1

Genomes and Molecular Markers

1.1 Introduction

Molecular ecology is a field of study that integrates molecular biology, genetics, ecology, and evolutionary biology to help us understand some of the many ways in which organisms interact with each other and with their environment. Over the past few decades, new techniques in molecular biology have allowed us to generate genetic data from an incredibly wide range of taxa. These data have revolutionized our understanding of ecology and evolution through investigations that include quantifying genetic diversity, retracing the routes that individuals follow across short and long distances, identifying cryptic species, reconstructing family relationships, inferring demographic processes – such as population declines – and identifying genes that enable evolutionary adaptations. Many of these topics have important real-world applications; for example, in conservation genetics, we might ask which populations have levels of genetic diversity that are low enough to justify human intervention, and in studies of biological invasions, we might use genetic data to identify the origin – and potentially the introduction route – of an invader.

Today, researchers can obtain genetic data from virtually all taxonomic groups, and from many different types of samples, at a fraction of the cost – and in the fraction of the time – compared to a couple of decades ago. For example, in 2003, it took scientists approximately 13 years and $3 billion to produce the first whole human genome sequence. Today, the complete human genome can be sequenced for less than $1000, and in 2022, a research group at Stanford received a Guinness World Record for sequencing a full human genome in five hours and two minutes (Rahlves 2022)! Generating whole-genome sequences (WGS) of non-model organisms takes longer than five hours, but can realistically be achieved within a few weeks. Whole genomes have now been sequenced from thousands of species of animals, plants, fungi, protists, and bacteria, and partial genomes have been sequenced from many more taxa (e.g. see https://www.ncbi.nlm.nih.gov/). However, although whole genomes provide a tremendous amount of data, they are not a requirement for research in molecular ecology; in some cases, only a small amount of genetic information can give us useful insights into an ecological question. Throughout this book, we will explore methodologies and examples that describe the types of ecological and evolutionary information that can be obtained from WGS, small sections of the genome, and everything in between.

This textbook is divided into seven chapters. The first chapter reviews some of the important theoretical underpinnings of molecular ecology: molecular markers, genomes, and some widely used lab methods. Chapter 2 asks how we can use genetic data to identify species and hybrids from different types of samples including environmental DNA (eDNA). Chapter 3 investigates some of the phylogenetic and phylogeographic approaches that can help us to understand the evolutionary relationships among taxa and some of the ways in which historical events have influenced the current distribution of species, with a particular emphasis on biological invasions. Chapters 4 and 5 focus on what may be considered more traditional population genetics by first investigating how we can quantify and understand the genetic variation within populations, and then looking at ways in which we can use genetic comparisons among populations and individuals to investigate movements across a landscape. Chapter 6 revisits many of the ideas from earlier chapters in the context of conservation genetics, and Chapter 7 discusses some key aspects of behavioral ecology that have incorporated genetic data, including dispersal, foraging, and mating systems.

1.2 DNA, RNA, and Proteins

We begin with a brief review of the basics: deoxyribonucleic acid **(DNA)**, ribonucleic acid (**RNA**), and **proteins**. DNA is the molecule that carries the genetic instructions for life. It is found in the cells of all living organisms and is responsible for passing genetic information from one generation to the next. **Nucleotides** are the building blocks of DNA. Each nucleotide consists of a nitrogen-containing base, a phosphate group, and a sugar molecule (deoxyribose). The four bases are adenine (A), guanine (G), cytosine (C), and thymine (T). Sugar-phosphate backbones link bases together to form strands of DNA that are arranged as two complementary sequences in which A always pairs with T, and G always pairs with C. These are held together by hydrogen bonds in a double helix formation (Figure 1.1). Each pair of complementary nucleotides is referred to as a **base pair** when describing sequences although only one strand of the sequence is written. For example, a sequence of 20 bp from a particular gene region is written as GATCATACACACTTCGGATT. **Prokaryotic** cells lack membrane-bound organelles including nuclei, which means that their DNA is arranged in a closed, double-stranded loop that lies free within the cell cytoplasm. In contrast, **eukaryotic** cells contain nuclei and other organelles that house DNA. Eukaryotic **nuclear DNA** (also referred to as **nrDNA**) is organized into **chromosomes** that are inside the nucleus of each cell. Each chromosome comprises a DNA molecule that is divided into functional units called genes. Genes are located at specific sites along a chromosome, and each of these sites is referred to as a **locus** (plural loci). The different forms of the same gene that occur at a particular locus are known as **alleles**. The additional eukaryotic organellar genomes are **mitochondrial DNA** (**mtDNA**), which is found in the **mitochondria** of plants and animals, and **chloroplast DNA** (**cpDNA**), which is found in chloroplasts – a type of plastid – in plants and algae.

DNA contains the instructions needed for an organism to develop, survive, and reproduce, and genomes are therefore very complex. All three types of genomes can be divided into coding and noncoding regions. Coding regions lead to **gene expression**, which occurs when genetic information is transferred from DNA to RNA to protein. As with DNA, RNA is made up of nucleotides but there are some differences: 1) the sugar molecule in RNA is ribose (deoxyribose in DNA); 2) the bases that are found in RNA nucleotides are A, G, C, and uracil (U, instead of T in DNA); and 3) RNA is mostly single stranded (whereas DNA is double stranded). **Transcription** occurs when the coding

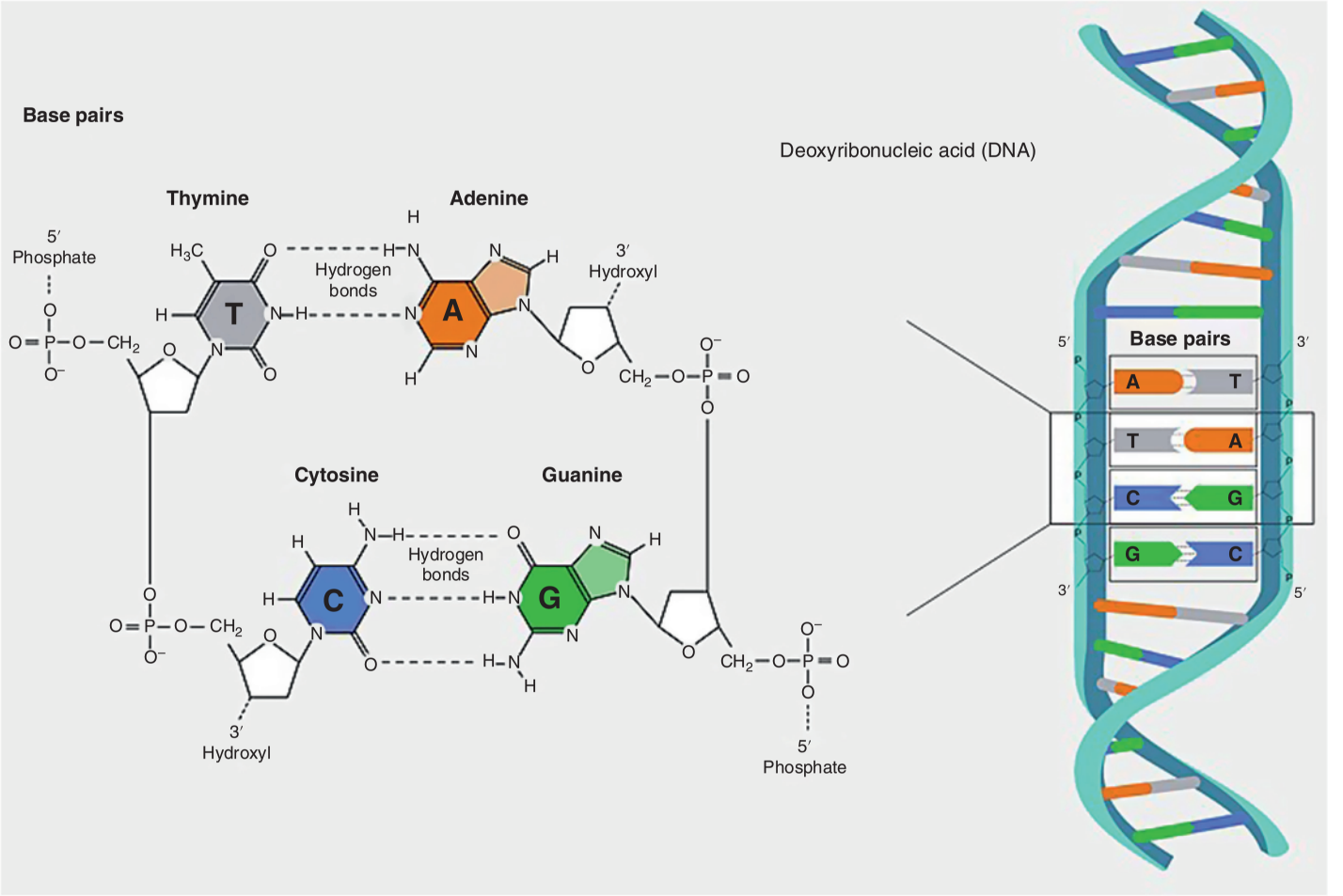

Figure 1.1 The structure of DNA. Each nucleotide consists of a nitrogen-containing base (G, A, T, or C), a phosphate group, and a sugar molecule (deoxyribose). Sugar-phosphate backbones link bases together to form strands of DNA that are arranged as two complementary sequences held together by hydrogen bonds in a double helix formation. *Source:* National Human Genome Research Institute/Wikimedia commons/Public Domain.

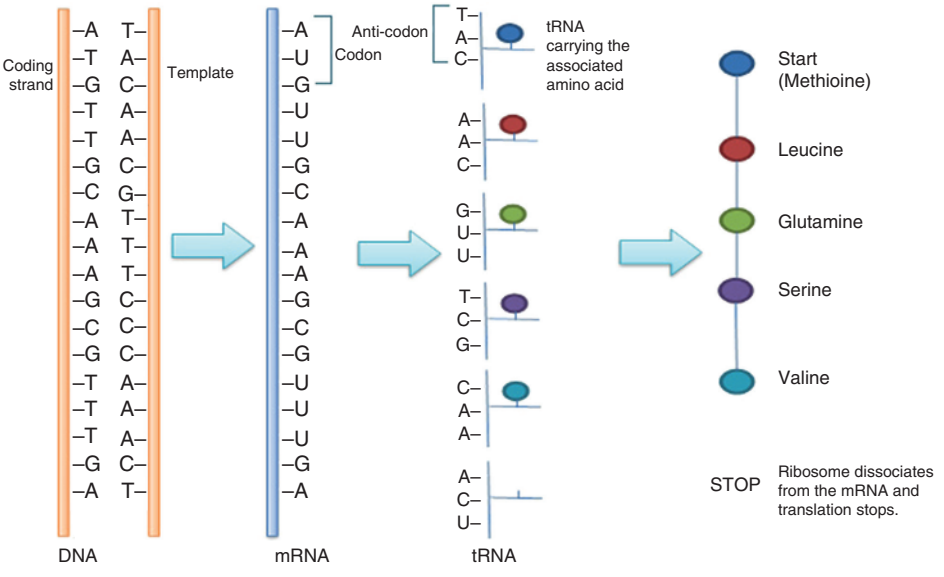

Figure 1.2 During gene expression, DNA is transcribed to mRNA (transcription), which is then translated into amino acids (translation). *Source:* Becky Boone/Wikimedia Commons/CC BY SA. 2.0.

region of DNA is transcribed into RNA, more specifically **messenger RNA** (**mRNA**), resulting in a single strand of RNA that is complementary to its corresponding DNA sequence; this is known as the primary transcript. In prokaryotes, mRNA does not undergo posttranscriptional modifications, whereas in eukaryotes, **RNA splicing** cuts out the noncoding **introns** and produces a mature mRNA that is complementary to the protein-coding **exons**. mRNA sequences are then translated into protein sequences following a process known as **translation** (Figure 1.2). Translation involves **transfer RNA** (**tRNA**), which carries the amino acids to the ribosomes, where protein synthesis is carried out by **ribosomal RNA** (**rRNA**). During translation, triplets of RNA bases (**codons**) encode amino acids. These amino acids make up protein chains known as **polypeptides**. Most – but not all – codons encode amino acids; the exceptions are "start" and "stop" codons, which act as signals to begin or end translation, respectively, and these provide important mechanisms for controlling gene expression. In some cases, different DNA sequences encode the same amino acid; for example, arginine (Arg) can be encoded by six different codons, and four different codons encode threonine (Thr) (Table 1.1).

1.3 Genomes

The previous section briefly introduced genomes, but these need more discussion because genome differences are very important in molecular ecology. The term genome can be used to describe the complete set of genetic material present in a cell, although in eukaryotes the term is more commonly used for either the nuclear genome, the mitochondrial genome, or the chloroplast genome. The largest portion of the genome is packaged as chromosomal DNA. In prokaryotes, chromosomal DNA is contained in a central area of the cell called the nucleoid, which is not surrounded by a nuclear

Table 1.1 Codons are three-letter chains of nucleotides that each encode one specific amino acid during translation. Most amino acids are encoded by multiple codons. There are also codons that initiate ("START") and terminate ("STOP") protein synthesis.

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

membrane. Bacteria and some other prokaryotes also have small, circular DNA molecules called plasmids, which are distinct from chromosomal DNA and replicate independently. Plasmids typically include a small number of genes that are involved in several important functions, such as genes which confer antibiotic resistance in bacteria. Eukaryotes collectively have three different types of genomes. The largest genome is nrDNA, which is found in the nucleus. In sexually reproducing organisms, this is inherited from both parents: a **diploid** organism (one with two sets of chromosomes) that resulted from sexual reproduction inherited one set of chromosomes from the maternal parent, and the other set from the paternal parent. This is known as **biparental inheritance** (inherited from both parents; but see also Box 1.1). Most eukaryotes also

Box 1.1 Haploid Sex Chromosomes

While the nuclear genome is typically described as having biparental inheritance, this is not entirely true: **sex chromosomes**, which have a role in the determination of sex, are normally uniparentally inherited. In some species, sex is determined by the environment (**environmental sex determination**); for example, the egg incubation temperature during early development determines the sex of crocodiles and many turtles and lizards. However, in many other species it is sex chromosomes that determine the sex of an individual. In most mammals the sex chromosomes are designated X and Y. Individuals with two copies of the X chromosome (**homogametic**, meaning two of the same type of sex chromosomes) develop into females, whereas individuals with one X and one Y chromosome (**heterogametic**, meaning two different types of sex chromosomes) develop into males. The sex chromosomes in birds and lepidopterans are designated Z and W, and for these taxa the females are heterogametic (ZW) and the males are homogametic (ZZ). In some other species including the nematode *Caenorhabditis elegans*, there is only one type of sex chromosome (X), and individuals with two copies of that chromosome (XX) develop into females, whereas individuals with only one copy (XO) develop into males. **Monoecious** plants – those with male and female reproductive organs on the same individuals – typically lack discrete sex chromosomes. **Dioecious** plants – those with male and female reproductive organs on different individuals – collectively have three systems that describe their sex chromosomes: XX/XY, XX/X0, and WZ/ZZ, in which XX and WZ are females and XY, X0, and ZZ are males.

In diploid taxa, individuals normally get one sex chromosome from their maternal parent and the other from their paternal parent. In mammals, females have two copies of the X chromosome, which means they pass on one copy of the X chromosome to each of their offspring, regardless of whether the offspring are male or female. Males have one X and one Y chromosome, which means that they pass on an X chromosome to half of their offspring and a Y chromosome to the other half. This determines the sex of the offspring: those that inherit an X chromosome from their male parent will be females (XX), and those that inherit a Y chromosome from their male parent will be males (XY). The Y chromosome is therefore uniparentally inherited in mammals – and more specifically, it is paternally inherited, which means it is inherited from the male parent. Furthermore, because males have only one Y chromosome, these chromosomes for the most part do not undergo recombination: there are two small regions at the tips of the Y chromosome that recombine with the X chromosome, but these two regions are separated by approximately 60 megabase pairs (Mbp = 1,000,000 base pairs) of non-recombining sequence.

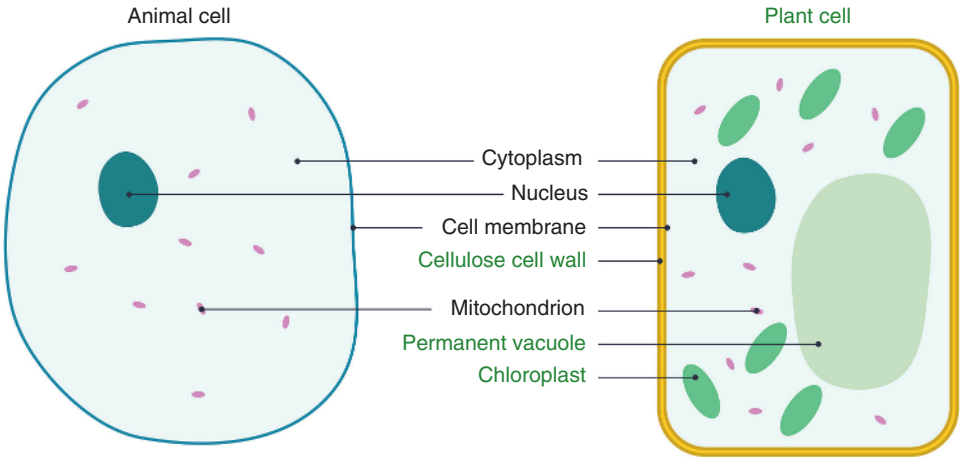

Figure 1.3 A simplified comparison of plant and animal cells, showing the nucleus, mitochondria, and – for plants – chloroplasts. Note that in each cell there is one nucleus (and hence one copy of the nuclear genome) but multiple mitochondria/chloroplasts (and hence multiple copies of the mitochondrial and chloroplast genomes). An additional important difference is the cell wall that is present in plants but not in animals. *Source:* domdomegg /Wikimedia commons/CC BY 4.0.

have mitochondria, the so-called powerhouses of the cell, in which the mitochondrial genome (mtDNA) is located. Plants and algae also have chloroplast genomes (cpDNA), which are found in chloroplasts, a type of membrane-bound plastid that produces energy following photosynthesis. Mitochondrial and chloroplast genomes are sometimes referred to as **organellar genomes.**

Mitochondria and chloroplasts are located outside the cell nucleus (Figure 1.3), and because each cell has multiple mitochondria and chloroplasts, it also has multiple, largely identical copies of the mitochondrial and chloroplast genomes. Mitochondrial and chloroplast genomes typically follow a pattern of **uniparental inheritance**, which means that they are inherited from only one of the parents. Although mitochondrial and chloroplast genomes are inherited independently of nuclear genomes, they are strongly interdependent because many of the proteins required for chloroplast and mitochondrial functions are encoded by nrDNA. The next three sections further discuss some of the important attributes of organelle and nuclear genomes. We will apply a somewhat historical lens to this section by first discussing organelle genomes (mtDNA and cpDNA), which were the most common sources of genetic markers in the early days of molecular ecology, and which continue to be widely used. We will begin with a general overview of organelle genomes and then discuss in more detail animal mtDNA, plant mtDNA, and plant and algal cpDNA.

1.4 Organellar Genomes

Organellar genomes (mitochondrial and chloroplast) have several traits in common that are relevant to molecular ecology. First, as noted above, mitochondrial and chloroplast genomes are uniparentally inherited in most taxa. Because the organelle genome is most commonly in the form of a single, circular molecule of DNA that is passed down only from one parent to the offspring, it undergoes little to no recombination (although see

exceptions discussed later). Therefore, barring mutation, an individual will have the same mitochondrial or chloroplast genome as their mother or father. In contrast, their nuclear genome will be a combination of alleles from both parents and will have undergone extensive recombination (assuming sexual reproduction). A lack of recombination in organellar genomes can make it easier for us to retrace the inheritance of organellar versus nuclear lineages through time. However, because organelle genomes are single molecules, they are effectively a single locus that includes many different genes that are inherited as a "package deal," and as we will see throughout this book, there are often advantages to obtaining data from multiple, independently inherited loci.

Uniparental inheritance also has important implications for the population size of organellar genomes. When we consider population sizes, we are often talking about the numbers of individuals from a particular species that exist within a given geographical region. However, because eukaryotic organisms each have two or three genomes, we can also think about the population sizes of these genomes. The following explanation refers to mitochondrial genomes, but the principle also applies to chloroplast genomes:

- In a diploid, sexually reproducing species, each individual has a nuclear genome that comprises two sets of chromosomes inherited from two parents.
- In contrast, each individual has one mitochondrial genome, which they normally inherit from only one parent (usually the mother).
- Imagine a population of 10 diploid individuals, comprising 5 females and 5 males: collectively they have 20 nuclear genomes (because diploid individuals each have two copies of the nuclear genome).
- In most cases, only the females pass down their mitochondrial genomes to the next generation, which means that the population has effectively five mitochondrial genomes (one for each female; we aren't counting the male mitochondrial genomes because they will not be passed along to the next generation).
- This means that the "population" size of mitochondrial genomes is ¼ that of the "population" size of nuclear genomes (20 nuclear genomes compared to 5 mitochondrial genomes).

The smaller population size of mitochondrial genomes compared to nuclear genomes makes it relatively sensitive to demographic events such as bottlenecks, which occur when the size of a population is temporarily reduced, for example, following a disease outbreak or an extreme weather event. Even if the population recovers quickly, it lost multiple genomes during the bottleneck. Bottleneck effects are typically more pronounced for mitochondrial genomes compared to nuclear genomes simply because there were far fewer mitochondrial genomes to begin with. A post-bottleneck population may have very low mtDNA variability, and as we will see in Chapter 3, this can help us reconstruct historical population bottlenecks.

One limitation to uniparentally inherited markers is that they reflect the evolutionary history of only one sex. For example, the chloroplast genome is maternally inherited (passed down from mothers to offspring) in most **angiosperms** (flowering plants) and therefore is not dispersed by pollen, which contains the male gametes. A comparison of the genetic similarity among populations of a flowering plant species based solely on cpDNA markers would therefore reflect the results of seed movement but not pollen movement. Similarly, the mitochondrial genome (also referred to as the **mitogenome**) is maternally inherited in mammals, and a comparison of the genetic similarity among populations of a mammalian species based solely on mtDNA markers would reflect female movement but not male movement. We will discuss this in more detail in Chapter 7.

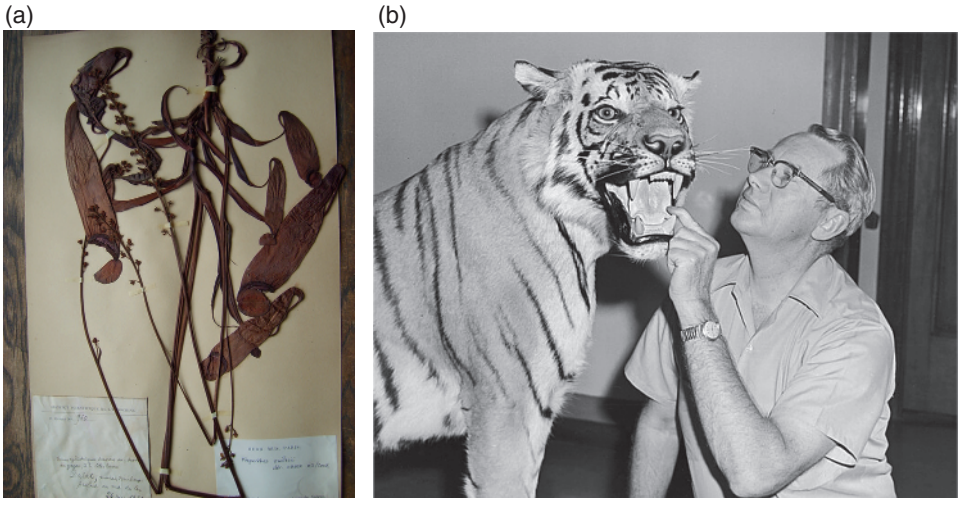

Figure 1.4 Herbarium specimen of *Nepenthes smilesii* deposited at the Museum National d'Histoire Naturelle in Paris, France. (a) Photographer: François MEY, Herbarium: Museum National d'Histoire Naturelle; (b) Dr. Wilmer W. Tanner, zoology professor and curator of the BYU Life Sciences Museum, examines a tiger trophy received by the museum in 1973. University Archives, L. Tom Perry Special Collections, Harold B. Lee Library, Brigham Young University. Herbaria and museums are treasure troves of DNA that can be extracted and used to study all sorts of questions that can be addressed using historical genetic data. *Source:* (a) François MEY/Wikimedia Commons/CC BY 3.0/https://commons.wikimedia.org/wiki/File:Nepenthes_smilesii_herbarium_specimen.jpg/ last accessed on October 15, 2025; (b) Brigham Young University/ Wikimedia Commons/CC BY-SA 3.0/ https://commons.wikimedia.org/wiki/File:Wilmer_W._Tanner_with_museum.jpg/ last accessed on Oct 28, 2025.

As noted above, there are many copies of organellar genomes in each cell, and their relatively small size and abundance greatly increase the likelihood of their recovery from degraded DNA. Herbaria, museums, sediment cores, ice cores, and archaeological digs can all be valuable sources of DNA for rare or hard-to-find taxa, or for addressing questions that benefit from historical genetic data (Figure 1.4); however, DNA from these sources is often highly degraded, and therefore data are more likely to be retrieved from organellar as opposed to nuclear genomes . For example, Ciucani et al. (2019) sequenced a fragment of mtDNA from 19 Late Pleistocene–Holocene wolf samples (bones and teeth) collected from archaeological sites in northern Italy. Their results provided insights into the genetic variability and the relationships among past and modern Eurasian and Italian wolf populations. In another study, Kurose et al. (2020) wished to identify the source populations of Himalayan balsam (*Impatiens glandulifera*), a weed that has invaded large areas of the United Kingdom. Biological control of this species uses a rust fungus, but not all lineages of this invasive plant are susceptible to the two strains of rust fungus that are currently used. Sequences of six chloroplast regions that were obtained from modern samples in the native and introduced ranges of Himalayan balsam, plus herbarium samples collected in the Himalayas between 1881 and 1956, allowed the authors of this study to identify the most likely geographic areas from which the invasive lineages originated and thus identify the most promising regions to look for new strains of rust fungus for future biocontrol in the United Kingdom. It is also possible to sequence entire organelle genomes from ancient material; for example, Scarsbrook et al. (2023) achieved this on samples from 19 modern, 6 historical/archival (1898–2011), and 16 Holocene subfossil New Zealand diplodactylid

geckos, which shed light on historical patterns of gecko range expansions and contractions. Additionally, next-generation sequencing (NGS) methods (Section 1.11.5) have facilitated the generation of nrDNA sequences from historical specimens.

To sum up, uniparental inheritance, low recombination, a small genome "population" size, and comparative ease of isolation from degraded samples are some of the features common to the organellar genomes of most species. Below, we will build on these ideas by considering some specific attributes of animal mtDNA, plant mtDNA, and plant cpDNA.

1.4.1 Animal Mitochondrial DNA

In most animals, mtDNA is a single, circular molecule of DNA that contains 13 protein-coding genes, 22 tRNA genes, and 2 rRNA genes. There is also a noncoding control region that plays an important functional role by regulating mitochondrial replication and transcription. The 13 protein-coding genes produce proteins necessary for oxidative phosphorylation, the process that harnesses energy in the form of ATP, and this makes mitochondria the powerhouses of cells. Animal mtDNA genomes typically range from 10 to 20 kilobase pairs (kb = 1000 bp) in size, although there are exceptions to this, with the tube anemone (*Isarachnanthus nocturnus*) holding the record for the largest animal mtDNA genome at an impressive 80,923 bp (Stampar et al. 2019). 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 1.5a).

mtDNA sequences have been widely used in molecular ecology partly because the conserved arrangement of genes allowed researchers to design **universal primers**. Primers determine which region(s) of DNA will be amplified in polymerase chain reactions (PCR), the process of replicating many thousands of copies of the target DNA sequence(s) (Section 1.10.2). Universal primers are those that anneal to short sequences flanking the target region in many different species and thus can be used to characterize that gene region from multiple taxa without prior knowledge of individual genome sequences. For example, Che et al. (2012) identified regions in the mitochondrial cytochrome c oxidase subunit 1 gene that are conserved across different amphibian taxa and designed primers that could be used to amplify this region in numerous species of caecilians, salamanders, and frogs. While there is no such thing as a truly universal primer (i.e. one that can be used for literally all taxa), there are numerous examples of universal primers that can be used to amplify target DNA from many different taxonomic groups (Figure 1.6).

In recent years, it has become easier to generate the sequence of the entire mitogenome, and this has shown us that recombination and gene rearrangements are more common in mtDNA than was previously thought. For example, a study that investigated mitochondrial gene rearrangements in 464 species of arthropods from 3 subphyla (Chelicerata, Myriapoda, and Crustacea) found novel mitochondrial gene arrangements in each major lineage, and the data suggested that 84% (390) of species – representing 43 out of the 47 orders that they analyzed – have at least one gene rearrangement (Sterling-Montealegre and Prada 2024). Mitogenome sequencing has also identified rearrangements in Antarctic notothenioids (a suborder of fish found mainly in Antarctica). Whole mitochondrial sequences from five species within this group (four ice fish species plus the cold-specialized *Trematomus borchgrevinki*) revealed duplications of the protein-coding gene *ND6*, two tRNAs, and the control region (Minhas et al. 2023). In another fish example, the mitogenome sequence of the daggertooth pike conger *Muraenesox cinereus* (a marine eel) showed the rearrangement/duplication of three different genes

compared with the mitochondrial gene arrangements of most vertebrates (Figure 1.7) (Zhang et al. 2021). It seems likely that as more whole mitogenome sequences are generated, we will continue to find examples of recombination and gene rearrangement in animal mitochondrial genomes. Nevertheless, recombination rates in mitochondrial genomes are much lower than those in nuclear genomes, and as we will discuss in more detail in Chapter 2, universal primers that amplify target regions of mtDNA continue to play a very important role in identifying species and communities.

Another reason for the popularity of animal mtDNA genetic markers is that many regions within this genome have higher mutation rates compared to many nuclear genes,

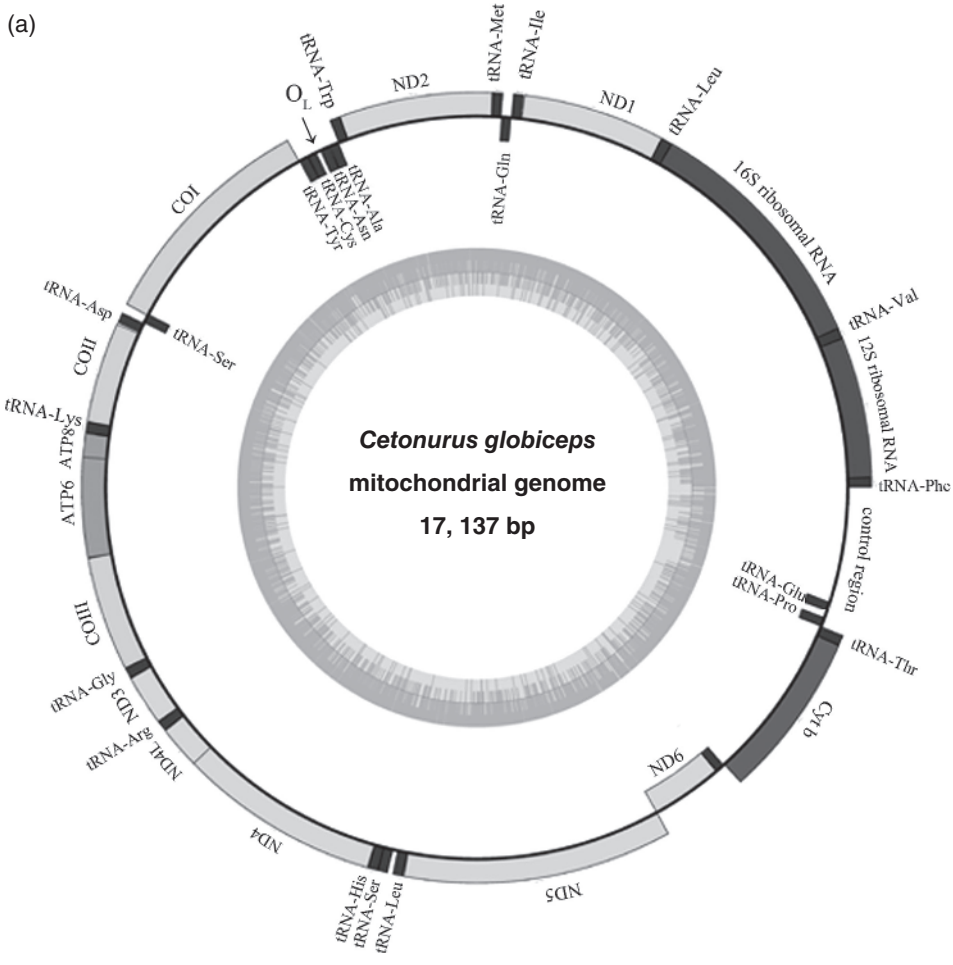

Figure 1.5 (a) Gene map and organization of the complete mitochondrial genome of the globose head whiptail *Cetonurus globiceps* (an Australian fish). The relatively large number of tRNA genes is typical of vertebrate mtDNA. Other common genes include subunits of NADH dehydrogenase (ND) and cytochrome oxidase (CO). *Source*: Shi et al. (2016)/PLOS/CC BY 4.0. (b) Gene map and organization of the complete mitochondrial genome of the terrestrial carnivorous plant *Utricularia reniformis*. The mitochondrial genes are shown in the outer circle. Genes are color-coded by their function in the legend. *Source:* Silva et al. (2017)/PLOS/CC BY 4.0. Note the roughly 50-fold size difference between these animal and plant mitogenomes (17,137 bp versus 857,234 bp).

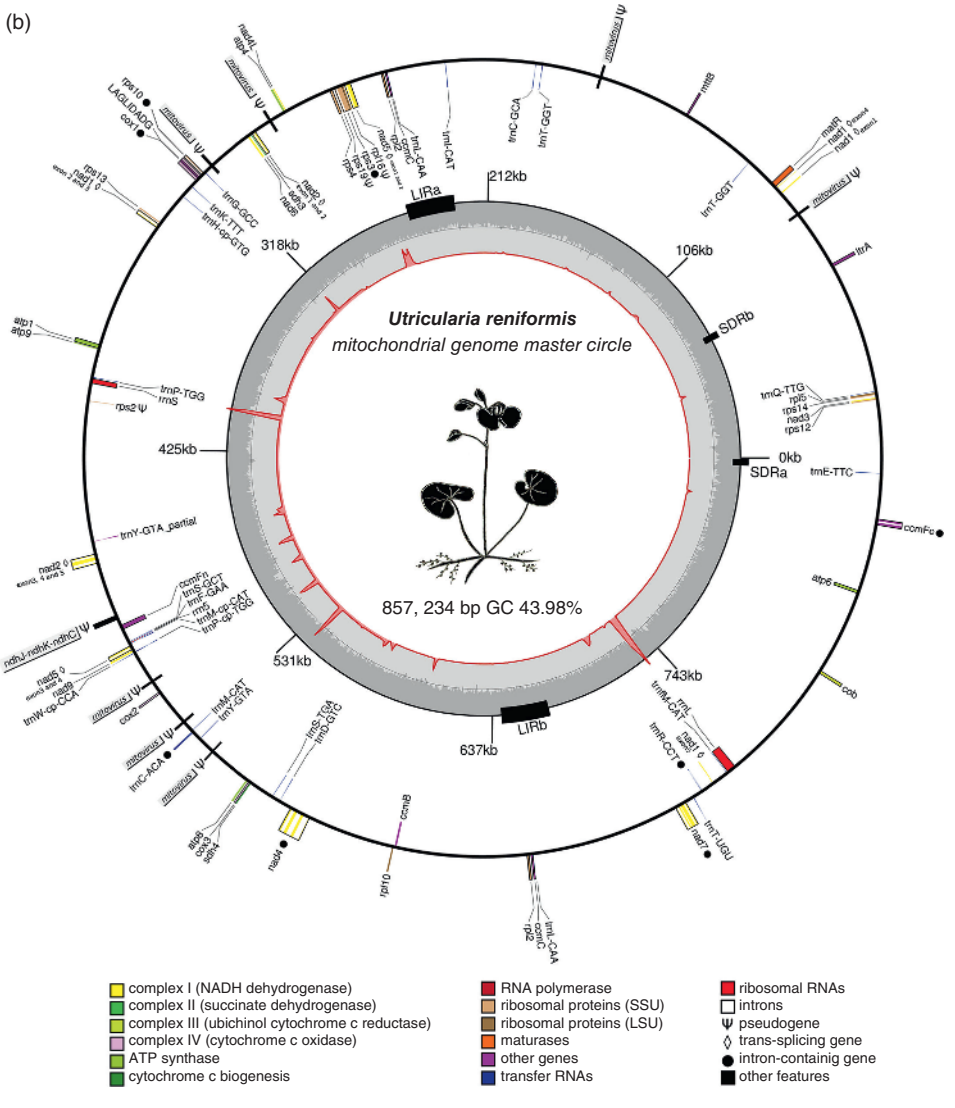

Figure 1.5 (Continued)

although these mutation rates vary considerably across taxa. One study that analyzed data from 4676 species of animals found that invertebrates including insects and arachnids had mutation rates that were between 2 and 6 times higher than their nuclear genome mutation rates, whereas mtDNA mutation rates in vertebrates –including reptiles and birds – were on average more than 20 times higher than nrDNA mutation rates (Allio et al. 2017). Variable mtDNA mutation rates can also be found within taxonomic groups; for example, while several members of a family of land snails showed mtDNA mutation rates that were between 2 and 5× higher than nrDNA mutation rates, in one species the mtDNA mutation rate was nearly 40× higher than the nrDNA rate (Allio et al. 2017).

In most animals, mtDNA is maternally inherited, meaning that it is passed down from mothers to their offspring (daughters and sons), and not from fathers to their offspring.

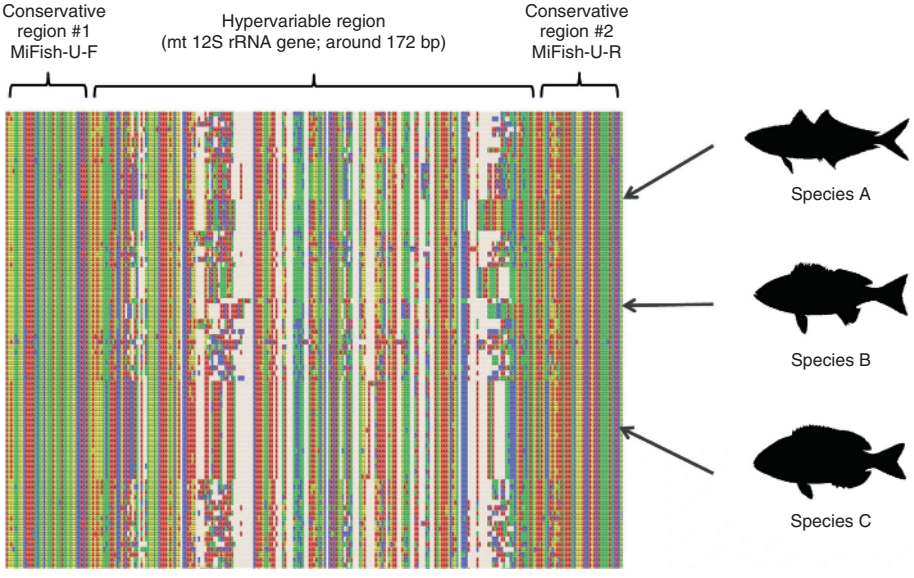

Figure 1.6 An alignment of a region of the mitochondrial 12SrRNA gene from multiple fish species (an alignment is a comparison of DNA sequences that shows the similarities among sequences from multiple individuals or species). Each row of letters represents the sequence of a different fish (e.g. Species A, B, and C) in which the nucleotides (G, A, T, and C) have different colors. When vertical bands show the same color across most or all rows, that means that most or all individuals have the same nucleotide at that position. Vertical bands that include multiple colors and/or white spaces reflect regions of the gene that are not conserved among species (i.e. they do not all have the same nucleotide at that site). This alignment shows two conserved regions at either end, and the authors of this study designed universal fish primers (MiFish-U-F and MiFish-U-R) complementary to these regions that they could use to amplify a section of 12SrRNA (a mitochondrial gene that encodes a small rRNA) from a large number of fish species. In contrast, universal primers could not be designed for the hypervariable region between the two conserved regions. *Source:* Miya et al. (2020)/Springer Nature/CC BY 4.0.

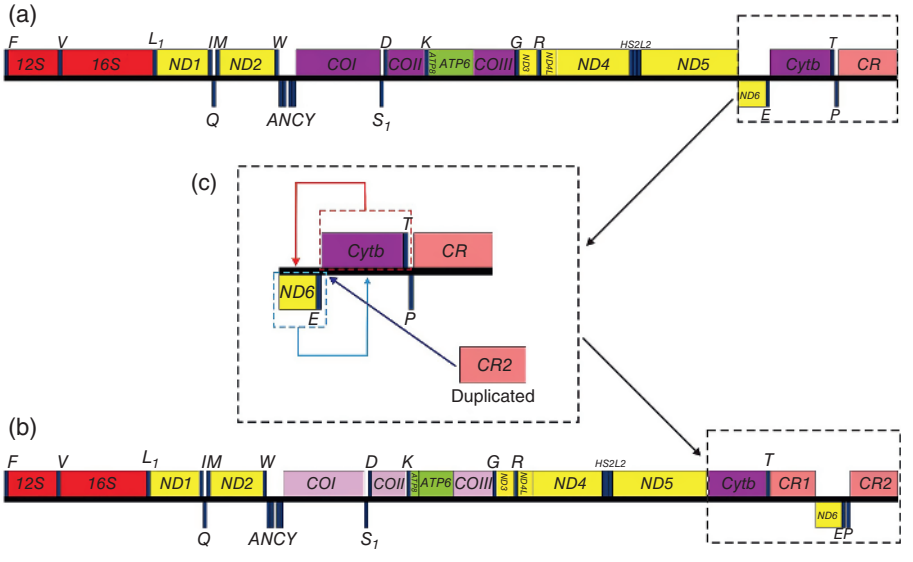

Figure 1.7 (a) A comparison between the mitochondrial gene arrangements of most vertebrates with (b) those of the daggertooth pike conger *Muraenesox cinereus* shows (c) the rearrangement/duplication of three different genes. *Source*: Zhang et al. (2021)/Springer Nature/CC BY 4.0.

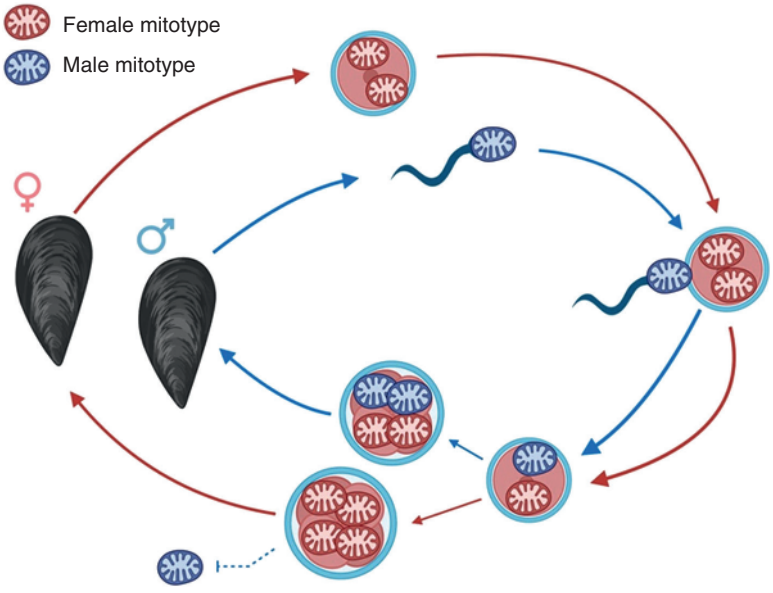

Figure 1.8 Diagram illustrating the process of doubly uniparental inheritance of mitochondrial DNA in blue mussels (*Mytilus* sp.). Mitotype refers to the two forms (M and F) of mitogenomes. Blue indicates the movement of M-type mtDNA and red indicates F-type mtDNA. Solid arrows represent the direction of the life cycle. The dotted line signifies the elimination of M-type mtDNA from an embryo, resulting in the formation of a female mussel. *Source:* Bramwell et al. (2024)/with permission of Elsevier.

However, in some cases paternal leakage occurs, which means that the sperm's mitochondria are not eliminated during fertilization, and the zygote has both maternal and paternal mitochondria. This is termed **heteroplasmy** (two or more mitochondrial genome variants in the same cell), in contrast to the more common **homoplasmy** (cells in which all mitochondrial genomes are the same, barring new mutations). One example of this was found in some hybrid offspring that resulted from interbreeding between loggerhead (*Caretta caretta*) and hawksbill (*Eretmochelys imbricata*) turtles and which had mitogenomes from both species (Vilaça et al. 2023). A more striking exception to the rule of uniparental inheritance of mtDNA in animals is **doubly uniparental inheritance** (**DUI**), which is found in some bivalve species including freshwater mussels in the order Unionida, plus some species in the orders Mytiloida (marine mussels), Veneroida (saltwater and freshwater bivalves including clams), and Nuculanoida (very small saltwater clams). In DUI, there are two distinct mitochondrial lineages, one associated with females (female or F-type mtDNA) and one associated with males (male or M-type mtDNA). The female type is transmitted through the eggs to all offspring, whereas the male type, which is present in the sperm and enters all eggs at the time of fertilization, is retained in the male – but not the female – offspring (Figure 1.8).

1.4.2 Plant Mitochondrial DNA (mtDNA)

Overall, plant and animal mitochondrial genomes have similar functions: they both generate energy in the form of ATP, and both are commonly maternally inherited, although paternal inheritance is the norm in some **gymnosperm** species (conifers, cycads, and ginkgos). However, there are substantial structural differences, starting with

the fact that mitochondrial genomes in plants are considerably larger and more complex than those in animals. While most animal mitogenomes are ~16.5 kb in size, most plant mitogenomes range in size from 200 to 700 kb. As always, there are exceptions; for example, the hemiparasitic mistletoe *Viscum scurruloideum* has a mitogenome that is 66 kb (Liu et al. 2014), whereas that of the flowering plant *Silene conica* is approximately 11,000 kb (Sloan et al. 2012), demonstrating a more than 150-fold size difference between species. A lot of this size variation can be explained by different amounts of noncoding DNA, including variable numbers of repeated sequences, and sequences that transferred from nuclear and chloroplast genomes. However, there is also considerable variation in the number of genes and introns: plant species harbor anywhere from 32 to 67 genes in their mitochondrial genomes and can have 18–25 introns (Mower et al. 2012). Considerable variation in noncoding sequences means that the gene number is not a good predictor of total mitochondrial genome size; for example, the terrestrial carnivorous plant *Utricularia reniformis* has a mitochondrial genome of ~857 kb which includes 70 transcriptionally active genes (42 protein-coding genes, 25 tRNA genes, and 3 rRNA genes; Figure 1.5b) (Silva et al. 2017), whereas *Bupleurum chinense*, a terrestrial plant in the family Apiaceae, has a smaller mitogenome (~435 kb) but a larger number (78) of transcriptionally active genes (39 protein-coding genes, 35 tRNA genes, and 4 rRNA genes) (Qiao et al. 2022). The complexity of plant mitogenomes is further illustrated by the fact that while many are in the form of circular molecules, others exhibit linear conformations and branched structures; for example, the ~5500 kb mitochondrial genome of Sitka spruce (*Picea sitchensis*) comprises a relatively small 168 kb circular segment of DNA, plus a branching structure of ~5400 kb (Jackman et al. 2020).

Another difference between animal and plant mitogenomes is their mutation rates: while mtDNA evolves more quickly than nrDNA in many animals (as discussed earlier), plants typically have slower evolutionary rates in mtDNA compared to their nrDNA. But once again, there are exceptions to this rule; for example, a few angiosperm clades have particularly high mutation rates in mtDNA (up to 5000× higher compared to closely related taxa) (Zwonitzer et al. 2024). Conversely, some plant taxa show a particularly low mtDNA mutation rate; for example, the mitochondrial genome of *Liriodendron tulipifera* has an overall mutation rate of only 0.035 nucleotide substitutions per site per billion years, which is ~2000 times slower than the mutation rate of human mtDNA (Richardson et al. 2013). Variation in mtDNA mutation rates in plants has been linked to both the size of the mitogenome and the number of mitogenomes per cell (Zwonitzer et al. 2024).

Despite having overall lower mutation rates, plant mitogenomes show much higher recombination rates than animal mitogenomes. Recombination often involves large (up to several kb) **repetitive sequences**, which are stretches of the same DNA sequence repeated multiple times. Together, recombination, size variation, and structural complexity mean that plant mtDNA markers are used relatively infrequently in molecular ecology. As a result, fewer whole mitogenomes from plants have been sequenced compared to animal mitochondrial genomes, plant nuclear genomes, and plant chloroplast genomes (discussed below).

1.4.3 Chloroplast DNA (cpDNA)

Within the cells of plants and algae are chloroplasts, a type of plastid that contains chlorophyll and performs photosynthesis. Chloroplasts also contain the chloroplast genome (cpDNA), which is much more widely used than the mitochondrial genome

for plant and algal studies in molecular ecology. Thousands of complete chloroplast genomes have now been sequenced, and the majority have revealed genome sizes between 100 and 200 kb (Wu et al. 2022). Most chloroplast genomes have 120–130 genes distributed across the large single-copy (LSC) region and the small single-copy (SSC) region, and these are primarily involved in photosynthesis-related transcription and translation. The LSC commonly ranges from 80,000 to 90,000 bp in length and contains a majority of the protein-coding genes, as well as some tRNA and rRNA genes. The SSC is commonly 15,000–25,000 bp in length and contains a smaller set of genes compared to the LSC region. In most species, the LSC and SSC are separated by two copies of an inverted repeat (IR), which is a segment of DNA that commonly ranges from 10,000 to 20,000 bp in length and that appears twice with the second version of the sequence complementary to the first (complementary sequences are shown in Figure 1.1, in which ATCG is complementary to CGAT) (Figure 1.9). Like mtDNA, cpDNA is uniparentally inherited: in most angiosperms (flowering plants), it is maternally inherited, whereas in most gymnosperms (conifers and cycads), it is paternally inherited (Table 1.2).

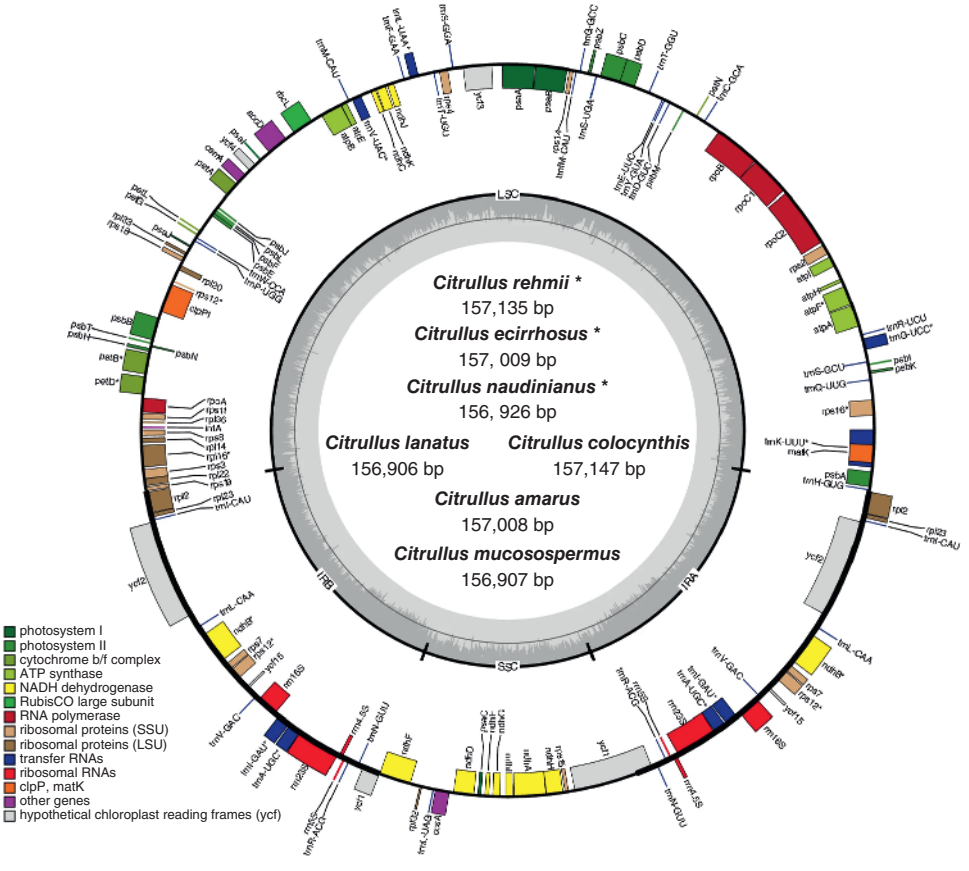

Figure 1.9 Complete chloroplast genome map showing conserved size and arrangement among seven *Citrullus* (watermelon) species. Boxes with different colors represent different functional groups of genes. The inner circle demarcates the large single-copy (LSC) region, the small single-copy (SSC) region, and the two inverted repeats (IRA and IRB). *Source:* Zhou et al. (2023)/Springer Nature/CC BY 4.0.

Table 1.2 The most common modes of inheritance of different genomes and genomic regions.

Genome	Most common mode of inheritance
Animals	
Nuclear genome (autosomal chromosomes)	Biparental
Nuclear genome (sex chromosomes)	Uniparental, e.g. Y chromosome paternally inherited in mammals, W chromosome maternally inherited in Lepidoptera and birds
Higher plants	
Nuclear genome (autosomal chromosomes)	Biparental
Nuclear genome (sex chromosomes)	Uniparental inheritance (paternal) in some dioecious plants
Mitochondrial DNA	Uniparental: Maternally inherited in angiosperms and many gymnosperms Paternally inherited in some gymnosperm families, e.g. Taxaceae and Cupressaceae
Chloroplast DNA	Uniparental: Maternal in most angiosperms Paternal in most gymnosperms Biparental in some plants

Compared to plant mitogenomes, the chloroplast genome has a more conserved gene order and a lower rate of recombination, making it a good target for universal primers (as discussed for animal mtDNA). The chloroplast genome also has overall higher mutation rates than plant mitochondrial genomes, but lower mutation rates than plant nuclear genomes. This was illustrated by one study that compared mitochondrial, chloroplast, and nuclear gene regions from 27 species of seed plants and found that the overall relative mutation rates of mitochondrial, chloroplast, and nuclear genes are 1:3:10 (e.g. the mutation rate in cpDNA was 3× faster than mtDNA and roughly 1/3 the rate of nrDNA). This ratio varied between taxonomic groups, being approximately 1:2:4 in gymnosperms and 1:3:16 in angiosperms, but in all cases, mitochondrial genes showed the lowest variability followed by chloroplast genes, with nuclear genes showing notably higher rates of evolution (Drouin et al. 2008). Relatively high mutation rates, low recombination, conserved gene order, and the availability of universal primers mean that the use of cpDNA markers in plants has been somewhat comparable to the use of mtDNA markers in animals.

1.4.4 Molecular Markers from Organellar Genomes: A Summary

The emergence of molecular ecology initially saw a heavy reliance on molecular markers from mtDNA in animal studies and cpDNA in plant studies; data from both genomes were relatively easy to generate, and this meant that they were widely used in studies employing genetic data to address taxonomic questions, calculate population genetic metrics, and investigate the evolutionary relationships among populations and species. These genomes are still being characterized for several reasons. First, it is often easier, faster, and cheaper to generate these types of data compared to nuclear data, and therefore

animal mtDNA and plant cpDNA data may be a more realistic option for labs that lack access to more specialized equipment or bioinformatic expertise. Second, and as we shall see in later chapters, data from mtDNA and cpDNA genomes may be sufficient for some types of questions in molecular ecology, for example, identifying taxa (discussed in Chapter 2). Nevertheless, data from organellar genomes are increasingly supplemented or replaced by data from nuclear genomes, as discussed in the next section.

1.5 Nuclear Genomes

Nuclear genomes are much, much larger than either mitochondrial or chloroplast genomes, although their sizes vary enormously across taxa. Note that the word "genome" is often used as a shortcut for the nuclear genome, with the prefix mitochondrial or chloroplast added when the smaller genomes are being discussed. Genome sizes are typically given as "C-values," which can be quantified as either units of mass or number of base pairs in a single set of chromosomes (Box 1.2). Genome sizes vary substantially across taxonomic groups (Figure 1.10). Bacteria and Archaea have relatively small genomes, normally ranging from approximately 140 kb to 15 Mbp, with most of the variability attributed to differences in the number of protein-coding genes (Han et al. 2013). Eukaryotes, on the other hand, have more than a 100,000-fold difference in their genome sizes (Wang et al. 2021), with little or no relationship between genome size and organismal complexity; instead, genome size variation is influenced considerably by

Box 1.2 Genome Size Databases

Interest in genome sizes has grown substantially in recent years partly because we now know that the genome size in plants can influence growth strategies, community composition, interactions with animals, and ecosystem dynamics (reviewed in Pellicer and Leitch 2020). Other studies have shown that relatively small genomes appear to have been influenced by the metabolic rate associated with powered flight in birds (Wright et al. 2014) and life history traits in frogs (Lamichhaney et al. 2021). In fungi, genome size is linked to ecological strategies related to resource exploitation and habitat preferences along soil fertility gradients (Zhang et al. 2023).

Our growing awareness of the importance of genome size is linked to the expanding databases that record genome sizes across taxa. These include the Plant DNA C-values Database, maintained by the Royal Botanic Gardens, Kew (UK), which by 2025 included C-value data for 12,273 species: 10,770 angiosperms, 421 gymnosperms, 303 pteridophytes (246 ferns and fern allies and 57 lycophytes), 334 bryophytes, and 445 algae (https://cvalues.science.kew.org/). As of 2025, the Animal Genome Size Database included genome sizes from 6534 species (3863 vertebrates and 2671 non-vertebrates) (https://www.genomesize.com/). Reference databases that include information about genomes sizes are expected to grow rapidly in future years; for example, the Vertebrate Genomes Project is an international collaboration that aims to generate reference genomes for all of the ~70,000 extant vertebrate species (https://vertebrategenomesproject.org/). These expanding databases are linked to the continued proliferation of genetic data because genome sizes can be estimated from WGS.

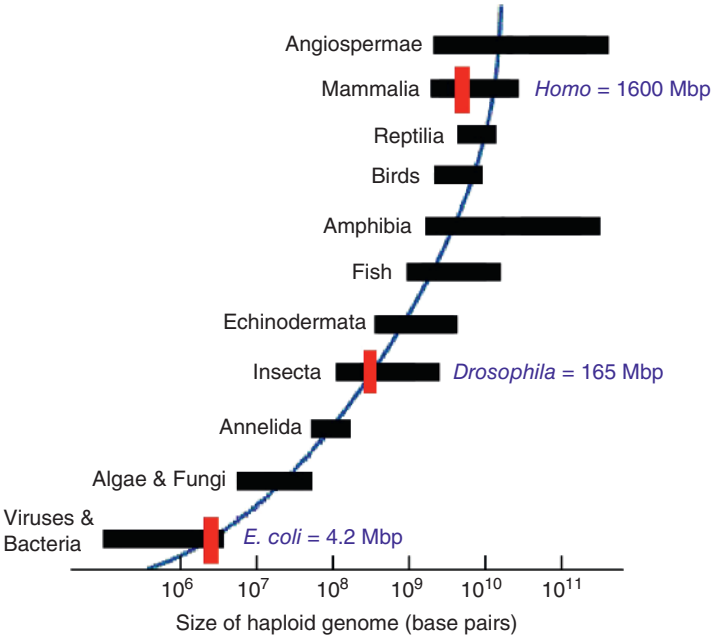

Figure 1.10 Genome sizes vary considerably across taxa. Genome size (C-value) can be measured as either the number of bp or the mass of DNA in a *haploid* chromosome set; in other words, it is based on a single set of chromosomes. The actual number of base pairs or mass of DNA in the cell of a *diploid* organism (an organism with two sets of chromosomes) is therefore twice that shown here.
Source: 2004 by Steven M. Carr.

repetitive regions such as **transposable elements** (**TEs**), also known as "jumping genes." TEs, which were discovered in the 1940s by Barbara McClintock, are DNA sequences that typically fall within the size range of 100–10,000 base pairs, although in some cases can be much larger than that. TEs can replicate themselves independently of the host genome and then move around the genome, and they now comprise a substantial proportion of most eukaryotic genomes. The importance of TEs to genome size variation was illustrated by a review of 74 flowering plant genomes – each from a different family – which found that in all species repeated regions (primarily TEs) made up a large proportion of the genome: the smallest contribution was 21.59% of the genome in the duckweed species *Spirodela polyrhiza*, and the largest contribution was 83.23% of the corn (*Zea mays*) genome (Wang et al. 2021). Similar ranges are found in vertebrates; for example, 6% of the genome comes from TEs in the pufferfish *Tetraodon nigroviridis*, compared to 55% in the zebrafish *Danio rerio* (Chalopin et al. 2015).

Regardless of the size of the genome, each harbors a tremendous diversity of DNA, including protein-coding genes, regulatory genes, large repetitive regions such as TEs (discussed earlier), smaller repetitive regions that include minisatellites (motifs of 10–100 bp repeated many times in succession) and pseudogenes (nonfunctional copies of functional genes). Until a few years ago, relatively few studies in molecular ecology incorporated nuclear genetic data. Most exceptions to this were studies that generated data from well-characterized genes such as regions within the ribosomal DNA complex (e.g. **internal transcribed spacers** – ITS1 and ITS2), genes within the **major**

histocompatibility complex (MHC; a gene family involved in immune response), and sex chromosomes in mammalian species. This has now changed with the increasing accessibility of genome-wide sequencing. Before discussing common approaches to generating nuclear genomic data in molecular ecology, we will look at some of the important characteristics of the nuclear genome, beginning with a review of key processes that generated – and continue to generate – genetic variation within and among species. Genetic variation is essential to every aspect of molecular ecology, and understanding the underlying processes is necessary before we can make sense of genetic data.

1.6 Novel Genotypes: Mutation and Recombination

Mutations create novel genetic material by altering DNA sequences. Many mutations occur during **DNA replication**, a process that uses existing DNA sequence as a template to create new DNA sequences. First, the hydrogen bonds that hold the two strands of the DNA double helix together are broken by the enzyme helicase, resulting in two single strands of DNA. These single strands are the templates along which new, complementary strands of DNA are synthesized. **DNA polymerase** is the enzyme that catalyzes the synthesis reaction by adding complementary nucleotides that pair G with C, and A with T. The synthesis of new strands occurs in the 5′-3′ direction (Figures 1.1 and 1.11). Once complete, the parent double-stranded DNA will have been replaced by

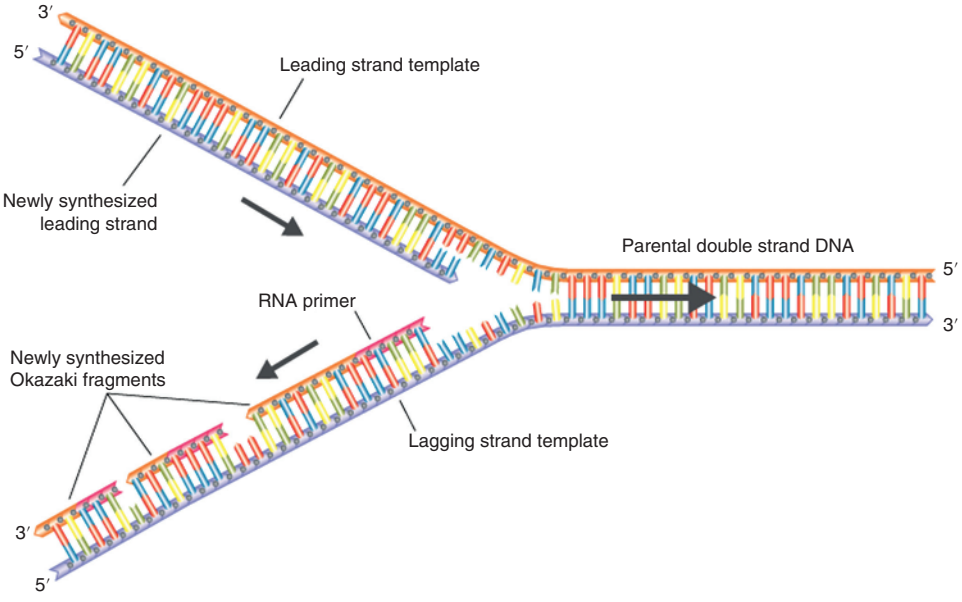

Figure 1.11 Graphic showing DNA replication. The arrows indicate the direction of synthesis, which occurs in the 5′-3′ direction. Continuous replication occurs along the leading strand because the 3′ to 5′ directionality of its template strand allows the polymerase (the enzyme that catalyzes replication) to follow the replication fork without interruption. The 5′ to 3′ directionality of the lagging strand means the polymerase must work backward from the replication fork, causing periodic breaks. This creates multiple fragments, known as Okazaki fragments, which are then connected by DNA ligase into a single, continuous strand of DNA. *Source:* HEE Genomics Education Programme/Wikimedia commons/CC BY 2.0.

two sets of DNA that each have one parent strand and one newly synthesized strand; this is known as semiconservative replication.

When errors occur during DNA replication, mutations can result. These include **transitions**, which arise when either one **purine** (A and G) is substituted for another or one **pyrimidine** (C and T) is substituted for another. Less common are **transversions**, which describe the replacement of a purine with a pyrimidine, or vice versa. As noted earlier (Table 1.1), redundancy in the genetic code means that nucleotide substitutions do not necessarily alter the encoded amino acid, and these are called **synonymous substitutions** because they result in different DNA sequences that encode the same amino acid. In contrast, substitutions that alter the encoded amino acid are known as **non-synonymous substitutions**. A **nonsense substitution** is one that introduces a new stop codon (Table 1.1); these are normally deleterious because they signal the end of transcription and hence lead to incomplete protein synthesis. Although many single-nucleotide substitutions have no phenotypic outcome, some have important effects. One example of this occurs in populations of Bananaquits (*Coereba flaveola*), a songbird living in Central and South America which can have either the widespread yellow morph with a dark back and yellow breast, or a less common melanic morph, which is almost entirely black. A single non-synonymous nucleotide substitution in the *melanocortin-1 receptor* (*MC1R*), a gene that helps regulate feather color, determines whether a bird has the yellow morph or the melanic morph (Theron et al. 2001). In addition to nucleotide substitutions, nucleotide insertions or deletions (collectively referred to as **indels**) can occur during DNA replication if errors by the DNA polymerase add or remove one or more nucleotides from a sequence. If this occurs in a coding region and changes the groups of three nucleotides that each encode an amino acid, it is called a **frameshift mutation**.

While small mutational changes such as those described above can be extremely important, another category of mutations known as **structural variants (SVs)** can profoundly impact genetic diversity. These include insertions, deletions, inversions, duplications, and translocations (Figure 1.12) and are sometimes described as genomic alterations that involve DNA regions longer than 50 bp – often much longer – although others have argued that SVs should collectively refer to all sizes of insertions, deletions, inversions, and duplications (Mérot et al. 2020). SVs are often categorized as either balanced SVs that neither increase nor decrease the amount of genetic material, or unbalanced SVs, which lead to either a loss or gain (duplication) of part of the genome. We are just beginning to learn about the importance of SVs, largely enabled by sequencing technologies that allow researchers to assemble sequences long enough to identify large-scale rearrangements. It is likely that most or possibly all genomes have been substantially influenced by SVs; for example, a comparison of genome sequences within and among seven songbird species in the genus *Corvus* identified 220,452 insertions, deletions, and inversions including a 2.25 kb insertion of a retrotransposon (a type of transposable element; Section 1.5) that leads to plumage differences among European crows by reducing expression of the *NDP* gene (Weissensteiner et al. 2020).

While mutations are key to the creation of genetic variation, another important process that leads to genetic differences among individuals is recombination. Here we are talking about recombination that occurs during meiosis – the process of cell division that reduces the complement of chromosomes by half and results in gametes (eggs and sperm). This is an essential part of sexual reproduction because it ensures that each

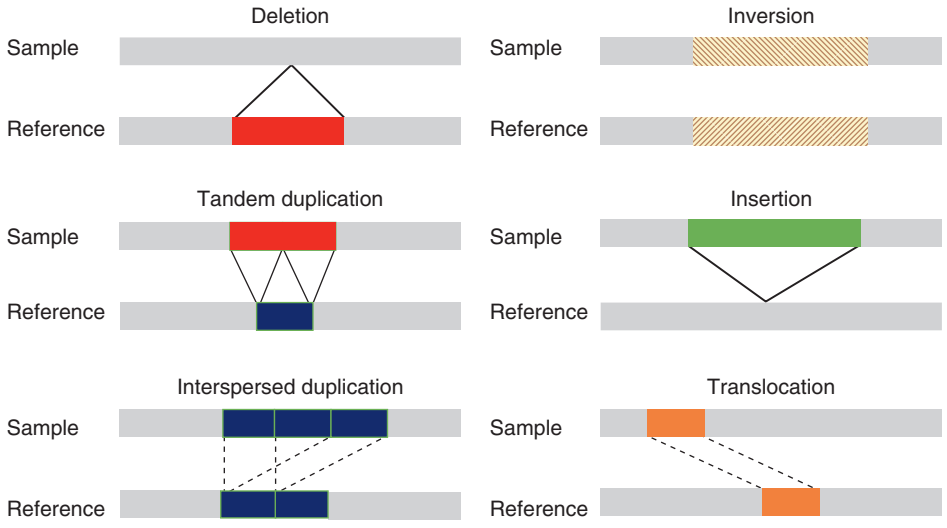

Figure 1.12 Schematic representation of different types of structural variants (SVs). Unbalanced SVs are represented by deletion, insertion, tandem duplication, and interspersed duplication, all of which can lead to the gain or loss of a part of the genome. Balanced SVs are represented by inversion and translocation, which alter the genome without any gain or loss. "Reference" refers to the initial sequence, and "Sample" refers to the sequence that has been altered by an SV. *Source:* Romagnoli et al. (2023)/Frontiers Media S.A/CC BY 4.0.

generation will not experience a doubling of chromosomes (Box 1.3). During the first stage of meiosis, pairs of chromosomes (one maternal and one paternal) that have the same genes in the same order (**homologous chromosomes**) align with one another. At this time, the arms of the chromosomes can overlap and temporarily fuse, causing a crossover that enables the exchange of genetic material between the maternal and paternal chromosomes. This is known as recombination (Figure 1.13) and ensures that parents and offspring have different combinations of genes. Recombination is more common among genes that are spaced further apart on the chromosome; genes that are close together are often "linked" (in **linkage disequilibrium**), which means that they are jointly inherited because they are too close together on the chromosome to become separated during recombination. Genes that are in linkage disequilibrium do not follow Mendel's law of independent assortment because they are not inherited independently of one another.

1.7 Epigenetic Marks

As summarized in Section 1.2, transcription and translation are the processes that enable DNA sequences to encode the proteins that underlie **phenotypes** (the observable characteristics of an individual including morphology, physiology, and behavior). Even though the genetic material is the same within every cell of an individual, different genes are expressed at different times because of cellular signals, developmental cues, and environmental stimuli; this is why different proteins are encoded in a nerve cell compared to a smooth muscle cell. Certain types of DNA mutations alter gene

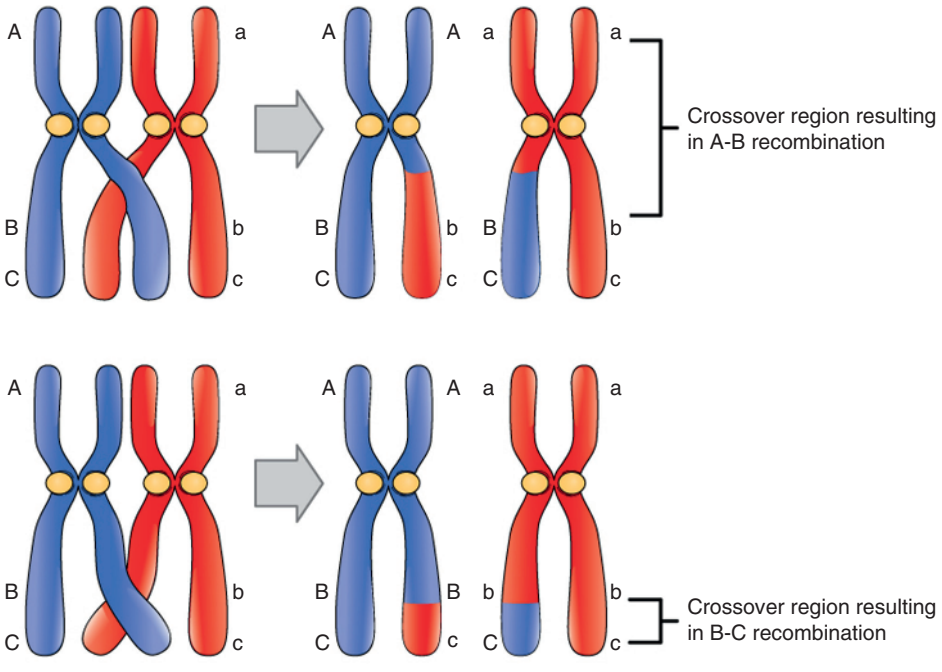

Figure 1.13 Two examples showing how recombination during meiosis can lead to different combinations of genes among individuals. *Source:* Rice University/CC BY 4.0.

Box 1.3 Chromosome Counts

Many – although by no means all – animals are diploid ($2n$), which means that they have two sets of chromosomes: one inherited from their mother and one inherited from their father. Meiosis results in gametes (eggs and sperm) that normally contain a single set of chromosomes, which ensures that when gametes fuse to form a zygote they will produce an offspring that is once again diploid, like its parents. The term "n" tells us how many different chromosomes there are in a "set" of chromosomes, which often includes a combination of **autosomes** and sex chromosomes. Each gamete in humans, for example, contains 22 autosomes and 1 sex chromosome (X or Y), which leads to $n = 23$ in humans. When the gametes fuse they create diploid offspring, which have one set of chromosomes from each parent, and their **karyotype** (the complement of chromosomes in a somatic cell) is $2n = 46$. In this case, $2n = 46$ comprises 22 pairs of autosomes and 2 sex chromosomes (either 2 X chromosomes in a female or 1 X and 1 Y chromosome in a male). See also Figure 1.14.

Some species of vertebrates (primarily fishes, reptiles, and amphibians), invertebrates (including some insects and crustaceans), and fungi are **polyploid**, which means that their cells each contain three or more sets of chromosomes. In addition, >90% of living flowering plants are either polyploid or had at least one polyploid ancestor (Adams and Wendel 2005; Cui et al. 2006). **Autopolyploid** individuals have >2 sets of chromosomes that all originated from a single ancestral species following errors in meiosis that led to 1 or more gametes with unreduced chromosome sets. **Allopolyploid** individuals also arose following meiotic errors, but the sets of chromosomes originated in two or more different species and led to polyploidization following interspecific hybridization (interbreeding between two different species) (Figure 1.15).

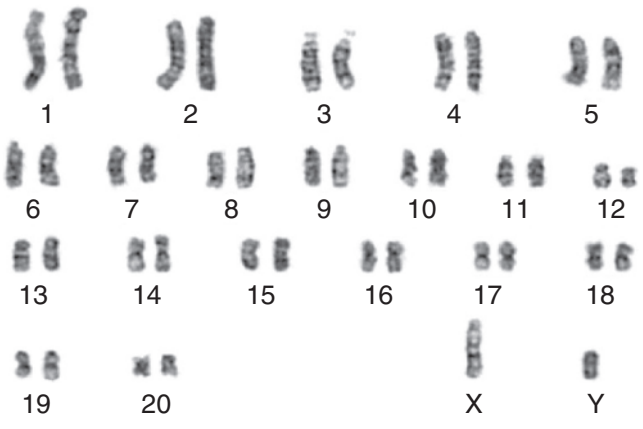

Figure 1.14 Karyotype of a rat (*Rattus norvegicus*). Based on this karyotype, we can tell that the rat is diploid (two sets of chromosomes) and has 20 pairs of autosomes and 1 pair of sex chromosomes (X and Y). The karyotype can therefore be described as $2n = 42$. A normal rat gamete (egg or sperm) would have the haplotype $n = 21$. *Source:* Hamanaka et al. (2011)/PUBLIC LIBRARY OF SCIENCE (PLOS)/CC BY 4.0.

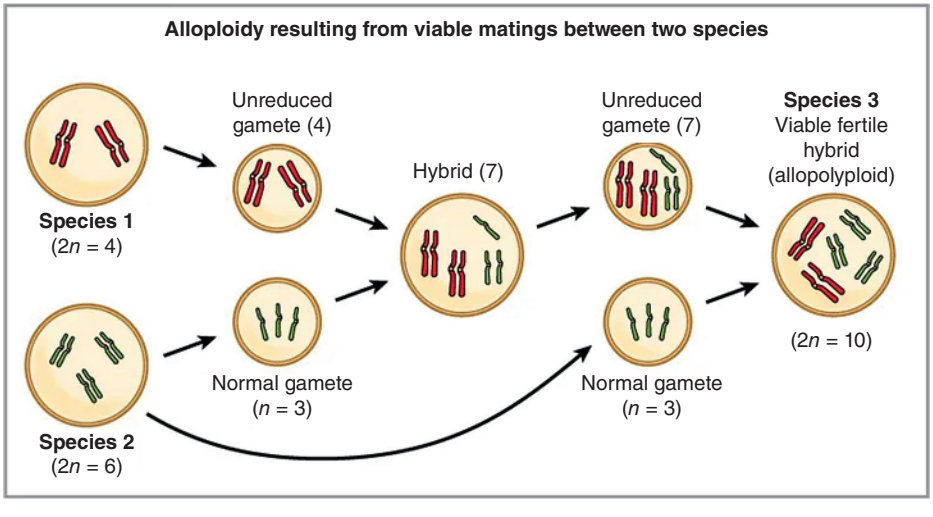

Figure 1.15 An example showing how two mating events between different taxa can lead to an allopolyploid. In this example, two diploid species interbreed. Species 1 produced an unreduced gamete ($2n$ instead of n) following nondisjunction during meiosis, whereas Species 2 produced the typical haploid gamete. The two species form a hybrid with seven chromosomes (four from Species 1 and three from Species 2). This hybrid in turn produced an unreduced gamete that, like its parent, also has seven chromosomes, and this unreduced gamete fuses with a normal gamete from Species 2. Their offspring is a viable allopolyploid with two copies of each of the five different chromosomes (which originated from two different species). *Source:* Clark et al. (2018)/Mary Ann Clark/CC BY 4.0.

expression, but expression can also be modified by **epigenetic** changes that do not alter DNA sequences ("epi" means above, epigenetic means above the genome). Epigenetic changes (also known as epigenetic marks) that occur throughout an individual's lifetime are essential to key developmental processes but can also be less predictably altered by environmental influences such as temperature or behaviors such as diet. One well-studied form of epigenetic modification is DNA **methylation**, which occurs when a methyl (CH3) group is added to a cytosine (C) nucleotide. Methylation most commonly occurs at cytosine–phosphate–guanine (CpG) dinucleotides (a cytosine nucleotide and a guanine nucleotide that are linked by a phosphate group), particularly in CpG islands, which are stretches of DNA with an abundance of CpG repeats (Figure 1.16). Excess

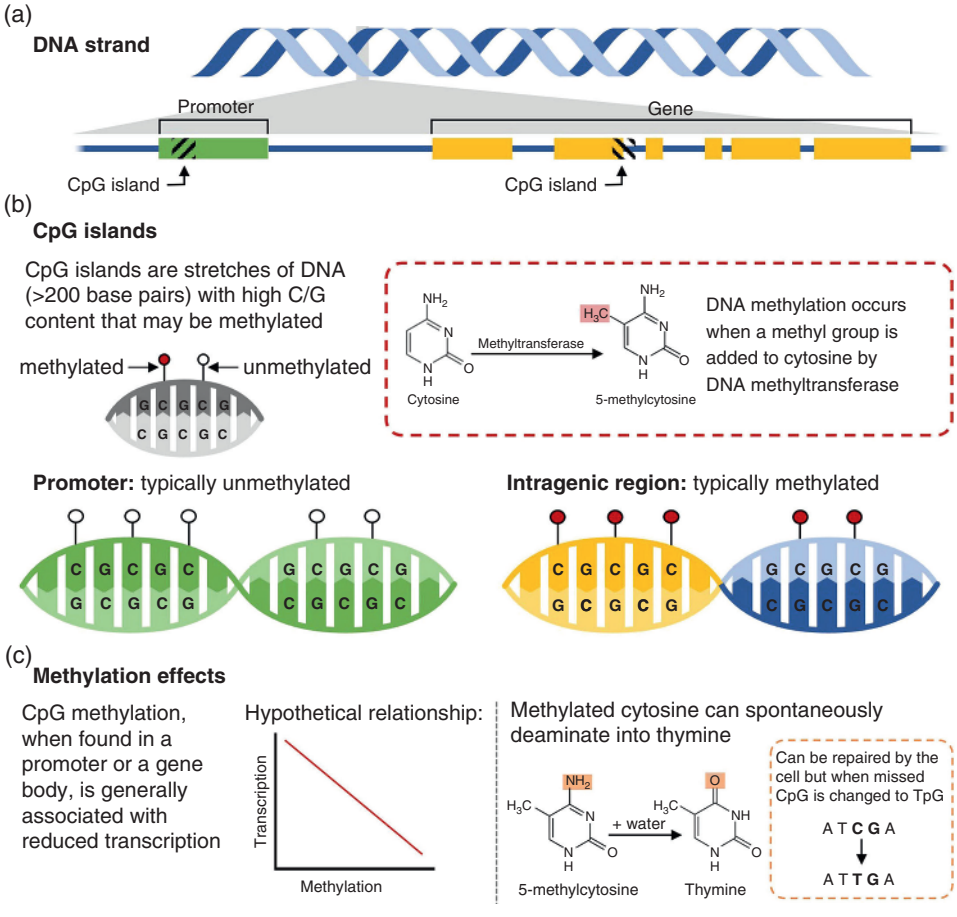

Figure 1.16 Methylation of cytosine–phosphate–guanine (CpG) islands is a form of epigenetic modification that influences transcription (gene expression). (a) CpG islands can be found within promoters (left) or in coding and noncoding regions (right). (b) In mammals, the CpG islands in promoters of genes that are transcriptionally active are unmethylated, while those in gene bodies are often methylated. (c) Methylation of CpG islands in promoters is generally associated with reduced transcription relative to unmethylated regions, and methylated cytosines can spontaneously deaminate into thymine, resulting in changes to the underlying DNA sequence. *Source:* Lamka et al. (2022)/Frontiers Media S.A/CC BY 4.0.

methylation typically reduces gene expression, thereby altering gene function. Histone modification is another type of epigenetic mark that arises when **histones** – proteins that act as spools around which DNA is wound – are modified in a way that causes chromatin (DNA + histone) to be wound either more or less tightly. This in turn makes DNA either more or less accessible to transcription and hence either increases or decreases gene expression. While we used to think that all epigenetic marks were acquired throughout an individual's lifetime, we now know that some epigenetic marks are heritable; in other words, they can be passed on from one generation to the next. We still have knowledge gaps surrounding epigenetics, but there is no doubt that epigenetic changes can be important modifications that impact ecology and evolution; for example, they can be associated with cold stress tolerance (Tang et al. 2018), heat stress tolerance (Dai et al. 2017), development (Metzger and Schulte 2017), sex determination (Ge et al. 2018), and timing of seasonal events such as reproduction (Viitaniemi et al. 2019).

1.8 RNA and Transcriptomes

Epigenetic modifications alter patterns of gene expression, which occurs when transcription converts DNA into mRNA, and translation then encodes the amino acids that comprise proteins (Section 1.2). Most RNA sequences are transcribed from protein-coding genes and are thus translated into proteins, although there are also some functional RNA molecules that do not produce proteins. The **transcriptome** is the set of all RNA transcripts and tells us which genes were expressed in a cell, tissue, or organism. In ecological research, the transcriptome can provide valuable insights into how organisms interact with their environment and respond to changing conditions by either **upregulating** (increasing gene expression) or **downregulating** (decreasing gene expression) different genes in response to altered cues that could include temperature, humidity, light, nutrient availability, and pollution. This information helps researchers understand how organisms adapt to their surroundings and cope with environmental stressors. For example, the water temperatures across the geographical range of spotted seatrout (*Cynoscion nebulosus*) range from near freezing to >30 °C. A transcriptomics study of spotted seatrout showed that individuals from warmer locations had a substantial transcriptomic response to acute cold stress and those from cooler locations had a substantial transcriptomic response to acute heat stress; together, these results show that changes in gene expression help spotted seatrout to tolerate a range of temperatures (Song and McDowell 2021). See also Box 1.4.

1.9 Proteins and Proteomics

Molecular ecology began with studies that were based on protein analyses, which predated our ability to analyze DNA and RNA from natural populations. In the 1960s, the field of population genetics emerged when researchers first characterized the proteins of natural populations of the fruit fly *Drosophila pseudoobscura* (Lewontin and Hubby 1966). These early genetic markers are usually referred to as **allozymes** – allelic forms (alleles) of the same protein-coding locus – and for several decades these were used extensively. Allozyme characterization provided an extremely important

Box 1.4 Epigenetics, Transcriptomics, and Phenotypic Plasticity

Phenotypic plasticity occurs when a single genotype can develop into multiple alternative phenotypes depending on environmental conditions (Figure 1.17). Epigenetic modifications can play a crucial role in phenotypic plasticity by regulating the expression of genes. Analysis of the transcriptome can identify which genes are being actively expressed and to what extent they are being expressed. These three ideas are interconnected: epigenetic modifications alter gene expression and hence the transcriptome, and this can result in different phenotypes without altering the underlying DNA sequences. Phenotypic plasticity can allow organisms to tolerate a range of environmental conditions such as temperature, nutrient availability, light intensity, and predation pressure. A link between epigenetic modifications, gene expression, and phenotypic plasticity was identified in barramundi (*Lates calcarifer*), a hermaphrodite fish in which individuals begin as males and subsequently change their sex to females. The timing of this male-to-female sex change is linked to the length of the fish: they must achieve a certain length before transitioning to female. This is described as the length-at-sex change, which varies among geographical regions and has been linked to differences in DNA methylation (a common type of epigenetic modification; see Section 1.7) plus variability in sea temperature and salinity. These associations led the authors to conclude that sex change is a phenotypically plastic characteristic that is epigenetically and environmentally mediated in barramundi (Adapted from Budd et al. 2022).

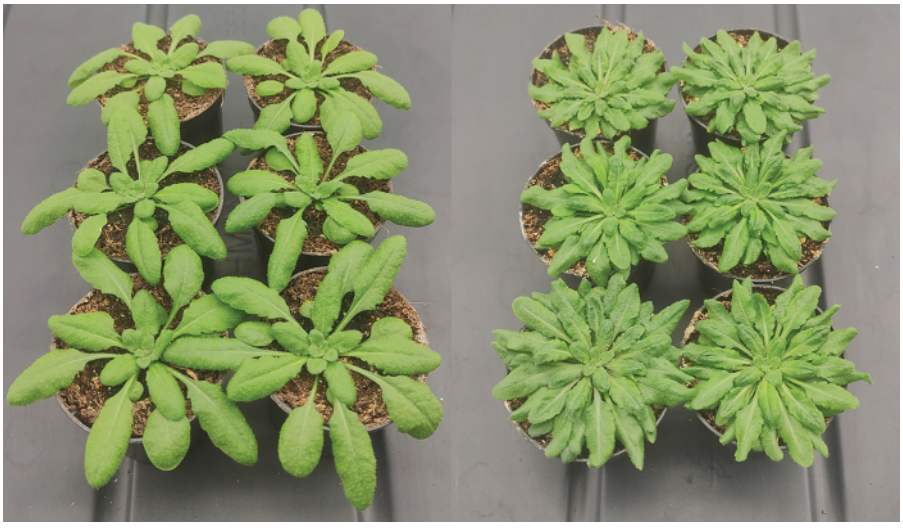

Figure 1.17 Phenotypic plasticity in *Arabidopsis thaliana*. These photos show the developmental differences between genetically identical plants that were grown under different light conditions. Phenotypic plasticity occurs when the same genotype can develop into different phenotypes depending on environmental conditions. *Source:* Andrea R Rudalima/ Wikimedia Commons/CC0 1.0/ https://commons.wikimedia.org/wiki/File:Phenotypic_plasticity_in_plants.png / last accessed on October 15, 2025.

breakthrough in the emerging field of molecular ecology because it allowed researchers to genetically characterize individuals from almost any species, although today very few studies characterize allozymes because of the subsequent development of much more informative - and less invasive - molecular markers (see Section 1.10).

These days, protein-based studies in molecular ecology are more likely to be based on **proteomics**, which is the large-scale study of proteins including their structures, functions, interactions, and expression levels within a biological system. Key aspects of proteomics include protein identification, which tells us which proteins are present in a given organism, tissue, or cell; protein quantification, which tells us how much of a protein is present; protein–protein interactions, for example, those that occur within cellular networks and pathways; and functional proteomics, which describes the roles and activities of proteins. One example of this helps to explain the coevolution of the venomous Northern Pacific rattlesnake (*Crotalus oreganus oreganus*) and its main prey species, the California ground squirrel (*Otospermophilus beecheyi*). Resistance proteins in blood serum can bind to venom and protect mammals that regularly interact with snakes. A proteomics study found that serum proteins from ground squirrels had higher binding affinities with the venom proteins of snakes from the same geographic area compared to snakes from more distant locations; therefore, coevolution between Northern Pacific rattlesnake venom toxicity and California ground squirrel venom resistance involves protein–protein interactions, which is one of the key components of proteomics (Gibbs et al. 2020).

As discussed earlier, the production of proteins is influenced by the underlying DNA sequences plus modifications such as epigenetics, which can influence which genes are expressed either as a result of inherited epigenetic marks or in response to changing environmental conditions. Yet another process that can help determine which proteins are being expressed is **alternative splicing**. The first step in transcription is the creation of a pre-mRNA molecule, which contains both exons (coding regions) and introns (noncoding regions). RNA splicing removes introns and joins exons together to form a mature RNA molecule; this is mediated by the spliceosome, which is a complex of RNA and protein molecules. In alternative splicing, different combinations of exons can be included or excluded from the final mRNA transcript (Figure 1.18). Alternative splicing can be an important source of variation and adaptation. One example of this was found in deer mice (genus *Peromyscus*), in which coat color – and hence visibility – is an important part of predator avoidance. Alternative splicing of the *Agouti* pigmentation gene in two deer mouse populations has led to locally adapted light-colored coats, despite the ancestral mice having dark-colored coats (Mallarino et al. 2017). Many other examples exist: with the increasing availability of genome-wide sequence data, researchers have discovered that almost all genes with multiple exons in model vertebrate species, and up to 70% of multi-exonic genes in plants, experience alternative splicing during transcription (Chaudhary et al. 2019; Merkin et al. 2012).

1.10 Commonly Used Methods in Molecular Ecology

Now that we have reviewed some of the key aspects of genetics that underlie molecular ecology, we will move on to general discussions of the methods that are commonly used in this field. In this section, we will discuss DNA extractions and PCR, both of which play essential roles in generating genetic data from natural populations.

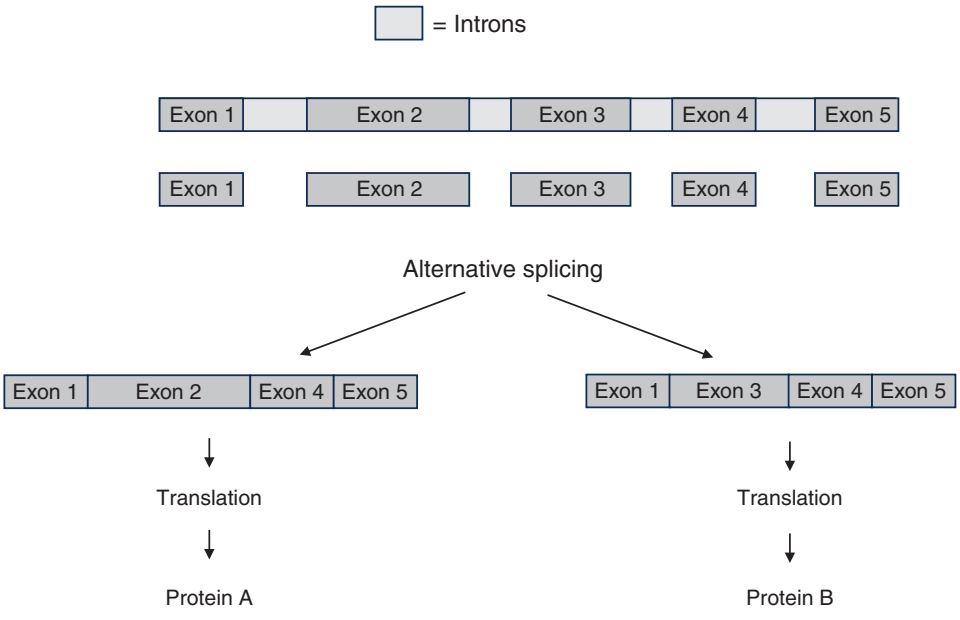

Figure 1.18 Overview showing one of the ways in which alternative splicing of the same gene can lead to different proteins.

1.10.1 DNA Extraction

Before DNA can be characterized, it needs to be extracted from organismal material such as whole organisms (e.g. bacteria and plankton), animal tissue (e.g. toenails, feathers, and blood), plant tissue (e.g. leaf, flower, root, and seed), or environmental samples (e.g. water, air, and soil). Samples must first be lysed – which means rupturing the cell wall or membrane – to release DNA from the cells. This includes chemical lysis (using detergents to disrupt cell membranes), enzymatic lysis (using proteases or lysozymes to digest proteins or cell walls), and mechanical lysis (using homogenization, grinding, or bead beating to physically break open cells). Lysis releases more than just DNA from the cells, and the next step is to separate DNA from other cellular components such as proteins, lipids, and RNA. Common methods for DNA purification include phenol-chloroform extraction, which uses organic solvents to separate DNA from proteins and other contaminants; silica-based spin columns (common to many DNA extraction kits), which selectively bind DNA while allowing contaminants to pass through; and magnetic bead-based purification, which uses magnetic beads coated with DNA-binding molecules to selectively capture DNA. The recovered DNA is resuspended in an appropriate buffer or solution for downstream applications. The concentration and purity of the extracted DNA can be determined using spectrophotometry or fluorometry.

1.10.2 Polymerase Chain Reaction (PCR)

One of the most important tools in molecular ecology is a thermocycler (also known as a thermal cycler or PCR machine). This is a critical piece of equipment because it allows researchers to start with a very small amount of DNA and make lots of copies of one or

more gene regions, which can then be characterized in different ways. **PCR** stands for **polymerase chain reaction**, a method that was developed by Dr. Kary Mullis in the 1980s, for which he became one of the recipients of the Nobel Prize for Chemistry. The importance of PCR to many biological disciplines including molecular ecology cannot be overstated.

PCR is most commonly done using extracted DNA, which provides the template that will be copied during PCR. The reactions include DNA, **primers** (also known as oligonucleotides), **DNA polymerase**, **dNTPs**, magnesium salt (usually MgCl2), and buffer. The primers, which are usually 15–35 bp long, provide the starting point for DNA synthesis. When a single region of DNA is targeted, the primers are each complementary to a DNA sequence that is either upstream (forward primer) or downstream (reverse primer) of the target sequence. As discussed below, some applications do not use primers that target specific DNA sequences, although regardless of the strategy, primers must always be added to the PCR cocktail because they provide the necessary starting point for DNA synthesis. Primers also determine which region(s) of DNA will be amplified. PCR reactions use a heat-stable polymerase, most commonly Taq polymerase, which was originally isolated from a bacterium called *Thermus aquaticus* that lives in hot springs, and therefore this enzyme is not deactivated at high temperatures. Deoxynucleotide triphosphates (dNTPs) are the building blocks used by DNA polymerase to synthesize new DNA strands. They consist of a deoxyribose sugar attached to a nitrogenous base ("N" means one of adenine, guanine, cytosine, or thymine: A, G, C, or T) and three phosphate groups to form dATP, dTTP, dCTP, and dGTP. $MgCl_2$ (or other salt) is the cofactor necessary for polymerase activity, and buffer is added to maintain the optimal pH and ionic conditions.

Once the cocktail is made and added to the template DNA, samples are placed in a PCR machine which cycles through different temperatures (hence the name thermal cycler). There are three main steps in PCR: denaturation, annealing, and extension (Figure 1.19). The denaturation step increases the temperature to ~94–98 °C, which breaks the hydrogen bonds between complementary base pairs and denatures double-stranded DNA into single-stranded DNA. The annealing step allows the primers to bind (anneal) to their complementary sequences on the single-stranded DNA template; this is normally accomplished at a temperature ~5 °C below the melting temperature of the primers (usually in the range of 40–65 °C). The third step is an extension, which describes the synthesis of new DNA. In this step, DNA polymerase uses the primers as starting points and the denatured DNA as a template to which it adds nucleotides (dNTPs) to the 3′ end of the primers, extending the DNA strands in a 5′ to 3′ direction by synthesizing strands that are complementary to the template. This is normally done at ~72 °C, which is the optimal activity temperature for Taq polymerase activity. PCR normally comprises 25–40 cycles of denature–anneal–extend. Each cycle generates two daughter strands for every parent strand, which means that the number of sequences increases exponentially throughout the PCR reaction. This results in many, many copies of the target DNA; for example, 35 cycles of amplifying a target region of DNA will generate ~68 billion copies of that fragment. This is sufficient to allow researchers to manipulate and visualize the DNA, as discussed below. This also means that researchers need only a very small starting amount of DNA in order to genetically characterize organisms (Box 1.5).

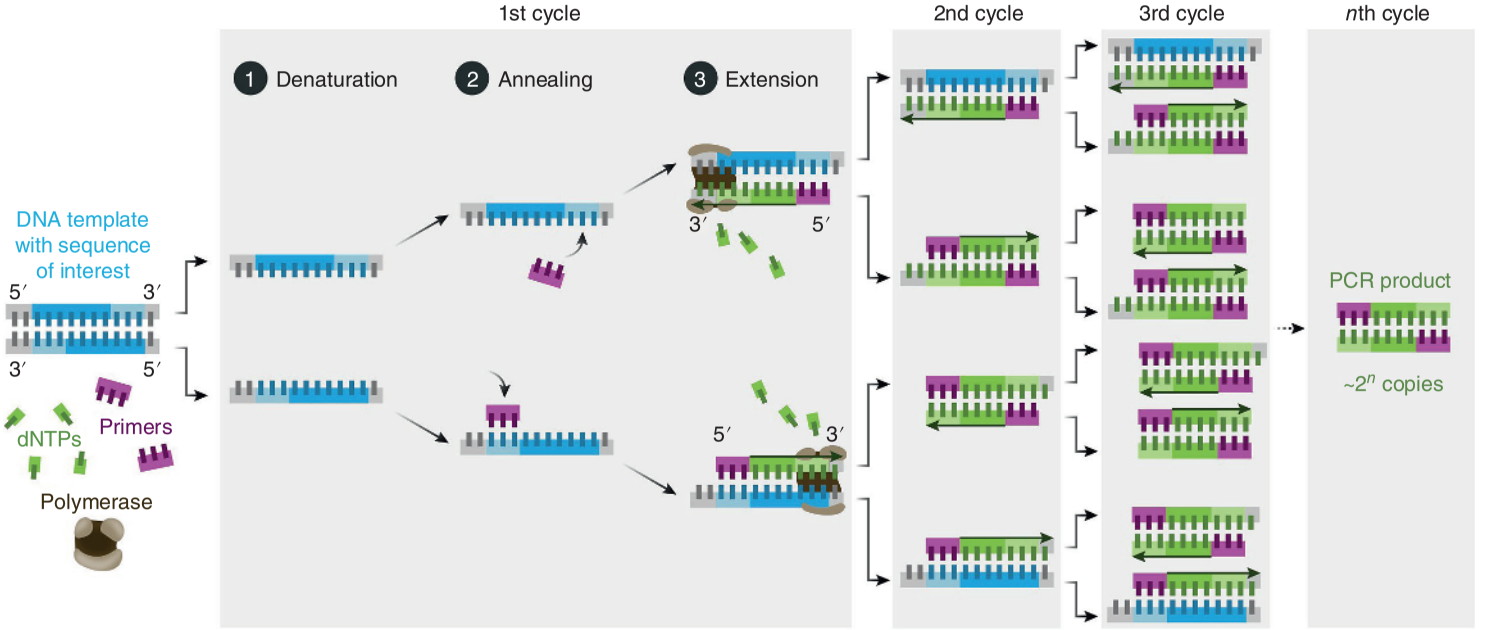

Figure 1.19 Overview of the polymerase chain reaction (PCR) which repeatedly cycles between three steps: 1) denaturation at ~94–98 °C, which breaks the hydrogen bonds between complementary base pairs and denatures double-stranded DNA into single-stranded DNA. 2) annealing (binding) of primers to complementary sequences on the single-stranded DNA template, and 3) extension of new DNA sequences in which DNA polymerase uses the primers as starting points and the denatured DNA as a template to which complementary nucleotides (dNTPs) are added to the 3′ end of the primers, extending the DNA strands in a 5′ to 3′ direction by synthesizing new strands of DNA. PCR normally comprises 25–40 cycles of denature–anneal–extend. Each cycle generates two daughter strands for every parent strand, which means that the number of sequences increases exponentially throughout the PCR reaction. *Source:* Enzoklop/Wikimedia commons/CC BY SA.

Box 1.5 Sources of DNA

DNA can be characterized from numerous different types of material thanks to PCR, which makes many copies of a very small amount of starting material. In molecular ecology, tissue samples collected from natural populations for the purpose of DNA extraction and amplification have included:

- Most vertebrates (one exception to this being mammals) have nucleated red blood cells (erythrocytes), which leads to an abundance of DNA in even a small volume of blood. Blood samples, e.g. collected via venipuncture, have therefore been frequently used as a DNA source from nonmammalian vertebrates.
- Swabs are often used to collect epithelial cells from small fish and freshwater mussels.
- Feathers are a noninvasive source of bird DNA.
- Hair samples can be noninvasively sampled from mammals e.g. by collecting naturally shed hair or from noninvasive hair traps.
- Fecal samples are another noninvasive source of DNA for many species, including mammals, birds, and reptiles. DNA extracted from feces includes donor, dietary (e.g. prey), and microbial DNA.
- Eggshell membranes are a potential source of DNA for birds and reptiles.
- DNA can be extracted from various plant tissues including leaves, stems, flowers, seeds, and roots.

DNA can also be extracted from museum and herbarium samples, which can be a valuable source of DNA for projects that wish to compare historical and current samples or even characterize the DNA of now-extinct species (Figure 1.20). DNA can be extracted from these specimens using specialized protocols optimized for degraded DNA. Again, a wide range of sample types can be used, including:

- Tissues such as muscle, skin, feathers, fur, or insect exoskeletons from taxidermy specimens.
- Dried or pressed plant parts such as leaves, flowers, stems, seeds, or fruits from herbarium specimens.
- Specimens preserved in alcohol or formalin such as fish, amphibians, reptiles, or invertebrates.
- Bones, teeth, or hair from skeletal specimens.

Finally, because organisms typically leave traces of themselves in their environments, DNA can also be extracted in the form of **eDNA** based on a wide range of sampling strategies, including:

- Water samples collected from lakes, rivers, streams, ponds, wetlands, estuaries, or oceans, which often yield eDNA from aquatic organisms including fish, amphibians, invertebrates, and plants.
- Soil samples often harbor eDNA from soil-dwelling organisms including bacteria, fungi, nematodes, arthropods, and plant roots.
- Benthic samples collected from the bottom of aquatic environments (e.g. mud, sand, and rocks) can include eDNA from benthic organisms such as macroinvertebrates, mollusks, or algae, and also from species that occur in the water column.

Figure 1.20 The greater one-horned rhinoceros (*Rhinoceros unicornis*). Liu et al. (2021) used a combination of historical museum samples and current-day samples from all five extant and three extinct rhinoceros species to show that while low genetic diversity is a long-term feature of the rhinoceros family, recent population declines resulting from human activity mean that today's populations have particularly low genetic diversity. *Source:* Prasan Shrestha/Wikimedia Commons/ CC BY-SA 4.0/ https://commons.wikimedia.org/wiki/File:The_Greater_one_horned_Rhinoceros_1.jpg / last accessed on October 15, 2025.

- Air samples can include eDNA from airborne biological material such as pollen, spores, fungal fragments, and microbial cells, and also from terrestrial and even aquatic organisms.
- Spider webs can capture a variety of eDNA sources including pollen, insects, fungal spores, and vertebrates.
- Leaf swabs can capture eDNA from a wide range of taxa that encountered those leaves including pollinators and herbivores.

eDNA will be discussed in detail in Chapter 2.

1.10.3 Quantitative PCR

All types of PCR generate multiple copies of the template DNA sequence(s), but a specific form of PCR known as **quantitative PCR** (**qPCR**, also known as **real-time PCR** or **RT-PCR**) also quantifies the starting amount of DNA in a reaction; in other words, it can tell you how much DNA is present in a particular sample. qPCR includes a fluorescent dye or probe in the reaction mixture that binds to the newly synthesized DNA strands, and as the amount of DNA increases during each PCR cycle, the strength of the fluorescence signal also increases at a rate that is proportional to the amount of starting DNA. This is because an excess of primers, polymerase, and dNTPs is added to the qPCR reactions, and therefore the amount of starting template DNA is the only factor limiting the number of DNA copies that are made in each cycle. The qPCR machine measures the amount of fluorescence emitted during each cycle of

amplification, and the starting amount of target DNA can be quantified by comparing the fluorescence to standard curves that are generated from known concentrations of DNA; this is referred to as absolute quantification. Relative quantification, on the other hand, does not directly quantify the amount of starting DNA but does allow users to determine the relative amount of starting DNA, for example, to conclude that sample A has more DNA than sample B.

qPCR with fluorescent probes is often used to characterize eDNA (Box 1.5). Because eDNA is extracted from an environmental sample such as soil or water, it is a composite of DNA from many different taxa including those that lived in, or passed through, that environment. Fluorescent probes in qPCR are often used to increase PCR specificity, in other words, to increase the likelihood that the PCR is amplifying DNA from the species of interest. This is because the probes anneal to a region of DNA in between the sequences to which the primers anneal, and if a probe is designed to anneal to a DNA sequence that is found only in the target species, no other DNA will be reflected in the fluorescent signal even if the primers anneal to nonspecific targets. Because eDNA is often highly degraded (fragmented), primers are typically designed to amplify relatively short fragments (commonly 50–400 bp), and the fluorescent probe is usually between 20 and 30 nucleotides long. Harper et al. (2018) added a probe to qPCR to increase rigor when amplifying an 81 bp fragment of the cytochrome b gene (a gene in the mtDNA genome) from eDNA that originated from the rare great crested newt *Triturus cristatus*. eDNA was extracted from water samples collected from multiple ponds, and qPCR-based detection of great crested newt eDNA from 50% of the sampled ponds identified sites that should be protected as great crested newt habitat (Harper et al. 2018).

Another application of qPCR is the quantification of gene expression. Provided that tissue samples are stored appropriately (e.g. by flash freezing or storing within a stabilizing buffer that protects RNA from degradation), researchers can extract RNA directly from samples using methods similar to those for DNA. The RNA can then be reverse-transcribed to make **complementary DNA (cDNA)**, which is a synthesized DNA molecule that is complementary to a specific RNA sequence. qPCR can then be used to quantify cDNA (and indirectly RNA), and this tells us whether – and to what extent – one or more genes were expressed in the tissue at the time of sampling. This approach was used as part of an investigation into the host–parasite interaction between honey bees (*Apis mellifera*) and the mite *Varroa destructor*. This is a relatively new interaction because *Varroa* switched host to parasitize *A. mellifera* only a few decades ago, and *A. mellifera* has so far evolved little resistance to this parasite. Multiple breeding programs have now selected for *Varroa*-resistant honey bee populations, with resistance in bees linked to two genes from the *ecdysone* biosynthesis pathway. This pathway describes a series of biochemical reactions in insects and other arthropods that synthesize products including hormones that are essential for the regulation of developmental processes such as molting, metamorphosis, and reproduction. qPCR of cDNA showed that in mite-resistant populations, the *ecdysone*-linked genes were upregulated (an increase in gene expression) in young honey bee pupae, a developmental period that is particularly susceptible to *Varroa* infestations. The authors of this study therefore concluded that upregulation of *ecdysone*-linked genes is a critical part of *A. mellifera* resistance to *varroa* mites (Conlon et al. 2019) (Figure 1.21).

Figure 1.21 Two honey bee (*Apis mellifera*) drone pupae (juvenile worker bees) with varroa mites (*Varroa destructor*). While many honey bee populations have no resistance to varroa mites, upregulation of key genes at an early stage of bee development in some populations can help them to resist the mites (Conlon et al. 2019). *Source:* Waugsberg/Wikimedia Commons/CC BY-SA 3.0/ https://commons.wikimedia.org/wiki/File:Drohnenpuppen_mit_Varroamilben_71a.jpg / last accessed on October 15, 2025.

1.11 Turning PCR into Data

The goal of PCR is to amplify one or more fragments of DNA so that there is sufficient DNA to allow for characterization. The next step is to characterize the DNA based on either the size or the DNA sequence of the amplified fragments.

1.11.1 Fragment Sizes

The easiest way to get information from amplified DNA (i.e. the post-PCR product) is to visualize the fragment on an electrophoresis gel, which allows us to estimate the size of the fragment. Gel electrophoresis is a technique that separates biological molecules – DNA, RNA, or proteins – based on their size. Electrophoresis gels most commonly use agarose as the gel, which is made by adding agarose powder to buffer and heating it until the powder dissolves. When partly cooled, the solution is poured into a tray that is surrounded by removable walls, and removable combs are added to create wells in the gel. Once the gel solidifies, the combs and walls are removed, and the gel is covered in buffer. A sample that includes some form of biological molecule (e.g. extracted DNA or PCR product) plus loading dye (for color and weight) is loaded into each well. An electrical field is then applied through the buffer, and because DNA molecules are negatively charged, they migrate toward the positive electrode at a rate that is correlated with their size, meaning that smaller molecules migrate through the gel more rapidly than larger molecules (Figure 1.22). Once DNA fragments have moved through the gel, they can be visualized using a dye such as ethidium bromide or, more commonly these days, a nontoxic fluorescent dye that binds to DNA molecules. Agarose gels have

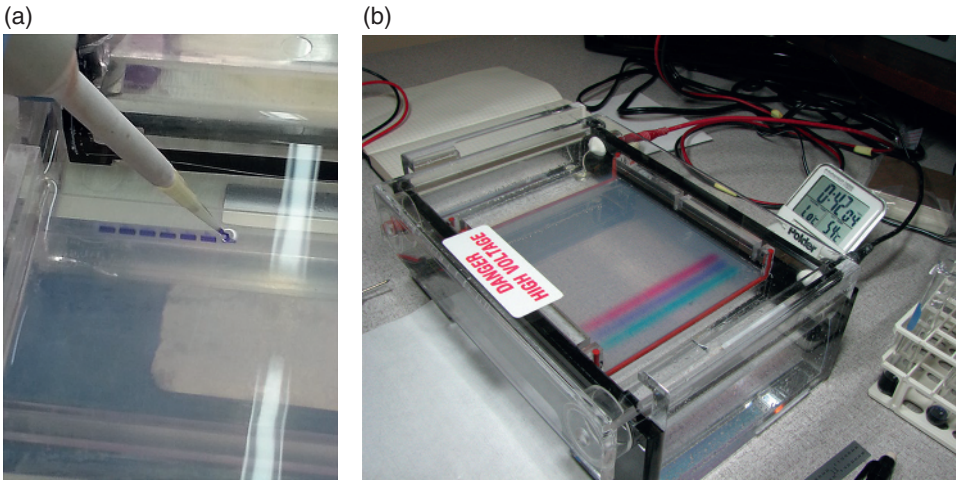

Figure 1.22 A standard agarose gel set-up used to visualize DNA fragments and estimate their sizes. (a) Here, the gel has solidified, and DNA is being loaded into the wells that are left behind once the combs have been removed and the gel has been submerged in buffer. The samples being loaded into the wells appear purple because of the loading dye that adds weight and color to the sample. (b) Here, the DNA has been loaded into wells, and an electrical charge has been added. DNA is negative and therefore migrates toward the red (positive) electrode. The colored stripes in this photo are created by the loading buffer as the DNA molecules migrate through the gel. *Source:* (a) Isabella Apriyana / Wikimedia Commons / CC BY 4.0 / https://commons.wikimedia.org/wiki/File:Loading_DNA_sample_into_to_agarose_gel_electrophoresis.jpg / last accessed on October 15, 2025; (b) Michael / Wikimedia Commons / CC BY 2.0 / https://commons.wikimedia.org/wiki/File:Electrophoresis_-_Moving_along_gel.jpg / last accessed on October 15, 2025.

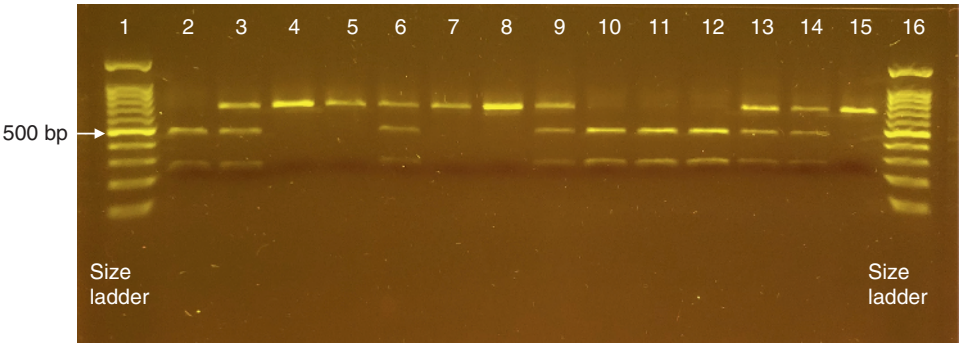

Figure 1.23 Gel electrophoresis image showing how the sizes of DNA fragments can be estimated by comparing them to ladders with known band sizes. Bands in this gel are visible because a fluorescent dye that binds to DNA was added to the gel. In this example, ladders that were loaded into lanes 1 and 16 show bands that increase in size by increments of 100 bp. PCR products from 14 different individuals were loaded into the remaining wells. The sample loaded into lane 2 has 2 bands of approximately 250 and 500 bp; the sample loaded into lane 3 has 3 bands of approximately 250, 500, and 750 bp; the sample in lane 4 has 1 band of approximately 750 bp; and so on.

multiple applications including determining whether a PCR reaction was successful, in which case one or more bands of the appropriate size are visible on the gel image. The size(s) of band(s) is estimated by comparing to a ladder with bands of known sizes (Figure 1.23).

1.11.2 Microsatellites

Microsatellite markers, also known as **simple sequence repeats** (**SSRs**) or **short tandem repeats** (**STRs**), are sequences of DNA in which a short (usually 1–6 bp) motif is repeated multiple times; for example, ATATATATATATATATATAT is a microsatellite with the motif AT that is repeated 10 times and can be written as $(AT)_{10}$. Microsatellites are prevalent in prokaryotes and eukaryotes, and while they are found in chloroplast and mitochondrial genomes, they are most commonly characterized from nuclear genomes. The alleles at a given microsatellite locus are normally identified from their sizes, which are primarily determined by the number of repeats in each allele (Figure 1.24). The microsatellite alleles from a particular locus are PCR-amplified using a pair of primers that anneal to the sequences that flank the repeats, and therefore the amplified product comprises the repeat motif plus flanking sequences. Note that microsatellite data sets typically include data from multiple microsatellite loci, but the alleles from each locus are amplified using primers that are specific to that locus. Microsatellite data are considered **codominant**, which means that in a diploid species, we can identify both alleles at each locus: these can be either two different alleles (**heterozygous**) or two copies of the same allele (**homozygous**). The agarose gels described earlier give only approximate sizes of amplified bands, which is insufficient to differentiate microsatellite alleles that may differ by only 1–2 bp. Instead, microsatellite allele sizes are determined using fluorescently labelled primers during PCR, followed by capillary electrophoresis that has lasers to detect size-separated DNA fragments.

Microsatellites are useful genetic tools because of their high mutation rates within the repeat motifs. These mutations often result from slipped-strand mispairing during DNA replication, which occurs when DNA polymerase temporarily disassociates from the template strand ("slips") and then reassociates in a slightly shifted position ("mispairing"), leading to either the gain or loss of a repeat. Nucleotide insertions or deletions in the flanking (non-repetitive) sequence are a less common reason for size differences between microsatellite alleles. Estimations of microsatellite mutation rates

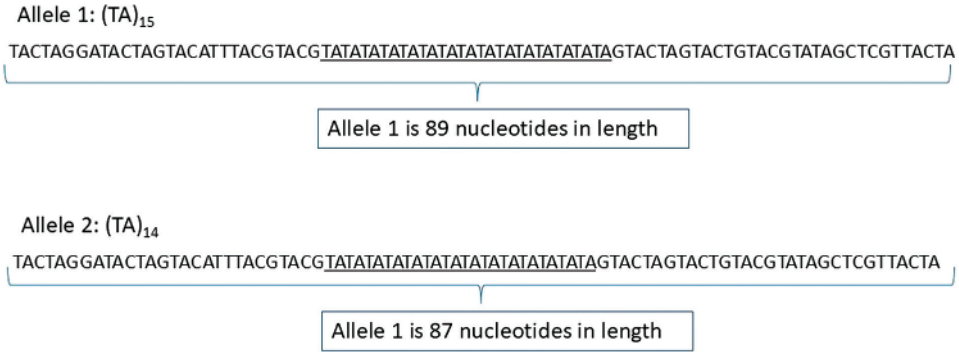

Figure 1.24 Microsatellite alleles are normally described on the basis of the allele length (the number of nucleotides). Primers anneal to sequences that flank the microsatellite motif, and the size of the amplified fragment most commonly depends on the number of repeats that are found within the repetitive (microsatellite) sequence. In this example, only one strand of DNA is shown for the sake of simplicity. The two alleles from the same locus have identical flanking sequences, but allele 1 has 15 repeats of TA, whereas allele 2 has only 14 repeats. As a result, allele 1 is 2 bp longer than allele 2. The primers that are used to amplify microsatellite alleles are specific to a single locus (i.e. they will anneal only to the sequences flanking that particular microsatellite).

range from approximately 10^{-6} (1 mutation event per 1,000,000 loci per generation) to 10^{-2} (1 mutation event per 100 loci per generation) across a wide range of taxa (see Bhargava and Fuentes 2010). This is much higher than a commonly cited typical mutation rate in eukaryotic coding DNA (10^{-9} per nucleotide per generation) (Ellegren 2000) although this too varies considerably, for example, a study of 68 species of mammals, fishes, birds, and reptiles found that there was up to a 40-fold difference in the per-generation mutation rate among species (Bergeron et al. 2023). Mutation rates also vary among microsatellite loci depending on factors that include the type of microsatellite motif, number of repeats, whether the microsatellite repeat is perfect or imperfect (imperfect repeats are interrupted, e.g. CACACATAGCACACA, in which the underlined TAG interrupts the CA repeats), location on the genome, taxonomic group, and even the sex of the individual. Nevertheless, microsatellites overall have much higher mutation rates than most other gene regions.

Microsatellite data are not particularly useful for inferring evolutionary events that occurred in the relatively distant past. Their rapid rate of mutation and their tendency to either increase or decrease in size (i.e. to either gain or lose a repeat) means that size **homoplasy** can occur. This can be illustrated by an example of 2 ancestral alleles at the same locus, one with 20 dinucleotide repeats (e.g. 20 repeats of AT), and the other with 18 dinucleotide repeats. If the larger allele loses 1 repeat and the smaller allele gains 1 repeat, then both mutations will result in alleles with 19 repeats. These 2 new alleles, each with 19 repeats, may appear to be two copies of the same ancestral sequence, but the evolutionary histories of the 2 alleles are in fact quite different. Size homoplasy means that ancestor–descendant relationships may be difficult to untangle from microsatellite data. On the other hand, the high mutation rates mean that microsatellite loci are often highly **polymorphic**; in other words, within populations and species there can be many different alleles at a given locus, which makes them useful for characterizing genetic diversity and relatedness.

1.11.3 Dideoxy Sequencing

As described earlier, microsatellite alleles are differentiated from one another based on the sizes of the amplified alleles. More precise genetic information can be obtained from DNA sequences. DNA sequencing is exactly what it sounds like: determining the sequence of nucleotides along a fragment of DNA. The first widespread method of DNA sequencing, known as dideoxy sequencing or Sanger sequencing, was invented by Frederick Sanger in the mid-1970s and led to his shared Nobel prize in 1980. This method involves synthesizing multiple copies of a strand of DNA using DNA polymerase plus nucleotides in a manner similar to that employed during PCR, but with two important differences: 1) only one primer is used as the starting point for synthesis (compared to the two primers that are used in PCR) so that the template DNA is copied in only one direction, and 2) to the reaction is added a proportion of each nucleotide that contain the sugar dideoxyribose (ddNTPs – in which "N" refers to G, A, T, or C) instead of deoxyribose (dNTPs, which are "normal" nucleotides). Dideoxyribose lacks a 3'-OH group, and this prevents the next nucleotide from being added to the growing DNA strand. This means that whenever a nucleotide with dideoxyribose (a ddNTP) is incorporated into the reaction, DNA synthesis is terminated. Each sequencing reaction occurs many times, and a correct ratio of dNTPs to ddNTPs will ensure that during the

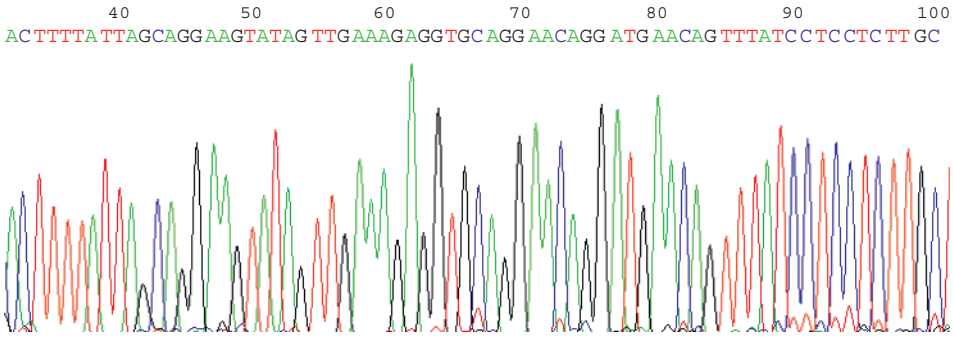

Figure 1.25 Part of a sequence electropherogram showing how each peak is assigned to a nucleotide based on its color: red = "T," blue = "C," black = "G," green = "A." The row of text at the top of the figure corresponds to the colored peaks below the sequence; in this case, the sequence would be read as ACTTTTATTAGC and so on.

reaction the template DNA will incorporate a ddNTP at every single site along the DNA sequence, resulting in every possible fragment size ranging from 1 bp of the target sequence to its maximum length.

The nucleotides that contain dideoxyribose (the ddNTPs) are labeled with different colored fluorescent dyes. When the sequencing reactions are run through 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. Imagine a sequence of 100 nucleotides. If the first nucleotide is an "A," one of the terminated reactions will have occurred when a ddATP was incorporated as the initial nucleotide, creating a fragment of 1 bp in length with a green peak. If the second nucleotide is a "C," one of the terminated reactions will have occurred when a ddCTP was incorporated as the second nucleotide, creating a fragment 2 bp in length with a blue peak. And so on. By substituting the appropriate base for each colored peak, the entire sequence can be read from a single image that is known as an electropherogram (Figure 1.25).

1.11.4 Single-Nucleotide Polymorphisms

Once DNA sequences have been generated, a common next step is to align sequences to see how similar they are among individuals, populations, or species. When comparisons are made among relatively closely related individuals (e.g. members of the same species or genus), many of the nucleotides will be the same across all compared sequences. Researchers typically focus on the positions that show variation, and these include **single-nucleotide polymorphisms**, or **SNPs** (pronounced snips). Most SNPs used in molecular ecology are biallelic markers, which means that they have two alternative states within the target species or population (Figure 1.26). When identifying SNPs within populations or species, researchers typically select only the ones at which the least common nucleotide is present in a minimum number of individuals (e.g. 5% or 10%). In other words, to be considered a SNP (as opposed to a rare mutation), each of the two allelic versions of a SNP (e.g. the T and the G, or the C and the A from Figure 1.26)

```
                  *                           *
Sequence 1: GATTATACGTATATGCCGGTATACGCGCGCCTTAACATAGTACGGTATAGCTCTGGTATT
Sequence 2: GATTATACGTATATGCCGGTATACGCGCGCCTTAACATAGTACGGTATAGCTCTGGTATT
Sequence 3: GATTAGACGTATATGCCGGTATACGCGCGCATTAACATAGTACGGTATAGCTCTGGTATT
Sequence 4: GATTATACGTATATGCCGGTATACGCGCGCATTAACATAGTACGGTATAGCTCTGGTATT
Sequence 5: GATTAGACGTATATGCCGGTATACGCGCGCATTAACATAGTACGGTATAGCTCTGGTATT
```

Figure 1.26 Alignment of a hypothetical 60 bp sequence from five different individuals of the same species. This alignment shows that most nucleotide positions are invariant (i.e. all individuals have the same nucleotide). However, there are two sites shown in bold that have two alternative nucleotides: at the first variable site some individuals have a "G" whereas others have a "T," and at the second site some individuals have an "A" and others have a "C." These two sites are biallelic (two allelic forms) single-nucleotide polymorphisms (SNPs).

must be present in at least 5% (or 10%) of individuals within the study population or species. SNPs are often identified from sequence data spanning relatively large sections of the genome, which is accomplished using high-throughput sequencing (HTS).

1.11.5 High-Throughput Sequencing

When SNPs were first used in studies of molecular ecology, the process for identifying them was rather laborious, but our ability to generate large amounts of sequence data – including SNPs – changed dramatically with the advent of HTS, also known as next-generation sequencing (NGS). HTS is a comprehensive term describing multiple technologies that generate large amounts of sequence DNA from across the genome in an increasingly rapid and cost-effective manner. A first step in HTS is the preparation of a library of samples compatible with the sequencing technology being used. There are many different methods for preparing libraries, but a common approach is to first fragment the DNA into pieces and then ligate (join) adapters (fragments of DNA with known sequences) to the ends of the fragments. In some methods, the next step is clonal amplification which generates clusters of identical DNA fragments to increase sensitivity and allow for parallel sequencing of multiple DNA fragments. To save resources, multiple libraries (e.g. libraries from multiple individuals or species) can be pooled together and sequenced in the same run – a process known as multiplexing. This is possible if index sequences, also known as "barcodes," that are unique to each individual, population, or species, are added along with the adapters. Bioinformatics broadly describes the computer-based analyses of large amounts of sequence data, and HTS requires considerable bioinformatics analyses to group sequences according to barcodes, assess the quality of sequences, align sequences, and concatenate (join) sequences. Although this takes much longer – and requires greater expertise – compared to identifying the sizes of microsatellite alleles or recording a short DNA sequence from an electropherogram, far more data are generated.

Technologies for NGS are constantly evolving (Figure 1.27), and the best approach depends on time, money, resources (e.g. which sequencing platforms are available), expertise, and the questions being asked. For example, larger numbers of short sequences (typically <300 bp) can normally be generated more quickly, more cheaply, and with fewer errors than long sequences (up to hundreds of thousands of bases), but it may be difficult to accurately align short sequences, particularly in regions of repetitive DNA. Technologies for generating longer sequences are constantly improving, which in turn

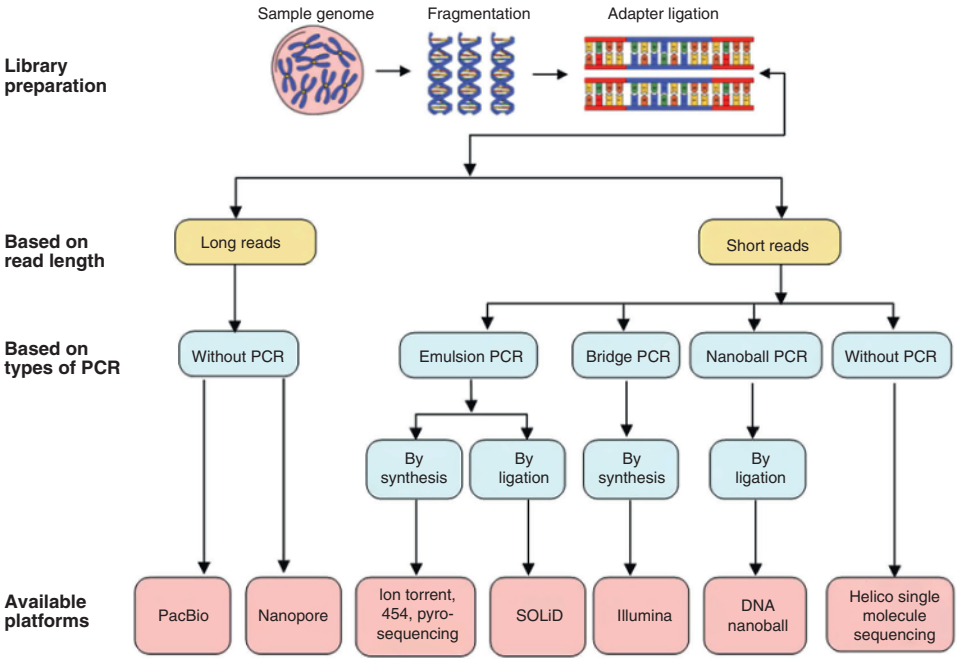

Figure 1.27 Overview of some HTS technologies that employ different platforms and principles. Methods are constantly evolving, with new approaches and equipment becoming available every year. *Source:* Satam et al. (2023)/MDPI/CC BY 4.0.

enhances our ability to align complex sequences such as lengthy SVs (Section 1.6). HTS can be used to sequence an entire genome (whole-genome sequencing), or to generate reduced-representation sequences; each of these is briefly discussed below, but keep in mind that technologies are continually updating, for example, see Lou et al. (2021), Bohmann et al. (2022), Hakimzadeh et al. (2024), Theissinger et al. (2023), and Uhlen and Quake (2023).

Reduced-representation sequencing generates sequence data from a subset of the genome by sequencing libraries that comprise a random pool of fragments from across the genome. Restriction-site associated DNA sequencing (RADseq) (Baird et al. 2008) and double-digest RAD sequencing (ddRAD) (Peterson et al. 2012) are two commonly used methods for reduced-representation sequencing. RADseq was the original method. This begins with the preparation of a library that uses a **restriction enzyme (RE)** to fragment the DNA samples. REs digest ("cut") DNA at specific recognition sequences; for example, EcoRI is an RE that recognizes the sequence GAATTC, and then cuts DNA between the G and the A nucleotides, leaving DNA fragments with overhanging single-stranded ends (so-called sticky ends; Figure 1.28). The sticky ends allow for the attachment (ligation) of adapters to the end of each fragment because the ends of these adapters are sequences that are complementary to the overhanging ends created by the enzymes. The adapters extend beyond the sticky ends: they have additional sequence to which the sequencing primers will anneal, plus a short barcode (typically 4–10 bp long depending on the number of groups) which can be unique to each designated group (e.g. individual or population) and used later for sample

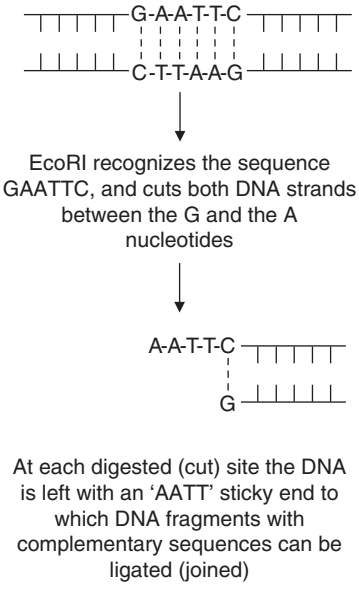

Figure 1.28 Some restriction enzymes, including EcoRI, digest DNA in a manner that leads to a "sticky end" (as opposed to a blunt end, which lacks overhanging, single-stranded DNA). The recognition site for EcoRI is GAATTC; the enzyme digests (cuts) DNA between G and A within this recognition sequence, and the digested (cut) DNA is left with an overhang of AATT.

identification. The fragments with their ligated adapters and unique barcodes can then be pooled and run out on an agarose gel using gel electrophoresis (Section 1.11.1), and a subset of fragments that typically falls within the 300–800 bp size range is selected by excising this portion of the gel and using a kit to isolate DNA from the agarose gel. Many protocols then use a second adapter which has divergent ends with two strands that are complementary for only part of their length. The reverse amplification primer cannot bind to the second adapter unless the first round of sequencing is successful, thus removing from the reaction any fragments that lack the forward adapter and associated barcode (Figure 1.29). The ddRAD sequencing protocol is similar, but as the name implies, it uses two REs. The first RE targets a common recognition sequence to which the barcoding primers are ligated. The second RE targets a rare recognition sequence and is used to improve the size selection step before the second set of adapters is ligated.

Returned sequences must first be demultiplexed, a step that assigns each fragment to the sample from which it originated based on its unique barcode sequences. After various steps of quality control, sequences can then be either aligned to a reference genome if one is available, or can be assembled based on overlapping sequences (Figure 1.30). Although reduced-representation sequencing generates data from only a subset of the genome, the sequences are for the most part randomly distributed across the genome and can be generated at a fraction of the cost of whole-genome sequencing, particularly in organisms with large and complex genomes. There have been some suggestions that ddRAD is a more precise method than RADseq (e.g. Ulaszewski et al. (2021)), and it may be increasing in popularity. However, it seems likely that new developments that either refine existing methods or introduce novel approaches will emerge in the near future.

Whole-genome sequencing (**WGS**) results in the DNA sequence of an entire genome; the default for WGS is normally the nuclear genome, whereas the terms whole mitochondrial or whole chloroplast genome sequences are normally used to specify

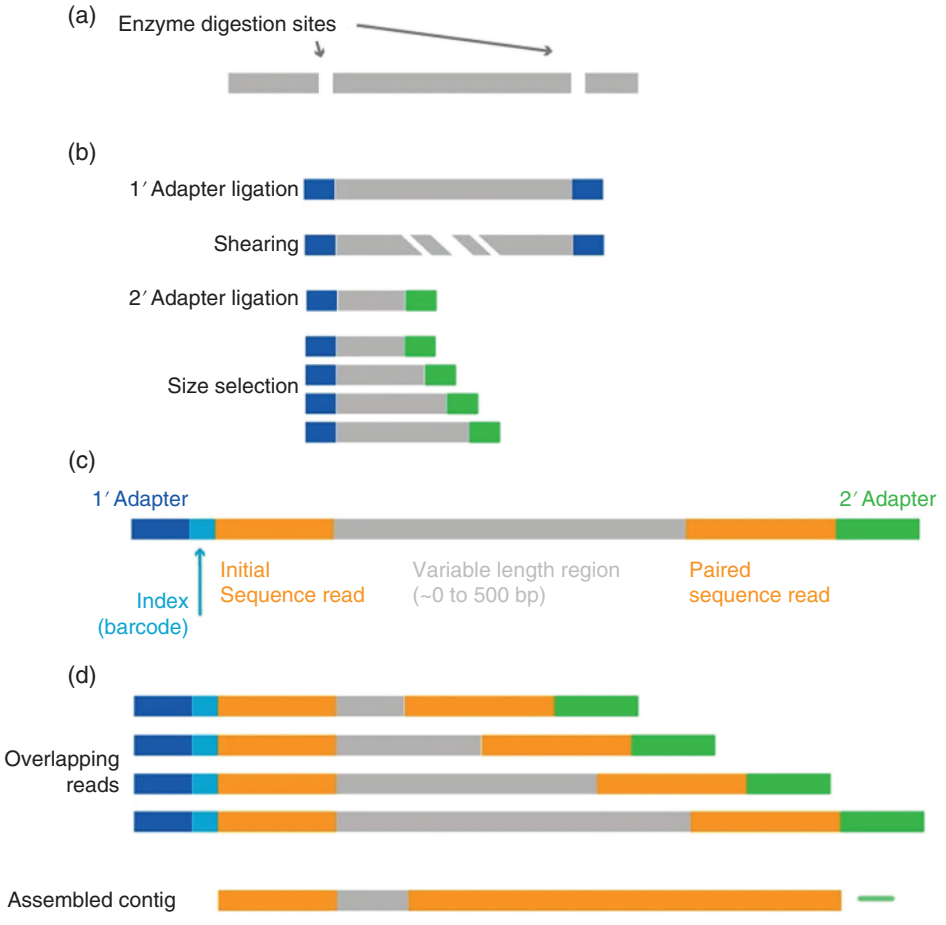

Figure 1.29 Paired-end RAD Sequencing Overview. (a) Genomic DNA is digested with a restriction endonuclease. (b) After ligation with a primary adapter, the fragments are sheared, then ligated with a secondary adapter. (c) A composite mixture of variable-length fragments is recovered from each RE digestion site. These fragments are size-selected, amplified, and sequenced on a next-generation DNA sequencing platform. (d) Genomic assembly is then completed bioinformatically. *Source:* Pegadaraju et al. (2013)/Springer Nature/CC BY 4.0.

```
                                                    TACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG
                                          AGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAG
GTATATGGGCTATAGCCGATACGTACG
      ATGGGCTATAGCCGATACGTACGGGGTATATAT
GTATATGGGCTATAGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG
        TGGGCTATAGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG
                                                GGGTATATATGGGCGCGTTTTCAGAGATGGATATG
                                                      ATATGGGCGCGTTTTCAGAGATGGATATG
GTATATGG
GTATATGGGCTATAGCCGA
GTATATGGGCTATAGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG
GTATATGGGCTATAGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG

GTATATGGGCTATAGCCGATACGTACGGGGTATATATGGGCGCGTTTTCAGAGATGGATATG
```

Figure 1.30 Representation showing how multiple reads that partially overlap are used to construct **contigs** (contiguous sequences). In this example, the contig is the underlined sequence which was assembled through a comparison of 12 partially overlapping sequences.

organellar genomes. WGS can be accomplished by fragmenting the entire genome into small pieces, then ligating adapters to the ends of those pieces. Unlike reduced-representation sequencing, the library includes fragments that collectively represent the entire genome. As with reduced-representation sequencing, the adapters contain the sequences to which the sequencing primers anneal, and when multiple samples are pooled (e.g. multiple species or multiple individuals), the adapters also contain individual barcodes so that sequences can be assigned to the correct individual. Sequences generated by whole-genome sequencing include multiple reads of each fragment, which are assembled into contigs that represent the consensus sequence for each fragment. One common trade-off in WGS is the number of samples versus the **depth**, which refers to the average number of reads per DNA fragment (e.g. an average depth of 6× means that each nucleotide was sequenced an average of six times). Higher coverage normally means less error; for example, if the depth is only 1× and there has been a sequencing error, that may not be detected; in contrast, if the depth is 6× and the error occurred in only one of the sequences, the majority sequence (i.e. that occurring in the remaining five sequences) will be considered correct and used in the alignment. While increasingly time- and cost-effective, WGS remains challenging for many labs, largely because of the complex bioinformatics analysis; this is particularly true for very large genomes, or if there are no existing reference genomes to which fragments can be compared. Nevertheless, the number of individuals for which we have complete genome sequences is rapidly increasing, and will continue to do so: to give just one example of a large-scale project, the Darwin Tree of Life Project (https://www.darwintreeoflife.org/) aims to sequence whole genomes from 70,000 species of eukaryotic organisms in Britain and Ireland.

1.12 Overview of Genetic Data

In molecular ecology, the term "molecular marker" is often used to describe the types of genetic data that are being generated, for example, nuclear microsatellites or mitochondrial SNPs. Given the wide range of molecular markers, their tremendous variation in terms of information provided, and their required costs, time commitment, equipment, and analytical skills, there are clearly many considerations before researchers decide on a protocol for generating genetic data from their samples. Short DNA sequences and microsatellite genotypes are relatively fast and easy to generate. Genome-wide SNPs, often identified through reduced-representation sequencing, are more time-consuming, and generating WGS is even more of a time investment. Both reduced-representation sequencing and whole-genome sequencing require specialist equipment that is not available to all labs, along with considerable bioinformatics skills and computer servers that can store large amounts of data and run complex analyses. However, maximum genetic information is obtained from WGS, including SNPs, gene sequences, SVs, etc.

The choice of marker also depends on the question that is being asked. Short DNA sequences are often used in taxonomic studies and to assay eDNA and will be discussed further in Chapter 2. Other common goals in molecular ecology include quantifying the genetic diversity of populations (Chapter 4) or describing the extent to which populations are genetically differentiated from one another (Chapter 5). Both of these goals can be achieved to varying degrees using either microsatellite loci (typically ~8–20 loci)

or SNPs (the number of these can vary from dozens to hundreds of thousands). Microsatellites were more commonly used than SNPs until a few years ago, but now SNPs are becoming increasingly popular, partly because SNP genotypes are highly repeatable and can be easily compared among studies. In contrast, microsatellite allele sizes often need to be recalibrated when the same loci are amplified in different labs, and as discussed earlier, microsatellites also have a higher tendency for homoplasy. However, SNPs are less variable than microsatellites. Consider a population of diploid birds. Each bird will have two alleles at each locus (either heterozygous or homozygous), which means that the variability of SNPs and microsatellites is the same within *individuals* (a maximum of two alleles). At the population level, however, the variability at a microsatellite locus will typically be much higher than the variability of a SNP locus. This is because most SNPs are biallelic (a maximum of two alleles), whereas the high mutation rate of microsatellites means that there can be a lot more than two alleles at the *population* level. For this reason, most studies based on SNPs need to describe genotypes based on a larger number of loci compared to studies that are based on microsatellites. In some cases, comparable information from microsatellites and SNPs can be obtained when investigating genetic diversity within populations and genetic similarity among populations, although in other cases, SNPs are more informative. For example:

- Paternity assignments are based on comparisons between the genetic profiles of offspring and their potential fathers (Chapter 7). Kaiser et al. (2017) found that in a black-throated blue warbler (*Setophaga caerulescens*) population, genetic fathers could be identified with equal precision when using either 6 microsatellite loci or 97 SNPs.
- A study of the fish species *Esox lucius* (pike) based on 10 microsatellites versus 1580 SNPs found overall higher estimates of genetic diversity based on the microsatellites, and comparable estimates of genetic differences among populations based on the two sets of markers (Sunde et al. 2020).
- An F1 (first generation) hybrid between two polecat species – *Mustela eversmanii* and *M. putorius* – could be identified from either 8 microsatellites or 164,644 SNPs, although the latter were more informative when looking for evidence of backcrossing (hybrids interbreeding with parent species) or advanced-generation hybrids (which result from interbreeding between hybrids) (Szatmári et al. 2021).
- Genetic structure and diversity in two amphibian species endemic to the Iberian Peninsula – *Hyla molleri* and *Pelobates cultripes* – were compared based on microsatellites (18 and 14 loci) versus SNPs (15,412 and 33,140 loci), and the authors concluded that the SNP dataset more accurately described genetic diversity and genetic similarities among populations compared to the microsatellite dataset (Camacho-Sanchez et al. 2020).
- Estimates of genetic diversity within brown trout (*Salmo trutta*) populations, and genetic differences among populations, were comparable based on 16 microsatellites versus 4876 SNPs; however, the latter were more informative for measures of individual heterozygosity, and for estimating the relatedness among individuals (Lemopoulos et al. 2019).
- Studies that look for adaptive genes (those that confer fitness advantages under particular environmental conditions) must almost inevitably survey substantial areas of the genome which cannot be achieved with microsatellites but can be done using

large numbers of SNPs. For example, one study that surveyed 241,371 SNPs across populations of white clover (*Trifolium repens*) found 213 SNPs that were embedded within potentially adaptive genes (Kuo et al. 2024).

On balance, considerably more information can be obtained from genome-wide SNPs compared to microsatellites, although microsatellites remain a more realistic option in some research labs and are sufficient for addressing some questions in molecular ecology.

Of course, maximum information comes from WGS even if only a few individuals are sequenced, provided researchers have the necessary time, money, equipment, and bioinformatic skills for this approach. For example:

- WGS from two brown trout (*S. trutta*) from each of eight freshwater lakes in Finland, and from each of two sampling periods (1970s versus current populations), were sufficient to investigate how genetic diversity, inbreeding (mating between relatives), and natural selection had changed over time (Kurland et al. 2024).
- WGS of 24 Komodo dragon (*Varanus komodoensis*) from the 5 main Indonesian islands determined which groups had enough genetic novelty to merit individual conservation plans, and provided detailed information about the link between historical population sizes and current genetic diversity (Iannucci et al. 2021).
- WGS of 129 gentoo penguins (*Pygoscelis papua*) from 12 colonies located at or near the most southern part of the species' geographical range allowed researchers to determine that these penguins historically expanded their range by dispersing rapidly in a stepping-stone pattern while maintaining genetic diversity (Herman et al. 2024).
- WGS of giant kelp (*Macrocystis pyrifera*) from 18 populations in Chile and California confirmed the genetic distinctiveness of 2 different types that occupy different niches and identified 83 genes that seem important to local adaptation (Gonzalez et al. 2023).

In conclusion, microsatellites and short targeted DNA sequences remain informative for some types of questions in molecular ecology and may be particularly useful for projects that have lower budgets or need to be completed within a short time frame. However, methods that generate genome-wide data are overall more informative and are increasingly affordable and accessible. Throughout the remaining chapters of this book, we will look at examples that are based on many different molecular markers, ranging from a handful of loci or a DNA sequence <100 bases in length, to entire genomes.

References

Adams, K.L. and Wendel, J.F. (2005). Polyploidy and genome evolution in plants. *Curr. Opin. Plant Biol.* 8(2): 135–141 https://doi.org/10.1016/j.pbi.2005.01.001.

Allio, R., Donega, S., Galtier, N., and Nabholz, B. (2017). Large variation in the ratio of mitochondrial to nuclear mutation rate across animals: implications for genetic diversity and the use of mitochondrial DNA as a molecular marker. *Mol. Biol. Evol.* 34(11): 2762–2772 https://doi.org/10.1093/molbev/msx197.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson, E.A. (2008) Rapid SNP discovery and genetic mapping using

sequenced RAD markers J.C. Fay (ed.). *PLoS One* 3 (10), e3376. https://doi.org/10.1371/journal.pone.0003376.

Bergeron, L.A., Besenbacher, S., Zheng, J. et al. (2023). Evolution of the germline mutation rate across vertebrates. *Nature* 615(7951): 285–291 https://doi.org/10.1038/s41586-023-05752-y.

Bhargava, A. and Fuentes, F.F. (2010). Mutational dynamics of microsatellites. *Mol. Biotechnol.* 44(3): 250–266 https://doi.org/10.1007/s12033-009-9230-4.

Bohmann, K., Elbrecht, V., Carøe, C. et al. (2022). Strategies for sample labelling and library preparation in DNA metabarcoding studies. *Mol. Ecol. Resour.* 22(4): 1231–1246 https://doi.org/10.1111/1755-0998.13512.

Bramwell, G., Schultz, A.G., Jennings, G. et al. (2024). The effect of mitochondrial recombination on fertilization success in blue mussels. *Sci. Total Environ.* 913: 169491 https://doi.org/10.1016/j.scitotenv.2023.169491.

Budd, A.M., Robins, J.B., Whybird, O., and Jerry, D.R. (2022). Epigenetics underpins phenotypic plasticity of protandrous sex change in fish. *Ecol. Evol.* 12(3): e8730 https://doi.org/10.1002/ece3.8730.

Camacho-Sanchez, M., Velo-Antón, G., Hanson, J.O. et al. (2020). Comparative assessment of range-wide patterns of genetic diversity and structure with SNPs and microsatellites: a case study with Iberian amphibians. *Ecol. Evol.* 10(19): 10353–10363 https://doi.org/10.1002/ece3.6670.

Chalopin, D., Naville, M., Plard, F. et al. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* 7(2): 567–580 https://doi.org/10.1093/gbe/evv005.

Chaudhary, S., Khokhar, W., Jabre, I. et al. (2019). Alternative splicing and protein diversity: plants versus animals. *Front. Plant Sci.* 10: 1–14 https://doi.org/10.3389/fpls.2019.00708.

Che, J., Chen, H.-M., Yang, J.-X. et al. (2012). Universal COI primers for DNA barcoding amphibians. *Mol. Ecol. Resour.* 12(2): 247–258 https://doi.org/10.1111/j.1755-0998.2011.03090.x.

Ciucani, M.M., Palumbo, D., Galaverni, M. et al. (2019). Old wild wolves: ancient DNA survey unveils population dynamics in Late Pleistocene and Holocene Italian remains. *PeerJ* 7: e6424 https://doi.org/10.7717/peerj.6424.

Clark, M.A., Douglas, M., and Choi, J. (2018). *Biology*. 2nd edition. Houston, TX: OpenStax.

Conlon, B.H., Aurori, A., Giurgiu, A.I. et al. (2019). A gene for resistance to the Varroa mite (*Acari*) in honey bee (*Apis mellifera*) pupae. *Mol. Ecol.* 28(12): 2958–2966 https://doi.org/10.1111/mec.15080.

Cui, L., Wall, P.K., Leebens-Mack, J.H. et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16(6): 738–749 https://doi.org/10.1101/gr.4825606.

Dai, T.M., Lü, Z.C., Liu, W.X. et al. (2017). The homology gene BtDnmt1 is essential for temperature tolerance in invasive Bemisia tabaci mediterranean cryptic species. *Sci. Rep.* 7(1): 3040 https://doi.org/10.1038/s41598-017-03373-w.

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49(3): 827–831 https://doi.org/10.1016/j.ympev.2008.09.009.

Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* 24(4): 400–402 https://doi.org/10.1038/74249.

Ge, C., Ye, J., Weber, C. et al. (2018). The histone demethylase KDM6B regulates temperature-dependent sex determination in a turtle species. *Science* 360(6389): 645–648 https://doi.org/10.1126/science.aap8328.

Gibbs, H.L., Sanz, L., Pérez, A. et al. (2020). The molecular basis of venom resistance in a rattlesnake-squirrel predator-prey system. *Mol. Ecol.* 29(15): 2871–2888 https://doi.org/10.1111/mec.15529.

Gonzalez, S.T., Alberto, F., and Molano, G. (2023). Whole-genome sequencing distinguishes the two most common giant kelp ecomorphs. *Evolution* 77(6): 1354–1369 https://doi.org/10.1093/evolut/qpad045.

Hakimzadeh, A., Abdala Asbun, A., Albanese, D. et al. (2024). A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses. *Mol. Ecol. Resour.* 24(5): e13847 https://doi.org/10.1111/1755-0998.13847.

Hamanaka, S., Yamaguchi, T., Kobayashi, T. et al. (2011). Generation of Germline-Competent rat induced pluripotent stem cells. *PLoS One* 6(7): e22008 https://doi.org/10.1371/journal.pone.0022008.

Han, K., Li, Z.F., Peng, R. et al. (2013). Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu. *Sci. Rep.* 3: 2101 https://doi.org/10.1038/srep02101.

Harper, L.R., Lawson Handley, L., Hahn, C. et al. (2018). Needle in a haystack? A comparison of eDNA metabarcoding and targeted qPCR for detection of the great crested newt (*Triturus cristatus*). *Ecol. Evol.* 8(12): 6330–6341 https://doi.org/10.1002/ece3.4013.

Herman, R.W., Clucas, G., Younger, J. et al. (2024). Whole genome sequencing reveals stepping-stone dispersal buffered against founder effects in a range expanding seabird. *Mol. Ecol.* 33(6): e17282 https://doi.org/10.1111/mec.17282.

Iannucci, A., Benazzo, A., Natali, C. et al. (2021). Population structure, genomic diversity and demographic history of Komodo dragons inferred from whole-genome sequencing. *Mol. Ecol.* 30(23): 6309–6324 https://doi.org/10.1111/MEC.16121.

Jackman, S.D., Coombe, L., Warren, R.L. et al. (2020). Complete mitochondrial genome of a gymnosperm, sitka spruce (*Picea sitchensis*), indicates a complex physical structure. *Genome Biol. Evol.* 12(7): 1174–1179 https://doi.org/10.1093/GBE/EVAA108.

Kaiser, S.A., Taylor, S.A., Chen, N. et al. (2017). A comparative assessment of SNP and microsatellite markers for assigning parentage in a socially monogamous bird. *Mol. Ecol. Resour.* 17(2): 183–193 https://doi.org/10.1111/1755-0998.12589.

Kuo, W., Zhong, L., Wright, S.J. et al. (2024). Beyond cyanogenesis: temperature gradients drive environmental adaptation in North American white clover (*Trifolium repens* L.). *Mol. Ecol.* 33(17): e17484 https://doi.org/10.1111/mec.17484.

Kurland, S., Saha, A., Keehnen, N. et al. (2024). New indicators for monitoring genetic diversity applied to alpine brown trout populations using whole genome sequence data. *Mol. Ecol.* 33(2): e17213 https://doi.org/10.1111/mec.17213.

Kurose, D., Pollard, K.M., and Ellison, C.A. (2020). Chloroplast DNA analysis of the invasive weed, Himalayan balsam (*Impatiens glandulifera*), in the British Isles. *Sci. Rep.* 10(1): 10966 https://doi.org/10.1038/s41598-020-67871-0.

Lamichhaney, S., Catullo, R., Keogh, J.S. et al. (2021). A bird-like genome from a frog: mechanisms of genome size reduction in the ornate burrowing frog, *Platyplectrum ornatum*. *Proc. Natl. Acad. Sci. U. S. A.* 118(11): e2011649118 https://doi.org/10.1073/pnas.2011649118.

Lamka, G.F., Harder, A.M., Sundaram, M. et al. (2022). Epigenetics in ecology, evolution, and conservation. *Front. Ecol. Evol.* 10: 871791 https://doi.org/10.3389/fevo.2022.871791.

Lemopoulos, A., Prokkola, J.M., Uusi-Heikkilä, S. et al. (2019). Comparing RADseq and microsatellites for estimating genetic diversity and relatedness—implications for brown trout conservation. *Ecol. Evol.* 9(4): 2106–2120 https://doi.org/10.1002/ece3.4905.

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(2): 595–609.

Liu, Y., Medina, R., and Goffinet, B. (2014). 350 my of mitochondrial genome stasis in mosses, an early land plant lineage. *Mol. Biol. Evol.* 31(10): 2586–2591 https://doi.org/10.1093/molbev/msu199.

Liu, S., Westbury, M.V., Dussex, N. et al. (2021). Ancient and modern genomes unravel the evolutionary history of the rhinoceros family. *Cell* 184(19) : 4874–4885.e16. https://doi.org/10.1016/j.cell.2021.07.032

Lou, R.N., Jacobs, A., Wilder, A.P., and Therkildsen, N.O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.* 30(23): 5966–5993 https://doi.org/10.1111/MEC.16077.

Mallarino, R., Linden, T.A., Linnen, C.R., and Hoekstra, H.E. (2017). The role of isoforms in the evolution of cryptic coloration in Peromyscus mice. *Mol. Ecol.* 26(1): 245–258 https://doi.org/10.1111/mec.13663.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338(6114): 1593–1599 https://doi.org/10.1126/science.1228186.

Mérot, C., Oomen, R.A., Tigano, A., and Wellenreuther, M. (2020). A roadmap for understanding the evolutionary significance of structural genomic variation. *Trends Ecol. Evol.* 35(7): 561–572 https://doi.org/10.1016/j.tree.2020.03.002.

Metzger, D.C.H. and Schulte, P.M. (2017). Persistent and plastic effects of temperature on DNA methylation across the genome of threespine stickleback (*Gasterosteus aculeatus*). *Proc. R. Soc. Lond. B Biol. Sci.* 284(1864): 20171667 https://doi.org/10.1098/rspb.2017.1667.

Minhas, B.F., Beck, E.A., Cheng, C.H.C., and Catchen, J. (2023). Novel mitochondrial genome rearrangements including duplications and extensive heteroplasmy could underlie temperature adaptations in Antarctic notothenioid fishes. *Sci. Rep.* 13(1): 6939 https://doi.org/10.1038/s41598-023-34237-1.

Miya, M., Gotoh, R.O., and Sado, T. (2020). MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fish. Sci.* 86(6): 939–970 https://doi.org/10.1007/s12562-020-01461-x.

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

Pegadaraju, V., Nipper, R., Hulke, B. et al. (2013). De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics* 14(1): 556 https://doi.org/10.1186/1471-2164-14-556.

Pellicer, J. and Leitch, I.J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytol.* 226(2): 301–305 https://doi.org/10.1111/nph.16261.

Peterson, B.K., Weber, J.N., Kay, E.H. et al. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7(5) : e37135.https://doi.org/10.1371/journal.pone.0037135

Qiao, Y., Zhang, X., Li, Z. et al. (2022). Assembly and comparative analysis of the complete mitochondrial genome of Bupleurum chinense DC. *BMC Genomics* 23(1) : 664. https://doi.org/10.1186/s12864-022-08892-z

Rahlves, A. (2022). *A New Frontier: Fastest Ever DNA Sequencing Technique Achieved England: Guinness World Records; 2022*. Guinness World Records https://www.guinnessworldrecords.com/news/commercial/2022/5/a-new-frontier-fastest-ever-dna-sequencing-technique-achieved-702401.

Richardson, A.O., Rice, D.W., Young, G.J. et al. (2013). The "fossilized" mitochondrial genome of Liriodendron tulipifera: ancestral gene content and order, ancestral editing sites, and extraordinarily low mutation rate. *BMC Biol.* 11(1): 29 https://doi.org/10.1186/1741-7007-11-29.

Romagnoli, S., Bartalucci, N., and Vannucchi, A.M. (2023). Resolving complex structural variants via nanopore sequencing. *Front. Genet.* 14: 1213917 https://doi.org/10.3389/fgene.2023.1213917.

Satam, H., Joshi, K., Mangrolia, U. et al. (2023). Next-generation sequencing technology: current trends and advancements. *Biology.* 12(7): 997 https://doi.org/10.3390/biology12070997.

Scarsbrook, L., Verry, A.J.F., Walton, K. et al. (2023). Ancient mitochondrial genomes recovered from small vertebrate bones through minimally destructive DNA extraction: phylogeography of the New Zealand gecko genus Hoplodactylus. *Mol. Ecol.* 32(11) : 2964–2984.https://doi.org/10.1111/mec.16434

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 B.-S. Yue (ed.). *PLOS ONE* 11(4): e0153666 https://doi.org/10.1371/journal.pone.0153666.

Silva, S.R., Alvarenga, D.O., Aranguren, Y. et al. (2017). The mitochondrial genome of the terrestrial carnivorous plant *Utricularia reniformis* (Lentibulariaceae): structure, comparative analysis and evolutionary landmarks. *PLoS One* 12(7): e0180484 https://doi.org/10.1371/journal.pone.0180484.

Singh, N.R. (2023). *Introduction to Genetics Copyright © 2023 by Natasha Ramroop Singh, Thompson Rivers University*. Kamloops, British Columbia: Thompson Rivers University.

Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P. et al. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biol.* 10(1): e1001241 https://doi.org/10.1371/journal.pbio.10.1241.

Song, J. and McDowell, J.R. (2021). Comparative transcriptomics of spotted seatrout (*Cynoscion nebulosus*) populations to cold and heat stress. *Ecol. Evol.* 11(1352): –1367 https://doi.org/10.1002/ece3.7138.

Stampar, S.N., Broe, M.B., Macrander, J. et al. (2019). Linear Mitochondrial Genome in Anthozoa (Cnidaria): a case study in Ceriantharia. *Sci. Rep.* 9(1): 6094 https://doi.org/10.1038/s41598-019-42621-z.

Sterling-Montealegre, R.A. and Prada, C.F. (2024). Variability and evolution of gene order rearrangement in mitochondrial genomes of arthropods (except Hexapoda). *Gene* 892: 147906 https://doi.org/10.1016/j.gene.2023.147906.

Sunde, J., Yıldırım, Y., Tibblin, P., and Forsman, A. (2020). Comparing the performance of microsatellites and RADseq in population genetic studies: analysis of data for pike (*Esox lucius*) and a synthesis of previous studies. *Front. Genet.* 11 https://doi.org/10.3389/FGENE.2020.00218.

Szatmári, L., Cserkész, T., Laczkó, L. et al. (2021). A comparison of microsatellites and genome-wide SNPs for the detection of admixture brings the first molecular evidence for hybridization between *Mustela eversmanii* and *M. putorius* (*Mustelidae, Carnivora*). *Evol. Appl.* 14(9) : 2286–2304.https://doi.org/10.1111/eva.13291

Tang, X., Wang, Q., and Huang, X. (2018). Chilling-induced DNA demethylation is associated with the cold tolerance of *Hevea brasiliensis*. *BMC Plant Biol.* 18(1): 70 https://doi.org/10.1186/s12870-018-1276-7.

Theissinger, K., Fernandes, C., Formenti, G. et al. (2023). How genomics can help biodiversity conservation. *Trends Genet.* 39(7): 545–559 https://doi.org/10.1016/j.tig.2023.01.005.

Theron, E., Hawkins, K., Bermingham, E. et al. (2001). The molecular basis of an avian plumage polymorphism in the wild: a melanocortin-1-receptor point mutation is perfectly associated with the melanic plumage morph of the bananaquit, *Coereba flaveola*. *Curr. Biol.* 11(8): 550–557 https://doi.org/10.1016/S0960-9822(01)00158-0.

Uhlen, M. and Quake, S.R. (2023). Sequential sequencing by synthesis and the next-generation sequencing revolution. *Trends Biotechnol.* 41(12): 1565–1572 https://doi.org/10.1016/j.tibtech.2023.06.007.

Ulaszewski, B., Meger, J., and Burczyk, J. (2021). Comparative analysis of SNP discovery and genotyping in *Fagus sylvatica* L. and *Quercus robur* L. using RADseq, GBS, and ddRAD methods. *Forests.* 12(2): 222 https://doi.org/10.3390/f12020222.

Viitaniemi, H.M., Verhagen, I., Visser, M.E. et al. (2019). Seasonal variation in genome-wide DNA methylation patterns and the onset of seasonal timing of reproduction in Great Tits. *Genome Biol. Evol.* 11(3): 970–983 https://doi.org/10.1093/gbe/evz044.

Vilaça, S.T., Maroso, F., Lara, P. et al. (2023). Evidence of backcross inviability and mitochondrial DNA paternal leakage in sea turtle hybrids. *Mol. Ecol.* 32(3): 628–643 https://doi.org/10.1111/MEC.16773.

Wang, D., Zheng, Z., Li, Y. et al. (2021). Which factors contribute most to genome size variation within angiosperms? *Ecol. Evol.* 11(6): 2660–2668 https://doi.org/10.1002/ece3.7222.

Weissensteiner, M.H., Bunikis, I., Catalán, A. et al. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nat. Commun.* 11(1): 3403 https://doi.org/10.1038/s41467-020-17195-4.

Wright, N.A., Gregory, T.R., and Witt, C.C. (2014). Metabolic 'engines' of flight drive genome size reduction in birds. *Proc. R. Soc. B: Biol. Sci.* 281(1779): 20132780 https://doi.org/10.1098/rspb.2013.2780.

Wu, Z.Q., Liao, X.Z., Zhang, X.N. et al. (2022). Genomic architectural variation of plant mitochondria—a review of multichromosomal structuring. *J. Syst. Evol.* 60(1): 160–168 https://doi.org/10.1111/jse.12655.

Zhang, K., Zhu, K., Liu, Y. et al. (2021). Novel gene rearrangement in the mitochondrial genome of *Muraenesox cinereus* and the phylogenetic relationship of Anguilliformes. *Sci. Rep.* 11(1): 2411 https://doi.org/10.1038/s41598-021-81622-9.

Zhang, H.Y., Bissett, A., Aguilar-Trigueros, C.A. et al. (2023). Fungal genome size and composition reflect ecological strategies along soil fertility gradients. *Ecol. Lett.* 26(7): 1108–1118 https://doi.org/10.1111/ele.14224.

Zhou, C., Wang, P., Zeng, Q. et al. (2023). Comparative chloroplast genome analysis of seven extant Citrullus species insight into genetic variation, phylogenetic relationships, and selective pressure. *Sci. Rep.* 13(1): 6779 https://doi.org/10.1038/s41598-023-34046-6.

Zwonitzer, K.D., Tressel, L.G., Wu, Z. et al. (2024). Genome copy number predicts extreme evolutionary rate variation in plant mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* 121(10): e2317240121 https://doi.org/10.1073/pnas.2317240121.