Molecular Ecology

Joanna R. Freeland
The Open University, Milton Keynes

John Wiley & Sons, Ltd

2
Molecular Markers in Ecology

Understanding Molecular Markers

In Chapter 1 we started to look at the extraordinary wealth of genetic information that is present in every individual, and to explore how some of this information can be accessed and used in ecological studies. We will build on this foundation by looking in more detail at some of the properties of the genetic markers that are used in molecular ecology. We will start with an overview of the different genomes, because the use and interpretation of all markers will be influenced by the way in which they are inherited. The second half of the chapter will be an overview of those molecular markers that are most commonly used in ecology. After reading this chapter you should understand enough about different molecular markers to suggest which would be applied most appropriately to general research questions.

Modes of Inheritance

Genetic material is transmitted from parents to offspring in a predictable manner, and this is why molecular markers allow us to infer the genetic relationships of individuals. This does not simply mean that we can use genetic data to determine whether two individuals are siblings or first-cousins. In molecular ecology, the calculation of genetic relationships often takes into account the transmission of particular alleles through hundreds, thousands or even millions of generations. In later chapters we will look at some of the ways in which both recent and historical genealogical relationships can be unravelled, but first we must understand how different genomic regions are passed down from one generation to the next. Not all DNA is inherited in the same way, and understanding different modes of inheritance is crucial before we can predict how different regions of DNA might behave under various ecological and evolutionary scenarios.

Molecular Ecology Joanna Freeland

Nuclear versus organelle

The offspring of sexually reproducing organisms inherit approximately half of their DNA from each parent. In a diploid, sexually reproducing organism for example, this means that within the nuclear genome one allele at each locus came from the mother and the other allele came from the father. This is known as **biparental inheritance**. However, even in sexually reproducing species, not all DNA is inherited from both parents. Two important exceptions are the **uniparentally inherited** organelle genomes of mitochondria (mtDNA) and **plastids**, with the latter including chloroplasts (cpDNA). These are both located outside the cell nucleus. Mitochondria are found in both plants and animals, whereas plastids are found only in plants. Organelle DNA typically occurs in the form of supercoiled circles of double-stranded DNA, and these genomes are much smaller than the nuclear genome. For example, at between 15 000 and 17 000 bp the mammalian mitochondrial genome is approximately 1/10 000 the size of the smallest animal nuclear genome, but what they lack in size they partially make up for in number – a single human cell normally contains anywhere from 1 000 to 10 000 mitochondria. Molecular markers from organelle genomes, particularly animal mtDNA, have been exceedingly popular in ecological studies because, as we shall see below, they have a number of useful attributes that are not found in nuclear genomes.

Animal mitochondrial DNA

Mitochondrial DNA is involved primarily in cellular respiration, the process by which energy is extracted from food. Animal mtDNA contains 13 protein-coding genes, 22 transfer RNAs and two ribosomal RNAs. There is also a control region that contains sites for replication and transcription initiation. Most of the sequences are unique, i.e. they are non-repetitive, and there is little evidence of either spacer sequences between genes or intervening sequences within transcribed genes. Although some rearrangement of mitochondrial genes has been found in different animal species, the overall structure, size and arrangement of genes are relatively conserved (Figure 2.1). In most animals, mitochondrial DNA is inherited maternally, meaning that it is passed down from mothers to their offspring (although there are exceptions; see Box 2.1).

There are several reasons why mtDNA markers have been used extensively in studies of animal population genetics. First of all, mtDNA is relatively easy to work with. Its small size, coupled with the conserved arrangement of genes, means that many pairs of universal primers will amplify regions of the mitochondria in a wide variety of vertebrates and invertebrates. This means that data often can be obtained without any *a priori* knowledge about a particular species' mitochondrial DNA sequence. Second, although the arrangement of genes is conserved, the overall mutation rate is high. The rate of synonymous substitutions in mammalian mtDNA

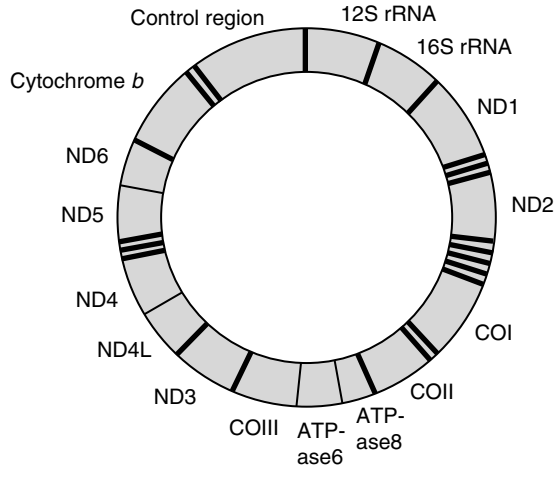

Figure 2.1 Typical gene organization of vertebrate mtDNA. Unlabelled dark bands represent 22 transfer RNAs (tRNAs). Gene abbreviations starting with ND are subunits of NADH dehydrogenase, and those starting with CO are subunits of cytochrome *c* oxidase

has been estimated at 5.7×10^{-8} substitutions per site per year (Brown *et al.*, 1982), which is around ten times the average rate of synonymous substitutions in protein-coding nuclear genes. The non-coding control region, which includes the **displacement (D) loop**, evolves particularly rapidly in many taxa. The high mutation rate in mtDNA may be due partly to the by-products of metabolic respiration and also to less-stringent repair mechanisms compared with those acting on nuclear DNA (Wilson *et al.*, 1985). Regardless of the cause, these high mutation rates mean that mtDNA generally shows relatively high levels of polymorphism and therefore will often reveal multiple genetic lineages both within and among populations.

The third relevant property of mtDNA is its general lack of recombination, which means that offspring usually will have (barring mutation) exactly the same mitochondrial genome as the mother. As a result, mtDNA is effectively a single **haplotype** that is transmitted from mothers to their offspring. This means that mitochondrial lineages can be identified in a much more straightforward manner than nuclear lineages, which, in sexually reproducing species, are continuously pooling genes from two individuals and undergoing recombination. The effectively clonal inheritance of mtDNA means that individual lineages can be tracked over time and space with relative ease, and this is why, as we will see in Chapter 5, mtDNA sequences are commonly used in studies of phylogeography.

Finally, because mtDNA is haploid and uniparentally inherited, it is effectively a quarter of the population size of diploid nuclear DNA. Because there are fewer copies of mtDNA to start with, it is relatively sensitive to demographic events such as bottlenecks. These occur when the size of a population is temporarily reduced, e.g. following a disease outbreak or a catastrophic event. Even if the population

recovers quickly, it will have relatively few surviving mitochondrial haplotypes compared with nuclear genotypes. As we will see in the next chapter, inferring past bottlenecks can make an important contribution towards understanding the current genetic make-up of populations.

Box 2.1 Mitochondrial DNA: exceptions to the rules

Uniparental inheritance and a lack of recombination have made mtDNA the molecular marker of choice in many studies of animal populations because these properties mean that, until a mutation occurs, all of the descendants of a single female will share the same mitochondrial haplotype. This means that genetic lineages can be retraced through time using relatively straightforward models. However, as with so many things in biology, there are exceptions to the rules of mitochondrial inheritance. For one thing, not all mitochondria in animals are inherited maternally. Instances of paternal leakage (transmission of mitochondria from father to offspring) have been found in a number of species, including mice (Gyllensten *et al.*, 1991), birds (Kvist *et al.*, 2003) and humans (Schwartz and Vissing, 2002). Nevertheless, the extent of paternal leakage is believed to be low in most animals with the exception of certain mussel species within the families Mytilidae, Veneridae and Unionidae, which follow double biparental inheritance. This means that females generally inherit their mitochondria from their mothers, but males inherit both maternal and paternal mtDNA. The males therefore represent a classic case of **heteroplasmy** (more than one type of mitochondria within a single individual).

Bivalves by no means represent the only group in which heteroplasmy has been documented, but its prevalence in some mussel species makes them an ideal taxonomic group in which to investigate another question that has been posed recently: do animal mitochondria sometimes undergo recombination? The lack of identifiable recombinant mtDNA haplotypes in natural populations has led to the assumption that no recombination was taking place, but there was always the possibility that recombination was occurring but was remaining undetected because it involved two identical haplotypes. The presence of more than one mtDNA haplotype in male mussels meant that recombination could, at least in theory, produce a novel haplotype, and this indeed has proved to be the case (Ladoukakis and Zouros, 2001; Burzynski *et al.*, 2003). These findings suggest that mtDNA recombination may be more common than was previously believed in the animal kingdom, although at the moment there is little evidence that

recombination, heteroplasmy or paternal leakage are compromising those ecological studies that are based on mtDNA sequences.

Plant mitochondrial DNA

As with animals, mtDNA in most higher plants is maternally inherited. There are a few exceptions to these rules, for example mtDNA is transmitted paternally in the redwood tree *Sequoia sempervirens* and biparentally inherited in some plants in the genus *Pelargonium* (Metzlaff, Borner and Hagemann, 1981). The overall function of plant and animal mitochondria is similar but their structures differ markedly. Unlike animal mtDNA, plant mitochondrial genomes regularly undergo recombination and therefore evolve rapidly with respect to gene rearrangements and duplications. As a result, their sizes vary considerably (40 000 – 2 500 000 bp). This variability makes it difficult to generalize; to take one example, the mitochondrial genome of the liverwort *Marchantia polymorpha* is around 186 608 bp long and appears to include three ribosomal RNA genes, 29 transfer RNA genes, 30 protein-coding genes with known functions and around 32 genes of unknown function (Palmer, 1991). Although the organization of plant mitochondria regularly changes, evolution is slow with respect to nucleotide substitutions. In fact, in most plant species the mitochondria are the slowest evolving genomes (Wolfe, Li and Shorg, 1987). The overall rates of nucleotide substitutions in plant mtDNA are up to 100 times slower than those found in animal mitochondria (Palmer and Herbon, 1988), and this low mutation rate combined with an elevated recombination rate means that plant mitochondrial genomes have not featured prominently in studies of molecular ecology.

That is not to say that there are no useful applications of mtDNA data in plant studies. For many plant species, dispersal is possible through either seeds or pollen, which often vary markedly in the distances over which they can travel. For example, if seeds are eaten by small mammals they may travel relatively short distances before being deposited. In contrast, pollen may be dispersed by the wind, in which case it could travel a long way from its natal site. Even if seeds are wind dispersed they are heavier than pollen and therefore still likely to travel shorter distances. Nevertheless, it is also possible that the opposite scenario could occur, for example seeds that are ingested by migratory birds may travel much further than wind-blown pollen. Tracking seeds and pollen is extremely difficult, but the different dispersal abilities of the two sometimes can be inferred by comparing the distributions of mitochondrial and nuclear genes. Because mtDNA is usually inherited maternally, its distribution will reflect the patterns of seed dispersal but will not be influenced by the spread of pollen, which contains only the paternal genotype.

Canadian populations of the black spruce (*Picea mariana*) grow in areas that were covered in ice until approximately 6000 years ago, so we know that populations must have been established since that time. Researchers found that current populations share the same mtDNA haplotypes but not the same nuclear alleles (Gamache *et al.*, 2003). This difference was attributed to the widespread dispersal of nuclear genes that were carried by wind-blown pollen, coupled with a much more restricted dispersal of mitochondrial genes that can be transported only in seeds. Because seeds usually do not travel very far, it is likely that only a few were involved in establishing populations once the ice had retreated, hence the lack of variability in mtDNA. On the other hand, the pollen that blew to these sites most likely originated in multiple populations, and this has led to a much higher diversity in nuclear genotypes. If this study had been based solely on data from either nuclear or mitochondrial DNA we would have an incomplete, and possibly misleading, picture denoting the dispersal of this coniferous species.

Plastids, including chloroplast DNA

The relatively low variability of plant mtDNA means that when haploid markers are desirable in plant studies, researchers more commonly turn to plastid genomes, including chloroplast DNA (cpDNA). Like mtDNA, cpDNA is inherited maternally in most **angiosperms** (flowering plants), although in most **gymnosperms** (conifers and cycads) it is usually inherited paternally. Chloroplast genomes, which in most plants are key to the process of photosynthesis, typically range from 120 000 to 220 000 bp (the average size is around 150 000 bp). Although recombination sometimes occurs, chloroplasts are for the most part structurally stable, and most of the size variation can be attributed to differences in the lengths of sequence repeats, as opposed to the gene rearrangement and duplication found in plant mtDNA.

In tobacco (*Nicotiana tabacum*), the cpDNA genome contains approximately 113 genes, which include 21 ribosomal proteins, 4 ribosomal RNAs, 30 transfer RNAs, 29 genes that are necessary for functions associated with photosynthesis and 11 genes that are involved with chlororespiration (Sugiura, 1992). A partial arrangement of chloroplast genes in the liverwort (*Marchantia polymorpha*) is shown in Figure 2.2. The average rate of synonymous substitutions in the chloroplast genome, at least in higher plants, is estimated as nearly three times higher than that in plant mtDNA (Wolfe, Li and Sharp, 1987), although this is still four to five times slower than the estimated overall rate of synonymous substitutions in plant nuclear genomes (Wolfe, Sharp and Li, 1989). However, this is an average mutation rate, and the use of cpDNA markers in plant population genetic studies has escalated in recent years following the discovery of highly variable microsatellite regions within the chloroplast genome (see below).

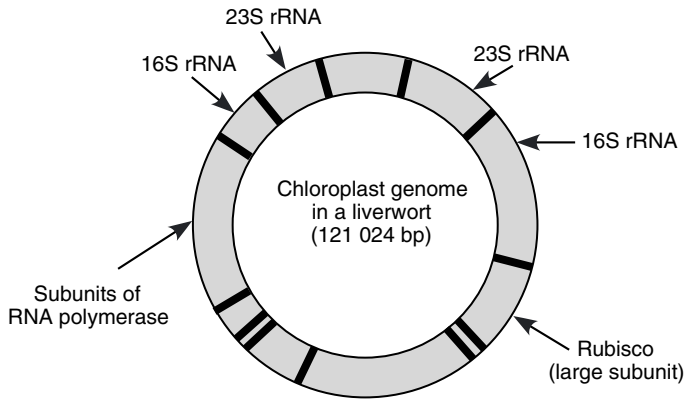

Figure 2.2 The genome of the chloroplasts found in the liverwort *Marchantia polymorpha* contains 121 024 base pairs (Ohyama *et al.*, 1986). These make up an estimated 128 genes, and the approximate locations of some of these are shown on this figure. The dark lines mark the locations of 12 of the 37 tRNAs

Even when variability is low, genetic data from chloroplasts continue to play an important role in ecological studies, in part because of their uniparental mode of inheritance. In the previous section, we saw how a comparison of data from mtDNA (dispersed only in seeds) and nuclear DNA (dispersed in both seeds and pollen) provided insight into the relative contributions that seeds and pollen make to the dispersal patterns of black spruce. Another way to approach this question in conifers, in which chloroplasts are inherited paternally, is to compare the distribution of chloroplast genes (dispersed in both seeds and pollen) with the distribution of mitochondrial genes (dispersed only in seeds). Latta and Mitton (1997) followed this approach in a study of Limber pine (*Pinus flexilis* James) in Colorado. The seeds of this species are dispersed by Clark's nutcracker (*Nucifraga columbiana*), which caches seeds within a limited radius of the natal tree. If these caches are subsequently abandoned, they may grow into seedlings. Pollen, on the other hand, is dispersed by the wind, and therefore should travel further than the seeds. As expected, the mitochondrial haplotypes of Limber pine were distributed over much smaller areas than the chloroplast haplotypes, once again supporting the notion of relatively widespread pollen dispersal (Latta and Mitton, 1997).

Haploid chromosomes

When discussing the inheritance of nuclear and organelle markers we usually refer to nuclear genes as being inherited biparentally following sexual reproduction. For the most part this is true, but sex chromosomes (chromosomes that have a role in the determination of sex) provide an exception to this rule. Not all species have sex chromosomes, for example crocodiles and many turtles and lizards follow

environmental sex determination, which means that the sex of an individual is determined by the temperature that it is exposed to during early development. Many other species follow **genetic sex determination**, which occurs when an individual's sex is determined genetically by sex chromosomes. This can happen in a number of different ways. In most mammals, and some **dioecious** plants, females are **homogametic** (two copies of the same sex chromosome: XX), whereas males are **heterogametic** (one copy of each sex chromosome: XY). The opposite is true in birds and lepidopterans, which have heterogametic females (ZW) and homogametic males (ZZ). In some other species such as the nematode *Caenorhabditis elegans*, the heterogametic (male) sex is XO, meaning that it has only a single X chromosome. **Monoecious** plant species typically lack discrete sex chromosomes.

In mammals, each female gives one of her X chromosomes to all of her children, male and female alike. It is the male parent's contribution that determines the sex of the offspring; if he donates an X chromosome it will be female, and if he donates a Y chromosome then the offspring will be male. The Y chromosome therefore follows a pattern of **patrilineal descent** because it is passed down only through the male lineage, from father to son (Table 2.1). Because there is never more than

Table 2.1 Usual mode of inheritance of different genomic regions in sexually reproducing taxa

Genomic region	Typical mode of inheritance
Animals	
Autosomal chromosomes	Biparental
Mitochondrial DNA	Maternal in most animals Biparental in some bivalves
Y chromosome	Paternal
Higher plant	
Autosomal chromosomes	Biparental
Mitochondrial DNA	Usually maternal
Plastid DNA (including chloroplast DNA)	Maternal in most angiosperms Paternal in most gymnosperms Biparental in some plants
Y chromosome	Paternal in some dioecious plants

one copy of a Y chromosome in the same individual (barring genetic abnormalities), Y chromosomes are the only mammalian chromosomes that are effectively haploid. In addition, like mtDNA, Y chromosomes for the most part do not undergo recombination. There are two small pseudo-autosomal regions at the tips of the chromosome that recombine with the X chromosome, but in between these are approximately 60 Mb of non-recombining sequence (Figure 2.3).

The mutation rate of Y chromosomes is relatively low. One study found that the variability of three genes on the Y chromosome was approximately five times lower

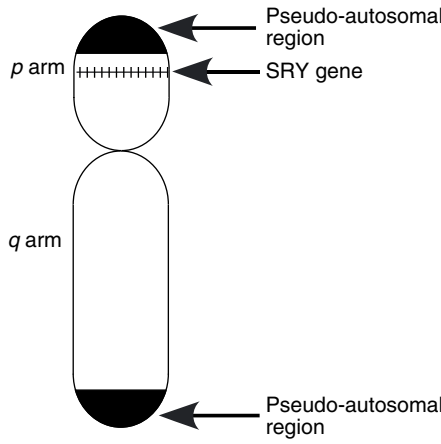

Figure 2.3 Mammalian Y chromosome. The SRY gene (sex-determining region Y) effectively converts an embryo into a male

than that of the corresponding regions of autosomal genes (Shen *et al.*, 2000). Reasons for this remain unclear, although it may be due in part to its smaller population size: the total number of Y chromosomes in any given species is a quarter of that for autosomes and a third of that for X chromosomes. This may seem initially confusing because the population size of mtDNA and Y chromosomes is essentially the same and yet mtDNA is relatively variable; however, Y chromosomes have the same repair processes that are found in other regions of nuclear DNA but are lacking in mtDNA. Despite relatively low levels of variability, the Y chromosome still has the potential to be a significant source of information because it is much larger than mtDNA and, unlike mitochondria, contains substantial amounts of non-coding DNA.

One important property of Y chromosomes is that they allow biologists to follow the transmission of paternal genotypes in animals in much the same way that chloroplast markers can be used for gymnosperms. A recent study (Zerjal *et al.*, 2003) found that a particular Y chromosome haplotype is abundant in human populations in a large area of Asia, from the Pacific to the Caspian Sea. Approximately 8 per cent of the men in this region carry it and, because of the high population density in this part of the world, this translates into approximately 0.5 per cent of the world's total population. The researchers who conducted this study suggested that this prevalent Y chromosome haplotype can be traced back to the infamous warrior Genghis Khan. Born in the 12th century, Khan created the biggest land-based empire that the world has seen (Genghis Khan means supreme ruler). He had a vast number of descendants, many of whom were fathered following his conquests, and apparently his sons were also extremely prolific. His policy of slaughtering millions of people and then reproducing en masse is one possible explanation for the widespread occurrence of a single Y chromosome haplotype in Asia today.

Identifying hybrids

It should be apparent from the examples in the previous section that there is often an advantage to using multiple molecular markers that have contrasting modes of inheritance. These potential benefits are further illustrated by studies of hybridization. **Hybrids** can be identified from their genotypes because hybridization results in **introgression**, the flow of alleles from one species (or population) to another. As a result, hybrids typically contain a mixture of alleles from both parental species, for example a comparison of grass species in the genus *Miscanthus* revealed that *M. giganteus* was a hybrid of *M. sinensis* and *M. sacchariflorus* because it had one ITS allele from each parental species and a plastid sequence that identified the maternal lineage as *M. sacchariflorus* (Hodkinson *et al.*, 2002). Identification of hybrids in the wild is often based on **cytonuclear disequilibrium**, which occurs in hybrids that have cytoplasmic markers (another name for mitochondria and chloroplasts) from one species or population and nuclear markers from another. Some examples of cytonuclear disequilibrium are given in Table 2.2. By using a combination of markers that have maternal, paternal or biparental inheritance, we may be able to identify which species – or even which population – the hybrid's mother and father came from.

Table 2.2 Some examples of cytonuclear disequilibrium in hybrids

Hybrid	Nuclear DNA	mtDNA or cpDNA	Reference
Freshwater crustaceans *Daphnia pulex* × *D. pulicaria*	*D. pulicaria*	*D. pulex*	Crease *et al.* (1997)
Grey wolf (*Canis lupus*) × coyote (*C. latrans*)	Wolf	Coyote	Lehman *et al.* (1991)
House mice *Mus musculus* × *M. domesticus*	*M. musculus*	*M. domesticus*	Gyllensten and Wilson (1987)
Northern red-backed vole (*Clethrionomys rutilus*) × bank vole (*C. glareolus*)	Bank vole	Northern red-backed vole	Tegelström (1987)
White poplar (*Populus alba*) × black poplar (*P. nigra*)	Black poplar	White poplar	Smith and Sytsma (1990)

Researchers used multiple markers to determine whether or not members of the declining wolf (*Canis lupus*) populations in Europe have been hybridizing with domestic dogs (*C. familiaris*), a question that has been open to debate for some time. One study that used mitochondrial markers to investigate this possibility found only a few instances of haplotype sharing between wolves and dogs and therefore concluded that hybridization between dogs and declining wolf populations was not a cause for concern (Randi *et al.* 2000). However, because mtDNA is

inherited maternally, dog haplotypes would appear in wolf populations only if hybridization occurred between female dogs and male wolves. In a more recent study, researchers used a combination of markers from mtDNA, Y chromosome and autosomal nuclear DNA to conduct a more detailed assessment of a possible wolf–dog hybrid from Norway. The mitochondrial haplotype in this suspected hybrid came from a Scandinavian wolf. The Y chromosome data suggested that the father of the hybrid was not a Scandinavian wolf, but did not provide enough information for researchers to discriminate between dogs or migrant wolves from Finland or Russia. However, data from autosomal chromosomes suggested that the most likely father was a dog. It was therefore the combination of uniparental and biparental markers that identified this specimen as a hybrid that had resulted from a cross between a female Scandinavian wolf and a male dog (Vilà *et al.*, 2003b).

Uniparental markers: a cautionary note

Although uniparental markers have many useful applications in molecular ecology, it is important to keep in mind that they also have some drawbacks. We have referred already to the high recombination rates of plant mtDNA and the low mutation rates of cpDNA, and whereas neither of these considerations is particularly relevant to animal mtDNA, there are other limitations that are common to all mtDNA and cpDNA markers. For one thing, organelles behave as single inherited units and are therefore effectively single locus markers. As we will see throughout this text, data from a single locus allow us to retrace the history of only a single genetic unit (gene or genome), which may or may not be concordant with the history of the species in question. This is particularly true of mtDNA, cpDNA and Y chromosomes because their reduced effective population sizes, relative to autosomal DNA, mean that their haplotypes have a greater probability of going extinct. This loss of haplotypes can cause researchers to infer an oversimplified population history or to underestimate levels of genetic diversity.

An additional drawback to uniparentally inherited markers is that they may not be representative of populations as a whole. We have seen already how the inheritance of markers can influence our understanding of the dispersal patterns of plants depending on whether we target mtDNA, cpDNA or nuclear genomes. The same can be true in animals if dispersal is undertaken by only one gender, e.g. if males disperse and females do not, the mtDNA haplotypes would be population-specific and mtDNA data may lead us to conclude erroneously that individuals never move between populations.

Another risk associated with mtDNA markers involves copies of mtDNA that have been translocated to the nuclear genome; these are known as **mitochondrial pseudogenes**, or **numts** (*nu*clear copies of *mt*DNA sequences). Once they have been transposed into the nucleus, these non-functional pseudogenes continue to

evolve independently of mtDNA. Problems will arise from this during PCR if the primer-binding sites have been conserved in the pseudogene, which then may be amplified in addition to, or instead of, the desired mitochondrial region. Although not evenly distributed across taxa, nuclear copies of mitochondrial DNA have been found in more than 80 eukaryotic species, including fungi, plants, invertebrates and vertebrates (reviewed in Bensasson *et al.*, 2001). Steps can be taken that often greatly reduce the likelihood of amplifying these mitochondrial pseudogenes, although one study that specifically set out to investigate this problem concluded that the high frequency of numts in gorillas, combined with their overall similarity to true mtDNA sequences, meant that the application of mtDNA analysis in this species should be undertaken only with extreme caution (Thalman *et al.*, 2004).

A final note about uniparentally inherited markers is that their applications are somewhat different in asexually reproducing organisms. Up to this point our discussion has centred, either implicitly or explicitly, on sexually reproducing organisms, but of course not all organisms reproduce in this way. A large proportion of prokaryotes, plus numerous eukaryotes (plants, invertebrates and some vertebrates), can reproduce asexually (Table 2.3). In the absence of sex, there is no distinction between genomes that are uniparentally and biparentally inherited. Note, however, that the picture is often complicated by the fact that many species are capable of both sexual and asexual reproduction. For example, reproduction of the grain aphid (*Sitobion avenae*) appears, for the most part, to be predominantly asexual in the north and sexual in the south of Britain, although

Table 2.3 Some methods of asexual reproduction

Method of reproduction	Examples
Vegetative reproduction (asexual reproduction from somatic cells). Includes:	
• Budding. An offspring grows out of the body of the parent	Hydra, planarians
• Fragmentation. Body of parent breaks into distinct pieces, each of which can form offspring	*Opuntia* cactus
• Rhizomes and stolons. Runners that give rise to new individuals	Strawberries
• Regeneration. If a piece of a parent is detached, it can grow and develop into a new individual	Echinoderms
Parthenogenesis (asexual reproduction via eggs). Includes:	
• Apomixis (mitotic parthenogenesis). The development of an individual from an egg that has not been fertilized, and which has a full complement of the mother's chromosomes. Because there is no involvement of a male gamete, it leads to the production of offspring that are genetically identical to the mother	Aphids, dandelions, flatworms, water fleas, rotifers, whiptail lizards
• Amphimixis (meiotic parthenogenesis). Female parent produces eggs by meiosis, which develop without uniting with a male gamete. The diploid state is restored either by a cell division that doubles the number of chromosomes or by fusion of the egg nucleus with another maternal nucleus	Bagworm moth *Solenobia*; nematodes (genus *Heterodera*)

this seems to depend partly on the climate (Llewellyn *et al.*, 2003). Research into the population genetics of species that have multiple reproductive modes often requires more data than can be obtained from a uniparentally inherited marker.

Molecular Markers

So far in this chapter we have learned that molecular markers will be influenced by the manner in which they are inherited. We will now take a more detailed look at the properties of some of the most common markers that are used to generate data from one or more genomes. An important point to remember is that molecular markers are simply tools that can be used for generating data; like anything else, the job can be done properly only if the correct tools are chosen. Since we cannot make an informed choice until we understand how the tools operate, we need to learn about some of the key characteristics of different markers. The other point to bear in mind while learning about molecular markers is that we are using them to answer ecological questions. In molecular ecology, genetic data are often combined with ecological data that have been obtained either in the wild or in controlled laboratory experiments, following, for example, observations of mating behaviour, census counts, capture–mark–recapture studies, comparisons of growth patterns under different environmental conditions, and descriptions of morphological characters. In this chapter we will limit ourselves to a general understanding of what markers are, and how they can be used in molecular ecology; methods for analysing data from different markers and more detailed discussions of how genetic data can be applied to ecological questions can be found in later chapters.

Several factors need to be taken into account when choosing a marker. First, it is important to consider the expected level of variability. Some genetic regions are expected to evolve more rapidly than others, and the desired variability will depend on the question that is being asked. Generally speaking, markers that allow us to differentiate between closely related organisms will need to be highly variable, whereas the relationships among more distantly related taxa may be resolved by less variable markers. This ties back to an earlier part of this chapter and also to Chapter 1, when we noted that mutation rates vary both within and between genomes. As we shall see, many types of markers can be applied to either nuclear or organelle genomes; however, different mutation rates combined with alternative modes of inheritance means that the decision about which DNA region or genome to target should not be taken lightly.

There are also practical concerns surrounding the choice of marker. Time, money and expertise are all relevant, and there is often a trade-off between precision and convenience. This brings us to the two main categories of markers that will be described in this chapter: **co-dominant** and **dominant**. Co-dominant markers allow us to identify all of the alleles that are present at a particular locus, whereas dominant markers will reveal only a single dominant allele. As a result,

co-dominant data are generally more precise than dominant data, although dominant markers usually require less development time and may therefore be a more convenient way to obtain data. In the rest of this section we will highlight some of the main features of markers in each of the two categories. Note that laboratory practice is, for the most part, beyond the scope of this text; readers looking for specific protocols are referred to the Further Reading section at the end of the chapter.

Co-dominant markers

In a diploid species, each dominant marker will identify one allele in a homozygous individual and two alleles in a heterozygous individual (Figure 2.4). This ability to distinguish between homozygotes and heterozygotes is one of the most important features of co-dominant markers because it means that we can calculate easily the **allele frequencies** for pooled samples (such as populations). Allele frequency simply refers to the frequency of any given allele within a population, i.e. it tells us how common a particular allele is. If we had a diploid population with 30 individuals then there will be a total of $30 \times 2 = 60$ alleles at any autosomal locus. If 12 individuals had the homozygous genotype *AA* and 18 individuals had the heterozygous genotype *Aa* at a particular locus, then the frequency of allele *A*

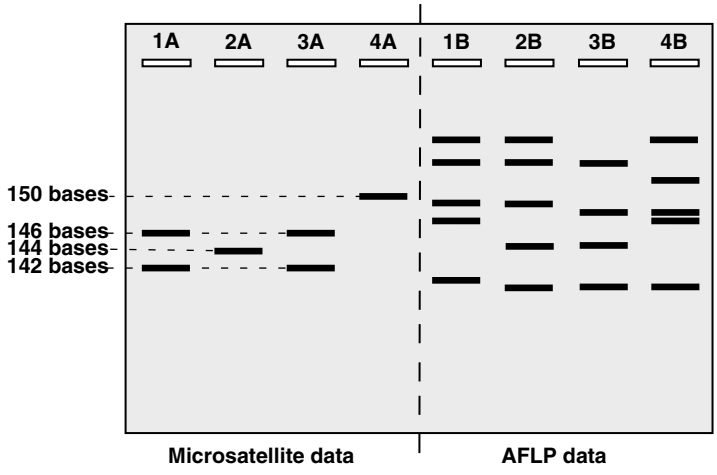

Figure 2.4 Gel showing the genotypes of four individuals based on one microsatellite (co-dominant) locus (1A–4A) and several AFLP (dominant) loci (1B–4B). According to the microsatellite locus, individuals 1 and 3 are heterozygous for alleles that are 142 and 146 bases long, whereas individuals 2 and 4 are homozygous for alleles that are 144 and 150 bases, respectively. Since there are two of each allele in this sample of eight alleles, the frequency of each microsatellite allele is 0.25. According to the AFLP marker, which screens multiple loci, all four individuals are genetically distinct but we cannot identify homozygotes and heterozygotes, nor can we readily calculate allele frequencies

is $[2(12) + 18]/60 = 42/60 = 0.7$ *or* 70 per cent and the frequency of allele *a* is $18/60 = 0.3$ or 30 per cent. As we will see in later chapters, numerous analytical methods in population genetics are based at least partially on allele frequencies.

It is important to note that although each co-dominant marker characterizes a single locus, most projects will use multiple co-dominant markers to generate data from a number of different loci so that conclusions are not based on a single, possibly atypical, locus. The main drawback to using these types of markers is that they tend to be a relatively time-consuming and expensive way to generate data, and in practice this can limit the number of loci that are genotyped.

Allozymes

In Chapter 1 we learned that allozymes were among the first markers to unite molecular genetics and ecology when they were used to quantify the levels of genetic variation within populations. Since their inception in the 1960s, allozymes have played an ongoing role in studies of animal and plant populations, although in recent years they have featured less prominently than DNA markers. Allozymes benefit from their co-dominant nature and may be more time- and cost-effective than some other markers because they do not require any DNA sequence information. However, as we noted in Chapter 1, they provide conservative estimates of genetic variation because their variability depends entirely on non-synonymous substitutions in protein-coding genes. In addition, allozymes are of limited utility when we are interested in the evolutionary relationships between different alleles; if an individual has allele B, there is no reason to believe that its ancestor had allele A, in other words it is not always possible to identify an ancestor and its descendant.

Another property of allozymes is that they are functional proteins and therefore are not always selectively neutral. This can be both an advantage and a disadvantage. A lack of neutrality can be a disadvantage if a marker is being used to test whether or not populations are genetically distinct from one another. The free-swimming larvae of the American oyster (*Crassostrea virginica*) can travel relatively long distances if swept along on ocean currents. Populations that are not connected by currents may therefore be genetically distinct from one other, a hypothesis that was tested in a genetic survey of populations located along the Atlantic and Gulf coasts of the USA (Karl and Avise, 1992). A comparison of mtDNA and six anonymous nuclear sequences clearly showed that populations around the Gulf of Mexico were, in fact, genetically distinct from those located along the Atlantic coast, a finding that is consistent with the expectation of very low dispersal between coastlines that are not connected by currents. Variation at six allozyme loci, on the other hand, revealed no genetic differences between the two geographical areas, presumably because natural selection has been maintaining the same alleles in different populations. If the researchers had looked at only allozyme data they probably would have concluded that larvae regularly travelled

between the Atlantic and Gulf coasts, a finding that would have been difficult to reconcile with the ocean currents in that region.

On the other hand, a non-neutral marker can be useful if we are looking for evidence of adaptation. Mead's sulphur butterfly *Colias meadii* showed some interesting patterns of variation in the glycolytic enzyme phosphoglucose isomerase (PGI), an enzyme involved in glycolysis, which provides fuel for insect flight (Watt *et al.*, 2003). Because flight ability is related to fitness, the allele that confers the best flight ability should be selected for, and therefore a level of genetic uniformity may be expected at the locus coding for PGI. This prediction was supported only partially by a comparison of PGI alleles from *C. meadii* that were sampled from lowland (below the tree line) and alpine (above the tree line) habitats in central USA. Populations showed a high level of genetic uniformity over several hundred kilometres *within* habitats but a marked and abrupt shift in allele frequencies *between* habitats.

Both of these trends are apparently driven by natural selection. *Colias* butterflies spend their adult life within a neighbourhood radius that seldom exceeds 1.5 km. These low levels of dispersal mean that genetic uniformity over hundreds of kilometres must be maintained by a selective force, in this case the relationship between PGI alleles and fitness. Selection also explains the contrasting allele frequencies between alpine and lowland habitats. Because the two habitats are delineated by the tree line, they may be expected to have different thermal (and other abiotic) properties. The activity of PGI varies with temperature, and the authors of this study suggest that alternative PGI alleles may be selected for under different thermal regimes (Watt *et al.*, 2003).

The markers that we will be discussing in the rest of this chapter all target variation in DNA as opposed to proteins. Although allozymes are often subject to selection pressures, DNA markers are more likely to be neutral because they often target relatively variable sequences that, in turn, are less likely to be selectively constrained. However, it is important to bear in mind that not all DNA markers are selectively neutral. In some cases we will specifically discuss non-neutral DNA markers. In other cases neutrality may be assumed, although it is always possible that an apparently non-functional region of DNA is subject to selective pressures that are acting on a genetic region to which it is linked – the so-called **hitch-hiking effect** (Maynard Smith and Haigh, 1974). A more detailed discussion of genetic markers and natural selection is included in Chapter 4.

Restriction fragment length polymorphisms

The first widespread markers that quantified variation in DNA sequences (as opposed to proteins) were **restriction fragment length polymorphism** (**RFLPs**). RFLP data are generated using **restriction enzymes**, which cut DNA at short (usually four to six base pairs), specific sequences. Examples of restriction enzymes

include *Alu*I, which cuts DNA when it encounters the sequence AGCT, and *Eco*RV, which cuts in the middle of the sequence GATATC. Digesting purified DNA with one or more restriction enzymes can turn a single piece of DNA into multiple fragments. If two individuals have different distances between two restriction sites, the resulting fragments will be of different lengths. The RFLPs therefore do not survey the entire DNA sequence, but any mutations that add or remove a recognition site for a particular enzyme, or that change the length of sequence between two restriction sites, will be reflected in the sizes and numbers of the fragments that are run out on a gel (Figure 2.5).

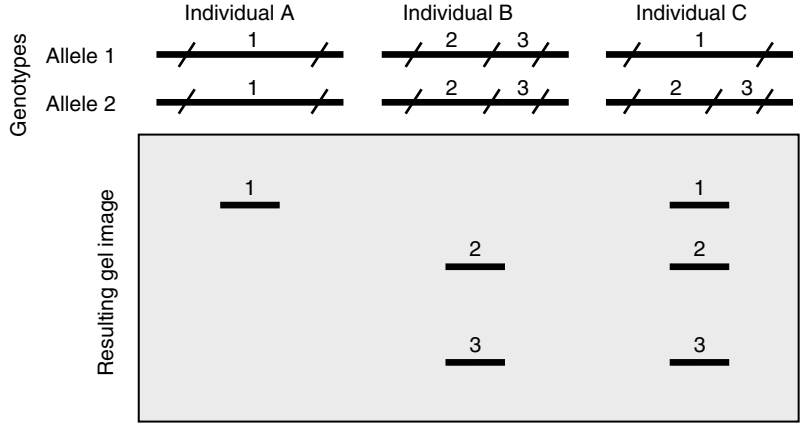

Figure 2.5 **Three different RFLP genotypes result from sequence differences that affect the restriction enzyme recognition sites (designated as /). At this locus, individuals A and B are homozygous for alleles that have two and three restriction sites, respectively. Individual C is heterozygous, with two restriction sites at one allele and three restriction sites at the other allele. The numbers of bands that would be generated by the RFLP profiles are shown in the resulting gel image**

Analysis of RFLPs can be done on either an entire genome (nuclear or organelle) or a specific fragment of DNA. The traditional method involves digesting DNA with one or more enzymes and then running out the fragments on a gel. These are then transferred onto a membrane that is placed in a solution containing multiple single-stranded copies of a particular sequence, all of which have been labelled radioactively or fluorescently (this is known as a **probe**). The single-stranded probe will hybridize to the bands that contain its complementary sequence, and these bands then can be identified from the radioactive or fluorescent label. The number of bands produced will depend on the region surveyed and the enzyme used. For example, a digestion with *Alu*I produces around 341 bands in tobacco chloroplast DNA, whereas *Eco*RV produces only 36 bands (Shinozaki *et al.*, 1986). The same enzymes generate approximately 64 bands and three bands, respectively, in human mitochondrial DNA (Anderson *et al.*, 1981). A comparison of the number and sizes of labelled bands among individuals

provides an estimate of overall genetic similarity. This method is useful for screening relatively large amounts of DNA but is fairly cumbersome and time-consuming.

A more straightforward method of generating RFLP data is to first amplify a specific fragment of DNA using PCR and then digest the amplified product with enzymes. The fragments then can be visualized after they are run out on a high-resolution gel. This technique is known as PCR-RFLP. Development of PCR-RFLP markers inevitably involves a period of trial and error during which different combinations of primers and enzymes must be screened before enough variable sites can be identified, but overall it is a fairly straightforward technique. In one study, PCR-RFLP markers were used to compare regions of the chloroplast genome of heather (*Calluna vulgaris*) collected from Western European populations (Rendell and Ennos, 2002). Four combinations of primers and enzymes revealed a total of eight mutations that collectively revealed twelve different haplotypes. The distributions of these haplotypes revealed high levels of diversity within populations and also substantial genetic differences among populations. The authors compared their results with an earlier study based on nuclear allozyme data and concluded that, unlike the earlier examples of coniferous trees in Chapter 1, seeds are more important than pollen for the long-distance dispersal of heather.

DNA Sequences

In Chapter 1 we saw how DNA sequences can be obtained from fragments of DNA that have been amplified by PCR. Although all genetic markers quantify variations in DNA, sequencing is the only method that identifies the exact base pair differences between individuals. This is an important feature of DNA sequencing because it leaves little room for ambiguity: by comparing two sequences we can identify exactly where and how they are different. As a result, sequencing allows us to infer the evolutionary relationships of alternate alleles. This is possible because, barring back-mutations, each mutation acquired by a specific lineage remains there, even after additional mutations occur. In other words, if an allele with a sequence of GGGATATACGATACG mutates to a new allele with a sequence of **C**GGATATACGATACG, then all descendants of the individual with the new allele will have a C instead of a G at the first base position, even if subsequent mutations occur at other sites along the sequence. Generally speaking, the more mutations that a pair of individuals has in common, the more closely related they are to one another, a concept that will be developed further in Chapter 5.

Sequence data were used to unravel the evolutionary history of the Hawaiian silversword alliance, a group of 28 endemic Hawaiian plant species in the sunflower family (Baldwin and Robichaux, 1995). These plants are of interest because collectively they demonstrate substantial morphological and ecological variation.

Throughout the Hawaiian archipelago they inhabit a range of wet (including sedgeland and forest) and dry (including grassland and shrubland) habitats, in contrast to their less ecologically diverse continental relatives. By comparing sequences from coding and non-coding regions of the nuclear ribosomal DNA genes, the authors of this study were able to identify both shared and unique mutations, which in turn allowed them to conclude which species are most closely related to one another. They were then able to reconstruct the events that led to the evolution of such a diverse group. The sequence data suggest that this group of species arose from a single ancestor that was dispersed, presumably by birds, to the Hawaiian archipelago some time in the past. By combining ecological and genetic data, the authors then could go one step further and conclude that shifts between wet and dry habitats occurred on multiple occasions. This would suggest that ecological diversification played an important role in the speciation of this alliance.

The variable rates of sequence evolution (Table 2.4) mean that we can use relatively rapidly evolving sequences for comparing closely related taxa, and more slowly evolving sequences for comparing distantly related taxa. In recent years, the choice of appropriate gene regions has been facilitated by the growing availability of sequence data. Nevertheless, although sequencing theoretically can be applied to any genomic region, our knowledge of chromosomal sequences is still inadequate

Table 2.4 **Evolutionary rates of some DNA sequences. All estimates are from Li (1997), with the exception of the value given for mitochondrial protein-coding regions in mammals, which is from Brown, George and Wilson (1979). To put the low values in perspective, recall that the diversity of 0.1 % in the human nuclear genome translates into a roughly three million base pair difference between individuals**

Type of sequence	Organism	Average divergence (% per million years)[a]
Nuclear DNA		
Non-synonymous sites	Mammals	0.15
	Drosophila	0.38
	Plant (monocot)	0.014
Synonymous sites	Mammals	0.7
	Drosophila	3.12
	Plant (monocot)	0.114
Introns	Mammals	0.7
Chloroplast DNA		
Non-synonymous sites	Plant (angiosperm)	0.004–0.01
Synonymous sites	Plant (angiosperm)	0.024–0.116
Mitochondrial DNA		
Non-synonymous sites	Plant (angiosperm)	0.004–0.008
Synonymous sites	Plant (angiosperm)	0.01–0.042
Protein-coding regions	Mammals	2.0

[a]These are estimates averaged over multiple loci.

in most taxa and there is a shortage of universal nuclear primers. Universal primers are more abundant for plant and particularly animal organelle genomes; in fact, animal mtDNA has been the source of data in most sequence-based ecological studies.

Although sequence data can be extremely informative, obtaining these data is quite expensive (although decreasingly so) and time-consuming. Development time will be longer if a number of sequences need to be screened before appropriately variable regions are identified. Furthermore, many studies benefit by having data from more than one genetic region, which further adds to the time and expense. Recently, however, a relatively new method known as **single nucleotide polymorphisms** (**SNPs**) has been gaining in popularity because it is specifically designed to target variable DNA bases in multiple loci.

Single nucleotide polymorphisms

Single nucleotide polymorphisms (SNPs) refer to single base pair positions along a DNA sequence that vary between individuals. Most SNPs (pronounced snips) have only two alternative states (i.e. each individual has one of two possible nucleotides at a given SNP locus) and are therefore referred to as biallelic markers. Although technically just another way of looking at sequence variation, SNPs are given their own classification because they provide a new approach for finding informative sequence data; DNA sequencing generally entails a comparison of sequences between individuals to see how much variation exists, whereas individual polymorphic sites must be identified before they can be classified as SNPs. Once we know that particular sites are variable, we can use these SNPs to genetically characterize both individuals and populations. Although still in its infancy, the use of SNPs as molecular markers seems to hold great potential.

There is no doubt that SNPs are widespread. In the human genome they account for approximately 90 per cent of genetic variation (Collins, Brooks and Chakravarti, 1998). A survey of multiple taxa including plants, mammals, birds, insects and fungi suggested that a SNP will be revealed for every 200–500 bp of non-coding DNA and every 500–1000 bp of coding DNA that are sequenced (Brumfield *et al.*, 2003, and references therein). This proved to be a conservative estimate in pied and collared flycatchers, *Ficedula hypoleuca* and *F. albicollis*, because when researchers screened around 9000 bp from each species, they discovered 52 SNPs in pied flycatchers and 61 SNPs in collared flycatchers (Primmer *et al.*, 2002). This translates into an average frequency of approximately one SNP per 175 bp and 150 bp for pied and collared flycatchers, respectively. This is encouraging, as the search for SNPs in the nuclear genome will initially be random in most non-model species.

SNPs can be identified in a relatively straightforward manner by sequencing PCR products that have been amplified using universal or species-specific primers,

or by sequencing anonymous loci such as those amplified by the multi-locus methods outlined below. By targeting multiple loci, researchers should be able to identify a number of SNPs distributed across multiple unlinked sites throughout the nuclear or organelle genomes. The mutation rates of SNPs appear to be in the order of 10^{-8}–10^{-9} (Brumfield *et al.*, 2003). This range is lower than the mutation rates of some other markers such as microsatellites (see below) and therefore the most promising application of SNPs in molecular ecology currently appears to be the elucidation of processes that occurred some time in the past. However, SNPs have been developed only recently, and as increasing numbers are characterized, SNPs are likely to prove suitable for a range of other applications, such as using SNP genotypes to identify individuals and to assess levels of genetic variation within populations (Morin, Luikart and Wayne, 2004).

Microsatellites

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), are stretches of DNA that consist of tandem repeats of 1–6 bp. An example of a microsatellite sequence is the dinucleotide repeat $(CA)_{12}$, which consists of 12 repeats of the sequence CA (CACACACACACACACACACACACA). In this case the complementary DNA sequence would have the microsatellite $(TG)_{12}$. Microsatellites are located throughout nuclear and chloroplast genomes and have also been found in the mitochondrial genomes of some species (Figure 2.6). The initial development of microsatellite markers can take considerable time and money. The usual approach is to clone random fragments of DNA

Figure 2.6 **Diagrammatic representation showing part of a chromosome across which three microsatellite loci are distributed (note that sequences are provided for only one strand of DNA from each chromosome). This particular individual is homozygous at locus 1 because both alleles are $(TA)_4$, heterozygous at locus 2 because one allele is $(TAA)_8$ and the other is $(TAA)_7$, and heterozygous at locus 3 because one allele is $(GC)_7$ and the other allele is $(GC)_8$**

into a library and then screen this library with a microsatellite probe (in much the same way as RFLPs are identified with probes). Clones that contain microsatellites are then isolated and sequenced, and primers that will amplify the repeat region are designed from flanking non-repetitive sequences (Figure 2.7). Once primers have been designed, data can be acquired rapidly by using these primers to amplify microsatellite alleles in PCR reactions. The PCR products then can be run out on a high-resolution gel that will reveal the size of each allele. The number of species from which microsatellite loci have been characterized is growing almost daily, and the sequences that flank microsatellite loci are often conserved between closely

TTGTCAAAGAGTTCAGCCGAATA CAATTTATTAAGTGAGCTTAATACAGTT
AGCGACGACAAAGGAAGAAGTACAAC<u>AGAGAGAGAGAGAGAGAGAGAGAG
AGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGA
GAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG</u>TGTAAAGATATAGGGAG
ACTAGCTAGAG**CCAAGCACTAAGATACAACACGC**

Figure 2.7 A DNA sequence that includes a microsatellite region that was isolated from the freshwater bryozoan *Cristatella mucedo* (Freeland *et al.*, 1999). The microsatellite, which is $(AG)_{53}$, is underlined. The flanking sequence regions in bold show the locations of the primers that were used to amplify this microsatellite in a PCR reaction

related species, which means that microsatellite primers sometimes can be used to generate data from multiple species (e.g. Lippe, Dumont and Berhatchez, 2004; Provan *et al.*, 2004).

Microsatellites mutate much more rapidly than most other types of sequences, with estimated mutation rates of around 10^{-4}–10^{-5} events per locus per replication in yeast (Strand *et al.*, 1993) and around 10^{-3}–10^{-4} in mice (Dallas, 1992). This is substantially higher than the estimated overall point mutation rate of around 10^{-9}–10^{-10} (Li, 1997). These high rates of mutation in microsatellites are ascribed most commonly to slipped-strand mis-pairing during DNA replication (Chapter 1), which, because it can result in either the gain or loss of a single repeat unit, has given rise to the **stepwise mutation model** (SMM; Kimura and Ohta, 1978). Alternatively, the **infinite alleles model** (IAM; Kimura and Crow, 1964) allows for mutations in which multiple repeats are simultaneously gained or lost, but it also assumes that any new allele size has not been encountered previously within a population. Mutation models are important for the analysis of genetic data, but reconciling particular models with the evolution of microsatellites is complicated by the fact that, although mutations often involve single repeats, multiple repeats are periodically gained or lost following a single mutation. At other times, insertions or deletions in the flanking sequences will alter the size of the amplified region. There is also considerable evidence suggesting that the mutation of microsatellites is influenced by the number and size of the repeat motif and also by the complexity of the microsatellite, e.g. whether it is composed of one or multiple repeat motifs (Estoup and Cornuet, 1999).

Microsatellite data are not particularly useful for inferring evolutionary events that occurred in the relatively distant past. Their rapid rate of mutation and their tendency to either increase or decrease in size means that **size homoplasy** may often occur. This can be illustrated by an example of two ancestral alleles at the same locus, one with 20 dinucleotide repeats and the other with 16 dinucleotide repeats. If the larger allele loses one repeat and the smaller allele gains one repeat, then both mutations will have led to alleles with 18 repeats. These two new alleles, each with 18 repeats, may appear to be two copies of the same ancestral sequence, but the evolutionary histories of the two alleles are in fact quite different. Size homoplasy means that ancestor-descendant relationships may be difficult to untangle from microsatellite data.

On the other hand, the high mutation rates of microsatellites mean that there are often multiple alleles at each locus, and this high level of polymorphism makes them suitable for inferring relatively recent population genetic events. East African cichlid fishes were therefore prime candidates for microsatellite analysis, because thousands of endemic species evolved in Lakes Malawi and Victoria within the last 700 000 years, and some species are believed to be only around 200 years old (Kornfield and Smith, 2000). Initial explorations of these species using mtDNA or allozymes revealed little information. The problem was that, even when polymorphic genetic regions were identified, the recency of speciation events meant that most alleles were still shared among taxa because there had not been enough time for species-specific alleles to evolve. In the 1990s, however, microsatellite markers identified much higher levels of variation within and among cichlid species (reviewed in Markert, Danley and Avnegard, 2001). As a result, researchers have been able to use microsatellite data to resolve some aspects of the evolutionary history of cichlid groups (Kornfield and Parker, 1997; Sültmann and Mayer, 1997), although their analyses were somewhat hampered by size homoplasy.

The variability of microsatellites means that, unlike some of the more slowly evolving gene regions, they can also be used to discriminate genetically between individuals and populations. This application has provided some interesting insights into cichlid mating systems. In one study, a combination of behavioural and microsatellite data was used to investigate the role of assortative mating in speciation. Although hybrids from the same lake are often fully fertile, females were found to consistently select males based on their highly divergent colour patterns, ignoring the overall shape similarity that might otherwise blur the division between species. Conclusions from the behavioural data were supported by microsatellite data, which identified the different morphs as genetically distinct taxonomic groups (van Oppen *et al.*, 1998). The high variability, co-dominant nature and increasing availability of microsatellites have made them one of the most popular types of markers in population genetics; however, their extensive development time means that there is still considerable support for dominant markers, to which we shall now turn.

Dominant markers

Dominant markers are also known as multi-locus markers because they simultaneously generate data from multiple loci (Figure 2.4). They typically work by using random primers to amplify anonymous regions of the genome, producing a pattern of multiple bands from each individual. Because they use random primers to amplify fragments of DNA, no prior sequence knowledge is required and therefore the development time may be relatively short. Furthermore, because dominant markers each characterize multiple regions of the genome, they often show reasonably high levels of polymorphism that can be useful for inferring close

genetic relationships; however, dominant markers are generally unable to resolve more distant evolutionary relationships. Perhaps the biggest drawback to these markers is that their dominant nature means that only one allele can be identified at each locus, and therefore heterozygotes cannot be differentiated from homozygotes. The presence of a band means that an individual is either homozygous (*AA*, with both alleles producing fragments) or heterozygous (*Aa*, with only the dominant allele producing a fragment) at that particular locus. Individuals that are homozygous recessive will not produce a band.

The inability to differentiate between homozygotes and heterozygotes makes it difficult to calculate allele frequencies from dominant markers. Methods for doing this do exist but several assumptions must be made, such as the population being in Hardy–Weinberg equilibrium, and as we shall see in the next chapter this is not always the case. The anonymity of dominant markers also can make it difficult to detect contamination and to compare data between studies. Nevertheless, as we will see below, dominant markers have been employed successfully in many studies of molecular ecology.

Random amplified polymorphic DNA

In 1990 a PCR-based technique known as **random amplified polymorphic DNA** (RAPDs) was introduced as a method for genotyping individuals at multiple loci (Welsh and McClelland, 1990; Williams *et al.*, 1990). RAPDs are generated using short (usually 10 bp) random primers in a PCR reaction. As there are about one million possible primers that can be made from ten bases, there is plenty of scope for detecting polymorphism by using a different primer in each reaction. A single primer is added to each PCR reaction, and multiple bands arise by chance when the primer happens to anneal to two reasonably proximate sites. The banding pattern for each individual will depend on where suitable primer binding sites are located throughout that individual's genome (Figure 2.8). Although RAPDs can provide a relatively quick and straightforward method for quantifying the genetic

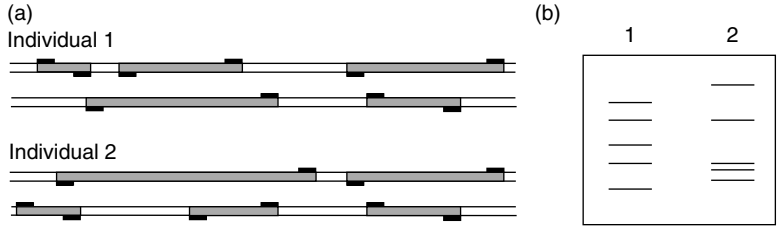

Figure 2.8 **(a) RAPD priming sites (indicated by black boxes) are distributed throughout the genome, although here only two partial chromosomes are represented. The sizes of the products (shaded in grey) that will be amplified during PCR will depend on the locations of these priming sites. (b) Diagrammatic representation of the gel that would follow PCR-RAPD screening of these two individuals. Recall that the rate at which a band migrates through the gel is inversely proportional to its size**

similarity of individuals, its reproducibility depends on stringent laboratory conditions. The amplification of bands will often vary, depending on a variety of factors including the starting concentration of DNA, the parameters that are used in the PCR cycle, the type of PCR machine used (results often vary between laboratories) and the particular brand of the reagents used. Furthermore, the random nature of RAPDs means that it can be difficult to identify bands that have been amplified from non-target DNA.

If the problems of reproducibility and contamination are overcome, RAPDs can be used to estimate the genetic similarity between two individuals based on the number of bands they share. However, despite their ease of use, a general lack of reproducibility, combined with their dominant nature, has decreased the popularity of RAPDs in recent years. The journal *Molecular Ecology* actively discourages researchers from submitting manuscripts that report population genetic studies that are based primarily on RAPD data, in part because other more reliable markers now are widely available. That is not to say that there is no role for RAPDs at all. As the *Molecular Ecology* editorial board points out, RAPDs can be useful in genetic mapping studies. They can also provide markers that are diagnostic of a given species or trait if RAPD screening identifies a band that consistently differs between two groups. This was the case in the toxic dinoflagellate *Gymnodinium catenatum*, which apparently was introduced to Tasmanian waters in the ballast water of cargo ships (Bolch *et al.*, 1999). Unique RAPD banding patterns from different populations allowed the authors of this study to eliminate Spain, Portugal and Japan as source populations, and led them to conclude that the populations causing algal blooms around Tasmania were restricted largely to local estuaries. They did not, however, have enough data to pinpoint the source of the original Tasmanian introductions.

Amplified fragment length polymorphisms

A more labour intensive, but also more reliable, method than RAPDs for generating PCR-based multi-band profiles is known as **amplified fragment length polymorphism** (**AFLPs**) (Vos *et al.*, 1995). AFLP markers are generated by first digesting DNA with two different restriction enzymes that cut the DNA so that one strand overhangs the other strand by one or a few bases, thereby producing overlapping ('sticky') ends. Meanwhile, short DNA linkers are synthesized so that one end of the linker is compatible with the overhanging sequence of DNA. The linkers are ligated to the original DNA fragments, leaving a collection of fragments that have the same DNA sequence at the end. These fragments can then be amplified using primers that anneal to the linker DNA sequence (Figure 2.9). Specificity of primers is usually increased by adding one to three nucleotides at one end of the sequence, because PCR requires a perfect match between the target sequence and the 3′ end of the primer. This results in the amplification of multiple

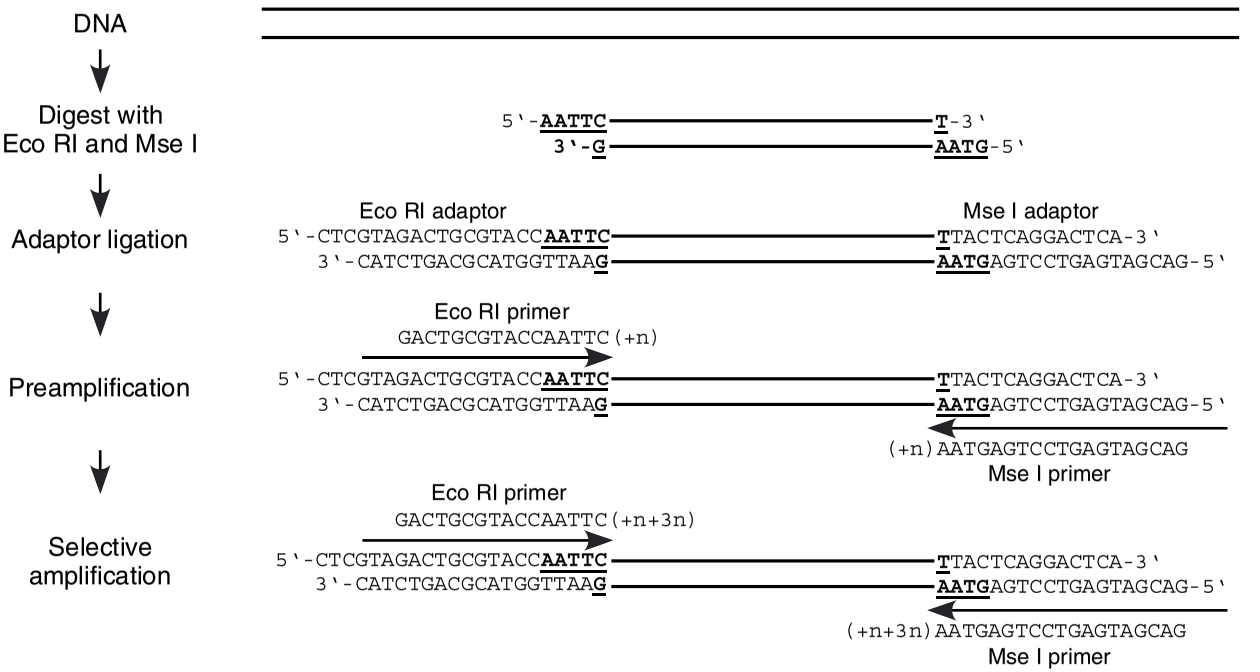

Figure 2.9 Schematic diagram showing how AFLP genotypes are generated. Digestion with two restriction enzymes produces sticky ends to which linkers can be ligated. During preamplification, the addition of a single base to the 3′ end of each primer will reduce the number of amplified fragments to 1/16 of the number of fragments that otherwise would be amplified. The addition of three more bases to the 3′ primer ends during selective amplification further reduces the chance of a perfect match between primers and target sequences, and as a result only 1/65 536 of the original set of fragments will be amplified

fragments that appear as a series of different-sized bands when run out on a gel. The pattern of bands will depend on the sequences that are immediately adjacent to the linkers, and also on the distances between the restriction sites. As with RAPDs, the generation of bands is essentially random; in contrast to RAPDs, however, this method has a much higher level of reproducibility and therefore has become a more popular method of dominant genotyping.

The genetic similarity of individuals and populations can be inferred from the numbers of AFLP bands that they have in common. Additional information can be obtained by modifying the standard AFLP method to study gene expression. By ligating linkers to digests of cDNA, researchers can compare the banding patterns of genes that have been expressed, as opposed to the entire genome. This method was used to compare two genetically distinct lines of the endoparasitic wasp *Venturia canescens* that differ in a number of ways, including oviposition behaviour, numbers of eggs laid and growth rates in the early stages of embryonic development. Researchers used the cDNA-AFLP method to compare gene expression in ovarian tissue between the two groups and found differences in a number of expressed genes, including some that apparently are involved in the regulation of protein degradation during stress responses (Reineke, Schmidt and Zebitz, 2003). This study provided some interesting suggestions about the importance of gene expression during early development.

Overview

A range of molecular markers are available for studying populations in the wild (see Table 2.5). Different types of markers will provide different sorts of information, depending, for example, on whether they are inherited biparentally or uniparentally, or are dominant or co-dominant. In the next chapter we will start to look at how various markers can be used to characterize genetically a single population, and in so doing we will start to discuss some of the different ways in which molecular data can be analysed.

Chapter Summary

- Different genomes are inherited in different ways. In sexually reproducing taxa, most of the nuclear genome is inherited biparentally. In most animals and higher plants, mitochondrial DNA is maternally inherited. Chloroplast DNA usually is inherited maternally in flowering plants and paternally in conifers.

- Animal mitochondrial markers are popular in molecular ecology because of their lack of recombination, high mutation rate, small effective population size and readily available universal primers.

Table 2.5 Summary of some of the properties of genetic markers

Genetic marker	Inheritance	Target genome	Development time[a]	Cost[b]	Comparison of data between studies	Suitability for inferring evolutionary relationships	Overall variability
Allozymes	Co-dominant	Nuclear	Low	Low	Limited	Limited	Low-moderate
PCR-RFLPs	Co-dominant	Nuclear Organelle	Moderate	Low	Limited	Limited	Low-moderate
DNA sequences	Co-dominant	Nuclear Organelle	Low-high	Moderate	Yes	High	Low-moderate
SNPs	Co-dominant	Nuclear Organelle	High	Moderate-high	Yes	High	Moderate
Microsatellites	Co-dominant	Nuclear Organelle	High	Moderate-high	Yes	Limited	High
RAPDs	Dominant	Nuclear	Low	Low	Limited	Limited	High
AFLPs	Dominant	Nuclear	Moderate	Moderate	Limited	Limited	High

[a]Assuming relevant markers have not been developed already for the species in question (or close relative). Note that some development is required for all markers.
[b]Cost here is relative, because all molecular work is expensive. Cost will be reduced if relevant markers have been developed already for the species in question.

- Plant mtDNA generally lacks recombination and has a relatively small effective population size, but mutation rates are much lower than in animals. Chloroplast DNA regularly undergoes recombination and therefore gene rearrangements are common.

- In mammals, the Y chromosome is the paternally inherited sex chromosome. Most of it does not undergo recombination, and the recent identification of variable regions means that this can be a useful marker for retracing male lineages.

- Obtaining data from markers with contrasting patterns of inheritance can be extremely useful for detecting past hybridization events and for differentiating between dispersal of pollen versus seeds in plants, and of males versus females in animals.

- Co-dominant markers provide locus-specific information that allows us to discriminate between homozygotes and heterozygotes and to calculate allele frequencies.

- Dominant data neither discriminate between homozygotes and heterozygotes nor provide accurate estimates of allele frequencies; however, choice of marker is affected by time, money and expertise, and initial development is often easier for dominant than for co-dominant markers.

- DNA sequence data are most suitable for inferring evolutionary histories, and SNPs are a newly emerging source of sequence data from multiple loci. Microsatellite loci are highly polymorphic and are appropriate for inferring recent events such as dispersal or mate choice, whereas multi-locus markers permit the rapid and simultaneous screening of several loci.

Useful Websites and Software

- Molecular Ecology Notes Primer database: http://tomato.bio.trinity.edu/home.html

- The Alaska Biological Science Centre website summarizing the cross-species utilization of microsatellite primers across a wide range of taxa: http://www.absc.usgs.gov/research/genetics/heterologous_primers.htm

- Microsatellite Analysis Server (MICAS) – an interactive web-based server to find non-redundant microsatellites in a given nucleotide sequence/genome sequence: http://210.212.212.7/MIC/index.html

- National Centre for Biotechnology Information science primer -- SNPs: variations on a theme: http://www.ncbi.nlm.nih.gov/About/primer/snps.html

- BioEdit software for the alignment and manipulation of sequence data: http://www.mbio.ncsu.edu/BioEdit/bioedit.html

Further Reading

Books

Avise, J.C. 2004. *Molecular Markers, Natural History and Evolution* (2nd edn). Sinauer Associates Sunderland, Massachusetts.

Goldstein, D.B. and Schlötterer, C. 1999. *Microsatellites: Evolution and Applications*. Oxford University Press, Oxford.

Hoelzel, R. 1998. *Molecular Genetic Analysis of Populations* (2nd edn). Oxford University Press, Oxford.

Review articles

Bonin, A., Bellemain, E., Eidesen, P.B., Pompanon, F., Brochmann, C. and Taberlet, P. 2004. How to track and assess genotyping errors in population genetics studies. *Molecular Ecology* **13**: 3261–3273.

Jobling, M.A. and Tyler-Smith, C. 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics* **4**: 598–612.

Mueller, U.G. and Wolfenbarger, L.L. 1999. AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution* **14**: 389–394.

Rokas, A., Ladoukakis, E. and Zouros, E. 2003. Animal mitochondrial DNA recombination revisited. *Trends in Ecology and Evolution* **18**: 411–417.

Sunnucks, P. 2000. Efficient genetic markers for population biology. *Trends in Ecology and Evolution* **15**: 199–203.

Review Questions

2.1. List some of the advantages and disadvantages of using mitochondrial markers in studies of animal population genetics.

2.2. Which genomes would you target, and why, if you wanted to:

(i) Compare pollen and seed flow in an angiosperm species.
(ii) Compare pollen and seed flow in a gymnosperm species.
(iii) Compare male and female dispersal in a mammal species.

2.3. Twenty-eight diploid individuals from the same population were genotyped at a single locus, and only two alleles (A_1 and A_2) were found. Ten individuals were homozygous

for A_1 and eight were homozygous for A_2. What are the frequencies of the two alleles in this sample?

2.4. Individuals 1 and 2 from the same population have the following sequences at a particular locus:

1. GATTATACATAGCTACTAGATACAGATACTATTTTTAGGGGCGTATGCTCGG
   ATCTATAGACCTAGTACTAGATACTAGGAAAACCCGTTGTGTCGCGTGCTGA

2. GATTATACATAGTTACTAGATACAGATACTATTTTTAGGGGCGTATGCTCGG
   ATCTATAGACCTAGTACTAGATACTAGGAAAACCCGTTGTGTCGCGTGCTGA

If these sequences are digested with two restriction enzymes – *Alu*I, which cuts DNA sequences of AGCT and *Rsa*I, which cuts DNA sequences of GTAC – how many bands will each sequence produce?

2.5. Two mtDNA lineages in a mouse population diverged around 500 000 years ago. If you are comparing a 500 bp stretch of DNA from a protein-coding region from two mtDNA haplotypes, approximately how many base pair differences would you expect to find, based on the data in Table 2.4?

2.6. What are some of the factors that need to be taken into account when deciding which molecular markers you would use in an ecological study?