Genomes

2

OM2.1 THE CELL NUCLEUS

The nucleus is the location of most of the genome of the cell. Enclosed within the nuclear membrane, the so-called "**nuclear envelope**," are large numbers of proteins and nucleic acids in a complex three-dimensional structure. DNA replication, gene transcription, and RNA processing occur within this organelle. Genes are transcribed by RNA polymerases, and the resulting RNA molecules are transported from the nucleus to the cytoplasm through **nuclear pores** – "molecular gates" in the nuclear envelope. The nuclear pores are also the routes by which proteins are imported from the cytosol to the nucleus. Imported proteins include RNA and DNA polymerases, histones, and transcription factors.

The nucleus comprises several distinct subdomains where the different activities carried out by the imported proteins take place (**Figure OM2.1-1A**). The genome is packaged together with proteins within different forms of chromatin. In most plants, the DNA of chromosomal centromeres and telomeres, and chromosomal regions rich in repeated DNA sequences rather than genes, are packed tightly within heterochromatin. Gene-rich regions of the chromosomes, where active transcription is occurring, are contained in euchromatin, which is less condensed and accessible to the transcriptional apparatus. The **nucleoplasm** is the fluid that surrounds the chromatin, where molecules such as proteins and RNA can diffuse and interact with each other.

The nucleolus is the site of transcription of genes that encode ribosomal RNA (rRNA) (**Figure OM2.1-1B**). These genes are arranged in tandem along a region of the chromosome called the **nucleolar organizing region**. Different parts of the nucleolus have distinct roles in the formation of ribosomal subunits. Transcription of the rRNA genes by RNA

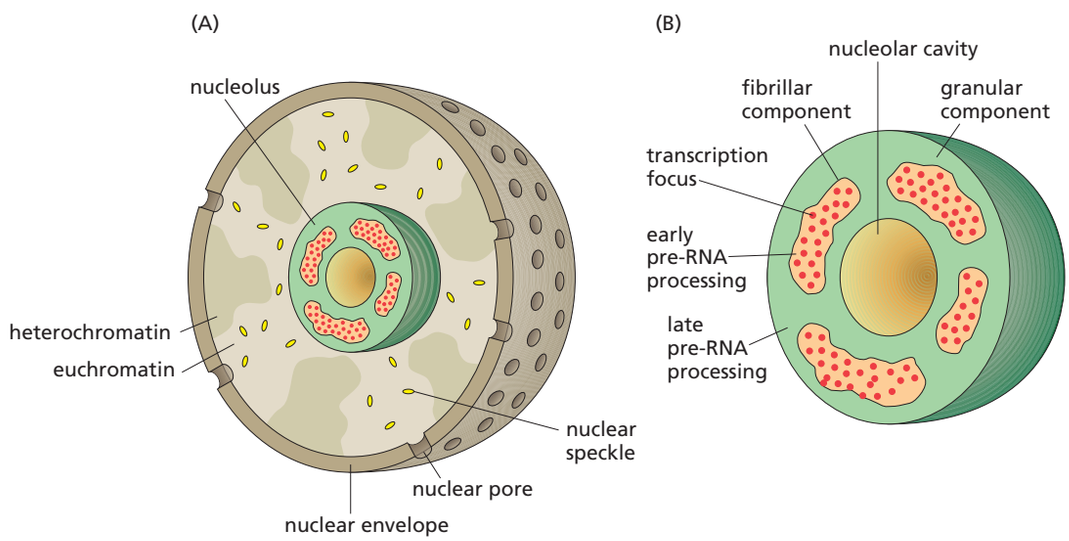

Figure OM2.1-1 Nuclear structure. (**A**) Internal structure of the nucleus showing the structures and subdomains discussed in the text. (**B**) Structure of the nucleolus in greater detail showing the sites of ribosomal RNA synthesis and processing (see also Figure 2.10).

polymerase I occurs at numerous foci within regions of decondensed chromatin known as the "dense fibrillar" component. Transcription produces a pre-RNA molecule (the 45S RNA), which is then processed to mature 18S, 5.8S, and 28S components of the ribosome. The first processing steps take place in the dense fibrillar material. The pre-RNA molecule moves to another region of the nucleolus, the "granular" component, in which the remainder of the processing occurs. The mature RNA molecules then associate with ribosomal proteins that have entered the nucleus, and the resulting ribosomal subunits are exported to the cytosol.

Smaller nuclear structures are thought to provide scaffolds for the assembly of complexes and small RNAs. As described in this chapter, newly made (*de novo*) transcripts of most protein-coding genes contain sequences called introns, which must be spliced (removed) to form mature transcripts that can be exported to the cytosol for translation into proteins. Introns are spliced by nucleases located on a particle called a **spliceosome**. The spliceosome complex contains five different small RNA molecules and at least 200 proteins. It is thought the complex is assembled in the **nuclear** "**speckle**" regions.

OM2.2 METHODS FOR INVESTIGATING GENE FUNCTION IN PLANTS

As described in this chapter, the rapid advances in DNA sequencing technology developed in the past 15 years have provided plant biologists with an unprecedented ability to determine the complete sequence of any plant genome rapidly and relatively cheaply. Coupled with significant advances in bioinformatics and advanced computational power, this information can be readily analyzed and the full complement of genes encoded within the genome discerned; putative functions can then be assigned to many of the predicted gene products. While it may be possible to determine the general nature and mode of action of the protein predicted by *in silico* translation (e.g. cellular growth and metabolism, signaling, or defense), the precise role of these genes in plant growth and development may not be obvious. For example, higher plants contain large numbers of **multigene families** such as transcription factors that control the expression of different genes at defined stages of the life cycle. Ultimately, plant biologists aim to determine the function of all genes in the genome and how gene networks function in different processes throughout the plant life cycle.

As in all areas of biological research, "model organisms" have been invaluable in enabling researchers to dissect specific aspects of the organism's biology, which can then be applied more generally to other organisms of that type. In higher plants several model systems have proved popular historically (e.g. maize, rice, tobacco, and tomato), each one chosen for its unique characteristics. The most popular model organism used in plant biology research is *Arabidopsis thaliana*. It has many features that make it an attractive model including its simple diploid genetics, a rapid life cycle allowing many generations of plants to be analyzed within a short timeframe, and its small size allowing large numbers of plants to be grown in a limited space.

This section discusses the strategies that plant biologists have used traditionally to identify genes and their function, and more recent technological advances available to dissect gene function affecting different aspects of plant form and function.

Forward genetics

One of the most widely exploited strategies to identify gene function in plants is **mutagenesis** using a strategy known as **forward genetics** (**Figure OM2.2-1**). In this strategy, random mutations are introduced into the genome using mutagens such as **ethyl methanesulfonate (EMS)**, a DNA alkylating agent that induces single base changes, or fast neutrons (FN), which induce small deletions in DNA. Imbibed seeds of Arabidopsis (which are multicellular) are first exposed to these mutagens at sublethal doses. As these mutagens act randomly each cell in the seed will contain a unique set of mutations. Since Arabidopsis is diploid (containing two copies or alleles of each gene) it is most unlikely that both alleles of any gene will be mutated simultaneously so relatively few effects are seen in the first generation of plants that develop following this treatment. Most often this strategy creates loss-of-function mutations in single alleles, so mutant phenotypes will only be observed in the **homozygous** state, i.e. in subsequent generations.

Individual cells that contain a unique set of mutations may divide to form clonal sectors in the developing plant that give rise to spores that form the male and female gametes. Following **self-pollination** of large numbers of mutagenized plants (hundreds to thousands in a typical forward screen), a proportion of the progeny obtained from individual plants may exhibit an interesting mutant phenotype (e.g. aberrant flower or leaf development, altered leaf starch turnover, loss of pathogen resistance, etc.) from which genes controlling important processes can potentially be identified. Controlled **backcrosses** of mutants of interest to wild-type plants are then performed over several

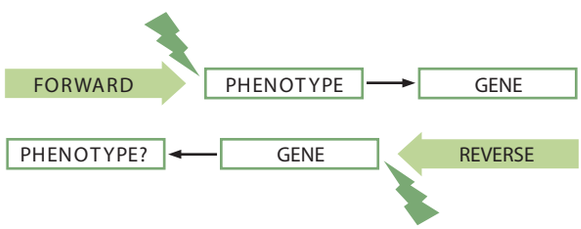

Figure OM2.2-1 Forward and reverse genetic approaches for investigating gene function. In forward genetic strategies, seeds are exposed to mutagenic agents (*green lightning bolt*) that induce gene mutations in a random manner. If an interesting phenotype is recognized in the next generation, and shown to be the result of a single gene, it can be isolated using a number of strategies. Reverse genetic strategies are more precisely targeted and focus on individual genes of potential interest. A number of technologies may be used to reduce target gene expression or inactivation including RNA silencing and gene editing, respectively. The effect on gene function can then be determined in plants containing the silenced or inactivated gene.

generations to confirm the stability of the mutant phenotype and to remove the numerous **unlinked** random mutations located around the genome following mutagenesis. The gene responsible for the mutant phenotype can then be targeted for cloning and sequencing using the "positional cloning" strategy described in the next section.

Gene isolation using positional cloning

Once a mutation conferring a particular phenotype is identified, and confirmed by genetic analysis to be due to a single gene mutation (i.e. a predictable proportion of wild-type and mutant individuals are present in the progeny from a controlled genetic cross), a strategy for gene isolation can be used. The positional cloning strategy is dependent on the availability of a high-density molecular **genetic map** of the organism (see this chapter and Section 8.2). Genetic maps are widely used in research and agriculture for numerous purposes. Crosses are performed between different **genotypes** of mutant and wild-type plants that are polymorphic for molecular markers across the whole genome. An F_2 population is created that segregates for the wild-type and mutant phenotypes in predictable Mendelian ratios, and also for genome-wide polymorphic **molecular markers**. By analyzing individual recombination events in large numbers of mutant progeny with genome-wide markers it is possible to identify markers most closely linked to the mutation and therefore to define the chromosomal location of the mutant gene more precisely. The markers that flank the target gene define a chromosomal region harboring candidate genes that can be readily identified from whole genome sequences. Identification of the gene responsible for the mutant phenotype can then be confirmed in a number of ways. These include comparison of the sequences of candidate genes in the original wild-type and mutant plants, and introduction of a wild-type copy of the gene into mutant plants to discover whether it can restore the wild-type phenotype. This can be achieved using transformation, most often with established transformation technology based on *Agrobacterium tumefacien*s (see Online Material 14.6).

So far, a strategy that exploits variation induced experimentally through mutagenesis has been described. However, plants also exhibit natural intraspecific variation (e.g. resistance to specific pathogens, response to abiotic stresses, etc.; see this chapter and Chapters 12, 13, and 15). When this variation results from single gene variants, the gene responsible can also be isolated using the positional cloning strategy described earlier. For example, *A. thaliana* exhibits a wide geographic range (from the Arctic to the Cape Verde Islands). Single gene traits that confer local adaptation of different populations (sometimes termed **ecotypes**) have been isolated and characterized. This has provided valuable insight into how genetic variation enables plants to adapt to diverse environments.

Reverse genetics

An alternative to the non-targeted approach to identifying gene function by forward genetic screens is a more targeted strategy; a **reverse genetics** approach (see Figure OM2.2-1). In addition to the rapid developments in genome sequencing and bioinformatics, researchers benefit from technologies that allow the introduction of defined mutations into single genes within the genome. In contrast to the forward genetics strategy that produces plants harboring multiple, randomly induced mutations, reverse genetics introduces precise mutations into single, user-defined target genes. The effects of such mutations on plant function can then be deduced in subsequent generations. These technologies may either induce down-regulation of mRNA expression in a target gene, or create loss- or gain-of-function mutations.

The cellular role of a target gene can potentially be determined from the effects on the plant of gene silencing, a collection of strategies developed from understanding of the cellular processes that regulate the level and translation of mRNAs levels and their translation (see Chapters 2 and 3, and Section 14.4). However, these technologies have several limitations, including their failure to suppress completely the expression of the target gene.

The most popular and versatile technology developed in recent years for introducing targeted gene mutations is genome editing. Of the three main genome editing technologies, the most versatile is known as CRISPR (which stands for clustered regularly interspaced short palindromic repeats). **CRISPR** technology was developed from the discovery of a system in bacteria and archaea that confers immunity to infection by bacteriophages (small viruses that infect these prokaryotic organisms). It is proving to be a versatile tool for studies of gene function in a wide range of organisms, and for applications in medicine including somatic gene therapy.

All genome editing technologies function by introducing a targeted double-stranded break in chromosomal DNA at a defined location within the target gene. CRISPR technology comprises a ribonucleoprotein complex that can be delivered into plants by a number of methods including *Agrobacterium* transformation. The RNA component is termed a single-guide RNA (sgRNA) and is comprised of two domains, one for binding to the protein component of the complex that has intrinsic nuclease activity, and a second domain of typically 21 bp that are complementary to the user-defined target gene sequence. When the RNA component of the complex anneals to its complementary target sequence in chromosomal DNA through conventional base pairing, the nuclease component of the complex then induces a double-stranded break.

Following double-stranded break formation, the natural cellular DNA repair processes catalyze the rejoining of the broken ends via one of two pathways. In the process of non-homologous end joining, the broken ends of DNA are repaired and rejoined in an imprecise way which can result in the deletion or addition of extra bases at the break site. If this happens within a protein-coding exon of a gene sequence it may induce a number of different types of mutation, most commonly a "**frameshift**" resulting in premature stop codons in the corresponding mRNA. This strategy is commonly used for inducing loss-of-function mutations in the target gene. Alternatively, the broken ends can be accurately repaired by another pathway known as homology-dependent repair. In homology-dependent repair the cell uses an identical double-stranded DNA template to repair the broken chromosome accurately. For example, if a break occurs after the **S phase** in the **cell cycle**, the replicated sister **chromatid** of the broken chromosome (see Section 8.2) is used as the repair template. For functional gene investigation purposes where a precise alteration to the gene sequence is required, homology-dependent repair can be exploited following a CRISPR-induced double-stranded break if a double-stranded DNA template containing the desired sequence is delivered to the target cells along with the CRISPR components. However, altering target gene sequences through this pathway is currently challenging in plants and awaits further developments before it can be used routinely. For this reason, genome editing has been primarily utilized for creating loss-of-function mutations.

Qualitative versus quantitative traits

Phenotypic characters conditioned by single genes may be referred to as qualitative traits. Qualitative traits are usually conferred by single **dominant genes**, i.e. in diploid plants only a single copy of the wild-type allele is required for the phenotype to be fully expressed so homozygous wild-type and heterozygous plants are phenotypically indistinguishable. Because of their dominant nature, these genes give rise to progeny that exhibit **discontinuous variation** in controlled genetic crosses. For example, consider one of Mendel's classic studies in *Pisum sativum* on the gene determining height (the "*Le*" gene). In a cross between a **pure-breeding** tall plant (genotype *Le*/*Le*) and dwarf plants homozygous for the mutant allele (genotype *le*/*le*), the F_1 progeny were self-pollinated to produce an **F_2 generation**. The F_2 progeny were either completely tall (genotypes 1/4 *Le*/*Le*, 1/2 *Le*/*le*) or dwarf in stature (genotype 1/4 *le*/*le*) in a phenotypic ratio of 3 tall : 1 dwarf. Only tall or dwarf progeny are observed with no plants of intermediate height (thus, discontinuous variation). A further example of a pea gene in this category – the *R* or *Rugosus* gene that determines seed shape – is described in the context of starch synthesis in Online Material 5.5. Genes in this category can be isolated and characterized using the positional cloning strategy described earlier.

Many traits have a more complex genetic basis that involves the interaction of multiple genes. These are referred to as quantitative traits, and the chromosomal regions where they are located as **quantitative trait loci (QTLs)**. Quantitative traits are often strongly influenced by the environment, so the phenotypic expression of these traits (p) is expressed as $p = g \times e$, where g is genotype and e is environment. As a simple example of a complex quantitative trait consider human height, which is determined by many interacting genetic factors and varies considerably due to an individual's environment, such as their health, nutrition, and other socio-economic factors. For this reason, height shows a wide range of values in the population in the form of a dumbbell-shaped normal distribution, i.e. it shows **continuous variation**.

In plants many traits, including ones relevant to crop plant improvement, are quantitative in nature. An example is grain yield in cereals. When the progeny of crosses derived from a "high-yielding" and "low-yielding" variety are analyzed, they exhibit a continuous distribution of yield levels across the range defined by (and sometimes exceeding) the extremes defined by the two parent varieties. This is due to the independent assortment during meiosis of multiple genes that affect yield. The progeny

from such crosses contain distinct combinations of the QTLs involved (see Section 8.2). Each QTL that contributes to the high-yield trait may influence it to a different extent (i.e. each QTL has a different quantitative effect, and may be classified as either a "major" or "minor" QTL). If the segregating progeny are each assigned a quantitative score for yield, these data can be analyzed in combination with genome-wide genotyping scores of the individual progeny. Statistical analyses of the data will identify molecular markers associated with specific QTLs and define their chromosomal locations. By utilizing carefully constructed genetic crosses the individual genes underlying the different QTLs can be isolated using the positional cloning strategy described. Where identification of the gene underlying a QTL is not essential, for example in a crop plant breeding program, it can instead be followed in segregating progeny using the molecular markers that flank it. This process, termed **marker-assisted selection**, has important applications in the crop improvement industry.

OM2.3 READING GENOME ANNOTATION

Genome annotation utilizes large amounts of information to determine the function and organization of plant genome sequences, including gene locations, the various sequence elements and their functions, mRNA transcripts, their encoded protein sequences, and in some cases gene regulatory sequences. As rich sources of information, annotated genomes have become essential tools in biological research. To make these tools effective, it is necessary to organize the information in a way that is easily accessible, following conventions that are recognized by researchers across species and disciplines.

Access to genome annotation is typically mediated by remote access to genome browsers in public databases. The genome is usually represented by lines corresponding to each chromosome, and researchers can zoom in to regions of interest to reveal the position and structure of genomic features, such as genes, transposons, and the underlying sequence. The annotated features are given unique codes, which normally start with the species initials, for example, "*At*" for *Arabidopsis thaliana* or "*Os*" for *Oryza sativa* (rice). The initials are followed by a chromosome number, the letter "g" (genomic sequence) and a number that corresponds to the order of annotated features along the chromosome. When the genome is viewed at a scale where individual genes become visible, their structures are represented by boxes and lines, corresponding to exons and introns, respectively. As an example, **Figure OM2.3-1** shows how a section of the Arabidopsis genome is presented in a typical genome browser.

Gene names and their corresponding graphs are associated with hyperlinks, which open new pages with detailed information about each gene, such as other names (often attributed based on features of the corresponding single gene mutant that provide clues to its normal function), links to literature describing gene function, features of the encoded proteins, etc. Hyperlinks are also used for downloading the relevant sequences for more detailed analysis. Gene names and their detailed descriptions can also be searched using keywords, so the genome information can be accessed either through its physical location, or through specific features of interest.

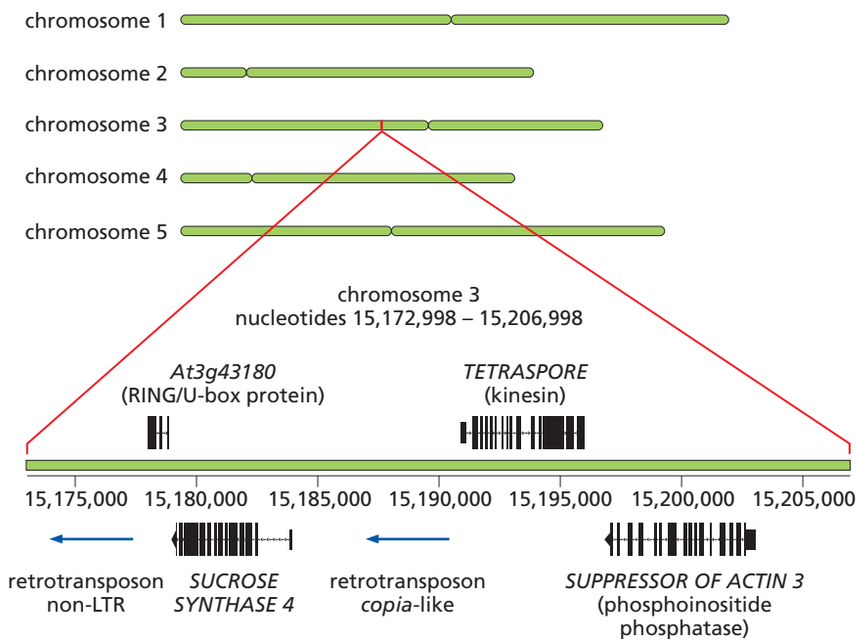

Figure OM2.3-1 Close-up view of a small region of the Arabidopsis genome from a genome browser. The five Arabidopsis chromosomes are represented as *green* bars, with constrictions corresponding to centromeres. A small section of chromosome 3 (i.e. *At3g*) is shown in the magnified view below. Immediately below this section are nucleotide coordinates along the chromosome in base pairs. Above and below the *green* bar, sets of *black* boxes and lines correspond to genes and *blue* arrows correspond to transposons. Within the genes, tall black boxes represent exons, which are separated by thin lines corresponding to introns. Shorter black boxes correspond to mRNA sequences that precede or follow the protein-coding sequences (i.e. the 5′ UTR and 3′ UTRs, respectively). Genes above and below the green line are transcribed right-to-left and left-to-right, respectively. Gene names are in italics, followed by the predicted function of the encoded protein (in brackets).

OM2.4 RECOMBINATION IN PLANT mtDNA CAN PRODUCE A RANGE OF SUBGENOMIC PRODUCTS

Plant mtDNA can contain a large number of repeated sequences in both direct and indirect orientations. The annotations in **Figure OM 2.4-1** are similar to those used in Figure 2.25 except that R shows direct repeats. A number of distinct homologous recombination events are possible, depending on which repeats are involved. Each event generates distinct subgenomic products. Plant mtDNA may thus be present in multiple forms within mitochondria, depending on the number and relative orientations of repeat elements within the genome, which can vary significantly between species.

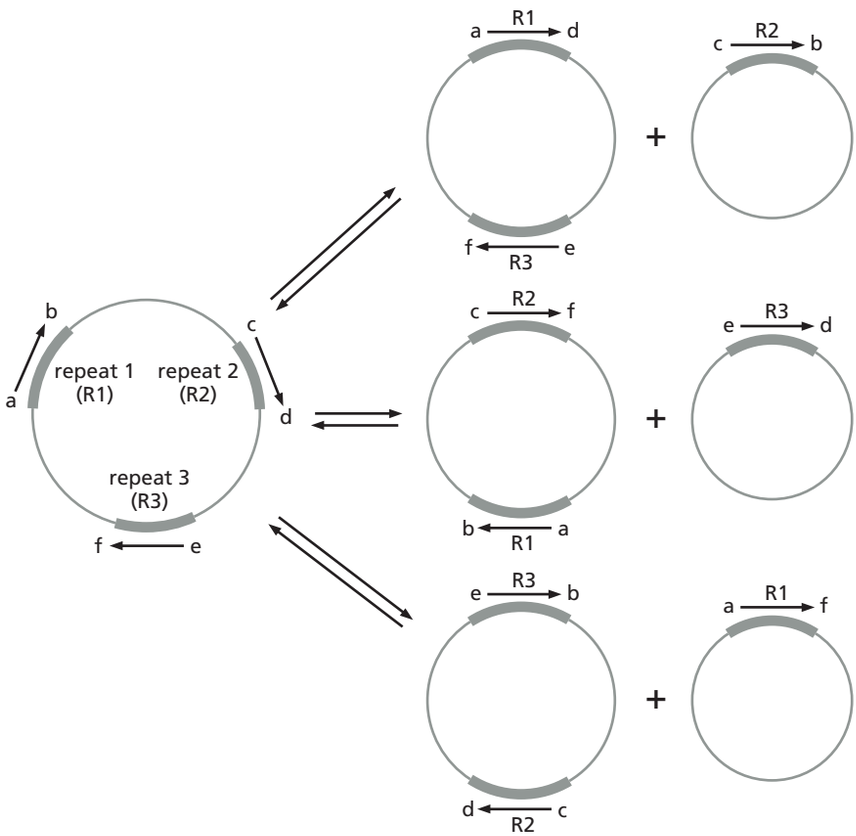

Figure OM 2.4-1 **Recombination in plant mtDNA.**