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

## Part 1

## Conceptual basis of population genetic approaches

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

## Part 1:

## Conceptual basis of population genetic approaches

### Contributors

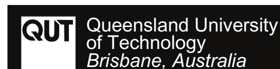
**Thuy Nguyen**
Network of Aquaculture Centres in Asia-Pacific

**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 1: Conceptual basis of population genetic approaches. NACA Monograph No. 1, 80p.

ISBN 978-974-88246-1-1

Printed by Scandmedia, Bangkok.

# Contents

Preface ... 5
Acknowledgements ... 7
Background ... 9
Target audiences ... 11
Aims, scope and format of the manual ... 12
Abbreviations ... 13

**Section 1. The fundamental nature of DNA ... 15**

**1.1 Basic DNA structure ... 17**
**1.2 Where does variation in DNA sequences come from? ... 18**

**Section 2. Genetic variation in nature ... 23**

**Section 3. Basic concepts in population genetics ... 29**

**Section 4. Natural selection ... 35**

**Section 5. Genetic drift ... 41**

**Section 6. Non-random mating and population structure ... 47**

**Section 7. Environmental influences on population processes ... 55**

**Section 8. Ecological influences on population processes ... 63**

**Glossary ... 69**
**Bibliography ... 79**

# Preface

The mandate of NACA is to support member governments in their endeavours to achieve long-term sustainability of inland fishery resource utilisation and aquaculture development. In this regard, NACA plays a major role in developing human capacity in aspects in the member countries.

In the current millennium, inland fisheries resource utilisation and aquaculture development have to go hand in hand with maintaining environmental integrity and biodiversity. Conserving biodiversity has become an important consideration worldwide. Nations that import aquaculture products, often stress that the production processes must not negatively affect natural biodiversity. Furthermore, conservation of biodiversity is an integral component of responsible fisheries and enshrined in the FAO Code of Conduct for Responsible Fisheries. Consequently, NACA as mandated by its Governing Council, is embarking on a program that attempts to sustain genetic diversity in relation to inland fisheries management and aquaculture development in the region.

One of the initial steps is to assist member nations to achieve the above broad objectives and to develop human capacity in the current methodologies used to assess genetic diversity and its applications to biodiversity issues in inland fishery resource utilisation and aquaculture development. This manual is produced to facilitate training processes that NACA will undertake in the ensuing years to enable the member nations to achieve the overall objectives in regard to maintaining biodiversity in relation to development of aquatic resources utilisation.

We accept the fact that a number of text books are available for reference in this field. Most however, are expensive for many users and some of the techniques provided in them are not always suitable for many of the molecular laboratories in the region. This has prompted us to prepare this manual, which is designed to be less expensive, more “user friendly” and of direct relevance to the region.

# Acknowledgements

T. T. T. Nguyen 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.

# Background

It has generally been accepted that aquaculture can contribute significantly to narrowing the gap between demand and supply for aquatic food supplies. Currently, aquaculture production is estimated to be 51.4 million tonnes annually, valued at US$60 billion. More importantly, developing countries, particularly in Asia, account for over 85% of current production. It is most likely that dominance of Asian countries in aquaculture production will be maintained into the foreseeable future.

With increasing developments in aquaculture however, the sector also has had to face public concern in regard to environmental effects. Aquaculture development with no regard for social and environmental issues is no longer acceptable to the public, be it in developed or developing countries. Aquaculture development needs increasingly to take into account environmental impacts. It is in this regard that maintaining and sustaining the environment has become paramount. Attention to genetic diversity and biodiversity in aquaculture development and aquatic resource management are therefore, crucial elements for sustainable environments.

Introduction of new species/strains can affect biodiversity via impacts on the native gene pool. New species/strains can hybridise with native stocks, and hence alter the natural genetic architecture. This may be expressed as a loss of valuable genetic material such as locally adapted genes or gene complexes or homogenisation of previously structured populations via flooding with exogenous genes. In Thailand, one example of such impacts is the outcome of hybridisation between the Thai walking catfish, *Clarias macrocephalus* and the African catfish *C. gariepinus* (Senanan et al., 2004). While the long-term impact of this hybridisation is still to be determined, there has been a general loss of genetic diversity in the native species. Similarly, it has been a suggested that hybrid *Clarias* are contributing to the decline of native *C. batrachus* in the Mekong Delta (Welcomme and Vidthayanon, 2003). A parallel situation appears to be occurring elsewhere in Viet Nam, but as yet no genetic analyses have been conducted (personal observation).

Stock enhancement is a common fishery practice in the freshwaters of many Asian nations, and is considered to be a means by which fish food supplies can be significantly enhanced (Petr, 1998 De Silva, 2004). Many enhancement practices, except those in China and perhaps India, are however dependent primarily on exotic species, with little understainding of their effects on genetic diversity in the native species. A limited study conducted in Thailand appears to indicate that stock enhancement together with escapees from aquaculture operations have brought

about a decrease in genetic diversity in the silver barb *Puntius gonionotus* populations (Kamonrat, 1996). Indeed, the observation itself indicates a need to step up the number of similar studies in the region to enable measures to be adopted that ensure levels of genetic diversity and biodiversity can be sustained for the long-term.

The major regional genetic program initiatives in Asia have thus far largely been confined to selective breeding programs, a much needed area of work for aquaculture development in the region. None of these programs were directly related however, to contributing to aquatic resource diversity. On the other hand, at recent regional workshops (Gupta and Acosta, 2001) in which most Asian nations were represented, ongoing and planned genetic related work was discussed and some consideration was made regarding biodiversity and conservation issues. Unfortunately, there were a very limited number of biodiversity related studies reported.

To date only a limited number of studies have addressed biodiversity issues in freshwater species in the region. These studies have raised however, important concerns regarding the potential negative impacts of aquaculture on biodiversity. Of particular concern is the ongoing practice of translocations and importation of exotic strains/species for culture. Senanan et al. (2004) and Na-Nakorn et al. (2004) have provided evidence that African catfish (*Clarias gariepinus*) genes have introgressed into native *C. macrocephalus* of wild and broodstock populations in Thailand, while Kamonrat (1996) demonstrated that a similar situation has resulted for silver barb *Puntius gonionotus*.

Another major concern is poor stock management practices in hatcheries, especially with respect to broodstock management, which may lead to losses of genetic variation in culture stocks due to genetic drift and inbreeding. Although the number of published works on this matter are limited (e. g. Eknath and Doyle, 1990) there is anecdotal evidence for genetic erosion of cultured stocks especially with regard to the major carp species.

Asian nations in the meeting of the International Network of Genetics in Aquaculture in 2000 (Gupta and Acosta, 2001) recognised that more attention needs to be paid to biodiversity and conservation issues. Thus while attention should be paid to genetic improvement of important cultured species, increasing awareness of the potential impacts of aquaculture and fisheries (and related activities) on biodiversity is also very important at this stage. There is a need to build the capacity of regional fisheries agencies in molecular genetic techniques to address this issue. Genetic diversity studies in the region should therefore focus on:

- Genetic improvement of important cultured species

- Assisting management practices in aquaculture operations, especially broodstock management
- Resolving taxonomic uncertainties, and phylogenetic relationships, especially for those species or populations that are endangered and/or commercially important
- Documenting patterns of natural genetic diversity and identifying management units
- Assessing genetic impacts of cultured stocks on indigenous stocks

In the light of the major aquaculture developments taking place in Asia, urgent attention is needed on biodiversity and genetic integrity issues of cultured as well as indigenous wild stocks; issues that are increasingly raised by the public and nations that import aquatic products. It is in this regard that there is a great need to build capacity in applied molecular genetic capabilities at the national and regional levels. This will allow characterisation of the genetic resources of relevant species important to aquaculture and inland fisheries in the respective nations/sub-region. Knowledge on the applications of molecular genetics to aquaculture and fisheries management will help reduce the negative impacts of many current activities on biodiversity, and allow development of suitable strategies for maintaining and sustaining diversity. It will also help to provide a useful guide to the identification and conservation of genetic integrity of aquatic species within the region.

## Target audiences

This manual is expected to enable NACA member country personnel to be trained to undertake molecular genetic studies in their own institutions, and as such is aimed at middle and higher level technical grades. The manual can also provide useful teaching material for specialised advanced level university courses in the region and postgraduate students.

The manual has gone through two development/improvement stages. The initial material was tested at a regional workshop and at the second stage feedback from participants was used to improve the contents.

# Aims, scope and format of the manual

The aim of this manual is to provide a comprehensive practical tool for the generation and analysis of genetic data for subsequent application in aquatic resources management in relation to genetic stock identification in inland fisheries and aquaculture.

The material only covers general background on genetics in relation to aquaculture and fisheries resource management, the techniques and relevant methods of data analysis that are commonly used to address questions relating to genetic resource characterisation and population genetic analyses. No attempt is made to include applications of genetic improvement techniques e.g. selective breeding or producing genetically modified organisms (GMOs).

The manual includes two 'stand-alone' parts:

- **Part 1 – Conceptual basis of population genetic approaches:** will provide a basic foundation on genetics in general, and concepts of population genetics. Issues on the choices of molecular markers and project design are also discussed.

- **Part 2 – Laboratory protocols, data management and analysis:** will provide step-by-step protocols of the most commonly used molecular genetic techniques utilised in population genetics and systematic studies. In addition, a brief discussion and explanation of how these data are managed and analysed is also included.

# Abbreviations

| | |
|---|---|
| **A** | Adenine |
| **AA** | Amino Acid |
| **AFLP** | Amplified fragment length polymorphism |
| **AMOVA** | Analysis of molecular variance |
| **ANOVA** | Analysis of variance |
| **C** | Cytosine |
| **DGGE** | Denaturing Gradient Gel electrophoresis |
| **DNA** | Deoxyribonucleic acid |
| **dsDNA** | Double stranded DNA |
| **G** | Guanine |
| **GD** | Genetic drift |
| **HWE** | Hardy-Weinberg Equilibrium |
| **IBD** | Isolation-by-distance or identical-by-descent |
| **kb** | 1000 nucleotide base pairs (kilobase) |
| **LHT** | Life history traits |
| **MDS** | Multidimensional scaling ordinations |
| **MHC** | Major histocompatability complex |
| **mRNA** | Messenger ribonucleic acid |
| **MSN** | Minium spanning network |
| **mtDNA** | Mitochondrial deoxyribonucleic acid |
| **MU** | Management units |
| **NCA** | Nested clade analysis |
| **nDNA** | Nuclear deoxyribonucleic acid |
| **Nm** | Effective number of migrants (where N= effective population size and m=mutation rate) |
| **NS** | Natural selection |
| **PCR** | Polymerase chain reaction |
| **RAPD** | Random amplified polymorphic DNA |
| **RE** | Restriction enzyme |
| **RNA** | Ribonucleic acid |
| **SCR** | Semi-conservative replication |
| **SSCP** | Single strand conformational polymorphism |
| **SSR** | Simple sequence repeats |
| **T** | Thymine |
| **TGGE** | Temperature gradient gel electrophoresis |
| **U** | Uracil |

# SECTION 1

## The fundamental nature of DNA

Traditional approaches in fisheries for identifying populations that should be managed separately (i.e. management units) have relied on documenting population life history traits including reproductive condition both temporally and spatially, breeding and feeding sites, population specific behaviours, and movement patterns to infer similarity or independence of gene pools. While the results often are in accord with subsequent population genetic analyses of the same populations this may not always be the case (i.e. observations of morphological similarity does not necessarily mean individuals belong to the same reproductive unit or observations of mating do not necessarily imply successful reproductive input into the population). Molecular analyses (either direct or indirect) have the capacity to directly test if morphological similarity corresponds with genetic similarity or breeding actually results in genetic exchange. This is because a large amount of essentially ecological and life history information is retained in the DNA and is expressed as variation in DNA sequences. So the basis of using Population Genetic approaches for identifying management units in fisheries is to understand the basic attributes of DNA, how it changes (evolves) and the limitations on storage of life history information in DNA sequences.

## 1.1 Basic DNA structure

DNA is a polymer and a macromolecule. It consists of three building blocks, Nitrogenous bases, a Pentose sugar (Deoxyribose in DNA and Ribose in RNA) and a Phosphate group. The three components are bound covalently and when joined are called a Nucleotide. There are four kinds of nucleotide present in any DNA strand. Essentially, the sugar and phosphate form the backbone of the molecule and the backbone is identical in all DNA and RNA molecules. The only potential difference between any two DNA or RNA molecules are the sequences of nitrogenous bases, so it is this sequence that encodes the genetic traits in an organism. There are four bases in both DNA and RNA: Thymine (T), Guanine (G), Adenine (A) and Cytosine (C) in DNA with Uracil (U) replacing (T) in RNA. For a long time the idea that all genetic diversity could be explained by sequence variation in four nucleotides was disputed because scientists could not comprehend how the diversity observable in nature could be explained by variation in only 4 bases. This was because biologists already new that there are 20 common Amino Acids (AAs) that are the building blocks of cells present in living organisms and it was difficult to see how four bases could encode the diversity of amino acids unless groups of bases were read together. The discovery of the genetic code whereby bases are read in groups of three bases (Codons) and then decoded into Amino Acids (AAs) solved this problem.

Other important aspects of DNA (RNA) structure to consider include; the Base-Pairing Rule whereby because of their chemical structure and the physical structure of the DNA molecule A binds to T (U) and C binds to G

and the fact that the bases in a DNA molecule are held on the inside of the helix and joined by a hydrogen bond. This allows for DNA replication, a necessary attribute for reproduction (both cellular and whole organism) and thus for near faithful transmission of genetic information from cell to cell and organism to organism. DNA replication is Semi-conservative (SCR) that implies that when DNA replicates the two strands separate with each old strand acting as a template for the production of a new strand that should have the reciprocal sequence to the strand that was used to generate it because replication occurs according to the base-pair rule (A - T and G – C). This is important to recognise because this attribute provides the basis for later proof-reading of new strands of DNA whereby the sequence along the new strand can be proof read by special enzymes to check to see if the correct base has been incorporated. Where an incorrect base has been incorporated in the new strand and this is detected by the repair enzyme relative to the old strand, it can be corrected. If however, a change occurs in both strands simultaneously then repair enzymes have no reference point to correct the change. The mechanism of DNA replication that occurs naturally in all cells forms the basis of a very powerful technique that was developed in the late 1970's/early 1980's called Polymerase Chain Reaction (PCR). Essentially, PCR mimics what happens naturally in the cell during DNA replication (*in vivo*) in a test tube (*in vitro*). We will discuss this in more detail later but to demonstrate the similarity, all of the ingredients placed into a test tube at the start of a PCR reaction are basically identical to what is present in the nucleus of a cell during DNA replication except we use an artificial short piece of DNA (called a Primer) to specify the sequence of DNA we wish to amplify. The other components are the same; target chromosomal DNA, a catalytic enzyme - DNA Polymerase, building blocks of new DNA strands – free nucleotides and a buffer to stabilise the reactions. By cycling the reaction repeatedly, millions of copies of the target sequence are generated so we can easily harvest it from the limited copies of other DNA sequences. So DNA replication provides us with a method for producing a specific target DNA sequence from a mix of all sequences in the cell.

## 1.2 Where does variation in DNA sequences come from?

When we compare the same DNA sequence from two individuals we may detect a different base at the identical point along the sequence. This difference is referred to as a Mutation or base-pair substitution. Mutations are the result of 'rare' errors during DNA replication but are a basic requirement for Evolution as a process of change because without mutation all DNA sequences would be identical to the first DNA sequence(s) that evolved originally. Potential for accumulating mutational change is an attribute of DNA and RNA molecules and is the basis for the differences we see in living and extinct organisms. Mutations can occur anywhere in the DNA (both in coding and non-coding

DNA) but where they occur in coding DNA they may produce changes in the AA sequence and be expressed as new phenotypes. If the mutation is present in some individuals in the population and not in others then the differential expression of the two phenotypes in a particular environment allows the environment to select the most appropriate form. The effect may be to change the relative frequency of the two different forms of the gene in the population over time. Mutation rates vary widely among DNA sequences in an organism's genome and relative mutation rate to a large extent is determined by what role (if any) a particular DNA sequence serves in the organism. So, the more important the role of an individual sequence is to the individual, the more slowly the sequence is likely to accumulate mutations and therefore evolve.

There are a number of different ways in which DNA can be modified by mutations, from simple base-pair substitutions involving individual nucleotides to changes in whole blocks of DNA, to loss or gain of a sequence or larger changes that could include loss or gain of one or more individual chromosomes or even whole chromosome sets (polyploidy). When they occur, their probability of long-term incorporation (survival) depends on their impact on Gene function. Non-coding DNA will tend to accumulate more mutations and so evolve faster than coding DNA because mutations in this type of DNA do not directly affect gene function. As a general rule, the more important the gene function, the lower the rate of mutation. The types of mutations most relevant to analyses of population structure are point substitutions e.g. GAG to GUG and deletions or insertions (Indels) of bases in a sequence e.g. GAG to GAGG.

Effects that mutations can have vary widely from no effect on the individual to death and there are no simple rules that we can apply to say what the likely impact of a particular type of mutation is going to be. The impact is determined by where they occur in the genome and what changes they produce. The simple fact is however that because mutations are random, when they occur in coding sequences they are likely to be deleterious (i.e. produce poor outcomes), simply because they are random changes to DNA. Ultimately the environment is the key however, as to whether a new mutation in coding DNA will provide better or poorer phenotypes.

Until the development of molecular technologies for examining variation in natural populations, the most common characters used to document variation were studies of external morphological phenotypes and while mutations in genes that code for morphological phenotypes can produce different outcomes, often these mutations do not survive because there can be strong functional constraints on many morphological traits. Thus morphological evolution can be a relatively slow process and populations may diverge genetically without any changes appearing in their external morphology. Many morphological traits are also polygenic,

meaning that they are the product of the combined effect of a few or more commonly many gene loci that may be expressed differently in different environments. This means that many systematic or population variation studies based solely on examination of variation in external morphological traits may underestimate the real extent of underlying genetic variation and hence population structure. Simply put, population studies based simply on morphology are unlikely to detect all of the significant population structuring that may exist in a species. Molecular systematic studies in contrast, are not limited in the same way.

Molecular markers can provide a more fundamental data set than morphology for examining relationships among populations and higher taxonomic levels. One important difference is that they are not complicated by any potential effect of the environment because they are fixed at fertilisation. If we target areas of DNA that do not encode phenotypes (i.e. non-coding DNA), these markers are usually neutral in respect to potential effects of Natural Selection (NS). Thus they should accumulate mutations at a constant rate determined by their locus specific mutation rates. What this means effectively is that the absolute number of mutations between homologous sequences in two individuals provides an absolute estimate of the time since they shared a common ancestor after allowing for the locus specific mutation rate. Where DNA sequences evolve neutrally, they allow phylogenetic relationships between individuals, populations, species etc. to be constructed without the need to consider complications of factors like transient impacts of natural selection or environmental effects on sequence divergence. Thus neutral markers can provide more fundamental information about phylogenetic relationships than can studies of morphology alone.

But molecular markers like morphological markers can also have associated problems that we need to address if we are going to use them constructively. For example, protein-based markers such as allozymes, often show low levels of polymorphism hence they may not be suitable for detecting genetic differentiation of organisms having weak population structure such as many marine organisms. Also, allozyme studies only detect a portion of the actual genetic variation because not all nucleotide changes lead to amino acid changes and not all amino acid substitutions result in electrophoretically detectable mobility differences. A major problem with DNA markers can be Homoplasy. Since there are only four potential character states at any point along a DNA sequence (A, T, C or G), eventually by chance mutations can change a single base many times, but we are only able to determine the base that is present at a particular location when we sequence that region (i.e. now), not what may have been there in the past. If we compare two individuals and find that they both share the same base at a particular point along a DNA sequence we interpret this as similarity due to common descent. If however, they share the same base due to

homoplasy we have no way of knowing this. Thus, it is essential to choose DNA markers carefully. Appropriate DNA markers need to evolve fast enough so that populations or species show differences, because without variation there is no basis for phylogenetic inference, but they must also evolve slowly enough so that there is little chance of character convergence (homoplasy) where we will score similarity, incorrectly. For any DNA marker there will be a point reached when homoplasy will become an issue, so we should choose a DNA marker appropriate for the time frame we are examining to reduce possible confounding effects of homoplasy. This point theoretically, is when sufficient evolutionary time has elapsed, given the mutation rate at the locus, for all four character states to have been expressed (A, T, C and G) at a single point with the final outcome being a return to a character state that previously occurred there. When this point is reached, we are likely to underestimate the real divergence time between two individuals with the same genotype.

The recognition that DNA sequences evolve at a constant rate as a function of their locus specific mutation rates, implies that the same gene sequence in two populations should evolve at the same rate. This is the basis for the idea of the 'Molecular Clock'. Essentially, what this means is that assuming population sizes are similar and remain constant over time, individuals in different populations will accumulate mutations at approximately the same rate and so the absolute minimum number of mutations between any two individuals in each population can be used to calculate the evolutionary time that has elapsed since they last shared a common ancestor, once we have calibrated for mutation rate. Put another way, the absolute number of mutational differences between the two individuals with the most similar genotypes in each population is a direct reflection of how closely related the two forms are. Once we have this information we can correlate estimated divergence times with past earth geological or climate history events that may have impacted on the evolution of the different forms.

An example of this is the evolution of the flightless ratite birds (Emus, Ostriches, Rheas, etc.). Ratites are an ancient order of birds that now are limited to a few relict species confined to the southern continents (Australia, Africa and South America, respectively). Molecular analyses confirm both the relationship between the three surviving families and the fact that they last shared a common ancestor in the Cretaceous. The simplest explanation for their evolution is that the common ancestor evolved when the three continents were part of a super continent (Gondwana) that also included Antarctica. This giant land mass that was fractured subsequently due to tectonic plate movement that lead to the sequential rafting of the three continents northward carrying their ancestral flightless ratites with them (first Africa, then South America and finally Australasia). The relatives of the ancient ratites (Emus, Ostriches

and Rheas etc.) are now living evidence for both the existence of giant super continents in the past but also for how Molecular Systematics can help to tease out the processes that have influenced modern biodiversity. Vicariance due to the splitting of the Gondwanan land mass has also been invoked to partially explain the distribution of freshwater galaxids among southern hemisphere continents (Waters et al., 2000).

# SECTION 2

## Genetic variation in nature

Genetic variation is essential for evolutionary change in populations. While we are well accustomed to detecting variation in our own species, we are less perceptive at detecting fine scale variation in other species. But the level of phenotypic variation we can detect in ourselves is also present in most other species. The amount of variation in a population will influence its relative rate of evolution. Thus large populations should contain higher levels of genetic variation than small populations. What this means, at least conceptually, is that populations with little or no genetic variation have little potential to respond to environmental disturbance and therefore have a higher probability of going extinct when environments change. As a consequence most extant populations are variable.

The genetic variation present in natural populations comes from three fundamental sources: Mutation, Genetic Recombination and Sexual Reproduction. All variation ultimately arises from mutation but this varation is mixed among individual chromosome strands by genetic recombination, and mixed among diploid individuals by sexual reproduction when individuals combine their gametes to produce a zygote. The extent of this variation means that no two individuals in a population (except for monozygotic twins or clones), will be genetically identical. Since most individuals are genetically unique and population size determines how much mutational variation can accumulate in a population, larger populations of sexually-reproducing individuals accumulate more genetic variation than will smaller populations.

Mutations are the ultimate source of genetic variation and occur randomly along DNA sequences. Mutation rates vary widely among DNA sequences (over 1000 fold among genetic loci). An optimal rate exists however, for each DNA sequence, so populations of a species and closely related species are likely to share the same mutation rate for each common DNA sequence.

A question that has long interested evolutionary biologists is 'how much genetic variation actually exists in natural populations?' For a long time (based on observations of external morphological variation in natural populations) biologists believed that genetic diversity in natural populations was relatively low because most con-specifics looked morphologically very similar. So, when biologists quantified variation in individuals in a population for morphological traits (e.g. colour variants), they tended to regard it as unusual and not typical of most traits that they examined. This led to a view that Evolution as a process in general was conservative and hence slow, and they argued that this could be explained by the fact that if mutations were random, then changes in genes were likely to produce poor (deleterious) outcomes in mutants and hence will be lost as a result of 'purifying selection'. This view however, developed at a time when we did not

know about non-coding DNA and had no molecular data on levels and patterns of genetic variation.

The modern view of how much genetic variation exists in natural populations is quite different to the 'Classical View" described above and resulted from the development and application of molecular analyses of genetic diversity in natural populations that commenced in the late 1960's. The early studies exploited the development of the technique of Allozyme Electrophoresis first developed for human disease diagnosis in the early 1960's. This was extended to molecular analyses of DNA markers in the late 1970's and early 1980's. What these studies have shown is that genetic diversity in most DNA sequences is in fact much higher than had been predicted from earlier morphological studies and so required a new explanation. This led to the idea that the high genetic variation evident in most DNA sequences could result from two evolutionary mechanisms; Natural Selection (NS) in the form of balancing selection and the random accumulation of neutral mutations.

Evolution via NS (or Darwinian evolution) results from the difference in relative fitness of the possible phenotypes present in a population. Evolutionary biologists now recognise a variety of different types of natural selection that can change gene frequencies in a population including: heterozygote advantage, effect of patchy environments, frequency dependent selection, epistatic interactions etc. to name but a few. For NS to directly influence allele frequencies at a locus however, the locus must be coding and produce different phenotypes. Neutral evolution in contrast, is where changes in allele frequencies occur simply as a consequence of accumulation of neutral mutations at a locus and their frequencies change as a result of random genetic drift (a function of population size) over time. This idea (developed by Kimura in the 1960's into the 'Neutral Theory of Evolutionary Change') argues that genetic drift acting in populations of different size (i.e. chance) will determine the fate of most individual mutations at a locus over time. Kimura showed that both coding and non-coding DNA sequences in theory, can evolve by genetic drift alone and that there is no absolute requirement for NS to influence gene frequencies at a locus for populations to diverge. Even though the effect of GD will be greater in small populations, over evolutionary time isolated populations are likely to diverge simply as a consequence of GD. This process will occur regardless of whether NS is affecting the frequency of alleles at the locus or not (unless strong balancing selection, i.e. heterozygote advantage, is present). Under this model the amount of genetic diversity at a locus is largely determined by a balance between how many new alleles are entering the population by mutation and the number being lost by genetic drift (the balance is referred to as 'mutation-drift equilibrium'). Since loss of alleles via genetic drift occurs more rapidly in small populations, large populations will be more genetically diverse simply because there are more

individuals in the population in which mutations can occur and fewer alleles will be lost by drift.

Thus developed two contrasting hypotheses that attempted to explain genetic change in populations over time: evolution via NS or neutral evolution. Both attempt to explain why there is so much genetic variation present in most natural populations, but do so from essentially opposing positions. Proponents of Darwinian evolution (i.e. evolution via natural selection) argue that genetic variation results from accumulation of mutants that produce phenotypes that have higher fitness than alternative phenotypes at the locus driven by a variety of different selective processes singularly or in concert. In contrast, proponents of the neutral model of evolutionary change argue that genetic variation accumulates in populations simply due to the fact that most new mutations entering a population are selectively neutral i.e. they affect fitness very little if at all and so it is chance and population size that are the most important factors that will affect their long-term fate. Since modelling has shown that both mechanisms in theory can change gene frequencies and there are practical examples in nature of both processes, this led to a debate as to which mechanism was responsible for the majority of observable genetic variation in nature. The so-called 'Neutralist - Selectionist debate' that is still to be finally resolved.

So is most genetic variation (read – evolution) in nature determined largely by adaptive or neutral processes? The first point to recognise is that different types of DNA are more likely to be affected by one or other mechanism. Non-coding DNA is likely to be influenced by neutral evolution because no phenotypes are expressed, so the only way that non-coding DNA can be affected by NS is if by chance it occurs in close proximity to a coding sequence on a chromosome and genetic variation at the non-coding sequence is influenced by selection acting at the adjacent coding locus, indirectly. This is called a 'hitch-hiking' effect. On the other hand, coding DNA can be affected by NS and the more important is the functional role of a coding locus, the more likely it is that NS has, is or will affect genetic variation at the locus over time.

While evolutionary biologists agree that genetic variation levels in nature are generally high and large populations contain on average more genetic diversity, the relative importance of NS and genetic drift in generating diversity remains a matter of ongoing debate.

# SECTION 3

# Basic concepts in population genetics

Basic concepts in population genetics are central to understanding the processes that influence development of population structure in natural populations. The science of population genetics focuses on heredity in groups of individuals and populations and aims to describe the genetic composition of populations and to document and understand the forces that change their genetic composition over time. Thus at its heart, population genetics seeks to understand the process of evolution.

The fundamental starting point to understanding population genetics is to recognise the relationship between DNA and the phenotype. Encrypted in the DNA of all organisms is the genetic information necessary to encode phenotypes, but first in eukaryotes it has to be transcribed into a carrier molecule (mRNA) and this molecule then moves to the cytoplasm of eukaryote cells where it can be translated into the encoded polypeptide chain at the ribosome. Essentially there are three steps in the process; transcription, translation and gene expression of which only gene expression can be influenced by external environmental factors. While mutations can affect any stage of the process, only mutations in the DNA have the potential to be passed among generations.

From a population genetic perspective, Evolution can be defined as any change in phenotypic frequency in a population over time. To demonstrate that evolution has occurred in a population we need to satisfy three requirements; first that the trait in question is heritable, secondly what evolutionary mechanism(s) may have caused the change and thirdly, how the evolutionary mechanism(s) are having their effect. In nature however, satisfying the three requirements is often not easy and so we may have to be satisfied by inference rather than direct evidence for some of the requirements.

If we believe that a phenotypic trait in a population may be evolving and wish to test the hypothesis and attempt to understand the process, we need a foundation (essentially a 'null hypothesis'). This is simply because we can never prove the hypothesis that evolutionary processes have changed the gene frequencies at the locus (or loci) coding for the trait, but we can refute the null hypothesis that evolutionary processes have not changed the gene frequencies. To set up this null hypothesis test we need to employ the Hardy-Weinberg Principle. Hardy and Weinberg were mathematicians in the 1930s that developed the basic mathematical platform for modern population genetics. They were interested in how gene frequencies can change in natural populations and recognised that before this process can be explored there has to be an *a priori* reason for focusing on a particular trait. Put simply, there is no reason to try to understand what forces are changing gene frequencies at a locus or how they have their effect without first having reasonable evidence that gene frequencies have in fact, changed! So they modelled the effects of gene frequency change in populations over time and developed what is now referred to as

the 'Hardy-Weinberg Equation' (H/W). The Hardy-Weinberg Principle can be defined as; 'in the absence of migration, mutation and natural selection, gene frequencies and genotypic frequencies remain constant in a large, randomly mating population'.

The Hardy-Weinberg equation essentially states the null hypothesis of gene frequency change, i.e. that if no evolutionary mechanisms are affecting the frequencies of alleles at a locus, then the frequencies should not change over time or among generations. Thus, if we have the necessary information and data to test the hypothesis for a locus of interest and we are able to refute the null hypothesis that no change in gene frequencies has occurred, then we are in a position to justify searching for a mechanism(s) that may be causing the change and to attempt to understand the process. Hardy and Weinberg recognised however, that there are qualifications on the attributes of populations in which their principle would hold. They defined this population as a 'Mendelian population' and recognised that it must have the following attributes; be diploid, sexual, outbreeding, randomly mating and large. Populations that satisfy these conditions are considered to have reached H/W equilibrium and this means that every reproductive individual has an equal chance of mating, all genotypes at a locus have equal fitness, each new generation is a random sample of the previous generation's gametes and no new alleles appear in the population. We now recognise that not all (if any!) natural populations will fully satisfy all of these attributes but unless we have evidence to the contrary we can assume that most large natural populations will approach H/W status. Characteristics of populations at H/W equilibrium are that allele frequencies at autosomal loci will not change across generations, genotype frequencies will also remain constant and if H/W equilibrium is disturbed, it can be re-established within one generation of random mating.

By inference populations that do not satisfy H/W equilibrium must be experiencing changes in gene frequency due to some evolutionary mechanism. An example would be changes in the frequency of a recessive allele that causes a genetic disease in humans across generations. When expressed in the homozygous state (rr), individuals suffer from the genetic disorder and may die before they can reproduce, thus when this happens, the frequency of the 'r' allele in the population will decline. Assuming that no other factors affect survival of the mutant recessive allele over time, we should expect that this allele will eventually go extinct because it is not favoured by natural selection. An example in humans is the mutant allele that causes haemophilia. Although inheritance of this allele is complicated by the fact that the locus is sex-linked and so is inherited in a different manner in males and females.

Where we have data on gene frequencies in a population at a locus we can test to see if the population conforms to H/W equilibrium by comparing

the distribution of genotypes against those expected if the population was at equilibrium. To do this we use the H/W equation i.e. in the simplest case, if the trait in question is determined by a single genetic locus with two alleles then if we let the frequency of the (**a**) allele equal **p** and the frequency of the alternative allele (**b**) equal **q** then:

$$p + q = 1$$

But in diploid organisms for most nuclear genes we inherit two copies of each gene, one from each of parent, so Hardy and Weinberg realised that their equation needed to take the diploid condition into consideration and to recognise that there are three ways that an individual can carry **a** and **b** alleles if they are diploid. To address this issue they expanded the equation to deal with diploid genotypes so that; $\mathbf{p^2}$ is the probability of receiving a copy of the a allele from both parents, **2pq** is the probability of being a heterozygote and $\mathbf{q^2}$ is the probability of receiving a copy of the **b** allele from both parents then:

$$p^2 + 2pq + q^2 = 1$$

The equation can also be expanded to deal with cases where there are more than 2 alleles at the locus, e.g. $p + q + r = 1$. Once we have data for the observed allele frequencies constituting the genotypes in the population we can then use the H/W equation to calculate the expected genotypic frequencies if the population was at equilibrium. The expected frequencies of genotypes can be compared with observed frequencies

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

$$\chi^2 = \sum \frac{(\text{Observed - Expected})^2}{\text{Expected}}$$

in a simple $\chi^2$ test with the $\sum$(observed – expected)$^2$ / expected (with degrees of freedom equal to number of genotypes minus number of alleles) (Example 1). If after completing such an analysis the result is that the population does not conform to H/W equilibrium then we can look for information that can help to identify the likely causative agent (evolutionary mechanism) with the results of the test providing some insight into the possible cause. For example, an excess of heterozygotes may indicate balancing selection (i.e. heterozygote advantage), whereas a deficiency of heterozygotes may reflect disruptive selection or non-random (assortative) mating.

While Darwin was aware of only a single class of causative agent for which he coined the term ‘natural selection’, modern evolutionary biologists recognise at least six different mechanisms that can cause populations to deviate from H/W equilibrium (mutation, migration/gene flow, non-random mating, genetic drift, natural selection and ‘molecular drive’). Of these; natural selection, genetic drift and migration/ gene flow are the mechanisms most commonly considered to be the most

important ones that can affect population structure, at least over shorter evolutionary time frames.

Example 1. Observed distribution and expected Hardy-Weinberg equilibrium distribution of genotypes can be summarised in the Table on the following page:

| | 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 (1 in the example), the expected number of the last class is set. Thus in the above example there is only one degree of freedom in the analysis.

Check the $\chi^2$ value of 0.37 at df = 1 in Table 1 we will have *P*-value > 0.05 and therefore we accept the null hypothesis of Hardy-Weinberg equilibrium in the population in our example.

**Table 1.** Chi-square Probabilities.

| | Probabilities | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| df | 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 at 0.05 level → | | | | | | | | ← Reject → | | |

# SECTION 4

## Natural selection

Charles Darwin and a colleague, Alfred Wallace, established that evolution could result from the effects of natural selection changing the frequency of genetically determined traits in nature. Since this was the only mechanism that had been proposed to drive evolutionary change in nature from the middle of the 19th century until the 1930's, it has attracted considerable interest from evolutionary biologists over time and continues to do so. Simply put, NS acts on heritable variation and is the relative ability of individuals with different phenotypes to survive and pass on their genes to their offspring. Where NS is affecting allele frequencies at a locus, over time individuals with superior phenotypes (and hence superior underlying genotypes) in a particular environment will tend to have more surviving offspring and so their alleles will increase in frequency in the population at the expense of individuals with poorer performing phenotypes. Differences in reproductive output was termed 'relative fitness' by Darwin and by this he meant that individuals in the population with high relative fitness would on average provide more surviving offspring to the next generation compared with another individual with a poorer phenotype. The comparison is always 'relative' because it is population specific and is made against the best-performing genotype in a particular environment. This is an important point because the best performing genotype may not always be the same genotype if populations of the species are found in different environments. We can estimate the relative fitness of different genotypes at a locus in a population where we have data on the average number of surviving offspring per genotype across at least two generations. Relative fitness varies from 0 to 1 because the calculation is made in such a way that the best performing genotype of all possible ones at the locus is always given a fitness value of 1 and poorer genotypes a value less than 1. A genotype that does not produce any surviving offspring across generations will have a relative fitness of 0.

Sometimes different genotypes can have equal fitness in a particular environment or multiple niches are available for different genotypes and where this occurs, multiple phenotypes may do well over time. This is called a 'balanced polymorphism' and is one way evolutionary biologists who support the notion that NS is the most important evolutionary mechanism, believe that NS can maintain high levels of genetic diversity in natural populations. Balanced polymorphisms may evolve for a number of different reasons including that across the natural distribution of the species there may be different habitat patches that favour different phenotypes (and hence influence the frequency of their underlying alleles). The 'peppered moth' is a classic example of this kind of balanced polymorphism because in polluted environments in Britain where it occurs, the dark morph of the moth is favoured because it is more cryptic to predators than is the light form. In contrast in pristine environments, where there is little air pollution from

coal dust, the light coloured form is favoured because the resting place moths use during the day (trunks of oak trees) are covered with lichens that are white and light grey in colour that provide more protection to the light coloured morph than to the dark morph. Lichens are very sensitive to air pollution particularly coal dust and so in polluted areas they do not thrive and so the trunks of oak trees are basically dark brown to black, the natural colour of the tree bark. An alternative way balanced polymorphisms may evolve, is where multiple niches are available in the same place (environment). Shell colour variation in English land snails (Genus *Cepaea*) can be influenced by this process. There are patches within a single habitat type where different colour morphs may be more cryptic (e.g. certain patches favour banded snails and other patches may favour un-banded snails, so overall both the alleles for 'banded' and 'un-banded' remain in the population. Thus selection can favour one allele in a single place or multiple alleles in the same place so that a balanced polymorphism evolves for a particular trait.

Relative fitness can vary temporally, geographically and ecologically for a population. If one or more of these effects are evident then polymorphisms will be common at the locus. Temporal variation in fitness is where fitness may be affected in different ways at different life history stages (e.g. eggs vs fingerlings) or with different seasons within a single life history stage. Geographical variation in fitness occurs where factors that determine the fitness of individuals may vary geographically across the natural distribution of the species, while ecological variation in fitness may occur where factors such as differences in fitness associated with substrate type, depth, canopy cover etc. can influence fitness.

Once evolutionary biologists had recognised that selection can act in a variety of ways, attempts were made to model the impact of selection on traits with different modes of inheritance. These models attempt to predict the outcome of selection. The most important factor in the models is the time to 'fixation' or when one allele (the one favoured by NS) reaches 100% and allelic variation at the locus is lost. Time to fixation will depend on the starting frequency of the allele favoured by NS, differences in relative fitness among genotypes and the mode of inheritance of the locus. For simplicity, most selection models assume that selection pressure remains constant over time but in reality this may not necessarily always be true. Most models are essentially H/W models that incorporate selection coefficients i.e. an estimate of how much advantage the favoured genotype has over alternative genotype(s) at the locus. There are four basic kinds of selection model; (a) selection against the recessive homozygote, (b) selection that favours the heterozygote, (c) selection against a single allele at a locus and (d) selection that acts against heterozygotes. A special case of type (a) is referred to as a recessive lethal where only the recessive homozygote is affected and always dies pre-reproduc-

tion, so the relative fitness of this genotype is 0. Cystic Fibrosis in humans used to be an example of this type of mutation until modern medicine devised ways to prolong the life of some affected individuals.

Once we have data on the mode of inheritance of the mutant we can use relative fitness estimates incorporated into a H/W Model to determine the likely time to fixation under different selection intensities and look at the effect of the change with different starting gene frequencies. While the outcomes can be very diverse, one obvious characteristic is that the time required to purge a recessive allele that produces even extremely poor fitness outcomes for a sufferer is much longer (in terms of generation time) than for an equivalent allele that shows dominant inheritance. Equally it will take a much longer time for a new recessive mutant that provides higher relative fitness than pre-existing allelic forms of the gene to reach fixation than an equivalent dominant favoured mutation. The simple explanation for these phenomena relate to the differences in the mode of inheritance and the fact that NS can only act when a mutation is expressed as a phenotype, so deleterious mutations can remain hidden in the population with no effect simply because an individual requires two copies of the gene to express the phenotype. This is one reason why so many mutations that cause 'nasty' genetic disorders can remain in the genomes of species for many thousands of generations. In contrast if a 'nasty' mutation shows dominant inheritance it can quickly be eliminated by natural selection as long as it does not provide better outcomes than pre-existing forms of the gene in certain environments as is the case for alleles that codes for 'sickle-cell' anaemia in humans. While individuals that express the sickle cell allele in either the homozygous or heterozygous state have lower fitness in most environments where humans occur, where *Plasmodium falciparum* malaria is a problem (i.e. many tropical and sub-tropical environments) heterozygotes that express sickle-cell anaemia have higher fitness than homozygous normal individuals because they have higher resistance to infection from the malarial parasite. Mortality due to malaria is higher in these areas than the lower reproductive capacity that is associated with sickle-cell anaemia so essentially the environment (presence or absence of malaria) changes the fitness of an allele from negative to positive.

As discussed earlier we now recognise that NS can take many forms and affect individuals and hence populations in a diversity of ways. The evolutionary effects of NS can often be very complicated and even in opposition when different types of NS act in concert on a population at the one time. For example sexual selection may be favouring alleles in males in a completely different way to females (e.g. favouring more conspicuous males that just happen to be more obvious to predators as well) while other forms of NS (e.g. NS favouring cryptic colouration to avoid predation is acting

on both sexes). What we measure is the cumulative effects of both types of NS on colouration patterns not the individual effects of each process. Estimating the relative effect of individual NS agents even when they are known can be very difficult, because to do this we need to know; how many different selective agents are affecting a trait at one time, what are their individual effects and how they interact. In most cases this is not possible, so we simply look at their cumulative impact on phenotypes over time.

It is obvious however, that NS can be a very powerful mechanism for evolutionary change in natural populations and in certain situations can influence how populations are structured in space and time. This is especially true when populations have been in the past or are currently isolated from other populations so that gene exchange is either restricted or completely disrupted for extended periods of time. When this occurs, isolated populations are likely to experience their local environments differently because conditions will not be identical and so local selective agents may produce unrelated changes in gene frequencies and so result in population divergence. Geographical speciation models argue that this is the simplest and most widely accepted way in which new species can evolve from ancestral types (isolation leading to populations experiencing local environments differently and hence NS driving their divergence).

# SECTION 5

## Genetic drift

The process of random genetic drift is a powerful evolutionary force and is central to our understanding of population genetics. Random genetic drift (GD) refers to the random fluctuations of allele frequencies from one generation to the next. Sometimes it is referred to as a sampling error of gametes between generations. In a randomly mating population, the expectation that two particular alleles coming together at fertilisation is a function of the relative frequencies of each allele in that population and therefore should conform to Hardy-Weinberg expectations. In essence, these expectations are rarely realised due to stochastic events that may affect random mating, for example, unequal offspring numbers from individual females.

Because of the random nature of genetic drift, it is impossible to predict absolutely the fate of a particular allele. The effects of genetic drift on an allele over time however, will be to either increase or decrease in frequency in the population. Given sufficient time, the allele in question will increase in frequency until it reaches fixation or alternatively decrease until it becomes extinct. Either way, the locus in question is heading towards a homozygous state. So even if we are unable to forecast the outcome of an individual allele, we can state that the overall effect of drift is to reduce genetic variation (push polymorphisms towards homozygosity).

In a population of N diploid individuals there are $2N$ alleles at any particular locus. Therefore, if a new mutation arises, it will start in the population with a frequency of $1/2N$. This is also roughly the probability of the new mutation being passed on to the next generation. It is also the probability of the new allele increasing in frequency to fixation. From this simple relationship it can be seen that the eroding force of genetic drift on genetic variability is purely a function of population size ($N$). The greater the population size, the smaller the effect. Similarly in small populations, a new mutation will initially exist at a relatively high frequency (relative to a large population of the same species) thereby having a greater chance of being passed onto the next generation. Effects of small population sizes on genetic drifts will be further explained later.

Kimura and Ohta (1971) calculated that it would take 4N generations for the frequency of a newly mutated allele to reach fixation in a population. That is, if $p$ is the frequency of allele **a**:

Time to fixation of allele **a** (i.e. $p$=1):

$$T_1(p)=4N$$

Conversely, the time it takes for allele **a** to be lost from the population is:

Time to loss of allele **a** (i.e. $p$=0):

$$T_0(p)=2\ln(2N)$$

With the ratio between them:

$$2N/[\ln(2N)]$$

It can be seen therefore, that in a population of 500 individuals, it takes approximately 145 times longer for the allele to go to fixation than it does for it to be lost from the population. It is easy to see therefore, that the majority of new genetic variants will be more likely to go extinct than to become established in the population

Given enough time, every locus will become homozygous as a single allele will have drifted to fixation at each gene locus within the population (i.e. a total lack of genetic variation). Because of the potentially huge timescale required for an allele to reach fixation, however, this rarely (if ever) occurs. During the time it takes for an allele to head towards fixation, new mutations are continuing to arise that are subject to the same pressure of drift (with their own respective probabilities of increasing in frequency). Hence, a population is in a constant heterozygous state for many loci with the persistence of the polymorphism dependent on population size.

In the absence of selection and assuming that each mutation results in a unique allele, the level of genetic variation (heterozygosity) can be considered as a balance between the force of drift (that erodes variability) and mutation (generating variability). Kimura and Crow (1964) defined the equilibrium of heterozygosity as:

$$H = 4N\mu/(4N\mu + 1)$$

where $N$ is the population size and $\mu$ is the mutation rate. It can be seen from this equation that if either population size or mutation rate is low then heterozygosity will also be low. This is intuitive as small population size results in elevated drift thereby reducing genetic variation.
A general rule regarding the relationship between these two opposing processes is that if $4N\mu$ is much larger than one, then mutation is the dominant process and heterozygosity is high. If $4N\mu$ is much lower than one, then drift is the dominant force and heterozygosity will be low.

Natural populations also rarely remain stable in size over time. Many populations at some stage experience a sudden crash in numbers (usually due to some extrinsic disturbance such as an outbreak of disease). A rapid decline in size is referred to as a 'Population Bottleneck' that has a two-fold effect on genetic variability. Firstly, only a relative few individuals manage to pass their alleles onto the following generation and what genetic diversity does survive is subject to a greatly elevated pressure of drift. Ultimately, the degree to which genetic variation is lost is a function of two factors: i) the magnitude of the bottleneck (i.e. how few individuals survive to reproduce)

and ii) the duration of the bottleneck (i.e. how long the population remains at a low number). It can be argued that the duration of the bottleneck has a greater impact with respect to the loss of genetic variability. That is, if a population that has undergone a bottleneck can recover numbers rapidly, then the loss of variation will be attenuated.

Another form of population bottleneck is when a few individuals colonise an environment previously unoccupied by the species. This is known as a 'Founder Event' with the new population being subject to the forces of drift in the same way as seen with a bottleneck. The founding population will lose genetic variability much faster than will the parent population.

So far we have discussed the fate of genetic variation due to the forces of drift within single populations. Genetic drift also plays an important role in leading isolated populations to become genetically differentiated. This concept is central to population genetic theory. Because the process of drift is random, alternative alleles within different populations will increase (or decrease) in frequency. Eventually, populations will become fixed at particular loci for different alleles (i.e. total differentiation). It should be recognised however, that due to the random nature of drift, two populations could also become fixed for the same allele by chance. The probability of this occurring reduces rapidly as the number of alleles at the locus increases. For example, for a locus where there are only two alleles there is a 50% chance of two populations drifting to fixation of the same allele. If there were five alleles present, then the probability would drop to a 20% chance. Also the probability of two populations going to fixation for the same allele would be very low when considered across many loci. Therefore, the net effect of drift is to cause populations to differentiate (diverge genetically).

We have seen the relationship between mutation and genetic drift and how their interaction determines genetic variation. These predictions are only valid under the neutral theory of evolution. Although many (most) point mutations are selectively neutral, particular mutations may bestow a significant fitness advantage on the individual. In this case, new mutations may increase in frequency at a much higher rate than may be expected under neutral theory. The question is, which evolutionary force is stronger? Balancing selection will tend to keep multiple alleles in relatively high frequencies at the locus under selection thereby maintaining high heterozygosity. This opposes the effects of drift that reduces variability. The relationship between these two evolutionary forces is once again a function of population size. The smaller the population the greater will be the probability that the effects of drift will outweigh the effects of selection. The general rule is that if $4Ns$ (where $s$ is the selection coefficient) is much less than one, then drift is the most important process determining variability. If $4Ns$ is much greater than one, then selection is likely to be the dominant force. It should be noted

however, that other forms of natural selection (e.g. directional selection) can also lead to reduced heterozygosity.

It is clear from the preceding discussion that the force of drift in influencing genetic variability depends on population size ($N$). Given this fact, it is probably important to briefly explore the value $N$. When we think of a population size, we merely see it as the number of individuals present at a location at a particular time. In terms of population genetics however, this value can be misleading. The important point is that we are interested in the probability of alleles being transferred successfully to subsequent generations. Therefore we are only interested in the number of individuals that contribute their genes to the next generation (i.e. individuals that breed successfully). These individuals constitute what is called the 'Effective Population Size' ($N_e$). In nearly all cases, the effective population size is significantly smaller than the census population size. So even a population that appears to be very large may in fact have a relatively small $N_e$ and therefore be subject to an elevated pressure of drift. The concept of $N_e$ is particularly significant to conservation genetics. That is, how small can the effective population be until levels of inbreeding reduce the overall fitness of the population?

Another important concept regarding $N_e$ is that it will vary depending on which gene locus we look at. All autosomal genes in the nuclear genome will follow the rule listed above. However, genes on the Y chromosome or in cytoplasmic genomes (e.g. mitochondrial or chloroplast DNA) will exist at lower $N_e$ than for nuclear genes. For example, mtDNA is maternally inherited (compared to bi-parental inheritance for nDNA) and is a haploid molecule (compared to diploidy of nDNA), therefore half the number of parents times half the ploidy results in a four-fold reduction in $N_e$. Therefore, the effects of drift will be four times greater on mtDNA genes than on nDNA genes in the same population. This concept and its implications for assessing population structure will be developed later.

Effective population size can also be influenced by other factors such as unequal sex ratios or particular breeding behaviours where one or a few males breed with many females. The simple formula for calculating $N_e$ in these cases is:

$$N_e = (4N_mN_f)/(N_m + N_f)$$

where $N_m$ is the number of breeding males and $N_f$ is the number of breeding females. $N_e$ is affected more (in terms of reduced numbers) by the rarer sex. This is because they constitute less than half of the breeding population (sometimes significantly so) yet they still contribute 50% of the gametes to the next generation.

# SECTION 6

## Non-random mating and population structure

A Gene Pool is the collection of genotypes present in all individuals that constitute a reproducing population, so essentially it comprises all individuals who potentially could exchange genes. Sometimes a gene pool is connected directly i.e. reproducing individuals can meet and exchange genes directly or exchange may be indirect via intermediates because individuals choose not to move large distances or individual dispersal distances are not large enough to allow contact with all members of the gene pool. While one assumption of the H/W theorem is that individuals within a gene pool mate at random, this is seldom if ever the case in nature both for intrinsic and extrinsic reasons. Thus individuals that belong to a discrete gene pool are often distributed as 'demes', local populations, subpopulations or populations and share more genes in common with members of their own sub-group than with the rest of the gene pool. When populations become subdivided by limitations on dispersal, the population will inevitably become subdivided as complete interbreeding may not be possible, so mating will not be at random. This results in genes being structured spatially across the natural distribution of the gene pool.

A number of different types of non-random mating have been recognised by evolutionary biologists. One form is 'inbreeding' where individuals share more genes by common descent than would be expected by chance. Different levels of inbreeding exist from one extreme of self-fertilisation where the population is essentially an assembly of clones. This situation is relatively rare in nature however, because self-fertilising species have little genetic variation and hence lose most of the advantages of diversity. Even species that are capable of self-fertilisation may not necessarily engage in it (e.g. some mollusc species). More common in nature is the situation where organisms within a gene pool practice some level of inbreeding (i.e. non-random mating). The consequences of this can be that populations will be structured spatially and/or temporally.

Inbreeding may result from both intrinsic (e.g. behavioural traits) and/or extrinsic factors (e.g. physical barriers to dispersal). If individuals mate assortatively, that is they either choose other individuals as mates that are phenotypically similar to themselves (positive assortative mating) or individuals that are phenotypically different to themselves (negative assortative mating), this can affect the level of inbreeding in the population. An example of positive assortative mating may be the fact that in human populations, individuals more often than at random select individuals of the opposite sex of similar height, while examples of negative assortative mating include mate choice in mice and self incompatibility factors in some plants. Female mice have been shown to select males with different odours to themselves as mates when the opportunity exists. Odour in mice is in part, determined by Major Histocompatability Complex genes (MHC) that provide a major component of the bodies defence system against disease, parasites and pathogens. It is thought

Once modelers had determined that inbreeding increases homozygosity they worked out that where genetic data were available, this effect could be used to estimate the level of inbreeding that was occurring in a population essentially by comparing the observed heterozygosity against that predicted under H/W equilibrium given the observed allele frequencies. Thus the probability that two alleles are inbred is given by the inbreeding coefficient (*F*), where *F* is the probability that two alleles in an individual are identical by descent. *F* or the inbreeding coefficient varies from 0 where the population is completely outbred to 1 when the population is completely inbred so the population will consist of only AA and aa homozygotes for a two allele system. We can estimate the relative level of inbreeding in a population using:

$$F_X = \Sigma[(\frac{1}{2})^{n_1+n_2+1} + (1+F_A)]$$

Where:

- $F_X$ is the inbreeding coefficient of the individual in question
- $F_A$ is the inbreeding coefficient of the common ancestor, and
- $n_1$ and $n_2$ are the number of generations from the sire and the dam to the common ancestor, respectively.

The statistics of inbreeding were developed by Sewall Wright in the 1922 and later who modelled the effect of various processes on gene frequencies in natural populations and related this to what was expected under H/W equilibrium. The result is that the modern statistics of inbreeding take his name i.e. Wright's (*F*) statistics and a variety of versions are available for analysis with genetic markers that possess different modes of inheritance and in theory mutate in different ways. The general equation is:

$$(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT})$$

Where:

- $F_{IT}$ is the correlation of uniting gametes relative to gametes drawn at random from the entire population
- $F_{IS}$ is the correlation of uniting gametes relative to gametes drawn at random from within a population and,

- $F_{ST}$ is the correlation of uniting gametes within subpopulations relative to gametes drawn at random from the entire population.

The statistic of real interest in studies of population structuring is $F_{ST}$ because in essence it measures the extent to which the populations under examination are subdivided, or put another way, how much gene flow is occurring among subpopulations.

$F_{ST}$ varies between 0 and 1 where an $F_{ST}$ of 0 implies that the populations under examination have the same set of alleles in identical frequencies and an $F_{ST}$ of 1 implies that the populations share no alleles in common. In practice $F_{ST}$ among populations is rarely larger than 0.5 and is often much less. Wright proposed for a simple two allelic system at a locus where $F_{ST} >$ 0.25 constitutes very great differentiation and within the range 0.15 to 0.25 this constitutes 'moderate differentiation'. The actual interpretation of $F_{ST}$ is more complex. An example is that recently it has been shown that for hypervariable genetic markers (e.g. microsatellites) that often possess many alleles per locus, $F_{ST}$ estimates among populations may be considerably lower than for traditional markers with fewer alleles per locus (e.g. allozymes).

that choice of a male with different odour type by female mice increases the probability that their offspring will be more heterozygous at MHC loci and this attribute may increase overall fitness of the offspring. Both inbreeding and positive assortative mating will increase homozygosity while negative assortative mating will increase heterozygosity above that predicted under the H/W model.

As discussed earlier, a major factor that keeps populations evolving as a unit is when they are connected by ongoing dispersal. In theory, the more effective dispersal that occurs (where individuals move among populations and reproduce in the new site), the more similar populations should be, genetically. Effective dispersal is called 'Gene Flow' and is another method apart from mutation by which new genes can enter a population. Gene flow is a very powerful force for homogenising gene frequencies among demes or populations and the more gene flow that occurs the lower will be the level of inbreeding. Essentially, gene flow is a force that opposes development of population differentiation and hence population sub-structuring. As gene flow increases it should also increase heterozygosity in the receiving population. This effect results from crosses among individuals from different populations that did not have identical gene frequencies at all loci at the start of the process. So gene flow and inbreeding are essentially opposing forces that largely determine the extent of population structure that

will evolve among demes. If gene flow is high among subpopulations, then population structuring will be low because inbreeding is reduced. If gene flow is low among subpopulations, then population structuring will be high because inbreeding will increase.

Once the relationship between inbreeding and gene flow was understood, interest focused on the diversity of potential patterns of population structure that could result in nature. So migration models were devised to describe patterns of population subdivision that were possible. Essentially because they are population genetic models, they describe the relative contribution that migrants make to demes that they enter (i.e. the extent of effective dispersal).

The simplest migration model is an 'Island Model' where subpopulations of equal size over a geographical area interact in such a way so that they can exchange genes with equal probability. An example could be subpopulations of a fish species confined to a large lake. The relationship between $F_{ST}$ and gene flow ($N_m$) for the island model is:

$$F_{ST} = 1/(1 + 4N_m)$$

A second kind of model is 'Isolation by Distance' where relative gene flow among subpopulations of one large population is affected by distance among subpopulations and/or possible alternative paths by which individuals can disperse. An example could be populations of a species of fish that occur widely across an ocean. While dispersal is possible either directly or indirectly via generational connections, individuals disperse more commonly at a relatively local scale so that subpopulation differentiation is greatest at the largest spatial scale. 'Stepping Stone Models' are mathematically more complex and describe situations where dispersal is only possible between adjacent populations and the greater the geographic distance between populations the less chance there is of gene flow, so there is genetic isolation by distance. In this case the relationship between $F_{ST}$ and gene flow is:

$$F_{ST} = 1/(1 + 4N_m)(2N_\mu/N_m)^{1/2}$$

Notice that the stepping stone model approaches the island model when populations become very large. Also, the stepping stone model is a function of not only gene flow but the mutation rate (μ) as well.

An example could be catadromous fish species that spend much of their life cycles in freshwater, but can have limited dispersal via the marine environment and hence reach neighbouring rivers. Complexity of stepping stone models can be increased by spatial and temporal effects of the environment and this will have consequential effects on the relative complexity of the mathematical equations used to describe the relationships.

The models discussed above attempt to describe the patterns of population structure that can exist in specific situations in nature. All rely on the association between gene flow and

As biologists began to apply migration models to aquatic species they quickly realised that in some instances (e.g. riverine freshwater systems) that the existing models were not adequate to explain all possible limitations on gene flow. Riverine systems are unique in that they can impose a hierarchical structure on potential for gene flow on species that are obligate in that environment (e.g. some freshwater invertebrates and fishes). This lead Meffe and Vrijenhoek (1988) to develop a specific model to address this situation, a model they called the 'Stream Hierarchy Model' (SHM). What this model attempts to describe is the fact that rivers and streams are essentially dendritic spatial systems for the organisms that are obligate users of them. Consequently, their patterns of genetic diversity should reflect the dendritic nature of the habitat with genetic diversity is likely to increase down the system because of water flow effects on relative dispersal and gene flow structured hierarchically. Thus gene flow is structured according to the following hierarchy, within stream > among streams > among drainages so that:

$$H_T = H_C + D_{CR} + D_{RS} + D_{ST}$$

Where:

- $H_C$ = within population diversity
- $D_{CR}$ = differences among populations in a river
- $D_{RS}$ = differences among rivers in a drainage
- $D_{ST}$ = differences among drainages

With the expectation that: $D_{CR} < D_{RS} < D_{ST}$

inbreeding, i.e. as gene flow increases, level of inbreeding should decline. This means that when we have data on how differentiated two or more populations are from each other, in theory this tells us how much gene flow is occurring (or has occurred historically between them). This information is used to calculate $N_m$, a statistic that equates to the number of migrants moving between population or the migration rate. $N_m$ is based on an Island model of population structure and estimates recurrent gene flow among subpopulations and is equivalent to 'the probability that an allele randomly chosen from the population comes from a migrant'. $N_m$ can be difficult to measure in nature, but if we know allele frequencies in both donor and recipient populations

before gene flow and the change in allele frequency in the recipient population after gene flow, then we can estimate $N_m$. This is because the change in allele frequency over time following gene flow is proportional to the difference in frequencies between donor and recipient populations. The outcome of modelling of the effect of different levels of gene flow among populations has shown quite clearly that even very limited gene flow is sufficient to keep populations essentially, genetically homogenous. As little as a single migrant per generation is sufficient in theory, to homogenise gene frequencies among populations. So only very limited dispersal is capable of restricting divergence that results from local selection and genetic drift effects.

# SECTION 7

## Environmental influences on population processes

In the previous sections we have discussed the genetic processes that operate at the population level that principally determine population structure (i.e. mutation, genetic drift, gene flow and selection). These processes however, must operate within a framework shaped by the environment (extrinsic factors) and the ecology and life history traits of the species (intrinsic factors). In fact it is rather meaningless to interpret population genetic data (especially for management purposes) in isolation without taking intrinsic and extrinsic factors into account. In this section we will look at the effect that the environment can play in shaping genetic variation in natural populations, with the emphasis on freshwater systems. Firstly we can disregard mutation, as the effect of the environment on the mutation process largely results in somatic mutations which are rarely heritable (e.g. solar radiation causing skin cancer).

The environment can either promote or inhibit gene flow among populations and as such a heterogeneous environment (as is usually the case) will result in varying levels of population connectivity. An important consideration is that the environment or habitat of a species is rarely stable over time and therefore its impact on shaping population structure will consist of a historical and a contemporary component. That is, how the environment is affecting structure today (on a ecological time scale), and how it affected population structure in the past (on an evolutionary time scale).

Change in the physical environment that has affected levels of gene flow among populations on an evolutionary time scale has been immense. For example, the continents that we know today once were part of super-continents (Pangaea, Gondwana). As the continents drifted apart (via plate tectonics) gene flow ceased, leaving populations isolated from each other (unless they were very good at swimming or flying). The separation of the super-continents happened so long ago that most species affected by it have since gone extinct or isolated populations have evolved into different species. There are however, still several closely related taxa that share a Gondwanan distribution (e.g. marsupials, ratite birds, lungfish). Because the separation occurred so long ago, it bears little application to intraspecific level processes.

On a more recent evolutionary time scale however, many events have shaped the population structure of extant species particularly during the Pleistocene. Many populations became isolated due to the expansion of ice sheets during the most recent ice age. In fact many populations still bear the genetic signatures of these vicariant events in North America and Europe even though levels of gene flow are significantly higher today than 10,000 years ago. Mountain uplift due to tectonics or volcanism (geomorpholoical change) also has resulted in much habitat fragmentation leading to population differentiation and many of these populations still remain isolated today. The rise and fall

of sea levels (eustasy) also connected and isolated landmasses and hence populations repeatedly. For example much of the terrestrial fauna shared among the Indonesian islands and between Australia and New Guinea can be explained through this process. Over this sort of time scale there was also significant fluctuations in temperature which played a significant role in shaping genetic variation in populations. Changes in temperature generally led to reduced habitat availability with intervening regions often inhospitable to dispersal.

All of the historical environmental fluctuations mentioned above have had significant impacts on the population structure of freshwater fauna through the modification of dispersal pathways. One of the most significant effects resulted from geomorphological change through the rearrangement of drainage channels (e.g. river capture). Under this scenario where a stream flowing to one river system is 'captured' and begins flowing in another direction, populations that had been connected through a high level of gene flow previously, became totally isolated while populations that may have been isolated started exchanging genes. The geomorphological evolution of drainage channels is generally seen as the primary factor that influences the distribution of most obligate freshwater species.

Sea level fluctuations have also have a significant influence on shaping population structure of freshwater fauna. For example, towards the end of the Pleistocene, low sea levels resulted in freshwater connection between Australia and New Guinea via 'Lake Carpentaria'. Several freshwater species (e.g. gudgeons, rainbowfish, freshwater prawns) still have a distribution that reflect this history. Another effect of eustatic change has been on river systems that are currently isolated by the marine environment but historically had a freshwater confluence at times of low sea level. This phenomenon also explains the distribution of genetic variation of a southeast Asian freshwater catfish (*Hemibagrus nemurus*) among currently isolated river drainages. During the Pleistocene, low sea levels resulted in freshwater confluences on the Sunda Shelf that facilitated interdrainage gene flow (Dodson *et al.* 1995).

Sometimes climate has changed so rapidly (e.g. temperature), that species fail to evolve *in situ* and are forced to move to more suitable habitat. This movement may take the form of latitudinal or altitudinal shifts. For example, freshwater crayfish in Australia (*Euastacus* sp.) historically had a widespread lowland distribution. As temperatures began to increase in the Miocene, they were forced to retreat further and further up mountains where cool moist conditions still remained. Eventually, connectivity among mountain top populations was cut as the intervening lowlands became uninhabitable for crayfish.

On an ecological time scale, there are also many factors that can affect levels of gene flow, either promoting

or restricting it. Firstly, it must be recognised that due to the nature of river systems, freshwater populations are expected to be highly structured especially among drainages. The terrestrial environment and the marine habitat that separate rivers, inherently dictate that gene flow will be highly restricted. Climatic fluctuations however, can overcome these barriers to dispersal. High rainfall can result in freshwater plumes around the mouths of rivers (e.g. the freshwater plume at the mouth of the Amazon River sometimes extends hundreds of kilometres into the Atlantic Ocean). Depending on the scale of the plume and the proximity of the neighbouring river mouths, connectivity among normally isolated rivers may exist and for a short period of time a small degree of dispersal may result. Also, flooding caused by high rainfall can lead to a high degree of connectivity amongst normally isolated drainages resulting in massive interdrainage dispersal events, especially in areas of low elevation (e.g. inland eastern Australia).

Within a single drainage there also exist several natural barriers to gene flow, some of which are influenced climatically. For example, headwater streams that are continuous during the wet season may be transformed into a series of isolated waterholes during the dry season. Also dispersal vectors such as water currents may change seasonally (e.g. Tonle Sap River, Cambodia). These processes can affect gene flow. Generally, most species have evolved to cope with seasonal environmental fluctuations but random catastrophic disturbances (e.g. tropical cyclones, earthquakes, volcanic eruptions) that alter the landscape will probably affect the qualities necessary for continued connectivity. For example, volcanic eruptions in New Zealand some 2,000 years ago with associated larva flows and ash deposits, resulted in small isolated populations of freshwater fish species with little or no potential for gene flow among them.

Other natural instream barriers to dispersal include waterfalls, rapids and cascades. It is not uncommon for upland populations to be totally isolated from downstream populations that are divided by a significant and rapid change in stream profile. Stream flow itself dictates that gene flow downstream is going to be significantly greater than in an upstream direction, unless species have evolved dispersal mechanisms to counteract this effect (e.g. positive rheotaxis). Another important barrier to dispersal in some freshwater systems is just physical distance. In extensive drainages such as the Mekong River, it is not physically possible for an individual to traverse the entire distance of the river in a single lifetime.

As can be seen from this discussion, there are many environmental factors that can influence gene flow (either promoting or restricting) that operate over various temporal scales. For management purposes, an important goal is to understand the magnitude of gene flow that is occurring today. As such, one of the challenges of population genetics is to be able to

differentiate the effects of historical versus contemporary gene flow on the observed population structure. Early population genetic studies based on allozymes were largely unable to accomplish this. Allozymes (and to a certain extent mtDNA haplotypic frequency data) can distinguish between high gene flow and total isolation, but the interpretation of situations in between these extremes can only ever be an educated guess. The development of more sensitive techniques (i.e. DNA sequencing, microsatellites) has provided tools that allow us to determine more confidently the relative contributions of both historical and contemporary processes to gene flow.

Finally, humans in recent times have had a substantial impact on gene flow in freshwater systems. Over the past few hundred years, anthropogenic modifications to natural water courses have been significant. Mostly these modifications such as dams, pollution and stream channel alteration have resulted in dispersal being restricted further. Because most anthropogenic disturbance has occurred relatively recently, any cessation of gene flow is unlikely to be detected in the data from a population genetics survey (although some population structuring has been recorded either side of some dams in the United States that have only been in existence for 50 years). On the other hand, human mediated gene flow via interbasin transfers of water or direct translocation of species among drainages can be detected if the newly mixed populations were genetically divergent in isolation (mtDNA is a particularly powerful marker for this application).

We know that the population process of random genetic drift is largely a function of population size – the smaller the population the greater will be the effect of drift. It is also commonly known that populations naturally fluctuate in size over time with much of the fluctuation a result of environmental influences. Once again, these environmental fluctuations have a historical and a contemporary component. For example, during the Pleistocene much of the freshwater habitat in the northern hemisphere was locked up as ice and what suitable habitat was left tended to fragment large populations, sending many subpopulations extinct. The surviving individuals existed as small populations in small habitat refugia. During this time much genetic variation would have been lost. Subsequent climatic warming re-opened much habitat allowing the small populations to rapidly increase their range and expand into areas previously unavailable to them with an associated increase in population size.

On an ecological time scale, natural seasonal shifts result in fluctuations of available habitat (as seen in the previous section). When habitat is reduced such as in the dry season, population size generally decreases. Similarly seasonal fluctuation can affect resource availability. If periods of poor habitat and low resources coincide, local populations may crash

and possibly become extinct. Because seasonal fluctuations are short lived, it is expected that the level of genetic variation in the population will be determined by the duration of poor conditions and hence the effects of drift when the population is at its smallest size. If only a few individuals make it through the bottleneck, regardless of the rate of recovery, genetic variation will have been lost and recovery can take a long time.

While most species are well adapted to their environments and have life history traits that are well suited to seasonal environmental fluctuations, catastrophic events can have a devastating effect on population numbers and even result in local extinctions. It may take many generations before population size recovers and many more before genetic variation reaches pre-disturbance levels. The amount of genetic variation a population can maintain over time is determined by the population size at its lowest level, not at the highest.

When new mutations arise, they are either beneficial, deleterious or neutral. Their relative fitness is purely a function of the environment. Much genetic variation may exist in a population that is essentially selectively neutral. A sudden change in environmental conditions however, may result in a particular genetic variant having a significant selective advantage (or disadvantage). This will lead to strong directional selection that will inherently reduce genetic variation. In association with genetic drift, localized differential selection pressures due to varying environments play a major role in influencing genetic differentiation among populations.

On the other hand, selection may act to homogenize allele frequencies among populations, even in the absence of gene flow. If the local environments of populations are similar, alleles may be under similar selection pressure (i.e. the same alleles are favoured in both populations) thereby creating a population structure that would be expected under a model of high gene flow. This highlights the necessity for choosing neutral markers for population studies that are capable of revealing any population structure that may be present.

# SECTION 8

# Ecological influences on population processes

It is difficult to discuss ecological influences on population processes without incorporating environmental factors because a species' life history traits (LHT) will adjust over time to local environmental conditions. However, certain LHTs will inherently influence the effects of gene flow and genetic drift.

Gene flow can be achieved by individuals at all life history stages (i.e. from fertilized eggs through to adult) or as gametes (eggs or sperm). Most species have evolved a dispersal phase in their life history in order to avoid inbreeding and competition with close relatives. Dispersal can either be of a passive or active nature. Passive dispersal is usually undertaken as gametes or as planktonic larvae, but exceptions do exist (e.g. some adult spiders disperse large distances in the wind by producing 'silk parachutes'). Passive dispersal has advantages because minimum energy is required, however a dispersal vector is required (e.g. a water current). The disadvantage of this form of dispersal is that the individual may end up in unsuitable habitat.

In most river systems, passive dispersal is always in a downstream direction. This presents a problem – how do upstream reaches of a stream remain colonized? Many freshwater taxa with a passive dispersal stage also have a compensatory behaviour at some stage of their life history. For example, many freshwater crustaceans display a positive rheotactic response as adults (i.e. they actively swim or walk against the current). Similarly, many insect species that have freshwater larvae, fly upstream as adults to breed. Some species, rather than compensating for down stream dispersal, have evolved physiological or behavioural traits that assist them to avoid displacement in the first place. Most freshwater crustaceans have an abbreviated larval phase thereby reducing the time in the plankton. Some species are dorso-ventrally flattened which makes them less 'visible' to the water current. Others still have the ability to adhere to the substrate or some species glue their eggs to the substrate. Behavioural adaptations include brooding of eggs or larvae, remaining at the edge of the stream (where the current has less velocity), hiding under large immovable objects (rocks, snags, etc.) and burrowing into the substrate.

Irrespective of compensatory, physiological or behavioural adaptations, the majority of gene flow in freshwater systems is in a downstream direction. Therefore downstream populations tend to act as 'sinks' for genetic variation and should display higher levels of diversity that populations further upstream. Furthermore, confluence sites should have a mixture of all alleles present in the river branches that lead to them. This effect is further accentuated if there is a barrier to dispersal such as a waterfall that significantly restricts upstream movement. Therefore, new mutations that arise in upstream populations can disperse downstream but new variants from downstream may not be found upstream.

Species that have evolved an active dispersal phase are particularly vulnerable to anthropogenically modified environments, especially migratory species. For example, dams or impoundments can interrupt long established dispersal pathways to breeding or feeding grounds. Disruptions to the natural life history of the species in this manner will result in a marked reduction in the potential for long-term population persistence.

Many species display sex-biased dispersal in their life history. That is, either males or females, but not both, are the principal dispersers. This has significant implications for mtDNA studies due to the maternal inheritance of the molecule. If dispersal is male mediated, then there is no effective dispersal of mitochondrial genes (i.e. gene flow is zero). Therefore a mtDNA survey may indicate strong genetic structuring while nuclear markers may reveal panmixia. A similar pattern may be seen in philopatric species (those that return to their natal site to reproduce). Even though these species may disperse over great distances (e.g. across oceans) if the female is philopatric, mtDNA gene flow is nil (e.g. this pattern is seen in sea otters).

Species evolve to maximise their reproductive output. This may be through breeding at a time of maximum resource quantity/quality, iteroparity (multiple breedings over time) or through modifying the reproductive allocation to suit prevailing environmental conditions. Irrespective of these adaptations to maintain high population numbers, fluctuations in environmental conditions (both predictable seasonal and catastrophic change) mean that most populations will go through declines and expansions (boom/bust). As discussed in previous sections, the severity and the duration of population declines will largely determine the level of genetic variation that can be maintained in the gene pool. The most extreme form of population size fluctuation is that of extinction and recolonisation. Depending on the source and magnitude of the recolonisation, genetic diversity may either increase or decrease, both within and among populations.

Evolution of LHT in some species has resulted in breeding systems where certain sexually mature individuals (usually the males) gain a high percentage of matings (e.g. harem system in many Pinniped species). This behaviour by itself reduces Ne significantly. From Section 5 we know that unequal numbers of breeding males and females will result in the effective population size being significantly smaller than the total number or breeders. In some breeding systems, selection has favoured mate choice (‘good genes’ hypothesis) where the reduction in $N_e$ is offset by the increased genetic quality of the offspring (e.g. co-operative breeding in birds).

An important outcome of size fluctuations is that many natural populations will rarely achieve mutation/drift

equilibrium, a condition that forms a common assumption underlying many statistical analyses.

# Glossary

**Aestivation:** Dormancy during summer or dry season.

**Allele:** An alternative form of a gene occurring at the gene locus.

**Allopatric:** Relating to the geographic distribution of populations/species with distributions that do not overlap.

**Allozymes:** Alternative forms of an enzyme coded for by different DNA sequences at a single genetic locus.

**Ancestral retention:** Isolated populations having the same allele from a time prior to isolation.

**Autosomal loci:** Gene sequences on non sex linked chromosomes.

**Balanced polymorphism:** Where multiple alleles exist at a single locus over evolutionary time.

**Balancing selection:** Process by which multiple alleles are maintained by selection at a coding locus.

**Base-Pairing Rule:** Where A binds to T (U) and G binds to C.

**Bonferroni correction:** Adjustment of the significance ($\alpha$) of a statistical test to reduce the probability of committing a Type I error through multiple comparisons.

**Bootstrapping:** Permutation method for testing the reliability of a node in a gene tree.

**Coding:** Region of DNA that can be transcribed and translated to produce functional polypeptide product.

**Co-dominant:** Locus where heterozygotes express a unique phenotype.

**Codon:** Nucleotide triplet in DNA or RNA that specifies the amino acid to be inserted in a specific position of a polypeptide.

**Confluence:** A place where two water channels join.

**Conspecific:** Of the same species.

**Cytoplasm:** All cell contents excluding nucleus.

**Denature:** Process of breaking the bonds between the two complementary strands of DNA through chemical or temperature stress.

**Dendritic:** Resembling a dendrite (nerve cell) which is characterised by many paths of connection.

**Diploid:** Referring to an organism having two sets of chromosomes, one from each parent.

**Directional selection:** Process by which one phenotype if favoured by selection produces an increase in relative frequency of the underlying allele.

**Dispersal:** The movement of individuals.

**Dispersal vectors:** Extrinsic entities that facilitate dispersal; physical (wind, water) or biological.

**DNA markers:** DNA sequences that can characterise individuals, populations, species, etc.

**DNA polymerase:** Enzyme that catalyses production of new DNA molecules.

**DNA replication:** Process of generating new DNA strands.

**Dominant:** Locus where heterozygotes are not detected from homozygote dominants.

**Duplex:** Single stranded DNA that has re-annealed to its complementary strand or another strand with different sequence (homoduplex and heteroduplex respectively).

**Effective population size:** The number of breeding adults in a population that contribute their genes to the next generation (Ne).

**Electrophoresis:** Procedure for separation of molecules (based on charge and/or structure) in an electric field.

**Epistatic interaction:** Interaction of nonallelic genes to produce phenotypes.

**Eukaryote:** Referring to the superkingdom that contains organisms whose cells contain membrane-bound nuclei and mitochondria (Protista, Fungi, Animalia, and Plantae).

**Eustasy:** Rise and fall of sea levels.

**Evolution:** Changes in genetic composition that occur within populations from one generation to the next.

**Evolutionary mechanism:** A process that results in changes in gene frequencies in populations.

**Extant:** Currently in existence, not extinct.

**Extrinsic factors:** Factors outside the basic nature of something.

**Fixation:** The point when an allele reaches a frequency of 100% in a population.

**Founder event:** The formation of a new population by one or a few individuals.

**Gamete:** Germ cell (egg, sperm) with a haploid genome.

**Gene:** A hereditary unit that occupies a specific location (locus) on a chromosome, the physical entity that is transmitted from parent to offspring.

**Gene expression:** When a coding locus produces a phenotype (protein).

**Gene flow:** Successful movement of genes among populations via dispersal of individuals or gametes.

**Gene frequency:** The frequency that an allele occurs at within a population.

**Gene function:** The role that a specific coding DNA sequence plays.

**Gene tree:** A branching diagram depicting the inferred relationships among a group of genes or other DNA fragments.

**Genetic Drift:** Random changes in gene frequency due to chance.

**Genetic recombination:** Reassortment of DNA sequences on homologous chromosomes at meiosis.

**Genome:** The total genetic material within a cell or individual.

**Genotype:** Genetic constitution of a cell or individual.

**Geographical speciation:** Where two species evolve from a common ancestor as a result of geographical isolation and independent evolution in different environments.

**Geomorphology:** Study of the evolution of physical landscapes.

**Gondwana:** Supercontinent (consisting of South America, Africa, Australia and Antarctica) existing ~200 million years ago.

**Haploid:** Single copy of each gene.

**Haplotype:** Genetic constitution of a haploid cell.

**Heritable:** Capable of being passed from one generation to the next.

**Heredity:** The genetic transmission of characteristics from parent to offspring.

**Heteroduplex:** Combining DNA's from different sources together.

**Heterogeneous:** Variable, made up of different elements.

**Heterologus:** Non-identical DNA sequences.

**Heteroplasmic:** Contains more than a single DNA sequence per cell.

**Heterosis:** Also known as hybrid vigour and heterozygote advantage. Occurs when the fitness of individuals with two different alleles at a locus is greater than the fitness of individuals with two identical alleles at a locus.

**Heterozygote:** An individual that carries two different alleles at a diploid locus.

**Homoduplex:** Single DNA strand bound to its mirror image.

**Homoplasmic:** Contains a single DNA sequence per cell.

**Homoplasy:** Denotes parallel or convergent evolution of DNA sequence information, same allelic state not resulting from descent from a common ancestor.

**Homozygous:** An individual that carries two copies of the same allele at a diploid locus.

**In vitro:** In an artificial environment outside an organism.

**In vivo:** Within a living organism.

**Indels:** Changes in DNA sequence, specifically insertion or deletion of nucleotides.

**Inbreeding:** Mating among closely related individuals.

**Intraspecific:** Pertaining to interactions among individuals of the same species.

**Intrinsic factors:** Factors belonging to the basic nature of something.

**Introgression:** Mixing of discrete entities/populations.

**Iteroparity:** The characteristic of breeding more than once in a lifetime.

**Life history trait (LHT):** Significant feature of the life cycle through which an organism passes.

**Lineage sorting:** Process of particular genetic variants going extinct over evolutionary time in a population through random drift.

**Locus:** The site that a gene or molecular sequence occupies on a chromosome (plural loci).

**Maternal:** Relating to the mother.

**Microsatellites:** Tandem repeats of short DNA motif's (non-coding DNA).

**Migration:** The mass directional movement of large numbers of individuals of a species as part of their life history.

**Miocene:** Geological time scale (epoch) from 23.8-5.3 million years ago.

**Mismatch distribution:** Distribution of the pairwise nucleotide differences among all DNA sequences in a sample, used for determining historical demographic change.

**Molecular clock:** Consistent accumulation/loss of mutations at a locus that occurs at the DNA level.

**Molecular drive:** A process of DNA ‘turnover’.

**Molecular systematics:** Use of DNA sequence data to characterise relationships among organisms.

**Morphology:** Shape, form, external structure or arrangement of an organism.

**Multidimensional scaling:** Multivariate statistical method that represents the multidimensional similarity of samples in two or three dimensions.

**Mutation:** Alteration in the arrangement or amount of genetic material of a cell.

**Mutation/drift equilibrium:** The increase in genetic variation in a population through mutation is offset by the reduction in genetic variation due to random drift.

**Mutation rate:** The frequency with which new mutations arise in a population.

**Natural selection:** Change in gene frequencies due to differences in individual fitness.

**Neutral markers:** DNA sequences that evolve solely due to genetic drift and mutation.

**Neutral Theory:** Theory proposed by M. Kimura to account for the high level of genetic variation in populations; most point mutations are selectively neutral.

**Niche:** The position or role of a plant or animal species within its community or environment.

**Nitrogenous bases:** Bases that code for variation at the DNA level.

**Non-coding:** Region of DNA that is not transcribed and translated.

**Nucleotides:** Building blocks of DNA molecules.

**Nucleotide diversity:** Measure of genetic variation, the mean number of base pair differences in a sample.

**Null alleles:** An allele that produces no functional product, a sequence that is not amplified in PCR-based analysis because a variation in the DNA sequence annealing to the 3' end of a primer results in nonamplification of the expected segment.

**Outbreeding:** Breeding with individuals from another sub-population.

**Pangaea:** Supercontinent (all continents joined) existing ~ 225 million years ago.

**Panmixia:** A single gene pool, no barriers to gene flow.

**Pentose sugar:** Part of DNA 'skeleton'.

**Permutation test:** Statistical procedure that repeatedly randomises a data set to create a null distribution for testing the significance of an parameter estimated from the data.

**Phenotype:** The physical manifestation of a genotype, eg. the colouration pattern of a fish.

**Philopatric:** Returning to the natal site to reproduce.

**Phosphate group:** Part of DNA 'skeleton'.

**Phylogenetic:** Relating to the hypothesised evolutionary relationships of individuals, populations or species.

**Phylogeography:** Analysis of genealogy, population genetics or evolution within a geographical context.

**Point substitutions:** Mutations at a single base pair site.

**Polygenic:** Refers to a trait or phenotype whose expression is the result of the interaction of numerous genes.

**Polymorphic:** The occurrence of different forms, stages, or types in individual organisms or in organisms of the same species, independent of sexual variations.

**Polyphyletic:** The term for a group of organisms, when despite their being classified together as one taxonomic category, it is thought that not all have descended from a common ancestor.

**Polyploidy:** The situation of cells or individuals having additional complete sets of chromosomes.

**Ploidy:** Number of sets of homologous chromosomes.

**Population bottlenecks:** Severe reduction in population size that reduces population genetic variation.

**Post hoc:** Planned after the fact.

**Plate tectonics:** Movement of continental plates.

**Pleistocene:** Geological time scale (epoch) from 1.8 million – 10,000 years ago.

**Primer:** A short oligonucleotide fragment from where nucleotide extension is initiated during PCR.

**Proof reading:** Process of scanning new DNA strands for replication "errors".

**Purifying selection:** Process of removal of deleterious alleles from gene pools.

**Recolonisation:** The arrival of a number of individuals to re-establish a population that had gone extinct.

**Refugia:** Places of suitable habitat generally surrounded by inhospitable habitat.

**Relative fitness:** Measure of relative reproductive success of different phenotypes.

**Restriction enzymes:** Prokaryote enzymes that cut DNA strands.

**Rheotaxis:** Active dispersal response when subject to a current; positive – against the current; negative - with the current.

**Ribosome:** Site of protein synthesis in cells.

**River capture:** Drainage rearrangement where a stream flowing into one river system is diverted (captured) by an adjacent system that has a higher erosional rate.

**Segregating site:** Nucleotide position in a series of homologous DNA sequences where a mutation has occurred.

**Semi-conservative replication:** Where new DNA molecules are synthesised using an 'old' strand as a template.

**Selection coefficient:** The relative fitness of a particular genetic variant(s).

**Selectively neutral:** Not affected by natural selection.

**Sex-biased dispersal:** Dispersal predominantly by one sex.

**Simple sequence repeats:** Microsatellites.

**Sink:** A population that accumulates genetic variation through gene flow from several other source populations.

**Somatic:** Of cells of the body as opposed to germ cells.

**Stochastic:** Involving chance or probability.

**Stock:** A group of organisms that shares the same genetic and demographic parameters.

**Stream profile:** 2 dimensional cross section of a stream reflecting elevation.

**Substrate:** Ground or other solid surface on which animals walk on or are attached to.

**Temporal:** Related to time.

**Transcription:** Process of encrypting DNA gene message onto a 'carrier' molecule (mRNA).

**Transition:** Point mutation where a purine base is replaced by another purine (A or T) or a pyrimidine is replaced by another pyrimidine (C or G).

**Translation:** Decoding of mRNA into polypeptide.

**Transversion:** A point mutation where a purine (A or T) is replaced by a pyrimidine (C or G) or vice versa.

**Vicariance:** Separation.

$\chi^2$ **Test:** A test that uses the chi-square statistic to test the fit between a theoretical frequency distribution and a frequency distribution of observed data for which each observation may fall into one of several classes.

**Zygote:** Offspring (2n) that results at fusion of egg (n) and sperm.

Also see additional glossary in “Glossary of biotechnology and genetic engineering”, FAO Rsearch and Technology Paper, No.7 at: http: //www.enaca.org/modules/wfdownloads/singlefile.php?cid=63&lid=769

# Bibliography

**De Silva, S. S. (2004).** Fisheries in inland waters in the Asian region with special reference to stock enhancement practices. Bangkok, Thailand.

**Dodson, J.J., Colombani, F. & Ng, P.K.L. (1995).** Phylogeographic structure in mitochondrial DNA of a South-East Asian freshwater fish, *Hemibagrus nemurus* (Siluroidei, Bagridae) and Pleistocene sea-level changes on the Sunda shelf. *Molecular Ecology* 4: 331-346.

**Eknath, A. E. & Doyle, R. W. (1990).** Effective population size and rate of inbreeding in aquaculture of Indian major carps. *Aquaculture* 85: 293-305.

**Gupta, M. V. & Acosta, B. O., eds. (2001).** Fish genetics research in member countries and institutions of the International Network on Genetics in Aquaculture.

**Kamonrat, W. (1996).** Spatial genetic structure of Thai silver barb *Puntius gonionotus* (Bleeker) population in Thailand. Halifax, Canada: Dalhousie University.

**Petr. T. (ed.) (1998).** Inland fishery enhancements. Papers presented at the FAO/DFID Expert Consultation on Inland Fishery Enhancements. Dhaka, Bangladesh, 7–11 April 1997. *FAO Fisheries Technical Paper. No. 374*. Rome, FAO. 1998. 463p.

**Kimura, M. and Crow J. F., (1964).** The number of alleles that can be maintained in a finite population. *Genetics* 49: 725-738.

**Rhymer, J. M. & Simberloff, D. (1996).** Extinction by hybridisation and introgression. *Annual Review of Ecological Systematics* 27: 83-109.

**Senanan, W., Kapuscinski, A. R., Na-Nakorn, U. & Miller, L. (2004).** Genetic impacts of hybrid catfish farming (*Clarias macrocephalus* x *C. gariepinus*) on native catfish populations in central Thailand. *Aquaculture* 235: 167-184.

**Waters, J. M., Lopez, J. A. & Wallis, G. P. (2000).** Molecular phylogenetics and biogeography of galaxiid fishes (Osteichthyes: Galaxiidae): dispersal, vicariance and the position of *Lepidogalaxias salamandroides*. *Systematic Biology* 49, 777-795.

**Welcomme, R. L. & Vidthayanon, C. (2003).** *The impacts of introduction and stocking of exotic species in the Mekong basin and policies for their control*. Cambodia: Mekong River Commission.

www.enaca.org