Choosing the right Molecular Genetic Markers for studying biodiversity: from molecular evolution to practical aspects

Anne Chenuil

Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects

Chenuil Anne
Centre d'Océanologie de Marseille Laboratoire DIMAR, UMR CNRS 6540-Université de la Méditerranée Chemin de la batterie des Lions, 13007, Marseille, France (E-mail: chenuil@com.univ-mrs.fr; Phone: +33-0491041617; Fax: +33-0491041635)

Key words: biodiversity, choice, molecular evolution, molecular marker, practice, substitution rate

Abbreviation: AFLP – Amplification fragment length polymorphism; ARRF – Anonymous rare-cutter restriction fragments; CAPS – Cleaved amplified polymorphic sequence; DALP – Direct amplification of length polymorphism; DGGE – Denaturing gradient gel electrophoresis; ETS – External transcribed spacer (of rDNA); FISH – Fluorescent *in situ* hybridization; ILP – Intron length polymorphism; ISSR – Inter-simple sequence repeat; ITS – Internal transcribed spacer; MGM – Molecular genetic marker; ORF – Open reading frame; PCR – Polymerizing chain reaction; RAPD – Random amplified polymorphic DNA; RFLP – Restriction fragment length polymorphism; RSCA – Reference Strand Conformation Analysis; rDNA – Ribosomal DNA; SNP – Single nucleotide polymorphism; SSCP – Single-strand conformation polymorphism; ssDNA – Single-strand DNA

Abstract

The use of molecular genetic markers (MGMs) has become widespread among evolutionary biologists, and the methods of analysis of genetic data improve rapidly, yet an organized framework in which scientists can work is lacking. Elements of molecular evolution are summarized to explain the origin of variation at the DNA level, its measures, and the relationships linking genetic variability to the biological parameters of the studied organisms. MGM are defined by two components: the DNA region(s) screened, and the technique used to reveal its variation. Criteria of choice belong to three categories: (1) the level of variability, (2) the nature of the information (e.g. dominance vs. codominance, ploidy, ...) which must be determined according to the biological question and (3) some practical criteria which mainly depend on the equipment of the laboratory and experience of the scientist. A three-step procedure is proposed for drawing up MGMs suitable to answer given biological questions, and compiled data are organized to guide the choice at each step: (1) choice, determined by the biological question, of the level of variability and of the criteria of the nature of information, (2) choice of the DNA region and (3) choice of the technique.

Introduction

The first markers used for genetic analysis were morphological traits transmitted by mendelian inheritance. More simply, any character genetically determined is considered as a genetic marker. Molecular genetic markers (MGMs) directly reflect the variation at the level of DNA.

Rapid technical as well as theoretical advances greatly modified the range of tools available for the study of biodiversity (Waser and Strobeck, 1998; Luikart and England, 1999; Sunnucks, 2000)

and, despite the amount of literature available (Avise, 1994; Dowling et al., 1996; Carvalho, 1998; Féral, 2002; Zhang and Hewitt, 2003), it may be difficult for those not familiar with molecular tools, population genetics or phylogenetic concepts to choose the right one. Methods detailed in this paper are those which allow one to reveal hitherto unknown variants, and are potentially applicable to any taxon (i.e., for which DNA sequences are not available), excluding therefore all diagnostic approaches *sensu lato* (e.g., fluorescence in situ hybridization, FISH; or single nucleotide polymorphisms, SNPs; (Amann et al., 1995; Kwok and Chen, 2003)).

To be properly made, the design of a MGM (or a set of MGMs) must follow three successive steps which correspond to the frame of the paper (Figure 1): (1) choice of the level of variability and of the criteria of the nature of information, (2) choice of the DNA region and (3) choice of the technique. The first step consists in determining the criteria to be fulfilled by the MGM in order to answer the biological question asked. These criteria can be separated in two categories, first, the level of variability, second, all criteria concerning the nature of the information (e.g., dominance/codominance, recombination, ...). Then, appropriate MGMs should be designed according to these criteria. Often, MGMs are confused either with a technique (e.g., single strand conformation polymorphism, SSCP), or with a DNA region (e.g. mt-DNA), but considering the DNA region and the technique as independent and complementary components of MGM definition is necessary to properly organize the information for guiding the choice of appropriate MGMs. In effect, the choice of the DNA region, which is the second step of MGM design, is the main determinant of the level of variability and also determines some key features of the nature of the information (ploidy, inheritance, availability of a database), whereas the choice of the technique, the third step, determines the other features of the nature of the information (codominance, possibility of assessing

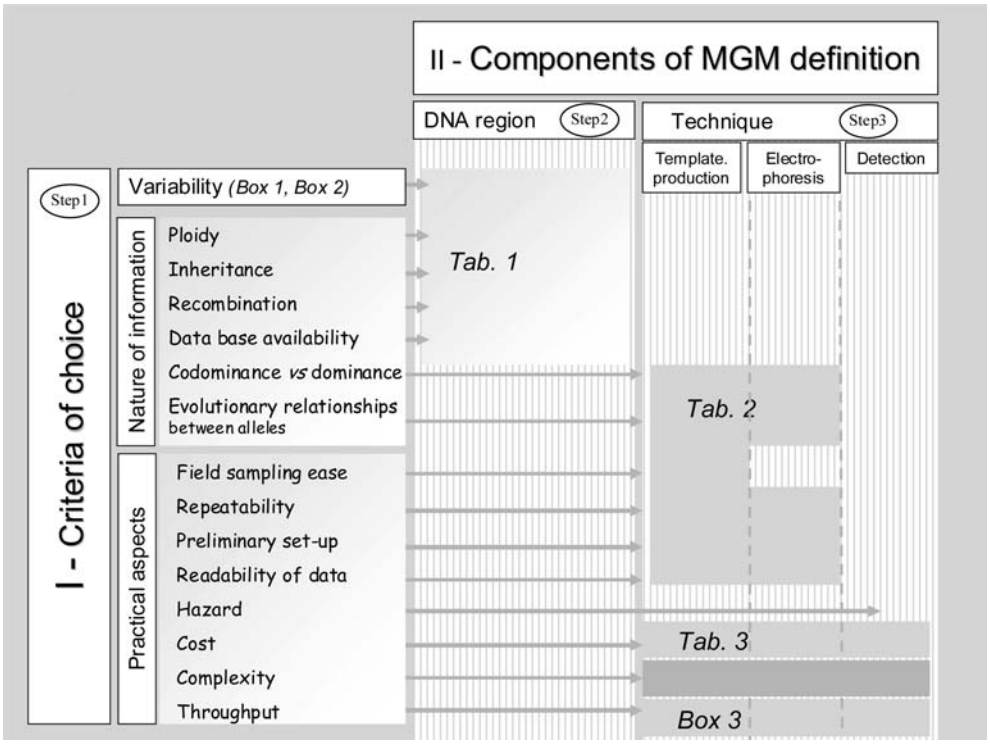

Figure 1. Flowchart diagram explaining how to use the paper for designing MGMs. It links the three categories of criteria of choice (first part of the paper) with the components of MGM definition, DNA region and techniques (second part of the paper). Relevant figures, tables and boxes are indicated. The three successive steps in the process of MGM design appear in circles. The grey arrows mean: ''is (mostly) determined by,'' for example, Ploidy and Inheritance are determined by the DNA region, ''Codominance (or not)'' is determined by the technique.

evolutionary relationships among alleles). I emphasize, and this will be demonstrated in the paragraph about the nature of the information, that using a combination of different types of MGMs is synergistic. Though this paper is not aimed at providing protocols, some ''bench'' details, which are not available in general reviews and appeared decisive in the building of the MGM, are given.

The criteria (first step)

Population genetics theory allows deducing biological parameters from genetic marker data. Table 1 gives an overview of the most common questions addressed using MGMs. Historically, the mathematical relationships, which were first used, were derived under the equilibrium assumption (i.e., the parameter value at generation

Table 1. Examples of classical biological questions at different biodiversity level, with the corresponding properties requested for MGMs about level of variability and nature of information, and most used markers

Biological issues/ biodiversity level	Level of variability	Nature of information required	Examples of most used markers
Intra-population			
Fine population structure, reproduction system, selfing rate	Medium to high	(N) codominant loci = (Multilocus) genotype[1]	Microsatellites, allozymes
Fingerprinting, parentage analysis	Very high	Codominant loci or numerous dominant loci[2]	Microsatellites (RAPD, AFLP)[2]
Demography (estimation of N_e)	Medium to high	Allele frequency in samples taken at different times[3]	Allozymes, Microsatellites
Demographic history	Medium to high	Allele frequency + evolutionary relationships[3]	Mt-DNA sequences
Inter-population			
Phylogeography, definition of evolutionary significant units (population structure)	Medium to high	Allele frequency in each population[3]	Allozymes, microsatellites (risk of size homoplasy)
Bioconservation	Medium	*But preferable with knowledge of:* Allele evolutionary relationships	Mt-DNA (if variable enough)
Inter-specific			
Close species	ca. 1%/my	Many characters. No variability within species if possible	Sequences of mt-DNA, ITS rDNA, ...
Different genera to families....	ca. 0.1%/my	Idem	Some LSU[4] rDNA domains (D1 < D2, D8), but also mt-DNA or SSU rDNA (Table 2)
Different classes to phyla	ca. 0.01%/my	idem	D1 of LSU rDNA, SSU rDNA sequences

[1] To compare observed proportions of heterozygotes to those expected assuming Hardy–Weinberg equilibrium, allowing us to detect departure from Hardy–Weinberg equilibrium, due to population admixture, non-panmixia or selection (mutation is negligible). Comparison among independent loci distinguishes patterns due to migration, which similarly affect all loci from those due to selection.
[2] One dominant marker yielding a high number of polymorphic fragments (each corresponding to a dominant locus) may provide finer resolution (exclusion probabilities) than few codominant loci (Gerber et al., 2000) when one parent is known.
[3] Methods using Multilocus genotypes are still less employed than monolocus ones though they are powerful for studying population admixture, migrant numbers, and demographic variations (Waser & Strobeck, 1998; Davies et al., 1999; Vitalis & Couvet, 2001).
[4] LSU: Large sub-unit. SSU: small sub-unit.

Table 2. Origin of the variation and consequences on the nature of the information for different DNA regions. Abbreviations are: "s" for substitution, "id" for insertions/deletions, "Inherit" for inheritance and "Recomb" for Recombination, + to + + + + relate to the size of the database, +/- means it depends upon locus, - means the database is extremely limited but may eventually develop, NH means there is no homology across taxa for these DNA regions thus no possibility of a cross-taxa database

Origin of the variation				Nature of the information			
DNA region	Nature of mutations	Variability[1] μ(mutation/l/g) K (%/my)	Ploidy: copy number	Inherit.	Recomb.	Database[2]	Reference
Autosomes			$2N$	M & F	Yes		
18S rDNA	S > > id	K = 0.006–0.04	$2N$	M & F	Yes	+ + + +	Sorhannus (1996)
28S rDNA domains	S, id	According domain				+ +	Qu (1986), Pélandakis & Solignac (1993)
ITS of rDNA	S, id	K = 0.15–0.4 %/my				+ + +	Després et al. (1992), Linder et al. (2000)
ETS of rDNA	id > S	Highest				+ (plants)	Linder et al. 2000
Exons (e.g. Allozymes[3])	S	Average K_s = 0.35–1.6 [4] Average K_a = 0.07–0.2[4]	$2N$	M & F	Yes	+/– +/–	Graur & Li 2000
Introns, non-coding	S, id	K = 0.32	$2N$	M & F	Yes	–	Graur & Li 2000
Microsatellites	id (2–6 bp)	μ = 10^{-4} [5]	$2N$	M & F	Yes	Rare	Ellegren (2000), Estoup & Angers (1998)
Minisatellites	id (10–200 bp)	μ = 0.4–7.10^{-2}	$2N$	M & F	Yes	NH	Bois (1999)
X chromosome (or Z)		< autosomes[6]	$2N_f + N_m$	M & F	Localized		
Y chromosome (or W)		> autosomes	N_m	M	Localized		
ZfX–ZfY exon	S	K_s = 0.7–3.5[7]				Mammals	Pamilo and Bianchi (1993)
ZfX–ZfY Intron	S, id	K < 0.1				Mammals	Slattery and O'Brian (1998)
*Sr*Y	S, id	K = 0.3				Mammals	Pamilo and O'Neill (1997), Nagai (2001)
CHD1Z-CHD1W		cf.reference				Birds	Fridolfsson and Ellegren (2000)
Microsatellites	id	μ = 3.10^{-3} (in Y chro.)				NH	Kayser et al. 2000
Mitochondrial genome ...							
... In Animals			N_f	F	No	+ + +	
Protein coding genes (all)	S	K_s = 2–3 K_a = 0.5–3.5	N_f	F	No	(variable)	Pesole et al. 1999

D-loop ETA	id > S	$K = 2$	N_f	F	No	Vertebrates,	Pesole et al. 1999, Crochet & Desmarais (2000)
D-loop Central		$K = 0.4$					
D-loop CSB		$K = 1.4$				Arthropods	
COI	S	$K = 0.2$ (annelids) $K = 1.5$ (sea urchin)	N_f	F	No	Vertebrates, Echinoderms, Arthropods, Molluscs, Annelids, ...	Chevaldonné et al. (2002) Lessios et al. (1999)
Cyt b	S	$K = 0.35$–1.4	N_f	F	No		Caccone et al. (1997)
16S	S > id	$K = 0.5$ (mammals) $K = 0.19$ (newts)	N_f	F	No		Caccone et al. (1997)
... In Plants	id > S	id: high s: very low	Variable	F or M	Rearrange[8]	+–	Palmer et al. (2000)
Chloroplast genome			Variable	Variable	No		
RbcL	S	Low (but variable)	Variable	Variable	No	++	Martin & Dowd (1991)
Introns, Spacers	S, id	= IGS = 3x *rcbL*	Variable	Variable	No	+	Gielly & Taberlet (1994)
Microsatellites		$\mu < 3$–8×10^{-5}	Variable	Variable	No	NH	Provan et al. (1999)
Random PCR/RFLP	?	High	$2N$	M & F	Yes	NH	

[1] Estimates are taken from references of last column. For micro- and minisatellites, mutation rates (μ) are given in the number of mutation per generation (g) and per locus (l). In other cases, nucleotidic substitution rates (K) are given (for coding DNA, synonymous (K_s) or non-synonymous (K_a)), in % substitution per million year. When no group is specified, estimates are from mammals, or *Drosophila*.

[2] The symbol "–" refers to nearly empty databases which may grow, "No" means that no database grouping homologous data from a variety of divergent taxa is possible. Only protostome and deuterostome phyla represented by several genera are listed but other phyla (e.g. sponges, cnidarians) form mitochondrial databases.

[3] Enzymatic electrophoresis only reveals about 1/3 of the differences existing in protein sequences.

[4] Lower (respectively higher) value is the average of 45 genes in mammals (resp. 32 in *drosophila*). Some exons have a large database (EF1, cyt c, histones) mostly in mammals. K_a values may exceed 0.8%/my (Ref. i of Box 1), but most genes (not undergoing positive selection) have $K_a < 0.35\%$/my.

[5] Estimates vary greatly among species. Generally dinucleotides are more variable than trinucleotides and tetranucleotides.

[6] There is evidence for a selectively favourable reduction in the mutation rate of the X chromosome in rodents (Ref. b of Box 1).

[7] Substitution rates are often higher in the male than female germ line even for "homologous" loci (Box 1).

[8] Sequences are rearranged but no recombination *sensu stricto* (i.e., sex) occurs.

n equals its value at generation $n + 1$) (e.g., F_{ST} and F_{IS} statistics). Though F_{ST} were used to infer genetic distances among populations, genetic data were generally not translated in quantitative estimates of biological parameters (e.g., selfing rates from F_{IS}, or migrant numbers from F_{ST}) but rather used to detect a phenomenon, or compare its strength among populations or species: for example, (i) a limit to gene flow between two populations is revealed by a significantly non-null F_{ST} or exact tests on allele distribution among populations, and (ii) the fact that a population is not at Hardy–Weinberg equilibrium is evidenced by a significantly non-null F_{IS} or relevant exact tests, suggesting either inbreeding or internal structure of the population considered. Then, technical progress and lowering costs of sequencing allowed us to obtain numerous DNA sequence data even for intra-specific studies. With the opportunity to infer genealogical relationships between variants (or alleles) and the development of the coalescent theory, it became easier to detect non-equilibrium processes and a variety of models may be built to estimate several biological parameters simultaneously (Templeton, 1998; Davies et al., 1999). All MGMs are not equally suitable to make different types of biological inferences. Two classes of criteria must be considered, the variability, and the nature of the information given.

Variability: origin and quantification at the DNA level

According to the level of biodiversity under study, a given level of variability of the marker is required. High to very high levels of variability are required for intra-population purposes (e.g., parentage analyses, reproductive systems, demographic history). Medium to high variability is adequate when distinct populations are compared (e.g., phylogeography, definition of evolutionary significant units). Phylogenetic studies require moderate to very low variability (ca. 1% per million year (%/my) for close species, ca. 0.1%/my when distinct genera/families are compared and ca. 0.01%/my for inferring relationships between different classes or phyla; inferred from Table 2). Excessive variability may lead to homoplasy (i.e., coexistence of identical variants of independent evolutionary origins).

Understanding how evolutionary forces (mutation, selection, drift and migration) create and remove variability at the DNA level (Box 1) helps us to choose the right molecular marker. Data of primary (nucleotide sequence) and secondary (folding of single strand DNA) structures of DNA sequences may give information on their variability. For instance, repetitive sequences are more mutable and therefore provide more variable markers. Moreover, if the sequence is apparently incompatible with protein coding frames (e.g., presence of long dinucleotide repeats, stop codons), there is an increased probability that selective constraints are weak, so that less mutations are eliminated by selection and more mutations contribute to polymorphism.

There are two fundamental classes of variability measures from MGM data, polymorphism (*e.g.* expected heterozygosity H_e, or its equivalent for haploid data, "haplotypic diversity") (which estimation only requires to know the frequency of all variants), and substitution rates K (which estimation requires sequence data) (Box 1). Nucleotide diversity (Pi) combines these two types of measures, using both sequence data and allele frequency data. Theory tells us that in the absence of selection (neutrality) these values only depend on the mutation rate of the DNA region characterized by the genetic marker and eventually (for polymorphism and diversity) on the effective size (Box 1). Mutation rates are generally unknown, but estimates can be inferred from neutral markers (Box 1). Two studies on fish provide a nice illustration of the double influence of the effective size of the population and the mutation rate of the DNA region on the level of polymorphism. Expected heterozygosity, H_e, is significantly smaller in freshwater than anadromous, and anadromous than marine fish populations, either with allozymic markers (Ward et al., 1994) (H_e are respectively 0.046, 0.052 and 0.059) or with microsatellite loci (DeWoody and Avise, 2000) (respective H_e : 0.54, 0.68 and 0.77); and H_e from allozymes were significantly smaller than from microsatellites. In practice, these measures of variability are simply deduced from the data given by the MGM (Box 1).

Information on evolutionary rates is useful for choosing an MGM because the relative evolutionary rates of different molecules are usually conserved across lineages. For example, small subunit rDNA of nuclear genomes always evolves about two orders of magnitude more slowly than 16 S mt-DNA, its mitochondrial counterpart

Table 3. Correspondence between the techniques and some features of the nature of the information and practical aspects. Technical pathways are given with symbols: E (extraction), L (ligation), D (D1: permanent denaturation using heat and chemical factor, D2: denaturation by heat then on ice, D3: denaturation by heat and slow cooling) PCR, SEQ (sequencing), RE (restriction enzyme digestion), MP (magnesphere purification), EP (electrophoresis), SBH (southern blotting plus hybridization). Nature of information symbols are: Cod (codominance), D (dominance), NR (non-relevant). For several practical aspects, marks from A (best case) to D (worst) are used. Set up indicates time and complexity to obtain markers and determine routine experimental conditions. Id: same as above

Technique Based on the method of separation of variants	Technical pathway (before detection)	Nature of information			Practical aspects		
		Codom. or Dom.	Evolutionary relationships betw. Alleles	Quality required for DNA extract[1]	Reliability Repeatability (critic. phase)	Set up difficulty	Data readability
DNA sequencing							
	E-PCR-SEQ-D1-EP	(NR)	YES	Standard	A	B	A
DNA conformation							
SSCP	*E-PCR-D2-EP*	Cod	Possible if each	Standard	B–C (*EP*)	C	B–C
DGGE	*E-PCR-EP*	Cod[2]	variant is sequenced	Id.	Id.	Id.	Id.
Duplex-heteroduplex	*E-PCR-D3-EP*	Id.	Id.	Id.	Id.	Id.	Id.
DNA fragment Length							
CAPS	*E-PCR-RE-EP*	Cod	NO	Standard	A–B	A	A–B
RFLP	*E-RE-EP-SBH*	D/Cod [3]	NO [9]	HMW Digest	B–D	A	B–C [4]
Microsatellites	*E-PCR-(D1)-EP*	Cod	YES [5]	Standard	B–C[6,7]	D (2–8 months)	B–C (small motif)
Length polymorphism (non-repetitive)	*E-PCR-(D1)-EP*	Id.	NO	Id.	A–C[7]	B Find locus	B–C
RAPD	*E-PCR-EP*	D [8]	NO [9]	Ct MMW	C–D (*E–PCR*)	B[10]	C–D
ISSR	*E-PCR-EP*	Id.	Id.	Id.	B–C (*E–PCR*)	Id.	C
AFLP	*E-RE-L-PCR1-2-(D1)-EP*	Id.	Id.	Ct Digest MMW	Id.	Id.	C–D
DALP	*E-PCR-EP*	Id.	Id.	Ct MMW	B (*E–PCR*)	Id.	C
ARRF	*E-D-L-MP-EP*	Cod	NO	High amounts	? (*E–PCR*)	C ?	B? (too recent)

[1] HMW'': High-molecular weight DNA, ''MMW'': medium molecular weight (DNA not too degraded), ''Ct'': constant quality among individuals, ''Digest'': DNA must be digestible.

[2] Often both homoduplexes of a heterozygote comigrate even in a denaturing gel, heteroduplexes are more easily detected (mismatched dsDNA migration is slow)

[3] In general RFLP from genomic DNA were used as dominant data but some minisatellites provide codominant markers. Use of this technique has decreased since microsatellites were developed.

[4] Generally easy (may be difficult if total genomic or organelle DNA), depending on the number and size of fragments, and the ploidy.

Table 3. (continued)

[5] Assuming models of mutation, one may compute relationships between alleles according to their sizes, but homoplasy is greater than for allele relationships deduced from non-repetitious sequence data.
[6] In microsatellites, risks of null alleles are greater than other markers for which primers are systematically designed in coding regions, and polymerase stuttering produces "phantom bands".
[7] There are risks of small allele dominance if allele sizes vary greatly. This affects reliability rather than reproducibility.
[8] Codominant markers can be obtained from fragments after isolation and sequencing, more easily with DALP and AFLP than RAPD since cloning is required when fragments are generated from a single primer.
[9] Distances between individuals or variants can be calculated from the number of shared bands but it is impossible to reconstruct evolutionary relationships between alleles, a necessary step for coalescence approaches.
[10] DALP using longer primers and not requiring digestion of the native extract DNA is less sensitive to DNA extract quality and to PCR conditions than RAPD or AFLP.

(Table 2) for a given species. Whenever possible, validation of evolutionary rates must be performed using independent (palaeontological, geological or biogeographical) information since these rates vary among lineages. Variation also occurs among sites for a given molecule (e.g., 18 S rDNA; Hillis and Dixon, 1991), some sites may be "saturated" (causing homoplasy) for a given species set even though a high proportion of sites are invariant and distances between sequences seem moderate when calculated globally (Tourasse and Gouy, 1997).

The nature of the information

The nature of the information provided by different MGMs is very variable, and the features of the nature of the information which are most desirable vary according to the biological question asked. Six features must be considered.

First, ploidy of the marker, which depends on its genomic localization, is crucial. (i) The effective number of copies is inversely related to the strength of genetic drift; haploid mitochondrial or chloroplastic DNA is more sensitive to genetic drift than diploid nuclear DNA (Box 1), hence it can reveal isolation between populations which occurred four times more recently than a nuclear DNA marker with the same mutation rate (Palumbi et al., 2001). (ii) Unambiguous sequence data are much simpler to obtain for haploid markers: for nuclear DNA regions, obtaining the nucleotide sequence of both alleles of a heterozygote individual requires cloning and multiple sequencing which forbids analysis of large samples, or the combined use of DNA conformation techniques which may not reveal all variants (see below). (iii) Diploid loci are the only ones able to provide the so-called codominant information (see next criterion).

Second, MGM provide either a molecular phenotype, that is, an array of presence /absence data of given fragments (dominant markers) or a genotype, that is, both alleles of diploid individuals are characterized (codominant markers), which is a more precise information. Diploid genotypic data (or codominant data) are necessary to estimate heterozygote deficiency, hence consanguinity.

Table 4. Approximate cost in €uros (for a small european laboratory in 2005) for all steps of most used techniques for 100 individuals. Prices depend upon the size of the laboratory, and the relative prices among techniques evolve rapidly. Facultative steps, proteinase K (alternative to grinding) and post-staining by other products than ethidium bromide (cost of ethidium bromide is negligible) are in italics. I consider that yellow tips are purchased in racks, not bulk, and that molecular weight markers giving regularly spaced fragments every 100 bp are used, eventually in addition to 20 bp spaced fragments, in two or three lanes per gel. Abbreviations "H" and "V" refer to "horizontal" and "vertical" gels. Three methods are compared for DNA extraction: chelex, classical phenol chloroform method and commercial kits (the Nucleon kit (APB biotech)). Yellow tips are included in cost estimation, as well as plastic PCR plates and plastic vials, except for electrophoresis loading where tips are not counted for 100 individuals since they may eventually be re-used after rinsing in the electrophoretic buffer. Using automatic sequencers, fragment size determination may be performed accurately for an approximate cost of 300 € per hundred samples including internal size standards, by private companies. PCR are in a volume of 20 µl

Technical task (for 100 samples)	Cost (€)
DNA Extraction	
Chelex method	60
Classic phenol method (CTAB)	26–30
Industry kit (Nucleon)	100
(alternative to grinding) Proteinase K	1–3
Enzymatic reactions	
PCR (Non-labeled)	25
PCR (Fluorescent "Abi" primers labels)	35
PCR (Radioactive labeling)	100
Sequencing (Industry) small quantities	800–1500
Sequencing (Industry) by plates of 96 samples	400-700
Restriction Digestion	10–16
Electrophoresis	
Routine Agarose 2%, 5 cm long, H	4–7
Routine Agarose 2%, 20 cm long, H	26–30
High resolution Agarose 3%, 20 cm long, H	60
High resolution Agarose 3%, 18 cm long, V	10–16
Polyacrylamide Urea 6%, 40 cm long, V	2–4
Polyacrylamide Urea 8%, 20 cm long, V	2–4
Automatic sequencer fragment analysis (Industry)	300
Post-staining by Gel star – Agarose gels 5 cm	6
Post-staining by Gel star – Agarose gels 20 cm	5–8
Post-staining by Gel star – all vertical gels	4
Yellow tips	4
Size marker (100 bp ladder)	5–8
Size marker (100 + 20 bp ladder)	10–18

Third, inheritance may be biparental (autosomes, X or Z chromosomes, chloroplast DNA in some species, mt-DNA in mussels), paternal (Y chromosome, chloroplast DNA in some plants), or maternal (W chromosome, mt-DNA in animals and many plants, chloroplasts in some plants) (Table 2). Combination in the same study of markers of different inheritance allows, for instance, the comparison of male and female migration (Prugnolle and De Meeûs, 2002) or male and female success in reproduction (Poteaux et al., 1999).

Fourth, recombination may exist (autosomes in most diploid eukaryotes, mt-DNA in plants), or not (species without crossing-over, mt-DNA of animals, part of the X, and Y chromosomes, chloroplastic DNA) in the DNA region characterized (Table 2). In the latter case, theoretical

Box 1. Evolutionary forces at the DNA level and measures of variability

Mutation	Relative importance of mutations of different nature (point, small or large indels) and their rate depend on (i) primary structure of the sequence (repetitive DNA is prone to slippage, unequal recombination and/or conversion), (ii) genome compartment (nuclear, mitochondrial, chloroplast) and (iii) chromosomal position [a b c d]. Direct data on mutation rates are extremely rare, since most estimates are deduced from substitution rates (cf. Theory, below).
Drift	Genetic drift is the random fluctuation in the frequency of genes that is due to the fact that population size is finite. Drift is stronger for smaller populations. According to its genomic localization, a DNA region is more or less strongly subjected to genetic drift for a given population effective size: for N_e reproducing individuals (N_{ef} females + N_{em} males) there is tranmission of $2N_e$ autosomes, ($2N_{ef}$ + N_{em}) X or ($2N_{em}$ + N_{ef}) Z chromosomes, N_{em} Y chromosomes or N_{ef} W chromosomes, N_{ef} mitochondrial genomes (in species with mitochondrial female inheritance), and N_{ef} or N_{em} chloroplast genomes (according to mode of inheritance).
Selection	Most mutations are either neutral (the majority) or deleterious (these ones are rapidly eliminated from the population and do not contribute to polymorphism), rare mutations are favorable (these ones will spread and reach a frequency of 1). Polymorphism is thus created mostly by neutral mutations, but also, in rare cases, by heterosis (the heterozygote genotype is fitter than the homozygotes) or balancing selection (the selective value varies because of a changing environment). Some DNA regions are not subjected to selection (neutral), while others evolve under selective constraints. A posteriori evidence is provided by comparisons of substitution rates (see below), but it is very problematical to identify constrained and neutral sequences a priori (ORF-looking sequences may be recent pseudogenes and non-protein coding sequences may be functional [e]). Pseudogenes on average evolve even more quickly than introns [f]. Constrained DNA sequences generally evolve more slowly than unconstrained ones for a given mutation rate. However, some particular genes such as those involved in biotic interactions (immunity, incompatibility,...) display higher substitution rates, with newly arisen variants being positively selected due to their low frequency (advantage of being rare). Confusion is often made between the role of selection and mutation at the DNA level, since direct assessment of mutation rate is very rare, and "more mutation" or "less purifying selection" have the same effect: increased polymorphism". Polymorphism is not produced by mutation alone, but from the joint action of mutation, drift, selection and migration. One common mistake among biologists ignoring the neutralist theory of evolution consists of the assumption that in DNA regions which have a known function all variants represent particular adaptations, even though the major part of the observed variation at the DNA level is neutral [g] (Box 2).
Migration	MGM from different genomic localizations are not equally sensitive to male and female migration (e.g. if males but not females disperse in an animal species where males are the heterogametic sex, gene frequencies from Y chromosome markers may be identical in all populations, but mt-DNA markers may reveal differentiation). However, one must note that other factors do differ among genomic compartments, and molecular evolution predictions are not straightforward (e.g. the mutation rate may be higher in the male germ line [h]).
Theory:	*Variability (He, K or Pi) expressed as a function of evolutionary forces* A substitution occurs when a mutation becomes fixed in a lineage

– The substitution rate, K, is a function of the distribution of the selective values of mutations s, their rate μ, and the effective size of the population or species, N_e.

– Mutations which are deleterious enough relative to genetic drift ($N_e s > 1$) are eliminated quickly and never reach fixation. Otherwise, some slightly deleterious mutations may eventually become fixed by random genetic drift.

– If μ is the rate of neutral mutations, $K = \mu$. For some non coding regions, all mutations may be neutral, allowing to estimate the mutation rate by the observation of divergence rates.

– Synonymous nucleotide mutations are those that do not change the encoded amino-acid and therefore should often be neutral. For many different genes, synonymous substitution rates (Ks) are very similar (10^{-9}/site/year) whereas non synonymous substitution rates vary widely, suggesting that Ks is actually the mutation rate [i] μ and that favorable mutations are negligible (see below).

– Favourable mutations become fixed at the rate: $K = 4N_e s\mu$ (if μ represents their rate). This formula is not very useful since observable data do not permit inference of the distribution of selective values among mutations.

– The expected polymorphism for a neutral marker also is a simple function of μ and N_e, which may be expressed by H_e , the proportion of heterozygotes expected in an isolated population at Hardy-Weinberg equilibrium: $H_e = 4N_e\mu/(1+4N_e\mu)$.

– The nucleotide diversity for a neutral marker is $Pi = 4N_e\mu$.

Practice: *Variability as a function of observed data from MGM*

Simple formulae allow to estimate K, He and Pi:

– To estimate K from the data, one may divide the evolutionary distance (D) between two sequences by their divergence time (T): $K = D/2T$. To calculate an evolutionary distance, from the observed proportions of differences between two sequences (distinct categories of differences are considered for some models) different models of mutation may be assumed, some of which attempt to correct for saturation, rate variation ...(chapter 5 in Page and Holmes, 1998 [j]).

– Estimating H_e from genetic marker data is straightforward (H_e is the proportion of heterozygotes expected in a population at the Hardy-Weinberg equilibrium displaying the observed allele frequencies):
$H_e = 1 - \Sigma\, p_i^2$. (p_i being the frequency of the i^{th} allele).

– Pi is the mean number of nucleotide differences per site between individuals of the sample.
$Pi = (\Sigma\, x_{ii.})/\, n_c$ (x_{ii} being number of differences per site between i and j, n_c the number of comparisons).

References

a Denver, D. R., K. Morris, M. Lynch, L. L. Vassilieva & W. K. Thomas, 2000. High direct estimate of the mutation rate in the mitochondrial genome of Caenorhabditis elegans. Science 289: 2342-2344.

b McVean, G. T. & L. D. Hurst, 1997. Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. Nature 386: 388-392.

c Nachman, M. W. & S. L. Crowell, 2000. Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297-304.

d Nachman, M. W., 2001. Singe nucleotide polymorphisms and recombination rate in humans. Trends Genet. 17: 481-485.

Box 1. (continued)

e Eddy, S. R., 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2: 919-929.

f Graur, D. & W.-H. Li (2000) Fundamentals of Molecular Evolution. Sinauer Associates, Inc., Sunderland, Massachussets

g Ohta, T., 1992. The nearly neutral theory of molecular evolution. Annu Rev Ecol Syst 23: 263-286.

h Huttley, G. A., I. B. Jakobsen, S. R. Wilson & S. Easteal, 2000. How important is DNA replication for mutagenesis ? Mol Biol Evol 17: 929-937.

i Kimura, M., 1986. DNA and the neutral theory. Phil Trans R Soc Lond B 312: 343-354.

j Page, R. M. & E. C. Holmes (1998) Molecular evolution. A phylogenetic approach., 1st edn. Blackwell-Science, Cambridge

deductions are simplified because there are less unknown parameters.

Fifth, there may be a universal database for an MGM (e.g. sequences of the small subunit ribosomal RNA is known from species of nearly all phyla). In such a case, information is homologous thus comparable to data obtained by the same marker in other taxa, variation of evolutionary rates can be tested and divergence times may eventually be estimated.

Sixth, and very important, evolutionary relationships between variants may be reconstructed (e.g., from sequence data, or, less reliably, from repeat numbers) provided an evolutionary model, depending on DNA region, is assumed. In such cases, data analyses are potentially much more powerful (Templeton, 1998). Furthermore, selective effects can be detected from the analysis of DNA sequences (Yang and Bielawski, 2000; Nielsen, 2001).

It is therefore clear that no ideal MGM exists, because these "ideal" properties are often mutually exclusive. For example, diploid codominant markers are necessary to assess consanguinity, but haploid markers are the best ones to infer evolutionary relationships among variants (since it requires unambiguous sequence information). Choosing a set of MGMs displaying complementary properties relative to the nature of the information (Buonaccorsi et al., 2001) is therefore synergistic. Though for a given type of marker, it is highly recommended to use several physically independent loci (e.g., various microsatellites or random amplified polymorphic DNA (RAPD) fragments), this condition obviously cannot be fulfilled for mt-DNA markers. In many cases, the use of well known mitochondrial regions (for animals) or chloroplastic regions for plants, combined with diploid codominant markers appears as a good solution. Presently, identification of polymorphic codominant markers for new taxa still requires preliminary research but with the growing number of sequenced genomes, more EPIC loci (Exon Primed Intron Crossing), working across high taxonomic levels should become available by identification of conserved intron positions and design of degenerate primers in the flanking exons. When the problem of " heterozygote sequencing" will be resolved (i.e., when the sequence of the two alleles of the same locus mixed as a result of polymerizing chain reaction (PCR) from an

Box 2. Selected versus neutral markers: a misleading distinction

	Markers from coding and non-coding DNA regions (*e.g.* allozymes and microsatellites) are often considered as "selected" and "neutral" markers respectively. This may be misleading because: *i*) microsatellites (or other non protein coding DNAs) may be physically linked to selected loci and in linkage disequilibrium with them, therefore microsatellites may reflect the evolution of a selected region and *ii*) different enzymatic alleles (allozymes) are in general selectively equivalent (cases of balanced polymorphism, where different alleles are favoured in different environments, are rare). Arguing that these two kinds of markers will reflect different evolutionary processes, *i.e.* neutral processes reflecting mutation, drift and migration, or selected processes reflecting adaptation is fallacious. The predictable difference is that in constrained DNA many mutations are eliminated by selection whereas in non constrained DNA all mutations contribute to polymorphism (but a microsatellite locus physically close to a constrained locus will display low polymorphism). For a given mutation rate, polymorphism may thus be higher in non constrained DNA regions (Box 1). Studies suggesting selection in allozymes and not in microsatellites for given species are rare[a] [b].
References	
a	Lemaire, C., G. Allegrucci, Y. Naciri, L. Bashri-Sfar, H. Kara & F. Bonhomme, 2000. Do discrepancies between microsatellite and allozyme variation reveal differential selection between sea and lagoon in the sea bass (*Dicentrarchus labrax*) ? Mol Ecol 9: 457-467.
b	Mitton, J. B., 1998. Molecular markers and natural selection, pp 225-241 in Advances in Molecular Ecology, edited by G. R. Carvalho. IOS Press, Amsterdam.

Box 3. How to increase throughput ?

	Electrophoresis methods and multiplexing are primordial factors.
Automatic sequencers	provide by far the highest throughput of all widespread electrophoresis and detection systems.
Fluorescence scanning	associated with wide electrophoresis plates also allows great time savings.
Some electrophoretic conditions	require much longer migrations than others: SSCP must run at very low power (unlike denaturing gels) and therefore may be about five times slower than separation by length (e.g. 10 hours instead of 2 hours).
	Smaller plates allow more rapid electrophoresis (less resistance, simpler cooling systems) but do not separate variants as well.
Fragment size.	Smaller fragments allow shorter migration time, and easier detection of small absolute size differences. Therefore, it is recommended to choose primers close enough to variable sites (for length or conformation variant separation).
Number of loci per lane.	Multiplexing can consist of mixing several primer pairs in a PCR (which requires careful optimization [a]) or migrating amplicons from several loci in the same lane. Fluorescence technology (scanner or automated sequencer) may allow the use of different fluorophores, allowing loci with overlapping allele sizes to be mixed. Random PCR techniques give multilocus phenotypes: the gain in information quantity may be balanced by time wasted in uneasy reading and interpreting those types of data (mutliple fragments of different intensity).
References	
a	Henegariu, O., N. A. Heerema, S. R. Dlouhy, G. H. Vance & P. H. Vogt, 1997. Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques 23: 504-511.

heterozygote can be determined simultaneously), the use of such powerful markers will spread rapidly.

Practical criteria

The practical criteria are not directly related to the biological questions and can be considered at the end of the process of designing MGMs.

Eight practical criteria are important (Figure 1, Tables 3 and 4, Box 2): (1) ease of field sampling, (2) repeatability of technique, (3) readability of the data (difficulty arise when the distinct DNA fragments of interest display intensity variation), (4) preliminary set-up, before routine conditions are established (requires variable amounts of time and money), (5) manipulation of hazardous products such as radioisotopes and mutagens (generally depends on the detection method), (6) technical complexity of routine typing, once primary set-up has been performed, (7) throughput (the number of samples which can be processed per day per person, once preliminary set-up has been completed) depends on the equipment of the laboratory) and (8) cost. The second part of the paper gives information to allow one to estimate these practical criteria, except for technical complexity, which strongly depends on laboratory equipment and personal preferences.

DNA regions and Techniques available to build your own MGMs

Choice of the DNA regions (second step)

The DNA region is the primary determinant of the variability of the MGM and determines several features of the nature of the information, which are detailed in Table 2. Estimates of variability which are theoretically independent of life history traits (*s.l.*) and contingent factors affecting the populations are more useful for choosing MGMs (Box 1). For this reason, neutral substitution rates or mutation rates, rather than estimates of polymorphism, are given in Table 2. The range of variation among DNA regions is much greater than among lineages. Comparative studies of evolutionary rates of various DNA regions in a given sample of taxa are still rare (Pesole et al., 1999; Rokas et al., 2002). DNA regions suitable for MGMs are known in any genome compartment (nuclear and cytoplasmic) and also, for few taxonomic groups, in sexual chromosomes. Small subunit rDNA was the marker of choice for phylogenetics at high taxonomic levels for a long time but several protein coding genes sequences are now available for a number of highly divergent taxa (Roger et al., 1999; Graur and Li, 2000; Rokas et al., 2002). Several regions of mt-DNA, with contrasting evolutionary rates, are intensively used in a diversity of animal groups (the regions forming the largest databases are reported in Table 2). Though two D-loop domains are famous for being the most rapidly evolving regions, synonymous changes or third codon positions of any mitochondrial gene display a similar variability (Pesole et al. 1999) while being easier to align since insertions and deletions are very rare and are multiples of three bases in mitochondrial coding regions. The 16 S rDNA has the largest database of the low variability mitochondrial regions (i.e., tRNAs and rRNAs). Numerous studies report rate variation between lineages (see Caccone et al. (1997) for vertebrates).

Some approaches reveal polymorphism from random target PCR (RAPD, amplification fragment length polymorphism, AFLP; DALP, ISSR): small primers are used to generate a pattern of presence/absence of fragments of different size, providing dominant markers. Alternatively, the DNA region characterized is *a priori* defined (i.e., between a pair of PCR primers encompassing a known nucleotide sequence). Several regions, coding or not, homologous between highly diverged species, are widely used (Table 2). Introns are particularly interesting since they are probably often selectively neutral and highly polymorphic. Choosing primers in the flanking constrained exon sequences (EPIC PCR) theoretically provides polymorphic markers working in diverged species and not subjected to null alleles (often due to non-binding of PCR primers). Introns often display insertions and deletions which facilitate their genotyping (size variation) and may be highly variable (Ohresser et al., 1997; Bierne et al., 2000). Some intron positions appear conserved across phylogenetically distant organisms or even among phyla (Palumbi, 1996; Jarman et al., 2002; Atarhouch et al., 2003) but these are likely to be under some sort of selective constraint and their polymorphism may be reduced, or they belong to multigenic families, impeeding genotype inference.

By contrast microsatellite loci are generally not conserved between diverging species. They are defined by their composition of tandem repeats of short motifs (one to four bases), and are famous for their high polymorphism.

Choice of the Techniques (third step)

The technique used to detect variation of the chosen DNA region determines two crucial elements of the nature of the information, codominance and the possibility of inferring evolutionary relationships among variants, and the practical criteria (Table 3). Four main phases are usually necessary to obtain the data: preliminary work to define the DNA region(s) to chose (see below), template production (extraction of DNA or allozymes and enzymatic reactions), electrophoresis, and detection. The phases of ''template production'' and ''electrophoresis'' influence the nature of information produced and two practical aspects: repeatability and preliminary set-up (Table 3). Detection methods determine hazards, and influence technical complexity, throughput (Box 3) and cost (Table 4). Main techniques available for MGMs are described by their technical pathway in Table 3. All technical phases are surveyed below, highlighting those which may present particular difficulties, or for which alternative choices correspond to different MGMs.

Phase 1: Preliminary work to define the DNA region used

Before starting the technical work, *sensu stricto*, choosing the DNA region requires searches, either in bibliographical databases or in banks of genes, and aligning DNA sequences in order to choose primers potentially conserved in the studied organism. In the case of microsatellites, determining primers requires previous isolation of sequences containing microsatellites which involves cloning and may be time consuming but may also be obtained from private companies (Zane et al., 2002).

Phase 2: Template production

Enzymatic extraction for allozyme markers requires fresh or frozen tissue. By contrast, DNA extraction, if followed by a PCR step, allows easy and non-invasive field sampling for relatively large organisms, and analysis of very small organisms, since minute amounts of tissue conserved in small volumes of ethanol can be used. DNA extraction can be very rapid and cheap (Chelex method, Walsh et al., 1991) although direct digestion of DNA extract by restriction enzymes and random target PCR methods may require more demanding extraction procedures. In AFLP (Vos et al., 1995), extracted DNA is digested (two restriction enzymes are generally used) and then linked to small adaptors (linkers) before PCR. Random PCR methods may generate relatively large fragments which may not be successfully amplified in DNA extracts which are too degraded (fragments of low-molecular weight). Restriction digestion may be inhibited by several compounds in DNA extracts.

PCR is performed in nearly all recent MGM techniques. For random target PCR, different types of primers may be used. PCR at low annealing temperatures with one short primer (around 10 bases, for RAPD) provides multiband patterns. Mis-priming may limit the repeatability of such PCR (Atienzar et al., 2000) and very constant experimental conditions are required from DNA extraction to detection to allow the comparison of profiles across experiments. In AFLP, primers are longer than in RAPD and correspond to the sequence of the linker, plus one to three bases at the 5′ end. Another method, direct amplification of length polymorphism (DALP, Desmarais et al., 1997) also uses long primer pairs and relatively high-PCR annealing temperatures, its reproducibility is excellent. AFLP provides more polymorphic bands than RAPD and DALP, but requires more steps, potential sources of error. In inter-simple sequence repeat (ISSR), primers are composed of a microsatellite sequence plus eventually one to three arbitrary bases in the 5′ or 3′ direction. Although less widely used, ISSR appear more reproducible than RAPD (probably because primers exceed 12 bases). Random target PCR techniques mostly give dominant markers, but have the advantage of rapidly producing a large number of polymorphic loci (each fragment), which compensate for the missing information in particular cases of parentage analysis (Gerber et al., 2000). After PCR, digestion by restriction enzymes, which generally does not require amplicon purification, may simply provide co-dominant markers, named cleaved

amplified polymorphic sequence (CAPS) (Konieczny and Ausubel, 1993). One may easily identify polymorphic sites by sequencing amplicons from a pool of individuals, and if a restriction enzyme corresponds to a polymorphic site, obtain a CAPS codominant marker (Laporte and Charlesworth, 2001). The advantage of this technique over SSCP (see below) is its reproducibility, the fact that electrophoresis may be run on a 2% agarose gel (easier handling) and the predictability of fragment positions from sequence data (no need to control electrophoretical conditions precisely).

For all MGMs obtained after PCR (particularly microsatellites) there are risks of null alleles and small allele dominance (Wattier et al., 1998). Marine invertebrates seem particularly prone to null alleles (Chenuil et al. (2003), and numerous unpublished reports of non-usable microsatellite loci), which is likely a consequence of their large effective sizes causing high H_e (Box 1). Primers located in coding regions are less prone to null alleles.

Anonymous rare-cutter restriction fragments (ARRF) is the only method providing multiple codominant markers, though allocating fragments to distinct loci may not always be straigthforward (McDonalds, website: *http://udel.edu/~mcdonald/arrf.html*). The procedure is roughly similar to AFLP (digestion and ligation to adaptors) except that the absence of PCR requires much higher DNA quantity, but allows distinguishing homozygotes and heterozygotes by twofold difference in fluorescence. Background is lacking to thoroughly evaluate this promising method.

Phase 3: Electrophoresis

Electrophoresis techniques discriminate variants by (i) their charge and mass (allozymes), (ii) their size in number of base pairs (microsatellites, RFLP, ILP, RAPD, AFLP, sequencing), (iii) their single-strand conformation (SSCP) or dsDNA conformation (reference strand conformation analysis, RSCA), or (iv) their denaturing grading gel electrophoresis profile (DGGE, (hetero)duplex analysis). In cases of heterozygosity at more than one site or for indels, sequencing does not allow to reconstruct diploid genotypes. Cloning is a time consuming solution not suitable to characterize populations. Electrophoresis techniques, which discriminate DNA by conformation differences (SSCP, Orita et al., 1989; Sunnucks et al., 2000; DGGE, Myers et al., 1987; hetero-duplex and duplex analysis, Hauser et al., 1998), potentially reveal variation of any nature, unlike size discrimination electrophoresis. Prior to electrophoresis *sensu stricto*, the denaturation step is crucial. It may be absent (agarose gel electrophoresis, non-denaturing polyacrylamide gels). There is permanent denaturation when DNA is heated in denaturing loading dye and run on denaturing gels (containing urea). Denaturation may be followed either by quick renaturation in ice (SSCP), which results in at least two single strand conformations (one folding for each strand) even for homozygous individuals, or by slow renaturation (duplex/heteroduplex), which allows heterozygote samples to form two heteroduplex dsDNA molecules in addition to the homoduplexes produced by PCR.

Except for capillary automated sequencers, electrophoresis is performed either in agarose gels or in more resolving gels: polyacrylamide gels, denaturing polyacrylamide-urea gels or special conformation sensitive matrices. For non-denaturing gels, voltage is limited by the risk of sample denaturation, dsDNA migrates much faster than single strand DNA (ssDNA; run in denaturing gels containing urea). Though denaturing gels are often used for microsatellites, non-denaturing polyacrylamide gels appear very convenient since they can be run on smaller apparati, are easier to cool and provide medium sized easily post-stained gels (convenient and cost effective alternative to radioactivity, fluorescence or silver staining). Unfortunately there is no precise relationship between migration distance and dsDNA size in polyacrylamide gels (Sambrook et al., 1989). High resolution agarose is promising according to White and Kusukawa (1997) (resolution may attain 2% for a 4% gel), but actually poses many problems at melting, casting and running and is expensive. For SSCP, the samples are run in polyacrylamide or special conformation-sensitive matrix gels. If the fragment is small (about 200–300 bp) typically more than 90% of point substitutions can be revealed. All variables (temperature, voltage, gel composition) influence the position of the variants unpredictably. For duplex/heteroduplex technique and DGGE, samples are run on a gradient gel of increasing denaturing composition (urea). Heteroduplex DNA denatures well before homoduplexes because of mismatches, and therefore gives

higher bands in such gels. Homoduplexes of different sequences also dissociate at different points according to their base composition in a roughly predictable way. About 95% of differences may be detected. For these conformation-sensitive techniques, electrophoresis conditions must be precisely determined if variants are to be compared among gels. In addition, profiles may be complex. In SSCP, even a homozygote produces at least two bands and additional bands are often encountered due to alternative conformations or presence of dsDNA. In duplex analysis or DGGE, a heterozygote may display four different bands. This may explain why SSCP in biodiversity studies is mostly applied to haploid DNA (mt-DNA) rather than used as a codominant marker. In another technique, RSCA, prior to a non-denaturing electrophoresis, PCR products are hybridized to a known homologous reference fragment or several reference fragments (labelled with different fluorescent molecules). The resulting dsDNA migrate according to their homology with the reference strand, that is, heteroduplex is slowed by mismatches compared to homoduplex. The analysis then focuses on dsDNA which have simpler patterns (only one band per allele) and require shorter migration times than for SSCP (Goldman and Madrigal, 1997). Size discrimination techniques may also be tricky when differences among variants are small (Table 3 Box 3) and the addition of size standards in each lane is often useful. In the case of microsatellites, phantom bands smaller (but also larger) than the actual allele by one, eventually two repeats are produced by polymerase stuttering. These fragments are generally less abundant than the actual allele, but may appear as intense if the signal is saturated (e.g., radioactive labeling).

Phase 4: Detection

This step determines the level and type of hazards. Radioactivity as well as post-staining using ethidium bromide or more recent dyes are potentially mutagenic methods. Either labeling (fluorescent or radioactive) or post-staining is used to visualize DNA. Labeling is performed either on the primer or by incorporation during PCR or sequencing. Alternatively, DNA is stained during or after electrophoresis with ethidium bromide, silver nitrate, or more recently, with several dyes more sensitive than ethidium bromide, some of which allow detection of ssDNA and SSCP fragments (e.g., Gelstar from CAMBREX Inc.). Radioactivity is the most sensitive, before fluorescence, silver nitrate, Gelstar and last, ethidium bromide. Oligonucleotide labeling is interesting (i) in RFLP, to allow detection of small molecular weight bands which otherwise would be much less visible than heavier fragments (but internal fragments will not be visualized), and (ii) when it is better to reveal only one DNA strand (e.g., SSCP, when patterns are too complex and some microsatellite loci, when run in denaturing gels, since complementary fragments do not perfectly comigrate). Polyacrylamide gels are seldom post-stained in population studies (though commonly for mutant diagnosis) though this technique avoids the necessity of managing decaying stocks of radioactivity, and the costs of fluorescence technology. Except equipment cost, fluorescence is the most convenient method. Though this service is not as widespread as sequencing, it is possible to send microplates of PCR products (one primer should be fluorescent) to private companies or technological platforms, and pay for fragment size determination (run in automated sequencers with internal size standards).

Methods of latest technology

High throughput methods (based or not on microarrays) progress rapidly but their development is biased towards diagnostic methods revealing already identified variants such as SNPs (Kwok and Chen, 2003). Most methods (not pyrosequencing) require that variation is biallelic and that the polymorphic site is surrounded by several invariant sites. As a consequence, even when several nuclear polymorphic sequences are known in a species (e.g., introns; internal transcribed spacer, ITS; rDNA; exons; ...) it may be difficult to find SNP sites suitable to be characterized by high throughput genotyping methods. When such loci are identified however, a custom genotyping service may now be proposed by industry or technological platforms, and reach competitive prices of the order of magnitude of a euro or a dollar per sample (in addition to the PCR cost).

Conclusion

This paper provides guidelines to choose a (set of) MGM(s) as follows (Figure 1). First, scientists identify the important criteria that must be fulfilled by the MGM according to the biological question addressed (first section, Table 1). Then, Table 2 guides the choice of the DNA region according to the criteria identified (level of variability, and the first four criteria of the nature of the information). Finally, the technique is chosen according to required features concerning the nature of the information and practical aspects, using Table 3 for most important choices, but also Box 3 and Table 4 for throughput and cost appraisal.

Acknowledgements

I am particularly grateful to Erick Desmarais who taught me and gave me information on some techniques, Michele Nishiguchi and Sigurd von Boletzky who thoroughly corrected the English of a previous version and Sara Via, Didier Aurelle, Patrick Berrebi and Yves Desdevises for their comments.

References

Amann, R.I, W. Ludwig & K.H. Schleifer, 1995. Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. Microbiol Rev 59: 143–169.

Atarhouch, T., M. Rami, G. Cattaneo-Berrebi, C. Ibanez, S. Augros, E. Boissin, A. Dakkak & P. Berrebi, 2003. Primers for EPIC amplification of intron sequences for fish and other vertebrate population genetic studies. Biotechniques 35: 676–682.

Atienzar, F., A. Evenden, A. Jha, D. Savva & M. Depledge, 2000. Optimized RAPD analysis generates high-quality genomic DNA profiles at high annealing temperature. Biotechniques 28: 52–54.

Avise, J.C., 1994. Molecular markers, Natural history and Evolution. Chapman & Hall, New-York-London.

Bierne, N., S.A. Lehnert, E. Bedier, F. Bonhomme & S.S. Moore, 2000. Screening for intron-length polymorphisms in penaeid shrimps using exon-primed intron-crossing (EPIC)-PCR. Mol Ecol 9: 233–235.

Bois, P.J.A., 1999. Minisatellite instability and germline mutation. Cell Mol Life Sci 55: 1636–1648.

Buonaccorsi, V.P., J.R. McDowell & J.E. Graves, 2001. Reconciling patterns of inter-ocean molecular variance from four classes of molecular markers in blue-marlin (*Makaira nigricans*). Mol Ecol 10: 1179–1196.

Caccone, A., M.C. Milinkovitch, V. Sbordoni & J.R. Powell, 1997. Mitochondrial DNA rates and biogeography in European newts (genus *Euproctus*). Syst Biol 46: 126–144.

Carvalho G.R., 1998. Molecular Ecology: Origins and Approach. In: Carvalho GR (ed) Advances in Molecular Ecology, IOS Press pp 1–16.

Chenuil, A., M. Le Gac & M. Thierry, 2003. Fast isolation of microsatellite loci of very diverse repeat motifs by library enrichment. Application in echinoderms. (Technical note). Mol Ecol Notes 3: 324–327.

Chevaldonné, P., D. Jollivet, D. Desbruyùres, R. Lutz & R. Vrijenhoek, 2002. Sister-species of eastern Pacific hydrothermal vent worms (Ampharetidae, Alvinellidae, Vestimentifera) provide new mitochondrial *COI* clock calibration. Cah Biol Mar 43: 367–370.

Crochet, P-A. & E. Desmarais, 2000. Slow rate of evolution in the mitochondrial control region of gulls (Ayes:Laridae). Mol Biol Evol 17: 1797–1806.

Davies, N., F.X. Villablanca & G.K. Roderick, 1999. Determining the source of individuals:multilocus genotyping in nonequilibrium population genetics. Trends Ecol Evol 14: 17–21.

Desmarais, E., I. Lanneluc & J. Lagnel, 1997. Direct amplification of length polymorphisms (DALP), or how to get and characterise new genetic markers in many species. Nucleic Acids Res 26: 1458–1465.

Després, L., D. Imbert-Establet, C. Combes & F. Bonhomme, 1992. Molecular evidence linking hominid evolution to recent radiation of Schistosomes (Platyhelminthes:Trematoda). Mol Phyl Evol 1: 295–304.

DeWoody, J.A. & J.C. Avise, 2000. Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. J Fish Biol 56: 461–473.

Dowling, T.E., C. Moritz, J.D. Palmer & L.H. Rieseberg, 1996. Nucleic Acids III:Analysis of fragments and restriction sites, pp. 249–320 in Molecular Systematics edited by D.M. Hillis, C. Moritz & B.K. Mable. Sinauer Associates, Inc, Sunderland, Massachusetts.

Ellegren, H., 2000. Microsatellite mutations in the germline:implications for evolutionary inference. Trends Genet 16: 551–558.

Estoup, A. & B. Angers, 1998. Microsatellites and minisatellites for molecular ecology:theoretical and empirical considerations, pp. 55–86 in Advances in molecular ecology edited by G.R. Carvalho. IOS Press, Amsterdam.

Féral, J-P., 2002. How useful are genetic markers in attempts to understand and manage marine biodiversity? J Exp Mar Biol Ecol 268: 121–145.

Fridolfsson, A-K. & H. Ellegren, 2000. Molecular evolution of the avian CHD1 genes on the Z and W sex chromosomes. Genetics 155: 1903–1912.

Gerber, S., S. Mariette, R. Streiff, C. Bodenes & A. Kremer, 2000. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol Ecol 9: 1037–1048.

Gielly, L. & P. Taberlet, 1994. The use of chloroplast DNA to resolve plant phylogenies:noncoding *versus* rbcL sequences. Mol Biol Evol 11: 769–777.

Goldman, J. & J. Madrigal, 1997. Complementary strand analysis:a new approach for allelic separation in complex polyallelic genetic systems. Nucleic acids Res 25: 2236–2238.

Graur, D & W.-H. Li, 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Inc, Sunderland, Massachussets.

Hauser, M.-T., F. Adhami, M. Dorner, E. Fuchs & J. Gössl, 1998. Generation of co-dominant PCR-based markers by duplex analysis on high resolution gels. Plant J 19: 117–125.

Hillis, D.M. & M.T. Dixon, 1991. Ribosomal DNA:molecular evolution and phylogenetic inference. Q Rev Biol 66: 411–427.

Jarman, S.N., R.D. Ward & N.G. Elliott, 2002. Oligonucleotide primers for PCR amplification of coelomate introns. Mar Biotechnol 4: 347–355.

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Kruger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. de Knijff, M. Stoneking & A. Sajantila, 2000. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet 66: 1580–1588.

Konieczny, A. & F.M. Ausubel, 1993. A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-bases markers. Plant J 4: 403–410.

Kwok, P.-Y. & X. Chen, 2003. Detection of Single Nucleotide Polymorphism. Curr Issues Mol Biol 5: 43–60.

Laporte, V. & D. Charlesworth, 2001. Non-sex linked, nuclear cleaved amplified polymorphic sequences in *Silene latifolia*. J Hered 92: 357–359.

Lessios, H.A., B.D. Kessing, D.R. Robertson & P.G. , 1999. Phylogeography of the pantropical sea urchin *Eucidaris* in relation to land barriers and ocean currents. Evolution 53: 806–817.

Linder, C.R., L.R. Goertzen, B.V. Heuvel, J. Francisco-Ortega & R.K. Jansen, 2000. The complete external transcribed spacer of 18S-26S rDNA:amplification and phylogenetic utility at low taxonomic levels in asteraceae and closely allied families. Mol Phyl Evol 14: 285–303.

Luikart, G. & P.R. England, 1999. Statistical analysis of microsatellite DNA data. Trends Ecol Evol 14: 632–638.

Martin, P.G. & J.M. Dowd, 1991. A comparison of 18S ribosomal RNA and rubisco large subunit sequences for studying angiosperm phylogeny. J Mol Evol 33: 274–282.

Myers, R.M., T. Maniatis & L.S. Lerman, 1987. Detection and localization of single base changes by denaturing gradient gel electrophoresis. Methods Enzymol 155: 501–527.

Nagai, K., 2001. Molecular evolution of Sry and Sox gene. Genetics 270: 161–169.

Nielsen, R., 2001. Statistical tests of selective neutrality in the age of genomics. Heredity 86: 641–647.

Ohresser, M., P. Borsa & C. Delsert, 1997. Intron-length polymorphism at the actin locus Mac-1:a genetic marker for population studies in the marine mussels *Mytilus galloprovincialis* Lmk. and M. *edulis* L. Mol Mar Biol Biotechnol 6: 123–130.

Orita, M., H. Iwahana, H. Kanazawa, K. Hayashi & T. Sekiya, 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86: 2766–2770.

Palmer, J.D., K.L. Adams, Y. Cho, C.L. Parkinson, Y.L. Qiu & K. Song, 2000. Dynamic evolution of plant mitochondrial genomes:mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97: 6960–6966.

Palumbi, A.R., F. Cipriano & M.H. Hare, 2001. Predicting nuclear gene coalescence from mitochondrial data:the three-times rule. Evolution 55: 859–868.

Palumbi, S.R., 1996. Nucleic Acids II:The polymerase chain reaction, pp. 205–247 in Molecular systematics edited by D.M. Hillis, C. Moritz & B.K. Mable.. Sinauer, Sunderland, Massachussetts.

Pamilo, P. & N.O. Bianchi, 1993. Evolution of the Zfx and Zfy genes:rates and interdependence between the genes. Mol Biol Evol 10: 271–281.

Pamilo, P. & R.J. O'Neill, 1997. Evolution of the Sry genes. Mol Biol Evol 14: 49–55.

Pelandakis, M. & M. Solignac, 1993. Molecular phylogeny of Drosophila based on ribosomal RNA sequences. J Mol Evol 37: 525–543.

Pesole, G., C. Gissi, A. De Chirico & C. Saccone, 1999. Nucleotide substitution rate of mammalian mitochondrial genomes. J Mol Evol 48: 427–434.

Poteaux, C., F. Bonhomme & P. Berrebi, 1999. Microsatellite polymorphism and genetic impact of restocking in Mediterranean brown trout (*Salmo trutta* L.). Heredity 82: 645–653.

Provan, J., N. Soranzo, NJ. Wilson, DB. Goldstein & W. Powell, 1999. A low mutation rate for chloroplast microsatellites. Genetics 153: 943–947.

Prugnolle, F. & T. De Meeû, 2002. Inferring sex-based dispersal from population genetic tools:a review. Heredity 88: 161–165.

Qu L.H., 1986. Structuration et evolution de I'ARN ribosomique 28S chez Ies eucaryotes. Etude systûmatique de la rûgion 5? terminale. Ph < D, Paul Sabatier de Toulouse.

Roger, A., O. Sandblom, W. Ford Doolittle & H. Philippe, 1999. An evaluation of elongation factor 1 alpha as a phylogenetic marker for eukaryotes. Mol Biol Evol 16: 218–233.

Rokas, A., J.A. Nylander, F. Ronquist & G.N. Stone, 2002. A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera:Cynipidae):implications for insect phylogenetic studies. Mol Phylogenet Evol 22: 206–219.

Sambrook, J., E.F. Fritsch & T. Maniatis, 1989. Molecular cloning. A laboratory manual. 2Cold Spring Harbor Laboratory Press, USA.

Slattery, J.P. & S.J. O'Brien, 1998. Patterns of *Y* and *X* chromosome DNA sequence divergence durign the Felidae radiation. Genetics 148: 1245–1255.

Sorhannus, U., 1996. Higher ribosomal RNA substitution rates in Bacillariophyceae and Dasycladales than in Mollusca, Echinodermata, and Actinistia-Tetrapoda. Mol Biol Evol 13: 1032–1038.

Sunnucks, P., 2000. Efficient genetic markers for population biology. Trends Ecol Evol 15: 199–203.

Sunnucks, P., A.C. Wilson, L.B. Beheregaray, K. Zenger, J. French & A.C. Taylor, 2000. SSCP is not so difficult:the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Mol Ecol 9: 1699–1710.

Templeton, A., 1998. Nested clade analyses of phylogeographic data:testing hypotheses about gene flow and population history. Mol Ecol 7: 381–397.

Tourasse, N.J. & M. Gouy, 1997. Evolutionary distances between nucleotide sequences based on the distribution of

substitution rates among sites as estimated by parsimony. Mol Biol Evol 14: 287–298.

Vitalis, R. & D. Couvet, 2001. Estimation of effective population size and migration rate from one-and two-locus identity measures. Genetics 157: 911–925.

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. Van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman & M.e.a. Kuiper, 1995. AFLP:a new technique for DNA fingerprinting. Nucleic Acids Res 23: 4407–4414.

Walsh, P.S., D.A. Metzger & R. Higushi, 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques 10: 506–513.

Ward, R.D., M. Woodwark & D.O.F. Skibinski, 1994. A comparison of genetic diversity levels in marine, freshwater, and anadromous fihes. J Fish Biol 44: 213–232.

Waser, P.M. & C. Strobeck, 1998. Genetic signatures of interpopulation dispersal. Trend Ecol Evol 13: 43–44.

Wattier, R., C.R. Engel, P. Saumitou-Laprade & M. Valero, 1998. Short allele dominance as a source of heterozygote deficiency at microsatellite loci:experimental evidence at the dinucleotide locus Gv1CT in *Gracilaria gracilis* (Rhodophyta). Mol Ecol 7: 1569–1573.

White, H.W. & N. Kusukawa, 1997. Agarose-based system for separation of short tandem repeat loci. Biotechniques 22: 976–980.

Yang, Z. & J.P. Bielawski, 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.

Zane, L., L. Bargelloni & T. Patarnello, 2002. Strategies for microsatellite isolation:a review. Mol Ecol 11: 1–16.

Zhang, D.-X. & G.M. Hewitt, 2003. Nuclear DNA analyses in genetic studies of populations:practice, problems and prospects. Mol Ecol 12: 563–584.