Chapter 1
Plant genetic material

1.1 DNA is the genetic material of all living organisms, including plants

Like all living organisms, plants use deoxyribonucleic acid (DNA) as their genetic material. DNA is a polymer that consists of alternating sugars and phosphates with nitrogenous bases attached to the sugar moiety. More specifically, the nucleotide building block of DNA is a deoxyribose sugar with a phosphate group attached to carbon 5 (C-5) and a nitrogenous base to carbon 1 (C-1). Phosphodiester bonds connect the C-5 phosphate group of one nucleotide to the carbon 3 (C-3) of another, creating the alternating sugar–phosphate backbone of the DNA molecule. This means that one end of the chain is terminated by a C-5 phosphate, and is known as the 5′ end, whereas the other end is terminated by a C-3 hydroxyl, and is known as the 3′ end (Figure 1.1a). The idea that DNA molecules have a polarity is one that will be revisited over and over throughout this book.

Only four nitrogenous bases are used in a DNA molecule. Two of these, **cytosine** (C) and **thymine** (T), have a single aromatic ring consisting of four carbons and two nitrogen groups, and are classified as **pyrimidines**. The other two, **adenine** (A) and **guanine** (G), each have a double ring consisting of a pyrimidine ring fused to a 5-membered, heterocyclic ring, and are classified as **purines**. The bases form a linear molecule, a strand of DNA that interacts with the nitrogenous bases on the other strand.

Two DNA polymers, or strands, together form the iconic double helix, a structure like a twisted ladder that has come to symbolize life and its historical continuity (Figure 1.1b). Even viruses, many of which have genomes of single-stranded nucleic acids, must eventually pass through a double-stranded stage to reproduce. The strands are held together by hydrogen bonds (H-bonds) between the nitrogenous bases, with two bonds between A and T, and three between G and C (Figure 1.1c). Since more H-bonds between bases hold them together more tightly, it is significantly easier to denature a DNA molecule with many A-T base pairs than one with many C-G base pairs. The pairing rules for DNA are largely inflexible: A forms H-bonds with T and G with C. The strands are arranged in antiparallel fashion, so that the 5′ end base of one strand pairs with the 3′ end base of the other, and vice versa.

The structure of DNA is not unique to plants, but rather is shared among all three domains of life (Eukarya, Bacteria, and Archaea), as well as by viruses. The patterns of covalent bonds and H-bonds can thus be studied in any organism, and indeed much of what we know about DNA structure was originally worked out in bacteria, which are unicellular and prokaryotic (lacking a nucleus).

The four bases (A, C, G and T) are not present in equal amounts and can vary between genomes, parts of genomes, and species. For example, the A+T content of the chloroplast genome, an organellar genome discussed later in Chapter 5, is variable, but generally greater than 50% of the total. In contrast, nuclear genes of many grasses are enriched in G+C, a bias that is particularly noticeable in maize.

Plant Genes, Genomes and Genetics, First Edition. Erich Grotewold, Joseph Chappell and Elizabeth A. Kellogg.

Companion Website: www.wiley.com/go/grotewold/plantgenes.

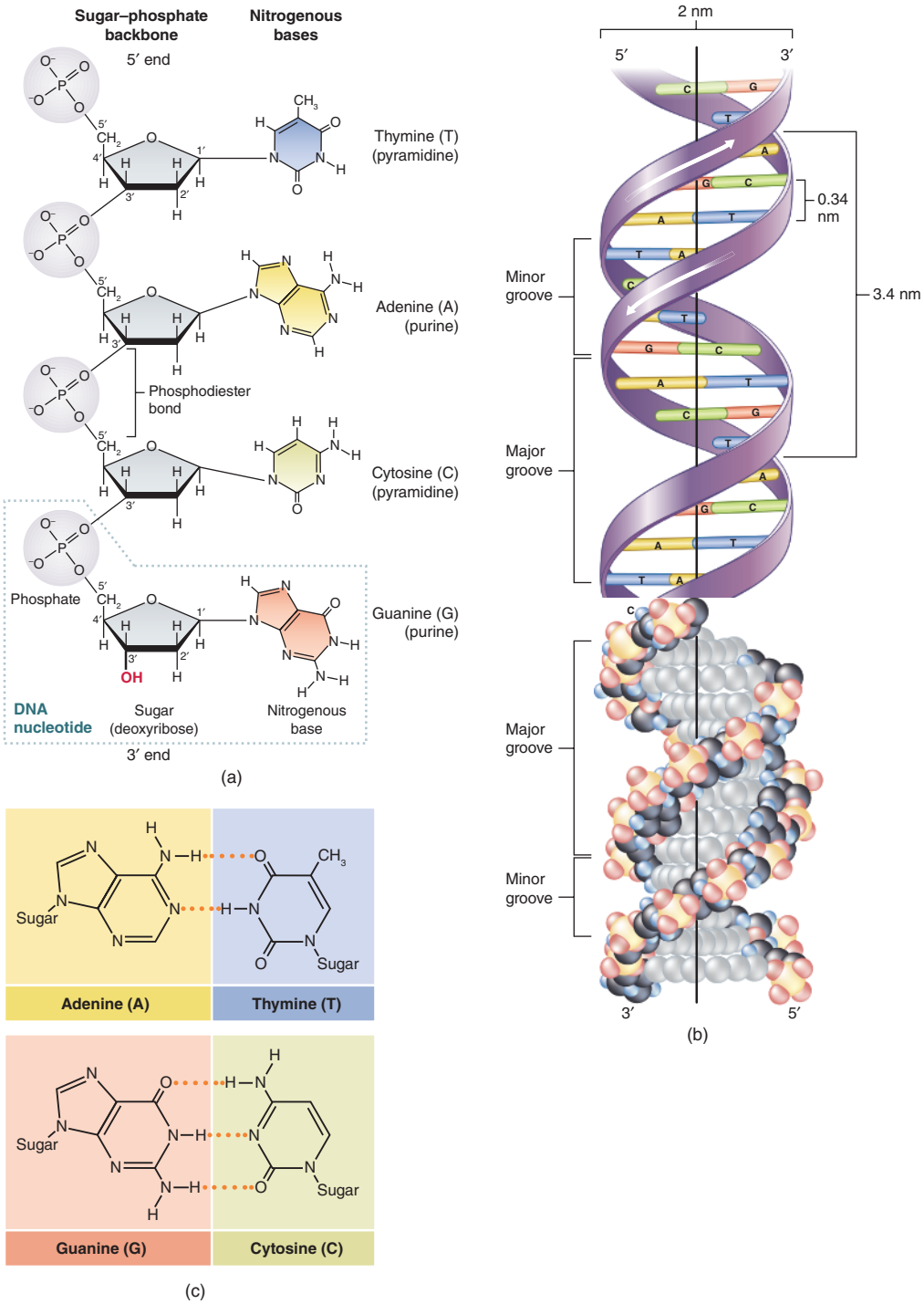


Figure 1.1 Structure of the DNA molecule. (a) Alternating phosphate and ribose groups make up the backbone of the DNA strand. The C-3 hydroxyl at the 3′ end is required for attaching new nucleotides to the chain, so the DNA strand always is extended from the 3′ end. The nitrogenous bases are attached to C-1 of the ribose. (b) The double helix, formed from two antiparallel strands of DNA. (c) The four nitrogenous bases and the hydrogen bonds between them. (a) and (c) Reece *et al.* (2011). (b) Adapted from Raven *et al.* (2011)

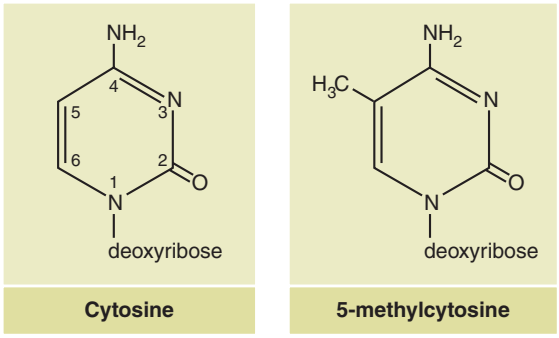


Figure 1.2 Structures of cytosine and 5-methylcytosine. http://en.wikipedia.org/wiki/File:Cytosine_chemical_structure.svg. By Engineer gena (Own work) [Public domain], via Wikimedia Commons

In plants, the nucleotide bases may be modified by attachment of methyl ($-CH_3$) groups to particular sites. A common position for DNA methylation is on C-5 of C (Figure 1.2), although adenine methylation is also possible, particularly in bacteria, Archaea, and unicellular eukaryotes. This common modification of the DNA is known to affect transcription, and will be discussed in more detail in Chapter 12. While methylation is also common in mammals, it is relatively rare in yeast and in the fruit fly (*Drosophila melanogaster*). Other insects, however, have extensive DNA methylation, as is the case of honeybees. Many aspects of biotechnology exploit the basic structure of DNA, as described in the box "Working with DNA."

Working with DNA: biotechnology takes advantage of the properties of the DNA molecule

The polymerase chain reaction (PCR), a method used extensively in biotechnology for generating large numbers of similar or identical DNA fragments, relies on repeatedly increasing the temperature to separate the DNA strands and then decreasing it to allow primers to bind. It is thus important to know the temperature at which the strands of particular DNA molecules separate; this is known as the melting temperature, or T_m, and corresponds to the temperature at which half the DNA molecules are single-stranded (melted) and half are double-stranded. The T_m is controlled by several factors, but a major one is the fraction of G-C base pairs. Because G-C pairs are held together by three H-bonds (rather than two as in A-T pairs), breaking them requires more energy input, such as higher temperatures. A rough equation for the T_m of a short (<20 base pairs, bp) strand of DNA is:

$$T = 2(\mathrm{A} + \mathrm{T}) + 4(\mathrm{G} + \mathrm{C})$$

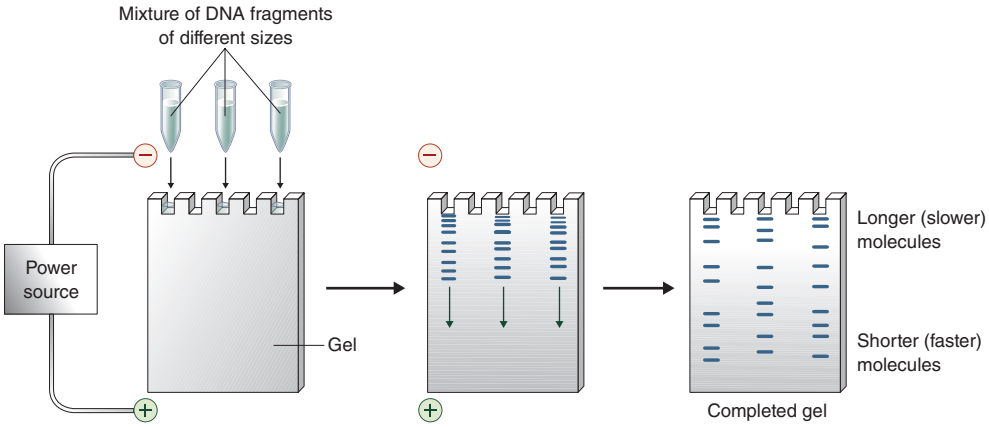
where T is the temperature in degrees Centigrade, A+T is the total number of A-T base pairs and G+C is the total number of G-C base pairs.

Assuming a random nucleotide distribution, this rough equation makes two assumptions. The first is that one strand of the DNA is bound to a membrane, as it would be for a Southern blot, and that the blot is being probed with a short oligonucleotide (a single-stranded DNA molecule, generally <100 nucleotides long; the Greek prefix *oligo-* means "few"). With one strand immobilized, the DNA melts at a lower temperature (about 8°C less) than it would in solution. The second assumption is that the concentration of salt (e.g., NaCl) is 0.9 M, and that there is no chemical in the solution that would interfere with the formation of H-bonds between the bases (such as formamide, $HCONH_2$). The melting temperature of DNA increases with the $\log_{10}$ of the concentration of salt. This means that the higher the salt concentration, the more stable the DNA heteroduplex. It also decreases linearly with the concentration of formamide or other similar small molecules that interfere with DNA H-bond formation. Thus, a more complete equation is:

$$T_m = 81.5 + 16.6 \log M + 41(XG + XC) - 500/L - 0.62F$$

where M is the molar concentration of cations (Na^+ in this case), XG and XC are the mole fractions of G and C in the DNA, L is the length of the DNA (actually just the shortest strand of DNA in the mix), and F is the concentration of formamide.

The strong negative charge of DNA created by the phosphate backbone (on the outside of the double helix, Figure 1.1) can also be used to sort DNA molecules by size using gel electrophoresis, one of the most common tools in molecular biology. DNA is placed in a well at one end of a gel matrix (commonly an agarose or polyacrylamide gel), and an electrical current is run through the gel. The negatively charged DNA will thus migrate toward the positive electrode, and because the gel acts as a molecular sieve, the rate of DNA migration is dependent upon its size (Figure 1.3a). Smaller DNA fragments run faster than larger ones; graphing the $\log_{10}$ of the molecular weight or fragment length against the distance traveled produces a line (Figure 1.3b).



Distance migrated (mm)	Fragment size (bp)
10.0	10180
11.5	9162
12.5	8144
13.3	7126
16.0	6108
19.0	5090
22.0	4072
27.0	3054
35.0	2036
39.5	1636
49.5	1018
61.0	517

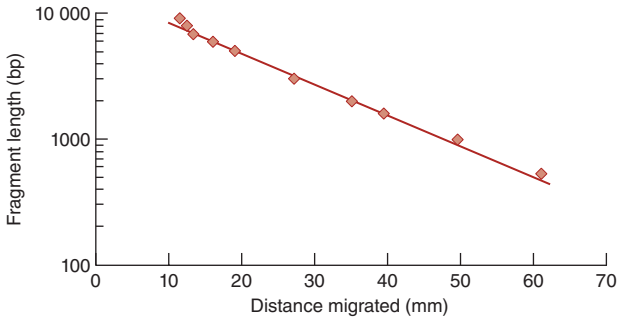
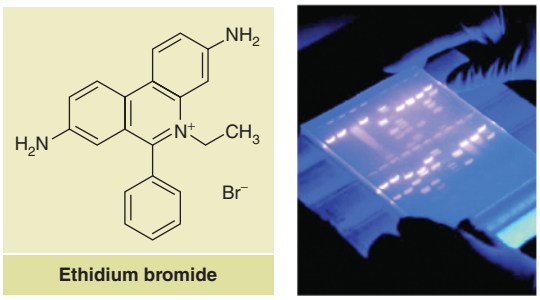
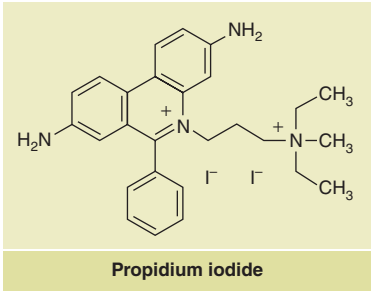


Figure 1.3 Gel electrophoresis, a powerful tool for biology. (a) DNA, suspended in an aqueous solution, is taken from a tube and placed in slots in a gel. An electrical current is applied to the gel. Because DNA is negatively charged, it moves toward the positive electrode, with smaller fragments moving more rapidly than larger ones. (b) Table and graph of the size of DNA fragments versus the distance migrated on a representative gel. Note that this relationship is not linear, so that the vertical axis of the graph is logarithmic. (b) Adapted from http://depts.noctrl.edu/biology/resource/handbook.htm

Certain chemicals bind to DNA and fluoresce when illuminated with an appropriate light source. For example, ethidium bromide has been widely used because it will intercalate into the double helix of the DNA; once there it will fluoresce under UV light. Thus, a common method of locating DNA on an agarose gel is to soak the gel in ethidium bromide and then place it on a UV light source. The DNA will then appear as pinkish bands (Figure 1.4a). Unfortunately, ethidium bromide will intercalate into the DNA of anything, including that of the biologist working with it. Because it can absorb light energy, it can damage the DNA; it is thus mutagenic. Its use is becoming less common because of its toxicity.



(a)



(b)

Figure 1.4 Chemicals that can be used to make DNA visible. (a) Ethidium bromide: (i) structure of ethidium bromide; (ii) an agarose gel stained with ethidium bromide and viewed with UV light; and (iii) ethidium bromide inserted into a DNA molecule (ai) http://www.sigmaaldrich.com, (aii) http://www.hamiltoncompany.com/products/syringes/c/893/. Reproduced with permission from Hamilton Company, (aiii)http://en.wikipedia.org/wiki/File:DNA_intercalation2.jpg. Image created by Karol Langner, via Wikimedia Commons. (b) Propidium iodide, from http://www.sigmaaldrich.com

Another fluorescent dye is propidium iodide, which, like ethidium bromide, intercalates into the DNA double helix (Figure 1.4b). Propidium iodide binds to DNA in a quantitative way, with one propidium iodide molecule per 4 or 5 bp; thus more DNA equals more propidium iodide binding, which equals more fluorescence. This direct relationship is used to estimate the genome size, which is the amount of DNA in the nucleus of a cell. Estimates of genome size using propidium iodide fluorescence are relatively rapid, so we have data on the genome size of many organisms, including flowering plants.

Methylated cytosines can be detected by a variety of methods. Restriction enzymes will cut DNA at particular sequences, but some will not cut DNA at 5-methyl cytosine, so that the methyl groups effectively protect the DNA from cleavage at these sites. Such methylation-sensitive restriction enzymes often have the same cut site sequence as methylation-insensitive enzymes. The restriction enzymes can thus be used sequentially. A restriction site that is cut with the methylation-insensitive enzyme and not with the methylation-sensitive one must be methylated. More recently, methods have been developed to find

all the methylated cytosines in a genome. One such method is bisulfite sequencing, in which a genome is treated with sodium bisulfite, which converts methylated cytosines to uracils (Figure 1.5a). When the DNA is amplified by PCR the uracils become thymines. By comparing the sequence of the genome (or genomic region) before and after the bisulfite treatment, the cytosines that were methylated can be identified. The non-methylated cytosines are not affected by the bisulfite treatment, and thus remain the same (Figure 1.5b).

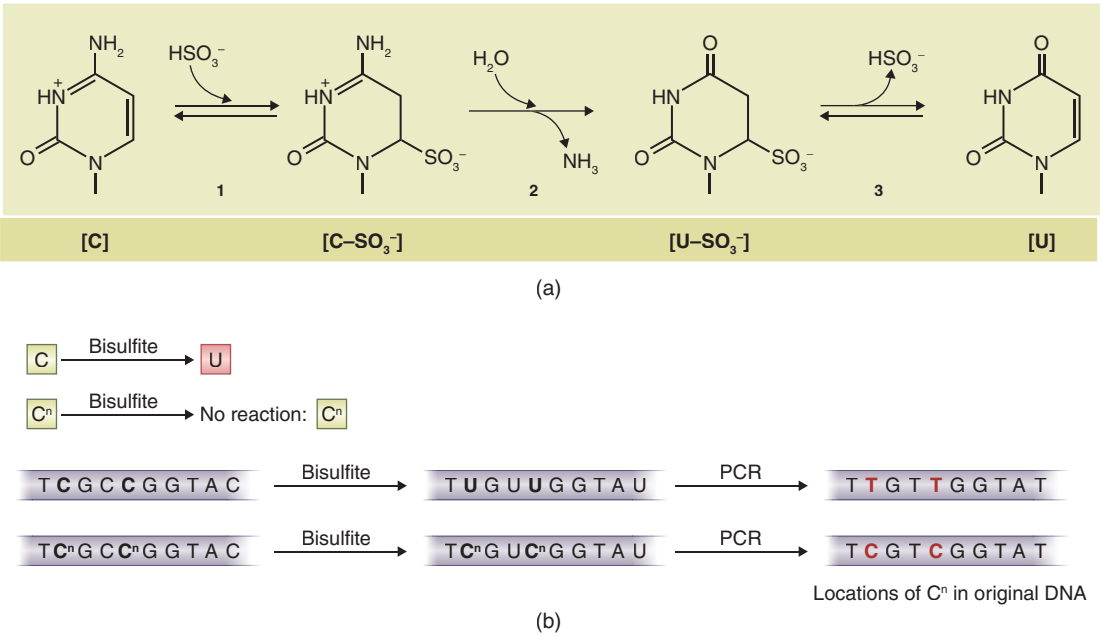


Figure 1.5 Bisulfite sequencing. (a) Chemical reactions necessary to remove the amino group from cytosine and convert it to uracil. This reaction will not occur if the cytosine is methylated. http://www.methylogix.com/genetics/bisulfite.shtml.htm. (b) Inferring the presence of 5-methylcytosine in a sequence of DNA. The DNA is treated with bisulfite and all non-methylated cytosines (C) are converted to uracil (U) (top line of sequences) whereas methylated cytosines are unchanged (bottom line of sequences). DNA is then amplified by PCR, which converts U to thymine (T), and the resulting sequence compared with the original. Any base that is C in the original but T in the amplified product is inferred to be non-methylated; conversely a base that is C in both the original and the amplified product must have been methylated. Redrawn from Hayatsu *et al.* (2008). Reproduced with permission of John Wiley & Sons, Inc.

Methylation-sensitive restriction enzymes are often used to generate genomic libraries, which are collections of DNA fragments cloned into autonomously replicating vectors such as a plasmid, phage (bacterial virus) or bacterial artificial chromosome (BAC).

1.2 The plant cell contains three independent genomes

The DNA in plant cells is found in the nucleus, the mitochondria and the chloroplasts. The latter two organelles are descendants of bacteria that were captured by a eukaryotic cell and have become endosymbionts; because many of the ancestral bacterial genes were transferred to the nucleus, the organelles can never revert to being free-living bacteria. Such microbial symbioses occur commonly. For example, cyanobacteria occur in the cells of some ferns, and in the stems of cycads. They co-occur with fungi to form lichens, although some lichens form from the symbiosis of a green alga and a fungus. There

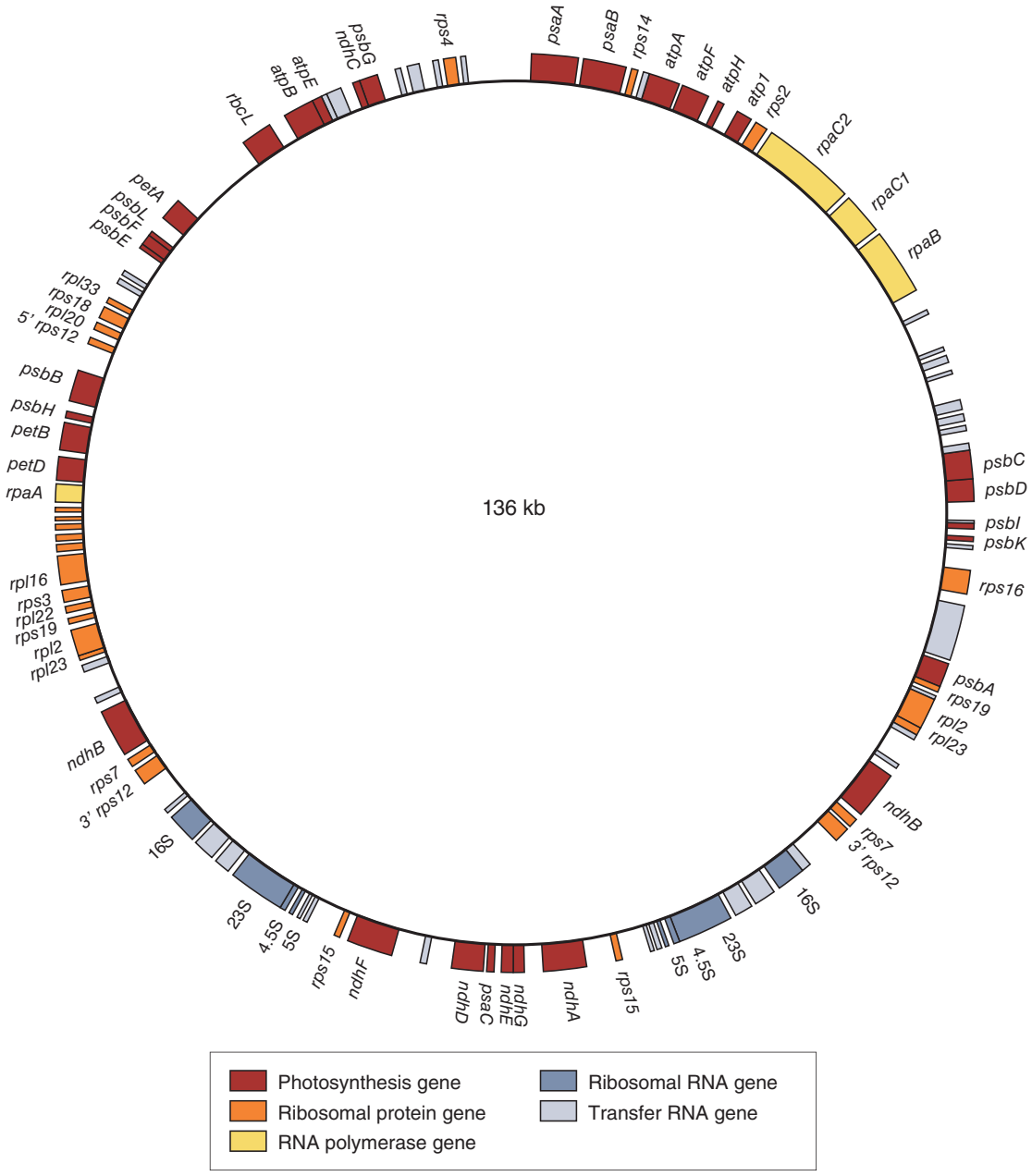


Figure 1.6 The chloroplast genome exemplified by that of rice. Genes are indicated by colored boxes. Note that there is a set of genes, from rps15 to rpl23, that appear twice but in inverse order. These regions are known as the inverted repeat. Redrawn from Brown, 2002

is even one recent example of a slug that has acquired plastids (Rumpho *et al.*, 2008). The extent of gene transfer between members of the symbiosis varies greatly in these examples, and hence the ability of the bacterial endosymbionts to function independently varies accordingly.

Plant mitochondria and chloroplasts have circular genomes, similar to those in Bacteria and Archaea (Figure 1.6). Their translational machinery (ribosomal RNA and proteins) is more similar to that in their bacterial ancestors than to the eukaryotic ribosomes encoded by the nuclear genome. The organellar genomes will be discussed in more detail in Chapter 5.

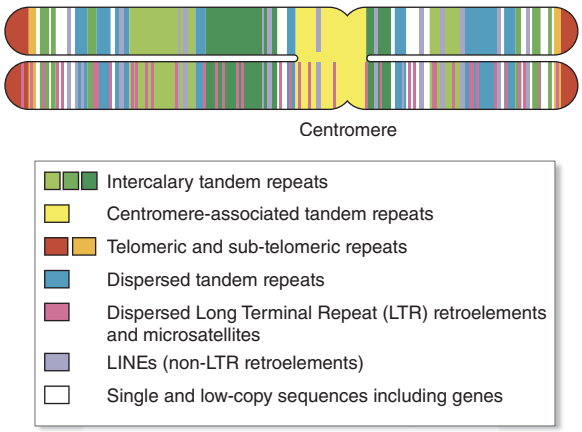


Figure 1.7 A chromosome, shown at metaphase, with the sister chromatids joined at the centromere. The two chromatids are identical in sequence, but different sorts of sequences are highlighted on each one for clarity. Schmidt and Heslop-Harrison (1998). Reproduced with permission of Elsevier

There is ample evidence from multiple plant species that the broad organization of plant genomes is similar to that of most eukaryotes. In all eukaryotes, including plants, the DNA of the nucleus is organized into linear chromosomes. The ends of the chromosomes are marked by distinctive structures, the telomeres, which have characteristic DNA sequences and have important roles to play in DNA replication. These will be discussed in more detail in Chapter 4 (Figure 1.7).

The other major landmark of eukaryotic chromosomes is the centromere. When chromosomes are viewed through a microscope, centromeres appear as constrictions that divide each chromosome into two segments – the chromosome arms. Centromeres in plants, as in all other eukaryotes, provide the point of attachment for the spindle apparatus during cell division. Centromeres will be discussed in more detail in Chapter 4 (Figure 1.7).

1.3 A gene is a complete set of instructions for building an RNA molecule

DNA is a coded set of instructions for making RNA. The fate of the RNA is hugely varied and is the subject of Part 2 of this book. In general, RNA may function as an independent regulatory molecule (e.g., micro RNA), or as piece of cellular machinery (e.g., transfer RNA, ribosomal RNA), or as a set of instructions for making a protein (e.g., messenger RNA, or mRNA), or some combination of these. In other words, RNA is a central molecule in the life of a cell, and DNA is simply a storage mechanism and blueprint for preserving and expressing the information in the form of RNAs. The genome is thus the full set of RNA-producing instructions, along with the information on when (in developmental or ecological time) and where (in what tissue) to use a particular set of instructions.

The RNAs that make proteins serve two distinct masters. First, and most familiar, are the mRNAs that make proteins to carry out all the cellular functions of the plant, including enzymes, structural proteins, receptors, and transcription factors. The other mRNAs serve the transposable elements, which are mobile pieces of DNA that move around the genome. In any given genome, there are thousands of transposable elements, each of which produces mRNAs that encode proteins that participate in transposon movement (transposition).

With this view of the genome, we may consider a gene as a complete set of instructions for making a particular RNA, including the non-coding DNA sequences that provide information on when and where the coding sequences should be transcribed. Under this very broad definition, the total number

Table 1.1 Genome sizes of selected plants for which genome sequences are available. Note that these are heavily biased toward plants with small genomes, which may affect our ability to generalize about plant genome structure and function. Note also that genome size does not correlate with any clear evolutionary relationships. For instance, Arabidopsis, a common experimental angiosperm species, has about the same genome size as *Chlamydomonas*, a green alga similar to algae that existed hundreds of millions of years prior to Arabidopsis

Species	Classification	Genome size (Mbp)	Estimated number of protein-coding genes
Selaginella moellendorffii (Banks *et al.*, 2011)	Lycophyte	106	22 285
Chlamydomonas reinhardtii (Merchant *et al.*, 2007)	Green alga	121	15 143
Arabidopsis thaliana (Arabidopsis Genome Initiative, 2000)	Angiosperm	125	27 025
Oryza sativa (Goff *et al.*, 2002; Yu *et al.*, 2002)	Angiosperm	420	39 045
Physcomitrella patens (Rensing *et al.*, 2008)	Moss	480	35 938
Populus trichocarpa (Tuskan *et al.*, 2006)	Angiosperm	485	41 335
Vitis vinifera (Velasco *et al.*, 2007)	Angiosperm	505	29 585
Sorghum bicolor (Paterson *et al.*, 2009)	Angiosperm	800	33 032
Zea mays (Schnable *et al.*, 2009)	Angiosperm	2300	39 475

of genes in any given genome is not known with any accuracy. Most commonly, when gene numbers are reported in the literature, the number includes only the mRNA-producing, protein-coding genes, excluding all those produced by the transposable elements. A few of these numbers are shown in Table 1.1, and are about the same order of magnitude (2–4 x 10^4) between different plants. However, these account for only a tiny fraction of the genome; they are vastly outnumbered by the protein-coding genes from the transposable elements and the genes encoding non-messenger RNAs. Even though they constitute only a minority of the genes in the genome, the mRNA producing, protein-coding, non-transposable element sequences are usually the ones simply called "genes" in the literature. We will follow this common usage here unless we specify otherwise.

1.4 Genes include coding sequences and regulatory sequences

Plant nuclear genes are similar in general structure to those of other eukaryotes. The overall architecture of a gene consists of two general components, the regulatory region and the coding or structural region of the gene (Figure 1.8). The regulatory region is responsible for controlling when a gene is transcribed into RNA. The regulatory region does not appear in the resulting mRNA, but directs the transcription machinery to start RNA biosynthesis (transcription) at a particular position, often but not necessarily, 3′ to the regulatory region. RNA transcription proceeds to the end of the gene generating a large precursor mRNA which will be processed in several ways. The portions of the gene that ultimately end up in the mature mRNA are known as the exons, whereas the portions of the RNA that get spliced out during processing are known as the introns, or intervening RNAs. In protein-coding genes, upstream (5′) of the coding sequence is a region that is transcribed into mRNA, but not translated to protein. This is referred to as the 5′ untranslated region (5′ UTR). A similar untranslated region occurs downstream of the last coding portion of the sequence and is known as the 3′ UTR. UTRs are part of the exons because they are present in the mature mRNA, even though they are not translated into proteins. The UTR regions of mRNAs are known to play roles in initiation of the translation process and stability of the mRNA. After transcription, the introns are spliced out of the messenger RNA, a 5′ cap is added to the 5′ UTR and a polyadenine (polyA) tail is added to the 3′ UTR. These RNA processing steps will be described in more detail in Chapter 13.

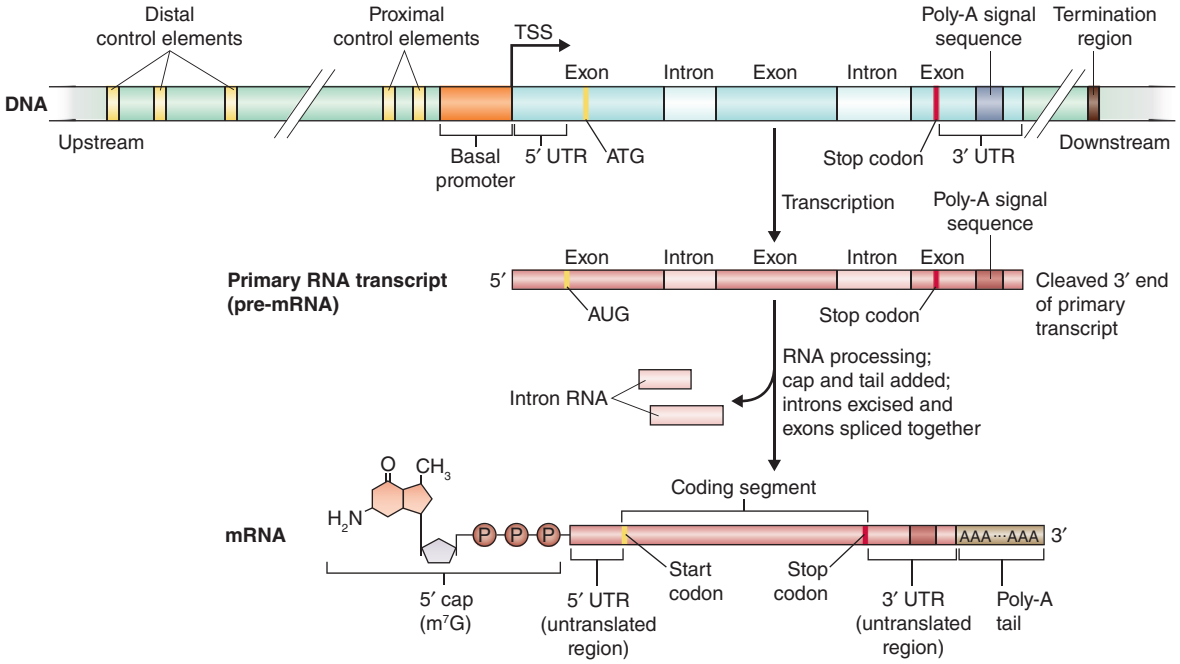


Figure 1.8 Steps in the formation of messenger RNA from DNA. The DNA for a single gene (a protein-coding sequence and all its control elements) is shown at the top. Note that the protein-coding sequence itself is a relatively small portion of the entire gene. The primary RNA transcript (pre-mRNA) also includes sequences upstream and downstream of the part that will be translated. The mature mRNA is formed by removing introns and adding a cap and a tail. Adapted from Campbell *et al*. (2005). While the structure shown is very common, the regulatory elements may occasionally be quite far away from the gene. Also the pre-mRNA may sometimes be spliced in several different ways, a process known as alternative splicing, which is discussed in more detail in Chapter 13

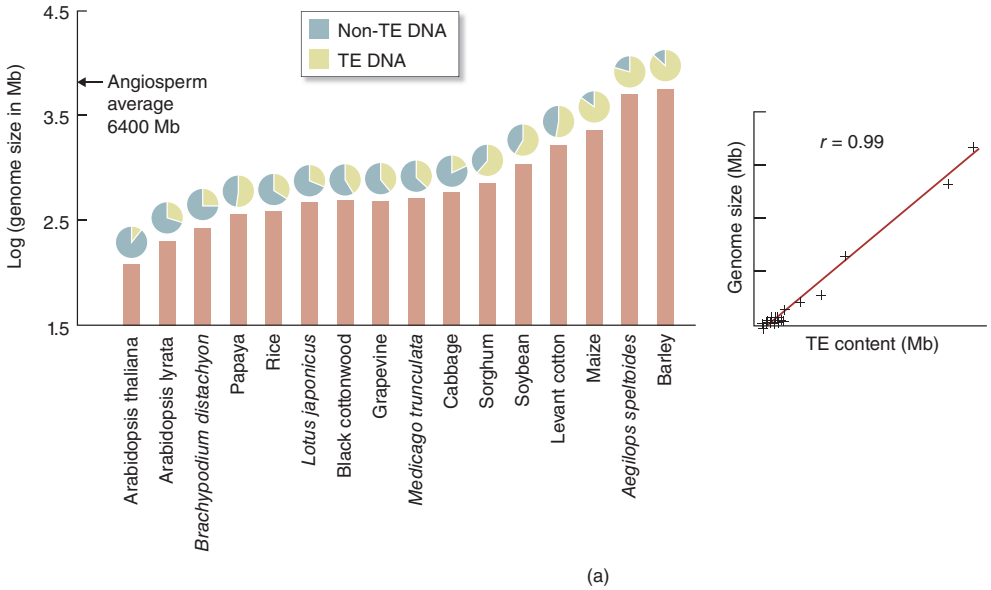
1.5 Nuclear genome size in plants is variable but the numbers of protein-coding, non-transposable element genes are roughly the same

In Bacteria, Archaea, and their mitochondrial and chloroplast descendants, most DNA is made up of sequences that encode RNA or proteins, and the coding sequences are barely separated from each other. In contrast, in the nuclear genome of eukaryotes, DNA encoding RNAs may account for only ~5% of the genome.

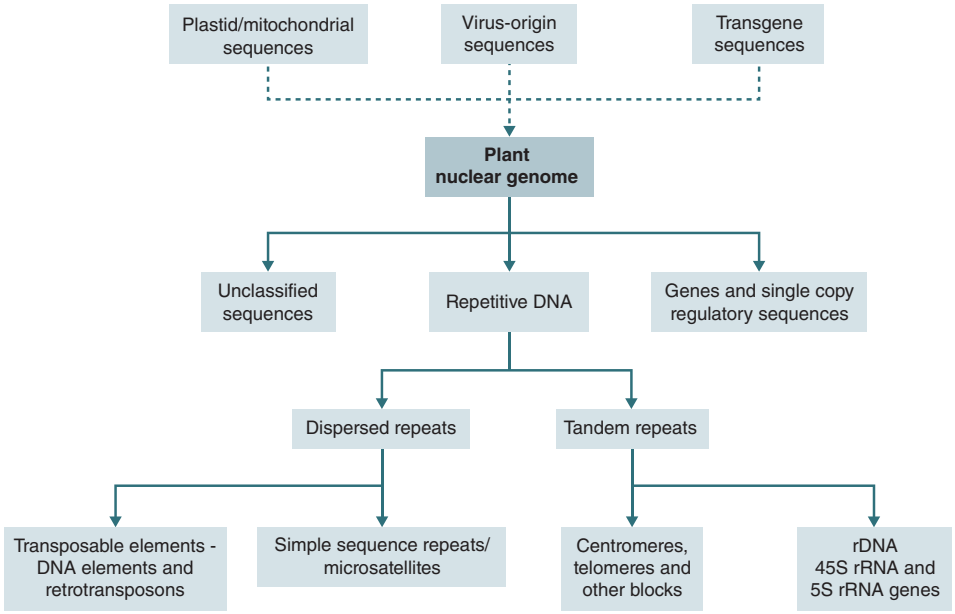
Individual plants and plant species differ in the amount of DNA in their genomes. DNA amounts are measured either as picograms (pg) of DNA per cell, or in numbers of base pairs (bp). Genome size is variable, with the largest genomes reported from the monocot *Paris japonica* (Melanthiaceae) (152.23 pg), and the smallest from the eudicot *Genlisea margaretae* (Lentibulariaceae) (0.063 pg) (Table 1.1).

Despite the variation in genome size, the number of protein-coding genes is surprisingly constant; in land plants it varies from approximately 22 300 for the lycophyte *Selaginella* (Banks *et al.*, 2011) to 35 900 for the angiosperm crop, maize (Schnable *et al.*, 2009). The difference in genome size thus cannot be accounted for by the genes themselves, but rather has to do with the size of the space between the protein-coding genes.

As can be seen from Table 1.1, the density of protein-coding, non-TE (non-transposable element) genes in the genome must be remarkably different among species. For example, while *Oryza sativa* (rice)



(a)



(b)

Figure 1.9 Types of sequences in the nuclear genome. (a) Genome sizes for flowering plants for which genome sequences are available. Pie charts show the relative amounts of DNA from transposable elements (TE-DNA) versus non-transposable elements (non-TE DNA). The linear relationship between the log of the genome size and the amount of transposable elements in the graph on the right shows that most of the difference in size is due to the difference in TE content. Tenaillon *et al.* (2010). Reproduced with permission of Elsevier. (b) DNA in the nucleus includes not only nuclear genes but also sequences from other sources such as organelles, viruses or transgenes inserted by humans. Nuclear DNA can then be classified into various categories. (b) Adapted from Heslop-Harrison and Schmidt (2007)

has about 39 000 protein-coding, non-TE genes spread across 400 Mbp of DNA, sorghum has about 33 000 in 800 Mbp of DNA, and maize has about 39 000 genes in a 2300 Mb genome. Clearly the difference must be the size of the "spaces" between the genes.

The word "spaces" is in quotes because in fact there are plenty of genes and regulatory sequences outside the protein-coding genes (Figure 1.7). Over the last decade it has become clear that this DNA (sometimes formerly dismissed as "junk") is a dynamic and active part of the genome and is as important for organismal function as the protein-coding fraction. Many of the discoveries about non-protein-coding genes have been made in plants.

Much of the space between the protein-coding non-TE genes is a complex mixture of repetitive sequences and transposable elements (Figure 1.9a). As noted above, transposable elements are mobile components of the genome with a propensity for creating repetitive bits of DNA sequence. The nature and dynamics of the transposable elements are discussed in the next chapter.

Other repetitive sequences are known as satellite DNA, which falls into several size classes (Figure 1.9b). Satellite DNA, as originally identified, corresponded to a set of repeats between about 150 and 180 bp now known to be located in and around the centromere. Minisatellites were described as repetitive sequences 10–100 bp long. Microsatellites are even smaller, generally consisting of repetitive units 2 or 3 bp long (e.g., ATATATATATAT); these are also known as Simple Sequence Repeats, or SSRs, and are widely used as markers for genetic mapping and for studies of population genetics. Most commonly, plant microsatellites are made up of A's and T's, whereas arrays of G's and C's are more common in animals. In all repetitive sequences, but particularly in microsatellites, the most common mutation is a change in the number of repeats, a process that is caused by replication slippage (Bhargava and Fuentes, 2010) (Figure 1.10). During DNA replication, the DNA polymerase continually dissociates from and reattaches to the DNA strand. Normally this does not create problems, but in a repetitive region the polymerase sometimes reattaches

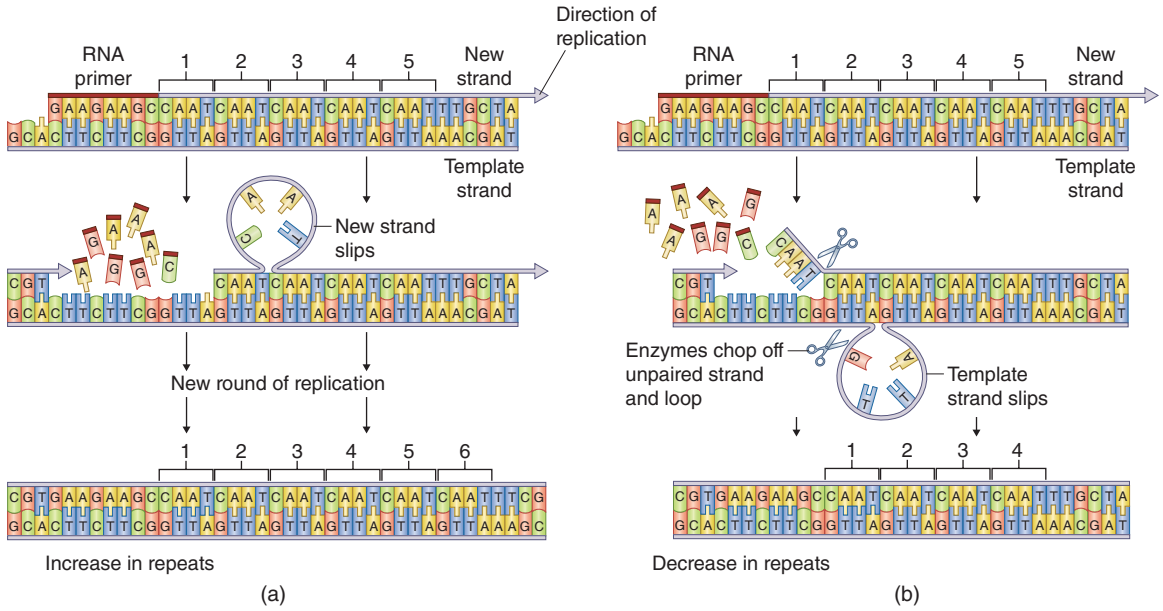


Figure 1.10 Replication slippage. DNA strands with short repeated sequences often lead to mistakes in replication. In this example, the sequence GTTA is repeated 5 times on the template strand. (a) If the new strand dissociates and reanneals such that the copy of repeat 1 anneals with the template for repeat 2, a loop of DNA is created in the copy strand, and repeat 1 on the template strand appears not to be copied. DNA polymerase copies repeat 1 again and the entire set of repeats is lengthened by one. (b) If the new strand slips and reanneals such that the copy of repeat 2 anneals with the template for repeat 1, a loop is formed in the template. This is recognized by DNA repair machinery and is removed; the set of repeats is shortened by one. Modified from Moxon and Willis, 1999

out of register, creating a loop or bubble in the DNA. Repair of this loop results in either loss or gain of one or more repeat units. The mutational process is not simple, and often microsatellites gain or lose several repeat units at once, rather than losses or gains of one at a time.

1.6 Genomic DNA is packaged in chromosomes

Chromosomes have been studied in plants since the nineteenth century. At the time, newly developed stains allowed microscopists to see for the first time colored bodies in the nucleus of some cells. The term chromosome is based on the Greek words *chroma* (color) and *soma* (body) to refer to these newly discovered structures. Once it became clear that the number of colored bodies was constant within an organism and often within a species, the chromosome count became a central piece of biological information. The number of chromosomes can be recorded either for the haploid gametes, in which case the number is referred to as n, or for the diploid (sporophyte) phase, in which case the number is $2n$. Thus for example in rice, $n = 12$ and $2n = 24$.

While the number of chromosomes is generally constant within a species, occasionally "extra" chromosomes appear in some plants but not others. The extra chromosomes are called **B-chromosomes**, or supernumerary chromosomes; they have no discernible effect on growth and development. B-chromosomes are smaller than normal chromosomes, and consist almost entirely of densely packed chromatin (known as heterochromatin). B-chromosomes do not segregate normally at meiosis, but instead accumulate preferentially in the cells that will become the gametes. Studies in maize find that pollen grains from a single plant can have from 0 to 20 B-chromosomes.

1.7 Summary

The structure of plant DNA is similar to that of all living things. The DNA molecule is a double helix, built of two antiparallel strands that are held together by hydrogen bonds. Methylation of cytosine residues is common and is now recognized as a structural feature that influences the function of the genome. The nucleotide composition of the DNA molecule determines its melting temperature, a parameter that is important for many biotechnology applications. The negative charge on the DNA can also be exploited for separation of molecules, and the ability of certain dyes to insert themselves into the double helix can be used to make the molecules visible to the human eye.

Plant DNA is found in the nucleus, the mitochondria and the chloroplasts, the latter two being ancient endosymbionts with genomes similar to those of their bacterial ancestors. Nuclear DNA is organized into linear chromosomes which have internal centromeres and are terminated by telomeres. The number of chromosomes varies between species, and chromosome number and genome size are not necessarily related. DNA provides all the instructions necessary to make RNA, from when the RNA is to be made to what protein will be encoded by the fully processed and mature form of the RNA. We define a gene as the full set of instructions for making a particular type of RNA. While much of the literature focuses on genes that will make protein-coding mRNAs, these are only a small component of the nuclear genome. Much of the rest of the genome is made up of genes that create proteins that are parts of transposable elements.

Genome size is variable among plants, even though the number of non-TE, protein-coding genes is fairly similar. Most of the differences in genome size are caused by differences in the number of transposable elements. Genome size also varies greatly in evolutionary time, so that closely related species may not have similar genome sizes.

1.8 Problems

1.1 You should be able to define the following terms: chromosome, centromere, transposable element, chloroplast, mitochondrion, methylation, bisulfite sequencing, microsatellite.

1.2 Calculate the approximate melting temperature of a DNA molecule with the sequence ATGGGCATAGCCGA.

1.3 Why is the value you calculated in Problem 1.2 an oversimplification?

1.4 What are two ways to determine which sites in a DNA molecule are methylated?

1.5 What is a gene? Show the architecture of a typical protein-coding gene in a diagram.

1.6 Draw the structure of a simple double-stranded DNA molecule with one strand having the nucleotide content of adenine and guanine.

1.7 How might DNA methylation perturb gene expression? And DNA replication?

1.8 Introns are removed from initial RNA transcripts with great fidelity. Can you come up with a design of DNA that could account for this fidelity? What features do you think would be necessary in the DNA to allow for this fidelity?

References

Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Banks, J.A., Nishiyama, T., Hasebe, M. *et al.* (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.

Bhargava, A. and Fuentes, F.F. (2010) Mutational dynamics of microsatellites. *Molecular Biotechnology*, **44**, 250–266.

Brown, T.A. (2002) *Genomes*, 2nd edn, Wiley-Liss.

Campbell, N.A. and Reece, J.B. (2007) *Biology*, 7th edn, Benjamin Cummings.

Goff, S.A., Ricke, D., Lan, T.-H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, **296**, 92–100.

Heslop-Harrison, J.S.P. and Schmidt, T. (2012) *Plant Nuclear Genome Composition*, John Wiley & Sons, Ltd.

Hyatsu, H., Shiraishi, M. and Negishi, K. (2008) Bisulfite modification for analysis of DNA methylation, in *Current Protocols in Nucleic Acid Chemistry*, John Wiley & Sons, Ltd, Unit 6.10.

Merchant, S.S., Prochnik, S.E., Vallon, O. *et al.* (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.

Moxon, E.R. and Wills, C. (1999) DNA microsatellites: agents of evolution? *Scientific American*, **280**, 94–99.

Paterson, A.H., Bowers, J.E., Bruggmann, R. *et al.* (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, **457**, 551–556.

Raven, P., Johnson, G.B., Mason, K.A. *et al.* (2011) *Biology*, 9th edn, McGraw-Hill.

Reece, J.B., Urry, L.A., Cain, M.L. *et al.* (2011) *Campbell Biology*, Benjamin Cummings.

Rensing, S.A., Lang, D., Zimmer, A.D. *et al.* (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.

Rumpho, M.E., Worful, J.M., Lee, J. *et al.* (2008). Horizontal gene transfer of the algal nuclear gene *Psbo* to the photosynthetic sea slug *Elysia chlorotica*. *Proceedings of the National Academy of Sciences USA*, **105**, 17867–17871.

Schmidt, T. and Heslop-Harrison, J.S. (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends in Plant Science*, **3**, 195–199.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Tenaillon, M.I., Hollister, J.D. and Gaut, B.S. (2010) A triptych of the evolution of plant transposable elements. *Trends in Plant Science*, **15**, 471–478.

Tuskan, G.A., DiFazio, S., Jansson, S. *et al.*, (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, **313**, 1596–1604.

Velasco, R., Zharkikh, A., Troggio, M. *et al.* (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One*, **2**, e1326.

Yu, J., Hu, S., Wang, J. *et al.* (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.