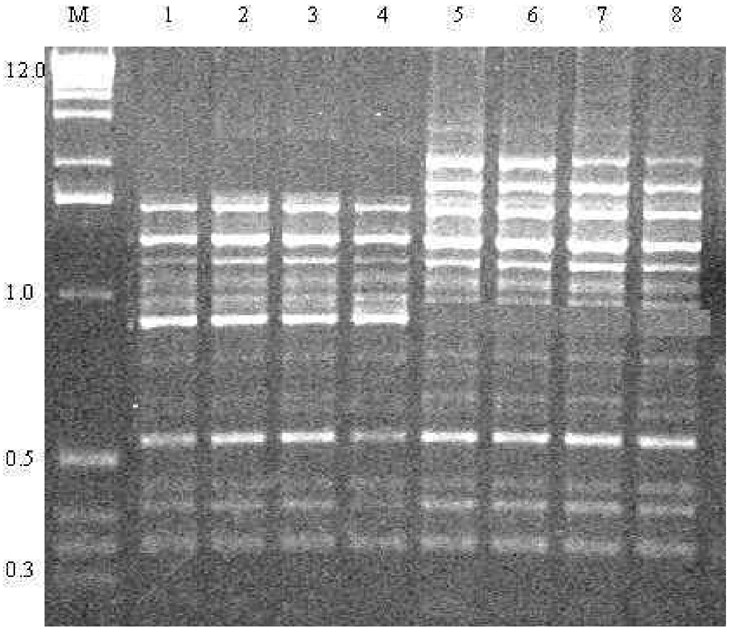
2.7.5. Scoring

- Designate a name or a number for each RAPD marker based on the molecular size and primer used.

- It is a good idea to keep the gel images for checking purposes. An example of a RAPD gel is shown in Figure 9.

Score RAPD bands using a binary system of 0 (in the absence of the band) and 1 (if the band is present). Remember that only sharp and clear bands are scored. Microsoft Excel would be useful for this purpose, because from this software we can export data to other genetic software easily.

Figure 9. An example of a RAPD gel. Bands are PCR products of two fish species (samples 1-4 are from the first species and samples 5-8 are from the second species), amplified using primer OPA20. There are many bands but only sharp bright bands should be scored (OPA20-1 to OPA 20-7).



2.7.6. Trouble shooting

Common problems with RAPD are summarised in Table 11.

2.8. Amplified fragment length polymorphism (AFLP)

AFLP was introduced by (Vos et al. 1995). The AFLP protocol involves the following steps: (1) DNA digestion with two different restriction enzymes (typically EcoR I and Mse I), (2) ligation of double-stranded adapters to the ends of the restriction fragments, (3) optional DNA pre-amplification of ligated product directed by primers complementary to adapter and restriction site sequences, (4) DNA amplification of subsets of restriction fragments using selective AFLP primers and labelling of amplified products, (5) separation of fragments via electrophoresis, and (6) scoring fragments as either presence or absence among samples.

The key feature of AFLP is the capacity for screening of many different DNA regions distributed randomly throughout the genome simultaneously. The high reliability of the technique is the result of combining the strengths of two methods, the replicability of restriction fragment analysis and the power of PCR (Vos et al. 1995). Thus, AFLP allows the detection of polymorphisms of genomic restriction fragments by PCR amplification. AFLP markers have proven useful for assessing genetic differences among individuals, populations and independently evolving lineages, such as species (Mueller & Wolfenbarger 1999).

2.8.1. Restriction digestion of genomic DNA

1. Prepare digestion mixture as per Table 12.

2. Heat one oven to 70°C and another to 37°C

Table 11. Common problems with RAPD.

Problem	Possible cause(s)
Single/monomorphic/very intense band	Product may be primer artefact - use different primer
Non-reproducible banding pattern	Too little or too much DNA. Keep the DNA concentration in the range 20-50 ng. Too little DNA may result in inefficient priming of real target sequences giving spurious bands as a result of primer artefact. Too much DNA can promote mis-match.
Inadequate separation of low or high molecular weight products	Agarose gel concentration not correct. Separate low/high molecular weight products on higher/lower concentrations of agarose gel. 2 % agarose gel (molecular biology grade) should be able to separate both low and high molecular weight products sufficient for RAPDs analysis.

Table 12. Preparation of digestion mixture.

Ingredients	Starting concentration	Final concentration	Volume / reaction (µl)
OnePhorAll (OPA, Parmacia)	10X	1X	5.00
Mse I (New England Biolabs)	4U/µl	5U	1.25
EcoR I (GibcoBRL) or Pst I (Promega)	10U/µl	5U	0.50
BSA (come with Mse I)			32.75
ddH_2O			
Genomic DNA	50-250 ng/µl		10
Total			50

3. Distribute 40 µl of cocktail in each labeled tube

4. Add 10 µl of DNA to each tube.

5. Vortex and briefly centrifuge.

6. Incubate at 37°C for 3 hours. Agitate every hour or so.

7. Inactive enzyme at 70°C for 15 min.

2.8.2. Adapter preparation

(Complete during or before digestion)

Eco RI adaptor (120 ligation recipe):

- Eco RI.1 oligo (1 µg/µl): 3.4 µl
- Eco RI.2 oligo (1 µg/µl): 3.0 µl
- OPA: 6.0 µl
- ddH_2O: 107.6 µl

Mse I adapter (120 ligation recipe):

- Mse I.1 oligo (0.5µg/µl): 64 µl
- Mse I.2 oligo (0.5 µg/ µl): 56.0 µl
- OPA: 7 µl

Mix in PCR tubes and run with the following thermal cycle:

- 65°C for 10 min
- 37°C for 10 min
- 25°C for 10 min
- Store at –20°C

2.8.3. Ligation of adapters

1. Make ligation mix as follows:
   - Eco RI adapter 1.00 µl
   - Mse I adapter 1.00 µl
   - T4 DNA ligase 10X buffer 1.00 µl
   - T4 DNA ligase (3U/ µl), Promega 0.33 µl
   - ddH_2O 6.7 µl

1. Add 10 µl of ligation mix to 50 µl of digested DNA. Vortex and briefly centrifuge.

2. Incubate at room temperature for 3 hours. Agitate every hour or so.

2.8.4. Pre-amplification reactions

1. Make reaction mixture as follows (ingredients are per reaction):
   - Eco RI + A oligo (50 ng/ µl) 0.5 µl
   - Mse I + C oligo (50 µl) 0.5 µl
   - DNTPs (5mM) 2.0 µl
   - 10X PCR buffer 2.0 µl
   - *Taq* polymerase (5U/ µl) 0.1 µl
   - $MgCl_2$ (50mM) 1.2 µl
   - ddH_2O 11.9 µl
   - Template DNA from restriction/ ligation 2.0 µl

1. Run under following thermal cycle:

1 cycle of:
- 94°C for 2 minutes

26 cycles of:
- 94°C for 1 minute
- 56°C for 1 minute
- 72°C for 1 minute

Final extension cycle:
- 72°C for 5 minutes
- Hold at 4°C

3. Transfer PCR product into new tubes with 100 µl sterile ddH_2O.

4. Blot testing can be used to test reaction success. Dot 2 µl of ethidium bromide (HARZARD!!!) and 3 µl of product on plexi-glass. Use 3 µl of cocktail as control. Visualise dots using UV box.

2.8.5. Selective amplification

1. Make reaction mixture as follows (ingredients are per reaction):
   - Eco RI + ANN oligo (50 ng/µl) 0.50 µl
   - Mse I + CNN oligo (50 ng/µl) 0.60 µl
   - DNTPs (5mM) 0.80 µl
   - 10X PCR buffer 2.00 µl
   - Taq polymerase (5U/µl) 0.08 µl
   - $MgCl_2$ (50mM) 1.20 µl
   - ddH_2O 13.82 µl
   - Template DNA from pre-selective PCR 1.00 µl

1. Run under following thermal cycle:

1 cycle of:
- 94°C for 2 minutes

12 cycles of:
- 94°C for 30 seconds
- 65°C for 30 seconds
- 72°C for 1 minute

23 cycles of:
- 94°C for 30 seconds
- 56°C for 30 seconds
- 72°C for 1 minute

Final extension of:
- 72°C for 2 minutes
- Hold at 4°C

3. Test product using dot blot if necessary

4. Combine 8 µl formamide-loading buffer and PCR product.

2.8.6. Gel electrophoresis

1. Acrylamide gel solution:
   - 42g urea
   - 10ml 10x TBE
   - 15ml 40% acrylamide

- ddH$_2$0 up to 100ml

1. Combine urea, TBE, and approximately 25ml water in a beaker. Stir with heat until urea dissolves.

2. Transfer solution to 100ml-graduated cylinder and add water up to 85ml. Transfer to vacuum flask.

3. Add acrylamide to flask and degas for approximately 10min.

4. Transfer solution to beaker and add 100µl TEMED and 500µl 10% fresh APS. Draw solution into syringe. Keep tip submerged at all times.

5. Place tube on syringe and turn it upward. Push air out of tube and pinch end of tube. Insert into Caster base.

6. Glass preparation (all glass must be scrupulously clean!)

7. Wipe integrated buffer chamber (IPC) unit with chem-wipe and ethanol.

8. Glass should be treated with Sigmacote about every 5 gels run or until top of gel sticks to long glass. Saturate a chem-wipe with Sigmacote and wipe vertically and horizontally. Wait five minutes and wipe glass three times with ethanol. Change gloves.

9. Wipe long glass with ethanol and chem-wipe.

10. In an Eppendorf tube combine 1ml of 95% EtOH 0.05% acetic acid and 2µl of bind silane.

11. Treat glass same as above description of Sigmacote. Use a great deal of pressure when wiping with EtOH. Change gloves.

12. In the event of a contamination of either Sigmacote or Bind silane on the respective glass, soak in 10%NaOH.

13. While horizontal, place spacers on IPC and long glass on top.

14. Erect vertically and clamp side braces.

15. Attach caster base insert pegs and turn. Be sure to do this while vertical and that you can see the space in between glass plates through caster base hole.

16. Check to see if comb will easily insert between glass. If not, adjust.

17. Lean clamps on top of tube racks.

18. Inject gel solution.

19. Insert combs and adjust unit to horizontal position.

20. Allow gel to polymerise for at least 1 hour.

21. Gel loading

22. Fill bottom tray with 1X TBE so about ½ inch of the bottom of gel unit is submerged. Fill IPC until ½ inch above short glass. Use needle and flush out well

23. Run gel at 75W for 1 hour.

24. Flush well again and insert comb without piercing gel.

25. Load 4.5 µl sample.

26. Run gel for 10min and then remove comb (good time to make fix/stop and developing solution).

27. Run gel for total of 2 hours and 50minutes. (Light blue dye should migrate 1 inch below bottom rib of IPC.

28. Insert tube in IPC and drain buffer.

29. Pull glass apart and wash IPC.

2.8.7. Silver staining

1. Separate plates while keeping the gel attached to short glass.

2. Fix the gel: Place gel in tray, cover with cold fix/stop solution and agitate well for 20 minutes. Gel may be stored in fix/stop solution overnight. Save fix/stop solution and place back in freezer.

3. Wash the gel: Rinse the gel 3 times for 2-3 min. each in ddH$_2$O using agitation. Lift gel from solution and allow to drain 10-20 seconds.

4. Stain the gel: Transfer the gel to staining solution and agitate well for 30 minutes.

5. Pour 1L of the developing solution into a tray. Transfer staining solution to beaker. Rinse tray and fill with ddH$_2$O.

6. Rinse gel for 5-10 seconds ONLY. Transfer to developing solution.

7. Agitate in developing solution until bands begin to appear. Transfer gel to remaining chilled developing solution for 2-3 minutes.

8. Fix the gel: add 1L of Fix/stop solution directly to developing solution and agitate for 2-3 minutes

9. Rinse gel twice for two minutes each in ddH$_2$O.

10. Dry gel on glass

Fix/stop solution

- 200 ml glacial acetic acid
- 1800 ml ultrapure water

Silver staining

- 2 g of silver nitrate ($AgNO_3$)
- 2 L ultrapure water

Immediately before use add:

- 3 ml (1 vial) of 37% formaldehyde

** *Developing solution*

- 60 g Sodium carbonate (Na_2CO_3)
- 2 L ultrapure water

**chill to 10°C. Place in freezer for approx 4 hours and stir to break up ice prior to use.

Immediately before use add:

- 3 ml of 37% formaldehyde
- 400 µl aliquot sodium thiosulfate (discard remaining)

2.8.8. Scoring

Score AFLP bands using a binary system of 0 (in the absence of the band) and 1 (if the band is present). Remember that only sharp and clear bands are scored. If scanning equipment is available, scan the gel and keep record together with manual scoring.

2.8.9. Gel preservation

1. Soak gel in 3% NaOH with gentle agitation for 30 to 60min, or until edge of corner of the gel starts coming loose. If gel does not come loose, tease a corner and pull gently. If it peels easily, gel is ready for transfer. Loosen edges with razor blade to facilitate transfer.
2. Carefully transfer gel to 3.5% acetic acid and soak for 3 min without agitation. Rinse in ddH_2O for 2 minutes without agitation.
3. Drain excess water from gel and smooth a sheet of chromatography paper over gel.
4. Very slowly pull edge or corner up while gel adheres to paper. Use a razor blade to persuade any lagging parts of the gel.
5. Cover gel with plastic wrap and dry on gel dryer at 70°C for 2 hours.

2.9. Microsatellites

Microsatellite loci can be identified by screening genomic libraries with probes made up of tandemly repeated oligonucleotides (Tautz, 1989; Hughes & Queller, 1993; Queller *et al*., 1993; Schlötterer & Pemberton, 1994) and then sequenced to identify conserved flanking regions for primer design. Loci identified in this way are analysed by amplifying the target region using PCR, followed by electrophoresis through a

Table 13. Oligo fragments.

EcoRI Linker 1	CTC GTA GAC TGC GTA CC
EcoRI Linker 2	AAT TGG TAC GCA GTC TAC
EcoRI +A	GAC TGC GTA CCA ATT CA
Pst I Linker 1	CTC GTA GAC TGC GTA CAT GCA
Pst I Linker 2	TGT ACG CAG TCT AC
Pst I +A	GAC TGC GTA CAT GCA GAC A
Mse I Linker 1	GAC GAT GAG TCC TGA G
Mse I Linker 1	TAC TCA GGA CTC AT
Mse I +C	GAT GAG TCC TGA GTA AC

polyacrylamide gel to allow resolution of alleles that may differ in size by as few as two base pairs.

A major disadvantage of microsatellites is that identifying appropriate regions from a genomic library for a new species can be time-consuming and expensive. Known primers are not usually useful for amplifying the same locus across related taxa unless the microsatellite region is flanked by highly conserved sequences where priming site are located (FitzSimmons *et al.* 1995). In addition, the presence of null alleles (alleles that do not amplify due to mutational changes in the priming site) can complicate the analysis as well. However, microsatellites have been extremely useful in fish and crustacean population studies and are quickly becoming the marker of choice for a variety of applications (Wright & Bentzen 1994; Xu *et al.* 1999).

In the last few years microsatellites have become one of the most popular molecular markers with applications in many different fields. High polymorphism and the relative ease of scoring represent the two major features that make microsatellites of large interest for many genetic studies. The major drawback of microsatellites is that they need to be isolated *de novo* from species that are being examined for the first time. High output microsatellite library screening requires an automatic sequencer which is only available in a few labs in Asian countries. However, in the case that primers used for microsatellites have been developed for different species, care should be exercised. Ideally target products should be sequenced to verify that they are 'real' microsatellites.

It is also common in the region that the size of gels used to screen variation are often too small and much of the 'real' variation is not detected, leading to potential problems with H/W conformation. Labs should therefore consider applying a more rigorous approach to silver staining microsatellite analysis, or use less technically demanding nuclear markers such as allozymes.

The protocol presented here can be applied in a small laboratory with minimum equipment.

2.9.1. Preparation of sequencing Dye

Mix 10 mg of bromophenol blue, 10 mg xylene, 200 µl 0.5 M EDTA and add formamide until a final concentration of 10 ml is reached. Store at 4°C.

2.9.2. Preparation of PCR cocktail

Each PCR reaction should contains the following ingredients which can be added directly to each tube, or mixed all ingredients except DNA template as a PCR cocktail. The ingredients for each reaction of 10 µl containing:

- 1 x PCR buffer (available with Taq polymerase)
- 1.0-2.0 mM $MgCl_2$
- 100 µM of mixed dNTPs
- 0.5 µM of reverse and forward primer
- 0.2 unit of *Taq* polymerase

- Add autoclaved ultra pure water up to the desired volume
- 5-10 ng of DNA template

In each PCR tube containing 1 µl DNA template, 9µl cocktail is added.

Note:

- Add 1 µl DNA template (diluted DNA) in all PCR microtubes with labels (use different pipette tips to prevent contamination). While preparing cocktail, the microtube should be kept on ice to prevent the reaction to start.

- *Taq* DNA polymerase should be added last to prevent reaction. It should be taken out from the deep freeze at the time of addition to the cocktail and returned to the freezer immediately after use.

Centrifuge tubes contained template and cocktail for 30 seconds at any rpm, do not centrifuge longer than 30 seconds because reaction may take place if centrifuge for longer period.

2.9.3. The PCR cycles

The PCR cycles are summarised in Table 14 below.

2.9.4. Check quality of PCR product using agarose gel

Before the PCR products are separated on acrylamide gel it is recommended to perform a preliminary check on agarose gel as follows.

1. Load 5 µl of each PCR product mixed with the stop dye in a well of the agarose gel. Same pipette tip can be use by washing it in 0.5 x or 1 x TBE buffer.

2. Perform electrophoresis using a voltage of 100 volt for around 30 minutes in small electrophoresis and 40 minutes in big electrophoresis.

3. Stain the gel in ethidium bromide for 10-15 minutes.

4. Observe presence of bands under a UV transluminator. If a single band is observed for each PCR product polyacrylamide gel electrophoresis can be performed.

Table 14: The PCR cycles.

Cycle	Temperature	Duration	No. of cycles
Predenature	94°C	3 min	1
Denature	94°C	30 sec	35
Annealing	appropriate temp	30 sec	
Extension	72°C	1 min	
Post extension	72°C	5 min	1
After the reactions are completed immediately add 5 µl of the stop dye to each tube and keep in 4°C until electrophoresis is started.			

2.9.5. Protocol for polyacrylamide gel electrophoresis

Glass plate preparation

1. Clean glass plates properly and wipe with absolute ethanol.
2. Wipe the inner side of the rear glass plate (a longer glass plate) with clear view (same solution for cleaning glass, window, etc.) to prevent gel from sticking on the glass plate.
3. Wipe the front (short) glass plate with glass bond chemical to enhance adherence of gel on the plate.
4. Leave the glass plates to air dry.
5. Place the long plate on the lab bench with the cleaner side upward and then carefully place one spacer on each long side.
6. Place the short glass plate on top of the long plate with the clean side facing the spacers. Make sure that the spacers reach the bottom part of the plates and align well with the long side of the plates.
7. Clip the long sides of the sandwiched plates with binder clips.
8. Carefully seal the bottom edge of the sandwiched plates using a piece of tape. Then remove the binding clip from each long side and seal with a piece of tape.

2.9.6. Polyacrylamide gel preparation

1. Take 60 ml of 4.5 % acrylamide (containing 6.75 ml of 40 % acrylamide gel, 25.2 g of urea, 6 ml of 10 x TBE buffer and add distilled water until to 60 ml) and pour into a beaker.
2. Preparation of 0.25 % ammonium persulphate solution: Weigh 0.025 mg of ammonium per sulphate in a microtube and add 100 µl of distilled water, wrap tube with aluminum foil to protect from light.
3. Add 60 µl of TEMED and 60 µl of 0.25 % ammonium persulphate solution to the mixture containing acrylamide and stir well.
4. Place the sandwiched plates at a 45° angle. Fill a syringe with the gel solution. Then slowly but continuously pour the gel into the space of the sandwiched glass plates until it is 3/4 full. Try to avoid air bubble.
5. Then slowly lower the sandwiched glass plate down and finally place it parallel to the bench. Allow the gel solution to spread to fill the upper edge of the sandwich. After pouring gel, insert the flat edge of a shark tooth comb 3-5 mm into the gel.

6. Then clip the open side with clips to prevent the gel from leaking. Clamp the long sides with the binding clips, 2 clips/side. Make sure that the clips clamp on middle of the spacer width.

7. Leave it to polymerize for 2 – 3 hours.

8. After polymerization, polyacrylamide gel can be used immediately.

2.9.7. Assembly of gel plate into the gel rig

1. Check whether the drain bottle at the back of the gel rig is empty and properly attached to the drain tube. Prepare running equipment set according to the instruction for each model.

2. Remove the clips and pieces of tape, gently rinse the sandwiched gel plate under running tapped water. Remove the excess polyacrylamide gel sticking outside the glass plates. Gently remove the comb, and rinse until all debris is washed out from the gel space. Then dry all the surfaces with paper towel.

3. Place the gel plates on the supporter in the lower buffer chamber with the short gel plate facing the aluminum plate of the sequencer. Close the sequencer properly according to the manual for each model.

4. Fill the upper chamber with 1 x TBE buffer up to the margin.

5. Pour 1 x TBE buffer in the lower buffer chamber until the gel bottom is properly submerge.

6. Insert the comb until the tip of the teeth just contact the gel surface. Check for bubbles, if present remove it by blowing buffer with a pipette into the groove.

7. Close the upper and lower chamber lids. Plug in the power lead assembly and start prerun by setting the power supply to 100 W and operate for 30 min to warm the gel.

2.9.8. Electrophoresis

1. About ten minutes prior the completion of the prerun, prepare samples for loading.

2. Flash centrifuge the samples and then denature at 94°C for 3 min and then immediately place the denatured samples on ice.

3. Load 2-3.5 µl of the sample into each sample well. Load size markers to a well ahead of the first sample and a well after last sample.

4. Plug in the HV lead assembly, adjust the power supply to 50 W and run for 1.30 hours.

5. When the electrophoresis is completed, turn off the power supply and disconnect the HV assembly.

6. Open the upper and lower buffer chambers and drain. Then uninstall the gel according to the manual for each sequencer model.

2.9.9. Chemical preparations

Preparation of M13 ladder

M13 ladder solution (10 µl) is prepared according to a protocol modified from Promega Corporation (undated) as follows:

- All preparation tubes must be placed on ice.

- Put 2 µl of each d/ddNTP (A, G, C, T) into a 0.2 ml PCR tube.

- Prepare a cocktail for PCR reaction as follows:
  - pGEM®-3Zf(+) control DNA (4µg): 1.2µl
  - DNA sequencing 5X buffer: 8 µl
  - pUC/M3 forward primer (4.5pMol): 1.8 µl
  - Nuclease-free water: 20 µl
  - Add *Taq* polymerase (5U): 1 µl

- Mix well with a vortex and add 8 µl of this solution to each of the tubes containing d/ddNTP. Then overlay the solution with a drop of mineral oil and flash spin.

Place the tubes in a PCR machine programmed as follows:

- One cycle of denature at 95°C for 2 min

- 60 cycles of:
  - Denature at 95°C for 30 sec
  - Annealing and extension at 70°C for 2 min

- Hold at 4°C.

- After the reaction is completed add to each tube 5 µl of a stop dye follwed by a flash spin. Prior to loading to electrophoresis gel the solutions must be denatured.

Developing solution

- A developing solution should be prepared in advance because it should be kept in a deep freezer until nearly freeze. To prepare the solution:
  - Weigh 30 gm sodium carbonate (Na_2CO_3) and put in a plastic jug.
  - Add 1 liter autoclaved distilled water and stir until dissolved.
  - Cover the jug with plastic wrap.
  - Store in a freezer.
  - Five minutes before use, add aliquote (200 to 250 µl) of sodium thiosulphate (10 mg of sodium thiosulphate in 1 ml of autoclaved distilled water, protect from light before use) and 1.5 ml of 37 % formaldehyde.

Fix/stop solution

- Add 100 ml of glacial acetic acid to 900 ml of autoclaved distilled water and mix.

- Pour it in a plastic bottle and close properly.

Staining solution

- Weigh 1 gram of silver nitrate and put in a plastic jug.
- Add 1 liter of autoclaved distilled water.
- Add 1.5 ml of 37 % formaldehyde to the solution.
- The preparation should be done quickly to protect from light and keep in the dark place before use.

2.9.10. Gel fixation and staining

1. Prepare two plastic trays and two plastic buckets.
2. Place the sandwiched plates on a laboratory bench. Then carefully separate the plates by inserting a small metal rod and lift. The gel would tightly adhere to the short plate.
3. Put the short plate with the gel on top in the tray. Then pour fix/stop solution to the tray, and shake for 20 minutes.
4. After 20 minutes remove the fix/stop solution and keep in a bucket for further use.
5. Pour 1 liter autoclaved distilled water into the tray (with the short plate) and shake for 2 minutes then remove.
6. Repeat it twice (all together 3 times).
7. Then stain the gel by pouring silver nitrate solution into the tray and shake for 30 minutes.
8. Remove silver nitrate by pouring into a designated bottle. Do not throw it elsewhere. Before discarding, the silver nitrate should be kept under sunlight.
9. Wash the stained gel with autoclaved distilled water 1 time (no longer than 5-10 seconds).
10. Pour developing solution over it, and shake until the bands appear and the color of the gel turns to light brown.
11. Add fix/stop solution and shake for 3 minutes.
12. Wash the plate twice with distillated water for 2 minutes each.
13. Keep the plate to dry over night at room temperature.

2.9.11. Scoring gel

Scoring of microsatellite genotypes is straightforward. The homozygotes produce a single band whereas the heterzygotes produce two bands. However, some problems may emerge for example dinucleotide microsatellite almost always produces stutter bands which may lead to miscoring of homozygotes. Some loci comprise null alleles which refer to alleles that do not give bands. Therefore heterozygotes are mis-scored as homozygotes.

2.9.12. Trouble shooting

Common problems with microsatellites are summarised in Table 15.

2.10. Temperature gradient gel electrophoresis (TGGE)

As the name suggests, this method relies on a temperature gradient to denature double-stranded DNA fragments that are differentially separated based on their respective melting profiles. Another method, Denaturing Gradient Gel Electrophoresis (DGGE) is similar to TGGE but uses a chemical gradient of increasing urea and formamide concentrations to denature the DNA instead of a temperature gradient. The procedure involves the electrophoresing of DNA through a polyacrylamide gel that is running parallel to a temperature gradient. The double-stranded duplexes of DNA migrate along the gel until they reach their respective melting points where the progress is greatly reduced when the dsDNA begins to unwind. The melting point of a specific fragment of DNA is a function of both the effect of base sequence on the helix structure and the electrophoretic mobility of the strand as it starts to unwind. Therefore DNA fragments with different base pair sequences tend to display different melting points and hence stop at differing points on the gel.

Figure 10. TGGE apparatus. Both heated (eg. 60°C) and cooled (eg. 20°C) water is pumped through opposite ends of the heating block creating a linear temperature gradient. Samples are loaded at the cool end of the gel and electrophoresed to a point in the gel where the temperature denatures the double-stranded fragments. The gel is then silver stained to visualise the bands.

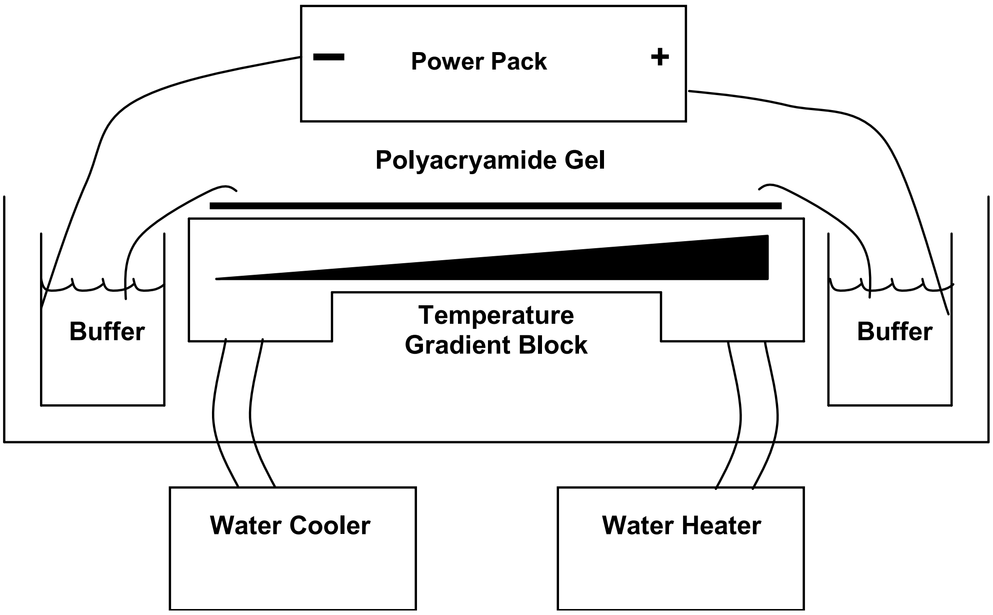


Table 15. Common problems with microsatellite.

Problem	Possible causes
For sequencing	
Faint or no bands	• Dirty template DNA • Insufficient template • Insufficient enzyme activity • Poor annealing of primer to template • Contamination of sequencing reaction with salt • Electrophoresis temperature too high • Samples not denatured before loading onto gel
Low band intensity at bottom of gel	• DNA concentration too low
Bands are fuzzy throughout the lanes	• Poor quality polyacrylamide gel • DNA sample contains two different templates, generating overlapping sequences
For staining	
Faint or no bands	• Improper rinsing following the staining • Poor quality water • Incorrect amount of sodium carbonate added to the developing solution • Too much sodium thiosulfate added to the developing solution
Low band intensity at bottom of gel	• Poor quality water
High background staining	• Developing solution too warm • Development performed too long • Insufficient fixation • Detergent residues present on glass plates may result in a brown background • Poor quality polyacrylamide gel • Poor quality sodium carbonate was used
Dark, swirling patterns on the gel surface	• Inadequate agitation during the staining steps • Inadequate rinsing before the development step
Yellow gel	• Improper fixing of the gel • Poor quality sodium carbonate was used
Gray gel	• The sodium thiosulfate was not added to the developing solution
Band stain yellowish-brown with poor contrast, as opposed to dark gray	• Dirty template DNA
Gel adheres to both plates	• Long glass plate contaminated with binding solution, or inadequate treatment of the long plate with clear view
Gel peels off the plate when dried	• Build-up of binding solution after multiple treatments • Acrylamide percentage in the gel was too high

Modified from Promega Corporation, undated. Instructions for use of products Silver SequenceTM DNA Sequencing System. Technical Manual No. 023.

To improve the resolution of the technique, TGGE is usually conducted in conjunction with Heteroduplex Analysis (TGGE/HA). Nuclear (diploid) DNA fragments can be heteroduplexed to themselves but because mtDNA is haploid, an extra reference DNA fragment (ideally from a moderately divergent conspecific individual) needs to be added. The heteroduplexing process involves heating both the reference and sample DNA together in order to reduce them to single strands. Upon cooling the strands recombine. Apart from the original double strands from the reference and sample fragments recombining to themselves (homoduplexes), mismatch pairings occurs with one strand from the reference and one strand from the sample also recombining (heteroduplexes). Where heteroduplex fragments have nonperfect complementary matches, they tend to have lower and more variable melting points than homoduplexes resulting in additional bands on the gel. TGGE is a reliable method for DNA fragments up to ~700 base pairs in length.

SECTION 3

Data analysis

3.1. Analysis of molecular data

One of the main goals in biodiversity conservation is the preservation of genetic diversity. Traditionally, the study of genetic diversity has fallen within the realm of population genetics, particularly in regard to comparing levels of genetic diversity within and among populations and in making inferences on the nature and intensity of evolutionary processes from the observed patterns of genetic diversity. Hence, there is a long tradition as well as a wealth of conceptual tools in population genetics for analysing, measuring and partitioning genetic diversity.

This summary on methods for analysing variation using molecular markers will start with a brief outline of the main population genetic concepts involved. These ideas were developed for simple situations, such as the one-locus two-alleles case, and were refined and generalised later. However, the main features are best understood by taking the simplest case, which in terms of a molecular marker, can be understood as an allozyme locus with only two alleles. In this situation, we are dealing with **co-dominant markers**, such as those generated by allozyme electrophoresis and microsatellite techniques, for which all possible genotypes (both homozygotes and the heterozygote) can be easily ascertained.

The manual will then move to the case of the richest possible markers in terms of the amount and quality of the information provided - **DNA sequences**. Once the direct analysis of nucleotide sequences has been developed, we will consider other markers which provide indirect estimates of nucleotide divergence between alternative alleles - **haplotypic markers**, such as RFLPs and SSCPs. Additionally, we consider **dominant markers**, such as RAPDs and AFLPs, for which the “presence” allele is dominant over the “absence” allele. Furthermore, wherever appropriate, some computer programs available for use in population genetics and analysis of molecular variation are introduced and discussed.

3.1.1. Co-dominant Markers

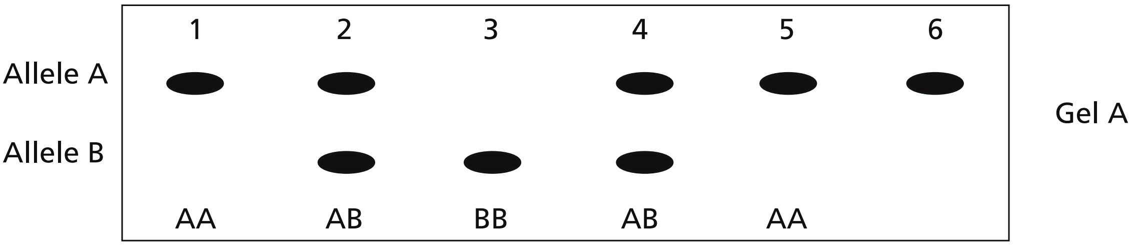
Allele / genotype nomenclature

The first assignment before analysing data is that a genotype should be interpreted from observed phenotype.

It may be appropriate here to provide some variation of allele nomenclature. **Allozyme alleles** are often named using alphabetical characters following alphabetical order with A being the slowest moving allele as shown in Figure 11. This is the simplest case, i.e. a monomeric enzyme with two alleles, which give phenotypes.

However, sometimes alleles can be named based on the distance, measured in mm, the protein produced from them migrates in the gel relative to the distance the protein produced from *the most common allele* migrates in the gel.

Figure 11. Hypothetical electrophopherograms of a monomeric enzyme.



For example:

- The most common allele migrates 10 mm
- Variant 1 migrates 8 mm
- Variant 2 migrates 12.5 mm
- Most common allele = (10*100)/10 = 100
- Variant 1 = (8x100)/10 = 80
- Variant 2 = (12.5*100)/10 = 125

Microsatellite alleles are, however, often named based on the name of primers used and the size of the allele. For example, assuming that the gel below is silver stained polyacrylamide gel with microsatellite alleles, which are amplified from 10 individuals of a fish species, using microsatellite primers developed for *Cyprinus carpio*, and the locus name is MFW-1. There are three alleles in the gel, with sizes of 156bp, 160bp, and 168bp, respectively. The genotypes assigned for individual 1 is MFW-1: 168/168, and individual 4 is a heterozygote at this locus with genotype 160/156.

How to calculate gene frequency

The genetic interpretation for phenotypes in Gel A is:

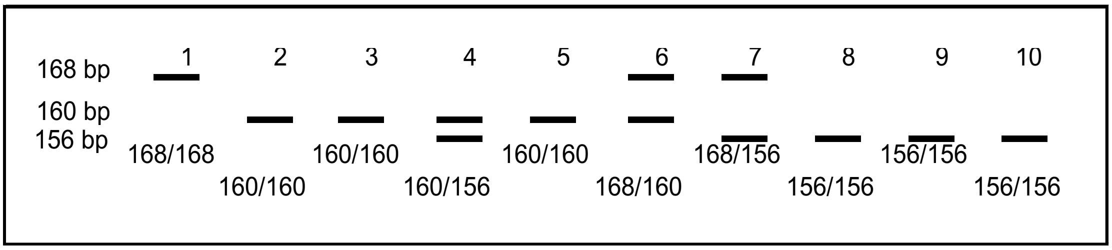
- 1, 5, 6 = AA (homozygote)
- 2, 4 = AB (heterozygote)
- 3 = BB (homozygote)

From Figure 11, frequency of allele A is:

$$f_A = \frac{2*3+2}{2*6} = 0.666$$

and frequency of allele B is:

$$f_B = 1 - f_A = 1 - 0.666 = 0.334$$



Some enzymes can be dimeric or tetrameric where product of different alleles interact with each other as shown in Figure 12.

Proportion of polymorphic loci (P)

Proportion of polymorphic loci (P) is estimated by using number of loci polymorphic divided by the total number loci examined.=

Example: Assuming the individuals depicted in Figure 11 and Figure 12 were also analysed for seven other loci all of which were monomorphic (not variable).

$$P = 3/(3+7) = 0.30$$

The frequency of an allele is given by:

$$\frac{2H_0 + H_e}{2N} \qquad (1)$$

where:

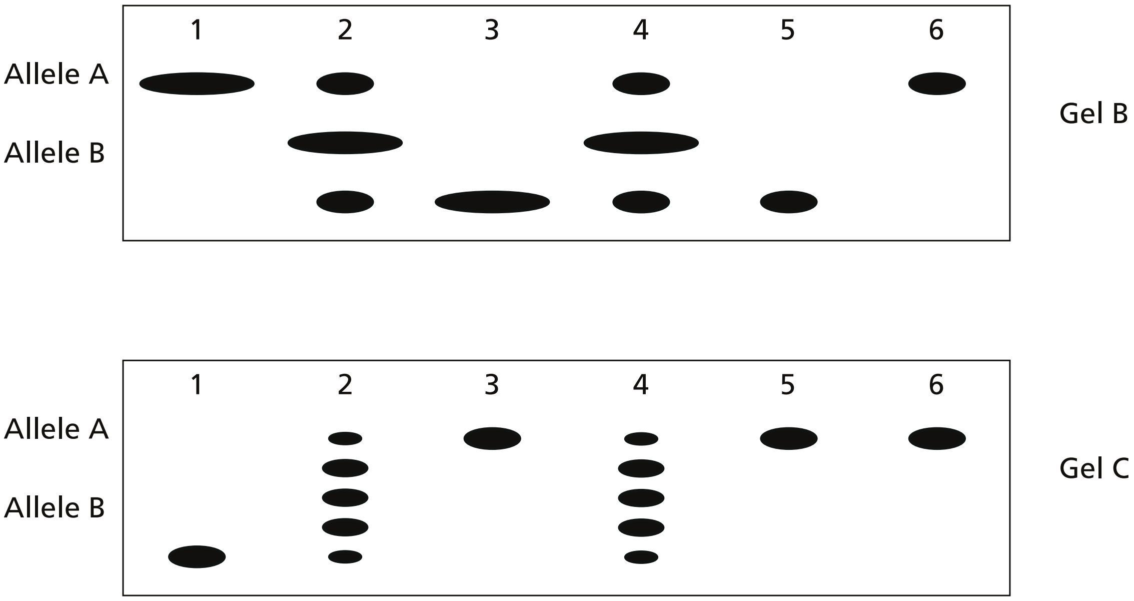
H_0 = number of homozygotes for that allele.

H_e = number of heterozygotes for that allele.

N = number of individuals examined.

The drawbacks of this parameter are that it does not consider how variable polymorphic loci are and is sensitive to the number of individuals examined and the number of loci screened. Probability of detecting low frequency variants increases with sample size. Therefore it is suggested that a criteria is set, in case of sample size less than 100, a locus is considered polymorphic

Figure 12. Hypothetical electrophopherograms of dimeric (Gel B) and tetrameric (Gel C) enzymes.



if frequency of the common allele does not exceed 0.95 (P_{95}). P_{99} is used when the sample size is larger than 100.

Average Expected Heterozygosity (H_e)

Average expected heterozygosity (H_e) is average proportion of loci at which an individual is expected to be heterozygous based on HWE.

Expected heterozygosity at a locus (h) is one minus the sum of the squared allele frequencies:

$$h = 1 - (p^2 + q^2) \text{ or } h = 1 - \sum p_i^2$$

If only two alleles exist at a locus then h = 2pq. These quantities are derived from expected Hardy-Weinberg genotypic proportions.

$$p^2 + 2pq + q^2 = 1 \text{ or } 2pq = p^2 + q^2$$

Example, Figure 12:

(Gel A): $h = 1 - (0.6662 + 0.3342)$
$= 0.445$

(Gel B): $h = 0.50$

(Gel C): $h = 0.445$

Average these values over all loci yields:

$H_e = (0.445 + 0.500 + 0.445 + 7[0])/10$
$= 0.139$

Each individual in the population is expected to be heterozygous at 13.9% of its loci.

H_e is mainly determined by loci with two or more alleles at an appreciable frequency (0.1 to 0.9). It is not sensitive to the loss of low frequency alleles.

Average number of alleles per locus (A_n)

Examples from Gel A, B and C plus seven monomorphic loci:

$$A_n = (7+6)/10 = 1.30$$

Testing for Hardy-Weinberg equilibrium

The Hardy-Weinberg equation is a key concept in population genetics that describes the relationship between gene and genotype frequencies. It states that "in the absence of migration, mutation and natural selection, gene frequencies and genotypic frequencies remain constant in a large, randomly mating population". Such a population may called "in Hardy-Weinberg equilibrium", i.e. the frequency of genotypes is dependent on the frequency of genes (also called "Hardy-Weinberg proportions"), and these are both constant over time. Thus, the Hardy-Weinberg equation is a statement of the null hypothesis that no evolutionary forces are acting on a large, randomly breeding population (the seven criteria listed on page 4). If we sample a population and discover that genotype frequencies are not in Hardy-Weinberg equilibrium (not in Hardy-Weinberg proportions) then we can conclude that one or more external forces are at work. The fishery manager

Average number of alleles per locus (A_n) is only calculated for co-dominant markers, and is estimated as the total number of alleles detected summed over all loci divided by the total number of loci examined.

$$A_n = \frac{\textit{Number of monomorphic loci + number of alleles at polymorphic loci}}{\textit{Number of loci analysed}} \qquad \textbf{(2)}$$

A_n is sensitive to the number of individuals analysed.

hopefully can determine what those forces are in order to manage the population.

The Hardy-Weinberg principle serves as a kind of a null hypothesis, i.e. there is random mating, no selection, no mutation and no migration occuring in the population studied. It tells us what to expect if all of the seven stated conditions are true in the population. If we sample a population and find that the genotype frequencies are different from Hardy-Weinberg expectations, then we can conclude that one or more of these assumptions is violated, or at least one evolutionary process is operating, or sometimes variations are not properly scored. This motivates us to study the population in more detail.

The usual way to compare a set of observed values to a set of expected values (based on some null hypothesis) is to use a **goodness of fit test**. The most commonly used goodness of fit test for Hardy-Weinberg conditions is the χ^2 **test**.

If a population conforms to Hardy-Weinberg equilibrium, then the frequencies of genotypes will be in ratio of p^2, $2pq$ and q^2, for a two allele polymorphism, where p is the frequency of allele A and q if the frequency of allele B.

We use Gel A as an example (please note that this is only an example with a small sample size, in practice we will have to generally deal with much larger number of samples).

Expected proportion of AA:
$p^2 = 0.6662 = 0.44$

Expected proportion of BB:
$q^2 = 0.3342 = 0.11$

Expected proportion of AB:
$2pq = 2 \times 0.666 \times 0.334 = 0.45$

Expected number of AA:
$0.44 \times 6 = 2.66$

Expected proportion of BB:
$0.11 \times 6 = 0.67$

Expected proportion of AB:
$0.45 \times 6 = 2.67$

Observed distribution and expected Hardy-Weinberg equilibrium distribution of genotypes can be summarised in the table below:

Table 16. Observed and expected HWE.

	Genotypes		
	AA	AB	BB
Observed	3	2	1
Expected	2.66	2.67	0.67
$(O - E)^2$	0.12	0.45	0.11
$(O - E)^2/E$	0.04	0.17	0.16

$\chi^2 = 0.04 + 0.17 + 0.16 = 0.37$

The degrees of freedom (df) in a test involving n classes are usually equal to *n*-1. That is, if the total number of individual (6 in this example) is divided among n classes (3 genotypic classes in the example), then once the expected numbers have been computed for *n*-1 classes (2 in the example), the expected number of the last class is set. Thus in the above example there are only two degrees of freedom in the analysis.

Check the χ^2 value of 0.37 at df = 2 in Table 17 we will have P value > 0.05 and therefore we accept the null hypothesis of Hardy-Weinberg equilibrium in the population in our example.

It is important to note that interpreting goodness of fit tests for Hardy-Weinberg equilibrium is not always straightforward. Detecting significant deviations requires large sample sizes and strong disturbing forces. Lack of significance cannot be interpreted to mean that evolutionary processes are not operating. Different processes may be acting in ways that are often not detectable with a goodness of fit test, or they may be too weak to be detectable with a given sample size.

For formal testing try one of the online tools (Online HWE and Association Testing; Online; HWE Test (Multi-allelic Markers), Genetic Calculation Applets (up to four allele), or freeware (**Arlequin v3.01**; **PopGene**; **GDA**; **TFPGA**). One important point is to choose an exact test for multiallelic markers because χ^2 tests are inappropriate when there are multiple alleles (Guo & Thompson 1992). It has been argued that even for large samples, a χ^2 test is inappropriate and exact tests should be used in assessment of HWE (Wigginton *et al.* 2005).

The Hardy-Weinberg Principle suggests that as long as the assumptions are valid, allele and genotype frequencies

The difference between the observed and expected values can be tested for statistical significance using a χ^2 **test for goodness of fit.**

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \qquad \textbf{(3)}$$

will not change in a population in successive generations. Thus, any deviation from HWE may indicate:

1. Small population size results in random sampling errors and unpredictable genotype frequencies (a real population's size is always finite and the frequency of an allele may fluctuate from generation to generation due to chance events).

2. Assortative mating which may be positive (increases homozygosity; self-fertilization is an extreme example) or negative (increases heterozygosity), or inbreeding which increases homozygosity in the whole genome without changing the allele frequencies. Rare-male mating advantage also tends to increase the frequency of the rare allele and heterozygosity (in reality, random mating does not occur all the time). Cryptic population stratification is another reason for departure from HWE.

3. A very high mutation rate in the population (typical mutation rates are < 10^{-5} per generation) or massive gene flow from a genotypically different population interfering with the allele frequencies.

Table 17. χ^2 **probabilities.**

df	Probabilities									
	0.95	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.016	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.42	2.37	3.67	4.64	6.25	7.82	11.35	16.27
4	0.71	1.06	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	0.15	1.61	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.64	2.20	3.83	5.35	7.23	8.56	10.65	12.59	16.81	22.46
7	2.17	2.83	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.13
9	3.33	4.17	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
11	4.58	5.58	8.15	10.34	12.90	14.63	17.28	19.68	24.73	31.26
12	5.23	6.30	9.03	11.34	14.01	15.81	18.55	21.03	26.22	32.91
13	5.89	7.04	9.93	12.34	15.12	16.99	19.81	22.36	27.69	34.53
14	6.57	7.79	10.82	13.34	16.22	18.15	21.06	23.69	29.14	36.12
15	7.26	8.55	11.72	14.34	17.32	19.31	22.31	25.00	30.58	37.70
20	10.85	12.44	16.27	19.34	22.78	25.04	28.41	31.41	37.57	45.32
25	14.61	16.47	20.87	24.34	28.17	30.68	34.38	37.65	44.31	52.62
30	18.49	20.60	25.51	29.34	33.53	36.25	40.26	43.77	50.89	59.70
50	34.76	37.69	44.31	49.34	54.72	58.16	63.17	67.51	76.15	86.66
	← Accept →								← Reject →	

4. Selection of one or a combination of genotypes (selection may be negative or positive). Selective elimination of homozygotes as in some autosomal dominant diseases, where homozygotes for the mutation may die in utero, is an example (in a very large sample, this could violate HWE). Similar to this selection, sampling error (selection bias) may also affect HWE if bias concerned ethnicity.

5. Unequal transmission ratio (transmission ratio distortion or segregation distortion) of alternative alleles from parents to offspring.

6. Different gene frequencies in males and females.

7. Gels have not been read correctly.

In most population genetic estimations (like linkage disequilibrium calculations), HWE is assumed. This means that genotype probabilities are determined by allele frequencies and nothing interferes with this. If this assumption is not met, the estimations will not be accurate. When HWE is assumed, this means that genotype probabilities are determined by allele frequencies, i.e., there is no transmission ratio distortion, selection against a genotype (lethality) etc. If HWE is violated, statistical methods using allele frequencies may not be valid and methods that use genotype frequencies should be preferred (Xu *et al*. 2002)

It has to be remembered that when HWE is tested, mathematical thinking is necessary. When the population is found in equilibrium, it does not necessarily mean that all assumptions are valid since there may be counterbalancing forces. Similarly, a significant deviation may be due to sampling errors (including **Wahlund effect**, see below and Glossary), misclassification of genotypes, measuring two or more systems as a single system, failure to detect rare alleles and the inclusion of non-existent alleles. The Hardy-Weinberg laws rarely holds true in nature (otherwise evolution would not occur). Organisms are subject to mutations, selective forces and they move about, or the allele frequencies may be different in males and females. The gene frequencies are constantly changing in a population, but the effects of these processes can be assessed by using the Hardy-Weinberg law as the starting point.

Wahlund effect: Reduction in observed heterozygosity (increased homozygosity) because of pooling discrete subpopulations with different allele frequencies that do not interbreed as a single randomly mating unit. When all subpopulations have the same gene frequencies, no variance among subpopulations exists, and no Wahlund effect occurs ($F_{ST} = 0$). **Isolate breaking** is the phenomenon that average heterozygosity temporarily increases when discrete subpopulations make contact and interbreed (this is due to a decrease in homozygotes). It is the opposite of **Wahlund effect**.

To calculate *D*, first one needs to calculate genetic identity (*I*). Nei's coefficient of genetic identity (I) between two taxa is given by:

$$I = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2 y_i^2}} \qquad \textbf{(4)}$$

Where x_i and y_i are the frequencies of the *ith* allele in population X and Y respectively.

Genetic distance

Many genetic distance measures have been proposed (Nei 1987; Reynolds 1981). One useful measure when dealing with genetic data was developed by Nei (1972; 1973), called standard genetic distance (*D*), which is illustrated opposite.

If in populations X and Y, the frequencies of alleles at a locus are as follows:

	A	B
Population X	0.46	0.54
Population Y	0.88	0.12

Then:

$$I = \frac{(0.46 * 0.88) + (0.54 * 0.12)}{\sqrt{[(0.46^2 + 0.54^2) * (0.88^2 + 0.12^2)]}} = 0.745$$

I=1 when X and Y are monomorphic for the same allele and *I*=0 when X and Y are monomorphic for different alleles.

How genetic variation is distributed

Determining population structure is an important aim of many population genetic studies. Many organisms naturally form subpopulations such as herds, flocks, schools etc. In addition, natural habitats are typically patchy especially inland waterbodies. When

The mean genetic identity (*I*) is the mean over all loci studied (including monomorphic ones) and is most conveniently calculated as:

$$\bar{I} = \frac{I_{xy}}{\sqrt{I_x I_y}} \qquad \textbf{(5)}$$

Where I_{xy}, $I_y$and I_y are the means, over all loci of $\sum x_i y_i$, $\sum x_i^2$ and $\sum y_i^2$ respectively.

Genetic distance is estimated by: $D = -\ln I$

there is population subdivision, there is almost inevitably some genetic differentiation among the subpopulations, e.g. difference in allele frequencies among the subpopulations. Genetic differentiation may result from natural selection favouring different genotypes in different subpopulations, but may also result from random processes in the transmissions of alleles from one generation to the next or from chance differences in allele frequencies among the initial founders of the subpopulations.

One of the main effects that population subdivision has on genetic diversity, is the reduction in observed heterozygosity compared with expected heterozygosity. The extent of reduction in observed heterozygosity can be used to quantify the level of genetic of differentiation between the subpopulations. This quantification was formalized in the first instance by Wright (1951) in a series of hierarchical *F*-statistics.

Assuming that there are three levels of population structure: Individuals (I), Subpopulations (S) and the Total population (T). *F*-statistics can be illustrated as:

- F_{IS} is inbreeding coefficient (formula 6): is the mean reduction in heterozygosity of an *individual* due to non-random mating within a subpopulation, i.e. a measure of the extent of genetic inbreeding within subpopulations. F_{IS} ranges from -1.0 (all individuals are heterozygous) to +1.0 (no observed heterozygotes).

- F_{ST} is fixation index (formula 7): is the mean reduction in heterozygosity of a *subpopulation* (relative to total population) due to genetic drift among subpopulations, i.e. a measure of the extent of

F-statistics can be estimated from molecular markers using the following equations:

$$F_{IS} = \frac{H_S - H_I}{H_S} \qquad \textbf{(6)}$$

$$F_{ST} = \frac{H_T - H_S}{H_T} \qquad \textbf{(7)}$$

$$F_{IT} \frac{H_T - H_I}{H_T} \qquad \textbf{(8)}$$

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 hterozygosity for the total population.

genetic differentiation among subpopulations. F_{ST} ranges from 0.0 (no differentiation) to 1.0 (complete differentiation - subpopulations are fixed for different alleles).

- F_{IT} is overall fixation index (formula 8): is the mean reduction of heterozygosity of an *individual* relative to *total* population.

Examples: The value F_{ST} = 1 was estimated for two subpopulations of a fish species. This means that there is absolute differentiation between the two subpopulations, with each fixed for a different allele. Simply speaking, this can be interpreted as 100% of the total genetic variation is *between* subpopulations, with *zero* variation present within subpopulations.

The value F_{ST} = 0.47 was estimated for two subpopulations of another fish species. That is, there is a substantial differentiation *among* subpopulations, and 47% of total genetic variation is distributed among subpopulations, with 53% of the variation *within* subpopulations.

Although F_{ST} has a theoretical range of 0 to 1, the observed maximum is usually less than 1. Wright (1978) suggests the following general guidelines for the interpretation of F_{ST} based on allozyme loci:

- F_{ST} = 0.00-0.05 may be considered as indicating *little* genetic differentiation

- F_{ST} = 0.05-0.15 indicates *moderate* genetic differentiation

- F_{ST} = 0.15-0.25 indicates *very large* genetic differentiation

- F_{ST} = > 0.25 indicates *extensive* genetic differentiation

The three *F*-statistics described above can be extended to include higher levels of hierarchy. For example, if we have a series of subpopulations of a fish species which naturally occur in three separate river basins, then the following F_{ST} related statistics could be estimated:

- F_{ST}: the variance among subpopulation relative to the total variance

- F_{SC}: the variance among subpopulations within groups (denoted as C)

- F_{CT}: the variance among groups relative to the total variance.

What F_{ST} may tell us about gene flow

A single population may split into two subpopulations at some point, which then each diverge randomly over time. However, some subpopulations may share some migrants (for example fish migrate during flooding seasons from one pool to another). If all subpopulations share migrants with all other subpopulations with equal chance, then there is a simple relationship between F_{ST} and migration (m):

$$F_{ST} = \frac{1}{4N_m + 1}$$

where N_m is the actual number of individuals that migrate. However, it is only a relative measure of migration between populations. Sometimes N_m is referred to as such as in Arlequin software.

3.1.2. Haplotypic markers (RFLPs, SSCPs, TGGEs)

In order to analyse genetic diversity using haplotypic markers, it is necessary to provide clear classification of the different kinds of haplotypic markers. Haplotypic genetic markers include those generated from techniques such as Restriction Fragment Length Polymorphism (RFLP) and Single Strand Conformation Polymorphism (SSCP). However, in the case of RLFP, it is important to distinguish between *restriction fragment data* and *restriction site data*.

- Restriction fragment data are those generated from RFLP using randomly chosen restriction enzymes to digest the whole genome, such as the mitochondrial genome without knowledge of the actual sequences, and for which only the size of the generated fragments are available.

$$S = \frac{2m_{XY}}{m_X\, m_Y} \qquad \textbf{(10)}$$

Where:

- m_X and m_Y are the number of restriction sites in sequence X and Y, respectively
- m_{XY} is the number of restriction sites shared by both sequences
- *Proportion of nucleotide differences* is estimated as:

$$p = 1 - \sqrt[r]{S} \qquad \textbf{(11)}$$

- The *nucleotide diversity* is estimated as:

$$\pi = [-\ln S]/r \qquad \textbf{(12)}$$

- The *number of substitutions per site*, *d*, is estimated as follows:

$$d = -\frac{3}{4}\ln\left(1 - \frac{3}{4}p\right) \qquad \textbf{(13)}$$

- Restriction site data are those generated by RFLP technique, for which the precise location of a recognised sequence for a restriction enzyme is known. It is common nowadays that the actual sequences of a gene are firstly generated for a number of representative individuals of several populations, and restriction enzymes are designed on the basis of sequence differences.

Data generated from SSCP technique, can be converted into haplotypes. Different conformation of DNA fragments are due to different mutations along the sequences. However, the recognition site are mostly single nucleotide. Therefore the method of analysing data is slightly different to restriction site data.

Diversity within population

For restriction site data, Levels of diversity within population is expressed by estimates of proportion of shared restriction sites, proportion of nucleotide differences, nucleotide diversity and number of substitutions per site.

Example:

Look at the two DNA fragments X and Y below:

(1) (2) (3)

X xxxxGAATTCxxxxxxxxxxxxxxxxxxxxGA**A**TTCxxxxxxxxGAATTCxxxxxxxxxxxx

Y xxxxGAATTCxxxxxxxxxxxxxxxxxxxxGA**T**TTCxxxxxxxxGAATTCxxxxxxxxxxxx

The sequence GAATTC is recognised by enzyme EcoR I and therefore the fragments will be cut wherever this sequence is present. Sequence X will be cut at 3 positions while sequence Y will be cut at only 2 positions as there is a mutation at position (2) as highlighted in **bold** letter. In this case, the value of S is calculated as:

$$S = \frac{2m_{XY}}{m_X + m_Y} = \frac{2*2}{3+2} = 0.8$$

The length of the recognition sequence r = 6. Therefore proportion of nucleotide differences:

$$p = 1 - \sqrt[r]{S} = 1 - \sqrt[6]{0.8}$$

And nucleotide diversity:

$$\pi = [-\ln S]/r = [-\ln 0.8]/6 = 0.223$$

$$d = \frac{\sum_{k} m_k r_k d_k}{\sum_{k} m_k r_k} \qquad \textbf{(14)}$$

Where:

- m_k = average number of bands for the restriction enzyme k
- r_k = length of the sequence recognised by the enzyme k
- d_k = estimated nucleotide divergence

The probability that two sequences, X and Y, share the same recognition sequence at a given site is denoted as *S*. The maximum likelihood estimator of *S* (Nei & Tajima 1983) is shown in formula (10).

If several enzymes with the same length of the corresponding recognition sequences are used, it is possible to still use the above expression simply by taking summations over all enzymes. However, if several enzymes with different lengths in their recognition sequences are employed, then it is convenient to follow the method proposed by Nei and Miller (1990) in order to weigh the data obtained with each enzyme class. When values of *d* have been estimated for each class of restriction enzymes, then a *combined estimate of d for all enzymes* is given in formula (14).

Between population diversity

This estimate of nucleotide divergence can be extended to be an estimate of inter-population nucleotide divergence for all classes of restriction enzymes.

The estimate of nucleotide divergence can be extended to be an estimate of inter population nucleotide divergence of all classes of enzymes as shown in formula (15).

This estimate of inter-population divergence includes both an inter-population and an intra-population component. If the interest is only to

$$d_{AB} = \frac{\sum_{k} m_k r_k d_k}{\sum_{k} m_k r_k} \qquad (15)$$

Where:

- m_k = average number of bands for the restriction enzyme k
- r_k = length of the sequence recognised by the enzyme k

focus on inter-population divergence, then it can be estimated as formula (16).

3.1.3. Population structure

When dealing with haplotypic data such as those generated from the RFLP technique, *F*-statistics can be done in a different way compared to those estimated from co-dominant markers such as allozymes or microsatellites. If RFLP data is collected by using restriction enzymes to digest the whole mitochondrial genome (e.g. without the knowledge on the actual sequences) then haplotypes diversity can be used (instead of heterozygosity in the case of co-dominant markers). Haplotype diversity is calculated using formula (16).

The table below shows an example how haplotypes diversity is calculated.

Now replace heterozygosity by the haplotype diversity in formula (6) to calculate F_{ST}.

For the data generated by using PCR-RFLP and DNA sequences where there is much more information that can be used to determine population structure rather than just using haplotypes diversity. This will be dealt with in Section 3.1.5.

Haplotype diversity:

$$H = 1 - \sum_{i}^{j} p_i^2 \qquad (16)$$

3.1.4. Dominant markers

Analysis of dominant data has been hampered by the failure of common methods to correct for the inability to detect al genotypes. This can result in a serious underestimation of the actual level of genetic diversity (Clark & Lanigan 1993). Recently, two methods for overcoming this difficulty and thus enabling the use of dominant data have been proposed (Clark & Lanigan 1993; Lynch & Milligan 1994).

The method proposed by Clark and Lanigan (1993) uses the frequency of the absence of a fragment in a population sample as an estimate of the population frequency of recessive heterozygotes (q^2) and then uses this value to correct for the relative detectability of individuals who have one versus two copies of a fragment. Once this correction has been taken

Haplotype	Population 1		Population 2	
	Frequency (p_i)	p^2	Frequency (p_i)	p_i^2
1	0.50	0.25	0.90	0.81
2	0.40	0.16	0.10	0.01
3	0.10	0.01	0.00	0.00
Sum		0.42		0.82
H		0.58		0.18

into account, data are treated in a very similar way to that already described for restriction fragment data. In the case of RAPD data, *r* corresponds to the primer length used for random amplification (usually *r*=10). However, there are a number of assumptions that have to be made as outlined by Clark and Lanigan (1993):

1. The amplification of a fragment depends strictly on the exact match between the oligonucleotide primer sequence and the genome template sequence.

2. Polymorphisms due to insertion/ deletion variation are rare.

3. Fragment of the same size in different population are locus specific.

4. Fragments can be identified unambiguously.

5. Nucleotide sequence diversity (π) should not exceed 10%.

Lynch and Milligan (1994) have adopted a different approach for analysing population structure using RAPDs. This approach assumes that alleles from different loci do not co-migrate to the same position in the gel, and that the researcher is capable of matching bands from different lanes within and among gels, and that each locus can be treated as a two-allele system, with a presence and an absence allele. Lynch and Milligan (1994) adopt the following estimate for the *gene frequency*, q, of the null allele at one locus (formula 17).

Once gene frequencies have been estimated, it is possible to estimate *gene diversity (or heterozygosity)* within a population:

$$H = 2pq = 1\text{-}p_i^2$$

Gene diversity, *H*, is the probability that two genes randomly chosen from population differ at a locus, is equivalent to the expected heterozygosity under Hardy-Weinberg equilibrium.

Other diversity indices can be estimated from dominant markers, These parameters are implemented in a free software package to analyse dominant data such as RAPDistance (Amstrong *et al*. 1995):

- Average gene diversity over loci (formula 18)

$$q = \frac{\sqrt{x}}{1 - \frac{\mathrm{var}(x)}{8x^2}} \qquad (17)$$

Where x is the frequency of null homozygotes (frequency if the absence phenotype on gel).

If loci have been sampled in population A, the average gene diversity in this population is:

$$h_{\mathrm{A}} = \frac{1}{X}\sum_{\mathrm{i}=1}^{X} h_{\mathrm{A_i}} \tag{18}$$

If *n* populations have been sampled, the average within-population gene diversity can be estimated as:

$$h = \frac{1}{n}\sum_{\mathrm{A}=1}^{\mathrm{n}} h_{\mathrm{A}} \tag{19}$$

The heterozygosity between populations A and B at the *i*th locus can be estimated by:

$$h_{\mathrm{AB(i)}} = q_{\mathrm{A_i}} + q_{\mathrm{B_i}} - 2q_{\mathrm{A_i}}q_{\mathrm{B_i}} \tag{20}$$

If there is no population subdivision, the gene frequencies in all sub-populations are the same, therefore $H_{AB} = H_A = H_B$, and the inter-population component of diversity can be estimated as:

$$H_{\mathrm{AB(i)}} = h_{\mathrm{AB(i)}} - \frac{h_{\mathrm{A(i)}} + h_{\mathrm{B(i)}}}{2} \tag{21}$$

Averaging over all X loci, the estimated mean gene diversity between poulations A and B is:

$$H_{\mathrm{AB}} \frac{1}{X}\sum_{\mathrm{i}=1}^{X} h_{\mathrm{AB(i)}} \tag{22}$$

And the mean between population gene diversity can be obtained by averaging over all pairs of populations:

$$H = \frac{2\sum H_{\mathrm{AB}}}{n(n-1)} \tag{23}$$

- Average gene diversity over populations (formula 19)
- Gene diversity at one locus (formula 20)
- Heterozygosity at one locus (formula 21)
- Average heterozygosity over all loci (formula 22)
- Average heterozygosity over all populations (formula 23)

In order to determine population structure using dominant markers, there are two options. The first option treats RAPDs as restriction data, and therefore *F*-statistics can be estimated as mention in Section 3.1.1 If the data showed evidence of Mendelian inheritance, then the principle used for co-dominant marker can be applied. It is suggested that when dealing with dominant markers such as RAPDs and AFLPs, it is ideal if an inheritance study can be conducted to test for Mendelian inheritance patterns of the alleles.

3.1.5. DNA sequences

Diversity indices

There are two different measures for the amount of genetic variation at the nucleotide level: The average number of pairwise nucleotide differences and the number of segregating (polymorphic) sites among a sample of sequences.

Number of segregating sites is the number of variable nucleotide sites in a sample of sequences. The disadvantage of this measure is that it does not incorporate the length of the sequence analysed and hence is not comparable across the data set.

Proportion of segregating sites is the number of segregating sites divided by the length of the sequences.

Consider the three sequences below:

- Seq1: AAATAGTCCT
- Seq2: AAACGGTCCT
- Seq3: AAACGGTTCT

The number of nucleotide differences between the three sequences can be presented in a table as follows:

Average number of pairwise nucleotide differences (*d*) is defined as:

$$d = \frac{\textit{Total number of nucleotide differences for all pairwise comparisons}}{\textit{Total number of pairwise comparisons}}$$

Or more generally, it can be estimated using the formula below:

$$d = \frac{2\sum_{i<j} d_j}{n(n-1)} \qquad (23)$$

- Where d_{ij} is the number of nucleotide differences between sequences i and j.
- *n* is the number of DNA sequences under comparison.

	Seq1	Seq2	Seq3
Seq1	*		
Seq2	2	*	
Seq3	1	1	*

The average number of pairwise nucleotide differences between the three sequences will be:

$$2(2 + 1 +1)/3(3-1) = 1.33$$

Nucleotide diversity (often denoted as π) equate to the average number of pairwise nucleotide differences between sequences divided by the length of sequences.

For example, consider the sequences above, $\pi = 1.33/10 = 0.133$

Population structure

For DNA sequences for the data generated from PCR-RFLP, Nei (1982) developed a similar measure of population differentiation as F_{ST}, but this time using a measure of nucleotide diversity (π) within a population, in place of heterozygosity or haplotype diversity. If we define π_{ij} as the genetic distance between haplotype i and haplotype j then the nucleotide diversity within the *total* population is:

$$\pi_T = \sum_{ij} p_i p_j \pi_{ij} \qquad (24)$$

where p_i and p_j are the overall frequencies of haplotypes i and j, respectively. That is, the distances between haplotypes pairs are simply weighted by how common they are, to arrive at an average. If we also define π_s as the average nucleotide diversity within *subpopulations*, then we can derive a familiar expression for an F_{ST}, like nucleotide measure of subpopulation differentiation:

$$\frac{\pi_T - \pi_S}{\pi_T} = \frac{\pi_B}{\pi_T} \qquad (25)$$

where π_B is the average nucleotide diversity between subpopulations

This statistic could also be called F_{ST} , but it was originally described by Nei (1982) as γ_{ST} . A related statistic derived by Lynch & Crease (1990) was called N_{ST} , one derived for mtDNA data by Takahata and Palumbi (1985), G_{ST}, and one by Excoffier et al. (1992, see below), Φ_{ST} (phi-st). Although each of these statistics for nucleotide data is calculated slightly differently, in reality they are all trying to estimate the same parameter - the proportion of nucleotide diversity among subpopulations, relative to the total – and their values are usually quite similar, particularly with large sample sizes. The same things can be said for all the different ways of calculating F_{ST} from allelic data (including G_{ST} and Φ_{ST}). Given the multitude of different descriptor variables used by different authors, and the fact that within the two classes they are trying to estimate essentially the same parameters, the convention is to refer to the allelic form of the statistic as F_{ST} , and the nucleotide diversity form as Φ_{ST}.

There is also a very simple conceptual relationship between the allelic F_{ST} and the nucleotide Φ_{ST}, shown below. In the allelic calculations (F_{ST}), it is assumed that all alleles are equidistant from each other, while in the nucleotide diversity calculations (Φ_{ST}), there are different distances between different alleles. (This can indeed be applied to any other type of distance calculation you might require, such as between microsatellite alleles). In fact, now we can actually calculate the allelic F_{ST} in exactly the same way we calculate Φ_{ST} (in AMOVA – see below) by simply making all the distances between alleles equal one (i.e., replace the pairwise distance matrix by a unity matrix). This is shown below. F_{ST} and Φ_{ST} values then can be estimated as shown in the example that follows.

As a result, the two values are different. This is because the two forms are really measuring different properties of the data.

One is not necessarily any 'better' than the other. Even in terms of simply detecting whether or not there is significant differentiation between subpopulations, it depends entirely on the data set as to which form of F_{ST} - allelic or nucleotide distance – is the more powerful statistically. So the bottom line is that it is useful to calculate both types of F_{ST} for one given data set.

For F_{ST}		Pairwise distance		For Φ_{ST}		Pairwise distance	
Allele	A	B	C		A	B	C
A				A			
B	1			B	1		
C	1	1		C	2	1	

Allele	Population 1	Population 2	Total
A	0.80	0.00	0.40
B	0.20	0.20	0.20
C	0.00	0.80	0.40
F_{ST} $H = 1 - \sum p_i^2$	1 - (0.80^2+0.20^2) = 0.32	1 - (0.20^2+0.80^2) = 0.32	1-(0.40^2+0.20^2+0.40^2) = 0.64
Φ_{ST} $\pi_T = \sum_{ij} p_i p_j \pi_{ij}$ $F_{ST} \frac{H_T - \overline{H}_S}{H_T} = \frac{0.64 - 0.32}{0.64} = 0.50$ $\Phi_{ST} = \frac{\pi_T - \overline{\pi}_S}{\pi_T} = \frac{0.48 - 0.16}{0.48} = 0.67$	0.80*0.20*1 = 0.16	0.20*0.80*1 = 0.16	0.40*0.20*1 = 0.08 +0.20*0.40*1 = 0.08 +0.40*0.40*2 = 0.32 **=0.48**

In population genetics studies, DNA sequences can be analysed using the ARLEQUIN software package, where important factors such as number of variable sites, haplotype frequency, pairwise number of nucleotide differences etc. can be obtained. Furthermore, analysis of population differentiation using *F*-statistics is also implemented, or population genetic structure using Analysis of Molecular Variance (AMOVA) can also be applied using this software.

3.2. Statistical tests

The rapid rate of increase of technical applications to population genetic studies has been paralleled by the number of statistical tests developed to analyse them and probably exceeded by the number of computer software programs available to perform the analyses. We certainly do not have the time or scope to investigate them all. Here, and in the next section, we only look at a few basic analyses (both quantitative and qualitative) that are useful for understanding population structure.

3.2.1. Neutrality tests

Several methods have been designed to use DNA polymorphism data (sequences and allele frequencies) to obtain information on past selection events. Most commonly, the ratio of non-synonymous (replacement) to synonymous (silent) substitutions (d_N/d_S ratio; see below) is used as evidence for overdominant selection (balancing selection) of which one form is heterozygote advantage. A classic example of this is the mammalian major histocompatibility complex (MHC) system genes and other compatibility system in other organisms: the self-incompatibility system of plants, fungal mating types and invertebrate allorecognition systems. In all these genes, a very high number of alleles are present. This can be interpreted as an indicator of some form of balancing (diversifying) selection. In the case of neutral polymorphisms, a single common allele and a few rare alleles are expected. The frequency distribution of alleles is also informative. A large number of alleles showing a relatively even distribution is against neutrality expectations and suggestive of diversifying selection.

Most tests detect selection by rejecting neutrality (observed data deviate significantly from what is expected under neutrality). This deviation, however, may also be due to other factors such as changes in population size or genetic drift. The original neutrality test was Ewens-Watterson homogeneity of neutrality based on a comparison of observed and predicted homozygosity calculated by Ewens's sampling formula which uses the number of alleles and sample size. This test is not very powerful.

Other commonly used statistical tests of neutrality include: Tajima's *D* (theta, θ), Fu and Li's *D*, *D** and *F*. Tajima's test (Tajima 1989) is based on the fact that under the neutral model estimates of the number of segregating/polymorphic sites and the average number of nucleotide differences are correlated.

If the value of *D* is very large or very small, the neutral 'null' hypothesis is rejected. **DnaSP** calculates *D* and its confidence limits (two-tailed test). Tajima did not base this test on the coalescent but Fu and Li's tests (Fu & Li 1993) are directly based on the coalescent. The tests statistics *D* and *F* require data from intraspecific polymorphism and from an outgroup (a sequence from a related species), while *D** and *F** only require intraspecific data. **DnaSP** uses the critical values obtained by Fu and Li (1993) to determine the statistical significance of *D*, *F*, *D** and *F** test statistics. **DnaSP** can also conduct the F_S test statistic (Fu 1996). The results of this group of tests (Tajima's *D* and Fu & Li's tests) based on allelic variation and/or level of variability may not clearly distinguish between selection and demographic alternatives (bottleneck, population subdivision) but this problem only applies to the analysis of a single locus (demographic changes affect all loci whereas selection is expected to be locus-specific which are distinguishable if multiple loci are analysed). Tests for multiple loci include the **HKA test** described by Hudson et al. (1997). This test is based on the idea that in the absence of selection, the expected number of polymorphic (segregating) sites within species and the expected number of 'fixed' differences between species (divergence) are both proportional to the mutation rate, and the ratio of them should be the same for all loci. Variation in the ratio of divergence to polymorphism among loci suggests selection.

For other tests and software to perform these statistics, see websites in **DnaSP**.

3.2.2. Linkage disequilibrium (LD)

Two 'alleles' can be present on the same chromosome (positive LD), or not segregate together (negative LD). As a result, specific alleles at two different loci are found together more or less than expected by chance. LD is the non-independence, at a population level, of alleles carried at different positions in the genome. In this case, the expected frequency of a two-locus haplotype can be calculated as the probability of the occurrence of two independent (or joint) events simply by multiplying their gene frequencies. The same situation may exist for more than two alleles. Its magnitude is expressed as the delta (Δ) value and corresponds to the difference between the expected and the observed haplotype frequency. If there is no LD, Δ will be zero (or not significantly different from zero), if there is positive LD it will be a positive value. It can also be negative if the two alleles tend not to occur together. The statistical significance of LD, which depends on the sample size, and the magnitude of LD are separate issues. The statistical significance is determined usually by Fisher test and the magnitude is determined by either Δ value or alternative measures. The magnitude can be normalised (for allele frequencies) to have the same range of values for any frequency.

Ideally, haplotype frequencies should be calculated from family data. Obviously, this gives the most accurate results. In practice, however, when family data are not available, Δ and two-locus haplotype frequencies are calculated from a sample of the population data by constructing 2x2 contingency tables for each allele pair. A contingency table for this purpose contains the individual (observed) values cross-classified by levels for two different attributes. A common 2x2 table constructed in genetic studies is as follows in the table below.

Counts for each combination of levels (presence or absence) of the two factors (alleles) are placed in each cell. The corresponding Δ_{ij} is estimated by the formula (usually in HLA studies):

$$\Delta_{ij} = (d/N)^{1/2} - [((b+d)/N)((c+d)/N)]^{1/2} \quad (26)$$

The haplotype frequency (HF_{ij}) equals G_{Fi} x $G_{Fj} + \Delta_{ij}$, where GF is the gene frequency (the proportion of the chromosomes carrying a particular allele). The haplotype frequency calculated with this formula from the population data compares reasonably well with estimates obtained directly from counting haplotypes constructed from family segregation data. This method generates a reliable estimate of a haplotype frequency with the exception of very small haplotype frequencies. Also for other parts of the genome, it has been reported that there is little or no advantage to constructing haplotypes from family data rather than unrelated individuals. The major point is that when using population data, genotyping errors become an issue. When genotyping a large number of markers, an error rate of only 1% will produce a large number of inaccurate haplotypes. Genotyping errors are not the only possible sources of accuracy problems. Other factors include sample size, allele frequency distributions and departures from HWE.

There are other measures of LD. Because the value of Δ depends on allele frequencies a normalisation of Δ is needed. This is achieved by taking into account the allele frequencies: normalised delta value (D') = Δ_{AB} / Δ_{max}. Δ_{max} is the lesser of $p_A p_b$ or $p_a p_B$ if Δ is positive or $p_A p_B$ or $p_a p_a$ if Δ is negative. Because the sign is arbitrary, |D'| is often used rather than Δ'. Therefore, D' (normalised LD) is scaled to remove allele frequency effects. In a large enough sample, D' = 1 that indicates complete LD and D' = 0 corresponds to no LD. |D'| is directly related to recombination fraction and its generalization to more than two loci is the only measure of LD not sensitive to allele frequencies. **ASSOCIATE** and **HAPLOVIEW** are some of the software that calculate D' values.

		Allele i		
Allele j		**Present (+)**	**Absent (-)**	**Row totals**
	Present (+)	*a* (+/+)	*b* (+/-)	*a*+*b*
	Absent (-)	c (-/+)	d (-/-)	*c*+*d*
	Column totals	*a*+*c*	*b*+*d*	*N*=*a*+*b*+*c*+*d*

Another linkage disequilibrium statistic is the square of the correlation coefficient (r^2) between the alleles at locus A and B: $r^2 = \Delta^2 / (p_A\ p_a\ p_B\ p_b)$ which can also be expressed as $r^2 = \Delta^2 / (p_A (1-p_A)\ p_B\ (1-p_B))$ (for two loci with two alleles each). The measure r^2 has several properties that make it more useful. In brief, for low allele frequencies r^2 has more reliable sample properties than |D'|.

LD estimates can be calculated using various software such as **ARLEQUIN**, **GDA**, **PopGene** amongst others.

Interpretation of LD Data: The patterns of LD observed in natural populations are the result of a complex interplay between genetic factors and the population's demographic history. LD is usually a function of distance between the two loci. This is mainly because recombination acts to break down LD in successive generations. When a mutation first occurs it is in complete LD with the nearest marker (D' = 1.0). Given enough time and as a function of the distance between the mutation and the marker, LD tends to decay and in complete equilibrium reaches a D' = 0 value. Thus, it decreases every generation of random mating unless some process opposes the approach to linkage 'equilibrium'. However, physical distance could account for less than 50% of the observed variation in LD. One genetic phenomenon that affects LD is gene conversion. Gene conversion is an important mechanism in the breaking down of allelic associations over short distances, i.e., decay of LD. Other factors that influence LD include changes in population demographics (such as population growth, bottlenecks, geographical subdivision, admixture and migration) and selective forces. Admixture (intermixture of populations) would cause LD if the mixing populations have different allele frequencies.
LD will also be erased faster in large populations than in small ones (chance in small populations maintain LD). Permanent LD may result from natural selection if some gametic combinations confer higher fitness than alternative combinations. Regional LD may also be variable according to haplotype. An example has been presented for HLA haplotypes. Haplotype-specific patterns of LD may reflect haplotype-specific recombination hotspots as has been shown for mouse MHC.

Note that LD has nothing to do with HWE and should not be confused with it.

3.2.3. Testing F_{ST} for significance

Once we have calculated our F_{ST} values, we need to know whether they represent significant population structure or not. Theoretically F_{ST} ranges from zero (no structure) to one (total differentiation). Any values that fall in between these extremes need to be statistically determined for significance. That is we are testing the null hypothesis that H_o: $F_{ST} = 0$ is not significantly different from 0.

There are several methods for determining the significance of F_{ST} values. Firstly, the estimated F_{ST} can be compared to a χ^2 distribution with the relationship between F_{ST} and the χ^2 critical value as:

$$\chi^2_{(crit)} = 2N^*F_{ST}(k\text{-}1) \qquad \textbf{(27)}$$

with degrees of freedom:

$$df = (k\text{-}1)(s\text{-}1)$$

As can be seen, this method is affected by sample size (N), the number of alleles present in the sample (k) and the number of population samples (s). Therefore the greater the number of samples and/or a more variable marker will increase the power of this test. Although this is a simple test (it can be calculated by hand), it is however, a relatively conservative test with a moderate chance of committing a Type II error (i.e. incorrectly accepting that there is no structure).

A more powerful method for determining the significance of F_{ST} is the Permutation Test. This method permutes (randomises) haplotypes among populations and calculates an F_{ST}. This process is repeated many times (e.g. 1000) resulting in a distribution of F_{ST} values. We can then place the F_{ST} observed from the original data into the distribution and make a judgement as to the level of significance. For example if 95% of the simulated F_{ST} values fall below our true F_{ST} value, then we can say that we are 95% confident that the estimated value is significantly different from zero (i.e. reject the null hypothesis at the α=0.05 level).

Another powerful test to determine whether significant differentiation exists among sample sites is the Exact Test of Raymond and Rousset (1995) which is analogous to Fisher's exact test and is based on a 'number of populations' x 'number of haplotypes' contingency table. This test employs a Markov chain random walk method that provides an unbiased estimate of the exact probability of incorrectly rejecting the null hypothesis (null hypothesis = there is no differentiation among sites) which is committing a Type I error. It is a particularly powerful test with small sample sizes and rare alleles. This is a 'whole of table' test that estimates the probability of observing a table less likely than the observed configuration under the null hypothesis of panmixia.

3.2.4. Estimating gene flow from mtDNA sequence data

Directly quantifying gene flow is extremely difficult (if not impossible) in natural systems because observing all migrants (dispersers) is often not possible and even if you can observe all dispersers you have no idea as to how many of them actually contribute their genes to future generations in the receiving population and in what proportion.

We can however, estimate gene flow (N_em) indirectly from inference gained from the relationship between genetic drift and gene flow at equilibrium (i.e. F_{ST}) where:

- N_em = number of migrants
- N_e = effective population size
- m = proportion of receiving population that is made up of migrants

The relationship between gene flow and population structure is:

$$F_{ST} = 1/(1+2N_em_f) \text{ (for mtDNA)} \quad \textbf{(28)}$$

Notice that here we are estimating effective female gene flow as only females can contribute their mtDNA genes to subsequent generations. As we can estimate F_{ST} from the data, we are able to estimate effective dispersal. Note that as N_em increases, F_{ST} decreases, as would be expected. This equation is specifically the relationship between gene flow and structure under the assumptions of the island model of gene flow. The relationship for the stepping-stone model is:

$$F_{ST} = 1/(1+2N_em_f)(\ 2N_e\mu\ /\ N_em_f)^{½} \quad \textbf{(29)}$$

The first thing to notice is that this model incorporates the effects of the mutation rate (μ). As population size (N_e) increases, the stepping-stone model approaches the island model. It has been argued that the difference between these two models in estimating gene flow is negligible. Most computer programs that calculate N_em in this way, do so under the assumptions of the island model.

There are several limitations however, to this method of estimating gene flow. 1) the gene flow model used (usually the island model) may not be appropriate for the system that you are working with; 2) the assumption of drift/gene flow equilibrium is usually violated; 3) it does not account for sex biased dispersal (i.e. for mtDNA we are only estimating female gene flow and dispersal may be male mediated; 4) it has been argued that the relationship between N_em and F_{ST} only holds true for global F_{ST} estimates and therefore is not relevant to pairwise comparisons (which is what we are really interested in); 5) it assumes dispersal is equal in both directions which is highly unlikely in freshwater systems.

Another potential problem with this method relates to the genetic marker chosen for the study. Any F_{ST} value less than 1 (absolute differentiation) will result in a positive estimate of gene flow. We know that F_{ST} is the partitioning of variation within and among populations so that a hyper-variable marker in a very large population may result in a surprisingly low level of differentiation. The lower the F_{ST}, the higher the estimate of N_em, even if two populations do not share any haplotypes in common. Furthermore, two populations that are totally isolated from each other today, may have haplotypes in common from a time when there was connectivity (ancestral retention). In this case, the methods for estimating gene flow above will be measuring historical rather than contemporary dispersal patterns. Being able to distinguish between these two scenarios is dealt with in section 3.3.1.

Because of the limitations of this method, it is strongly suggested that estimates of the number of migrants never be treated as absolute values (although this is tempting for management purposes). At best, one could infer differences in dispersal rates by orders of magnitude (e.g. N_em between populations A and B is 10 times greater than between populations A and C).

3.2.5. The Stream hierarchy model of gene flow

The island and stepping-stone models are not really appropriate for F_{ST} and gene flow analyses in freshwater systems for several reasons, including the dendritic nature of river systems and the unequal bi-directional gene flow which is idiosyncratic to individual species. For this reason there has not been a simple equation that describes the relationship between F_{ST} and gene flow in these systems. However a gene flow model has been proposed (Stream Hierarchy Model (SHM)) to explain the distribution of genetic variation in freshwater systems were variation can be partitioned at different hierarchical levels within the system (Meffe & Vrijenhoek 1988). In this model the total genetic diversity (H_T) is partitioned into several components:

$H_T = H_C + D_{CR} + D_{RS} + D_{ST}$ (30)

- H_C = variation within populations
- D_{CR} = differences among populations in a river
- D_{RS} = differences among rivers in a drainage
- D_{ST} = divergence among drainages
- Under the SHM it is expected that $D_{CR} < D_{RS} < D_{ST}$

Analysis of Molecular Variance (AMOVA) (Excoffier et al., 1992) is a method of analysis that is ideal for investigating population structure and gene flow in a hierarchical fashion in freshwater systems. It is analogous to the standard Analysis of Variance (ANOVA), a standard parametric statistical procedure for partitioning error (variation about the estimate of the mean) within and among treatments. Similarly, AMOVA can partition genetic variation within and among populations. Although this is essentially what normal F_{ST} analysis does, AMOVA can partition genetic variation at

Table 18. AMOVA table showing the partitioning of genetic variation at different hierarchical levels.

Source of variation	d.f.	Sum of squares	Variance components	%variation
Among groups	4	7.768	0.13830	15.46
Among populations within groups	4	2.686	0.00560	1.63
Within populations	92	70.099	0.76195	82.91
Total	**100**	**80.552**	**0.89464**	

spatially different hierarchical levels (i.e. hierarchical F_{ST}). Hence, genetic variation can be partitioned into

- F_{IS} = among individuals within populations
- F_{SC} = among populations within region (rivers)
- F_{CT} = among regions within total (among river drainages)

Due to the restricted nature of dispersal within and among river drainages, the Stream Hierarchy Model predicts that $F_{CT} >> F_{SC}$. The significance of these parameter estimates can be determined using a permutations test. Of course this method relies on a hierarchical sampling strategy as described later.

3.2.6. Isolation by distance

Sometimes in river systems the only barrier to dispersal is distance. In vast drainages such as the Mekong River, it may be physically impossible for an individual to traverse the geographic distance between two populations within a single lifetime. If this is the case, then a signature of isolation by distance (IBD – not to be confused with 'identical by descent') may result. To test for this pattern, we generally use a Mantel's test that is an extension of a standard parametric correlation analysis. Here we are looking for a correlation between levels of genetic differentiation (F_{ST} – the dependent variable) and geographical (or stream) distance (the independent variable). A positive relationship will indicate IBD (i.e. the greater the geographic distance separating two populations, the higher the F_{ST}). The reason that we cannot use a standard correlation is that the basic assumption of the test of independence of the samples is violated. This is because all estimates of both genetic and geographic distance are pairwise measures. Therefore the same population will be included in many of the data points in the analysis.

To test for IBD we use a permutation procedure similar to that described above. Firstly we calculate our pairwise F_{ST} matrix (other measures of genetic distance may be used) and a pairwise geographic distance matrix and perform a normal correlation test to get our correlation coefficient (*r*). The geographic matrix is then randomised and another correlation coefficient is calculated. This process is repeated many times until we have a distribution of simulated *r* values. We then place the observed *r* value into the distribution and determine the level of significance as described previously.

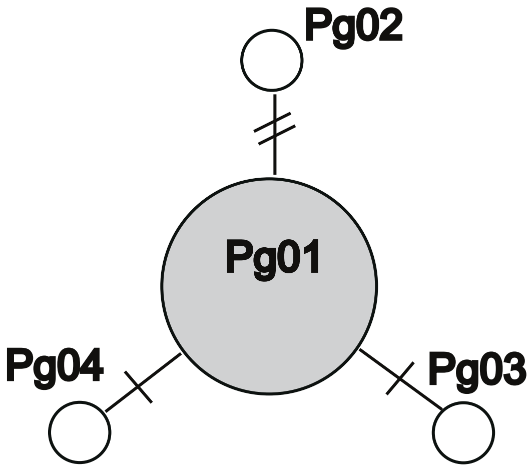
3.3. Presentation of data

An important component of population data analysis is to be able to display the data is a way that makes it easy to understand and to assist in communicating the results to managers. The methods can be categorised as either qualitative or quantitative or sometimes a combination of both. The most common way of displaying mtDNA data is by gene trees of which there are several types.

3.3.1. Minimum spanning networks

A minimum spanning network (or cladogram) displays the relationship among unique haplotypes in the sample (see Figure 13), relying on the number of base pair differences. This method is particularly appropriate for intraspecific studies using mtDNA analysis and can provide significant qualitative insight into both population and evolutionary processes that may have influenced observed patterns.

Figure 13. A minimum spanning network of four mtDNA 16S rRNA haplotypes obtained from 16 individuals of the critically endangered Mekong giant catfish, *Pangasianodon gigas* **(Na-Nakorn et al. 2006).**



Because of the way mtDNA evolves in the context of population processes, we can make several fairly strong assumptions. For example, haplotypes that are internal to the network tend generally to be older (ancestral, e.g. Haloptype Pg01 in the above figure), higher in frequency in the sample and geographically widespread. On the other hand, tip haplotypes are generally more recently evolved (derived) and are therefore younger and usually represented in fewer individuals.

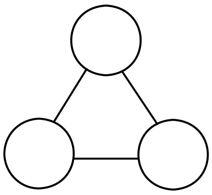
If we relate an individual haplotype's position in the network to its geographic distribution we have some power (albeit qualitative) to be able to differentiate between historical gene flow or ancestral retention and contemporary gene flow. For instance, if internal haplotypes are widespread but tip haplotypes are restricted then historical gene flow is a more likely scenario. On the other hand, widespread internal and tip haplotypes indicate contemporary gene flow. The study of the relationship between population genealogies and their geographical distributions is referred to as phylogeography.

Attempts have been made to test statistically the relationship between haplotype position in the network and their respective spatial distribution. One common method is Nested Clade Analysis (NCA) (Templeton *et al* 1995). By nesting the cladogram, the relative ages of the haplotypes and clades that they are nested within, can be determined. A statistical approach is used to determine whether the haplotypes (or their clades in which they are nested) are distributed over a significantly small or large distance. This can then make it possible to differentiate between restricted and long distance dispersal, isolation by distance, historical range expansion or allopatric fragmentation. Although this analysis is grounded in a strong statistical framework, it still relies on a largely qualitative interpretation of the results. A few limitations exist with network analysis. The most obvious is that the data are not used to their full potential. Simply using the number of base pair differences to determine the relationship among haplotypes excludes the opportunity to incorporate more complex mutation models that may better reflect the evolutionary history of the population. Secondly, homoplasy caused by multiple substitutions at a single nucleotide site can mask the true relationship among haplotypes. For example the following DNA sequences produce the network below.

ATCAG

ACCAG

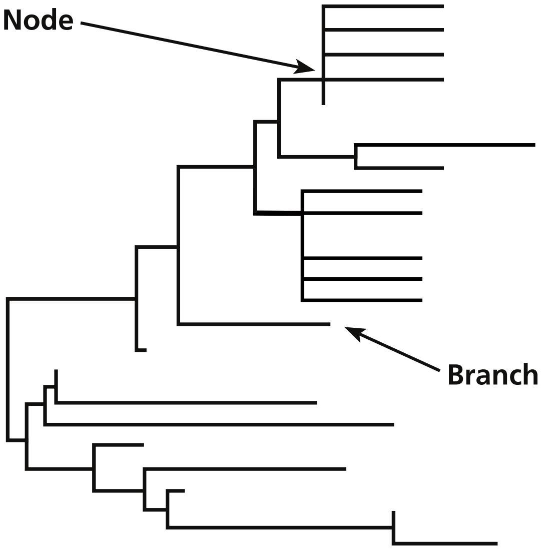
AGCAG



In this case it is impossible to discern which haplotypes are tip and which are internal. Also homoplasy can make it difficult to group closely related haplotypes with any statistical probability. Furthermore, incomplete sampling may result in a haplotype being labelled as a tip when in fact it is internal (and possibly ancestral to the whole cladogram).

3.3.2. Neighbour-joining trees

The other common method for displaying mtDNA data is by reconstructing genealogies from the sequence data. There are several methods for building gene trees, some require considerable computational power. However, one of the most appropriate for intraspecific studies is also one of the simplest. The neighbour-joining method clusters haplotypes based on the level of distance/similarity that allows for unequal rates of molecular change among haplotypes resulting in a tree with varying branch lengths. The basic method is to calculate the pairwise distances among all haplotypes. These distances are then scaled based on the distances among all other pairwise comparisons. The haplotypes that share the lowest scaled value are joined first through node 1 (a node is the internal point in a tree where two or more branches converge). The next haplotype is then joined to the tree through node 2 with the distance between node 1 and node 2 calculated and so on until all haplotypes are joined. Although this is a simple interpretation of the algorithm for building a neighbour-joining tree, we will not go into further detail here but the method is readily available in many computer programs.



An advantage that the neighbour-joining method has over networks is that the appropriate distance method can be applied and that the confidence of haplotype groupings can be determined statistically. The most common form of testing the haplotype groupings (clades) statistically is bootstrapping. Bootstrap values provide an estimate of how well a particular node in the tree is supported. Bootstrapping has become a general term for a different permutation tests relating to gene trees. For DNA data, bootstrapping involves randomly removing a nucleotide base or a number of bases from every haplotype sequence and generating a neighbour-joining tree. The base positions are then replaced and others removed to build another tree. This process is repeated (usually >500 times). Finally the tree from the total data set is calculated. The bootstrap values are the percentage of time that a particular node was supported during the permutations. Although neighbour-joining trees are more statistically rigorous, one disadvantage is that it is not always easy to differentiate between ancestral and recently derived haplotypes.

3.3.3. Population trees

Usually we are more interested in the similarities among populations (especially when we are seeking to determine population structure) than that among haplotypes. The neighbour-joining method is also useful for this purpose. The same algorithm is performed as mentioned above but the matrix consists of pairwise population distances. Once again there are several population distance methods available of which we will look at two here.

The first is Nei's D_A which is a net genetic divergence among populations where:

$$D_A = \pi_{12} - \frac{\pi_1 + \pi_2}{2} \qquad \textbf{(31)}$$

This method is sensitive to differential drift pressure between populations and is therefore a good measure when there is little variation in population sample size.

Another measure is the Carvalli-Sforza chord distance D_{CS} which standardises distances with respect to drift and therefore is more appropriate when drift is the main process causing differentiation:

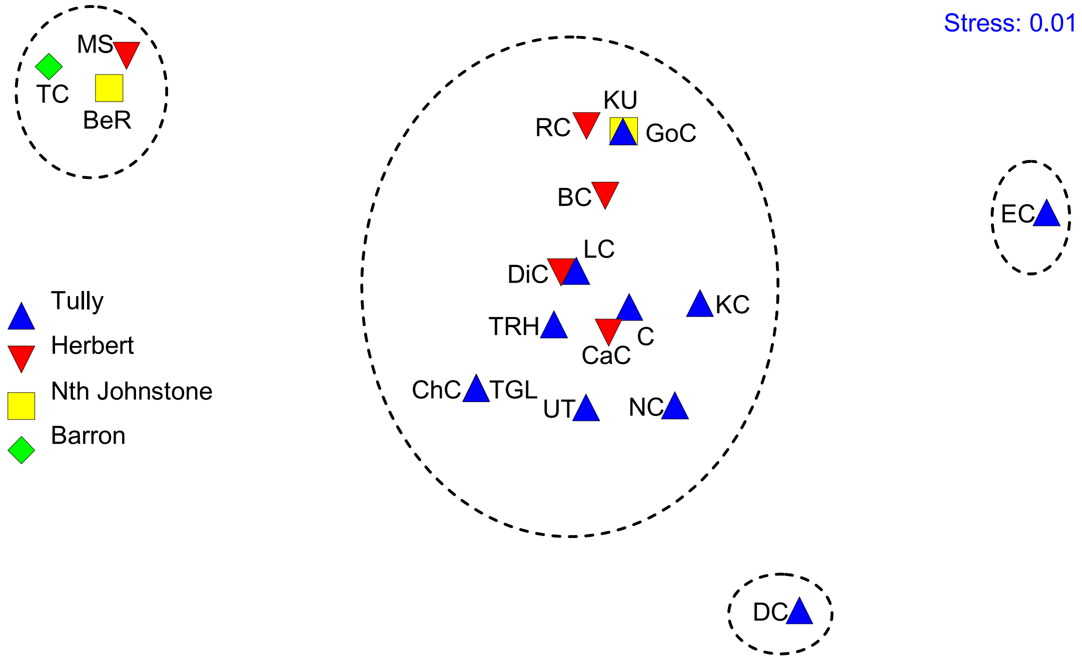
$$D_{CS} = \frac{2\sqrt{2}}{\pi}\sqrt{1-\cos\theta} \qquad \textbf{(32)}$$

where:

$$\theta = \sum \sqrt{x_i y_i}$$

A limitation of the population tree method is that some populations may be forced into groups erroneously. This may happen if one population shows some affinity with one of two closely related populations but not to the other.

Figure 14. An example result of MDS analysis.



An unbiased way of visualising populations' affinities to each other is using multivariate techniques. Multidimensional scaling ordinations (MDS) are a non-biased method that calculates a pairwise similarity matrix between *n* populations in *n*-dimensional hyperspace. It ranks the values and then expresses this ranking in a lower order plot, usually visualised in two or three dimensions. The MDS configuration is constructed to preserve the similarity ranking as Euclidean distances in the lower-dimensional plot. How well the true relationship among populations is represented in the lower-dimensional plot is tested by the measure of 'stress' (a measure of goodness of fit).

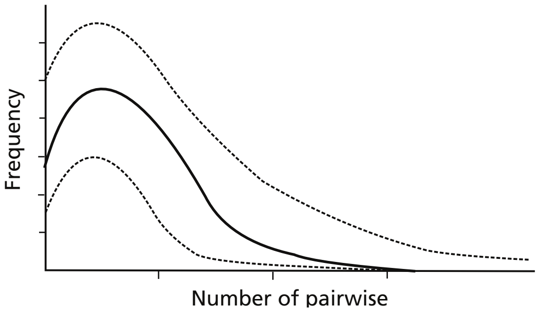
The general 'rule of thumb' of stress is

- <0.05 = an excellent representation with no prospect of misinterpretation
- <0.10 = a good ordination with no real prospect of misleading interpretation
- <0.20 = a potentially useful representation – too much reliance should not be placed on the details
- >0.30 = points are close to being arbitrarily placed in the lower-dimensional ordination space

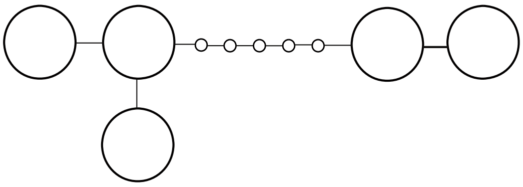
3.3.4. Historical inference

In previous sections we have discussed the importance of differentiating between historical and contemporary processes affecting the observed population structure. Historical demographic fluctuations (expansions or bottlenecks) can influence the 'neutrality' of the observed data and confound interpretations (especially from tests that assume neutrality).

Many populations have undergone significant size changes in their past. One way to test for a historical expansion is by using the mismatch distribution. A mismatch analysis is simply the frequency distribution of all nucleotide pairwise differences between all individuals in the sample. It has been shown that particular historical demographic events will leave a signature in the distribution. For example, a population that has undergone an expansion will tend to retain more newly evolved haplotypes than under stable conditions because as the population expands, the effects of genetic drift are reduced. This will provide a MSN with one or a few internal tips (usually with a relative high frequency) and many low frequency tip haplotypes resulting in what is commonly referred to as a 'star phylogeny'. Under the expansion model, the mismatch distribution from this star phylogeny should approximate a smooth Poisson curve.



Based on the mode of this distribution and the underlying mutation rate of the marker used in the study, the timing of the expansion event can be estimated. A population that has remained large and stable over time would result in a multimodal (ragged) distribution due to the process of random lineage sorting. Lineage sorting refers to the long term effect of drift within a population where many internal haplotypes have gone extinct (see below) resulting in clade structure within a single population.



This is just one of many examples of how the analysis of mtDNA in association with the concept of a molecular clock can elucidate important historical events influencing current population structure. In the example of the mismatch distribution above, tests for neutrality would reject the null hypothesis of neutral evolution. Instead of inferring that our marker is under selection, we can propose that it is actually a neutral marker but that demographic influences have influenced the observed structure.

3.3.5. Statistical inference and interpreting the data

The section on experimental design dealt with the statistical power of a test and the probability of incorrectly accepting or rejecting a hypothesis. The final content of this section concerns the pitfalls of over-interpreting our results.

Firstly, we will quickly revisit the probability of committing a Type I error. In many ecological analyses, there is a need to make multiple comparisons from the data set. It has been argued that this will result in an overall reduction in the significance level of the test (i.e. α increases). The reasoning is that some tests will return a significant result by chance that equates to committing a Type I error where the null hypothesis is rejected incorrectly (e.g. out of 100 tests, 5 may be found significant by chance). To overcome this problem, a Bonferroni Correction for multiple comparisons is usually implemented. The rule of thumb for the correction is to divide the *apriori* significance level by the number of tests performed (e.g. α=0.05 and 20 tests, then α=0.05/20 = 0.0025). This correction is usually invoked when we are estimating values within a population (e.g. HWE tests or neutrality tests). Some maintain that it is also applicable to pairwise F_{ST} although an exact test of differentiation (whole of table test) largely removes the necessity. Currently however, the application of the Bonferroni test in ecological studies is contentious.

Whether to correct for multiple tests or not raises an important point. Even though we may be confident that we have correctly accepted or rejected the hypothesis under investigation, how does this relate to the true biological processes underlying our analyses? All estimated parameters that contain variation are just that – an estimate. As such, there will always be a certain degree of error around that estimate. Secondly, we will never be 100% sure that our sample actually represents the population truthfully. For example, the absence of a haplotype from a sample does not mean that it is absent from the population, but simply we may not have sampled intensely enough to capture it. Section 12 highlighted certain aspects of gene flow estimation that places limitations on the interpretation of the estimated values (e.g. non-equilibrium, inappropriate models). Many genetic statistical tests are conservative in nature and therefore Type II errors are always possible. An acceptance of the null hypothesis does not mean it is true, but only that there was insufficient power to reject it. Additionally, a rejection of the null hypothesis does not necessarily mean that the alternative hypothesis is true. Finally mitochondrial DNA, although an ideal marker in many respects, is a single locus and as such, inherently decreases the power of any test.

Taking the factors listed above into account, there is always a risk of over-interpreting data. In a good population study, the genetic data should complement the ecological data rather than dictating the inference made from it, especially where the data is used for making management decisions.

Table 19. Commonly used programs for population genetic analysis.

Program	OS	Author	URL	Data
TFPGA	Windows	Miller M. P.	http://www.marksgeneticsoftware.net/tfpga.htm	Co-dominant, dominant, haploid
Arlequin	Windows/ Mac	Schneider S., Kueffer J. M., Roessli D., Excofier L.	http://lgb.unige.ch/arlequin/	Co-dominant, haplotypic, DNA sequence
GenePop	DOS	Raymond M., Rousset F.	http://wbiomed.curtin.edu.au/genepop/	Co-dominant
PopGene	Windows	Yeh F. C.	http://www.ualberta.ca/~fyeh/	Co-dominant, dominant, haploid
GenAlEx	Windows	Peakall R., Smouse P.	http://www.anu.edu.au/BoZo/GenAlEx/	Co-dominant, dominant, haploid
REAP	DOS	McElroy D.	http://bioweb.wku.edu/faculty/mcelroy/	Haplotypic

Table 20. The major features of the listed programs.

Feature	TFPGA*	Arlequin*	GenePop*	PopGene	GenAlEx
Diversity					
Observed heterozygosity	✓	✓		✓	
Expected heterozygosity	✓	✓		✓	
No. of alleles per locus		✓		✓	
Proportion of polymorphic loci	✓	✓		✓	
Hardy-Weinberg equilibrium	✓	✓		✓	
Population structure					
F-statistics		✓	✓	✓	✓
AMOVA		✓			✓
Homogeneity	✓		✓		
Migration		✓	✓	✓	
Linkage equilibrium					
Two locus	✓	✓	✓		
Genetic distance					
Nei's	✓	✓		✓	✓
Rogers'	✓				✓
Pairwise FST	✓	✓			✓
Clustering					
UPGMA	✓			✓	
Neutrality test		✓		✓	

* Performing exact tests for significance.

3.4. Commonly used software for data analysis

Many software programs have been developed to analyze data in relation to molecular population genetic analysis. Their easy access, implementation of sophisticated, powerful statistical techniques and user-friendly interface make them an attractive alternative to performing calculation on spreadsheets or writing a program by oneself. Although there are a number of programs available for cost-free downloading from the internet, only some of the most commonly used ones are mentioned here.

As most of software packages provide instructions about data formatting and related methodologies, this manual will not go into these aspects. Here, we provide a review on the application of some of the common software programs, so that users will have an idea of which program will be best suited for their data.

See Table 19 for the availability of some programs. The common features of the listed programs are summarised in Table 20.

Section 4

Project design

In population genetic studies, project design is probably the most critical step for several reasons. Generating molecular genetic data requires expensive chemicals and facilities, it may also involve destructive sampling of animals and therefore careful planning is recommended in order to obtain the best information at the least cost if possible. A project often goes through several steps of planning such as identification of problems to be addressed, conducting a pilot study to identify suitable markers, followed by the development of sampling strategies, collecting and preserving samples, generating and analysing data.

4.1. Hypothesis testing and identification of problems

Naturally, the design of any study will depend on the question(s) that the researcher wants to answer. The more specific the question, the more rigorous a design can be achieved. For example, the question 'does fish species X display population structure in this river system?' is fairly broad and does not impart strong guidelines for an appropriate design. Furthermore, this question can be answered without providing significant insight into the ecological processes underlying the species distribution (usually the impetus for the study in the first place). By simply changing our question to 'at what spatial scale does genetic structuring exist for species X?' or 'does migration between points A and B result in introgression of genetic material between the respective populations?' we already have some sense of the spatial sampling strategy required. The more concise the question or questions being asked, the greater the chance that an appropriate design will be formulated, and consequently we will have a greater confidence in the ecological interpretations from the resulting data.

Problems associated with biodiversity in relation to aquaculture and fisheries often can be addressed by studying genetic variation within and/or among populations. These populations under study could be either wild or farmed stocks. Population genetics can be useful in:

- Identification of reduction in genetic diversity associated with inbreeding in farmed/ domesticated stocks
- Resolving population structure
- Defining management units within species
- Detecting hybridisation
- Study genetic interaction between farmed and wild stocks in the cases of escapement or stock enhancement programs
- Comparing levels of genetic variation between a farmed stock with the wild counterpart
- Developing suitable restocking strategies

- Understanding species biology (mating patterns, dispersal and migration).

4.2. Statistical inference

Once we have our concise questions, we need to be able to test them in a statistical manner. That is, we need an experimental design that will allow us to either accept or reject our hypotheses with a certain level of probability that minimizes errors in interpreting the data. Using the question above, 'does migration between points A and B result in introgression of genetic material between the respective populations?' we can formulate specific testable hypotheses:

Null Hypothesis N_0: There is no genetic differentiation among populations A and B (i.e. panmixia)

or

N_0 : F_{ST} between A and B is not significantly different from zero

The goal of the experimental design is to provide sufficient power to confidently accept or reject the null hypothesis. The statistical power of the analyses will depend on quantitative factors such as how many sample sites, how many individuals per site, how many genetic loci were assayed and even how many base pairs of DNA were included per locus.

These factors need to be taken into account to reduce error in the statistical results. The errors that can arise fall into two categories: Type I and Type II errors. Type I or α error arises if the null hypothesis is rejected when it is actually true (e.g. inferring structure when there is none). The level of Type I error that we are willing to accept in many biological/ecological analyses is commonly 5% (i.e. $\alpha = 0.05$). That is, we are willing to accept that the null hypothesis will be incorrectly rejected 5% of the time.

Type II or β error occurs when we fail to reject the null hypothesis when it is actually false. Although the probability of committing a Type I error is the pre-specified significance level (α), the probability value of committing a Type II error is unspecified and generally unknown. However the power of the test can be defined as $1 - \beta$ (the probability of rejecting the null hypothesis when it is truly false). Even though it is difficult to quantify β, the relationship between the two error types is known. For a given sample size the value of α is inversely related to β. The lower the chance of committing a Type I error, the higher the chance of committing a Type II error. The only way to reduce these error rates simultaneously is to increase the sample size. In other words, the greater the sample size the less likelihood there is of making incorrect conclusions. This fact needs to be taken into account when designing the study. As will be seen in later sections, the capacity for making errors (especially Type II) in population genetic studies is considerable.

4.3. Pilot study

Population genetic studies often involve large sample sizes and are usually expensive; as such a small-scale pilot study is desirable to identify suitable markers before large-scale screening of genetic variation. Although there is no exact rule of thumb as to how many individuals or populations that should be used in the pilot study, it is recommended that about 10-20 individuals of at least two populations (originally from different geographic areas) be used to screen for polymorphisms. A reasonable sample size of 10-20 individuals for each population will increase the possibility of detecting intra-population variation, while small sample sizes (2-5 individuals per population) may be enough to provide an idea on inter-population variation, depending on the variability of the markers.

It is recommended that one should consult colleagues and the relevant literature before conducting a pilot study. There may be some established markers and methods that are available which should be adopted rather than repeating the same development process, when lots of loci need to be screened, and much more resources are needed.

Choice of markers depends on several factors. The first factor is the status of species under study. For example, allozyme electrophoresis may not be suitable for studying species that are rare or endangered if destructive sampling is required, although many allozymes can be screened using body tissues such as body slime or finclips (Mather and Ruscoe, 1992). The second factor is the availability of funding; sequencing is rather expensive and therefore should be used when the budget allows, otherwise only small number of individuals are sequenced, then the results can be used to develop new primers for SSCP or restriction enzyme for RFLP.

There are no clear answers of how many loci should be screened. The rule of thumb is that more the better. However, more often than not that research budgets are limited and therefore number of loci employed also limited. Again, one should consult colleagues and literature on relevant species to understand which loci are likely to be variable and the level of that variation. This and the question being asked will help determine an approximate number of loci. For example, the question, “Is there population sub-structuring in this sample?” may require analyses of many loci, whereas the question, “Is there more than one species in this sample?” would require analysis of only one or a few diagnostic loci.

4.4. How many samples?

The discussion above clearly demonstrates the need for maximising numbers to gain credible results. A perfect experimental design would require that every individual was sampled and that every bit of DNA in each individual was sequenced. Although this would totally eliminate error (at least that generated from the experimental design), it is clearly not possible. Constraints (financial, time, resource, logistical) exist that prevent a perfect design. Therefore, it is desirable to design a study that fits within these constraints but still provides statistically powerful results.

Unfortunately, there are few absolute benchmarks with which to decide on sample size. The number of sample sites required is reflected by the question that you are asking. Ultimately, we assume that our spatial sampling is representative of the entire system under investigation and that the scale is fine enough to detect the effects of population processes.

The number of individuals sampled per site should be representative of the greater population or deme from which the sample is taken. That is, the allelic frequencies in the sample should be the allelic frequencies in the real population. There are specific calculations to determine relevant samples sizes needed to obtain a certain statistical power (see table opposite) but this is dependent on the relative allelic frequencies in the population. But it is clear that high power to detect small differences in allele frequency requires large sample sizes. As allele frequencies in the study populations are generally unknown prior to the sampling, these values cannot be implemented in the experimental design. Therefore, maximising the number of individuals within constraints is advisable. A general rule of thumb for required sample size where $\alpha = 0.05$ and $\beta = 0.50$ (i.e. a power of 0.5) is

$$2n = 1/F_{ST}$$

so that to detect an F_{ST} value of 0.01 you would require only 50 individuals (Slatkin and Barton 1989). Once again this is a posthoc test and is not really usable in the design stage (unless you can establish the desired level of differentiation prior to sampling) but rather a test to determine whether sampling was adequately undertaken.

The number of diploid individuals in each of two samples required to detect a given difference in allele frequency (Δp) given the actual frequencies of the allele in the population (p) and the power desired (1-β).

When using mtDNA a good rule of thumb to use is a sample size of 30 individuals. It has been shown that a sample size of 30 will provide a 95% chance of detecting all haplotypes that exist in the population at a frequency of 0.10 or more.

The current philosophy of achieving high statistical power in population genetic studies is through increasing the number of loci assayed rather than increasing the number of individuals per sampled population. However a balance is required here as lab work tends to be more costly and labour intensive. It should be noted that mtDNA is a single circular molecule with no recombination (i.e. a single locus), so that increasing the number

		p				
Power	Δp	0.55	0.70	0.80	0.90	0.95
50%	0.05	760	645	492	276	146
	0.10	190	162	123	69	50
	0.20	48	40	31	25	50
	0.50	6	9	13	25	50
80%	0.05	1154	1319	1006	564	299
	0.10	389	332	252	141	76
	0.20	99	82	64	27	50
	0.50	16	14	13	25	50
90%	0.05	2081	1766	1345	756	400
	0.10	520	444	337	189	102
	0.20	132	110	85	50	50
	0.50	22	20	14	25	50

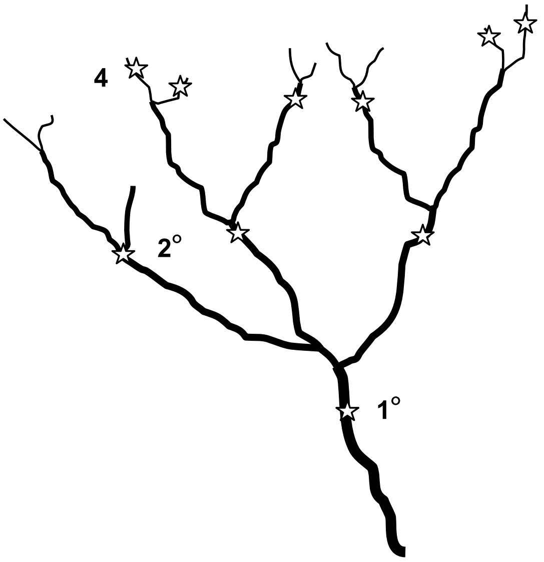
of mitochondrial genes assayed (and subsequently the number of nucleotide bases) does not significantly increase the statistical power to detect genetic differentiation.

4.5. Spatial sampling in freshwater systems

The first consideration when developing a spatial sampling design requires some knowledge of the target species' ecology. Even if a large sample is taken at a site (e.g. >100 individuals) there is a real possibility that all individuals may belong to one or a few families. As such the allelic frequencies in the sample may not reflect that of the population that they are drawn from. This problem occurs repeatedly in freshwater population studies, especially in species that school or where related offspring tend to cluster due to low dispersal and large samples can be obtained in a single scoop of a net. Because of this, samples are often incorrectly referred to as populations, which can be misleading when the analyses are interpreted.

Therefore, the spatial design at which samples are taken needs to be at a scale that will be sensitive to the underlying population processes influencing the degree of genetic differentiation. The best way to achieve this goal in freshwater systems is via a hierarchical sampling scheme. This involves taking representative samples at different spatial scales within the river system starting at the lowest spatial scale (e.g. between pools within a stream). Naturally there are countless pools in the headwaters of a river drainage and it is not possible to sample all of them. Take a subset of pools within the drainage so that you can get an estimate of the 'among pool' variation. Then move to the next hierarchical level, for example between headwater streams and take a subset of samples in order to estimate the 'among stream' variation. Continue in this way until the whole catchment is sampled (see Figure 15 below). This design can further be expanded to investigate genetic differentiation among river drainages. The same design in each drainage would be ideal, but if we already have an estimate of finer scale structuring from one drainage, then only a few samples spread around the second drainage are necessary to estimate the 'among drainage' component of genetic differentiation.

Figure 15. A stylised depiction of a hierarchical sampling scheme.



Should you be aware of specific barriers to dispersal, such as waterfalls, then more intense sampling in these areas may be warranted. It is often a good idea to proceed cautiously as the number of samples in this scheme will quickly multiply. A pilot study may be prudent in order to reduce the overall cost of the exercise. For example, if you investigate the level of differentiation at the largest spatial scale, and you find that there is no significance, is there any point in continuing on to smaller spatial scales? It is strongly suggested that if this is the case however, that a reduced sample from finer spatial scales is assayed. This is because factors other than population processes (i.e. gene flow and drift) may have influenced population structure (e.g. historical drainage rearrangement).

4.6. Temporal sampling

For specific questions, sampling schemes may be required to be more rigorous over a temporal scale rather than a spatial scale. For example, the question may be 'is the exploitation of a fishery reducing genetic variation?' In this case we are interested in the partitioning of genetic variation among sampling times rather than among sites. In these instances, it is important to maintain the same experimental design among sampling times (i.e. sample size, genetic markers used, statistical analyses employed) to be able to make valid conclusions concerning temporal fluctuations in genetic variation.

Another consideration is that of the timing of the sampling with respect to the life history of the species under investigation. Depending on the time of year, you may be inadvertently sampling non-breeding individuals. Only some knowledge of the ecology of the species will allow representative sampling of the breeding population.

4.7. Which DNA Marker?

Intuitively we may expect that the more alleles (haplotypes) present for a particular marker, the greater the statistical power to detect genetic differentiation among populations. However, this is not always the case. Remember that F_{ST} analysis partitions genetic variation within and among populations. The more variation partitioned within populations, the less there is remaining to be partitioned among populations (i.e. reducing F_{ST}). So a highly variable marker may indicate panmixia when in fact gene flow is highly restricted (Type II error).

There are no right or wrong markers to use in population studies, but some general rules can be applied. The level of genetic variation within a population is a function of both population size (large population - low effect of genetic drift) and the mutation rate (generating variability). So for large populations, it would be wise to choose a more slowly evolving marker (e.g. an rRNA gene). Conversely, in small populations the fast evolving control region may be more appropriate. If we are unsure of the possible size of the population, then a protein coding gene with a moderate mutation rate would be a good place to start. Although a certain degree of trial and error will be required, many studies already exist that may provide insight into the best marker to use.

4.8. Field trip planning, tissue collection and storage

Preparation for collecting tissues in the field is a very important step. Often field trips are expensive and labour intensive, and as such careful planning would help to reduce cost, time and energy. Field trip preparation should take into account characteristics of species to be collected, and the environment where they live. Required equipments and facilities such as boat, nets, waders, dissecting kits, liquid nitrogen containers etc. should be well planned for. Also, if there is more than one on going project at a time, consider the possibility of combining the sampling for more than one species in one trip to reduce travel cost.

It is also important to plan for storage packages. Field trips often conducted in the remote areas where a market is not available to purchase plastic bags, aluminum foil or tubes. If it is planned to collect whole specimens of fish, plastic bags may be convenient. This is mostly related to protein-based techniques in which tissues from different organs are required. In case of DNA-related work, small amount of tissue is sufficient and therefore small vials and good sealing cap will be fine. Sample labeling is extremely important, as samples will become useless without correct labels. It is advisable to use waterproof and non-smearing ink permanent marker pens to label the samples. Also, try to test the stability of ink in different environment such as freezing in ultracold freezers, liquid nitrogen etc. before taking to the field.

Care should be taken when dissecting tissues from animals. It may be good to apply anesthetic drugs to immobilize the animal before dissection. It is also important to note that in some countries, a permit is needed before the field sampling. Handling animal should follow international and national regulations. Destructive sampling should be avoided as much as possible.

We may repeat here that for projects involving allozyme electrophoresis, whole specimens may be needed to obtain tissues from different organs such as liver, muscle, eyes, heart are required. However, for DNA-related work then only small amounts of tissue are required, for fish for example, a small piece (one square centimeter) of finclip or a few scales should be sufficient, or for crustacean species, a leg may be enough. In the latter case, the animal can be released back into their original place (rivers, ponds, tanks). It is also noted that an extra amount of tissue may be useful, as some projects may require more than one marker and therefore more tissue will be required.

Storage conditions in the field depend largely on the techniques to be applied in the laboratory. If tissue samples are going to be used for allozyme electrophoresis, specimen should be either kept alive, or frozen in liquid nitrogen or dry ice before transferring to laboratory. In laboratory, specimen these should be stored in ultra-cold freezers, at -80°C as some enzyme may deteriorate at higher temperatures. If tissues are for DNA analysis, then they can be frozen (-20°C) or preserved in ethanol 70% or higher concentration at room temperature. Ethanol may evaporate overtime, and therefore care should be taken to ensure there is sufficient ethanol to cover the tissue. Fish scales can be kept dry at room temperature; this is perhaps most convenient in case samples are requested from elsewhere, where ethanol may be a constraint as most courier companies do not accept consignments with ethanol.

Conclusion

It is important for us to reiterate and recapitulate what we endeavoured to do and why we did what we did in the foregoing sections, and what their basic applications are in modern aquaculture and inland fisheries management and conservation. Prior to the application of molecular genetic analysis, often there were no techniques available to ascertain what constitutes a good hatchery broodstock, how much genetic variation in a potentially important species for aquaculture, how genetics in a population changes over time, what genetic effects result from escapement/stock enhancement/ restocking programs. We have learnt to address these questions by characterising genetic variation in population(s) under consideration.

Genetic characterisation either individually or collectively give you a quantitative measure of the genetic diversity of the fish you are dealing with. This information is very basic and pivotal to maintaining diversity. These parameters provide an indication of the genetic "healthiness" of your broodstock, enable you to build up a suitable broodstock from the very beginning, and/or provide clues to developing an appropriate stock enhancement program, and/or develop suitable conservation measures recognising the management units with a known genetic identity.

Needless to say the study of population genetics is founded in mathematics. However, our lives have been made easier in the present times in that most of the mathematical analyses are incorporated into menu driven software packages, which are extremely user friendly. However, even though you may not go deep into the mathematical equations it is imperative that you understand the manner in which you have to interpret the final results. The interpretations of course will also be very much dependent, as had been pointed out earlier, on the objectives you set out to address initially. As such, the most important aspect will be to plan your study - recognize the needs and therefore the objectives and then set out to complete it effectively.

References

Amstrong, J., A. Gibbs, R. Peakall, and G. Weiller. (1995). RAPDistance programs; Version 1.03 for the analysis of patterns of RAPD fragments.

Carlson, J. E., L. K. Tulsieram, J. C. Glaubitz, V. W. K. Luk, C. Kauffeldt, and R. Rutledge. (1991). Segregation of random amplified DNA markers in F1 progeny of conifers. *Theoretical and Applied Genetics* 83: 194-200.

Clark, A. G., and C. M. S. Lanigan. (1993). Prospects for estimating nucleotide diveregence with RAPDs. *Molecular Biology and Evolution* 10: 1096-1111.

Erlich, H. A., editor. (1989). *PCR Technology: Principles and Applications for DNA Amplification*. Stockten Press, New York, USA.

Fetzner, J. W. J., and K. A. Crandall. (2001). Genetic Variation. In D. M. Holdich, editor. *Biology of Freshwater Crayfish*. Blackwell Science, Oxford.

FitzSimmons, N. N., C. Moritz, and S. S. Moore. (1995). Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. *Molecular Biology and Evolution* 12: 432-440.

Fu, Y. X. (1996). New statisical test of neutrality for DNA samples from a population. *Genetics* 143.

Fu, Y. X., and W. H. Li. (1993). Statistical test of neutrality of mutations. *Genetics* 133: 693-709.

Garcia, D. K., and J. A. H. Benzie. (1995). RAPD markers of potential use in penaeid prawn (*Penaeus monodon*) breeding programs. *Aquaculture* 130: 137-144.

Guo, S. W., and E. A. Thompson. (1992). Performing the exact test of Hardy-Weinberg proporttion for multiple alleles. *Biometrics* 48.

Hadrys, H., M. Balick, and B. Schierwater. (1992). Amplifications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Molecular Ecology* 1: 55-63.

Hillis, D. M., C. Moritz, and B. K. Mable, editors. (1996). *Molecular Systematics*. Sunderland, Sinauer Associates.

Hudson, R. R., M. Kreitman, and M. Aguadé. (1997). A test of neutral molecualr evolution based on nucleotide data. *Genetics* 116: 153-159.

Karl, S. A., and J. C. Avise. (1992). Balancing selection at allozyme loci in oysters: implications from nuclear RFLP's. *Science* 256: 100-101.

Landsmann, R. A., R. O. Shade, J. F. Shapira, and J. C. Avise. (1981). The use of restriction endonulease to measure mitochondrial DNA sequence relatedness in natural population. III Techniques and potential applications. *Journal of Molecular Evolution* 17: 214-226.

Lewis, P. O., and A. A. Snow. (1992). Deterministic paternity exclusion using RAPD markers. *Molecular Ecology* 1: 155-160.

Lynch, M., and E. G. Milligan. (1994). Analysis of population genetic structure with RAPD makers. *Molecular Ecology* 3: 91-99.

Mather, P.B., Ruscoe, W.A. (1992). Use of cellulose acetate electrophoresis and non-destructive sampling procedures for identification of potential gene markers. *The Progressive Fish-Culturist* 54: 246-249.

Mueller, U. G., and L. L. Wolfenbarger. (1999). AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution* 14: 389-393.

Na-Nakorn, U., Sukmanomon, S., Nakajima, M., Taniguchi, N., Kamonrat, W., Poompuang, S., Nguyen T. (2006). MtDNA diversity of the critically endangered Mekong River giant catfish (*Pangasianodon gigas*) and closely related species: Implications for conservation. *Animal Conservation* 9: 483-394.

Nei, M. (1972). Genetic distance between populations. *The American Naturalist* 106: 283-292.

Nei, M. (1973). The theory and estimation of genetic distance in N. E. Morton, editor. *Genetic Structure of Populations*. University Press of Hawaii, Honolulu.

Nei, M. (1982). Evolution of human races at gene level in B. Bonne-Tamir, editor. *Human Genetics, Part A: the unfolding Genome*. Columbia University Press, New York.

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbis University Press, New York.

Nei, M., and J. C. Miller. (1990). A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. *Genetics* 125: 873-879.

Nei, M., and F. Tajima. (1983). Maximum likelihood estimation of the number of nucleotide subsitutions from restriction data. *Genetics* 105: 207-217.

Parker, P. G., A. A. Snow, M. D. Schug, G. C. Booton, and P. A. Fuerst. (1998). What molecules can tell us about populations: choosing and using a molecular marker. *Ecology* 79: 361-382.

Reynolds, J. (1981). *Genetic Distance and Coancestry*. North Carolina State University, Raleigh, NC, USA.

Richardson, B. J., P. R. Baverstock, and M. Adams (1986). *Allozyme Electrophoresis: A Handbook for Animal Systematics and Population Studies*. Academic Press, Australia.

Riedy, M. E., W. J. Hamilton, and C. F. Aquadro. (1992). Excess of non-parental bands in offspring from known pedigree assayed using RAPD PCR. *Nucleic Acids Research* 20: 918.

Ruzzante, D. E., C. T. Taggart, and D. Cook. (1996). Spatial and temporal variation in the genetic composition of larval cod (*Gadus morhua*) aggregation: cohort contribution and genetic stability. *Canadian Journal of Fisheries and Aquatic Sciences*.

Ryman, N., and F. Utter (1987). *Population Genetics & Fishery Management*. University of Washington, Washington, USA.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239: 487-491.

Scott, M. P., K. M. Haymes, and S. M. Williams. (1993). Parentage analysis using RAPD PCR. *Nucleic Acids Research* 20: 5493.

Swofford, D. L., G. J. Olsen, P. J. Wadell, and D. M. Hillis. (1996). Phylogenetic Inference in D. M. Hillis, C. Moritz, and B. K. Mable, editors. *Molecular Systematics*. Sinaeur Associates, Inc., Sunderland, MA.

Tajima, F. (1989). Statiscal method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

Tegelström, H. (1992). Detection of DNA fragments. Pages 89-113 in A. R. Hoelzel, editor. *Molecular Genetic Analysis of Populations: A Practical approach*. Oxford University Press, Oxford, UK.

Vos, P., R. Hogers, M. Bleeler, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23: 4407-4414.

Wigginton, J. E., D. J. Cutler, and G. R. Abecasis. (2005). A note on exact test of Hardy-Weinberg equilibrium. *American Journal of Human Genetics* 76: 887-893.

Williams, J. G. K., A. R. Kubelik, K. J. Livak, J. A. Rafalski, and S. V. Tingey. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acid Research* 18: 6531-6535.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics* 15: 323-354.

Wright, S. (1978). *Evolution and the Genetics of Populations. IV. Variability within and among Natural Populations*. University of Chicago Press, Chicago.

Wright, S., and P. Bentzen. (1994). Microsatellites: genetic markers for the future. *Reviews in Fish Biology and Fisheries* 4: 384-388.

Xu, J., A. Utrner, J. Little, E. A. Bleeler, and D. A. Meyers. (2002). Positive results in asscociation studies with departure from Hardy-Weinberg eliqubrium: hint for genotyping error? *Human Genetics* 111: 573-574.

Xu, Z., A. K. Dhar, J. Wyrzykowski, and A. Alcivar-Warren. (1999). Identification of abundant and informative microsatellites for shrimp (*Penaeus monodon*) genome. *Animal Genetics* 30:150-156.

Annex 1: List of commonly used buffers in allozyme electrophoresis

	Amount per gel (ml)†		Electrode	Running Conditions		
Buffer	Stock*	H_2O	(Stock:H_2O)	Current/Gel	Voltage	Time
TC6	6.2	440	1 : 4	45-55ma	~200V	4hr
TC8	14	440	1 : 4	45-55	~150	6
TEB	23.4	440	1 : 5	45-55	~300	4
TG	50	400	1 : 4	45-55	~300	4
TM	9	440	1 : 4	75	~100	8-10
				120(+ice)	<200	5-6
TG	50	450	1 : 9	75-80	~450	4-5
LiOH	48	425	1 : 4	**70(+ice)	~400	4
Poulik	45	400	1 : 4	**70(+ice)	~300	4

† for 200 mm x 200 mm x 12 mm gel (suitable for 4 slices).

* See Annex 2.

** The current for discountinuous buffers will drop during the run; the voltage may be increased to compensate.

Annex 2: Stock solutions of buffers used in allozyme electrophoresis

TC6

Per 1 litre final volume:

- 66.6g Tris (Sigma 7-9)
- 39.5g Citric acid (monohydrate)
- pH6; Store in refrigerator

TEB

Per 1 litre final volume:

- 109g Tris
- 30.9g Boric acid
- 7.4g Disodium EDTA
- Store at room temperature

TG

- 30g Tris
- 144g Glycine
- Make up to 1 litre.
- Store at room temperature

LiOH gel buffer

Per 1 litre final volume:

- 27.2g Tris
- 7.5g Citric acid
- 200 ml Electrode stock buffer
- Store in refrigerator

Poulik gel buffer

- Per 1 litre final volume
- 92.1g Tris
- 10.5g Citric acid
- Store in refrigerator

TC8

Per 1 litre final volume:

- 104g Tris
- 39.5g Citric acid
- pH8
- Store in refrigerator

TM

Per 1 litre final volume:

- 60.5g Tris
- 58g Maleic acid
- 18.6g EDTA
- 10g $MgCl_2$
- 25g NaOH
- Adjust to pH7.4 with NaOH Store at room temperature

LiOH electrode buffer

Per 1 litre final volume:

- 6.3g LiOH*
- 59.4g Boric acid
- Store at room temperature

*Carefully - the dust is irritating to breath

Poulik electrode buffer

Per 1 litre final volume:

- 92.8g Boric acid
- 12g NaOH

Annex 3: Staining recipes for common enzymes

Each of the assay recipes which follow are in amounts suitable for staining two gels; the width of the gel is about 10mm, and 16mm long (i.e. 2 x 8cm). Standard procedures for all assays are given below.

Assay buffers:

Tris-HCl pH 8

- 70 ml 1M *Tris*
- 30 ml 1M HCl
- make to 1000 ml with H_2O

Phosphate pH 7

- 100 ml 1M Na_2HPO_4 (14.2 g)
- 98 ml 1M NaH_2PO_4 (14.3 g)
- make to 1000 ml with H_2O

Agar overlay recipes

Heat 2% agar suspension in water in a flask in beaker of boiling water until agar solution is clear. This solution can be prepared in advance, and kept in a stoppered flask in the 60˚C oven.

Mix contents of enzyme assay recipe in 25 ml buffer (including any substrate solutions).

Add 25 ml of the melted agar, mix and pour onto gel slices.

Standard recipe amounts

- MTT: 4 mg
- PMS: About 0.2 mg - tiny amount on tip of spatula.
- **Add PMS last,** just before agar. All assays with PMS must be incubated in the dark.
- NAD: 3 mg
- NADP: 2.5 mg
- G6PD: 10 µl of a 1000 units/ml solution.
- **Add G6PD just before adding PMS.**

Acid phosphatase (Acph)

- α-napthyl acid phosphate 50 mg
- 0.1 M Na acetate buffer pH5 100 ml
- Incubate gel in substrate solution 30 minutes at 37˚ C then add 25 mg
- Fast black K in 20 ml H_2O.

Adenosine deaminase (Ada)

- Adenosine 30 mg
- MTT
- Xanthine oxidase 1 unit

- Nucleoside phosphorylase 5 units add just before PMS
- PMS
- Phosphate buffer 25 ml
- Agar overlay

Adenylate kinase (Ak)

- ADP 10 mg
- Glucose 20 mg
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD
- Hexokinase 10 mg
- ***T*ris-HCl** buffer 25 ml
- Agar overlay

Alcohol dehydrogenase (Adh)

- Ethanol 5 ml
- MTT
- PMS
- NAD
- **Tris-HCl** buffer 20 ml
- Agar overlay

Aldehyde oxidase (Ao)

- Benzaldehyde 1 ml (N.B. Benzaldehyde stinks; complete the assay in the fume hood with the extractor fan on. The enzyme displays high levels of activity and will stain at room temprature (ie. leave it in the fume hood!).
- MTT
- PMS
- NAD
- **Tris-HCl** buffer 25 ml
- Agar overlay

Alkaline phosphatase (Alph)

- β-napthyl acid phosphate 50 mg
- TEB electrode buffer 100 ml
- Incubate gel in substrate solution 30 minutes at 37°C, then add 25 mg fast black K in 20 ml H_2O.

ARGININE PHOSPHOKINASE (Apk)

- Phosphoarginine 10 mg
- ADP 10 mg
- Glucose 100 mg
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD
- Hexokinase 10 mg
- **Tris-HCl** buffer 25 ml
- Agar overlay

Catalase (Cat)

- **Solution A**:
  - H_2O_2 1 ml of 28% solution
  - Phosphate buffer 15 ml
  - H_2O 100 ml

- Incubate in solution A for 30 minutes at room temperature. Bubbles should appear, indicating catalase activity.

- Rinse with distilled H_2O (**take care to wear gloves** and hold gel at top corners).

- **Solution B**:
  - KI 30 ml of 0.2 M solution HCl
  - 3 ml of 1 M solution
  - H_2O 70 ml

- Score within 2 min. Catalase activity is indicated by white bands on dark blue background.

Creatine kinsase (Ck)

- Phosphocreatine 50 mg (try more if that doesn't work)
- ADP 10 mg
- Glucose 20 mg
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD
- Hexokinase 10 mg
- **Tris-HCl** buffer 25 ml
- Agar Overlay

Esterase (Est)

- Substrate solution 2 ml (substrate solution contains : β-napthyl acetate 1 g & acetone 100 ml)
- **Phosphate** buffer 100 ml
- Incubate gels in substrate solution for 20 minutes at 37˚C, then add 50 mg fast blue BB in 20 ml H_2O. If enzyme is very active, add the dye directly to the substrate solution.

Glucose-6-phosphate dehydrogenase (G6pd)

- Glucose-6-phosphate 40 mg
- MTT
- PMS
- NADP
- MgCl
- **Tris-HCl** buffer 25 ml
- Agar Overlay

Glutamate-oxaloacetate transaminase (Got or AAT)

- Substrate solution 25 ml:
  - α-ketoglutaric acid 365 mg
  - L-aspartic acid 1331 mg
  - EDTA 0.5 g
  - Na_2HPO_4 14.2 g
  - H_2O to 500 ml (should be pH 7.4)
- H_2O 25 ml
- Incubate gells in substrate solution for 30 minutes at 37˚C, then add 50 mg fast blue BB in 20 ml H_2O.

Glutamate dehydrogenase (Gdh)

- 0.1 M sodium glutamate 10 ml
- MTT
- PMS
- NAD (use NADP in some species)
- **Tris-HCl** buffer 15 ml
- Agar overlay

α-glycerophosphate dehydrogenase (α-Gpd)

- 0.1 M α-glycerophosphate 10 ml
- MTT
- PMS
- NAD
- **Phosphate** buffer 15 ml
- Agar overlay

Guanine deaminase (Gda)

- Guanine 30 mg
- MTT
- PMS
- Xanthine oxidase 20 μl (1 unit)
- Nucleoside phosphorylase 10 μl (5 min.)
- **Phosphate** buffer 25 ml
- Agar overlay

Hexokinase (Hk)

- Glucose 2 g
- ATP 25 mg
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD

- **Tris-HCl** buffer 25 ml
- Agar overlay

Isocitrate dehydrogenase (Idh)

- Na_3 isocitrate 40 mg
- MTT
- PMS
- NADP
- MgCl
- **Tris-HCl** buffer 25 ml
- Agar overlay

Lactate dehydrogenase (Ldh)

- Substrate solution 10 ml (5 ml for vertebrates):
  - DL lactic acid (85% solu.) 10.6 ml
  - 1 M Na_2CO_3/water 49 ml
  - Water to 100 ml
  - Adjust to pH 7 with 0.5 M Na_2CO_3
  - MTT
  - PMS
  - NAD
  - **Tris-HCl** buffer 15 ml (20 ml for vertebrates)
  - Agar overlay

Leucine aminopeptidase (Lap)

- L-leucyl-β-napthylamide 40 mg
- dissolved in 1 ml dimethyl formamide (or 1 or 2 drops of acetone)
- **Phosphate** buffer 100 ml
- Incubate gel in substrate solution 30 minutes at 37˚C, then add 25 mg fast black K in 20 ml H_2O

Malate dehydrogenase (Mdh)

- Substrate solution 10 ml:
  - L-malic acid 13.4 g
  - 2 M Na_2CO_3 49 ml
  - Water to 100 ml
  - Adjust to pH 7 with Na_2CO_3
  - MTT
  - PMS
  - NAD
  - **Tris-HCl** buffer 15 ml
  - Agar overlay

Malic enzyme (Me)

- As for Mdh, but substitute NADP for NAD

Mannose-6-phosphate isomerase (Mpi)

- Mannose-6-Phosphate 20 mg
- MTT
- PMS
- NAD
- G6PD
- Phosphoglucose isomerase 20 µl (20 units) add just before PMS
- **Tris-HCl** buffer 25 ml
- Agar overlay

Nucleoside phosphorylase (Np)

- Inosine 30 mg
- MTT
- PMS
- Xanthine oxidase 20 µl (1 unit) add just before PMS
- **Phosphate** buffer 25 ml
- Agar overlay

Peptidase (Pep)

- Peptide 20 mg. Peptides used:
  - L-leucyl-glycylglycine (LGG)
  - L-leucyl-proline (LP)
  - L-leucyl-l-tyrosine (LT)

- Dissolve LT and o-Dianisidine in
  - 2 drops of 0.1 M HCl
  - o-dianisidine 5 mg
  - L-amino acid oxidase 5 mg
  - Horseradish peroxidase 5 mg
  - **Phosphate** buffer 25 ml
  - Agar overlay

Phosphoglucomutase (Pgm)

- Glucose-1-phosphate 60 mg (try more if needed)
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD
- **Tris-HCl** buffer 25 ml
- Agar overlay

6-phosphogluconate dehydrogenase (6pgd)

- Na_3 6-phosphogluconic acid 20 mg
- NADP
- MTT
- PMS
- **Tris-HCl** buffer 25 ml
- Agar overlay

Phosphoglucose isomerase (Pgi or Gpi)

- Fructose-6-phosphate 40 mg
- MTT
- PMS
- NAD
- $MgCl_2$
- G6PD
- **Tris-HCl** buffer 25 ml
- Agar overlay

Sorbitol dehydrogenase (Sdh)

- Sorbitol 200 mg (or more if needed up to 1 gm)
- MTT
- PMS
- NAD
- **Tris-HCl** buffer 25 ml
- Agar overlay

Superoxide dismutase (Sod)

- MTT
- PMS
- **Tris-HCl** buffer 25 ml
- Agar overlay

Annex 4: Practical exercises

Co-dominant markers

Problems

There is a fish species that have been used for aquaculture in the last 20 years or so. A national broodstock center was established and the first batch of broodstock were domesticated from the wild population of a river nearby. The center artificially breeds the fish and distributes fingerlings to farmers for grow-out. Some farmers have also successfully bred the fish, recruiting broodstock from what they breed and recently observe slow growth, loss of productivity as well as high level of deformity. The environmental authority also claims there is evidence that farmed escapees have interbreed with wild fish.

Several questions are raised:

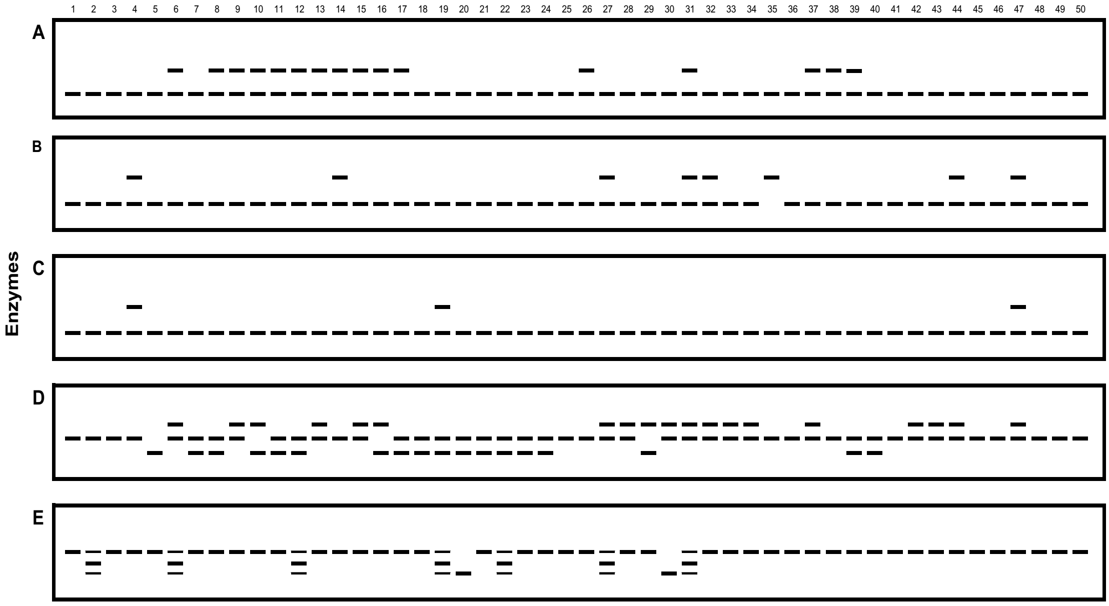
1. Is the slow growth of fish in the farm due to genetic problems? How do you find it out?

2. Assuming the problem is genetic, how do you confirm this?

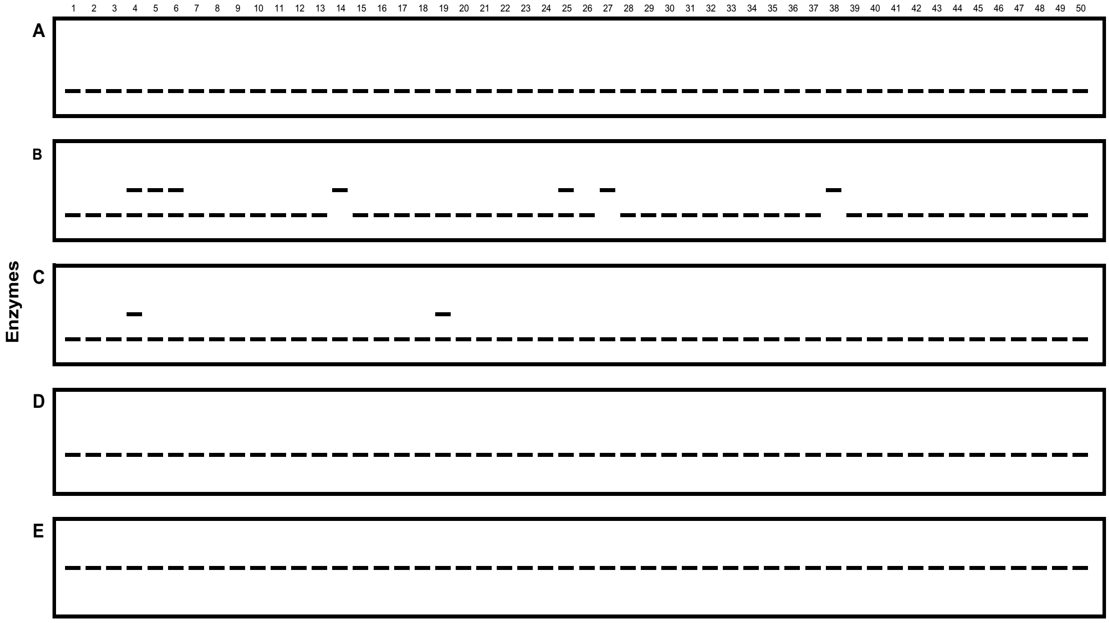
3. How do you find evidence to support or reject the environmental authority's claim?

4. Fifty individuals from each of the three populations of fish were sampled and scored for five enzyme loci as represented below. The three populations sampled were from (1) the national broodstock centre, (2) the river nearly, and (3) a farm which had originally obtained only a small number of brood fish from the national broodstock centre 20 years ago. How do you use the data to answer the above questions?

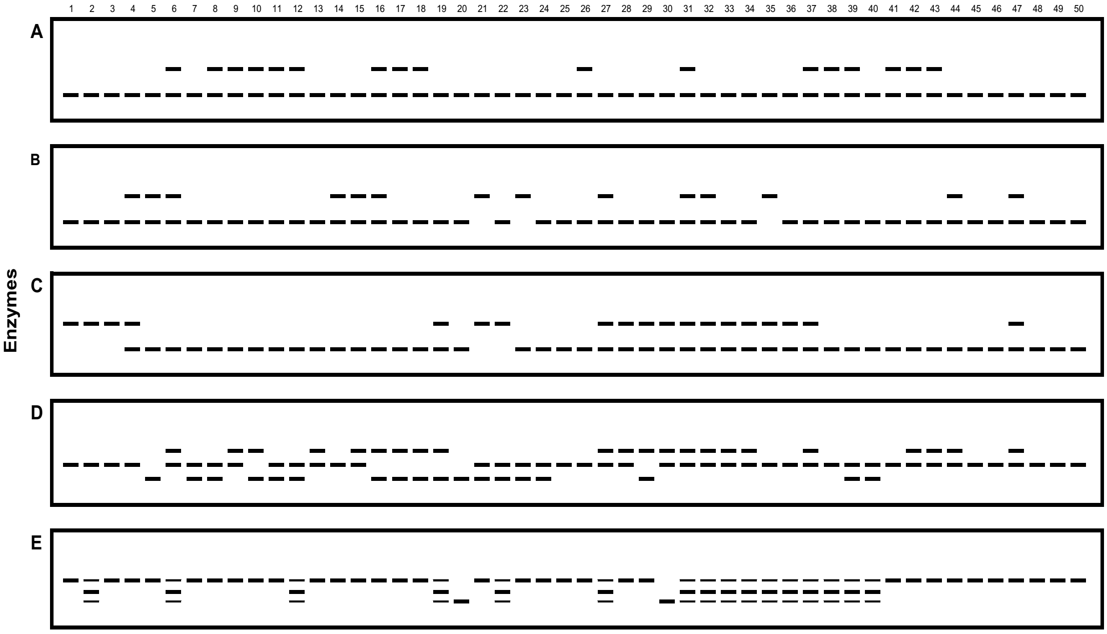
Population 1: River.



Population 2: Farmer’s broodstock.



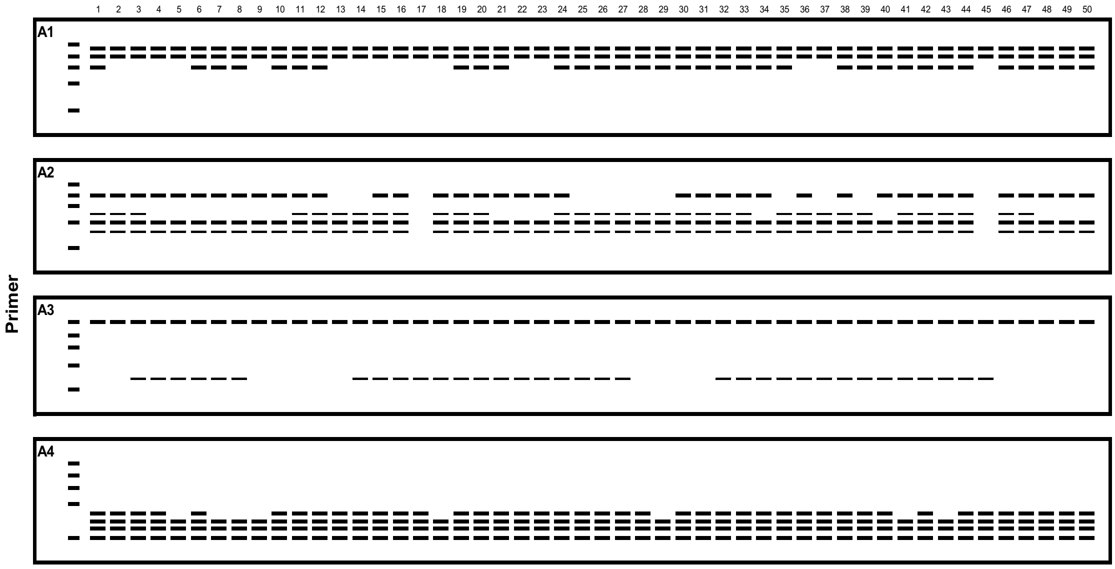
Population 3: National Broodstock Centre.



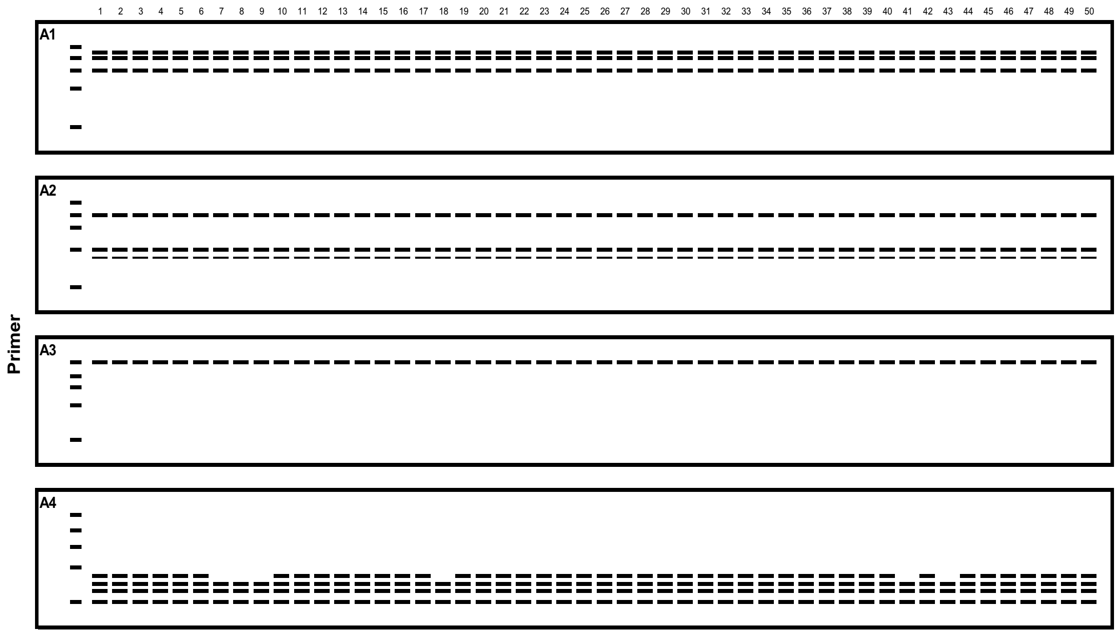
Dominant markers

The same fifty individuals from 3 populations above were analysed and scored for variation using 5 sets of RAPD primers as represented below. Would these data help to answer the same questions?

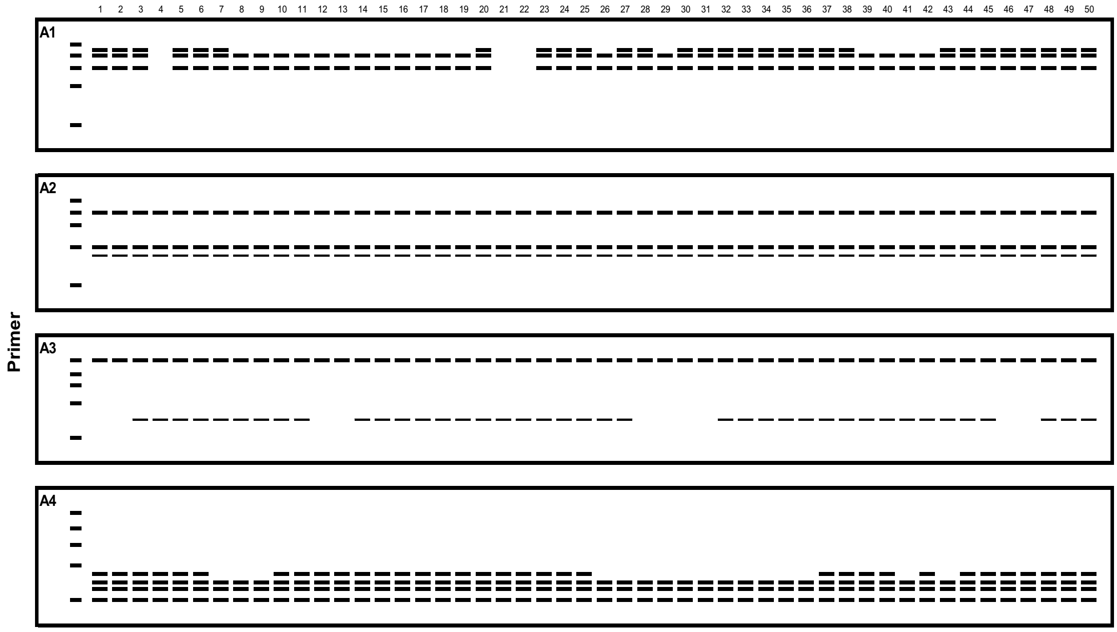
Population 1: River.



Population 2: Farmer’s boodstock.



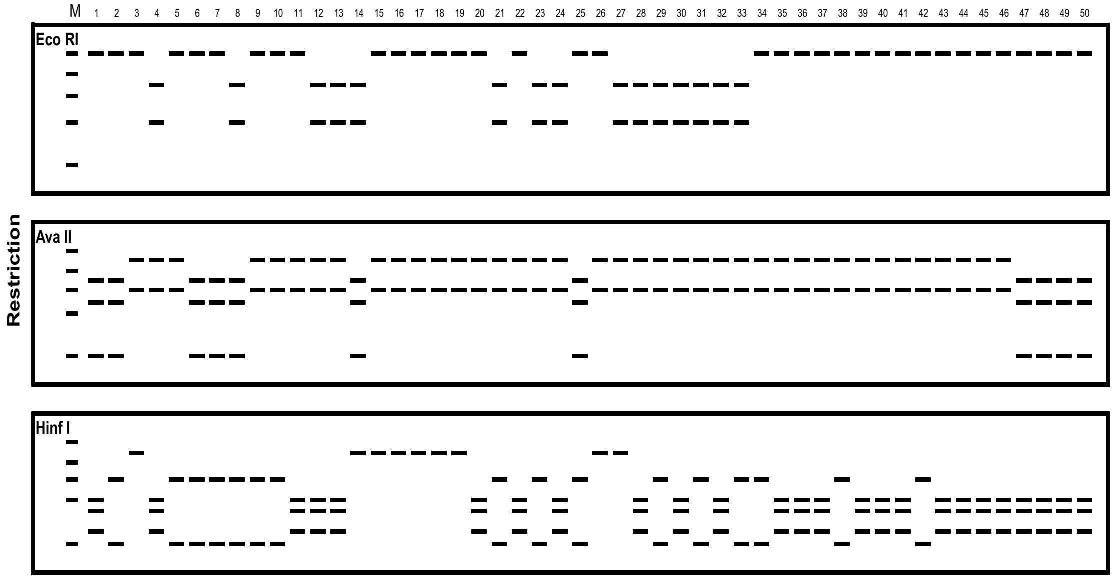
Population 3: National Broodstock Centre.



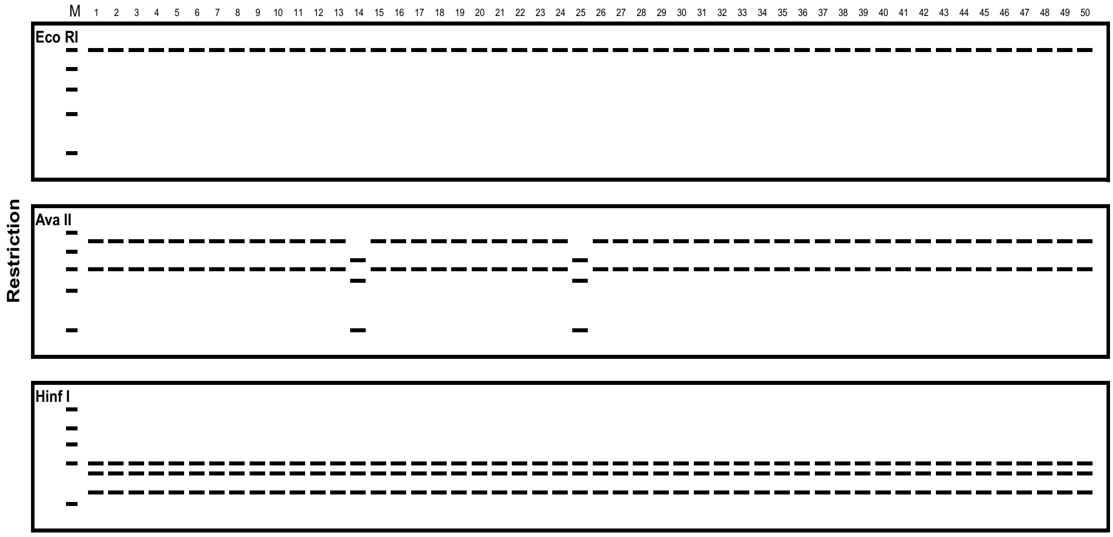
Haplotypic markers

The same individuals of fish above were used for RFLP analysis. The D-loop region of the mtDNA was amplified for all 150 individuals. PCR products were subjected to digestion using 3 restriction enzymes, namely EcoR I, Ava II annd Hinf I. After digestion, the restriction digested products were run on agarose gels and stained with ethidium bromide. The gel images were taken as below. Do these data show similar results found using allozymes and RAPDs?

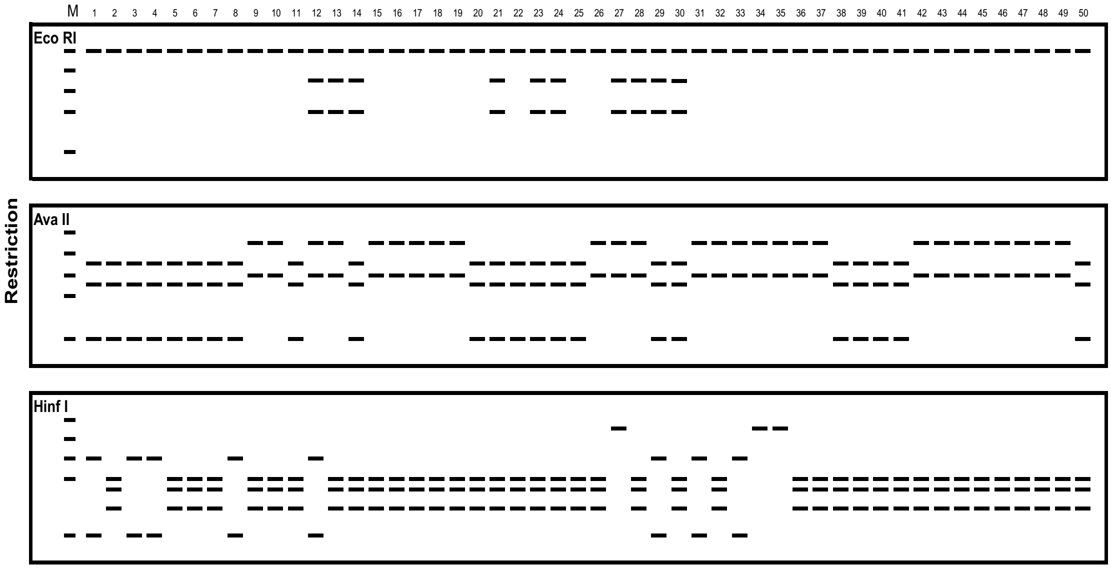
Population 1: River.


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Eco RI
Ava II
Hinf I
Restriction

Population 2: Farmer's broodstock.


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Eco RI
Ava II
Hinf I
Restriction

Population 3: National broodstock center.


M 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
Eco RI
Ava II
Hinf I
Restriction

www.enaca.org