1

Molecular Genetics in Ecology

What is Molecular Ecology?

Over the past few decades, molecular biology has revolutionized ecological research. During that time, methods for genetically characterizing individuals, populations, and species have developed at a truly impressive rate, and continue to provide us with a wealth of novel data and fascinating new insights into the ecology and evolution of plants, animals, fungi, algae, and bacteria. Molecular markers allow us, among other things, to quantify genetic diversity, track the movements of individuals, measure inbreeding, identify the remains of individuals, characterize new species, and retrace historical patterns of dispersal. More recently, increasingly sophisticated genomic techniques have provided remarkable insight into the functioning of different genes, and the ways in which evolutionary adaptations (or lack thereof) can influence the survival of organisms in changing environments. All of these applications are of great academic interest, and are also frequently used to address practical ecological questions such as which endangered populations are most at risk from inbreeding, or how much hybridization has occurred between genetically modified crops and their wild relatives. Every year it becomes easier and more cost-effective to acquire molecular genetic data, and laboratories around the world can now regularly accomplish previously unthinkable tasks such as describing entire communities based on nothing more than remnant DNA extracted from water samples, or comparing a suite of functional genes between individuals from different populations.

This third edition of *Molecular Ecology* has been substantially overhauled because of the tremendous leaps and bounds that have occurred in this field over the past few years. Arguably the most important development of the past decade has been the introduction and increasing cost-effectiveness of high throughput sequencing; this technology was initially limited to a few labs with hefty research budgets, but is now accessible to a large community of researchers who are able to obtain sequence data sets about which they could previously only dream (Figure 1.1). When this book was first published in 2005, a major reason for the excitement surrounding molecular ecology was the ease with which researchers could obtain genetic data from natural populations. While this is still true, the main difference between then and now is that studies conducted prior to 2005 were based on a handful of loci (gene regions), whereas molecular ecology studies are now often based on much larger numbers of loci, or in some cases

Molecular Ecology, Third Edition. Joanna R. Freeland.

Companion website: www.wiley.com/go/freeland/molecularecology

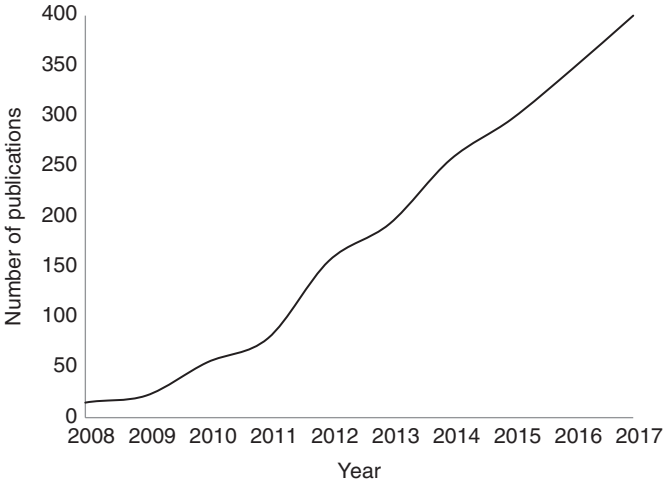


Figure 1.1 Numbers of results, by year, of a search in Web of Science that queried "next-generation sequencing" or "high throughput sequencing" and restricted results to the categories of "ecology," "biodiversity conservation," or "marine freshwater biology."

entire genomes. As a result, we now have greater insight into virtually all of the topics covered in this book, including population genetics, evolutionary change, conservation genetics, and behavioral ecology. This first chapter introduces high throughput sequencing (HTS) as a topic that will be revisited in subsequent chapters. Other technologies that are becoming increasingly widespread in ecological studies, and which will be discussed in later chapters, include environmental DNA (eDNA) assays, metabarcoding, transcriptomics, and epigenetics. We will begin in this chapter by reviewing some principles of genetics and some widely used techniques that are essential to our understanding of molecular ecology.

DNA, RNA, and Protein

This section will provide a short review of the relationship between DNA, genes, and proteins, because this background is necessary in order to understand how molecular markers can be used to address ecological questions. **Prokaryotes**, which lack cell nuclei, have their DNA arranged in a closed double-stranded loop that lies free within the cell's cytoplasm. Most of the DNA within the cells of **eukaryotes**, on the other hand, is organized into **chromosomes** that can be found within the nucleus of each cell and which comprise the nuclear genome (also referred to as **nuclear DNA**, or **nrDNA**). Each chromosome includes a single DNA molecule that is divided into functional units called genes. The site that each gene occupies on a particular chromosome is referred to as its **locus** (plural **loci**). At each locus, different forms of the same gene may occur, and these are known as **alleles**.

Each allele is made up of a specific sequence of DNA. DNA sequences are determined by the arrangement of four nucleotides, each of which has a different chemical constituent known as a base. The four DNA bases are adenine (A), thymine (T), guanine (G),

and cytosine (C), and these are linked together by a sugar-phosphate backbone to form a strand of DNA. In its native state, DNA is arranged as two strands of complementary sequences that are held together by hydrogen bonds in a double helix formation. No two alleles have exactly the same DNA sequence, although the similarity between two alleles from the same locus can be very high.

The function of some genes is to encode a particular protein, and the process in which genetic information is transferred from DNA to RNA to protein is known as **gene expression**. The DNA sequence of a protein-coding gene determines the structure of the protein that is synthesized. The first step of protein synthesis occurs when the coding region of DNA is transcribed into **ribonucleic acid** (**RNA**) through a process known as **transcription**. The result of transcription is a **primary transcript**, which is a single strand of RNA complementary to DNA sequences. RNA is made from the same bases as DNA with the exception of uracil (U), which replaces thymine (T). In prokaryotes, this transcript is also the **messenger RNA (mRNA**). In eukaryotes, the **introns** (noncoding segments of DNA) are excised following a process known as RNA splicing, producing a mature mRNA that is complementary to the **exon** (protein-coding) DNA template. mRNA sequences are then translated into protein sequences following a process known as **translation** (Figure 1.2). Translation is possible because each RNA molecule can be divided into triplets of bases (known as **codons**), most of which encode one of 20 different **amino acids**; these are the constituents of proteins (Table 1.1).

Specific combinations of amino acids give rise to **polypeptides**, which may form either part or all of a particular protein or, in combination with other molecules, a protein complex. If the DNA sequences from two or more alleles at the same locus are sufficiently different, the corresponding RNA triplets will encode different amino acids, and this will lead to alternative forms of the same protein. However, not all changes in DNA sequences will result in different proteins. Table 1.1 shows that there is some redundancy in the **genetic code**, for example leucine is specified by six different codons. This redundancy means that it is possible for two different DNA sequences to produce the same polypeptide product. The genetic code also signals "start" and "stop" functions: a stop codon (UAA, UAG, or UGA) signals the end of transcription, whereas a start codon (AUG, which encodes the amino acid methionine) marks the beginning of translation. These stop and start codons are therefore critical to gene functioning because they provide one of the mechanisms that controls gene expression. Gene expression can also be influenced by physical modifications to DNA molecules, which will be discussed below in the section on epigenetics.

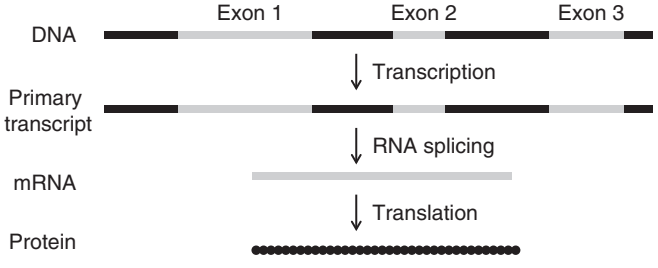


Figure 1.2 DNA codes for RNA via transcription. The mature mRNA transcript is then used as a template that is translated into a protein. This is known as the central dogma of molecular biology.

Table 1.1 The eukaryotic nuclear genetic code (RNA sequences). A total of 61 codons specify 20 amino acids. An additional three stop codons (UAA, UAG, UGA) signal the end of translation. This genetic code is almost universal, although minor variations exist in some microbes and also in the mitochondrial DNA (mtDNA) of animals and fungi.

		Second letter					
		U	**C**	**A**	**G**		
First letter	**U**	*UUU* } Phenylalanine *UUC* } (*Phe*) *UUA* } Leucine *UUG* } (*Leu*)	*UCU* *UCC* } Serine *UCA* } (*Ser*) *UCG*	*UAU* } Tyrosin *UAC* } (*Tyr*) *UAA* STOP *UAG* STOP	*UGU* } Cysteine *UGC* } (*Cys*) *UGA* STOP *UGG* Tryptophan (*Trp*)	U C A G	**Third letter**
	C	*CUU* *CUC* } Leucine *CUA* } (*Leu*) *CUG*	*CCU* *CCC* } Proline *CCA* } (*Pro*) *CCG*	*CAU* } Histidine *CAC* } (*His*) *CAA* } Glutamine *CAG* } (*Gln*)	*CGU* *CGC* } Arginine *CGA* } (arg) *CGG*	U C A G	
	A	*AUU* *AUC* } Isoleucine *AUA* } (*Ile*) AUG Methionine(Met), START	*ACU* *ACC* } Threonine *ACA* } (*Thr*) *ACG*	*AAU* } Asparagine *AAC* } (*Asn*) *AAA* } Lysine *AAG* } (*Lys*)	*AGU* } Serine *AGC* } (*Ser*) *AGA* } Arginine *AGG* } (arg)	U C A G	
	G	*GUU* *GUC* } Valine *GUA* } (*Val*) *GUG*	*GCU* *GCC* } Alamine *GCA* } (*Ala*) *GCG*	*GAU* } Asparticacid *GAC* } (*Asp*) *GAA* } Glutamicacid *GAG* } (*Glu*)	*GGU* *GGC* } Glycine *GGA* } (*Gly*) *GGG*	U C A G	

Allozymes

Ecology is a branch of biology that is primarily interested in how organisms in the wild interact with one another and with their physical environment. Historically, these interactions were studied through field observations and experimental manipulations. These observations and experiments typically included descriptions of **phenotypes**, which are based on one or more aspects of an organism's morphology, physiology, biochemistry, or behavior (Figure 1.3). What we may think of as traditional ecological studies have greatly enhanced our knowledge of many different species, and have made invaluable contributions to our understanding of the processes that maintain ecosystems. However, prior to the 1960s we knew very little about the genetics of natural populations; more specifically, we had little to no understanding of the genetic variation of populations, the genetic similarities among populations and species, the links between phenotype and **genotype** (an individual's complete set, or subset, of genes), and the roles of functional genes. For example, when individuals from the same population or species had different phenotypes, it was often unclear whether these were the result of genetic differences, or instead reflected **phenotypic plasticity**; the latter arises when a single genotype can develop into multiple alternative phenotypes depending on environmental conditions (Figure 1.4). In the 1960s our understanding of population genetics began to change when a method known as starch gel **electrophoresis** of allozymic proteins allowed biologists to obtain direct information about some of the genetic properties of individuals, populations, species, and higher taxa. These protein markers, most commonly referred to as **allozymes**, were used extensively for several decades and were responsible for a hugely important breakthrough in the emergent field of molecular ecology because they allowed researchers to genetically characterize individuals from almost any species.

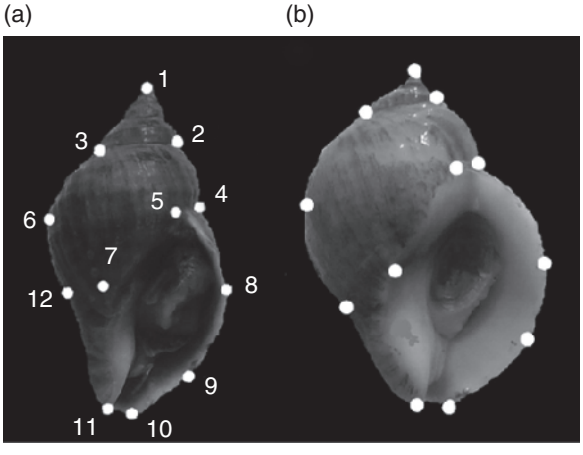


Figure 1.3 An example of ecological and evolutionary data based on a comparison of *Nucella lapillus* phenotypes. The numbered points on the shells were used to collect data for morphometric analysis. The shell on the left was collected from a site that was sheltered to a fairly large degree from wave action, whereas the more robust shell on the right was collected from a site that was exposed to strong wave action. A combination of reciprocal transplants, common garden experiments, and heritability assessments allowed the authors of the study to conclude that differences in shell shape can be explain by a combination of genes and phenotypic plasticity. *Source:* figure reproduced from Pascoal et al. (2012).

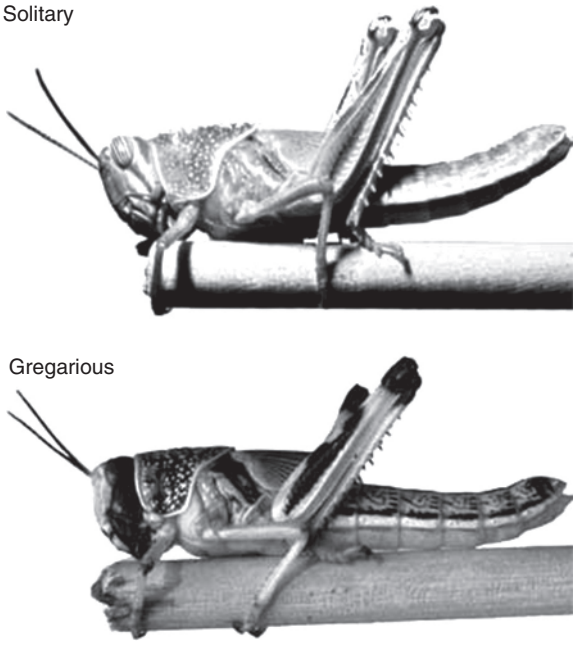


Figure 1.4 The desert locust (*Schistocerca gregaria*) exhibits an extreme form of phenotypic plasticity. At low population densities, green solitary locusts (top) are cryptic in both behavior and appearance, and avoid other locusts. When populations become crowded, locusts transform into the highly active brown/yellow gregarious phase (bottom), forming dense migratory swarms that can be catastrophic to crops. There are many differences between the two forms, including larger brains in gregarious locusts (Ott and Rogers 2010). Differences between the two forms can be explained by different patterns of gene expression (Badisco et al. 2011). *Source:* Photo attributable to NASA. https://commons.wikimedia.org/wiki/File:DesertLocust.jpeg.

Allozymes are allelic forms (alleles) of the same protein-coding locus (Lewontin and Hubby 1966), and were the first type of molecular marker that allowed us to infer levels of genetic variation of individuals and populations. At the individual level, a **diploid** organism that has two copies of the same allele at a particular locus is **homozygous** at that locus, and lacks variation at that particular locus (although may have variation at other loci). At the population level, a relatively large number of alleles within a population (i.e. the total of all the different alleles from all the sampled individuals) means that the population has relatively high levels of genetic variation. This distinction between individuals and populations will be made repeatedly throughout this book, as it is fundamental to many applications of molecular ecology. Keep in mind that data are usually collected from individuals, but if the sample size from any given population is sufficiently large then we often assume that the individuals collectively provide a good representation of the genetic properties of that population.

It is difficult to overstate the importance of allozymes as the first tool for quantifying genetic diversity in individuals and populations from many different taxonomic groups (Allendorf 2017). However, although tremendously important in their day, allozymes are now seldom used in molecular ecology (Figure 1.5) for a number of reasons. First, not all variation in protein-coding DNA sequences translates into variable protein

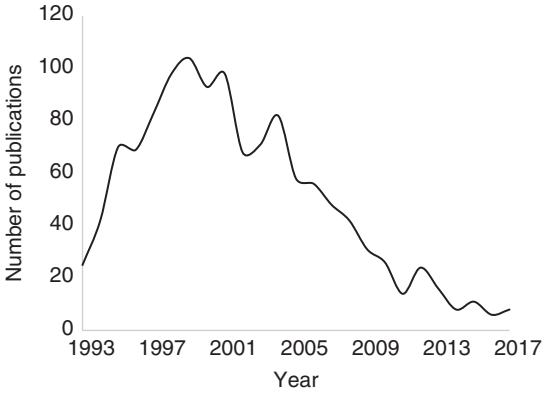


Figure 1.5 Numbers of results, by year, of a search in Web of Science that queried "allozymes" and restricted results to the categories of "ecology," "biodiversity conservation," or "marine freshwater biology."

products because, as noted earlier, the same amino acid can be encoded by multiple sequences. Second, a wealth of information is contained within every organism's **genome**, and allozyme studies capture only a small portion of this because they provide information only from protein-coding genes; to give just one example, the human genome project revealed that <2% of the ~3.3 billion nucleotides that comprise the human genome encode proteins (Lander et al. 2001). Third, in many cases the acquisition of allozyme data is both cumbersome and inhumane, because organisms often have to be killed before adequate tissue can be collected, and this tissue must then be stored at very cold temperatures (< −70 °C), which is a logistical challenge in most field studies. All of these drawbacks can be overcome by using appropriate DNA markers and sequences, which are now by far the most common source of data in molecular ecology because they have the potential to provide an endless source of information while allowing a more humane approach to sampling organisms of interest.

DNA: An Unlimited Source of Data

Even very small organisms have extremely complex genomes. The unicellular yeast *Saccharomyces cerevisiae* is so small that around 4 billion of them can fit in a teaspoon, but it nevertheless has a genome size of around 12 megabases (Mb; 1 Mb = 1 million base pairs) (Goffeau et al. 1996). The largest recorded genome belongs to *Paris japonica*, a rare Japanese plant with a genome size of approximately 149 billion base pairs (Pellicer et al. 2010). Clearly there is tremendous variation, with genome sizes varying up to 200 000-fold in eukaryotes (Gregory 2001), and up to 7000-fold in animals (Palazzo and Gregory 2014). Furthermore, genome size in eukaryotes seems to bear no relationship to organismal complexity (Hjelmen et al. 2017): even closely related species can have dissimilar genome sizes, for example in the fruitfly genus *Drosophila*, *D. orena* has a genome that is 1.6 times larger than that of *D. melanogaster* (Boulesteix et al. 2005).

Regardless of the size of the genome, each harbors a tremendous diversity of DNA. This diversity is partly attributable to the different types of protein-coding genes.

These include gene families, which are sets of several similar genes that arose following duplication of a single original gene. A well-studied example of a gene family is the **major histocompatibility complex** (MHC), which encodes a set of cell surface proteins that play an essential role in the immune response of vertebrates because they allow organisms to recognize foreign molecules. In contrast, **single copy nuclear DNA** (scnDNA) occurs only once in a genome. In addition to diverse protein-coding genes, the majority of DNA does not encode a protein. Non-coding DNA includes functional elements that are responsible for critical functions such as gene regulation. There are also introns (intervening sequences; Figure 1.2) and **pseudogenes**; the latter were derived from functional genes but have undergone mutations that prevent transcription. Another substantial source of non-coding DNA is the integration of retroviral elements and transposons throughout the genome (reviewed in Sultana et al. 2017). Some DNA sequences can also be defined as repeat motifs, which are short, highly repetitive sequences that include **minisatellites** (motifs of 10–100 bp repeated multiple times in succession) and **microsatellites** (repeat motifs of 1–6 bp; discussed in more detail in Chapter 2).

Although the structures and functions of genes vary between species, they are typically conserved among members of the same species, although that does not mean that all members of the same species are genetically alike. Variations in both coding and non-coding DNA sequences mean that no two individuals have exactly the same genome. This is because DNA sequences are structurally altered by events during replication that include recombination, duplication and mutation, and are functionally altered by epigenetic changes. It is worth examining in some detail how these processes occur, because if we remain ignorant about the mechanisms that generate DNA variation then our understanding of genetic diversity will be incomplete.

Mutation and Recombination

Novel genotypes arise from two processes: **mutation** and **recombination**. Most mutations occur during **DNA replication**, when the sequence of a DNA molecule is used as a template to create new DNA or RNA sequences during critical processes such as reproduction, gene expression, or cell growth. During replication, the hydrogen bonds that join the two strands in the parent DNA duplex are broken, thereby creating two separate strands that act as templates along which new DNA strands can be synthesized. The mechanics of replication are complicated by the fact that the synthesis of new strands can occur only in the 5′–3′ direction (Figure 1.6). Synthesis requires an enzyme known as **DNA polymerase**, which adds single nucleotides along the template

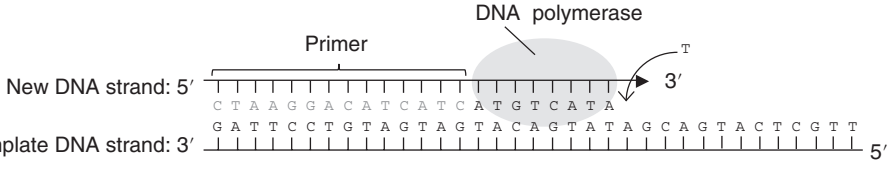


Figure 1.6 During DNA replication, the enzyme DNA polymerase adds nucleotides one at a time following a denatured (single-stranded) DNA template strand to make a new strand that grows in a 5′–3′ direction. In eukaryotes, replication is bi-directional and can be initiated at multiple sites by a primer (a short segment of DNA, shown in gray text in this figure).

strand in the order necessary to create a complementary sequence in which G is paired with C, and A is paired with T (or U in RNA). Successive nucleotides are added until the process is complete, by which time a single parent DNA duplex (double-stranded segment) has been replaced by two daughter duplexes that each comprise an old strand and a new strand.

Errors in DNA replication can lead to **nucleotide substitutions** if one nucleotide is replaced with another. These can be either **transitions**, which involve changes between either **purines** (A and G) or **pyrimidines** (C and T), or **transversions**, which arise when a purine is replaced by a pyrimidine or vice versa. Generally speaking, transitions are much more common than transversions. When a substitution does not change the encoded amino acid, it is known as a **synonymous substitution** because the DNA sequence has been altered, but the encoded product remains the same (Table 1.1). Alternatively, **non-synonymous substitutions** occur when a nucleotide substitution changes the encoded amino acid, in which case the function of that stretch of DNA may be altered. **Nonsense substitutions** arise when a nucleotide substitution results in a stop codon, and when this happens not all of the DNA sequence will be transcribed and therefore the encoded protein will be incomplete. Although single nucleotide changes often have no phenotypic outcome (i.e. when synonymous), they have the potential to be highly significant. Sickle-cell anemia in humans is the result of a single base pair change that replaces a glutamic acid with a valine, a mutation that is generally fatal in individuals that are homozygous for the sickle-cell allele.

Errors in DNA replication also include **nucleotide insertions** or **deletions** (collectively referred to as **indels**), which occur when one or more nucleotides are either added to, or removed from, a sequence. If an indel occurs in a coding region it often shifts the reading frame of all subsequent codons, in which case it is known as a **frameshift mutation**. When this happens, the gene sequence usually becomes dysfunctional. Mutations can also involve **slipped-strand mispairing**, which sometimes occurs during replication if the daughter strand of DNA becomes temporarily dissociated from the template strand. If this occurs in a region of a repetitive sequence such as a microsatellite repeat, the daughter strand may lose its place and re-anneal to the "wrong" repeat. As a result, the completed daughter strand will be either longer or shorter than the parent strand because it contains a different number of repeats (see also *Microsatellites*, Chapter 2).

Mutations are by no means restricted to one or a few nucleotides. **Gene conversion** occurs when one allele at a locus apparently converts the other allele into a form like itself. In the 1940s, Barbara McClintock discovered another example of gene alterations, **transposable elements**, which are sequences that can move to one of several places within the genome. Not only are these particular elements relocated, but they may also take with them one or more adjacent genes, resulting in a relatively large-scale rearrangement of genes within or between chromosomes. Transposable elements can interrupt function when they are inserted into the middles of other genes, and can also replicate so that their transposition may include an increase in their copy number throughout the genome. Genetic material can also move between individuals of either the same or different species following a process known as **horizontal gene transfer.** This is a completely different process from the heritable transmission of genetic material from parents to offspring (which is sometimes referred to as vertical transfer), and is one of the main processes that facilitates the spread of antibiotic resistance genes among bacteria.

The other key process that alters DNA sequences is recombination. Most individuals start life as a single cell, and this cell and its derivatives must replicate many times during the growth and development of an organism. This type of replication is known as **mitosis**, and involves the duplication of an individual's entire complement of chromosomes – in other words, the daughter cells contain the same number and types of chromosomes as the parental cells. Mitosis occurs regularly within **somatic** (non-reproductive) cells. While necessary for normal body growth, mitosis would cause difficulties if it were used to generate reproductive cells. Sexual reproduction typically involves the fusion of an egg and a sperm to create an embryo. If the egg and the sperm were produced by mitosis then they would each have the full complement of chromosomes that were present in each parent, and the fused embryo would therefore have twice as many chromosomes as either of its parents. This number would double in each generation, rapidly leading to an unsustainable amount of DNA in each individual. This is circumvented by **meiosis**, a means of cellular replication that is found only in **germ** cells (cells that give rise to eggs, sperm, ovules, pollen, and spores). In **diploid** species (Box 1.1), meiosis leads to **gametes** that have only one set of chromosomes (n), and when these fuse they create a diploid (2n) embryo. During meiosis, recombination occurs when **homologous chromosomes** exchange genetic material. This leads to novel combinations of genes along a single chromosome (Figure 1.7), and is an important contributor to genetic diversity in sexually reproducing **taxa**.

Epigenetic Marks

As discussed above, the most important source of genetic variation is changes in nucleotide sequences that result from DNA mutations, with recombination also playing a pivotal role in the generation of genetic diversity. However, variation in organismal phenotypes can also be influenced by **epigenetic** changes that alter gene expression without altering gene sequences. Epigenetic means above the genome, and includes DNA **methylation**, which occurs when a methyl (CH_3) group is added to a cytosine nucleotide. Excess methylation typically leads to a reduction in transcription and hence gene expression. A second important mechanism of epigenetic change is **histone** modification. Histone proteins act as spools around which DNA is wound, and modifications to histones cause chromatin (DNA + histone) to be more or less tightly wound, which in turn makes DNA more or less accessible to transcription and hence either increases or decreases gene expression. In medical genetics there is considerable interest in how epigenetic changes can influence disease. Researchers used to think that all epigenetic marks were "erased" in embryos, and that all epigenetic change was acquired throughout an individual's lifetime; however, we now know that some epigenetic marks can be passed on from one generation to the next. This so-called epigenetic inheritance is not well understood, for example we don't know for how many generations it can persist, but it provides a fascinating addition to the much better understood pattern of heritable mutations in DNA sequences. Although initially the focus of medical genetic studies, epigenetics is of growing interest in ecology because in some cases epigenetic modifications that alter patterns of expression across different genes may help individuals from the same species to tolerate different environments, even if their overall genetic similarity is high. In Chapter 2 we will look at how molecular markers can help us to infer epigenetic changes among individuals sampled from wild populations.

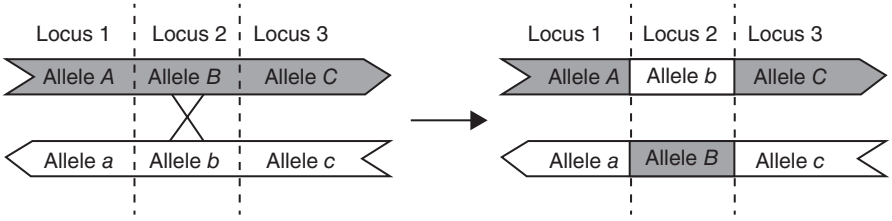


Figure 1.7 An example of recombination at the gene level, showing how the gene sequence at chromosome 1 can change from ABC to AbC. Recombination typically leads to novel combinations of alleles along a chromosome.

Box 1.1 Chromosomes and Polyploidy

The **karyotype** (the complement of chromosomes in a somatic cell) of many species includes both **autosomes**, which usually have the same complement and arrangement of genes in both sexes, and **sex chromosomes**. The number of copies of the full set of chromosomes determines an individual's ploidy. **Diploid** species have two sets of chromosomes (2n), and if they reproduce sexually then one complete set of chromosomes will be inherited from each parent. Humans are diploid, and have 22 pairs of **autosomes** and two **sex chromosomes** (either two X chromosomes in a female or one X and one Y chromosome in a male), which means that their karyotype is 2n = 46 (22 autosomes plus one sex chromosome, multiplied by two because they are diploid). The total number of chromosomes varies between species (Figure 1.8). **Polyploid** organisms have more than two complete sets of chromosomes. In **autopolyploid** individuals, all chromosomes originated from a single ancestral species after chromosomes failed to separate during meiosis. In this way, a diploid individual (2n) can give rise to a tetraploid individual (4n), which would

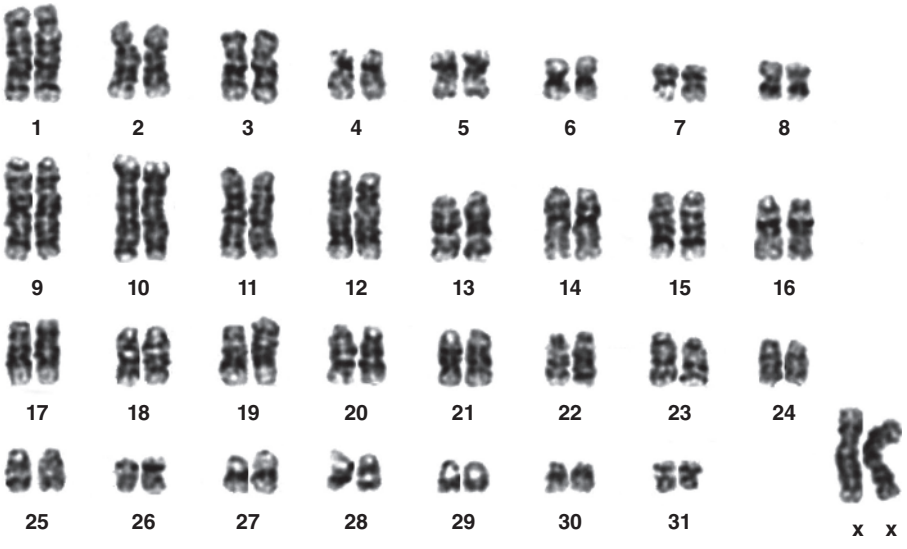


Figure 1.8 Karyotype of female nine-banded armadillo (*Dasypus novemcinctus*). The normal karyotype in this diploid species, as shown here, includes 31 pairs of autosomes and one pair of sex chromosomes (in this case two X chromosomes), which is reported as 2n = 64. *Source:* figure reproduced from Svartman et al. (2006).

have four copies of the original set of chromosomes. This contrasts with **allopolyploid** individuals, which have chromosomes that originated from multiple species following hybridization. The creation of new polyploids sometimes results in the formation of new species, although a single species can comprise multiple races, or **cytotypes**.

Polyploidy is very common in flowering plants, and also occurs to a lesser degree in fungi, vertebrates (primarily fishes, reptiles, and amphibians), and invertebrates (including insects and crustaceans). Polyploidy is of ecological and evolutionary interest for a number of reasons, including the fact that polyploids are overrepresented in invasive plant populations. One thought is that polyploids can be successful biological invaders because even though small colonizing populations can often experience inbreeding, which in turn can lead to a reduction in fitness, the extra chromosomes associated with polyploidy could provide a buffer against low genetic diversity and make some non-native species more immune to the effects of inbreeding. For example, introduced polyploid spotted knapweed (*Centaurea stoebe*) populations were less likely to suffer from inbreeding depression than native diploid populations of comparable population sizes (Rosche et al. 2017). Another example of how polyploidy can be relevant to ecology comes from the fact that polyploid and diploid conspecifics are often able to co-exist through habitat partitioning. Diploid and tetraploid populations of the food-deceptive orchid *Anacamptis pyramidalis*, for example, are reproductively isolated from one another for a number of reasons, one reason being their occupation of different microhabitats, which can most likely be explained by the availability of soil mycorrhiza (Pegoraro et al. 2016). There will be other examples throughout this text that show the relevance of ploidy to molecular ecology.

Genomes

When we talk about DNA, there may initially be a tendency to assume that we are talking about nuclear DNA that is inherited from both parents, leading to the well-known idea that individuals inherit half of their DNA from their mother and half of their DNA from their father. Indeed, the offspring of sexually reproducing organisms do 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** (**mitochondrial DNA**, or **mtDNA**) and **plastids** (**chloroplast DNA**, or **cpDNA**). Mitochondria and chloroplasts are located outside the cell nucleus (Figure 1.9). Mitochondria are found in both plants and animals, whereas chloroplasts, a type of plastid, are found only in plants. Organelle DNA typically occurs in the form of super-coiled 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. However, what they lack in size they partially make up for in number: a single human cell normally contains anywhere from 1000 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.

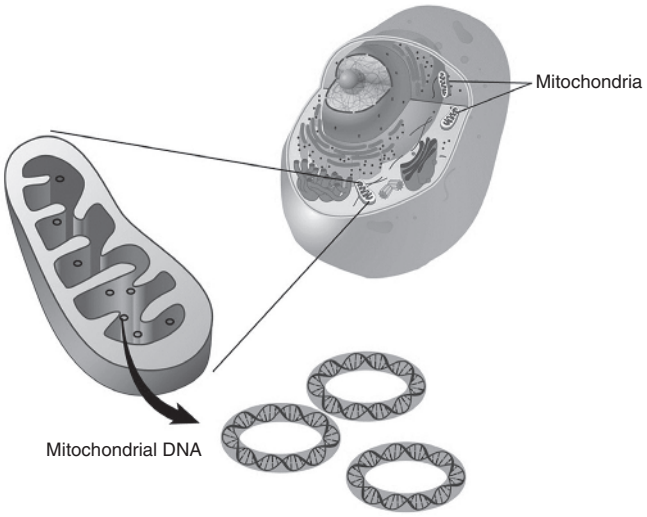


Figure 1.9 Mitochondria are organelles that are the sites of energy production within cells. Inside mitochondria is the small circular chromosome known as mitochondrial DNA (mtDNA), which normally follow maternal inheritance, in other words are passed down from mothers to offspring (sons and daughters). Figure attributed to National Institutes of Health.

Mitochondrial DNA (mtDNA)

Mitochondrial DNA produces RNA and proteins that are central to cellular respiration, the process by which energy is extracted from food. Animal mtDNA genomes typically range from 16 kb to 18 kb in size (kb = kilobase = 1000 bases), and contain 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, in other words they are non-repetitive, and there is little evidence of either spacer sequences between genes, or intervening sequences within 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 1.10). In most animals, mitochondrial DNA is **maternally inherited**, in other words it is passed down from mothers to their offspring (daughters and sons) but not from fathers to their offspring.

The overall function of plant and animal mitochondria is similar, but their structures differ markedly. The most obvious difference is their size: in contrast to the average size of ~16.5 kb in animals, plant mtDNA genomes range in size from 200 to 2500 kb. The increased and variable sizes are mostly a result of repeated sequences, large intron and non-coding sequences, plus integrated sequences from nuclear and chloroplast genomes. Gene and intron content also vary considerably among mtDNA for plant species, as evidenced by whole cpDNA genome sequences that have identified between 32 and 67 genes, and between 18 and 25 introns (Mower et al. 2012).

Chloroplast DNA (cpDNA)

Plant mtDNA sequences tend to be fairly conserved within species and this fact, combined with its relatively complex and variable structure, means that when haploid

Figure 1.10 Gene map and organization of the complete mitochondrial genome of the globose head whiptail *Cetonurus globiceps* (an Australian fish). The relatively large number of transfer RNA (tRNA) genes is typical of vertebrate mtDNA. Other common genes include subunits of NADH dehydrogenase (ND) and cytochrome oxidase (CO). *Source:* figure reproduced from Shi et al. (2016).

markers are desirable in plant studies, researchers more commonly turn to cpDNA genomes. Plastids, which include chloroplasts, are organelles in plants and algae that manufacture and store a range of chemical compounds that are used by the cell. Those that contain chlorophyll (green plastids) can carry out photosynthesis. In molecular ecology, we typically refer to plastid DNA as chloroplast DNA (cpDNA). Like mtDNA, cpDNA is maternally inherited in most **angiosperms** (flowering plants), whereas in most **gymnosperms (**conifers and cycads) it is usually paternally inherited. Chloroplast DNA genomes, which in most plants are key to the process of photosynthesis, typically range from 120 000 bp to 220 000 bp (the average size is around 150 000 bp; Figure 1.11). In many photosynthetic plants the cpDNA genome comprises 70 (gymnosperms) to 88 (liverworts) protein-coding genes and 33 (most eudicots) to 35 (liverworts) structural RNA genes, for a total of 100–120 different genes (Wicke et al. 2011). Although recombination sometimes occurs, chloroplasts are for the most part structurally stable, and

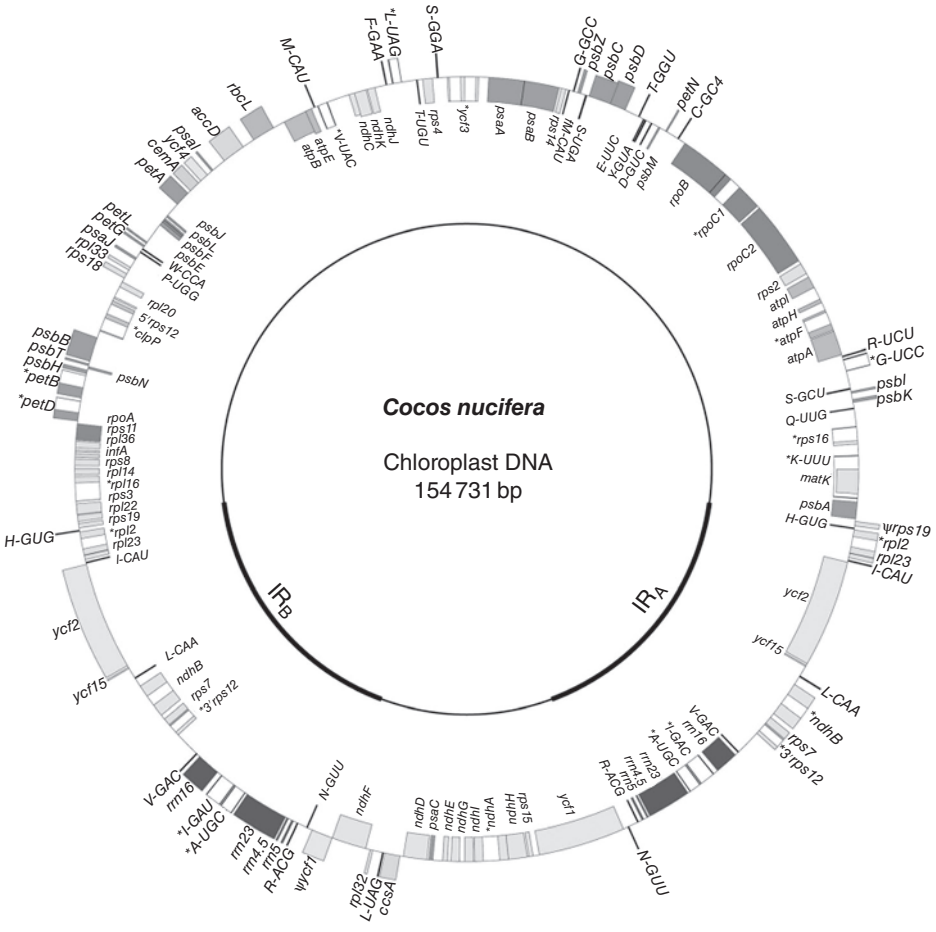


Figure 1.11 Map of the coconut palm (*Cocos nucifera*) chloroplast genome, which has a total size of 154 731 bp. Thick lines of the small inner circle indicate inverted repeats (IRs). Genes with introns are marked with "*". Pseudo genes are marked with "Ψ". *Source:* figure reproduced from Huang et al. (2013).

most of the size variation can be attributed to differences in the lengths of sequence repeats, as opposed to the gene rearrangement and duplication that is found in plant mtDNA.

As with mtDNA, the cpDNA genome has been sequenced in its entirety from multiple species, and this has greatly increased our knowledge about cpDNA genome size and structure. To give just a couple of examples, the cpDNA genome of the alga *Chlorella sorokiniana* comprises 109 811 bp, which encode a total of 109 genes, including 74 protein-coding genes, 3 rRNAs, and 31 tRNAs (Orsini et al. 2016). The goosegrass (*Eleusine indica*) chloroplast genome is somewhat larger at 135 151 bp, and has 108 unique genes, of which 76 are protein-coding genes, 28 are tRNAs, and 4 are rRNAs (Zhang et al. 2017). (Jiang et al. 2016) demonstrated some of the benefits of obtaining sequences from the entire cpDNA genome when inferring the evolutionary relationships among plants: they compared whole-genome cpDNA sequences at the

intraspecific, interspecific, and intergeneric levels in Cycades, which include some of the world's most threatened plant species. The chloroplast genome was 162 094 bp in length, and included 87 protein-coding, 37 tRNA, and 8 rRNA genes. The whole-genome comparison allowed them to find informative variable regions even in species with highly restricted geographic distributions.

Haploid Chromosomes

As noted above, when discussing the inheritance of nuclear and organelle markers we usually refer to nuclear genes as being biparentally inherited. 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 to which it is exposed during the early stages of development. Many other species follow **genetic sex determination**, which means that an individual's sex is genetically determined 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; Figure 1.8), 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 passes on 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 the offspring 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** (in other words it is **paternally inherited**) because it is passed down only through the male lineage, from fathers to sons (Table 1.2). Because there is never more than one copy of a Y chromosome in each set of chromosomes (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 pseudoautosomal regions at the tips of the chromosome that recombine with the X chromosome, but these two regions are separated by approximately 60 Mb of non-recombining sequence (Figure 1.12).

Polymerase Chain Reaction

A wealth of information in the genome is of no use to molecular ecologists if it cannot be accessed and quantified, and after 1985 this became possible thanks in large part to Dr. Kary Mullis, who invented a method known as **polymerase chain reaction** (PCR) (Mullis and Faloona 1987). This was a phenomenal breakthrough that allowed researchers to isolate and amplify specific regions of DNA from the background of large and complex genomes. The importance of PCR to many biological disciplines including molecular ecology cannot be overstated, and its contributions were recognized in 1993 when Mullis was one of the recipients of the Nobel Prize for Chemistry.

Table 1.2 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 plants	
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

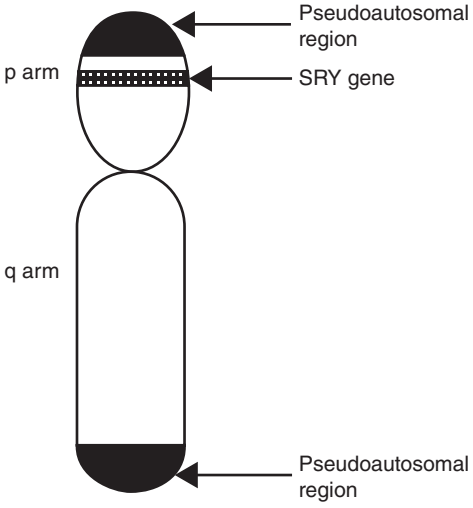


Figure 1.12 Mammalian Y chromosome. The SRY gene (sex-determining region Y) essentially converts an embryo into a male. The pseudoautosomal regions are the only regions that recombine with the X chromosomes during DNA replication.

The beauty of PCR is that it allows us to selectively amplify one or more genomic regions with relative ease. This is most commonly done by first isolating total DNA from a sample, and then using paired **oligonucleotide primers** to repeatedly amplify one or more target DNA regions until there are enough copies to allow subsequent manipulation and characterization. The primers, which are usually 15–35 bp long, provide a necessary starting point for DNA synthesis, and they must each be complementary to a stretch of DNA that flanks the target sequence so that they will anneal to the desired site and provide an appropriate starting point for replication.

Each cycle in a PCR reaction has three steps: denaturation of DNA, annealing of primers, and extension of newly synthesized sequences (Figure 1.13). The first step,

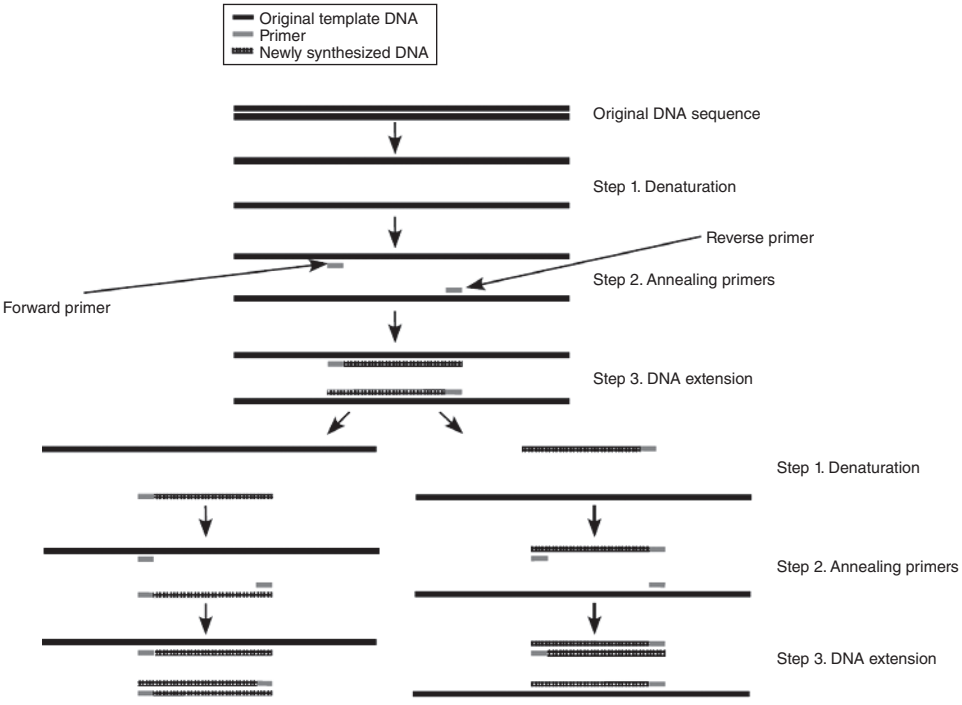


Figure 1.13 The first two cycles in polymerase chain reaction (PCR). Solid black lines represent the original DNA template, short gray lines represent the primers, and hatched lines represent DNA fragments that have been newly synthesized.

denaturation, is done by increasing the temperature to approximately 94 °C so that the hydrogen bonds will break and the double-stranded DNA will become single-stranded template DNA. The temperature is then dropped to a point, usually between 40 and 65 °C, that allows the primers to anneal to complementary sequences that flank the target sequence. The final stage uses DNA polymerase and the free nucleotides that have been included in the reaction to extend the primer sequences, generally at a temperature of 72 °C. Nucleotides are added in a sequential manner, starting from the 3′ primer ends, following the same method that is routinely used for DNA replication *in vivo* (Figure 1.6). Each PCR cycle generates two daughter strands for every parent strand, which means that the number of sequences increases exponentially throughout the PCR. A typical PCR follows 35 cycles, enough to amplify a single template sequence into 68 billion copies!

PCR uses a heat-stable polymerase, most commonly *Taq* polymerase, so-called because it was originally isolated from a bacterium called *Thermus aquaticus* that lives in hot springs. *Taq* is not deactivated at high temperatures, and therefore it needs to be added only once at the beginning of the reaction, which then runs in computerized thermal cyclers (PCR machines) that repeatedly cycle through different temperatures. Some optimization is generally required when using new primers or targeting the DNA of multiple species, for example altering the annealing temperature or using different salt concentrations to sustain polymerase activity. However, once the optimization is

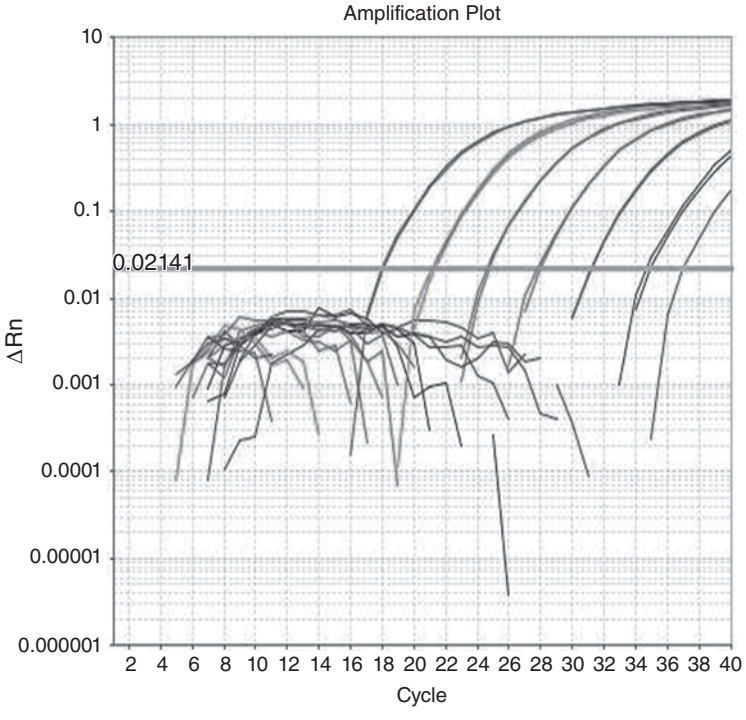
complete, all the researcher has to do is set up the reactions, program the machine, and come back when all the cycles have been completed, which usually takes 1–3 hours. By this time copies of the target region will vastly outnumber any background non-amplified DNA, and the final product can then be characterized in one of several different ways, some of which will be outlined later in this chapter, and also in Chapter 2.

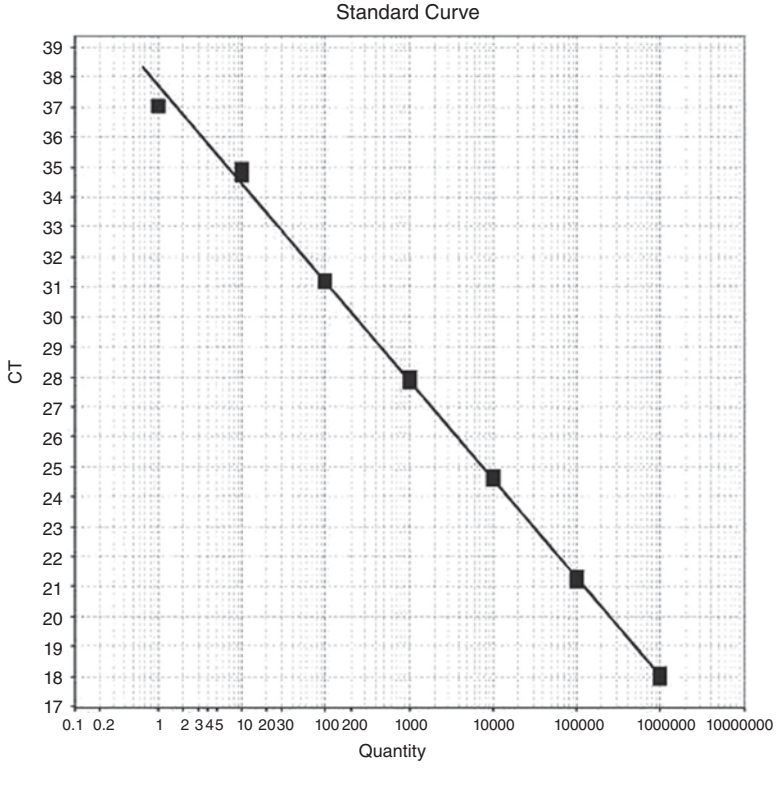
Quantitative PCR

As discussed above, the most common use of PCR is to generate a large amount of one or more target regions of DNA that can then be used to create genetic profiles of individuals, populations, and species. One thing conventional PCR cannot do, however, is supply us with accurate estimates of the amount of DNA that is present in a particular sample. This is because there is no correspondence between the amount of template at the start of the reaction, and the amount of DNA that has been amplified by the end of the reaction. In the 1990s, however, a technique known as **quantitative PCR** (qPCR, also known as **real-time PCR** [RTPCR]) was developed, and this does allow researchers to quantify the amount of DNA in a particular sample.

qPCR allows users to monitor a PCR reaction in "real time," in other words as it occurs, instead of waiting until all of the cycles in a PCR reaction have finished. qPCR reactions use the same components as those used in "standard" PCR, with one exception: the fragments produced in qPCR are labeled with either fluorescent dye probes or DNA binding dyes and can be quantified after each cycle. An excess of primers and polymerase is added to the qPCR reactions so that the amount of template DNA is the only factor that limits the numbers of DNA copies that are made in the initial rounds of PCR; as a result, the fluorescent signal incorporated into the amplicons is directly proportional to the amount of starting DNA. There is a correlation between the first significant increase in the amount of PCR product, and the total amount of the original template. qPCR can quantify DNA or RNA in either an absolute or a relative manner. Absolute quantification determines the number of copies that have been made of a particular template, usually by comparing the amount of DNA generated in each cycle to a standard curve based on a sample of known quantity (Figure 1.14). Relative quantification allows the user to determine which samples have more or less of a particular gene product.

There are several ways in which qPCR can benefit ecological studies. One advantage to qPCR is that it allows researchers to quantify the sensitivity of their assays, in other words determine how much DNA must be present in a sample before it is detected through PCR. Brandl et al. (2015) used qPCR to determine the sensitivity of assays that screened gut contents for specific fish species, with the goal of determining the extent to which introduced fish species are preying on native fishes in the San Francisco Estuary–Delta. qPCR can also be used to quantify gene expression: because RNA is transcribed only during gene expression, the amount of RNA in a sample is indicative of the amount of gene expression that is taking place. Researchers can extract RNA from organisms and reverse-transcribe (i.e. make DNA from RNA, the reverse of transcription) to make **complementary DNA** (**cDNA**). By quantifying cDNA with qPCR, it is possible to determine the extent to which the gene of interest was being expressed in the organism at the time of sampling. This may allow researchers to identify conditions that


Amplification Plot
10
1
0.1
0.02141
0.01
0.001
0.0001
0.00001
0.000001
ΔRn
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Cycle


Standard Curve
39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17
CT
0.1 0.2 1 2 3 4 5 10 20 30 100 200 1000 10000 100000 1000000 10000000
Quantity

lead to either **upregulation** (increased gene expression) or **downregulation** (decreased gene expression). qPCR of cDNA was used to quantify RNA in a study of coral reef organisms known as octocorals. Coral reefs are increasingly suffering from anthropogenic impacts such as increasing pH and temperature of oceans around the world, and Shimpi et al. (2016) used qPCR to compare the transcription levels of a suite of reference genes to see if gene expression altered in response to environmental stressors. One of the genes that they investigated was heat shock protein 70 (hsp70). Heat shock proteins, also known as stress proteins, are ubiquitous proteins that play a number of crucial roles including the protection of cells from stress, and in the octocoral genus *Sinularia* these were upregulated during thermal stress, and downregulated during low-pH stress. This study improved our mechanistic understanding of the ways in which octocorals respond to stress responses, and helped us to understand how they may respond to future abiotic changes in the ocean.

Sources of DNA

PCR and qPCR each require only a very small amount of starting template DNA, which means that we can genetically characterize individuals from an amazingly wide range of samples, most of which can be collected without causing lasting harm to the organism from which they originated. Broadly speaking, there are three different categories of DNA that are used in molecular ecology. The first type of DNA is genomic DNA that is extracted from either whole organisms (e.g. microbes, small invertebrates), or from organismal material such as blood, hair, feathers, leaves, roots, and other sources. The second category of DNA is community DNA, which describes a pool of genomic fragments extracted at the same time from multiple species and individuals that have been removed from their habitat, e.g. a group of microorganisms that were collected in a soil sample (Creer et al. 2016). The third broad category is eDNA, which refers to the extraction of remnant DNA from an environmental sample such as soil or water without the need to first isolate organisms (Taberlet et al. 2012).

Because only a tiny amount of DNA is necessary for a successful PCR reaction, lethal sampling of animals is no longer necessary before individuals can be genetically characterized. Examples of non-lethal and in some cases non-invasive samples that have been

Figure 1.14 Generation of a standard curve against which quantitative PCR (qPCR) amplifications can be compared in order to quantify the amount of DNA that was added to the qPCR cocktail. In this case the standard curve was based on a series of 10-fold concentration standards, in other words a series of samples to which known amounts of DNA had been added. (A) Amplification plot depicting fluorescent output (ΔR_n) at each qPCR. Each pair of amplification curves (seven in total) represents a different concentration of DNA standard ranging from 10^6 copies/reaction (far left) to 10^0 copies/reaction (far right), each of which was run in duplicate. Also shown are the quantification (C_q) threshold (ΔRn = 0.02141) below which amplified DNA is considered "noise," and typical background fluorescence (erratic lines below threshold). (B) qPCR standard curve generated during the same reaction as in (A). Each square corresponds to one amplification curve; C_T is the quantification cycle, and "quantity" is the DNA copy number per reaction (manually input during setup). By comparing qPCR reactions with unknown DNA quantities to these standard curves, it is possible to estimate the number of DNA copies that are present in a particular sample, hence qPCR. Figure attributed to Charise Currier.

successfully used for DNA analysis include fecal DNA to characterize elusive species such as large felids (Mesa-Cruz et al. 2016); fecal or spider web DNA to characterize small species such as spiders that in many cases would otherwise have to be killed when sampled (Blake et al. 2016; Sint et al. 2015); DNA from regurgitates, feces, or whole gut contents to identify prey species (Kamenova et al. 2018); bird feathers (Kleven et al. 2016); partial antennae from bees (Oi et al. 2013); body mucus swabs from fish or freshwater mussels (Cho et al. 2016; Le Vin et al. 2011); and hair samples collected in the wild using non-invasive hair snares (Kendall et al. 2009). See Beja-Pereira et al. (2009) for a review of how non-invasive genetic sampling has been used in wildlife research. When working with small samples, however, particular care must be taken to avoid contamination, because very small amounts of target DNA can easily be overwhelmed by "foreign" DNA. In addition, there may be challenges associated with inconsistent amplification of DNA from degraded samples such as those recovered from hair or scats which require either extensive confirmation (e.g. through repeat genotyping of single samples) or statistical adjustments that allow for genotyping errors (Knapp et al. 2009).

From a practical perspective, the storage of samples destined for PCR is relatively easy during field trips, because samples for PCR analysis can be stored either as dried specimens (e.g. leaf samples dried in envelopes that contain desiccating beads), or in small vials of 70–95% ethanol or buffer that can be kept at room temperature. The DNA in freshly harvested blood or tissue will remain in good condition provided it is quickly placed into suitable buffer or ethanol, but improperly stored DNA will rapidly degrade into fragmented DNA molecules. DNA extracted from a non-living sample, such as fecal material or a museum specimen, will already be degraded. If the DNA fragments in a degraded sample are smaller than the size of the target DNA region, then PCR amplification will be impossible; therefore, when working with degraded samples, relatively short DNA sequences should be targeted (Box 1.2).

PCR that simultaneously targets multiple species can be used to survey community DNA extractions from bulk samples of organisms, and this has been done for a range of taxonomic groups including fungal communities in soil, (Schmidt et al. 2013), invertebrate communities collected in traps (Ji et al. 2013), or planktonic microbial communities in lakes (Poretsky et al. 2014). eDNA, the third category, refers to DNA that has been shed into the environment by decaying bodies, mucus, blood, leaves, pollen, seeds, urine, feces, skin, and other types of organismal material. Most organisms leave a DNA signature when they pass through an environment, and this DNA can be used as evidence of their current or recent presence without actually having to see or handle the organism itself. Throughout this book, assume that discussions of DNA refer to genomic DNA extracted from individual organisms, unless otherwise specified. Community and eDNA will be discussed in more detail in Chapter 3.

Getting Data from PCR

Fragment Sizes

Once a particular gene region has been amplified from the requisite number of samples, it must be characterized in some way that allows the researcher to assign a genotype to each individual. The simplest way to do this is from the size of the amplified product,

Box 1.2 Museums and Herbaria: A Treasure Trove of Biological Data

Museums and herbaria around the world collectively harbor an estimated 3 billion specimens collected from approximately 2 million species from geographical locations around the world (Wheeler et al. 2012). In many cases, collections of a single species span decades, thereby presenting the potential for temporal sampling of populations without the need to travel back through time. Tremendous insight into phenotypic evolutionary changes has been obtained from temporal comparisons involving museum specimens. For example, the threespine stickleback fish (*Gasterosteus aculeatus*) from Lake Washington in Washington, USA, evolved greater amounts of protective plating over their bodies over a period of five decades, potentially in response to increased predation pressures (Kitano et al. 2008). Garden tiger moths (*Arctia caja*) evolved longer, narrower hindwings and narrower forewings over the past century, possibly driven by a need for increased dispersal abilities (Anderson et al. 2008) (Figure 1.15). Collections can also provide a wealth of long-term information from genetic data, for example the same garden tiger moth study described above extracted DNA from legs that had been removed from museum specimens, and used DNA sequencing to determine that the decline in moth numbers over the past few decades was accompanied by a decline in genetic diversity.

Other studies that have extracted and amplified DNA from animals preserved in museum collections include one that found mutant resistance alleles in insects that predated applications of the pesticide malathion, and therefore provided evidence of preadaption that could help to explain the rapid evolution of insecticide resistance in blowflies (Hartley et al. 2006). In conservation genetics, temporal declines in the genetic diversity of species-at-risk have been found in multiple species including red grouse (*Lagopus lagopus*) (Freeland et al. 2007) and white-headed ducks (*Oxyura leucocephala*) (Jacobs and Latimer 2012). Studies such as these are possible because the advent of PCR allowed

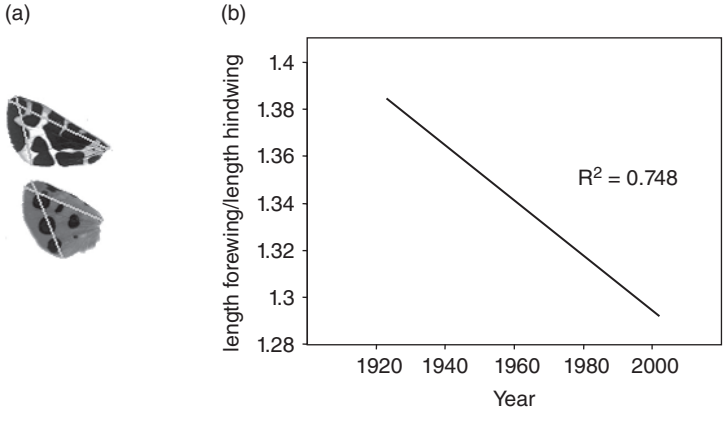


Figure 1.15 Forewing and hindwing images from garden tiger moths (*Arctia caja*) (a) that had been collected from around the UK over decades and kept in multiple museum collections revealed a historical decline in both forewing and hindwing width relative to wing length, and an increase in hindwing length (b). This trend may reflect selection over time for individuals that are increasingly able to disperse over relatively long distances following habitat fragmentation, and was accompanied by a decrease in genetic diversity (Anderson et al. 2008).

researchers to obtain sufficient amounts of target DNA from a very small amount of starting tissue such as a small piece of skin, the base of a feather, a leaf fragment, or a fish scale.

Although a potentially valuable source of data, DNA in museum and herbarium collections is typically degraded, often to a substantial degree (reviewed in Burrell et al. 2015). As a result, successful DNA amplification and sequencing of archived samples can be patchy, and has so far been limited largely to targeting gene regions that are present in high copy number numbers, most commonly mtDNA or cpDNA regions because both of these genomes are copied many times in each cell. However, the technology associated with HTS (see below) yields more data from archived specimens compared with earlier sequencing technologies because these more recent methods are designed to use short, fragmented DNA molecules as sequencing templates. Using HTS, researchers have been able to sequence the entire nuclear genome of a 43-year-old *Arabidopsis thaliana* (Brassicaceae) herbarium specimen, and substantial coverage of fungal specimens up to 82 years old (Staats et al. 2013). In another study, researchers used HTS to generate ~4 Mb of sequence data from early twentieth-century alpine chipmunks (*Tamias alpinus*) museum skins, and determined that genetic diversity was not reduced by a climate-related range retraction in the high Sierra Nevada area of California, USA (Bi et al. 2013).

For extinct species, museum collections represent the only source of data. Anmarkrud and Lifjeld (2017) were able to sequence the entire mitochondrial genome from each of 11 extinct bird species. The potential for future genetic work based on existing collections is therefore more promising than ever. Museum collections can also alleviate some of the limitations that are based on fieldwork logistics, for example the Natural History Museum in London, the American Museum of Natural History, and the Smithsonian's Natural History Museum collectively harbor >30 000 primate samples (Burrell et al. 2015). Archived biological material therefore holds tremendous promise for genetic research in taxonomy (including of extinct species) and conservation biology, and can provide a viable alternative to long-term experiments when investigating the historical effects of processes such as fluctuating population sizes or adaptation to changing environments.

which can be quantified by running out the completed PCR reaction on an agarose gel. The gel solutions are made by adding agarose powder to buffer and heating it until the solution becomes liquid. The solution is then poured into a tray that is edged by removable "walls," with combs that are left in place while the gel cools and becomes solid. At this point the combs are removed, leaving wells at one end of the gel. The gel "walls" are then removed and the gel is covered with buffer DNA, samples are loaded into the wells, and an electrical field is applied. DNA molecules are negatively charged, and will therefore migrate toward the positive electrode.

The speed at which fragments of DNA migrate through electrophoresis gels depends primarily on their size, with the largest fragments moving most slowly. Once DNA fragments have segregated across a gel they can be visualized using a dye such as ethidium bromide that binds to DNA molecules and can be seen with the human eye when illuminated with short-wavelength ultraviolet light. The sizes of DNA bands can then be extrapolated from ladders that consist of DNA fragments of a known size (Figure 1.16). Agarose gels are useful for determining whether PCRs successfully amplified a fragment of DNA that is approximately the expected length; however, precise estimates of band sizes must be obtained from more sophisticated genotyping equipment such as

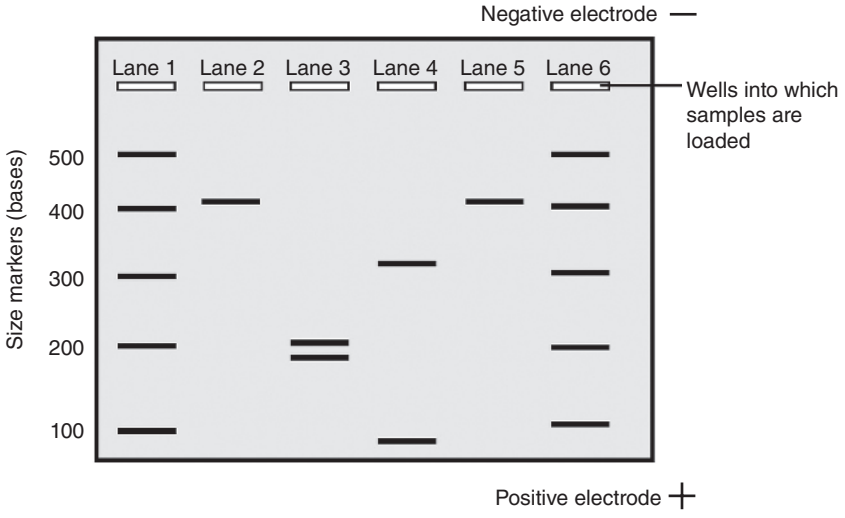


Figure 1.16 A representation of an agarose gel through which DNA fragments have been run. Lanes 1 and 6 are size markers – note that the smaller fragments migrate through the gel more rapidly than the larger fragments. The samples in lanes 2 and 5 each have a single band of just over 400 bases long. The sample in lane 3 has two bands that are both close to 200 bases long, and in lane 4 the two bands are around 100 and 300 bases long.

that used to size microsatellite alleles (Chapter 2). If the amplified products are of variable sizes then we may be able to assign individual genetic identities (see *Microsatellites*, Chapter 2). However, it is often the case that sequences with different compositions will be of the same length, in which case we must sequence the PCR products before we can identify different genotypes.

DNA Sequencing

The first widespread method of DNA sequencing, known as dideoxy sequencing or Sanger sequencing, was invented by Frederick Sanger in the mid-1970s (this work helped him to win a shared Nobel Prize in 1980). His protocol was designed to synthesize a strand of DNA using a DNA polymerase plus single nucleotides in a manner analogous to PCR, but with two significant differences. First, only a single primer is used as the starting point for synthesis so that sequences are built along the template in only one direction. Second, some of the nucleotides contain the sugar **dideoxyribose** instead of deoxyribose, the sugar normally found in DNA. Dideoxyribose lacks the 3′-hydroxyl group found in deoxyribose, and without this the next nucleotide cannot be added to the growing DNA strand; therefore, whenever a nucleotide with dideoxyribose is incorporated into the reaction, synthesis will be terminated.

Dideoxy sequencing can be done in four separate reactions, each of which include all four nucleotides in their deoxyribose form (dNTPs), and a small amount of one of the nucleotides (G, A, T, or C) in its dideoxyribose form (ddNTP). Incorporation of the ddNTPs will eventually occur at every single site along the DNA sequence, resulting in fragment sizes that represent the full spectrum from 1 bp of the target sequence to its maximum length. Different fragment sizes will be generated by each of the four

reactions, and in manual sequencing the products of each reaction are run out in separate but adjacent lanes on a gel. Fragments can be visualized in a number of ways including silver staining or the use of radioactive labels (isotopes of sulfur or phosphorus) that can be developed on x-ray films following a process known as autoradiography. All of the fragment sizes in a given lane indicate positions at which the dideoxyribose bases for that particular reaction were incorporated. For example, if the reaction containing the dideoxy form of dATP contains fragments that are 1 bp, 5 bp, and 10 bp long, then the first, fifth, and tenth bases in the sequence must be adenine (A). The fragments from each of the four reactions can be pieced together to recreate the entire sequence.

Although manual dideoxy sequencing was the norm for a number of years, it has now been largely replaced by automated sequencing. Many brands and models of automated sequencers are currently available, but the principle remains the same in all. The different fragments that make up a sequence are generated in the same way as described above, but the nucleotides that contain dideoxyribose are labeled with different colored fluorescent dyes. This modification means that reactions do not need to be kept separate in the same way as they do with manual sequencing, because different colors represent the size fragments that were terminated by each type of ddNTP. When these reactions are run out on automated sequencers, lasers activate the color of the fluorescent label of each band (typically black for G, green for A, red for T, and blue for C). Each color is then read by a photocell and stored on a computer file that records the fragments as a series of different colored peaks. By substituting the appropriate base for each colored peak, the entire sequence can be read from a single image (Figure 1.17).

High Throughput Sequencing

As mentioned earlier, a major reason for many of the most exciting recent developments in molecular ecology is the growing accessibility of **high throughput sequencing** (**HTS**), also known as **next-generation sequencing** or **massively parallel sequencing**. This technology first emerged around 2005, but until recently was largely restricted to a few specialist labs with large budgets; however, rapid developments in technology and affordability over the past few years mean that HTS is now accessible to a large community of users. In a nutshell, HTS simultaneously sequences millions of DNA molecules and therefore generates far more sequence data than was previously possible or perhaps even imaginable. HTS is actually a collective term that refers to a number of technologies, but they all share two main steps: library fragmentation/amplicon library preparation, followed by the detection of the incorporated nucleotides (Glenn 2011). NGS technologies can be classified into two main categories: (i) PCR-based technologies (sometimes referred to as second-generation technologies), and (ii) "single-molecule" sequencing (SMS) technologies, which do not include a PCR amplification step prior to sequencing (sometimes referred to as third-generation technologies). Detailed explanation of these technologies is beyond the scope of this textbook, but can be found in a number of review articles (e.g. Bleidorn 2016; Glenn 2011; Heather and Chain 2016; Pareek et al. 2011; Reuter et al. 2015).

There are trade-offs among the technologies with respect to bias (e.g. bias generated during fragment amplification that is part of the second-generation technologies), cost, and coverage (e.g. lengths of generated sequences, known as reads). Generally speaking, second-generation sequencers generate massive amounts of short reads, whereas third-generation sequencing technology is faster, and generates longer reads that are thus easier

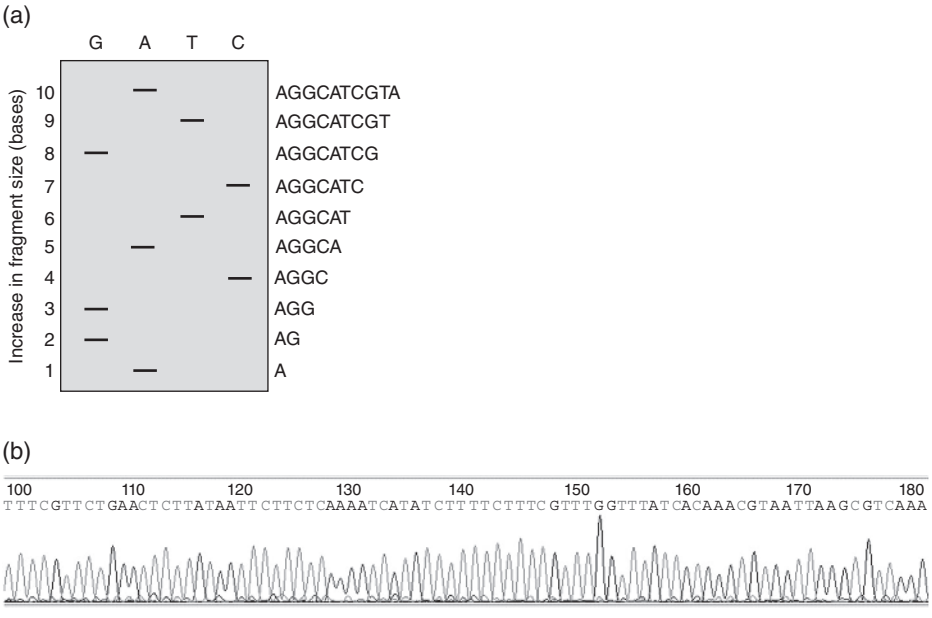


Figure 1.17 (a) A representation of a sequencing gel that followed the original method of dideoxy sequencing. Reactions were loaded into the lanes labeled G, A, T, and C, depending on which nucleotide was present in the dideoxyribose form. Because the smallest fragments migrate most rapidly, we can work from the bottom to the top of the gene to generate the cumulative sequences that are shown on the right-hand side. (b) Example of a sequence electropherogram, which is the end product of automated sequencing. Normally, the peaks that represent "T" are in red, "C" peaks are in blue, "G" peaks are in black, and "A" peaks are in green. This example is a fragment of the rps16-trnQ intergenic spacer in the cpDNA genome of water soldier (*Stratiotes aloides*). The sequence is read in the order that it is generated, so in this case it would be written as TTTCGTTCTGAA and so on (the row of text at the top of the figure corresponds to the colored peaks below).

to assemble (Bleidorn 2016; Miyamoto et al. 2014). However, the error rates are higher and the output rates are lower (in terms of numbers of reads) in third-generation compared with second-generation sequencing, and thus the second-generation technologies, most notably those platforms manufactured by Illumina, will likely be widely used for a number of years to come (Bleidorn 2016). Nevertheless, it is worth noting that technologies that were at the vanguard of second-generation sequencing are already being phased out, for example the widely used technology of 454 pyrosequencing is no longer supported by Roche, its distributing company. Perhaps the most common application of HTS to ecological studies follows methods that generate large amounts of sequence data from a subset of the genome. These methods are commonly referred to as genotype-by-sequencing (GBS) or RAD sequencing, and will be discussed in Chapter 2. Additional examples of research studies based on HTS will be provided throughout the text.

Regardless of the technology, HTS can also be used to characterize gene expression, for example when comparing which genes are being expressed under different environmental conditions. As discussed in the earlier section on qPCR, cDNA that is made from RNA template can be used to quantify gene expression. The mRNA molecules present in an individual reflect the collective products of gene expression, known as the transcriptome, and provide important insight into functional genes. HTS of the

transcriptome (**transcriptomics** is the study of transcriptomes and their functions) allows researchers to survey many or all expressed genes, and allows us to determine what genes – and what variants of each gene – are expressed in particular tissues, life stages, environments, and so on. The transcriptome is of course much smaller than the total genome as it reflects only those genes that are being expressed, and it should therefore be much easier to use HTS to characterize the transcriptome compared with reconstructing the entire genome. Prior to HTS transcriptomics, researchers were faced with a "needle in a haystack" approach to identifying potentially relevant genes, for example when the goal was to determine which genotype was responsible for a particular phenotype, but it is now possible to compare a large number of expressed genes between individuals, or between the same individual at different life stages. Jones et al. (2015) used HTS to compare the transcriptomes of the invasive moth pest the cotton bollworm (*Helicoverpa armigera*) under different levels of flight activity. They identified a suite of expressed genes that appeared to be linked to the physiological adaptations required for long-distance flight, including some genes important to the mobilization of lipids as flight fuel, some genes that contribute to the development of flight muscle structure, and other genes that help to regulate hormones as part of migratory physiology.

The previous example (cotton bollworm) was based on an experimental approach to varying flight durations. Another general approach is to obtain transcriptome data from organisms sampled under natural conditions, which, if combined with ecological and meteorological data, can provide an "ecological transcriptome" approach (Richards et al. 2009). Kobayashi et al. (2013) used this method when investigating the phenomenon of "general flowering," which occurs when numerous – sometimes hundreds – of plant species in tropical South-East Asia flower synchronously at irregular intervals that range from less than 1 year to several years (Kobayashi et al. 2013; Sakai et al. 2006). Kobayashi et al. (2013) used HTS of transcriptomes to test the hypothesis that drought is the trigger for this spectacular event. They collected bud samples from a single tree of *Shorea beccariana* (Dipterocarpaceae) over a time series that spanned pre-flowering and flowering, and compared the transcriptome over this time. They also used long-term meteorological data to identify the level of drought that in the past has induced general flowering. Their data showed pronounced transcriptional changes prior to flowering in a floral pathway integrator gene, *SbFT*, and in a floral repressor gene, *SbSVP*. In addition, drought-responsive and sucrose-induced genes showed changes in transcription across their time series. These data supported the hypothesis that drought is a trigger for general flowering, and additionally provided some mechanistic explanations for how flowering can be rapidly induced.

Overview

In this chapter we summarized why the application of molecular data to ecological studies has been so important. We have also been introduced to different genomes, and have considered why DNA is so variable, both within and among species. Now that we know how techniques such as PCR, qPCR, and sequencing (including HTS of both genomes and transcriptomes) allow us to tap into some of the information that is stored within genomes, we will build on this information in the next chapter by taking a more detailed look at the properties of the different genomes and genetic markers that are used in molecular ecology.

Chapter Summary

- Before the emergence of molecular ecology it was very difficult to obtain genetic data from wild populations, and biologists often had to rely on visible polymorphisms. Phenotypic data are useful for many things, although phenotypic plasticity often obscures the relationships between phenotypes and genotypes.
- The first studies to link molecular genetics and ecology were based on allozyme data. Proteins are encoded by DNA, and therefore reflect some of the variation in DNA sequences. Although undoubtedly a major breakthrough, relevance to only a small portion of the genome, combined with logistical challenges, mean that allozymes have now been almost entirely superseded by DNA markers.
- Many different types of gene regions exist within a genome, both coding (repetitive and single copy) and non-coding (including introns and pseudogenes) regions.
- Genetic diversity is continually generated through recombination and mutations, which include slipped-strand mispairings, nucleotide insertions/deletions, and nucleotide substitutions.
- These days, laboratories routinely use PCR to selectively amplify specific regions of DNA, a technique that allows researchers to genetically characterize individuals by generating enough copies of a particular segment of DNA to allow subsequent manipulation and characterization.
- Because PCR requires very little starting material, DNA extracted from small fragments of organisms and their remains allows non-invasive and non-lethal sampling. DNA for PCR amplification has often been extracted from feces, feathers, hair, leaves, fish scales, mucus, museum samples, and so on; this permits the humane sampling of wild organisms.
- Unlike "regular" PCR, qPCR allows researchers to quantify the number of DNA copies that were present in each sample.
- Gel electrophoresis allows us to separate and identify the fragments that are amplified by PCR on the basis of their sizes.
- DNA sequencing reactions generate the precise sequences of amplified DNA products. New technologies such as next-generation sequencing are rapidly changing the amount of sequence data that researchers have access to and will have significant implications for all aspects of genetics; for example, it is no longer uncommon for studies in molecular ecology to be based on thousands of loci or even entire genomes or transcriptomes.

References

Allendorf, F.W. (2017). Genetics and the conservation of natural populations: allozymes to genomes. *Mol. Ecol.* 26: 420–430. https://doi.org/10.1111/mec.13948.

Anderson, S.J., Conrad, K.F., Gillman, M.P. et al. (2008). Phenotypic changes and reduced genetic diversity have accompanied the rapid decline of the garden tiger moth (*Arctia caja*) in the U.K. *Ecol. Entomol.* 33: 638–645. https://doi.org/10.1111/j.1365-2311.2008.01013.x.

Anmarkrud, J.A. and Lifjeld, J.T. (2017). Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Mol. Ecol. Resour.* 17: 334–341. https://doi.org/10.1111/1755-0998.12600.

Badisco, L., Ott, S.R., Rogers, S.M. et al. (2011). Microarray-based transcriptomic analysis of differences between long-term gregarious and solitarious desert locusts. *PLoS One* 6: e28110. https://doi.org/10.1371/journal.pone.0028110.

Beja-Pereira, A., Oliveira, R., Alves, P.C. et al. (2009). Advancing ecological understandings through technological transformations in noninvasive genetics. *Mol. Ecol. Resour.* 9: 1279–1301. https://doi.org/10.1111/j.1755-0998.2009.02699.x.

Bi, K., Linderoth, T., Vanderpool, D. et al. (2013). Unlocking the vault: next-generation museum population genomics. *Mol. Ecol.* 22: 6018–6032. https://doi.org/10.1111/mec.12516.

Blake, M., McKeown, N.J., Bushell, M.L.T., and Shaw, P.W. (2016). DNA extraction from spider webs. *Conserv. Genet. Resour.* 8: 219–221. https://doi.org/10.1007/s12686-016-0537-8.

Bleidorn, C. (2016). Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Syst. Biodivers.* 14: 1–8. https://doi.org/10.1080/14772000.2015.1099575.

Boulesteix, M., Weiss, M., and Biémont, C. (2005). Differences in genome size between closely related species: the *Drosophila melanogaster* species subgroup. *Mol. Biol. Evol.* 23: 162–167. https://doi.org/10.1093/molbev/msj012.

Brandl, S., Schumer, G., Schreier, B.M. et al. (2015). Ten real-time PCR assays for detection of fish predation at the community level in the San Francisco Estuary-Delta. *Mol. Ecol. Resour.* 15: 278–284. https://doi.org/10.1111/1755-0998.12305.

Burrell, A.S., Disotell, T.R., and Bergey, C.M. (2015). The use of museum specimens with high-throughput DNA sequencers. *J. Hum. Evol.* 79: 35–44. https://doi.org/10.1016/j.jhevol.2014.10.015.

Cho, A., Morris, T., Wilson, C., and Freeland, J. (2016). Development of species-specific primers with potential for amplifying eDNA from imperilled freshwater unionid mussels. *Genome* 59: 1141–1149.

Creer, S., Deiner, K., Frey, S. et al. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods Ecol. Evol.* 7: 1008–1018. https://doi.org/10.1111/2041-210X.12574.

Freeland, J.R., Anderson, S., Allen, D., and Looney, D. (2007). Museum samples provide novel insights into the taxonomy and genetic diversity of Irish red grouse. *Conserv. Genet.* 8: 695–703. https://doi.org/10.1007/s10592-006-9217-7.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* 11: 759–769. https://doi.org/10.1111/j.1755-0998.2011.03024.x.

Goffeau, A., Barrell, B.G., Bussey, H. et al. (1996). Life with 6000 genes. *Science* 274: 546–567. https://doi.org/10.1126/science.274.5287.546.

Gregory, T.R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol. Rev. Camb. Philos. Soc.* 76: 65–101. https://doi.org/10.1111/j.1469-185X.2000.tb00059.x.

Hartley, C.J., Newcomb, R.D., Russell, R.J. et al. (2006). Amplification of DNA from preserved specimens shows blowflies were preadapted for the rapid evolution of insecticide resistance. *Proc. Natl. Acad. Sci. U. S. A.* 103: 8757–8762. https://doi.org/10.1073/pnas.0509590103.

Heather, J.M. and Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics* 107: 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003.

Hjelmen, C.E., Johnston, J.S., Bzeznik, B. et al. (2017). The mode and tempo of genome size evolution in the subgenus *Sophophora*. *PLoS One* 12: e0173505. https://doi.org/10.1371/journal.pone.0173505.

Huang, Y.-Y., Matzke, A.J.M., Matzke, M. et al. (2013). Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). *PLoS One* 8: e74736. https://doi.org/10.1371/journal.pone.0074736.

Jacobs, B.S. and Latimer, A.M. (2012). Analyzing reaction norm variation in the field vs. greenhouse: comparing studies of plasticity and its adaptive value in two species of *Erodium*. *Perspect. Plant Ecol. Evol. Syst.* 14: 325–334. https://doi.org/10.1016/j.ppees.2012.04.002.

Ji, Y., Ashton, L., Pedley, S.M. et al. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.* 16: 1245–1257. https://doi.org/10.1111/ele.12162.

Jiang, G.-F., Hinsinger, D.D., and Strijk, J.S. (2016). Comparison of intraspecific, interspecific and intergeneric chloroplast diversity in Cycads. *Sci. Rep.* 6: 31473. https://doi.org/10.1038/srep31473.

Jones, C.M., Papanicolaou, A., Mironidis, G.K. et al. (2015). Genomewide transcriptional signatures of migratory flight activity in a globally invasive insect pest. *Mol. Ecol.* 24: 4901–4911. https://doi.org/10.1111/mec.13362.

Kamenova, S., Mayer, R., Coissac, E. et al. (2018). Comparing three types of dietary samples for prey DNA decay in an insect generalist predator. *Mol. Ecol. Resour.* 18: 966–973. https://doi.org/10.1111/1755-0998.12775.

Kendall, K.C., Stetz, J.B., Boulanger, J. et al. (2009). Demography and genetic structure of a recovering brown bear population. *J. Wildl. Manag.* 73: 3–17.

Kitano, J., Bolnick, D.I., Beauchamp, D.A. et al. (2008). Reverse evolution of armor plates in the threespine stickleback. *Curr. Biol.* 18: 769–774. https://doi.org/10.1016/j.cub.2008.04.027.

Kleven, O., Aarvak, T., Jacobsen, K.-O. et al. (2016). Cross-species amplification of microsatellite loci for non-invasive genetic monitoring of the snowy owl (*Bubo scandiacus*). *Eur. J. Wildl. Res.* 62: 247–249. https://doi.org/10.1007/s10344-016-0986-0.

Knapp, S.M., Craig, B.A., and Waits, L.P. (2009). Incorporating genotyping error into non-invasive DNA-based mark–recapture population estimates. *J. Wildl. Manag.* 73: 598–604. https://doi.org/10.2193/2007-156.

Kobayashi, M.J., Takeuchi, Y., Kenta, T. et al. (2013). Mass flowering of the tropical tree *Shorea beccariana* was preceded by expression changes in flowering and drought-responsive genes. *Mol. Ecol.* 22: 4767–4782. https://doi.org/10.1111/mec.12344.

Lander, E.S., Linton, L.M., Birren, B. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860–921. https://doi.org/10.1038/35057062.

Le Vin, A.L., Adam, A., Tedder, A. et al. (2011). Validation of swabs as a non-destructive and relatively non-invasive DNA sampling method in fish. *Mol. Ecol. Resour.* 11: 107–109. https://doi.org/10.1111/j.1755-0998.2010.02909.x.

Lewontin, R.C. and Hubby, J.L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics* 54: 595–609.

Mesa-Cruz, J., Brown, J., Waits, L., and Kelly, M. (2016). Non-invasive genetic sampling reveals diet shifts, but little difference in endoparasite richness and faecal glucocorticoids, in Belizean felids inside and outside protected areas. *J. Trop. Ecol.* 32: 226–239. https://doi.org/10.1017/S0266467416000213.

Miyamoto, M., Motooka, D., Gotoh, K. et al. (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics* 15: 699. https://doi.org/10.1186/1471-2164-15-699.

Mower, J., Sloan, D., and Alverson, A. (2012). Plant mitochondrial genome diversity: the genomics revolution. *Plant Genome Diversity* 1: 123–144. https://doi.org/10.1007/978-3-7091-1130-7_9.

Mullis, K.B. and Faloona, F.A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalysed chain reaction. *Methods Enzymol.* 155: 335–350. https://doi.org/10.1016/0076-6879(87)55023-6.

Oi, C.A., López-Uribe, M.M., Cervini, M., and Del Lama, M.A. (2013). Non-lethal method of DNA sampling in euglossine bees supported by mark–recapture experiments and microsatellite genotyping. *J. Insect Conserv.* 17: 1071–1079. https://doi.org/10.1007/s10841-013-9582-8.

Orsini, M., Cusano, R., Costelli, C. et al. (2016). Complete genome sequence of chloroplast DNA (cpDNA) of *Chlorella sorokiniana*. *Mitochondrial DNA* 27: 838–839. https://doi.org/10.3109/19401736.2014.919466.

Ott, S.R. and Rogers, S.M. (2010). Gregarious desert locusts have substantially larger brains with altered proportions compared with the solitarious phase. *Proc. R. Soc. B Biol. Sci.* 277 (1697): 3087–3096. https://doi.org/10.1098/rspb.2010.0694.

Palazzo, A.F. and Gregory, T.R. (2014). The case for junk DNA. *PLoS Genet.* 10: e1004351. https://doi.org/10.1371/journal.pgen.1004351.

Pareek, C.S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J. Appl. Genet.* 52: 413–435. https://doi.org/10.1007/s13353-011-0057-x.

Pascoal, S., Carvalho, G., Creer, S. et al. (2012). Plastic and heritable components of phenotypic variation in *Nucella lapillus*: an assessment using reciprocal transplant and common garden experiments. *PLoS One* 7: e30289. https://doi.org/10.1371/journal.pone.0030289.

Pegoraro, L., Cafasso, D., Rinaldi, R. et al. (2016). Habitat preference and flowering-time variation contribute to reproductive isolation between diploid and autotetraploid *Anacamptis pyramidalis*. *J. Evol. Biol.* 29: 2070–2082. https://doi.org/10.1111/jeb.12930.

Pellicer, J., Fay, M.F., and Leitch, I.J. (2010). The largest eukaryotic genome of them all? *Bot. J. Linn. Soc.* 164: 10–15. https://doi.org/10.1111/j.1095-8339.2010.01072.x.

Poretsky, R., Rodriguez-R, L.M., Luo, C. et al. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* 9: e93827. https://doi.org/10.1371/journal.pone.0093827.

Reuter, J.A., Spacek, D.V., and Snyder, M.P. (2015). High-throughput sequencing technologies. *Mol. Cell* 58: 586–597. https://doi.org/10.1016/j.molcel.2015.05.004.

Richards, C., Hanzawa, Y., Katari, M.S. et al. (2009). Perspectives on ecological and evolutionary systems biology. In: *Annual Plant Reviews* (eds. G.M. Coruzzi and R.A. Gutíerrez), 331–351. Oxford: Wiley Blackwell.

Rosche, C., Hensen, I., Mráz, P. et al. (2017). Invasion success in polyploids: the role of inbreeding in the contrasting colonization abilities of diploid versus tetraploid

populations of *Centaurea stoebe* s.l. *J. Ecol.* 105: 425–435. https://doi.org/10.1111/1365-2745.12670.

Sakai, S., Harrison, R.D., Momose, K. et al. (2006). Irregular droughts trigger mass flowering in aseasonal tropical forests in Asia. *Am. J. Bot.* 93: 1134–1139. https://doi.org/10.3732/ajb.93.8.1134.

Schmidt, P.-A., Bálint, M., Greshake, B. et al. (2013). Illumina metabarcoding of a soil fungal community. *Soil Biol. Biochem.* 65: 128–132. https://doi.org/10.1016/j.soilbio.2013.05.014.

Shi, X., Tian, P., Lin, R. et al. (2016). Characterization of the complete mitochondrial genome sequence of the globose head whiptail *Cetonurus globiceps* (Gadiformes: Macrouridae) and its phylogenetic analysis. *PLoS One* 11: e0153666. https://doi.org/10.1371/journal.pone.0153666.

Shimpi, G.G., Vargas, S., and Wörheide, G. (2016). Evaluation and validation of reference genes for qPCR analysis to study climate change-induced stresses in *Sinularia* cf. *cruciata* (Octocorallia: Alcyonidae). *J. Exp. Mar. Bio. Ecol.* 483: 42–52. https://doi.org/10.1016/j.jembe.2016.06.001.

Sint, D., Thurner, I., Kaufmann, R., and Traugott, M. (2015). Sparing spiders: faeces as a non-invasive source of DNA. *Front. Zool.* 12: 3. https://doi.org/10.1186/s12983-015-0096-y.

Staats, M., Erkens, R.H.J., van de Vossenberg, B. et al. (2013). Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. *PLoS One* 8: e69189. https://doi.org/10.1371/journal.pone.0069189.

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* 18: 292–308. https://doi.org/10.1038/nrg.2017.7.

Svartman, M., Stone, G., and Stanyon, R. (2006). The ancestral eutherian karyotype is present in *Xenarthra*. *PLoS Genet.* 2: e109. https://doi.org/10.1371/journal.pgen.0020109.

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L.H. (2012). Environmental DNA. *Mol. Ecol.* 21: 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x.

Wheeler, Q.D., Knapp, S., Stevenson, D.W. et al. (2012). Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Syst. Biodivers.* 10: 1–20. https://doi.org/10.1080/14772000.2012.665095.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W. et al. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76: 273–297. https://doi.org/10.1007/s11103-011-9762-4.

Zhang, H., Hall, N., McElroy, J.S. et al. (2017). Complete plastid genome sequence of goosegrass (*Eleusine indica*) and comparison with other *Poaceae*. *Gene* 600: 36–43. https://doi.org/10.1016/j.gene.2016.11.038.