CMFRI

Course Manual

Winter School on

Recent Advances in Breeding and Larviculture of Marine Finfish and Shellfish

30.12.2008 -19.1.2009

Compiled and Edited by

Dr. K. Madhu, Senior Scientist and Director, Winter school

&

Dr. Rema Madhu, Senior Scientist and Co-ordinator

Central Marine Fisheries Research Institute

Central Marine Fisheries Research Institute

(*Indian Council of Agricultural Research*)

P.B.No.1603, Marine Drive North Extension,

Ernakulam North ,P.O.

Cochin, KERALA – INDIA - 682018

APPLICATION OF MOLECULAR GENETIC MARKERS IN FISHERIES & AQUACULTURE

46

A.Gopalakrishnan,
Principal Scientist, National Bureau of Fish Genetic Resources (NBFGR), Cochin Unit, CMFRI Campus, Kochi, Kerala - 682 018.
E -mail : nbfgrcochin@eth.net

Introduction

Establishment of genetic markers is the prerequisite for stock structure analysis. The markers can detect genetic variations and they can be explained and analysed within the limits of genetic principles. Based on their mode of transmission and evolutionary dynamics genetic markers can be categorized into (**1**) protein markers such as allozymes and (**2**) DNA markers such as (2.1) mitochondrial DNA and (2.2) nuclear DNA markers like randomly amplified polymorphic DNA (RAPDs) and variable number of tandem repeats (VNTRs) loci such as minisatellites and microsatellites.

Types of molecular markers and their principles

All organisms are subject to mutations as a result of normal cellular operations or interactions with the environment, leading to genetic variation (polymorphism). In conjunction with selection and genetic drift, there arises genetic variation within and among individuals, species, and higher order taxonomic groups. For this variation to be useful to geneticists, it must be (1) heritable and (2) discernable to the researcher, whether as a recognizable phenotypic variation or as a genetic mutation distinguishable through molecular techniques. At the DNA level, types of genetic variation include: base substitutions, commonly referred to as single nucleotide polymorphisms (SNPs), insertions or deletions of nucleotide sequences (indels) within a locus, inversion of a segment of DNA within a locus, and rearrangement of DNA segments around a locus of interest. Through long evolutionary accumulation, many different instances of each type of mutation should exist in any given species, and the number and degree of the various types of mutations define the genetic variation within a species. DNA marker technology can be applied to reveal these mutations. Large deletions and insertions (indels) cause shifts in the sizes of DNA fragments produced upon digestion by restriction enzymes, and are among the easiest type of mutations to detect, mainly by electrophoresis of the fragments on an agarose gel; smaller indels require DNA sequencing or more elaborate electrophoretic techniques to determine smaller changes in size. Inversions and rearrangements that involve restriction sites can be easy to detect because they disrupt the ability of a restriction enzyme to cut DNA at a given site and thus can produce relatively large changes in DNA fragment sizes. Point mutations are more difficult to detect because they do not cause changes in fragment sizes. Several marker types are highly popular in genetics. In the past, allozyme and mtDNA (restriction fragment length polymorphism (RFLP)) markers have been popular genetics research. More recent marker types that are finding service in this field include, mtDNA sequence information, randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), microsatellite, single nucleotide polymorphism (SNP), and expressed sequence tag (EST) markers.

Type I versus type II markers and polymorphic information content (PIC)

Molecular markers are classified into two categories: type I are markers associated with genes of known function, while type II markers are associated with anonymous genomic segments. Under this classification, most RFLP markers are type I markers because they were identified during analysis of known genes. Likewise, allozyme markers are type I markers because the protein they encode has known function. RAPD markers are type II markers because RAPD bands are amplified from anonymous genomic regions via the polymerase chain reaction (PCR). AFLP markers are type II because they are also amplified from anonymous genomic regions. Microsatellite markers are type II markers unless they are associated with genes of known function. EST markers are type I markers because they represent transcripts of genes. SNP markers are mostly type II markers unless they are developed from expressed sequences

(eSNP or cSNP). Indels are becoming more widely used as markers since they often are discovered during genomic or transcriptomic sequencing projects; they can be either type I or type II markers depending on whether they are located in genes. The significance of type I markers was not fully appreciated in the early stages of aquaculture genetics, though it is becoming clear that these markers are extremely important. In addition to their functions as markers in population studies, type I markers are becoming very important in studies of genetic linkage and QTL mapping. Type I markers have utility in studies of comparative genomics, genome evolution, candidate gene identification, and enhanced communication among laboratories. Due to evolutionary constraints on the genome, many genes and their organization are conserved among species. Comparative genomics deals with the similarity and differences found among genomes. Much time, money, and effort can be saved in developing markers for use in aquaculture genetic studies if genetic information is already available for closely related species. To date, full understanding of aquaculture genomics depends heavily on information from well-studied species such as human, mouse, and zebra fish. Type I markers serve as a bridge for comparison and transfer of genomic information from a map-rich species into a relatively map-poor species. Such interspecific comparisons can also be made based on type II markers, but the extent to which the comparison can be made is limited to closely related taxa. The requirement for such comparisons lies in sequence conservations. For the most frequently used microsatellite markers, such comparative studies depend on conservation of the flanking sequences used for the design of PCR primers. In contrast, sequence conservation within genes are high, allowing type I markers to serve as anchor points for genomic segments to be compared among species. For instance, if 15 genes are located between type I markers A and B in zebra fish, it is likely that the majority of the 15 genes also reside between markers A and B in catfish, even though the exact number of genes, gene order, and orientation are not necessarily identical. Currently, large insert bacterial artificial chromosome (BAC) libraries are already available for several leading fish species in aquaculture genomics including channel catfish (http://bacpac.chori.org/catfish212.htm; Quiniou *et al*., 2003), tilapia (Katagiri *et al.,* 2001), Atlantic salmon (http://bacpac.chori.org/salmon214.htm), and rainbow trout (Thorgaard *et al.,* 2002). In general, type II markers such as RAPDs, microsatellites, and AFLPs are considered to be non-coding and therefore selectively neutral. Such markers have found widespread use in population genetic studies whose characterizations of genetic diversity and divergence within and among populations are based on assumptions of Hardy–Weinberg equilibrium and selective neutrality of the markers employed. Type II markers also have proven useful in genetics for species, strain and hybrid identification, in breeding studies, and more recently as markers linked to QTL. The usefulness of molecular markers can be measured based on their polymorphic information content (PIC). PIC refers to the value of a marker for detecting polymorphism in a population. PIC depends on the number of detectable alleles and the distribution of their frequencies, and equals 1 minus the sum of the square of all allele frequencies. For instance, the PIC of a microsatellite marker with two alleles of frequency 0.5 each should be $1-[(0.5)^2+(0.5)^2] = 0.5$, while PIC for a microsatellite marker of two alleles with allele frequencies of 0.9 and 0.1 is 0.18. Thus, the greater the number of alleles, the greater the PIC; and for a given number of alleles, the more equal the allele frequencies, the greater the PIC. Comparison of PIC values can give researchers a rough idea of the power of the various marker types discussed below to address specific questions in genetics.

Allozymes

Isozymes are functionally similar and separable forms of enzymes encoded by one or more loci. Isozyme products of different alleles at the same locus are termed as *allozymes*. The most important quality of allozyme data is the co-dominant nature of inheritance of gene products and thus genetic interpretation (genotype) of the phenotype is facilitated because all products are normally visible and not masked by dominance of one over another. Other advantages include function of most of the proteins are known and extensive database is available for many fish species. Allozyme electrophoresis has been used in defining genetic markers for stock identification on the basis of differences in allelic frequencies between stocks in many species. Using allozyme markers, it is possible to determine whether a population is a random mating one with equilibrium genotypes frequencies or sample comprises of an assembly of genetically distinct units. Their allele frequencies primarily respond to mutation, gene flow and drift. One of the limitations of enzyme variants as genetic markers is the low level polymorphism observed in some species and populations. The extensive allozymes studies undertaken on fish stocks have not only proven valuable for estimating population divergence, but also have focussed attention on the underlying evolutionary forces that promote differentiation.

Sarcoplasmic proteins (water-soluble proteins)

The soluble proteins of the sarcoplasm, located within the sarcolemma are referred to as sarcoplasmic proteins. Among them, some albumins and so called myogens; to which belong most of the glycolytic enzymes are the real water-soluble proteins. (The other fractions of sarcoplasmic proteins are soluble in low salt concentrations). The genetic differences between species are more pronounced in this than in other group of proteins, as they are responsible for widely divergent enzymatic transformations in the muscle cell. Hence, the separation patterns of profiles obtained on electrophoresis or isoelectric focusing (IEF) can be used for the unequivocal identification of the species.

Myofibrillar proteins (salt-soluble proteins)

They are salt soluble proteins present in the myofibrils of the muscle fibre. Of the different myofibril proteins, myosin and tropomyosin find application in fish species identification by electrophoresis. Fish myosin, similar to myosin of other vertebrates, is a hexameric protein consisting of two identical heavy chains and four light chains, of which two of them are identical. Electrophoretic pattern of heavy chains from different species are similar whereas that of the light chains is different for different species. Hence, an electropherogram of myosin light chain isolated from fish muscle is used for species identification. Electrophoresis of most of the fish muscle tropomyosin gives a single band whose electrophoretic mobility is different for different species. Tropomyosin is a heat stable protein that can be extracted from heat-treated fish products, thus useful in identifying the species of fish of the product by studying the SDS-electrophoretic pattern of tropomyosin.

Eye-lens Proteins

The soluble proteins of the eye-lens have great value in taxonomic studies, because they are synthesized only one cell type present in the eye as a single layer. Three saline soluble eye lens proteins are distinguishable by electrophoretic and immunological techniques. There are alpha, beta and gamma crystallines in order of decreasing electrophoretic mobilities, each of which constitutes a family of similar, but no identical proteins. Protein with alpha-crystallin characteristics have been found in all vertebrate species and regarded as a classical organ-specific protein. The beta- and gamma- crystallin patterns are species-specific and can be used to resolve taxonomic disputes using ultra-thin IEF technique.

Isoelectric Focusing (IEF) of proteins

Isoelectric focusing uses a polyacrylamide gel with large pore size containing a mixture of polyamino, poly carboxylic acids with different isoelectric point (pI) s. These form a stable pH gradient along the gel in an electric field. Strong acid applied at the anode and strong base in the cathode contain and stabilize the gradient. Proteins migrate under the influence of their charge until they reach the point in the gel where the pH is equivalent to their iso-electric point and so their charge is neutralized. At the iso-electric point, proteins in the electrical field do not migrate to either of the poles. High resolutions are achieved permitting separation of proteins differing only by 0.01 pI and the technique using muscular and eye lens proteins is highly useful in generating species-specific profiles of both finfish and shellfish.

Two - dimensional (2D) electrophoresis

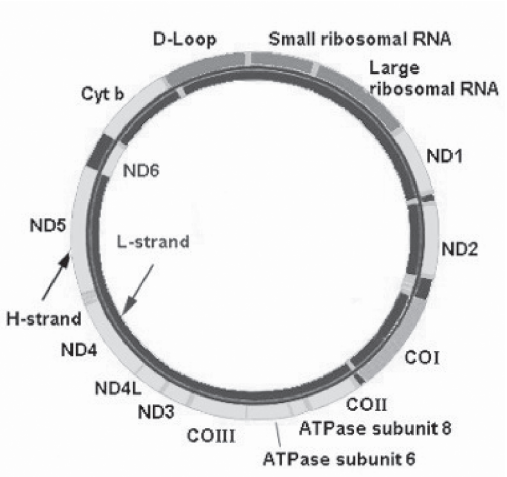
The techniques of isoelectric focusing and polyacrylamide gel electrophoresis have been combined to produce two-dimensional separation of proteins. This technique is increasingly used now a days and its great resolving power is due to the use of two independent properties of proteins. The proteins are first separated by isoelectric focusing (this is the first dimension), which separates proteins according to their charge (isoelectric point). The proteins are subsequently separated by SDS-PAG electrophoresis (this is the second dimension) at right angles, which separates proteins according to their size (molecular weight). This technique results in a series of spots distributed throughout the polyacrylamide gel.

DNA markers

The development of DNA amplification using the polymerase chain reaction (PCR) technique has opened up possibility of examining genetic changes in populations over the past 100-years or more even using archive material. In PCR reaction, a DNA sequence can be amplified many thousand folds to provide sufficient product for restriction analysis or direct sequencing. Once appropriate primers are available, large number of individuals can be assayed quickly thus facilitating large population screening for variability.

Mitochondrial DNA (mt DNA)

The mitochondrial genome is a small and double stranded circular DNA molecule. It is haploid *i.e.* each mitochondrion contains only one type of mt DNA which is cytoplasmically inherited, these making it predominantly maternally transmitted. It is non-recombinant because there is little or no paternal contribution of mtDNA in organisms and no recombination have been reported. It has high mutation rate as compared to single copy nuclear DNA (scnDNA). These factors in combination reduce the effective population size for mt DNA to one fourth in comparison to nuclear DNA. Small effective population size results in greater genetic differentiation between isolated gene pools, making it an attractive marker for studying population specificity. Mitochondria provide the primary source of cellular ATP in eukaryotes via the process of oxidative phosphorylation. In animals, extranuclear mitochondrial genomes are typically circular, and with few exceptions, code for 13 subunits of the oxidative phosphorylation machinery as well as genes for two rRNA subunits and 22 tRNAs. Mutations in mitochondrial DNA (mtDNA) have a number of known deleterious effects. At least 50 base substitutions and hundreds of insertion/deletion mutations have been identified in human mtDNA, with effects ranging from degenerative diseases to aging to cancer. In addition to their role as the powerhouse of the cell, mitochondria are also involved in regulating programmed cell death (apoptosis) and mutagenic reactive oxygen species are generated in the process of energy production. Pathologies can result directly from the loss of ATP production in affected tissues, the build-up of oxygen radicals due to downstream blockage of the oxidative phosphorylation pathway, or unregulated apoptosis. Hundreds of mitochondria and thousands of mtDNAs are inherited maternally through the cytoplasm of the oocyte. If a zygote receives more than one form of mtDNA (heteroplasmy), different forms can be randomly distributed to daughter cells during cell division and, over many cell generations, can drift to high or low frequencies in various cell lineages. Thus, if one of the mutant forms is deleterious, disease may affect lineages where it reaches sufficiently high frequency. Somatic mutations in mtDNA appear to behave similarly and may be a significant source of mitochondrial disease. Given the central role of mitochondria in cell physiology, mutations (either inherited or somatic) are probably responsible for many developmental abnormalities. Although a high frequency of mutant mtDNA molecules is likely to be lethal during embryogenesis, oocytes with moderate to low levels of heteroplasmy occur at detectable levels. Mitochondrial mutations probably affect a number of both general and tissue-specific developmental processes; however, the role of mitochondria in early development has not been



well characterized. Structurally, most animal mitochondrial genomes contain the same 37 genes, and among vertebrates the gene order is highly conserved. Vertebrate mitochondrial genomes are typically ~16 kb and are extremely compact with no introns and few, if any, intergenic spacers. The only significant non-coding sequence is the control region, which is involved in regulating transcription and replication and is usually <5% of the total genome size. Since its endosymbiotic origin around 1.5 billion yr ago, a substantial fraction of original mitochondrial genes have moved to the nucleus. The products of many of these genes remain essential for oxidative phosphorylation or housekeeping functions and are selectively transported into the mitochondria after translation in the cytoplasm. A result of this evolutionary trend is that some mitochondrial abnormalities are due to mutations in genes that now reside in the nucleus and are inherited in a Mendelian fashion, rather than through the maternal, haploid inheritance of mtDNA. Knowledge of gene location (and thus mode of inheritance) is therefore essential for accurate characterization of the developmental–genetic basis of mitochondrial abnormalities. Mitochondrial genomes from several species have now been sequenced. Their small size and relative autonomy from the nucleus makes mitochondrial genomes valuable windows on the process of genome evolution and with respect to cytonuclear interactions. Mitochondrial sequences have also proven to be of great utility in molecular phylogenetic studies, and complete genome sequences have provided valuable insights into deeper-level phylogenetic problems.

In contrast to the largely independent sources of information from different allozyme loci, the mitochondrial DNA molecule is effectively a single locus with composite genotypes equivalent to alleles. Earlier studies of mt DNA variation required large tissue samples and time consuming protocols but use of mt DNA probes and PCR amplification of selected regions have made examination of mt DNA much faster and easier. However, the contribution of mitochondrial studies to stock concept in fisheries although similar to allozymes in nature has been less informative overall.

Nuclear DNA Markers

Random Amplified Polymorphic DNA (RAPD)

The principle behind Randomly amplified polymorphic DNA (RAPD) analysis is that at low annealing temperatures or high magnesium concentrations, a primer is likely to find many sequences within the template DNA to which it can anneal. Depending on the length and complexity of genome of an organism, there can be numerous pairs of these sequences and they will be arranged inversely to and within about two kilobases of each other. Considering this, PCR will amplify many random fragments that can vary in sizes when different species, subspecies, populations or individuals are analysed and this will constitute the basis of identification.

RAPD analysis has several advantages. These include relatively shorter time (1-2 days) required to complete analysis after standardization; no need of prior information on the genome of an organism; availability of series of primers for analysis; minimal operational cost requirement; relatively smaller amount (»20 ng) of high molecular weight DNA; simpler protocol and involvement of non-invasive sampling for tissue analysis. However, the application and interpretation of RAPD – PCR in population genetics is not without technical problems and practical limitations. The main negative aspect of this technique in is the necessity of extensive standardization to obtain reproducible results. In addition, most of the RAPD polymorphisms segregate as dominant markers and individuals carrying two copies of an allele cannot be distinguished from individuals carrying one copy of an allele. The limited sample size in each population and the specific RAPD primers utilized can also have an influence over the results. Even then RAPD technique is used in microbes, plants and animals for resolving taxonomic ambiguities and stock identification.

RAPD-PCR technique can also generate species-specific, sex-specific and population specific fragments. These fragments are useful in developing specific "**Sequence Characterized Amplified Region** (**SCAR**) **Markers**". For this, SCAR primers need to be synthesized from specific RAPD fragments. Usually, fragments above 1000 bp and less than 300 bp are not considered to develop SCAR markers owing to difficulties arising from co-migration and the lesser possibility of designing suitable primers from smaller fragments. The identified fragments are excised from the gel, purified and sequenced; and based on the sequence information, suitable SCAR primers are synthesized. These primers will amplify only specific fragments that are useful in settling taxonomic disputes and identifying sex or distinct

populations. However, to identify specific RAPD fragments, screening of large number of samples and RAPD primers are required.

Multilocus fingerprinting

Protein or mtDNA markers are based on changes in DNA sequence generally as a result of point mutations involving base substitutions. Recently attention has turned to another type of variation that of differences in number of repeated copies of a segment of DNA called Variable Number of Tandem Repeat (VNTR) loci. On multilocus DNA fingerprinting, the length variation is surveyed at many VNTR loci simultaneously. Due to the large number of loci examined and the extremely variable nature of this particular class of repeated DNA, each profile of bands (the so called "fingerprinting") is usually highly informative and individual specific. Numerous probes are available that hybridize to different VNTR loci processing similar repeat unit sequences. Even probes that cross-hybridize in distant taxa are also available. However, multilocus fingerprinting is not usually the method of choice for population level applications as the DNA profiles are often very complex and it is usually not possible to estimate allelic frequencies, necessary in many population analyses. In addition, they often do not generate reproducible results. Even under carefully controlled conditions, the intensity and presence of bands can also vary between gels.

Minisatellites and single locus VNTR profiling

Minisatellites are tandemly repeated DNA, consisting of shorter repeat units (10 – 64 base pairs) which are repeated from two to several hundred times at a locus. In single locus VNTR profiling, allelic variation is surveyed at individual minisatellite loci using two methods. The first method involves restriction endonuclease digestion of genomic DNA, separation of fragments by electrophoresis through agarose gels and southern blotting onto DNA binding membranes. Membranes are then probed with denatured labelled DNA from a single VNTR locus, preferably the unique flanking region. The second method is to PCR amplify the locus using primers flanking the array. The PCR products are separated by standard gel electrophoresis and visualized. Single locus minisatellite profiling is commonly used for maternity and paternity assessment of offspring but in population applications it is much more difficult to estimate the possible range of allele sizes likely to be encountered. Also intra-allelic duplication or deletion, inter-allelic recombination and gene conversion occur at some minisatellite loci causing the variations highly complex. Due to this, devising realistic mutation modes for such loci will be exceedingly difficult.

Microsatellites

Microsatellites are repeated DNA sequences having a unit length of 1-6 base pairs tandemly repeated minimum 6 times usually; maximum several times at each locus. They are also known as "Short Tandem Repeat (STR)" DNA or "simple sequence". Individual alleles at a locus differ in the number of tandem repeats of the unit sequence owing to gain or loss of one or more repeats and they as such can be differentiated by electrophoresis according to their size.

There are four types of microsatellites

1. Perfect : Perfect tandem repeat sequences.
2. Imperfect : Tandem repeat sequences with intervening sequences.
3. Compound : More than one kind of repeats, adjacent ones.
4. Complex : More than one kind of repeats, with intermediary sequences.

Based on the number of base pairs in a repeat unit, microsatellites can be again classified into mono (*e.g.* C or A), di (*e.g.* CA), tri (*e.g.* CCA), tetra (*e.g.* GATA) repeat unit microsatellites. The most common ones are dinucleotide repeats. Though they are widely employed in stock identification studies, appearance of stutter bands often cause problems in scoring alleles. Tetra-nucleotide microsatellites are gradually replacing dinucleotide loci as the preferred genetic marker for stock analysis.

Properties of Microsatellites: Several features of VNTR render them invaluable for examining fish population

structure. Microsatellites are codominant in nature and inherited in Mendelian fashion, revealing polymorphic amplification products from all individuals in a population. They contain information, which are directly related to the effective number of alleles at each locus. PCR for microsatellites can be automated for identifying simple sequences repeat polymorphism. Small amount of samples of blood or alcohol preserved tissue is adequate for analysing them. Because they are highly variable in nature, abundant variants are ensured for characterisation of populations. However, sample size in excess of 50 may be required to represent the genotype frequencies. The microsatellites are non-coding and therefore variations are independent of natural selection. These properties make microsatellites ideal genetic markers for defining heterozygosity, genetic diversity and distance measures.

Applications of Microsatellites: They are very abundant, so sufficient markers can be readily developed for any research objective. Some microsatellite exhibit high levels of allelic variation, thus can be used for species that show low overall level of variation with other markers like allozymes or mitochondrial DNA and populations that are inbred or have experienced several bottlenecks and recently derived or geographically proximate populations where genetic differentiation may be limited. They are also used in pedigree analysis. Not all microsatellites display extremely high levels of allelic variation. It may be from di-allelic to more than a dozen. Thus selection of microsatellite markers can be done for any given research problem e.g. low number of alleles i.e. 3 to 5 for population studies whereas with more alleles for aquaculture genetic studies.

Expressed sequence tags (ESTs)

Expressed sequence tags (ESTs) are single-pass sequences generated from random sequencing of cDNA clones. The EST approach is an efficient way to identify genes and analyze their expression by means of expression profiling. It offers a rapid and valuable first look at genes expressed in specific tissue types, under specific physiological conditions, or during specific developmental stages. ESTs are useful for the development of cDNA microarrays that allow analysis of differentially expressed genes to be determined in a systematic way, in addition to their great value in genome mapping. For genome mapping, ESTs are most useful for linkage mapping and physical mapping in animal genomics such as those of cattle and swine, where radiation hybrid panels are available for mapping non-polymorphic DNA markers. A radiation panel is composed of lines of hybrid cells, with each hybrid cell containing small fragments of irradiated chromosomes of the species of interest. Typically, the cells from species of interest are radiated to break chromosomes into small fragments. The radiated cells are unable to survive by themselves. However, the radiated cells can be fused with recipient cells to form hybrid cells retaining a short segment of the radiated chromosome. Characterization of the chromosomal break points within many hybrid cell lines would allow linkage and physical mapping of markers and genes. In spite of its popularity in mammalian genome mapping, radiation hybrid panels are not yet available for any aquaculture species. Development of radiation hybrid panels from aquaculture species is not expected in the near future, given the fact that physical mapping using BAC libraries can provide even higher resolution and the fact that BAC libraries are already available from several aquaculture species. Therefore, ESTs are useful for mapping in aquaculture species only if polymorphic ESTs are identified. Additionally, ESTs can be mapped to physical maps by hybridization, and integration of physical and genetic linkage maps would in turn anchor the ESTs to the linkage maps. Likewise, ESTs can be mapped to genetic linkage maps if they are found to be associated with microsatellites. In this context, microsatellite-containing ESTs are rich resources of type I markers.

Single nucleotide polymorphism (SNP)

Single nucleotide polymorphisms or SNPs (pronounced “snips”) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. Such sequence differences due to base substitutions have been well characterized since the beginning of DNA sequencing in 1977, but the ability to genotype SNPs rapidly in large numbers of samples was not possible until the application of gene chip technology in the late 1990s. SNPs are again becoming a focal point in molecular marker development since they are the most abundant polymorphism in any organism, adaptable to automation, and reveal hidden polymorphism not detected with other markers and methods. For example a SNP might change the DNA sequence A**A**GGCTAA to A**T**GGCTAA. For a variation to be considered a SNP, it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement

of cytosine (C) with thymine (T). SNPs can occur in both coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but scientists believe others could predispose people to disease or influence their response to a drug. Theoretically, a SNP within a locus can produce as many as four alleles, each containing one of four bases at the SNP site: A, T, C, and G. Practically, however, most SNPs are usually restricted to one of two alleles (most often either the two pyrimidines C/ T or the two purines A/G) and have been regarded as bi-allelic. Obviously, their PIC is not as high as multi-allele microsatellites, but this shortcoming is balanced by their great abundance. SNP markers are inherited as co-dominant markers.

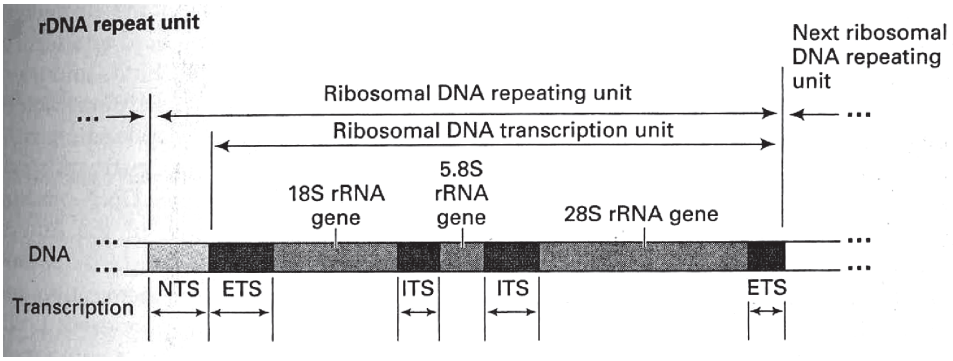
SNPs have properties and a density in the human genome that makes them attractive as markers or tools for identification of genes in as yet uncharacterized parts of the genome that may have some relation to a specific disease. There are great expectations that SNPs will be useful in identifying candidate genes that contribute to population-wide, polygenic diseases. At present several initiatives are ongoing to exploit the information contents of genetic variability. Their purpose is to (1) identify genes that contribute to disease, (2) identify gene targets for development of new therapeutic principles and (3) identify genes that may predict outcome from therapy.

DNA sequencing has been the most accurate and most-used approach for SNP discovery and genotyping. Random shotgun sequencing, amplicon sequencing using PCR, pyrosequencing and comparative EST analysis are among the most popular sequencing methods for SNP discovery. Each approach has its advantages and limitations, but all are still useful for SNP genotyping, especially in small laboratories limited by budget and labor constraints. Despite technological advances, SNP genotyping is still a challenging endeavor and large-scale analysis however, depends on the availability of specialized, expensive and cutting-edge equipments. Another consideration is the expense of genotyping in relation to sample sizes. Microarray (gene chip) technology and quantitative PCR are particularly useful in medical and clinical settings where large numbers of samples (thousands of individuals per locus) are involved and that can justify the cost involved in the development of the gene chips and hybridization probes.

The exploitation of SNPs is likely to proceed in several phases: In the first phase a sufficiently dense map of SNPs will be created which will eventually cover the entire genome. The physical location of SNPs will be determined in a similar way to micro-satellite markers. SNPs will, at least initially, be selected based on how informative they might be as genetic markers. SNPs in low proportions (<10 %) will be less informative than SNPs at higher frequency (30-50 %) in a given population. In a second phase, the relative frequencies of SNPs covering a large portion of the human genome will be correlated to specific diseases by comparing allelic frequencies in healthy and diseased populations. This information will focus further analysis to a smaller part of the genome thus continuously increasing the resolution of useful SNPs. In a third phase genetic variability will be studied in more detail utilizing SNPs. Typically, a limited number of genes (10-100) which link with specific diseases will be investigated for genetic variability. In this phase genetic variability present at a lower frequency is likely to be more informative. Genes contributing to disease are expected to be identified in an iterative process. The majority of SNPs are likely to be located in non-coding regions (SNP) but SNPs located in the coding regions (cSNP) will also be detected. The identification of SNPs located in coding regions of genes (cSNPs) will be particularly important since they may represent a genotype/phenotype relationship in specific diseases.

ITS (Internal Transcribed Spacers)

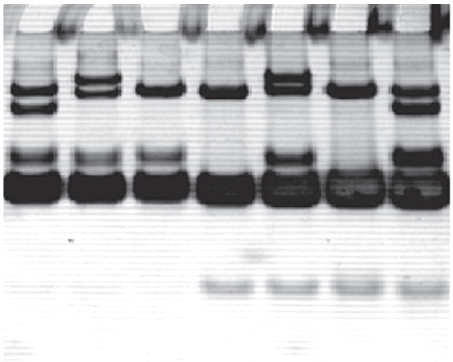
Eukaryotic ribosomal RNA genes (known as ribosomal DNA or rDNA) are found as parts of repeat units that are arranged in tandem arrays, located at the chromosomal sites known as nucleolar organizing regions (NORs). Each repeat unit consists of a transcribed region (having genes for 18S, 5.8S and 28S rRNAs and the external transcribed spacers i.e. ETS1 and ETS2) and a non-transcribed spacer (NTS) region. In the transcribed region, internal transcribed spacers (ITS) are found on either side of 5.8S rRNA gene and these are described as ITS1 and ITS2.The length and sequences of ITS regions of rDNA repeats are believed to be fast evolving and therefore may vary. Universal PCR primers designed from highly conserved regions flanking the ITS and its relatively small size (600-700 bp) enable easy amplification of ITS region due to high copy number (up to-30000 per cell) of rDNA repeats. This makes the ITS region an interesting subject for evolutionary/phylogenetic investigations as well as biogeographic investigations.



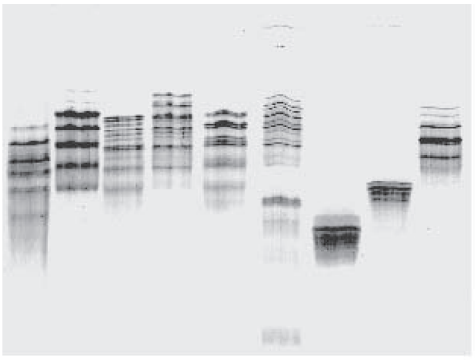
Chloroplast DNA (cp DNA)

Chloroplasts are replicative organelles and contain a number of copies of a double-stranded circular DNA chromosome which is also known as **cpDNA.** The number of copies of this chromosome in each chloroplast varies between cells; 20 – 30 in old leaves to 100 in young leaves. Chloroplast chromosomes lie within the stroma and a number of features of their structure resemble prokaryotic chromosomes. They are circular DNA molecules which, unlike nuclear chromosomes, are not complexed with histones. Replication of the chloroplast genome and its distribution between daughter proplastids is a complex and ill-defined process. Chloroplasts genomes have been classified into 3 types. Two groups of land plants namely the gymnosperms, Pinaceae and a group of legumes (including peas and broad bean) have chloroplast chromosomes without an inverted repeat – IR (Group I). Most land plants, including all other angiosperms have chloroplast genomes containing a large (6 – 76 kb) inverted repeat; these are Group II genomes. The alga *Euglena* has three tandem repeats in its Group III chloroplast genome.

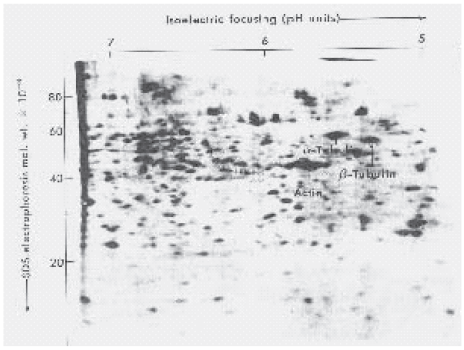
Chloroplasts are maternally inherited, that is there is essentially no transmission of chloroplasts through the male pollen gamete; thus the mode of inheritance of cpDNA is uniparental and the genome is typically non-recombinant and effectively haploid. The cpDNA of plants has been a focus of research in plant molecular evolution and systematics. Several features of this genome have facilitated molecular evolutionary analyses. First, the genome is small and constitutes an abundant component of cellular DNA. Second, the chloroplast genome has been extensively characterized at the molecular level providing the basic information to support comparative evolutionary research. And third, rates of nucleotide substitution are relatively slow and therefore provide the appropriate window of resolution to study plant phylogeny at deep levels of evolution. Chloroplast genomes contain between 120 – 140 genes and of ~160 kbp size (120×10^3 to 200×10^3 base pairs in higher plants and 180×10^3 base pairs in green algae). It is identified as one of the relatively stable genome with marked conservation of gene content and a substantial conservation of structural organization. Conservation of gene content and a relatively slow rate of nucleotide substitution in protein-coding genes has made the chloroplast genome an ideal focus for studies of plant evolutionary history. This has led to determining the DNA sequence of the cpDNA gene *rbcL* encoding the large subunit of ribulose-1, 5-bisphosphate carboxylase (RuBisCo). Unlike the animal mitochondrial genome, cpDNA of several plants contain introns. The functional



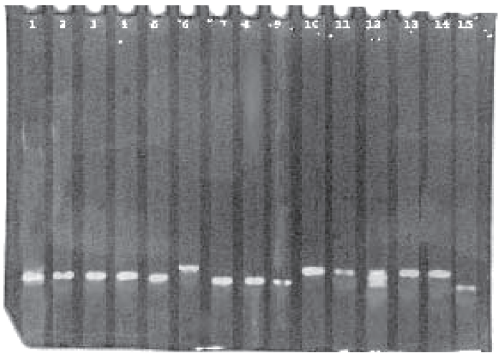
Allozyme (Esterase) pattern in PAGE



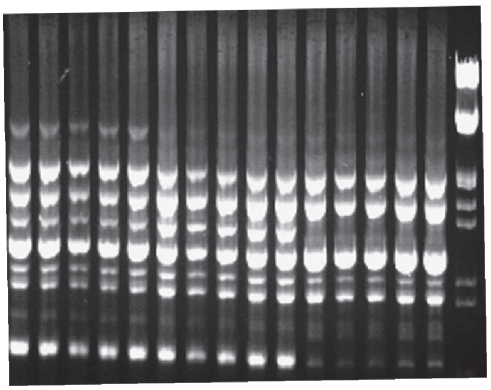
Ultra-thin IEF of fish haemoglobin



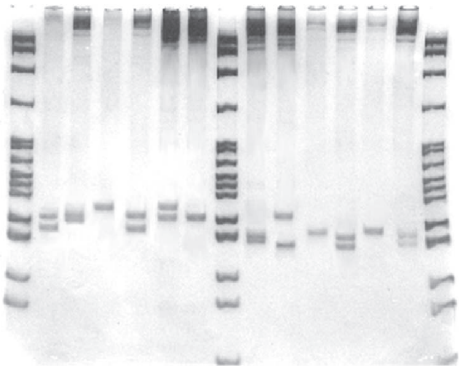
2D gel electrophoresis of frog oocytes
(IEF and SDS PAGE at right angles)



Allozyme (SOD) pattern in PAGE



Agarose (1.5%) electrophoresis RAPD
pattern of fish DNA with Operon primer

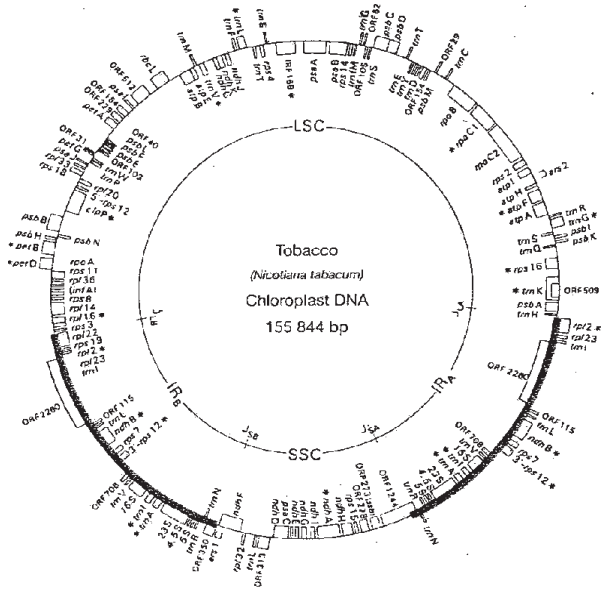


Microsatellite pattern of fish DNA
in PAGE with silver staining.

Images of different types of gel electrophoresis.

categories of cpDNA are (i) DNA regions that do not code for tRNA, ribosomal RNA (rRNA) or protein (referred to as "noncoding DNA"); (ii) protein-coding genes; and (iii) chloroplast introns. The chloroplast is highly condensed compared with eukaryotic genomes; for example, only 32% of the rice genome is non-coding. Most of this non-coding DNA is found in very short segments separating the functional genes. The genetic information for the synthesis of much of chloroplast encoded proteins is present in the chloroplast genome, with four genes for rRNA present in the large inverted repeat (IR). The other cp protein synthesis genes identified include genes for ribosomal proteins, 30tRNA genes and a RNA polymerase gene together with protein synthesis coupling and elongation factors.

The chloroplast genome shares many features with animal mtDNA and the two have been referred to as 'natural counterparts'. Its conserved gene order, the widespread availability of primers and a general lack of heteroplasmy and recombination, have made the chloroplast genome an attractive tool for phylogenetic studies of plants. Furthermore, its uniparental mode of inheritance (usually maternal in angiosperms and paternal in gymnosperms) makes it possible to elucidate the relative contributions of seed and pollen flow to the genetic structure of natural populations by comparing nuclear and chloroplast markers. Regions in cpDNA such as *trH-psbA* spacer, *rbcL*, *matK*, *rpoC*1, *rpoB*, *accD* and *YCF*5 are identified as the most promising regions in the cpDNA for DNA barcoding in plants and universal primer pairs for these regions have been developed (barcoding@kew.org). An interesting feature of cpDNA is the occurrence of polymorphic mononucleotide microsatellites (cpSSR) that are increasingly used in population genetics and understanding crop plant evolution and domestication. Unlike, nuclear microsatellites, cpSSRs are uniparentally inherited (some species have maternal inheritance of the chloroplast and others paternal), nonrecombinant and all loci are linked. The genotyping of cpSSRs will result in haplotypes that will be composed of the combination of alleles found at each cpSSR locus. Chloroplast microsatellites are fast evolving and typically consist of mononucleotide motifs that are repeated 8 to 15 times. Levels of polymorphism in cpSSRs are quite variable across loci and across species, and some loci have been found to be monomorphoic in all species study. Many studies have demonstrated high levels of intraspecific variability of cpSSRs; hence, they represent potentially useful markers at the population level in plants.



Gene map of the tobacco chloroplast genome. Genes shown inside the circle are transcribed clockwise, genes on the outside are transcribed anticlockwise. Asterisks denote split genes. The major open reading frames are included. IRF, intron-containing reading frame; IR, inverted repeat; LSC, large single-copy region; SSC, small single-copy region; J, junctions between IR and LSC and SSC. (From Sugiura, M. (1992) The chloroplast genome. *Plant Molecular Biology*, 19, 149–168.)

☆☆☆☆☆