

GENOMES AND THE FLOW OF BIOLOGICAL INFORMATION



INTRODUCTION

The biological world is a maze of interconnections on a multitude of scales: atoms join to form molecules – molecules cluster to form cells – cells interact to form tissues – and tissues aggregate to form an organism. The interplay continues beyond the organism, as organisms form populations, populations inhabit ecosystems, and ecosystems unite to form the world around us.

Through all the layers of connectivity runs a common thread: the communication and onward passage of information – from cell to cell, from organism to organism and, ultimately, from generation to generation. This information is stored, at the most fundamental level, in each living cell in our body. But how does this information – no more than a static repository of data – come alive to direct the processes that constitute life?

The answer lies in the concerted action of molecular components, which cooperate in a series of ingenious processes to bring the information deposited in each of us, in our **genome**, to life. These components and processes lie at the heart of molecular biology, and are the key players in the chapters that follow.

Before embarking on our journey to discover these components and processes, however, we begin by considering some of the fundamental biological themes on which the study of molecular biology relies. You may be familiar with some of these themes, and the concepts and ideas introduced, from earlier studies. If so, look on this chapter as a refresher before later chapters lead you into more unfamiliar – but ultimately fascinating – topics.

1.1 THE ROOTS OF BIOLOGY

The diversity of life is unified by some common themes

Life on earth is remarkably diverse, ranging from the relatively simple unicellular organisms, such as bacteria, to more complex multicellular organisms, such as plants and humans. Despite their diversity at the macroscopic level, where distinctive features of organisms are readily apparent with the naked eye (Figure 1.1), the core molecular features of all organisms are remarkably similar. Indeed, we can think of life on Earth as being unified by a number of key themes.

- First, a living organism must be distinct in some way from its environment, that is, it must be defined by a physical barrier, which serves to separate 'organism' from 'environment'. This separation allows the internal environment of an organism to be carefully regulated, and ultimately distinguishes 'self' from 'non-self'.
- Second, a living organism must be able to store information in a stable way, and also have a way of using this information to determine its characteristic features – its structural composition, how it functions, and so on.
- Third, a living organism must be able to reliably replicate and pass information from one generation to the next. **Replication** lies not only at the heart of the propagation of a **species** from generation to generation, but also at the heart of the growth and repair of a multicellular organism within a single generation. Without replication, not only could an organism not reproduce to yield



Figure 1.1 The life forms on earth are diverse and interconnected. (a) A monarch butterfly on a milkweed plant (Louise Murray/Science Photo Library), (b) bacteria on the scales of an insect, (c) a California Alligator Lizard carrying multiple ticks (Dante Fenolio/Science Photo Library), (d) a clownfish surrounded by the tentacles of a sea anemone (Andrew J. Martinez/Science Photo Library), (e) an oxpecker bird sitting on an African buffalo (Dr P. Marazzi/Science Photo Library), and (f) a fungus on a tree (Ron Bass/Science Photo Library).

offspring, but a multicellular organism could not develop from an embryonic state to a fully formed individual. The processes of replication and transmission must remain faithful to the entity being replicated, just as a recipe – itself an information store – ensures that a given dish can be prepared in a consistent way time after time.

- Finally, living organisms require a source of energy from their surroundings to grow and reproduce; this energy is used to drive the biological processes that keep an organism alive. There are substantial differences in how energy is harvested by various organisms, for example, plants get their energy from the sun whereas humans get energy from food. Nevertheless, the core molecular mechanisms used by cells in all organisms to grow and propagate are remarkably similar (that is, they are highly conserved). For example, the same metabolic cycles are used in all organisms to transform sugars into the energy currency of the cell, adenosine triphosphate (ATP).

In modern-day organisms, the physical barrier that separates individuals from their surroundings (and so defines ‘self’) is a lipid-based membrane, while a molecular species known as **nucleic acid** stores the biological information, and allows for ready replication of this information (thus conferring ‘self’ identity). These two basic molecular features come together – the lipid membrane surrounding the nucleic acid – to form the cell, which is the building block of organismal life (Figure 1.2). Some organisms are composed of a single cell (unicellular) whereas others are formed of many cells (multicellular). However, the cell is not the only fundamental building block of life; common building blocks exist at the molecular level too.

Living organisms are constructed from common molecular building blocks

It is thought that the entire diversity of modern-day life derives from a common original life form often referred to as the last universal common ancestor (LUCA) or **progenote**, whose molecular components have been conserved from generation to generation, and from species to species, as different organisms have evolved. In support of this, when we compare modern organisms, we find that the core building blocks of all organisms are the same.

The four basic classes of molecules from which living organisms are constructed are (Figure 1.3):

- nucleic acids
- proteins
- lipids
- carbohydrates.

These molecules reflect a further unifying theme – of large biological molecules being constructed from smaller, repeated subunits.

Nucleic acids are the informational molecules and typically have a simple, linear structure composed of the monomer building blocks (the nucleotides) that are the instructions for life, and thus for the propagation of life. The information stored in

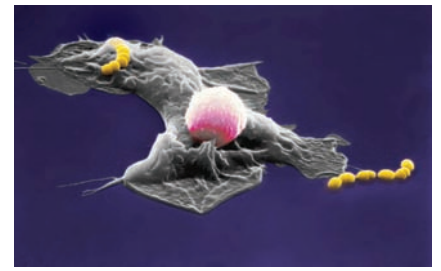


Figure 1.2 All cells share the same basic components. False-colored scanning electron micrograph showing chains of *Streptococcus* bacteria (orange) and two human cells: a macrophage (gray) and a lymphocyte (pink). Both of the human cells are part of the immune response, and act together to eliminate the bacteria. Despite their different sizes, all three types of cell share the basic components of a cell: a membrane that isolates each individual cell; a genome that is the ‘blueprint’ of each cell; a means to divide and reproduce; and a means by which to utilize energy. Electron micrograph courtesy of James A Sullivan.

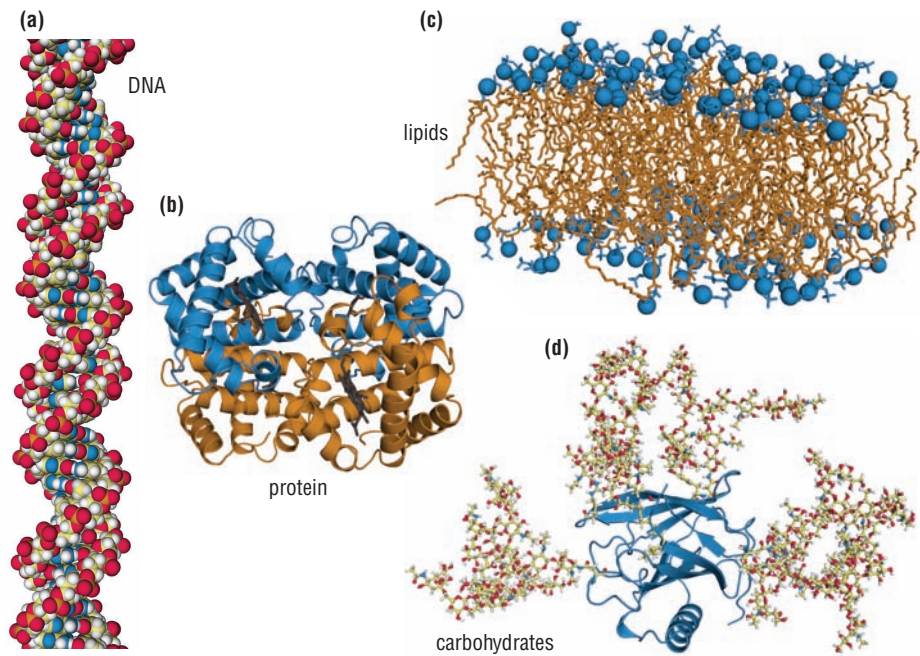


Figure 1.3 All cells are composed of four basic classes of molecules. (a) Nucleic acids (DNA and RNA) store information and serve as the interpreters of these instructions for life. (b) Proteins carry out most of the structural and functional tasks. (PDB 1A00). (c) ‘Water-fearing’ lipids form the membranes that comprise the physical barrier to the outside of the cell. (d) Carbohydrates provide energy and often adorn extracellular proteins.

these molecules directs the synthesis of proteins (a ‘coding’ role), and, in so doing, specifies the structure and function of the cell, as noted below. In addition, nucleic acids play essential non-coding roles within the cell as functional components of molecular machines and by regulating how the information they contain is used.

Proteins are also linear polymers, composed of distinct building blocks known as **amino acids** that typically fold into complex three-dimensional structures. Proteins are often referred to as the workhorses of the cell, carrying out most of the critical structural and functional tasks.

Lipids are ‘water-fearing’ molecules that self-assemble to form the membranes that present a physical barrier to the exterior milieu, thus defining self. Lipids are also constructed from repeated subunits – multiple fatty acid molecules join together to form the lipid membrane.

Carbohydrates are ‘water-loving’ molecules that play a wide range of roles in the cell, from providing energy to increasing the solubility of otherwise insoluble proteins. Carbohydrates continue to reflect the theme of large molecules being composed of smaller subunits: they are polymeric molecules constructed from simple sugar monomers.

➔ We learn more about the structure and chemical nature of these four classes of biological molecules in Chapter 2.

From these four components are built the macromolecular machines that are responsible for all cellular transactions – from copying the biological information stored in nucleic acids, to producing functional proteins, to harvesting energy from the environment.

The flow of biological information supports the maintenance of life

The title of this chapter refers to the flow of biological information – a phenomenon that is made possible through the components and processes that comprise ‘molecular biology’. As we have mentioned above, storage and passage of

information is one of the unifying characteristics of living organisms. But why is the flow of biological information necessary? Ultimately, a flow of information is needed to ensure that the offspring of an organism remains faithful to its parents in terms of the characteristics it exhibits. In other words, a parent must somehow act as the blueprint for its offspring. Without such a flow of information, a cat could not reliably give birth to another cat, or a human could not reliably give birth to another organism that was recognizably human.

The maintenance of life relies on the process of reproduction – both in the context of the growth of a specific tissue through rounds of cell division, and in the development of new offspring from a parent. In both cases, for the offspring to be a faithful reproduction of its parents, there must be a means of storing information (and using it), and a mechanism through which this information can be copied and inherited by the next generation.

Heritable information is stored in a simple polymer of monomer building blocks

Information in all cells is stored in a repeating polymer structure called deoxyribonucleic acid (DNA), which is easily copied and transmitted to subsequent generations through processes we will explore later in the book. DNA is a type of nucleic acid, and is composed of four building blocks known as **nucleotides**: guanosine (G), adenosine (A), thymidine (T), and cytidine (C). These nucleotides are strung together to form an extended linear polymer. The resulting strings of letters specify our genetic identity: they encode all functional molecules that must be synthesized for the cell and ultimately the organism to survive and propagate.

A particularly important chemical feature of these four distinct nucleotides is that the faces of A and T fit together, as do those of G and C, that is their shapes are complementary. When two different nucleic acids have complementary sequences – that is, where A on one strand is mirrored by T on another, and where C on one strand is mirrored by G on the other – the close association of these pairs of nucleotides allows the two sequences to form extended double-stranded structures with a characteristic double-helical conformation (Figure 1.4). This double-helical structure is the basic form in which the genetic information is stored.

Each DNA strand in a double helix has directionality: one end of the strand is denoted 3' and the other 5'. When two complementary strands come together they do so in an antiparallel fashion – that is, the 3' end of one strand aligns with the 5' end of the other, and vice versa. (In essence, the two strands lie 'head to toe'.)

Duplication of the information stored in DNA is facilitated by the simple complementary shape correspondence – a complementary, fully informational strand can be made by separating the two strands and using the sequence of nucleotides in each of the two original strands to direct the synthesis of two new strands, as depicted in Figure 1.5. Such strand replication follows the same rules of complementarity described above: an A in one of the original strands directs the addition of a T in a new strand (and vice versa), and a C in the original strand directs the addition of a G in the new strand (and vice versa). It is important to



Figure 1.4 The DNA double helix is a repeating polymer that stores information in cells. DNA is composed of the nucleotide building blocks guanosine (G), adenosine (A), thymidine (T), and cytidine (C). The distinct feature that G pairs with C and A pairs with T allows a strand of DNA to be copied.

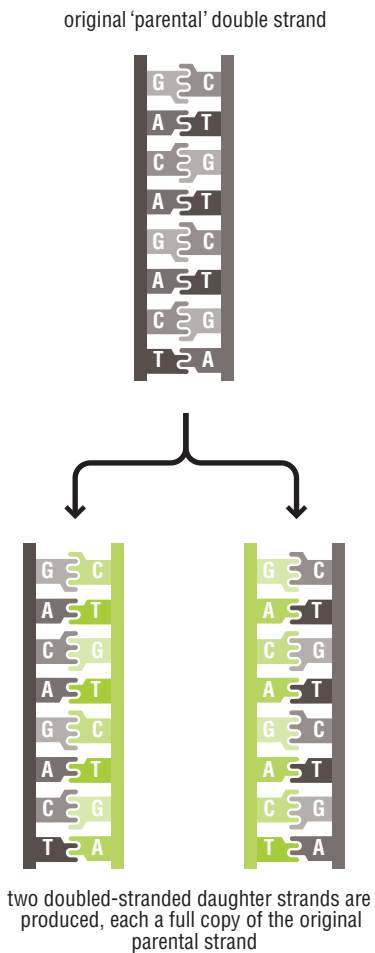


Figure 1.5 The process of replication is common to all organisms. The complementary nature of the two DNA strands allows the DNA to be copied.

note the word 'complementary' above: when a nucleic acid strand is replicated, the strand produced as a result is not identical to the original strand, but is complementary to it.

DNA is a very stable nucleic acid, making it the nearly ubiquitous material for the storage of information in modern biology. However, the chemically similar but less stable nucleic acid, ribonucleic acid (RNA), is occasionally used. (This does not mean that RNA is a mere bit-player on the biological stage: we will see throughout this book the myriad functional roles of RNA in the workings of an organism.)

1.2 THE GENOME: A WORKING BLUEPRINT FOR LIFE

A linear sequence of nucleotides in DNA comprises the genetic makeup of an organism, called the **genome** of an organism. This repository stores all information needed to specify cellular function and can be considered a blueprint for life. In the past decade, researchers across the world have determined the complete genome sequence of numerous organisms, from thousands of bacterial species to the human genome sequence. These approaches have generated great excitement in all areas of biology and medicine as analysis of these sequences has greatly increased our understanding of function and evolution.

Transmission and maintenance of the genome is essential for life

As we note above, reproduction by all organisms depends on their ability to make a copy of their blueprints for life to hand down to their progeny. At the cellular level, the copying of the genetic information (replication), the separation (or '**segregation**') of the copies, and their subsequent transfer into the cells of the next generation is required even if the recipient is to remain part of the same organism – for example, a new skin cell, or muscle cell. We refer to this copying and segregation process as the transmission of the genome. In molecular terms, this means that the nucleic acid, typically DNA, must be copied in its entirety, without many mistakes being made, in order to maintain the integrity of the genome – that is, the accuracy of the information contained within it. The process of segregation must also happen correctly, so that each daughter cell receives a full complement of DNA.

The molecular processes involved in transmission and maintenance of the genome are highly conserved from bacteria to human beings. DNA is first copied or replicated by core machinery that separates the complementary strands of the double helix and simultaneously makes copies of each individual strand. The molecular machine, or enzyme, responsible for synthesizing the new DNA copies is known as **DNA polymerase**, and as we shall see, is very highly conserved. Following this DNA replication process, the duplicated copies of the DNA are brought together to ensure that when the actual physical division of the cell occurs, the genome will be accurately partitioned (or segregated) between the

two progeny cells (Figure 1.6). Throughout each of these processes, there are elaborate mechanisms in place to ensure that replication and segregation are extremely accurate.

The genome is composed of genes and intergenic regions

The genome – the total DNA content of the cell – contains the instructions for cellular life. Within the genome, much of the critical information is found in discrete regions referred to as **genes**. A gene is typically defined as a region of DNA that controls a discrete, hereditary characteristic, and as such usually specifies the production of a functional product (a protein or RNA molecule). Different types of sequence are typically found within a given gene: some of the sequence directly encodes the product of the gene, whereas other sequences within a gene are important for specifying how and when to make the product. These regions are often referred to as coding and regulatory regions, respectively. Genes can vary in length from fewer than 100 nucleotides for certain functional RNAs, for example, to 2.4 million nucleotides for the human dystrophin gene (which is defective in muscular dystrophies).

It is worth noting that we give a ‘typical’ definition of a gene above. In reality, giving a clear-cut definition for a gene – something that has never been as straightforward as it might appear – is becoming increasingly difficult. Why is this? As more genomes are sequenced and studied in detail, we are finding that far more of the genome appears to be expressed (that is, actively ‘used’ in some way) than previously anticipated. So, we may need to broaden our view of what a ‘gene’ really is to reflect the ever-expanding proportion of the genome that is active – beyond those areas that simply code for a single functional product.

Going further, recent innovations are allowing for breathtaking amounts of sequence information to be rapidly generated and this new information coupled with functional analysis is changing our view of biology. We are now asking questions such as ‘How discrete an element is a gene?’ and ‘Are all RNA transcripts worthy of being called a gene?’ It is an exciting time to be a molecular biologist and this book sets the stage for understanding new discoveries and how they impact our understanding of evolution and biology as a whole.

The genome is physically organized into chromosomes

The large amount of DNA that constitutes the genome must be packaged in an organized way that facilitates the various key events in the life of a cell, including the replication of the DNA and the expression of the genes. The individual genes are arranged in a linear array, either densely or less densely packed with intergenic DNA spacers (segments of DNA between the genes themselves). All of this genomic DNA is packaged into one or more functional units called **chromosomes**. Some organisms have multiple, independent linear chromosomes, while others have a single, circular chromosome.

In eukaryotic chromosomes, the DNA is wrapped tightly around packing proteins (referred to as histones) to assume a highly ordered three-dimensional structure (Figure 1.7). This highly compacted form of the DNA is optimal for storage as it is

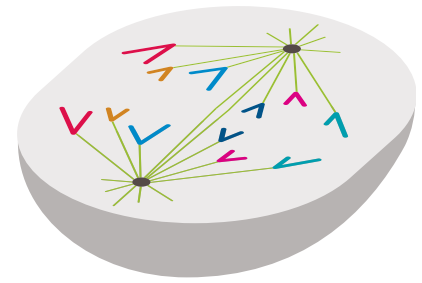


Figure 1.6 Replicated chromosomes are partitioned into two progeny cells by a process called segregation. To ensure accurate partitioning, the replicated chromosomes (having the same shape and color in this figure) are first brought together before they are pulled away into the progeny cells.

→ We learn about replication in more detail in Chapter 6 and about chromosomal segregation in Chapter 7.

➔ Chromosomal features are the focus of Chapter 4.

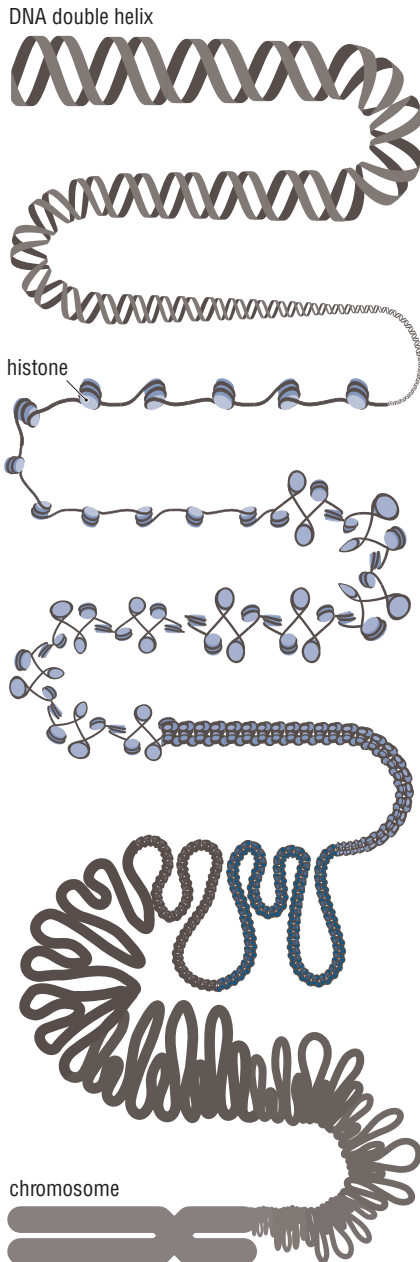


Figure 1.7 DNA is packaged into chromosomes. The double-stranded DNA strands are wound around packing proteins (histones in eukaryotes) and then further wound to assume a highly ordered and compact three-dimensional structure.

very space efficient. However, highly compacted DNA does not facilitate information retrieval. Instead, specialized mechanisms drive the opening of this tightly wound package to allow access to the DNA strands by various molecular components as needed to mediate gene expression. In addition to chromosomes, some cells have **plasmids**, which are extra-chromosomal pieces of DNA that can also carry important genetic information.

Given that the genome contains all of the genes of an organism, how then are the genes assembled? Are the genes placed in a particular order or in a particular location on the chromosome? Are they densely or sparsely distributed? What can we learn about an organism by examining overall DNA composition and structure? As we shall see, the overall composition of genomes varies widely across biology. For example, in small genomes, such as a typical bacterium (*Escherichia coli* has 4.6 million base pairs (Mbp) of DNA), the genes are rather tightly packed together, sometimes even overlapping one another to minimize overall genome size. By contrast, the human genome comprises 3200 Mbp of DNA and is composed predominantly (98%) of intergenic regions that do not appear to contain gene units. The composition of these intergenic regions is not uniform nor is the distribution of genes throughout the genome. As more genomes have been sequenced, and more studies of **gene expression** are performed, it has become very clear that we still have much to learn about the impact of overall genome structure on the fitness of the organism.

Gene number and arrangement vary greatly among organisms

The number of genes in an organism ranges widely – from 500 to 800 in some bacteria to a grossly estimated 30 000 in humans. Although in general, single-celled bacteria have fewer genes than multicellular organisms, as shown in Figure 1.8, there is no simple correlation between genome size, gene number, and organism complexity. As an example, although at the organismal level, zebrafish and lungfish appear to be very similar, the genome sizes are substantially different: the zebrafish has a genome of 2000 Mbp while that of the lungfish is 65 times larger with 130 000 Mbp. Furthermore, the nematode *Caenorhabditis elegans* and the fruit fly *Drosophila melanogaster* are certainly less complicated than humans, but possess only slightly fewer genes (around 20 000 in *C. elegans*, and around 15 000 in *D. melanogaster*). What are some potential explanations for this conundrum?

First, it is clear that the number of different products (RNAs or proteins) that can be generated from a given gene is large. So, even though the genome of a given organism may feature a seemingly modest number of genes, the same genome may, in fact, encode a huge range of different functional products – more, perhaps, than a different organism that possesses a larger number of genes.

Further, it is clear that the density of gene packaging varies widely from organism to organism, such that some genomes feature relatively more intergenic regions than others. But why is this important? Intergenic regions were previously thought to be little more than ‘junk’, acting solely as ‘spacers’ between genes. However, we now think that, far from being junk, many intergenic regions play critical roles in regulating gene expression. For example, we shall learn in Chapter 14 that much

Approximate genome size and gene number for representative organisms		
element	gene number	genome size
<i>Mycoplasma genitalium</i>	~480 genes	580 000 bp
<i>Escherichia coli</i>	4600 genes	4 600 000 bp
<i>Saccharomyces cerevisiae</i>	~ 6000 genes	12 100 000 bp
<i>Schizosaccharomyces pombe</i>	~ 5000 genes	14 000 000 bp
worm (<i>Caenorhabditis elegans</i>)	~23 000 genes	98 000 000 bp
fruit fly (<i>Drosophila melanogaster</i>)	~27 000 genes	130 000 000 bp
duckweed (<i>Arabidopsis thaliana</i>)	~29 000 genes	157 000 000 bp
zebrafish (<i>Danio rerio</i>)	~13 000 genes	2 000 000 000 bp
human (<i>Homo sapiens</i>)	~32 000 genes	3 200 000 000 bp
marbled lungfish (<i>Protopterus aethiopicus</i>)		130 000 000 000 bp
amoeba (<i>Amoeba dubia</i>)		670 000 000 000 bp

Figure 1.8 Approximate gene number and genome size for representative organisms.

of the DNA of certain organisms is composed of **transposable elements** that can (or once could) move from place to place within a genome. While this DNA was for some time thought to be functionally unimportant, new information on its critical function in evolution and in gene expression is emerging. Further, despite earlier expectations that a majority of the human genome is functionally silent, recent studies indicate that a significant fraction of the human genome is actively being transcribed.

Therefore, it is becoming clear that the value of the genome, in terms of governing the processes that underpin biological complexity, comes from more than just the genes from which a genome is composed. Instead, it comes from a combination of gene number, the production of multiple gene products from the information carried in a single gene, and the regulatory activity of the genome's intergenic regions.

Molecular conservation is highlighted by DNA sequencing

We noted at the start of this chapter how the diversity of life is unified by some common themes, with commonalities extending to the molecular building blocks used to construct living organisms, and the biological processes through which life is maintained. Direct proof of the remarkable extent of molecular conservation in all cells comes from analysis of the genome of an organism.

With the relatively recent advance of whole-genome sequencing has come new understanding of the extent to which commonalities exist between different organisms, as well as of the incredible complexity of modern-day life. This new understanding has come from the examination and comparison of whole-genome sequences from different organisms, from which similarities in genome sequence have been revealed. Such commonalities in genome sequence correspond, at least in part, to commonalities in the components and processes that operate at the molecular level in the many organisms studied (applying the logic that, if two genomes have the same (or very similar) sequences, then what results from their expression (for example, proteins) must also be similar). The new

information revealed by genome sequencing has allowed us to expand on earlier views originally based on systematic analysis of **model organisms**, key reactions, and specific molecular features.

In this book we will focus on one particular area of cellular metabolism – on understanding key molecular mechanisms critical to the storage and copying of genetic information (the genome) and the expression of this information in its functional form.

Throughout the book, emphasis will be placed on the commonalities of these mechanisms in all organisms, but we will also highlight instructive differences. The impact of whole-genome sequencing on our understanding of molecular evolution and function cannot be overstated, and promises to continue to reveal fundamental surprises about both biological commonality and diversity on earth.

1.3 BRINGING GENES TO LIFE: GENE EXPRESSION

We have now seen how genetic information is captured by the sequence of nucleotides in nucleic acids. But how is this information interpreted by the organism in which it is stored? How is it converted from a static sequence of nucleotides into dynamic biological activity, termed gene expression? It is now that we begin our exploration of molecular biology to discover the molecular machinery and processes that bring the genome to life.

Genes encoded by DNA must be ‘expressed’

While the blueprint of the cell is encoded in the DNA, the implementation of the blueprint is accomplished through the ‘expression’ of the genes into their functional forms. The gene products consist of two types of macromolecule, RNA and protein. These two molecules are principal structural and functional components of the cell, and are responsible for reactions that harvest energy, synthesize cellular components, and facilitate the import and export of materials. In short, these molecules implement cellular function.

In order for the individual genes to be expressed, the packets of information need to be transformed into the specified products. It is this transformation that lies at the heart of this book, and is a process we will gradually unravel in the chapters that follow. Genomes are the blueprints for life, but it is through molecular biology – the array of molecular components and processes – that this blueprint is used to direct how life at the level of a given organism proceeds.

The first stage in the expression of all genes is the synthesis of a copy of the DNA region of interest in the form of RNA through a process called **transcription**. Once the RNA copy of the gene has been synthesized, it is either ‘translated’ into the corresponding protein sequence or used directly, such that the RNA molecule produced by transcription has its own specific biological function, as illustrated in Figure 1.9. Let us now consider each of these processes (transcription and **translation**) in a little more detail.

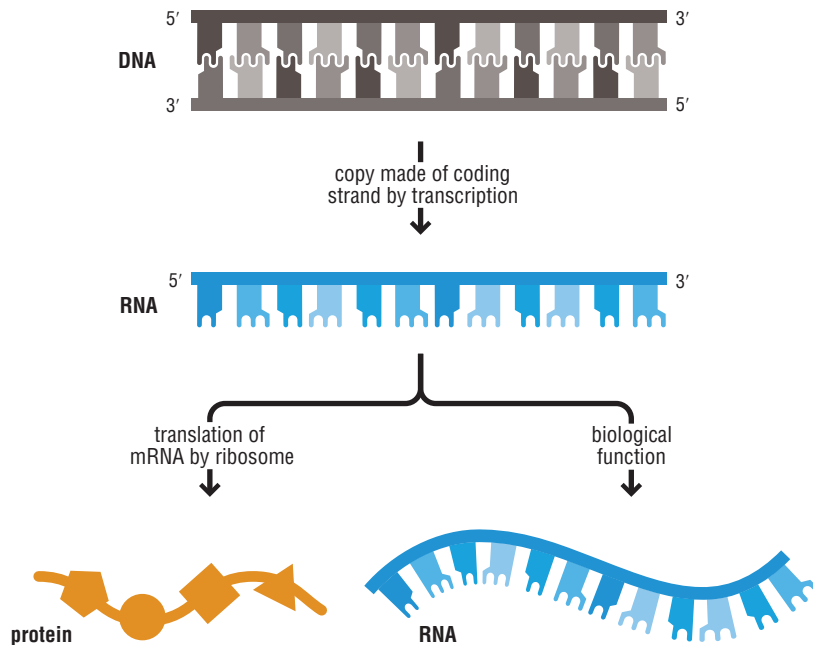


Figure 1.9 Two functions of RNA. RNA copies of DNA can serve as messenger RNAs (mRNAs) to direct the synthesis of a protein (left) or they can have intrinsic functions (right).

Transcription is the first step in gene expression for the production of both non-coding RNA and proteins

The process of copying the DNA sequence of interest into the corresponding RNA is referred to as transcription. The molecular machine that is responsible for carrying out transcription is called **RNA polymerase**, and is closely related to the DNA polymerase that is responsible for replicating the genome. The start point and end point of transcription along a particular DNA template strand are specified by sequences within the DNA template that are interpreted by RNA polymerase as a signal to start or stop.

So, the RNA polymerase catalyzes the synthesis of an RNA strand by joining together nucleotide monomers (single nucleotides) in a sequence dictated by the DNA template that the RNA polymerase is 'reading'. The process of joining together nucleotides to form a chain-like polymer is an example of a polymerization reaction. How are nucleotides added in the correct sequence to the growing RNA strand by RNA polymerase? The physical basis for the polymerization reaction is the complementarity in shape between a nucleotide on the template DNA and an incoming nucleotide that is destined for the growing RNA chain, as depicted in Figure 1.10. This complementarity mirrors that seen for the process of DNA replication, with one important difference: as in DNA replication, a C on the DNA strand directs the addition of a G on the growing RNA strand and vice versa; a T on the DNA strand directs the addition of an A on the growing RNA strand, but an A on the DNA strand directs the addition of a U (uridine, a nucleotide unique to RNA) on the growing RNA strand.

The initial product of a transcription event is referred to as the primary transcript. For some genes, the final functional gene product is an RNA species derived directly from the primary RNA transcript; such RNA species often possess regulatory functions. For other genes, the primary RNA product is an intermediate that will be used to direct the synthesis of a protein product (called translation, as

➔ The molecular details of the transcription process are described in greater detail in Chapter 8.

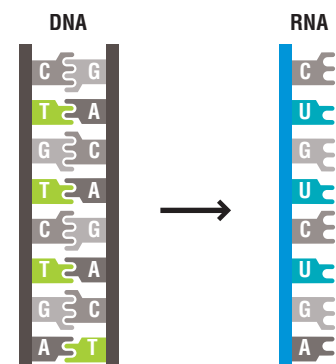


Figure 1.10 The process of transcription is common to all organisms. The complementary nature of the two DNA strands allows the DNA to be copied into a complementary strand of RNA. When DNA is copied into RNA, a uridine (rather than a thymidine) is incorporated opposite A.

➔ The post-transcriptional RNA processing events are described in greater detail in Chapter 9.



Figure 1.11 Proteins vary considerably in shape, size and function. Proteins can range in size from only a few amino acids to thousands of amino acids and have roles in every cellular process, ranging from roles as signaling molecules to receptors and structural scaffolds in the cell. In the examples shown here, the 148 amino acid calmodulin protein (PDB 1UP5) binds calcium (indicated by yellow balls), and the complex formed by the three proteins comprising the hemagglutinin protein (PDB 3HMG) mediates binding of influenza virus to the mammalian host cell and ultimately membrane fusion and entry.

➔ We explore the features of the genetic code more fully in Chapter 10.

discussed below). These two classes are commonly referred to as non-coding and coding RNAs, respectively; the key distinction is that non-coding RNAs are not translated to give a protein product, whereas coding RNAs are.

This versatility of RNA is thought to have been central to the origin of life, and is still key to the functioning of modern-day cells. These ideas will be discussed more thoroughly in Chapters 2, 9, and 10. For both non-coding and coding RNAs, the primary transcript RNA is typically subjected to a series of events that trim away and modify the RNA in a refinement process broadly referred to as RNA processing, to yield the final RNA molecule.

The RNA sequence is next ‘translated’ to make functional proteins

Many genes do not encode a functional RNA but instead encode a protein. Proteins are composed of linear strings of 20 distinct building blocks, the amino acids, which are encoded by the sequence of nucleotides comprising the RNA transcript template. Proteins built from this diverse set of building blocks (relative to the four quite similar nucleotide building blocks of DNA and RNA) can form a wide range of structures that in turn perform quite diverse roles in the cell (Figure 1.11).

The nucleic acid code for protein synthesis is composed of the four different nucleotides, which are assembled in long linear arrays that are ‘read’ three nucleotides at a time to specify which of the 20 different amino acids should be incorporated at a given point in the protein, a process termed translation (Figure 1.12). Proteins are not directly synthesized from the DNA copy of the gene, but are instead synthesized from the transcribed RNA copy of the gene, which acts as an intermediate in the overall synthesis pathway. Such RNAs that function as coding molecules to direct protein synthesis are called messenger RNAs (mRNAs). Distinctive sequences in the mRNA specify the start and stop sites for protein synthesis, while the successive triplet nucleotide sequences in between encode the linear sequence of amino acids that comprise the final protein product.

The term ‘translation’ makes sense when one understands that this is the process by which the sequence of nucleotides within a nucleic acid is ‘translated’ into the sequence of amino acids in a protein. The code that correlates triplet nucleotides with the encoded amino acid is referred to as the genetic code (Figure 1.13). The nucleotide triplets are called codons. We see, for example, that the codon UUU is translated as phenylalanine whereas AUG is translated as methionine. All codons dictate a particular outcome, either the incorporation of an amino acid or the halting of translation (a ‘stop’ site), and so we say that the code is non-ambiguous.

Translation is carried out by an elaborate macromolecular machine known as the **ribosome**, which comprises both RNA and protein molecules. At the core of the process is a functional RNA called transfer RNA (**tRNA**), that acts as a physical link between the coding information (in the form of a triplet nucleotide sequence that reads the codons) and the appropriate amino acid. Figure 1.12 shows how the tRNA functions as the actual decoder that makes translation possible: in essence, a population of tRNAs sequentially ‘reads’ the codons in a processive march along the mRNA template, recruiting amino acids to the growing protein chain in the order specified by the mRNA template. These incoming amino acids are then

chemically linked together as the march proceeds. The ribosome is the structural scaffold that facilitates this molecular march.

Most proteins that emerge from the ribosome are not fully functional. Different portions of the protein product may be removed (cleaved or spliced out) and additional moieties may be added. These processes are collectively referred to as post-translational processing.

The functional repertoire of the cell is much more complex than the genome

A direct consequence of the fact that RNA and protein products can be modified following their initial synthesis, is that the complete set of functional RNAs and proteins is far more extensive than the set that is simply encoded by the genome. While the human genome encodes an estimated number of 30 000 genes, for example, the number of different human gene products is estimated in the range of one million and it is likely that this is an underestimate. Moreover, we are beginning to understand that certain classes of molecule have escaped notice for a variety of reasons, including low expression levels and sizes that are difficult to identify. However, recent technical advances are leading to more systematic approaches to gene identification. There has been a recent explosion in known gene products exemplified by the discovery of several new classes of small, functional RNA. These provide a compelling example of how the overall gene count is likely to remain in flux for the foreseeable future. These issues will be discussed in more detail in Chapter 15. It is the exquisite control of the production of an increasingly large number of cellular components that allows for the incredible complexity that we see in the living world.

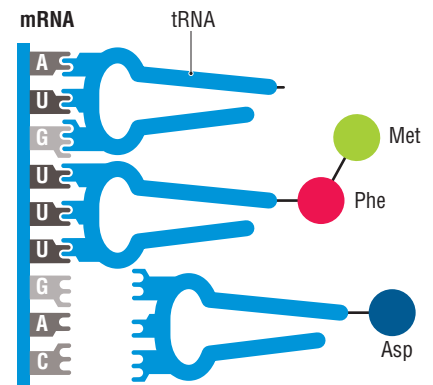


Figure 1.12 The process of translation is common to all organisms. The nucleic acids in RNA are read three at a time to specify which amino acid should be incorporated.

➔ The process of translation is explored in more detail in Chapter 10 and post-translational processing is described in Chapter 11.

1st position (5' end)	2nd position				3rd position (3' end)	Amino acids	Abbreviations	Codons	
	U	C	A	G					
U	Phe	Ser	Tyr	Cys	U	Alanine	Ala	A	GCA GCC GCG GCU
	Phe	Ser	Tyr	Cys	C	Cysteine	Cys	C	UGC UGU
	Leu	Ser	stop	stop	A	Aspartic acid	Asp	D	GAC GAU
	Leu	Ser	stop	Trp	G	Glutamic acid	Glu	E	GAA GAG
C	Leu	Pro	His	Arg	U	Phenylalanine	Phe	F	UUC UUU
	Leu	Pro	His	Arg	C	Glycine	Gly	G	GGA GGC GGG GGU
	Leu	Pro	Gln	Arg	A	Histidine	His	H	CAC CAU
	Leu	Pro	Gln	Arg	G	Isoleucine	Ile	I	AUA AUC AUU
A	Ile	Thr	Asn	Ser	U	Lysine	Lys	K	AAA AAG
	Ile	Thr	Asn	Ser	C	Leucine	Leu	L	UUA UUG CUA CUC
	Ile	Thr	Lys	Arg	A	Methionine	Met	M	AUG
	Met	Thr	Lys	Arg	G	Asparagine	Asn	N	AAC AAU
G	Val	Ala	Asp	Gly	U	Proline	Pro	P	CCA CCC CCG CCU
	Val	Ala	Asp	Gly	C	Glutamine	Gln	Q	CAA CAG
	Val	Ala	Glu	Gly	A	Arginine	Arg	R	AGA AGG CGA CGC
	Val	Ala	Glu	Gly	G	Serine	Ser	S	AGC AGU UCA UCC
	Val	Ala	Glu	Gly	U	Threonine	Thr	T	ACA ACC ACG ACU
					C	Valine	Val	V	GUA GUC GUG GUU
					A	Tryptophan	Trp	W	UGG
					G	Tyrosine	Tyr	Y	UAC UAU

Figure 1.13 The genetic code.

Gene expression is regulated in both time and space

Every step in the overall process of gene expression – be it transcription, RNA processing, translation, or protein modification – represents an opportunity for regulation. Each gene can be expressed when and where the product is specifically needed by the organism, a process broadly referred to as **gene regulation**. As we shall see throughout this book, organisms commit tremendous cellular resources to gene regulation.

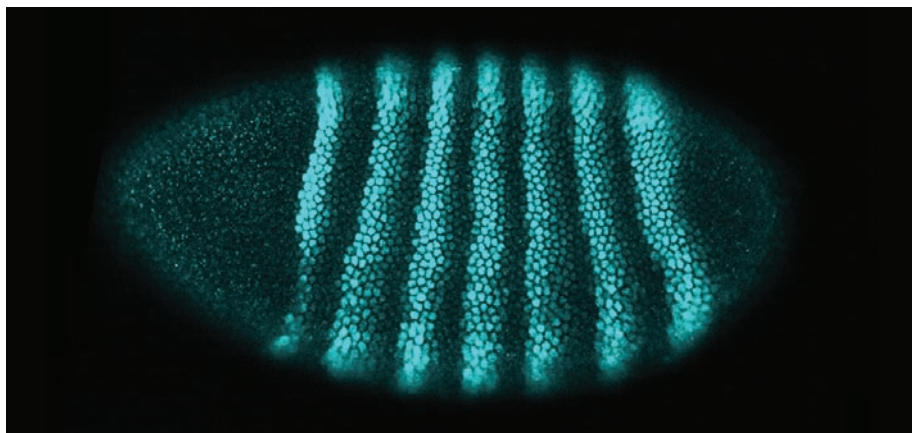
Differences in cellular behavior or identity can be brought about by differences in when or at what level a given gene product is expressed. For example, different tissues are defined by the different cells from which they are formed, and different cells are defined by the particular elements of the genome – the particular subset of genes – they express. Therefore, a liver cell will express a different set of genes than that from a nerve cell; these differences in gene expression mean that the two cells are biologically distinct.

A view of the impact of spatial regulation of single genes on the correct development of an organism is provided by the segmentation of the *Drosophila* embryo during early development. Correct segmentation of the organism depends on the localized expression of several genes including *even-skipped*. Figure 1.14 shows how the expression of *even-skipped* is localized in discrete bands of cells along the length of the embryo, laying down the pattern for the segmentation of the embryo later in development. When this localized expression is disrupted, the patterning of the embryo fails to occur as it should.

1.4 CELLULAR INFRASTRUCTURE AND GENE EXPRESSION

The cell is far from being an empty vessel in which molecular components tumble around freely. Instead, it is compartmentalized into distinct structures, each with its own function, which act in concert to mediate the processes essential to life. In this section we learn in broad strokes how the process of gene expression – from retrieval of information from the genome to its expression as gene products – is accommodated within the cell, and explore the cellular compartments in which these processes take place.

Figure 1.14 Expression of *even-skipped* gene in a *Drosophila* embryo. The *even-skipped* gene is expressed in discrete bands of cells along the length of the embryo at an early stage of *Drosophila* development and helps to lay down the pattern for the segmentation of the fly.
Figure © Sabrina Desbordes.



An internal membrane-bound compartment in the cell contains the genome in eukaryotes

At the beginning of this chapter, we discussed how the cellular membrane is key to life because it defines self and non-self. The membrane encapsulates the genetic information critical for self-propagation together with the functioning molecules of the cell. In addition to the exterior membrane shared by all organisms, more complex organisms have additional membrane-bound compartments in their interior. Many unicellular species, including members of the bacterial and the archaeal kingdoms, do not have internal membranes within their cells; because of this distinction they are referred to as **prokaryotes** (meaning ‘before the kernel’ or pre-nuclear). It should be noted, however, that these cells are not completely without compartmentalization. For example, the chromosomal DNA is sequestered in a region called the **nucleoid** (Figure 1.15).

By contrast, higher-complexity organisms (including those that are unicellular and multicellular) contain an internal membrane-bound structure called the **nucleus** where the genetic information is contained. The name eukaryote literally translates as ‘well kernel,’ and defines a kingdom of organisms having a nuclear compartment that contains their genetic information. The contents of the cell lying outside the nucleus are collectively referred to as the **cytoplasm**.

Eukaryotic cells feature multiple membrane-bound compartments with specialized functions

While the nucleus is the defining compartment within eukaryotic cells, other membrane-bound compartments play equally important roles in cellular life. These compartments are generally referred to as organelles; the key organelles in eukaryotic cells are illustrated in Figure 1.16. The **mitochondria** are known as the powerhouses of the cell and contain the macromolecular machinery responsible for deriving chemical energy from food precursors. **Chloroplasts**

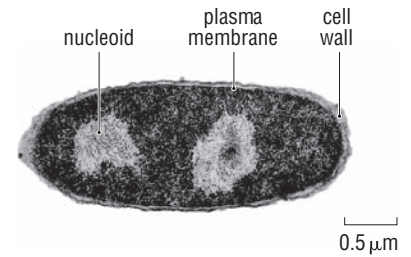


Figure 1.15 Electron micrograph of an *E. coli* cell. The nucleoid containing the DNA and the plasma membrane and cell wall surrounding the cell are labeled. From Sinauer *et al.* Biozentrum, University of Basel. Science Photo Library. Photo Researchers, Inc.

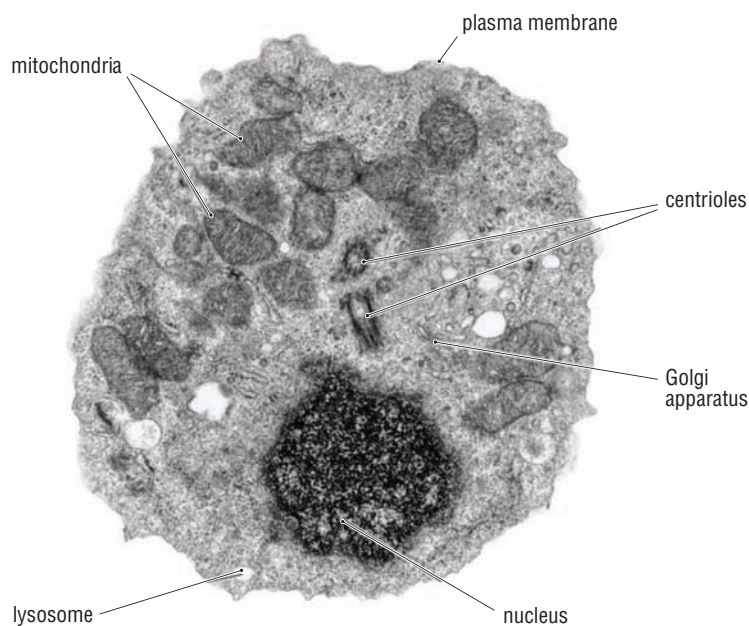


Figure 1.16 Electron micrograph of a human lymphocyte cell. This is the same type of cell that is shown in pink in Figure 1.2. The nucleus, Golgi apparatus, centrioles, mitochondria, lysosomes, and plasma membrane are labeled. From Dr Gopal Murti/Visuals Unlimited.

are organelles found in plant cells that house the machinery for harvesting energy from sunlight for the production of sugars within the cell. Both mitochondria and chloroplasts contain DNA unique to these organelles that encodes certain components critical to their function. This feature led scientists to hypothesize that these organelles originated as **endosymbionts**, organisms living within another cell that eventually became wholly dependent on their host for existence.

Other well-characterized membrane-bound compartments in the eukaryotic cell include the **endoplasmic reticulum** and the **Golgi apparatus**, which are essential for the production of proteins that are inserted into the membrane or secreted from the cell. Certain other organelles, which will not be discussed further in this book, are those devoted to the metabolic and catabolic needs of the cell. These include the lysosome and the peroxisome.

In addition to the membrane-bound compartments that we describe here, the eukaryotic cell also depends on other subcellular localization mechanisms both in the nuclear and cytoplasmic regions of the cell. These regions, which generally were first visualized by microscopy, include structures denoted P-bodies, stress granules, and neuronal particles in the cytoplasm, as well as the nucleolus and Cajal bodies in the nucleus.

What advantages does compartmentalization confer upon a cell? Why does it help biological processes occur in the way that they do?

Compartmentalization is important for facilitating chemical reactions

If we look at biological diversity, we see that overall cell size increases in the eukaryotic lineage when compared with simpler bacterial cells. All cellular reactions depend on interactions between molecules; by localizing various components together, their effective cellular concentrations are increased and so reactions can proceed more efficiently. These principles will be described in more detail in Chapter 2.

In addition to the chemical benefits of increased concentrations of cellular components, cells also benefit from the level of organization that accompanies compartmentalization. Just as we organize our own lives and make them more efficient by compartmentalizing our possessions – we store food items together in the kitchen ready for the preparation of meals, we gather books, class notes, and laptop on a desk ready for study – by co-localizing related processes with shared substrates, intermediates or macromolecular components, cellular efficiency is increased. Cell biological approaches provide striking images of detailed subcellular localization patterns for various cellular components that highlight the potential of such compartmentalization (Figure 1.17).

Compartmentalization is crucial for regulating gene expression in eukaryotic cells

The natural consequence of the fact that the DNA in eukaryotic cells is housed in a membrane-bound compartment, the nucleus, is that there has developed a large amount of regulatory control of gene expression based on this physical

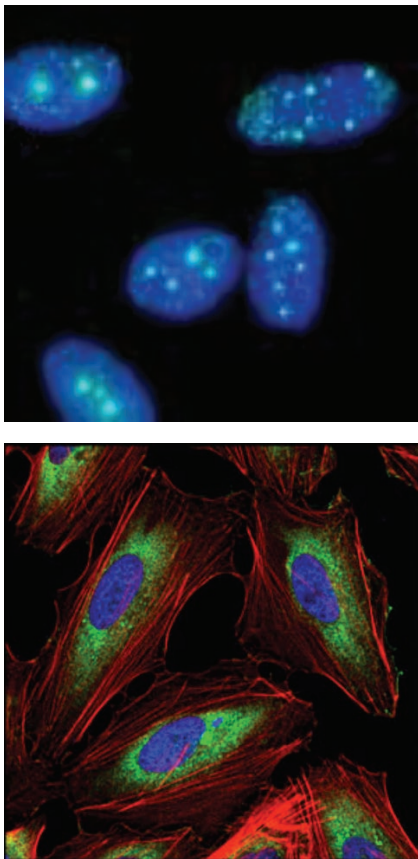


Figure 1.17 Subcellular localization allows for more efficient reactions and regulation. An RNA polymerase (light blue) is localized to specific regions in the nucleus (stained dark blue with dye that binds to DNA) of a eukaryotic cell, while ribosomes (a ribosomal protein is stained green) are localized to the cytoplasm surrounding the nucleus (blue). This separates transcription and translation, providing an opportunity for regulation.

separation from the rest of the cell. Transcription of the DNA (to make RNA) takes place in the nucleus where the genome is housed, whereas translation of the RNA (to make proteins) takes place in the cytoplasm. This difference in location offers an immediate opportunity for regulation: just as you can regulate movement from one room to another in a house by locking (or unlocking) the door between them, so the physical barrier presented by the nuclear membrane provides an opportunity to regulate passage from the nucleus to the cytoplasm (and vice versa), such that passage can only occur if certain conditions are met.

For example, we know that in eukaryotic cells, transcripts that are produced in the nucleus are significantly modified following transcription (RNA processing), but that this modification occurs before the transcript is exported from the nucleus to the cytoplasm. In fact, transcripts cannot be exported from the nucleus unless they have been fully processed. This feature thus provides the cell with an opportunity for quality control – the blocking of movement of the transcript from the nucleus to the cytoplasm across the nuclear membrane – which prevents inadequately processed mRNAs from encountering the translational machinery. In bacteria, where there is no nucleus to sequester the genetic information, transcription and translation are spatially coupled, leading to distinct forms of regulation in these organisms.

An additional complexity faced by gene expression is the production of proteins destined for different locations – for example, for the cytoplasm, for insertion into the cellular membranes or beyond the cell itself. How can proteins be targeted to their correct locations?

As we will discover, proteins that are destined for the membrane or for export from the cell have physical properties that are incompatible with passage through the membrane without specific assistance. In bacteria, where there are no internal organelles, proteins destined for secretion are directly targeted to the extracellular membrane as they are translated. Eukaryotic cells appear to have elaborated on this plan, but instead of directly targeting proteins to the exterior of the cell, translation occurs on the membrane of an extensive organelle system known as the endoplasmic reticulum. This organelle is really a network of tubules where membrane and secreted proteins are translated, folded, modified and then sent on their way.

The final stop in the maturation of many proteins from genes is the Golgi apparatus, an organelle that modifies the nearly completed protein with sugars before packaging the protein for final export. Proteins are sent to the Golgi apparatus from the endoplasmic reticulum, and are ultimately sent to the cell membrane for export.

While this book will not focus on the cell biological aspects of gene expression, it is important to remember the critical role played by cellular structure in all biological processes.

1.5 EXPRESSION OF THE GENOME

In Section 1.2 we explored the expression of individual genes. However, genes do not operate in isolation. Just as individual humans in a community interact, and influence the behavior of others, so genes operate in networks, where the activity

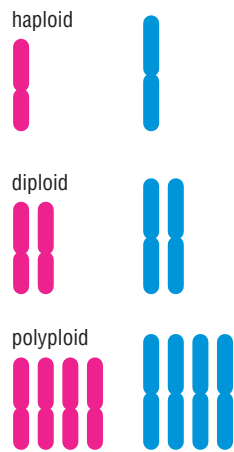


Figure 1.18 Cells can have one or more copies of their chromosomes. Haploid cells only contain a single copy of the two distinct red and blue chromosomes and diploid cells have two copies. Some cells are polyploid and have multiple copies.

of one gene can quite often influence the activity of another. In this section we look beyond the expression of individual genes to ask how whole communities of genes – that is, genomes – are expressed.

The output of the genome is dictated by the interplay of genes

The expression of the information in a genome determines the physical properties of an organism. The visual features and properties of an organism are referred to as the **phenotype**, while the collective DNA sequence that determines the phenotype is referred to as the **genotype** of the organism. The output of the genome, the phenotype, is dictated by which genes are expressed, at what times, and at what levels. Each organism is thus the product of a combinatorial process dictated by a great many genes working together.

An additional complexity is that organisms can carry one or more copies of the individual genes (Figure 1.18). For example, yeast can survive indefinitely as a **haploid** organism, carrying a single copy of its genome. More complex multicellular organisms by contrast typically have two copies of their genome and are referred to as **diploid**. Some organisms have more than two versions of each of their genes and are said to have higher ploidy – for example, the frog species *Xenopus laevis*, commonly used for biochemical experiments, is **tetraploid** – it has four copies of each gene.

The importance of ploidy is that, in a haploid organism, a single version of a gene is expressed whereas in a diploid organism there is typically a mixed population of gene products derived from the expression of two different copies of the gene. Why can this be important? If one copy of a gene is defective, such that a functional gene product is not produced when the gene is expressed (or is produced in reduced amounts), a diploid organism has the advantage of a second copy of the gene to compensate for the defective one. By contrast, a haploid organism does not have the benefit of a ‘back-up’ gene, putting it at a greater risk of experiencing phenotypic defects if its only copy of a given gene is mutated.

Genetic analysis has been essential in deciphering the natural function of genes

We see above how different genes are expressed to produce different RNA or protein products, and that it is the concerted action of all the genes in a genome that directs the development and function of an organism. But how do we know what individual genes actually do? How can we say that gene *x* drives process *y*?

Much of our understanding of biology comes from what are referred to as genetic approaches, which probe the natural function of genes by comparing so called **wild-type** (normal) properties, the typical or most common form, with non-wild-type (mutant) ones. In classical genetic approaches, scientists observed organisms under various environmental conditions following the imposition of DNA damage, looking for visible phenotypic changes. That is, they looked to see the consequence of different genes being rendered inactive or altered. When such mutant (variant) organisms were found, then the specific changes in the genome that specified the difference between mutant and wild-type were pursued in order to provide information on potential gene function. So, for example, if a wild-type

fruit fly had red eyes, but a defect in gene *a* produced a mutant with white eyes, one could deduce that the function of gene *a* was to determine eye color in some way (Figure 1.19).

This general approach is typically referred to as ‘**forward genetics**’. For example, Figure 1.20 shows a single mutant *Arabidopsis thaliana* plant seedling surrounded by a group of wild-type plants. In the example shown, the phenotype of the mutant plant is an elongated stem while the genotype turns out to be a **mutation** in a gene encoding a blue light photoreceptor. A specific change in the gene for the blue photoreceptor is manifested as this obvious change in the growth properties of the plant. We will sometimes refer to these two different versions of the gene as the wild-type and mutant **alleles** of the gene. An allele is thus a version of a gene.

With a more complete understanding of genomes and their complete composition, and with sophisticated tools of gene manipulation, it has become possible to specifically disrupt a gene of interest in some organisms and to study the phenotypic consequences. This powerful approach is typically referred to as **reverse genetics** and has been used with wide success to decipher gene function.

Genotypic changes have phenotypic consequences

In forward genetic studies, the phenotype of the mutant organism is typically tracked by the scientist with the ultimate goal being to identify the specific changes in the DNA (mutation) that resulted in the phenotype in question. This type of analysis in principle establishes causality in that a certain change in the DNA results in the physical properties being followed.

Mutations that cause differences from the wild-type organism are typically categorized as recessive or dominant. These terms take on meaning when one remembers that organisms do not always have just a single copy of a given gene. If, in a diploid, the phenotype of the mutant gene product is masked by the presence of the wild-type version, then the mutation is said to be recessive. A common example of a **recessive mutation** is a change in an enzyme that reduces its activity, but which is compensated for by the wild-type version that provides sufficient activity. By contrast, if a mutant gene product has obvious phenotypic consequences when the wild-type product is present, then the mutation is said to be dominant. A common example of a **dominant mutation** might be a change in a structural protein that disrupts higher-order protein structure, such that filament formation is globally disrupted by the presence of the mutant protein. Figure 1.21 shows schematically how recessive and dominant mutations impact mouse coat color.

What is the physical nature of the mutations that can cause such phenotypic changes? The mutations can be the result of very small changes in the gene (i.e. the alteration of a nucleotide), the rearrangement of large regions of the chromosome, the insertion of new segments of DNA or the **deletion** or excision of small or large sections of the genome (Figure 1.22). These changes can result in direct changes to the gene product encoded by a given gene, or in disruptions of overall gene expression (where the actual identity of the gene product is not changed). Mutations that eliminate the function of a gene are termed **null** or **loss of function mutations**. These types of mutation typically are associated either with a change in the product that completely eliminates function or a large disruption in the DNA sequence that eliminates gene expression.



Figure 1.19 *Drosophila* mutant with a defective pigmentation gene. Wild-type fruit flies have red eyes (left), while mutant flies carrying a defective *white* gene have white eyes (right). Dr Jeremy Burgess/Science Photo Library.

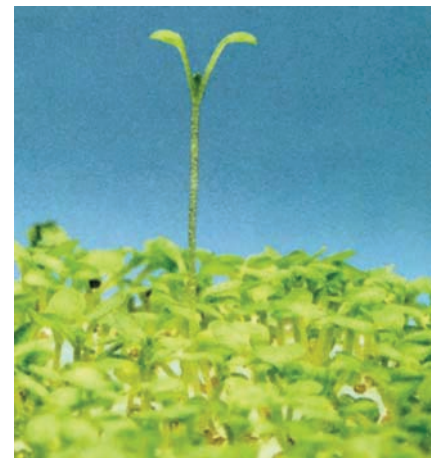


Figure 1.20 *Arabidopsis* mutant unable to sense blue light. The stem (hypocotyl) extension of wild-type *Arabidopsis* seedlings is inhibited by blue light. The growth of a mutant carrying a defective blue light receptor *cry* gene is not inhibited by blue light so it stands out among the wild-type seedlings.

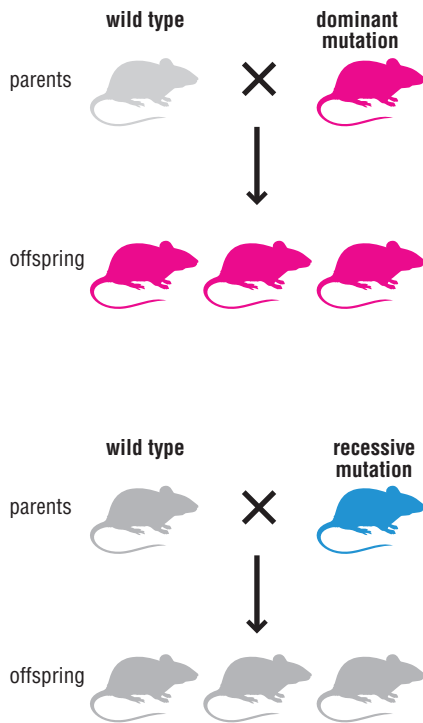


Figure 1.21 Effects of dominant and recessive coat color mutations in mouse. In the cross between one parent (gray) contributing the wild-type gene and another parent (pink) contributing a gene with a dominant mutation, all offspring will have the phenotype of the parent carrying the dominant mutation. In the cross between one parent (gray) contributing the wild-type gene and another parent (blue) contributing a gene with a recessive mutation, all offspring will have the phenotype of the parent carrying the wild-type gene.

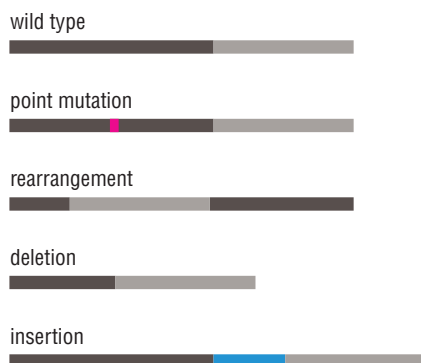


Figure 1.22 Types of mutation. The sequence of DNA can be mutated by a single nucleotide change (point mutation, red), shuffling of different regions of DNA here indicated by the two shades of gray (rearrangement), excision of DNA (deletion) or insertion of DNA (blue).

Single nucleotide changes in the DNA sequence can directly result in changes in the identity of the encoded gene product (either RNA or protein). In protein coding genes a change of one nucleotide can be sufficient to change a codon from one which encodes a particular amino acid to one encoding a different one. Such changes in amino acid identity are referred to as **missense mutations** (Figure 1.23). If the amino acid substitution results in gene function being completely lost the missense mutation can be thought of as a null mutation. By contrast, other substitutions may result in the production of partially functional gene products.

In some cases, there can be a change in the DNA sequence that introduces a premature stop signal such that only truncated species of the protein product are made. Such changes are typically referred to as **nonsense mutations** (Figure 1.23). Sometimes there are changes in the DNA sequence that do not result in obvious phenotypic changes. Such changes are typically referred to as **silent mutations** (Figure 1.23).

The origin of disease can often be traced to mutations in the organism

In multicellular organisms, there are two general types of cell:

- **somatic cells** – these constitute the building blocks of the body of an organism and simply transmit their genetic information to daughter cells in the same organism
- **germline cells** – these are involved in the production of the progeny and transmit the genetic information to the next generation of the organism.

Consequently, mutations found in the germline cells typically affect the progeny of the organism whereas those found in somatic cells affect the organism itself.

One common consequence of gene mutation is the onset of disease, whereby normal biological function is disrupted in some way. Some diseases are the consequence of variation in a single gene and are said to be **monogenic**. A well-known example of a monogenic disease is phenylketonuria, in which a mutation in a gene encoding a specific liver enzyme can lead to severe mental retardation. In other cases, disease states are thought to result from variations in multiple genes and are said to be **polygenic** (though the term ‘multifactorial’ is also commonly used). Some of the most prevalent diseases, such as heart disease, Alzheimer’s, arthritis, and diabetes are thought to be polygenic.

The presence of a specific mutation does not always lead to a disease state, however. For example, mutations in genes called *BRCA1* and *BRCA2* are associated with breast cancer, but not all individuals carrying these mutations develop cancer during their lifetime. The **penetrance** of a given mutation is defined as the percentage of individuals carrying a particular mutant genotype that exhibit the mutant phenotype. The risk of developing a mutant phenotype when carrying a mutant genotype is determined by both the genetic makeup of an organism and environmental influences, such as diet and exposure to different environmental conditions.

The relative extent to which the genotype (nature) and the environment (nurture) influence the phenotype associated with a specific mutation has long been debated and undoubtedly will continue to be a subject of controversy. In the example of the *BRCA1* and *BRCA2* mutations, other genetic factors clearly influence whether individuals are more or less susceptible to developing cancer, but environmental

factors such as diet and smoking can also affect the penetrance of these mutations. As we shall see in Chapter 15, whole-genome sequencing is providing critical information for the genetic mapping of complex diseases such as cancer.

1.6 EVOLUTION OF THE GENOME AND THE TREE OF LIFE

The genome of a particular species is not a static, unchanging entity. If it were, there would not be the remarkable variety of life in the world around us. Instead, genomes are dynamic: they undergo gradual, incremental change over time, which results in progressive change in the organism whose development and function they are orchestrating. The progressive change in an organism's phenotype over many generations, underpinned by a gradual change in that organism's genotype, is described by the process of evolution, as we now discuss.

The evolution and diversification of life is gradual

The argument in favor of a common ancestor (LUCA, see above) is based most simply on the fundamental similarities in the molecular features of all organisms, for example, the usage of nucleic acids to store genetic information and the genetic code that specifies the amino acid sequence of proteins. It is highly unlikely that similar mechanisms would have evolved independently in many different organisms. Instead, it is far more likely that these common mechanisms were passed down from our ancestors into each new species as they evolved.

If common features have been conserved between species, how did the great diversity of life on earth derive from the last common ancestor? If we see such similarities between species, how could the differences that delineate different species arise at all? Evolution is the underlying principle of modern biology that explains this conundrum. It is the process by which organisms gradually change over successive generations through a process of random change or mutation in the genome, coupled with natural selection, the outcome of competition. Let us consider a hypothetical example to illustrate this process (Figure 1.24).

Imagine a population that shares a common genome: all members of the population have the same genotype (genomic composition) and, hence, the same phenotype. This genome can only be passed on to future generations if an organism

wild type	ATG	AAT	ATT	CGA	GAT
	Met	Asn	Ile	Arg	Asp
missense	ATG	AAT	ACT	CGA	GAT
	Met	Asn	Thr	Arg	Asp
nonsense	ATG	AAT	ATT	TGA	GAT
	Met	Asn	Ile	stop	Asp
silent	ATG	AAT	ATC	CGA	GAT
	Met	Asn	Ile	Arg	Asp

Figure 1.23 Multiple types of point mutation. Point mutations can lead to the incorporation of the wrong amino acid (missense), premature termination (nonsense), or not affect the sequence (silent) of the translated protein.

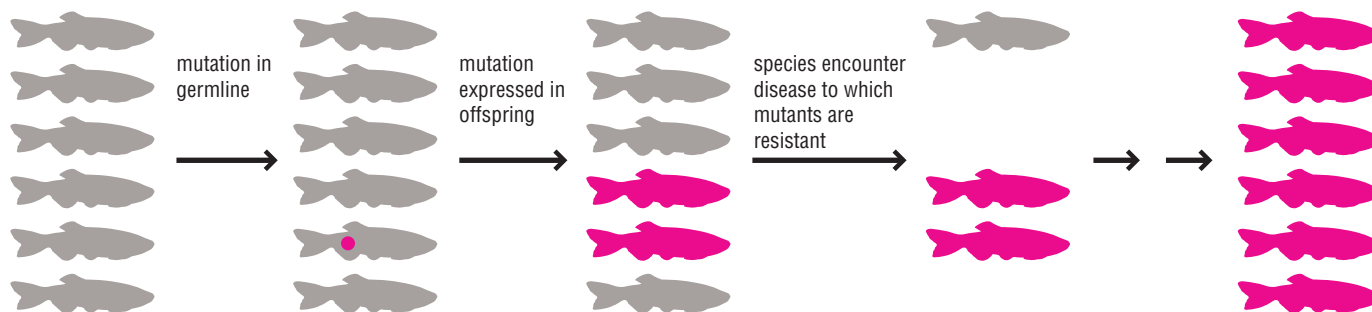


Figure 1.24 Ever-larger populations of individuals carrying specific DNA changes due to selection. Initially all the fish have the same genotype. Then a random mutation arises in the germline cells of one of the fish. By chance, the offspring (pink) of this fish have increased resistance to a particular disease. Since the offspring are more likely to survive, they are also more likely to reproduce, leading to an increase in the number of fish carrying the DNA change.

survives long enough to reproduce. In this case, however, all of this population will be equally healthy (because they have the same genome) and so all should be equally likely to reproduce, and have their genome transmitted to the next generation.

Now, imagine that one or two of the population develop a mutation in their germline cells. This mutation would not affect them directly (because it is in their germline and not their somatic cell line). However, the mutation *will* be expressed when the genome is transmitted to the next generation.

Let us assume the mutation confers upon the individual carrying it a particular resistance to disease. The next generation will therefore contain a minority of individuals who show unusually high disease resistance. If the population as a whole is then exposed to a disease in the surrounding environment, it is those individuals who carry the mutated (unusually effective) disease-resistance gene who are most likely to survive – and are therefore most likely to reproduce. Consequently, there is an increased chance of this mutated gene being passed on to the *next* generation.

Over time, environmental influences, such as disease, will continue to ‘select for’ the mutant individuals and, over time, an ever-larger proportion of the population will contain the mutant gene, until it becomes the norm. This is the process of natural selection in action. As a consequence of natural selection, as generations come and go, we see our original species evolving into one that is more disease-resistant than it was before. Notice how evolution is not a rapid process – it is a gradual change in genotype which, if beneficial to the population, becomes increasingly established, driving a concomitant shift in the observed phenotype for the population.

So we see how mutations that result in increased success for the organism will be propagated or selected; such selection eventually leads to the emergence of a reproductively isolated group of organisms or species. A species is defined as a group of organisms capable of interbreeding and producing fertile offspring.

All organisms can be grouped into three domains of life based on sequence comparisons

We have learned that the collection of genes in an organism, and their expression, determines the phenotype and the identity of an organism. By comparing the sequences of specific genes or whole genomes from different organisms, we can learn something about the likelihood that a given property will be similar among different organisms, which tells us how closely these organisms are related in evolutionary terms.

As shown in an alignment of the sequences of the same region of the ribosomal RNA genes from many different organisms (Figure 1.25), the sequences for individual genes can be very similar in certain regions while varying more extensively in others. The extent of relatedness can be quantitated and these results then represented in a branched model called a **phylogenetic tree**. An example of a phylogenetic tree based on the comparison of ribosomal RNA genes is shown in Figure 1.26. In these models, the length of the line reflects the relatedness of the gene at one end of the line to that at the other end, and thus by extension, the relatedness of these organisms. Therefore, an organism that is separated from its most recent ancestor by a short line is more closely related to its ancestor than a second organism that is separated from the same ancestor by a longer line.

<i>T. acidophilum</i>	GGCCUUUAGUACGAGAGGAACAAGGG
<i>E. coli</i>	UGCUCUAGUACGAGAGGACCGGAGU
<i>B. subtilis</i>	UGUCUUUAGUACGAGAGGACCGGGAU
<i>E. histolytica</i>	ACAACUCAGUACGAGAGGAACCGUUG
<i>S. cerevisiae</i>	UGAACUUAGUACGAGAGGACAGUUC
<i>C. elegans</i>	CCUGCUUAGUACGAGAGGAACAGCGG
<i>D. melanogaster</i>	CCUGCGUAGUACGAGAGGACCGCAG
<i>R. norvegicus</i>	GGUGCUCAGUACGAGAGGACCGCAC
<i>O. sativa</i>	UCAACCUAGUACGAGAGGACCGUUG
<i>A. thaliana</i>	UCAACCUAGUACGAGAGGACCGUUG
<i>H. sapiens</i>	CCUGCUCAGUACGAGAGGACCGCAG

Figure 1.25 Similarity between sequences from distantly related organisms. The alignment of sequences corresponding to ribosomal RNAs from distantly related organisms shows regions of similarity and variation.

Trees constructed from comparisons of different genes can differ, and so phylogenetic trees are typically based on comparisons of multiple genes. The precision of such classification approaches far exceeds more traditional methods of classification that depended on visually apparent physical features such as whether or not cells have nuclei.

Sequence comparisons based on many genes indicate that there are three distinct groups, or branches, of the tree into which all of organismal life can be split. These broad groups are referred to as the **three domains of life** and include the bacteria, archaea, and eukaryotes.

- The bacterial domain encompasses the largest number of organisms and continues to grow as the environment is explored. These organisms are single celled and have minimal subcellular compartmentalization.
- Archaeal organisms share features with organisms in the bacterial domain in that they are unicellular and do not have a nuclear structure. However, core archaeal proteins typically show more similarity to the corresponding proteins found in eukaryotic organisms. Thus, early classification identified these organisms as bacterial, based on their physical features, but sequence comparisons allowed for the identification of this separate domain of life as recently as 1977.
- Eukaryotic organisms are distinguished by more extensive compartmentalization than found in either bacteria or archaea, including the nucleus that holds their DNA, and by possessing a cytoskeleton that helps to structure and organize

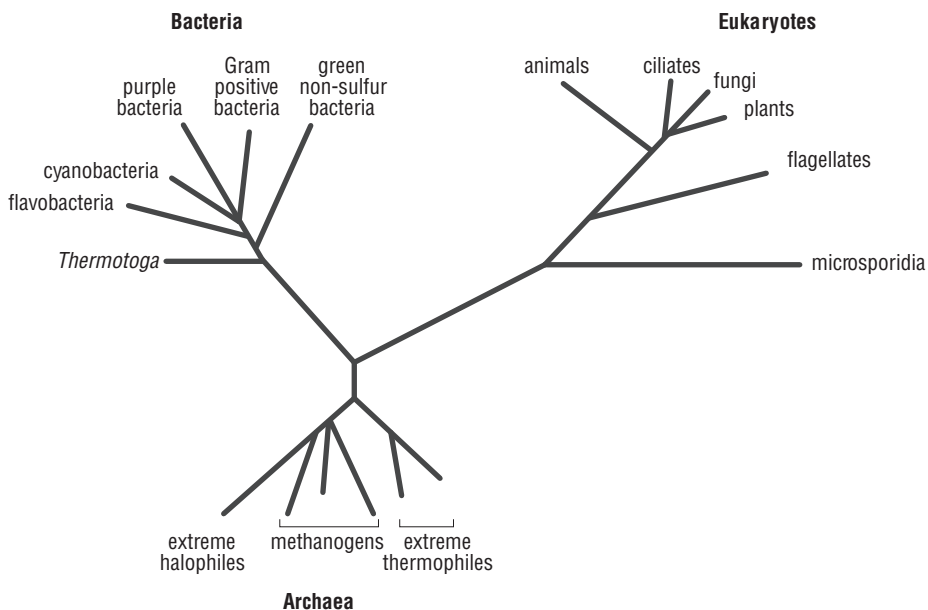


Figure 1.26 Relatedness of organisms depicted by a phylogenetic tree. The tree shows the relatedness of the sequences encoding ribosomal RNAs from bacteria, archaea, and eukaryotes. The length of each of the lines corresponds to the relatedness of the genes. From Woese, *Microbiological Reviews* 1987; 51:221–271.

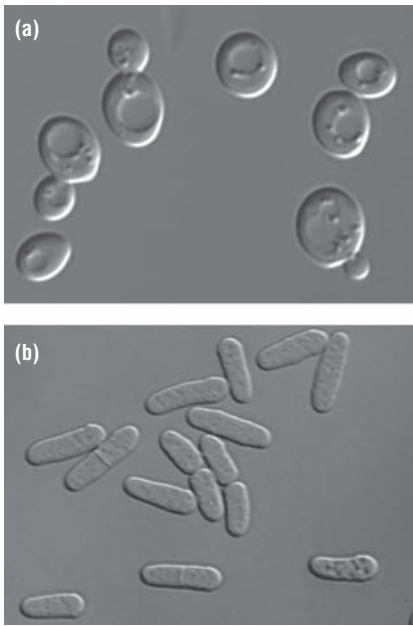


Figure 1.27 Commonly studied fungi *S. cerevisiae* and *S. pombe*. As seen in the micrograph, *S. cerevisiae* is oval shaped and divides by budding, whereas *S. pombe* is rod shaped and divides by fission. From Morgan D.O. 2007 *The Cell Cycle: Principles of Control*.

their larger cell volume. All multicellular organisms, including animals, plants, and fungi, are eukaryotic, but not all eukaryotes are multicellular: certain fungi such as yeast, algae, and slime molds are single-celled eukaryotes.

This tree of life is an organizing principle for biological studies and we will see below how biologists have chosen a range of organisms covering some of this diversity to focus on explorations of molecular function.

Studies of model organisms have been key to understanding many biological processes

The study of organisms that have features of agricultural or medical importance, yet are easy to manipulate under laboratory conditions, has been critical to the understanding of the many processes in biology. These organisms, often referred to as model organisms, are typically directly evaluated and manipulated, subjected to mutational studies, or are used to prepare samples for biochemical analysis.

Two commonly studied bacteria, *E. coli* and *Bacillus subtilis*, belong to two distinct groups of the bacterial domain referred to as Gram-negative and Gram-positive, respectively, based on their differential staining with a particular violet dye (Gram stain). Two commonly studied fungi belonging to the eukaryotic domain are *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Figure 1.27). While both are fungi, these two organisms exhibit significant differences in their biology – including differences in something as essential as their mode of cell division, where budding is seen in *S. cerevisiae* and fission in *S. pombe*. Careful comparison of these genomes suggests that *S. pombe* is more closely related to higher eukaryotes such as humans.

Multicellular eukaryotes that are commonly studied are the fruit fly *D. melanogaster*, the worm *C. elegans*, the frog *X. laevis*, the zebrafish *Danio rerio* and the plant *A. thaliana*. The predominant mammalian model organism is the mouse *Mus musculus* (Figure 1.28). But what gives an organism its ‘model’ status? Most of the model organisms were chosen for practical reasons: they have advantages in terms of their biological composition, development, or behavior that make them particularly amenable to study. For example, many of the organisms have

Figure 1.28 Commonly studied multicellular eukaryotes. Model multicellular eukaryotes commonly studied in biology include the fruit fly *D. melanogaster* (Hermann Eisenbeiss/Science Photo Library), the worm *C. elegans* (Sinclair Stammers/Science Photo Library), the frog *X. laevis* (Zigmund Leszcynski/Photolibary Group.com), the zebrafish *D. rerio* (Oxford Scientific (OSF) Photolibary Group), the plant *A. thaliana* (Dr Jeremy Burgess/Science Photo Library), and the mouse *M. musculus* (Oxford Scientific (OSF)/Photolibary Group).



fast generation times: the average generation time of a fruit fly is two weeks, while an *Arabidopsis* plant is able to produce seed a month after germination. All of the model organisms mentioned above are also relatively small and can thus be cultivated under typical laboratory conditions.

In addition, each of the organisms mentioned above has different strengths and weaknesses regarding the types of analysis that can be readily performed. For example, *X. laevis* is a key organism for studying development in a biochemically tractable system since the developing oocyte is of large enough size to permit direct manipulation. However, genetic approaches are not very well developed in this system. By contrast, *D. melanogaster*, *C. elegans*, and *S. cerevisiae* are probably best known for their potential as genetically tractable organisms where unbiased genetic screens are readily performed.

It bears remembering that model organisms are simply that – they allow us to build models (or informed estimates) of how biological systems operate but do not allow us to formulate exhaustive descriptions of exactly what is going on in a single investigation. It is only by successive rounds of investigation, sometimes in different model organisms, each perhaps only revealing a small new insight, that our understanding is incrementally increased. As these insights build up, so we refine our understanding, and build gradually more detailed, and more refined, best guesses of the reality of life. Equally importantly, model organisms do not always provide a definitive indication of biological processes throughout the kingdoms of life; it may not be possible to extrapolate an observation made in one organism to a seemingly similar organism.

The study of viruses has provided tremendous insights into biological function

Viruses are particles composed of a nucleic acid-based genome surrounded by a protein coat. Although viruses contain their own genomes, they cannot replicate on their own. Instead, viruses replicate within another organism where they usurp components of the host cell machinery to replicate or express their own genomes (Figure 1.29). Viruses have been found to be associated with all organisms and are present in an astounding variety and number. For example, it is estimated that 10 000 000 **bacteriophages** (the term for viruses that infect bacteria) are present in 1 mL of coastal seawater.

Viruses use quite varied mechanisms to usurp the host cell machinery. Therefore, studies of the mechanisms used by viruses to infect and replicate in their host cells have time and again provided fundamental insights into basic cellular functions, and the strategies and mechanisms used by organisms to bring their genomes to life. So we can learn much even from entities that we would not classify as model organisms.

The molecular components of different organisms are described using different nomenclature rules

As a consequence of the large body of work carried out with the model organisms mentioned above, many of the examples cited in the subsequent chapters will be from studies involving them. In citing such examples we need to be aware that, despite the common molecular components and processes that operate in

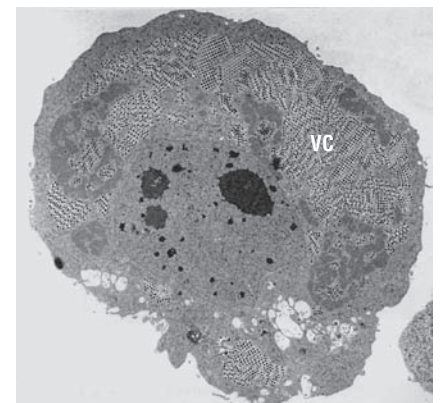


Figure 1.29 Electron micrograph of a virus-infected cell. The picture illustrates the relative size of the crystalline-like virus particles (VC) relative to the eukaryotic cell.

different organisms, the ways in which these components are described varies. In particular, the conventions used for naming genes, proteins, and mutant alleles in different organisms – the nomenclature used – vary from organism to organism. For example, in some organisms, a wild-type gene name is written in uppercase letters, whereas the mutant gene name is written in lowercase letters. Sometimes, the gene name is in italics, and the name of the protein resulting from the expression of that gene is written in roman (non-italic) type. Throughout this book, we observe the most widely accepted nomenclature rules for each organism, and these are summarized in the table on page ii at the front of the book.

Also, we need to be aware that homologous genes and proteins in different organisms – that is, genes and proteins with conserved sequences (and, hence, conserved functions) – may have different names. For example, the eukaryotic homolog of the bacterial sliding clamp β -protein involved in DNA replication is called PCNA (proliferative cell nuclear antigen, a name that derived from how the factory was originally identified). Such differences can make it more challenging to see patterns and similarities, but they simply reflect the rich history and diversity of research that underpins our current knowledge of the field.

We have now completed our brief tour of biological systems, and the journey from genetic information to the living organism. We have seen how our genome directs our development and function, and how changes in our genome propagate changes in our physical character as evolutionary forces are brought to bear. We have also had our first glimpse of the molecular mechanisms that retrieve the information from our genome, and, by transforming information into activity, bring the genome to life.

In the next two chapters we move from the level of the cell to the level of atoms and molecules, to consider the chemical infrastructure of life, before beginning our more detailed exploration of molecular biology, and the principles of genome function.

* SUMMARY

The unifying features of life

- The diversity of life is unified by some common themes: a living organism must be distinct from its environment, have a way of storing biological information, be able to replicate and transmit the information to subsequent generations, and have a source of energy from the environment to grow and reproduce.
- The molecular building blocks of all modern organisms, nucleic acids, proteins, lipids, and carbohydrates, are conserved and are constructed from small, repeating subunits.

The blueprint of life: the genome

- Biological information is typically stored in the nucleic acid DNA and is referred to as the genome.

- A gene is typically defined as a region of DNA that controls a discrete hereditary characteristic. The number and arrangement of genes within a genome varies from species to species; gene number is a poor indicator of biological complexity.

Gene expression

- The first stage of gene expression is transcription, in which an RNA copy of the gene is synthesized. The product of transcription may be a functional RNA, which does not encode a protein (a non-coding RNA), or a messenger RNA (mRNA), which does encode a protein (a coding RNA).
- mRNAs are translated to produce a protein product in a process mediated by the ribosome. During translation, the nucleotide sequence of the mRNA is deciphered in

three-base triplets called codons; each codon represents a particular amino acid (or tells the ribosome to stop translation).

- A single gene may encode more than one RNA or protein product such that the estimated 30 000 human genes are estimated to encode more than one million gene products.
- The regulation of gene expression in both time and space is extensive and occurs at every level of expression.

The cellular basis of gene expression

- Eukaryotic cells feature membrane-bound compartments called organelles, such as the nucleus which contains the genome. Bacteria and archaea lack organelles, though the genome is sequestered in a region of the cell called the nucleoid.
- Subcellular compartmentalization provides opportunities to increase the efficiency of cellular events and to regulate biological processes.

Expression of whole genomes

- The sequence of an organism's genome is its genotype. The overall expression of an organism's genome determines its phenotype (its physical features and properties).

- Null or loss of function mutations eliminate the function of a gene; missense mutations change the identity of a single amino acid within a protein product; and nonsense mutations result in the premature termination of translation to give a truncated protein product.

Evolution of the genome

- All modern-day life forms evolved from a single common ancestor.
- Random mutations that confer advantageous traits upon an organism will be selected for – that is, there will be an increased likelihood of such mutations being transmitted from generation to generation. This is the basis of the evolution of organisms through the process of natural selection.
- At the highest level, sequence similarities and differences allow us to divide all of organismal life into three domains – bacteria, archaea, and eukaryotes.

Studying genome function

- The study of model organisms provides insights into molecular components and processes, helping us to understand how genomes function.
- Viruses exploit the host cellular machinery to ensure their own propagation. Consequently, the study of viruses has helped to illuminate how the host machinery operates.