Part One
Basic Principles of Human Genetics

COPYRIGHTED MATERIAL

CHAPTER 1

DNA Structure and Function

Introduction

The 20th century will likely be remembered by historians of biological science for the discovery of the structure of DNA and the mechanisms by which information coded in DNA is translated into the amino acid sequence of proteins. Although the story of modern human genetics begins about 50 years before the structure of DNA was elucidated, we will start our exploration here. We do so because everything we know about inheritance must now be viewed in the light of the underlying molecular mechanisms. We will see here how the structure of DNA sets the stage both for its replication and for its ability to direct the synthesis of proteins. We will also see that the function of the system is tightly regulated, and how variations in the structure of DNA can alter function. The story of human genetics did not begin with molecular biology, and it will not end there, as knowledge is now being integrated to explain the behavior of complex biological systems. Molecular biology, however, remains a key engine of progress in biological understanding, so it is fitting that we begin our journey here.

Key Points

- DNA consists of a double-helical sugar–phosphate structure with the two strands held together by hydrogen bonding between adenine and thymine or cytosine and guanine bases.
- DNA replication involves local unwinding of the double helix and copying a new strand from the base sequence of each parental strand. Replication proceeds bidirectionally from multiple start sites in the genome.
- DNA is complexed with proteins to form a highly compacted chromatin fiber in the nucleus.
- Genetic information is copied from DNA into messenger RNA (mRNA) in a highly regulated process that involves activation or repression of individual genes.
- mRNA molecules are extensively processed in the nucleus, including removal of introns and splicing together of exons, prior to export to the cytoplasm for translation into protein.
- The base sequence of mRNA is read in triplet codons to direct the assembly of amino acids into protein on ribosomes.
- Some genes are permanently repressed by epigenetic marks such as methylation of cytosine bases. These include most genes on one of two X chromosomes in cells in females and one of the two copies of imprinted genes.

Human Genetics and Genomics, Fourth Edition. Bruce R. Korf and Mira B. Irons.
© 2013 John Wiley & Sons, Ltd. Published 2013 by John Wiley & Sons, Ltd.

Deoxyribonucleic Acid

Mendel described dominant and recessive inheritance before the concept of the gene was introduced and long before the chemical basis of inheritance was known. Cell biologists during the late 19th and early 20th centuries had established that genetic material resides in the cell nucleus and **DNA** was known to be a major chemical constituent. As the chemistry of DNA came to be understood, for a long time it was considered to be too simple a molecule – consisting of just four chemical building blocks, the bases **adenine**, **guanine**, **thymine**, and **cytosine**, along with sugar and phosphate – to account for the complexity of genetic transmission. Credit for recognition of the role of DNA in inheritance goes to the landmark experiments by Oswald Avery and his colleagues, who demonstrated that a phenotype of smooth or rough colonies of the bacterium *Pneumococcus* could be transmitted from cell to cell through DNA alone. Elucidation of the structure of DNA by James Watson and Francis Crick in 1953 opened the door to understanding the mechanisms whereby this molecule functions as the agent of inheritance (Sources of Information 1.1).

Sources of Information 1.1 Mendelian Inheritance in Man

Dr. Victor McKusick and his colleagues at Johns Hopkins School of Medicine began to catalog genes and human genetic traits in the 1960s. The first edition of the catalog *Mendelian Inheritance in Man* was published in 1966. Multiple print editions subsequently appeared, and now the catalog is maintained on the Internet as "Online Mendelian Inheritance in Man" (OMIM), located at www.omim.org.

OMIM is recognized as the authoritative source of information about human genes and genetic traits. The catalog can be searched by gene, phenotype, gene locus, and many other features. The catalog provides a synopsis of the gene or trait, including a summary of clinical features associated with mutations. There are links to other databases, providing access to gene and amino acid sequences, mutations, and so on. Each entry has a unique six-digit number, the MIM number. Autosomal dominant traits have entries beginning with 1, recessive traits with 2, X linked with 3, and mitochondrial with 5. Specific genes have MIM numbers that start with 6.

Throughout this book, genes or genetic traits will be annotated with their corresponding MIM number to remind the reader that more information is available on OMIM and to facilitate access to the site.

DNA Structure

DNA consists of a pair of strands of a sugar–phosphate backbone attached to a set of **pyrimidine** and **purine** bases (Figure 1.1). The sugar is deoxyribose – ribose missing an oxygen atom at its 2′ position. Each DNA strand consists of alternating deoxyribose molecules connected by phosphodiester bonds from the 5′ position of one deoxyribose to the 3′ position of the next. The strands are bound together by hydrogen bonds between adenine and thymine bases and between guanine and cytosine bases. Together these strands form a right-handed double helix. The two strands run in opposite (antiparallel) directions, so that one extends from 5′ to 3′ while the other goes from 3′ to 5′.

The key feature of DNA, wherein resides its ability to encode information, is in the sequence of the four bases (Methods 1.1).

✓ Methods 1.1 Isolation of DNA

DNA, or in some cases RNA, is the starting point for most experiments aimed at studying gene structure or function. DNA can be isolated from any cell that contains a nucleus. The most commonly used tissue for human DNA isolation is peripheral blood, where white blood cells provide a readily accessible source of nucleated cells. Other commonly used tissues include cultured skin fibroblasts, epithelial cells scraped from the inner lining of the cheek, and fetal cells obtained by amniocentesis or chorionic villus biopsy. Peripheral blood lymphocytes can be transformed with Epstein–Barr virus into immortalized cell lines, providing permanent access to growing cells from an individual.

Nuclear DNA is complexed with proteins, which must be removed in order for the DNA to be analyzed. For some experiments it is necessary to obtain highly purified DNA, which involves digestion or removal of the proteins. In other cases, relatively crude preparations suffice. This is the case, for example, with DNA isolated from cheek scrapings. The small amount of DNA isolated from this source is usually released from cells with minimal effort to remove proteins. This preparation is adequate for limited analysis of specific gene sequences. DNA preparations can be obtained from very minute biological specimens, such as drops of dried blood, skin cells, or hair samples isolated from crime scenes for forensic analysis.

Isolation of RNA involves purification of nucleic acid from the nucleus and/or cytoplasm. This RNA can be used to study the patterns of gene expression in a particular tissue. RNA tends to be less stable than DNA, requiring special care during isolation to avoid degradation.

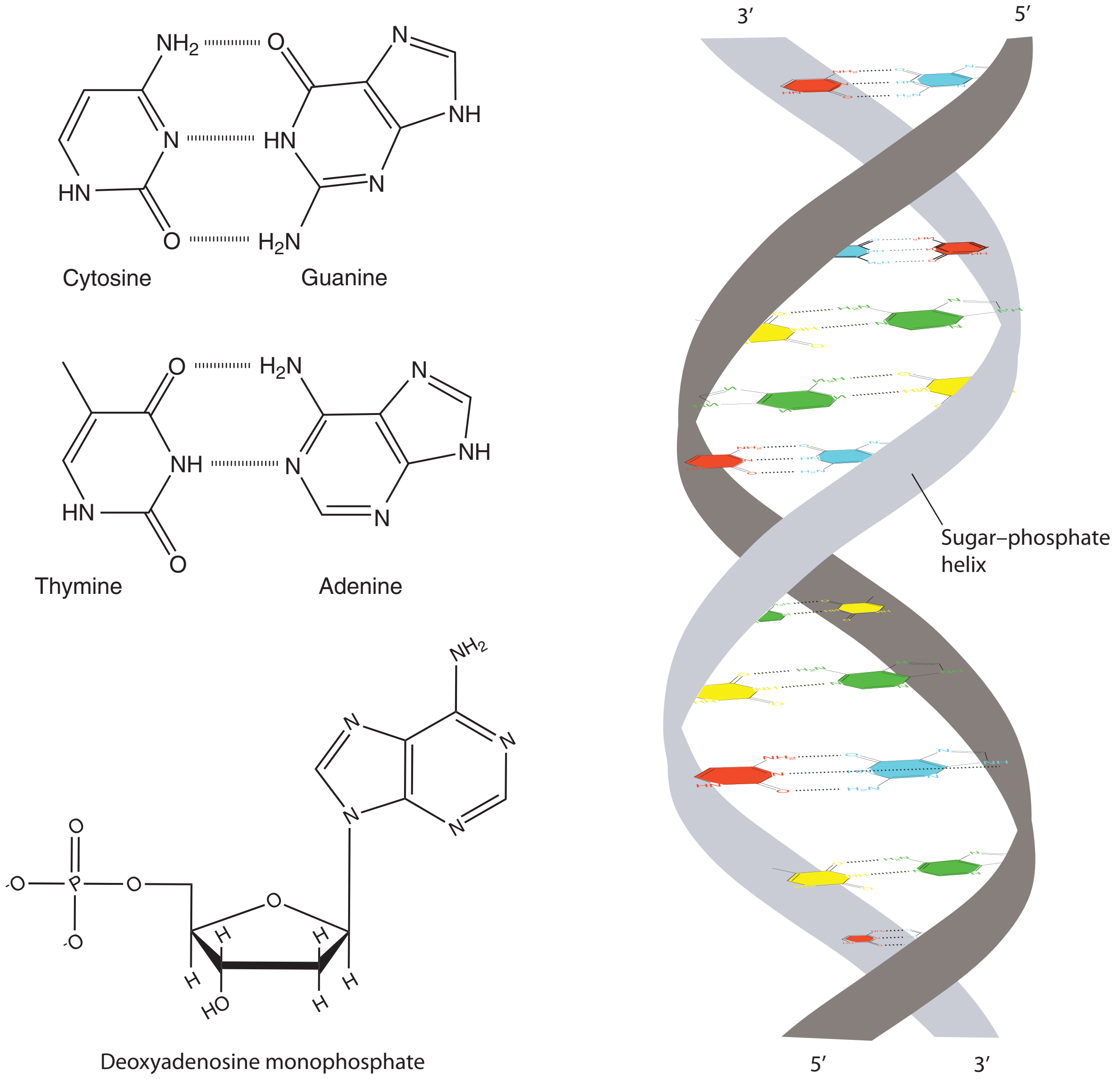

Figure 1.1 Double helical structure of DNA. The double helix is on the right-hand side. The sugar–phosphate helices are held together by hydrogen bonding between adenine and thymine (A–T) bases, or guanine and cytosine (G–C) bases. Deoxyadenosine monophosphate is shown at the bottom left.

The number of adenine bases (A) always equals the number of thymines (T), and the number of cytosines (C) always equals the number of guanines (G). This is because A on one strand is always paired with T on the other, and C is always paired with G. The pairing is noncovalent, due to hydrogen bonding between complementary bases. G–C base pairs form three hydrogen bonds, whereas A–T pairs form two, making the G–C pairs slightly more thermodynamically stable. Because the pairs always include one purine base (A or G) and one pyrimidine base (C or T), the distance across the helix remains constant.

DNA Replication

The complementarity of A to T and G to C provides the basis for DNA replication, a point that was recognized by Watson and Crick in their paper describing the structure of DNA. DNA replication proceeds by a localized unwinding of the double helix, with each strand serving as a template for replication of a new sister strand (Figure 1.2). Wherever a G base is found on one strand, a C will be placed on the growing strand; wherever a T is found, an A will be placed; and so on. Bases are positioned in the newly synthesized strand by hydrogen

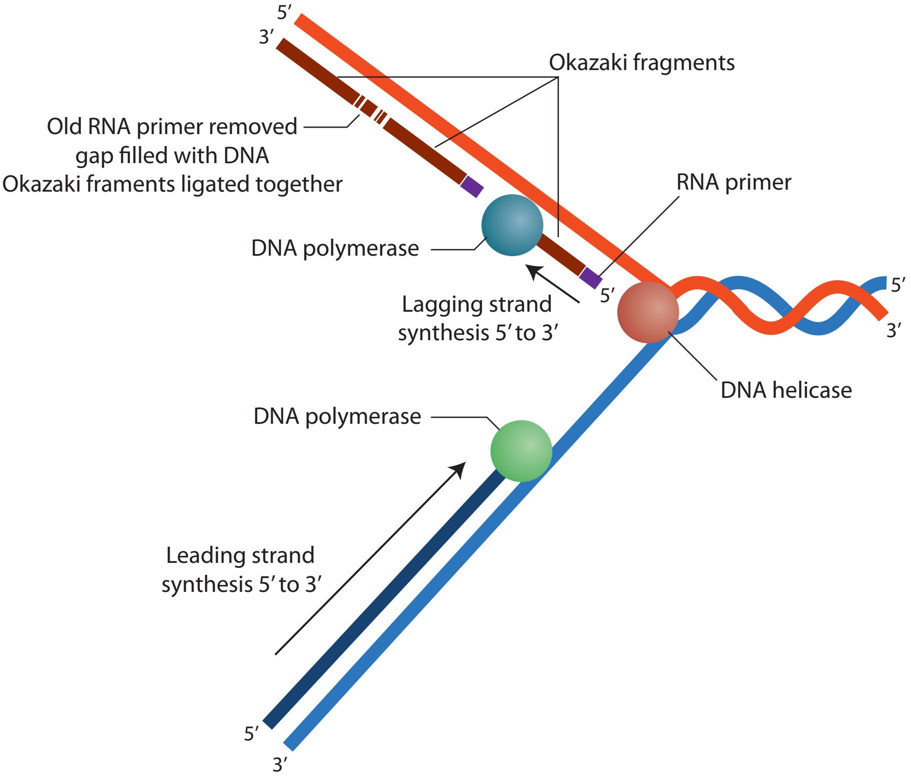

Figure 1.2 DNA replication proceeds in a 5′ to 3′ direction. This occurs with the direct addition of bases to the leading DNA strand. In the other direction (the lagging strand), replication begins with the creation of short RNA primers. DNA bases are added to the primers, and short segments, called Okazaki fragments, are ligated together. The DNA at the replication fork is unwound by a helicase enzyme.

bonding, and new phosphodiester bonds are formed in the growing strand by the enzyme **DNA polymerase**. This is referred to as **semiconservative replication**, because the newly synthesized DNA double helices are hybrid molecules that consist of one parental strand and one new daughter strand. Unwinding of the double helix is accomplished by another enzyme system, **helicase**.

DNA replication requires growth of a strand from a preexisting primer sequence. Primer sequences are provided by transcription of short RNA molecules from the DNA template. **RNA** is a single-stranded nucleic acid, similar to DNA except that the sugar molecules are ribose rather than deoxyribose and uracil substitutes for thymine (and pairs with adenine). During DNA replication, short RNA primers are transcribed and then extended by DNA polymerase. DNA is synthesized in a 5′ (exposed phosphate on the 5′ carbon of the ribose molecule) to 3′ (exposed hydroxyl on the 3′ carbon) direction. For one strand, referred to as the **leading strand**, this can be accomplished continuously as the DNA unwinds. The other strand, called the **lagging strand**, is replicated in short segments, called **Okazaki fragments**, which are then enzymatically ligated together by DNA ligase. Distinct polymerases replicate the leading and lagging strands. The short RNA primers are ultimately removed and replaced with DNA to complete the replication process.

The human genome consists of over 3 billion base pairs of DNA packaged into 23 pairs of chromosomes. Each chromosome consists of a single, continuous DNA molecule, encompassing tens to hundreds of millions of base pairs. If the DNA on each chromosome were to be replicated in a linear manner from one end to another, the process would go on interminably – certainly too long to sustain the rates of cell division that must occur. In fact, the entire genome can be replicated in a matter of hours because replication occurs simultaneously at multiple sites along a chromosome. These origins of replication are bubble-like structures from which DNA replication proceeds bidirectionally until an adjacent bubble is reached (Figure 1.3).

One special case in DNA replication is the replication of the ends of chromosomes. Removal of the terminal RNA primer from the lagging strand at the end of a chromosome would result in shortening of the end, since there is no upstream

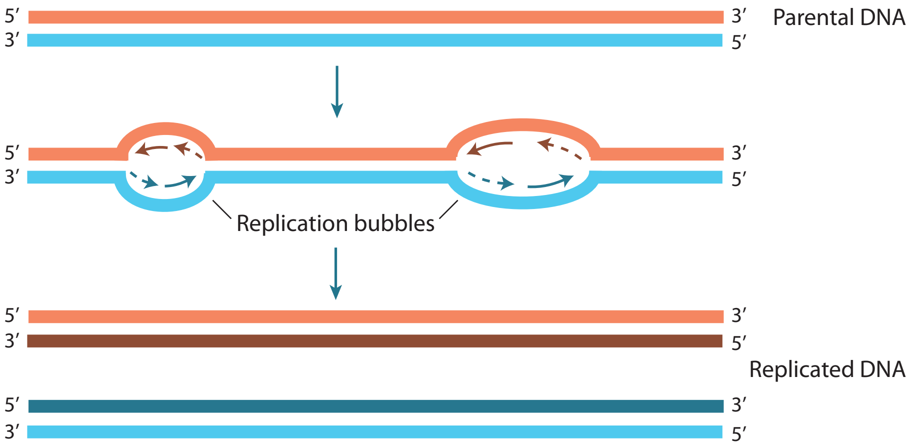

Figure 1.3 DNA replication proceeds bidirectionally from multiple start sites, creating replication bubbles.

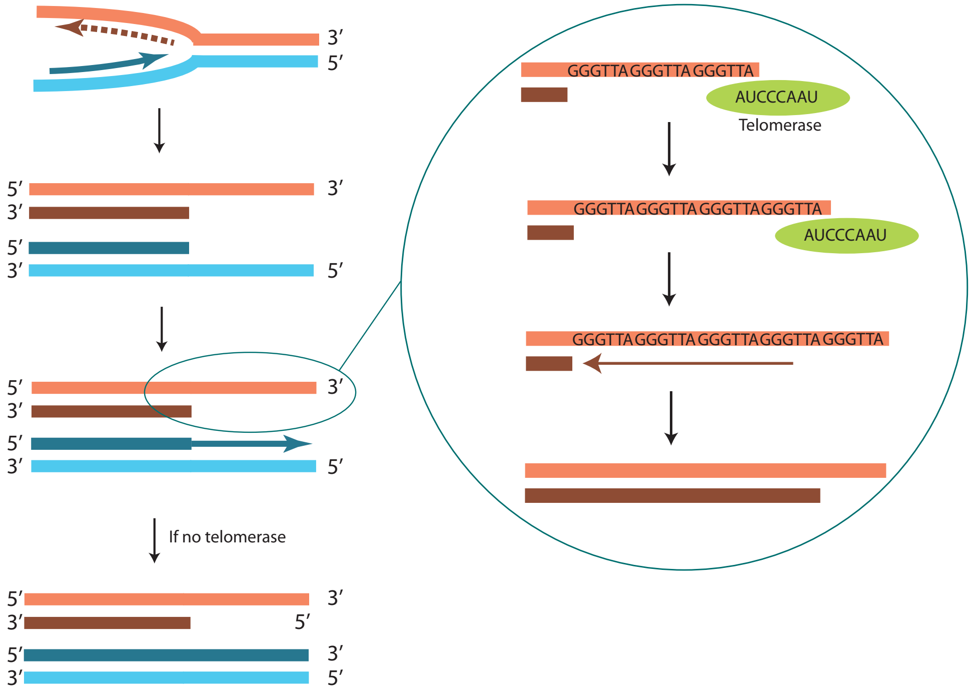

Figure 1.4 The lagging strand cannot replicate the end of the chromosome, since there is no site at which the RNA polymerase can bind (as it would be off the end of the chromosome). This is shown to the left, where the replicated chromosome has a shortened 5′ end on one strand. If telomerase is present (right), the 3′ end is extended, using an RNA template intrinsic to the enzyme. The lagging strand is then replicated using a DNA polymerase. Extension of the telomere by telomerase prevents erosion of the end of the chromosome during replication.

primer for DNA polymerase to replace the short RNA primer. This problem is circumvented by action of an enzyme called telomerase, which uses an RNA template intrinsic to the enzyme to add a stretch of DNA onto the 3′ end of the lagging strand (Figure 1.4). The DNA sequence of the **telomere** is determined by the RNA sequence in the enzyme; for humans the sequence is GGGTTA. Each chromosome end has a tandem repeat of thousands of copies of the telomere sequence that is replicated during early development. Somatic cells may replicate without telomerase activity, resulting in a

Clinical Snapshot 1.1 Dyskeratosis Congenita

Eddy is a 4-year-old boy brought in by his parents because of recurrent cough. He has had two bouts of pneumonia that required treatment with antibiotics over the past 2 months. Now he is sick again, having never stopped coughing since the last episode of pneumonia. His parents have noted that he has had low energy over the past several weeks. His examination shows a fever of 39°C and rapid respirations with frequent coughing. His breath sounds are abnormal on the right side of his chest. He also has hyperkeratotic skin with streaky hyperpigmentation. His finger and toenails are thin and broken at the ends, and his hair is sparse. A blood count shows anemia and a reduced number of white blood cells. A bone marrow aspirate is obtained, and it shows generalized decrease in all cell lineages. A clinical diagnosis of dyskeratosis congenita is made.

Dyskeratosis congenita consists of reticulated hyperpigmentation of the skin, dystrophic hair and nails, and generalized bone marrow failure (Figure 1.5). It usually presents in childhood, often with signs of pancytopenia. There is an increased rate of spontaneous chromosome breakage seen in peripheral blood lymphocytes. Dyskeratosis congenita can be inherited as an X-linked recessive (MIM 305000), autosomal dominant (MIM 127550), or autosomal recessive (MIM 224230) trait. The X-linked form is due to mutation in a gene that encodes the protein dyskerin (MIM 300126). Dyskerin is involved in the synthesis of ribosomal RNA and also interacts with telomerase. The autosomal dominant form is due to mutation in the gene *hTERC* (MIM 602322). *hTERC* encodes the RNA component of telomerase. The autosomal recessive form can be due to mutations in *NOLA2* (MIM 606470) or *NOLA3* (MIM 606471); both encode proteins that interact with dyskerin. The X-linked recessive form is more severe and earlier in onset than the dominant form. Both forms are associated with defective telomere functioning, leading to shortened telomeres. This likely leads to premature cell death and also explains the spontaneous chromosome breakage. The phenotype of the X-linked form may also be due, in part, to defective rRNA processing.

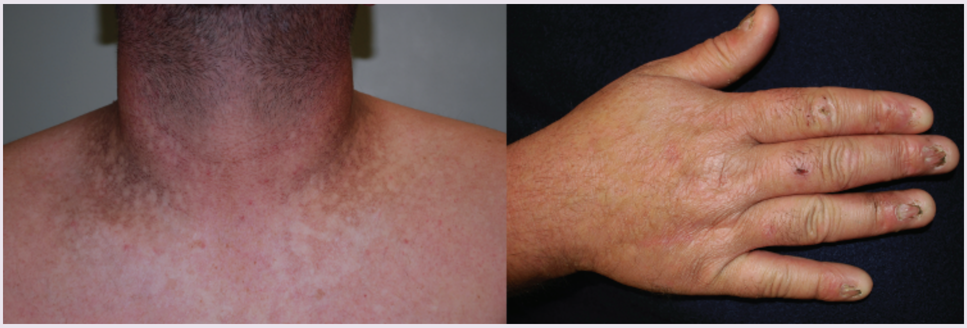

Figure 1.5 Skin and nail changes in an individual with dyskeratosis congenita. (Courtesy of Dr. Amy Theos, University of Alabama at Birmingham.)

gradual shortening of the ends of the chromosomes with successive rounds of replication. This may be one of the factors that limits the number of times a cell can divide before it dies, a phenomenon known as **senescence** (Clinical Snapshot 1.1).

Chromatin

The DNA within each cell nucleus must be highly compacted to accommodate the entire genome in a very small space. The enormous stretch of DNA that comprises each chromosome is actually a highly organized structure (Figure 1.6). The DNA double helix measures approximately 2 nm in diameter, but DNA does not exist in the nucleus in a "naked" form. It is complexed with a set of lysine- and arginine-rich proteins called histones. Two molecules of each of four major histone types – H2A, H2B, H3, and H4 – associate together with about every 146 base pairs to form a structure known as the **nucleosome**, which results in an 11 nm thick fiber. Nucleosomes are separated from one another by up to 80 base pairs, like beads on a string. This is more or less the conformation of actively transcribed chromatin but, during periods of inactivity, some regions of the genome are more highly compacted. The next level of organization is the coiling of nucleosomes

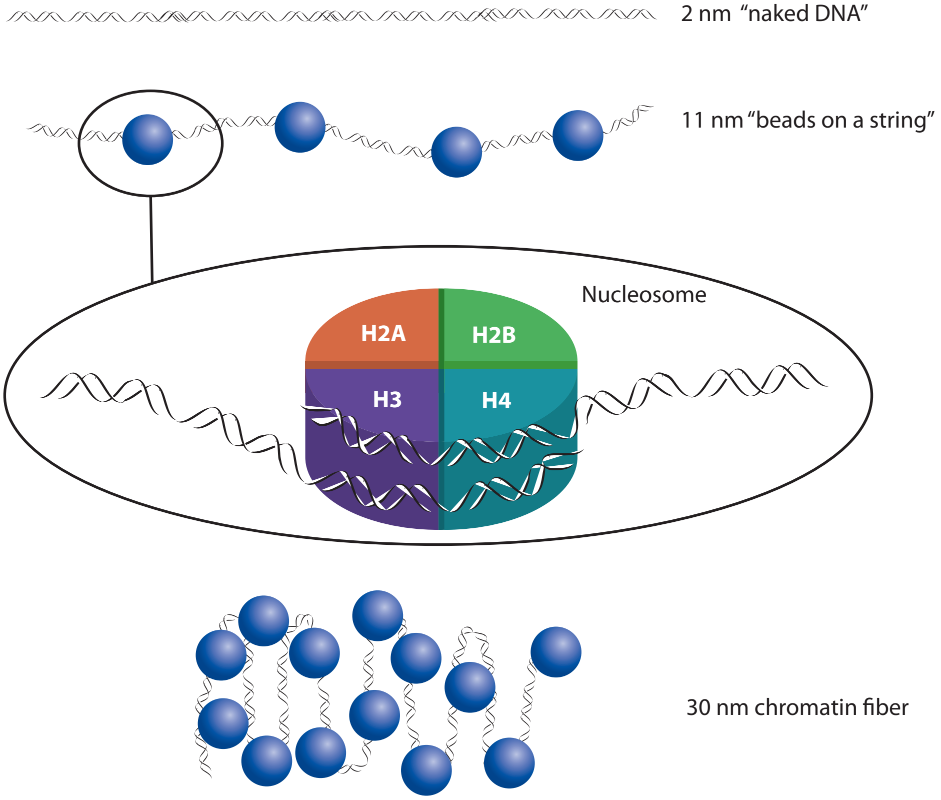

Figure 1.6 Levels of organization of chromatin. The DNA double helix has a width of approximately 2 nm. The fundamental unit of chromatin is the nucleosome, which consists of 146 base pairs of DNA wound around a core consisting of two copies of each of the four histone proteins (H2A, H2B, H3, and H4). These are arrayed as beads on a string. The diameter of a nucleosome is 11 nm. The nucleosomes are, in turn, condensed into a structure measuring 30 nm. This is further coiled and condensed to compose a metaphase chromosome.

into a 30 nm thick chromatin fiber held together by another histone, H1, and other nonhistone proteins. Chromatin is further compacted into the highly condensed structures comprising each chromosome, with the maximum condensation occurring during the metaphase stage of mitosis (see Chapter 6).

Gene Function

The basic tenet of molecular genetics – often referred to as the central dogma – is that DNA encodes RNA, which in turn encodes the **amino acid** sequence of proteins. It is now clear that this is a simplified view of the function of the genome. As will be seen in Chapter 4, much of the DNA sequence does not encode protein. A large proportion of the genome consists of noncoding sequences, such as repeated DNA, or encodes RNA that is not translated into protein. Nevertheless, the central dogma remains a critical principle of genome function. We will explore here the flow of information from DNA to RNA to protein.

Transcription

The process of copying the DNA sequence of a gene into **messenger RNA (mRNA)** is referred to as **transcription**. Some genes are expressed nearly ubiquitously. These are referred to as **housekeeping genes**. They include genes necessary for cell replication or metabolism. For other genes, expression is tightly controlled, with particular genes being turned on or off in particular cells at specific times in development or in response to physiological signals. Genomic studies are now being applied to analysis of gene regulation and are revealing remarkable details on the structure and function of regulatory elements in the human genome (Hot Topic 1.1).

Hot Topic 1.1 Encode Project

The sequencing of the human genome (see Chapter 2) has made it possible to identify the full complement of genes, but does not in itself reveal how these genes are regulated. In 2003 a project was launched to characterize the entire set of coding elements in the genome, including identification of all genes and their corresponding regulatory sequences. The project is referred to as ENCODE, which is an acronym for *encyclopedia of DNA elements*. A pilot study was published in 2007 and more complete results were published in a series of papers in 2012.

The project has revealed a number of surprises. A total of 20687 protein-encoding genes (not representing the full complement) were identified, each with, on average, 6.3 alternatively spliced transcripts. Among RNAs that do not encode protein, there were 8801 small RNAs and 9640 long non-coding transcripts. More than 80% of the genome was found to be transcribed into some kind of RNA in at least one cell type, a much higher level of transcriptional activity than expected. Ninety-five percent of the genome was found to be within 8 kb of a region involved in DNA/protein interaction. Nearly 400000 regions were found to have features of enhancers and over 70000 had features of promoter activity. Genetic variants associated with common disease (see Chapter 5) were found to be enriched in regions with functional elements that do not encode protein. It has become clear that much of the genome is dedicated to encoding elements responsible for fine control of gene expression in distinct cell types.

Gene expression is regulated by proteins that bind to DNA and either activate or repress transcription. The anatomy of elements that regulate gene transcription is shown in Figure 1.7. The **promoter region** is immediately adjacent to the transcription start site, usually within 100 base pairs. Most promoters include a base sequence of T and A bases called the TATA box. In some cases there may be multiple, alternative promoters at different sites in a gene that respond to regulatory factors in different tissues. Regulatory sequences may occur adjacent to the promoter or may be located thousands of base pairs away. These distantly located regulatory sequences are known as enhancers. Enhancer sequences function regardless of their orientation with respect to the gene.

DNA-binding proteins may serve as repressors or activators of transcription, and may bind to the promoter, to upstream regulatory regions, or to more distant enhancers. Activator or repressor proteins are regulated by binding of specific ligands. Ligand binding changes the confirmation of the transcription factor and may activate it or inactivate it. The ligand is typically a small molecule, such as a hormone. Many transcription factors form dimers, either homodimers of two identical proteins or heterodimers of two different proteins. There may also be **corepressor** or **coactivator** proteins. Some transcription

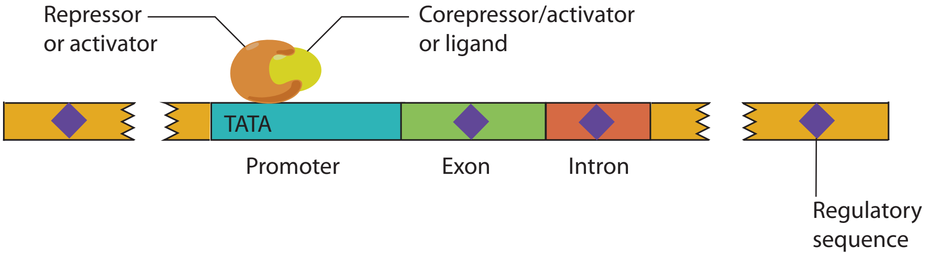

Figure 1.7 Cis-acting elements regulating gene expression. Transcription starts at the promoter by the binding of an RNA polymerase. Control of gene expression occurs via the binding of transcription factors upstream of the transcription start site at the TATA box. These factors can be either activators or repressors, and may bind coactivators or corepressors. Other DNA regulatory elements can occur within exons or introns or at some distance upstream or downstream of a gene.

factors stay in the cytosol until the ligand binds or some other activation process occurs, at which time they move to the nucleus to activate their target gene(s). In other situations, the transcription factors reside in the nucleus most of the time and may even be located at the response element sequences, but without the ligand they are inactive or even repress transcription.

Transcription begins with the attachment of the enzyme RNA polymerase to the promoter (Figure 1.8). There are three major types of RNA polymerase, designated types I, II, and III. Most gene transcription is accomplished by RNA polymerase II. Type I is involved in the transcription of **ribosomal RNA (rRNA)**, and type III transcribes **transfer RNA (tRNA)** (see the "Translation" section of this chapter). The polymerase reads the sequence of the DNA template strand, copying a complementary RNA molecule, which grows from the 5′ to the 3′ direction. The resulting mRNA is an exact copy of the DNA sequence, except that uracil takes the place of thymine in RNA. Soon after transcription begins, a 7-methyl guanine residue is attached to the 5′-most base, forming the cap.

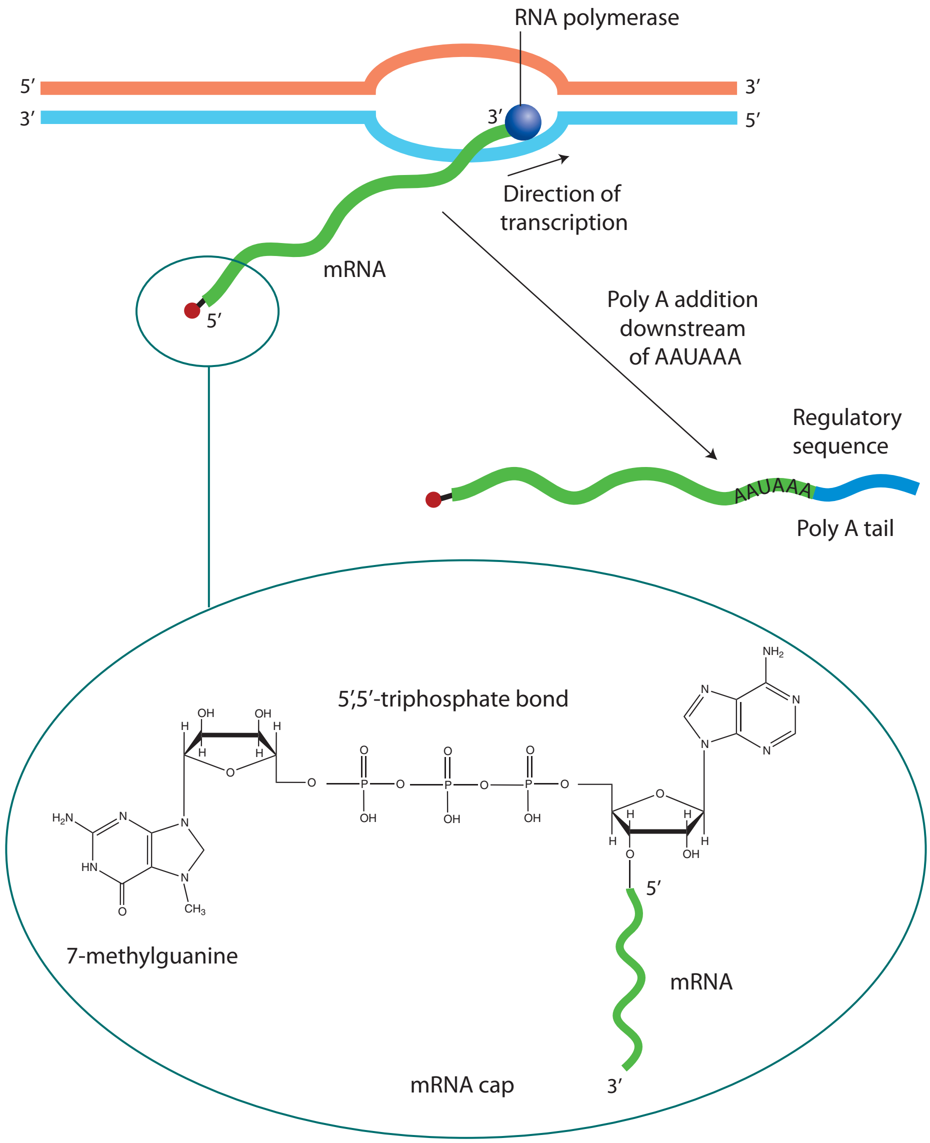

Figure 1.8 Transcription involves copying RNA from one strand of DNA. The reaction is catalyzed by RNA polymerase. A 7-methylguanosine cap is added to the 5′ end of most mRNA molecules before transcription is completed, and a poly-A tail is added enzymatically to the 3′ end downstream of an AAUAAA sequence.

Transcription proceeds through the entire coding sequence. Some genes include a sequence near the 3′ end that signals RNA cleavage at that site and enzymatic addition of 100 to 200 adenine bases, the poly-A tail. Polyadenylation is characteristic of housekeeping genes, which are expressed in most cell types. Both the 5′ cap and the poly-A tail stabilize the mRNA molecule and facilitate its export to the cytoplasm.

The DNA sequence of most genes far exceeds the length required to encode their corresponding proteins. This is accounted for by the fact that the coding sequence is broken into segments, called **exons**, which are interrupted by noncoding segments called **introns**. Some exons may be less than a hundred bases long, whereas introns can be several thousand bases in length; therefore, much of the length of a gene may be devoted to introns. The number of exons in a gene may be as few as one or two, or may number in the dozens. The processing of the transcript into mature mRNA requires the removal of the introns and splicing together of the exons (Figure 1.9), carried out by an enzymatic process that occurs in the nucleus. The 5′ end of an intron always consists of the two bases GU, followed by a consensus sequence that is similar, but not identical, in all introns. This is the **splice donor**. The 3′ end, the **splice acceptor**, ends in AG, preceded by a consensus sequence.

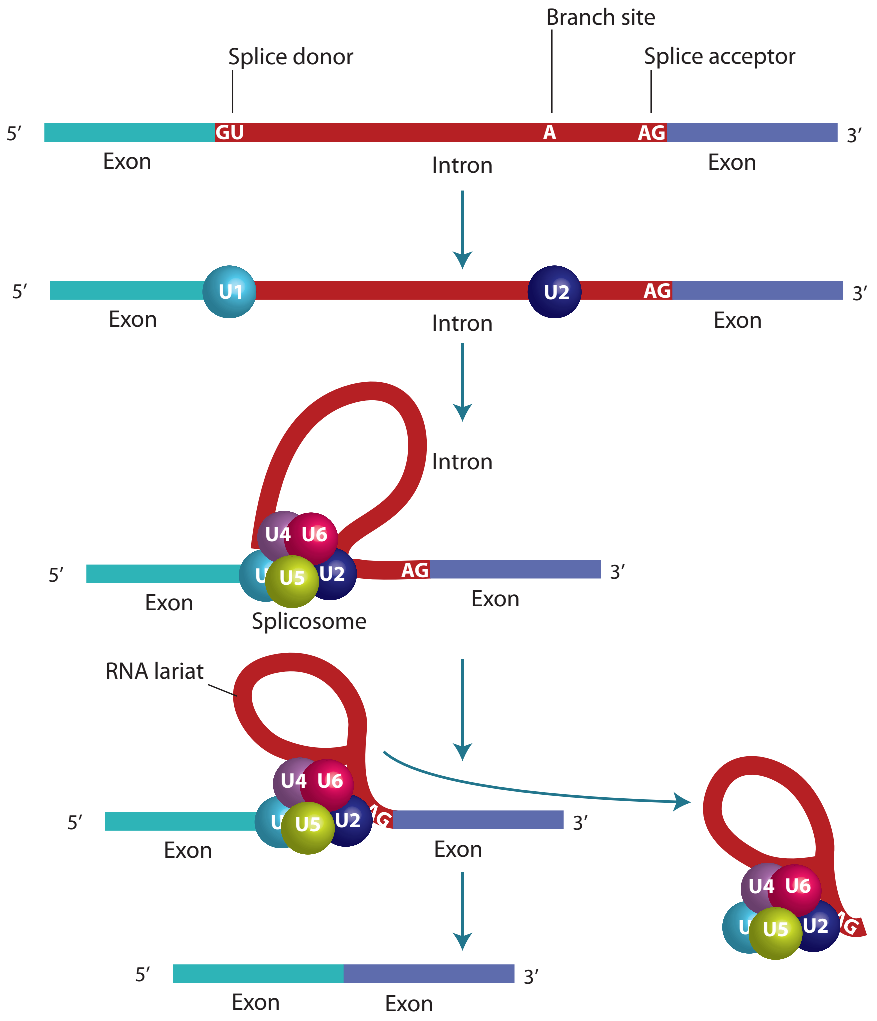

Figure 1.9 RNA splicing begins with the binding of specific ribonucleoproteins (U1 and U2) to the splice donor and branch site. These two sites are then brought together by other components of the splicosome. The donor site is then cut, and the free end of the intron binds to the branch point within the intron to form a lariat structure. Then the acceptor site is cleaved, releasing the lariat, and the exons at the two ends are ligated together.

The splicing process requires complex machinery composed of both proteins and small RNA molecules (**small nuclear RNA**, or **snRNA**), consisting of fewer than 200 bases. snRNA is also transcribed by RNA polymerase II. The splice is initiated by binding of a protein–RNA complex to the splice donor, at a point within the intron called the **branch point**, and the splice acceptor. First the RNA is cleaved at the donor site and this is attached in a 5′–2′ bond to the branch point. Then the acceptor site is cleaved, releasing a lariat structure that is degraded, and the 5′ and 3′ ends are ligated together. The splicing process also requires the function of proteins, **SR proteins**, which are involved in selecting sites for the initiation of splicing. These proteins interact with sequences known as **splice enhancers** or **silencers**. The splicing process is vulnerable to disruption by mutation, as might be predicted from its complexity (see Chapter 4).

The RNA-splicing process offers a point of control of gene expression. Under the influence of control molecules present in specific cells, particular exons may be included or not included in the mRNA due to differential splicing (Figure 1.10). This results in the potential to produce multiple distinct proteins from the same gene, adding greatly to the diversity of proteins encoded by the genome. Specific exons may correspond with particular functional domains of proteins, leading to the production of multiple proteins with diverse functions from the same gene. Some mRNAs are subject to RNA editing, wherein a specific base may be enzymatically modified. For example, the protein apolipoprotein B exists in two forms, a 48 kDa form made in the intestine and a 100 kDa form in the liver. Both forms are the product of the same gene. In the intestine the enzyme cytidine deaminase alters a C to a U at codon 2153, changing the codon from CAA (encoding glutamine) to UAA (a stop codon). This truncates the peptide, accounting for the 48 kDa form.

MicroRNA

Gene regulation is not limited to control at the level of gene transcription. There is another level of posttranscriptional control that involves RNA molecules that do not encode protein, referred to as **microRNAs (miRNAs)** (Figure 1.11). Several hundred distinct miRNAs are encoded in the human genome. These are transcribed by RNA polymerase II, are capped, and have poly-A tails added. miRNAs form hairpin structures through base pairing. The enzyme Drosha trims the hairpins, which are then exported to the cytoplasm, where the enzyme Dicer further cleaves them. After further processing, single-stranded miRNA molecules associate with a protein complex called the RNA-induced silencing complex (RISC). The RISC then binds to mRNA molecules by base sequence complementarity with the miRNA. This leads to either cleavage and degradation of the mRNA, or reduced rate of translation and eventual degradation of the mRNA. The overall effect is for miRNA to reduce the quantity of protein produced from a transcribed gene. Any specific miRNA might bind to many different mRNAs, leading to coordinated reduction in gene product from a large number of genes.

A similar process also occurs when double-stranded RNA is introduced into the cell by viral infection. The Dicer enzyme cuts the invading RNA into shorter fragments, called small interfering RNA (siRNA), which attach to the RISC. The RISC then binds to additional double-stranded RNA molecules and causes their degradation. This is referred to as RNA interference, and represents a kind of immune system

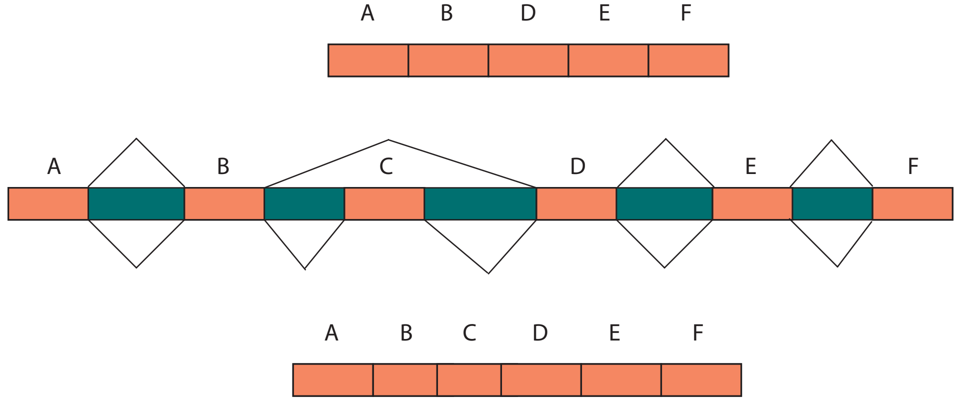

Figure 1.10 Alternative splicing. Splicing out each intron results in the inclusion of exons A–F in the mRNA. Alternatively, a splice can be made directly between exons B and D, skipping exon C. This results in the production of a distinct protein, missing the amino acids encoded by exon C.

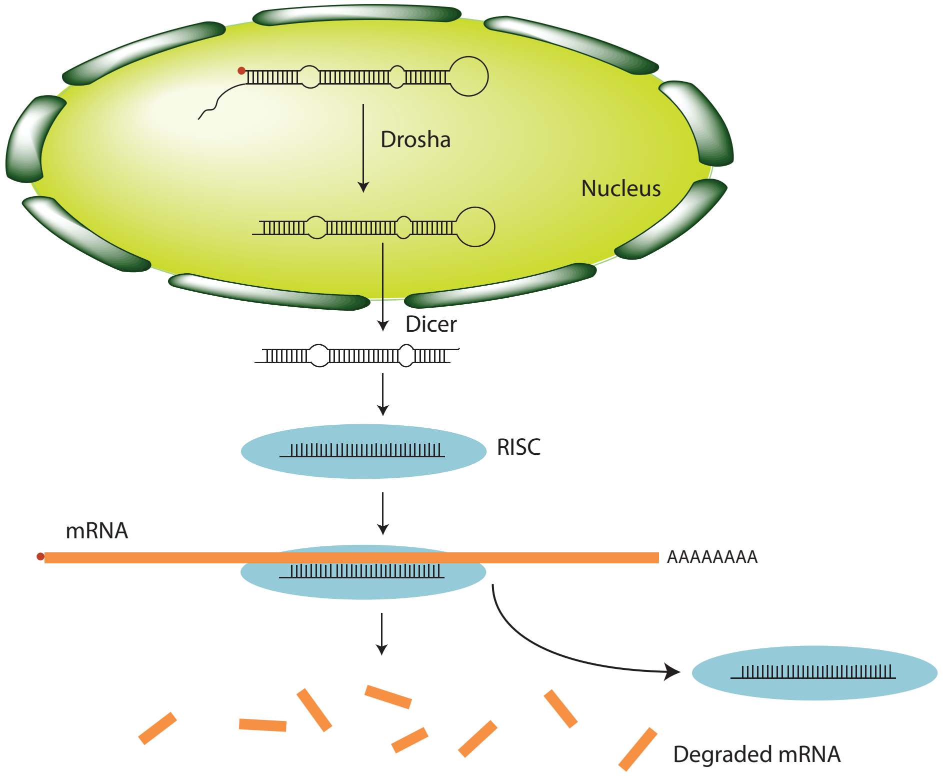

Figure 1.11 Mechanism of action of microRNA (miRNA). miRNA is transcribed and capped, and poly-A is added. The transcripts form hairpins due to internal base complementarity. The enzyme Drosha cleaves the miRNA in the nucleus, after which it is exported to the cytoplasm and further cleaved by Dicer. A single strand of the miRNA then associates with the RNA-induced silencing complex (RISC). Binding of the miRNA to a molecule of mRNA results in cleavage or, if the base pairing is not perfect, a reduced rate of translation.

within the cell. The process has also been used experimentally, wherein siRNA molecules are introduced into the cell to selectively interfere with translation of targeted mRNAs.

Translation

The mature mRNA is exported to the cytoplasm for translation into protein. During translation, the mRNA sequence is read into the amino acid sequence of a protein (Figure 1.12). The translational machinery consists of a protein–RNA complex called the **ribosome**. Ribosomes consist of a complex of proteins and specialized rRNA. The eukaryotic ribosome is composed of two subunits, designated 60S and 40S ("S" is a measurement of density, the Svedberg unit, reflecting how the complexes were initially characterized by density gradient centrifugation). Each subunit includes proteins and an rRNA molecule. The 60S subunit includes a 28S rRNA, and the 40S subunit an 18S rRNA. Ribosomes can be free or associated with the **endoplasmic reticulum (ER)**, also known as the rough ER.

The mRNA sequence is read in triplets, called codons, beginning at the 5′ end of the mRNA, which is always AUG, encoding methionine (although this methionine residue is sometimes later cleaved off). Each codon corresponds with a particular complementary anticodon, which is part of another RNA molecule, tRNA. tRNA molecules bind specific amino acids defined by their anticodon sequence (Table 1.1). Protein translation therefore consists of binding a specific tRNA to the appropriate codon, which juxtaposes the next amino acid in the growing peptide, which is enzymatically linked by an amide bond to the peptide. The process ends when a stop codon is reached (UAA, UGA, or UAG). The peptide is then released from the ribosome for transport to the appropriate site within the cell, or for secretion from the cell. A leader peptide sequence may direct the protein to its final destination in the cell; this peptide is cleaved off upon arrival. Posttranslational modification, such as glycosylation, begins during the translation process and continues after translation is complete.

The process of translation consists of three phases, referred to as initiation, elongation, and termination. Initiation involves the binding of the first amino acyl tRNA, which always carries methionine, to the initiation codon, always AUG. A set of proteins, referred to as **elongation factors**, are involved in the

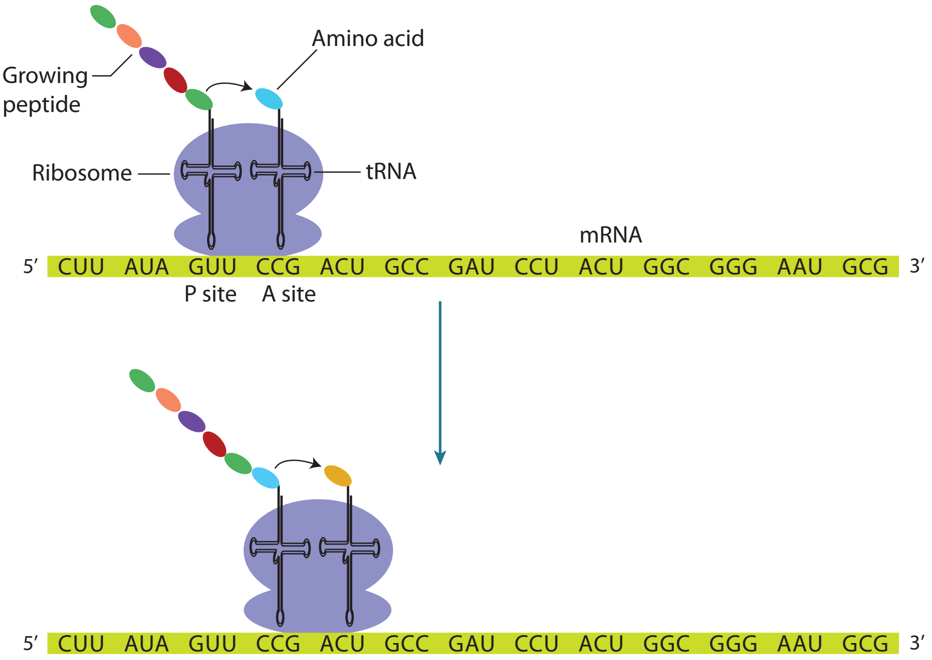
Growing peptide
Amino acid
Ribosome
tRNA
mRNA
5′ CUU AUA GUU CCG ACU GCC GAU CCU ACU GGC GGG AAU GCG 3′
P site
A site
5′ CUU AUA GUU CCG ACU GCC GAU CCU ACU GGC GGG AAU GCG 3′

Figure 1.12 The process of protein translation. Translation takes place at the ribosome, which binds to the mRNA. Specific amino acyl tRNA molecules bind to the mRNA by base pair complementarity between a triplet codon on the mRNA and an anticodon on the tRNA. A peptide bond is formed between the growing peptide and the next amino acyl tRNA, transferring the growing peptide and elongating it by one amino acid. This continues until a stop codon is reached.

Table 1.1 The genetic code. A triplet codon is read from the left column, to the top row, to the full triplet in each box. Each codon corresponds with a specific amino acid, except for the three stop codons (labeled “Ter”). Most amino acids are encoded by more than one codon.

	T	**C**	**A**	**G**
T	TTT Phe (F)	TCT Ser (S)	TAT Tyr (Y)	TGT Cys (C)
	TTC “	TCC “	TAC “	TGC “
	TTA Leu (L)	TCA “	TAA Ter	TGA Ter
	TTG “	TCG “	TAG Ter	TGG Trp (W)
C	CTT Leu (L)	CCT Pro (P)	CAT His (H)	CGT Arg (R)
	CTC “	CCC “	CAC “	CGC “
	CTA “	CCA “	CAA Gln (Q)	CGA “
	CTG “	CCG “	CAG “	CGG “
A	ATT Ile (I)	ACT Thr (T)	AAT Asn (N)	AGT Ser (S)
	ATC “	ACC “	AAC “	AGC “
	ATA “	ACA “	AAA Lys (K)	AGA Arg (R)
	ATG Met (M)	ACG “	AAG “	AGG “
G	GTT Val (V)	GCT Ala (A)	GAT Asp (D)	GGT Gly (G)
	GTC “	GCC “	GAC “	GGC “
	GTA “	GCA “	GAA Glu (E)	GGA “
	GTG “	GCG “	GAG “	GGG “

process, which also requires adenosine triphosphate (ATP) and guanosine triphosphate (GTP). The ribosome binds to the mRNA at two successive codons. One is designated the **P site** and carries the growing peptide chain. The other is the next codon, designated the **A site**. Elongation involves the binding of the next amino acyl tRNA to its anticodon at the A site. This delivers the next amino acid in the peptide chain, which is attached to the growing peptide, with peptide bond formation catalyzed by **peptidyl transferase**. The ribosome then moves on to the next codon under the action of a **translocase**, with energy provided by GTP. When a stop codon is reached, a release factor protein–GTP complex binds and the peptidyl transferase adds an OH to the end of the peptide, which is then released from the ribosome under the influence of proteins called **release factors**.

Epigenetics

Individual genes may be reversibly activated or repressed, but there are some situations in which genes or sets of genes are permanently silenced. This occurs as a result of chemical modifications of DNA that do not change the base sequence, and also involves chemical changes in the associated histone proteins that result in compaction of the chromatin. Gene silencing is characteristic of one of the two copies of the X chromosome in females and on the maternal or paternal copy of imprinted genes. It also can occur on other genes in specific tissues and may be subject to environmental influences.

At the DNA level, gene silencing is accompanied by methylation of cytosine bases to 5-methylcytosine (Figure 1.13). This occurs in regions where cytosine is following by guanine (5′–CpG–3′) near the promoter, sites referred to as **CpG islands**. Methylated sites bind protein complexes that remove acetyl groups from histones, leading to transcriptional repression. The silencing is continued from cell generation to generation because the enzymes responsible for methylation recognize the 5-methylcytosine on the parental strand of DNA and methylate the cytosine on the newly synthesized daughter strand.

X-chromosome inactivation provides a mechanism for equalization of gene dosage on the X chromosome in males, who have one X, and females, who have two. Most genes on one of the two X chromosomes in each cell of a female are permanently inactivated early in development (Figure 1.14). The particular X inactivated in any cell is determined at random, so in approximately 50% of cells one X is inactivated and in the other 50% the other X is inactivated. Regions of homology between the X and Y at the two ends of the X escape inactivation. These are referred to as **pseudoautosomal** regions. The inactive X remains condensed through most of the cell cycle, and can be visualized as a densely staining body during interphase, called the **Barr body**.

Initiation of inactivation is controlled from a region called the **X inactivation center (Xic)**. A gene within this region, known as Xist, is expressed on one of the two X chromosomes early in development. Xist encodes a 25 kb RNA that is not translated into protein, but binds to sites along the X to be inactivated. Subsequently, CpG islands on this chromosome are methylated and histones are deacetylated.

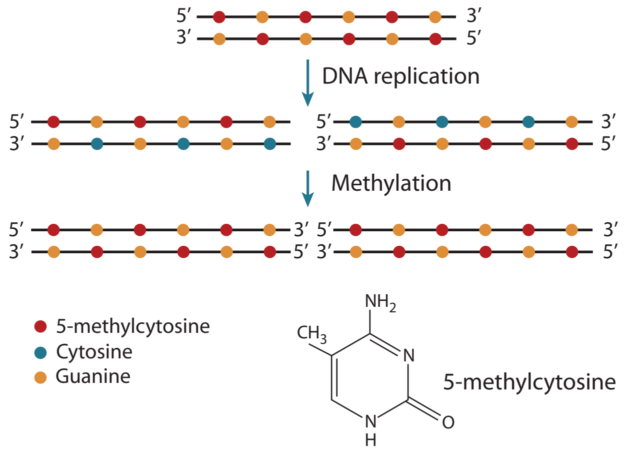

Figure 1.13 Cytosine bases 5′ to guanines are methylated at CpG islands. When DNA is replicated, unmethylated cytosines are first inserted into the daughter strand. A DNA methyltransferase enzyme recognizes the 5-methylcytosine on the other strand and methylates the cytosines on the newly synthesized daughter strand.

Genomic **imprinting** involves the silencing of either the maternal or paternal copy of a gene during early development (Figure 1.15). Like X chromosome inactivation, imprinting is probably accomplished through methylation of specific chromosome regions. The methylation "imprint" is erased in germ cells, so the specific gene copy to be inactivated is always determined by the parent of origin, regardless of whether that particular gene copy was active or inactive in the previous generation. Genomic imprinting applies to only a subset of genes, although the full extent of imprinting is not yet known.

Although methylation at a site tends to repress transcription, the effects of methylation can be complex (Figure 1.16). For example, there can be regulatory signals sent from one locus to another on a chromosome. If the signal is repression and the locus is methylated, this might disinhibit the target locus. Alternatively, a protein might bind to a site that blocks a signal transmitted from one locus to another. Methylation might prevent binding of the protein, permitting the interaction between the two loci. There are many clinical disorders that have been attributed to defects in imprinting. We will discuss these in Chapter 3.

Aside from having a role in X chromosome inactivation, epigenetic changes appear to be involved in silencing genes as a component of normal development or physiological responses

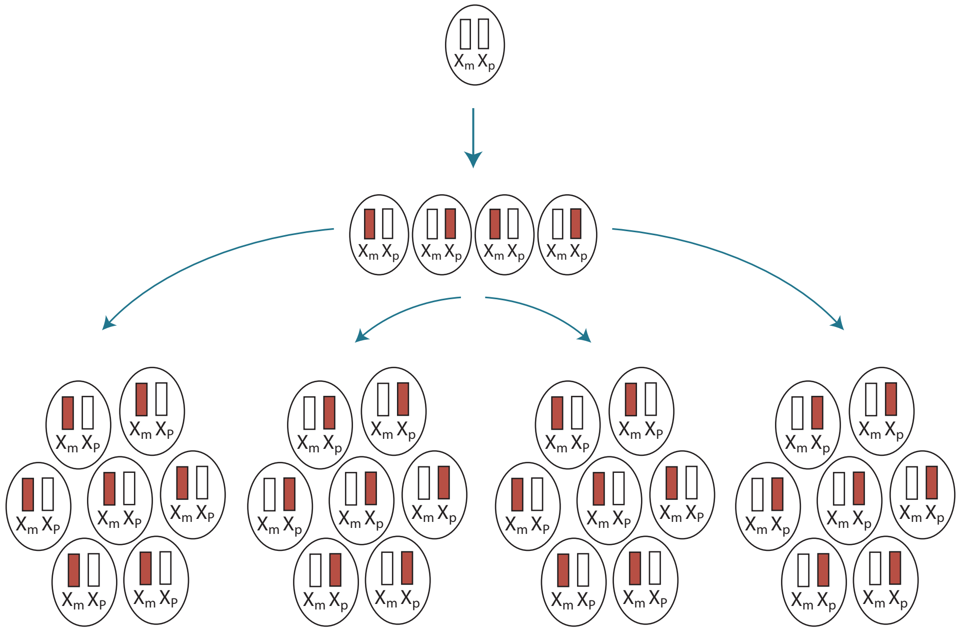

Figure 1.14 X chromosome inactivation. In the zygote, both the maternally and paternally derived X chromosomes (X_m and X_p) are active. Early in development, one of the two X chromosomes in each cell is inactivated (indicated as the red chromosome). This X chromosome remains inactive in all the descendants of that cell.

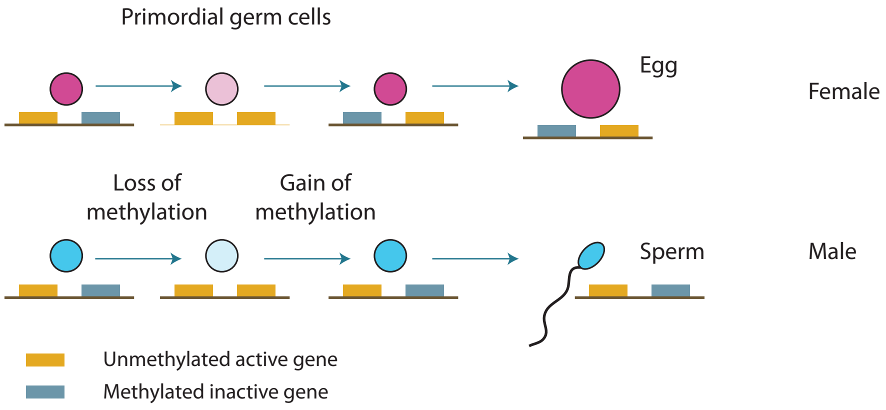

Figure 1.15 Concept of genomic imprinting. An imprinted gene will be methylated and inactive in somatic cells. In this example, a pair of adjacent genes is shown, one methylated if maternally inherited (left gene) and the other methylated if paternally inherited (right gene). In this illustration, both primordial germ cells (after meiosis) inherited the paternal allele; hence, the left-hand gene is unmethylated and the right-hand one is methylated. The methylation imprint is erased as the germ cell develops, and then reestablished according to whether the primordial germ cell is in a male (in which case the right-hand gene is methylated) or female (in which case the left-hand gene is methylated).

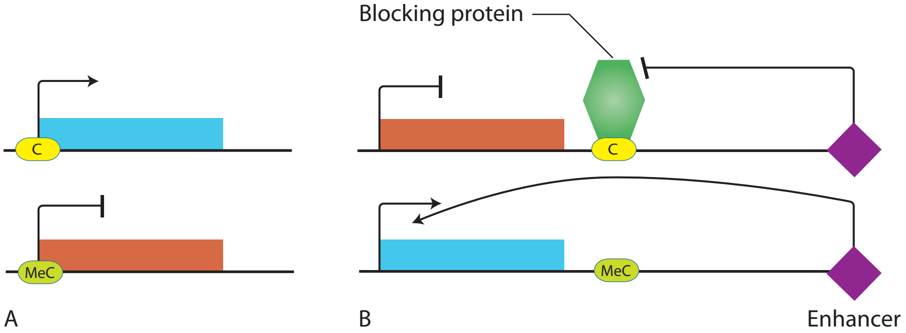

Figure 1.16 Consequences of methylation on gene expression. (A) Methylation of a CpG island at the promoter of a gene results in transcriptional repression. (B) The sequence at the right contains an enhancer that activates the gene at the left. This activation is blocked by a protein that binds to a DNA sequence between the two loci. The blocking protein does not bind to methylated DNA, so the gene at the left is expressed only if the sequence between the two loci is methylated.

to the environment. There is evidence that fetal nutrition *in utero* may influence the later risk of type 2 diabetes due to epigenetic marks laid down during fetal development, or that the response of an individual to stress may be mediated by epigenetic marks that occur during infancy in response to nurturing. The contribution of epigenetics to health and disease is an emerging important area of research on susceptibility to common disease.

Conclusion

More than half a century of research in molecular biology has resulted in a detailed picture of the mechanisms of gene structure and function. Much of the remainder of this book will be devoted to exploration of the implications of dysfunction at the level of the gene or groups of genes and their interactions with the environment. We will see also that genetics research is moving to a new level of integration of basic molecular mechanisms, toward formation of a picture of how entire cells and organisms function. It is important to realize, however, that some fundamental molecular mechanisms, such as the role of small RNAs and genomic imprinting, have been discovered only within the past decade or so. Even as the effort toward larger scale integration goes forward, there remains much to be learned about the fundamental molecular mechanisms at the level of the gene.

REFERENCE

ENCODE Project Consortium 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* vol. 489, pp. 57–74.

FURTHER READING

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P 2007, *Molecular biology of the cell*, 5th edn, Garland Science, New York.

Krebs JE, Goldstein, ES, Kilpatrick ST 2009, *Lewin's genes X*, Jones & Bartlett, Sudbury, MA.

Lodish H, Berk A, Kaiser CA, Krieger M, Scott MP, Bretscher A *et al.* 2007, *Molecular cell biology*, 6th edn, Freeman, New York.

Find interactive self-test questions and additional resources for this chapter at **www.korfgenetics.com**.

Self-Assessment

Review Questions

1.1 The two strands of DNA separate when heated, and the temperature at which separation occurs is dependent on the base content. Specifically, DNA with a higher proportion of G–C base pairs tends to "melt" at a higher temperature than molecules with a higher A–T content. Why is this?

1.2 What is the role of transcription in DNA replication?

1.3 Consider this gene sequence. What is the base sequence of the mRNA that would be transcribed from this gene, and what is the amino acid sequence of the peptide that would be translated?

5′–promoter–ATG GTT GAT AGT CGT TGC CGC GGG CTG TGA–3′
3′–promoter–TAC CAA CTA TCA GCA ACG GCG CCC GAC ACT–5′

1.4 There are more proteins than there are genes. What are some of the mechanisms that account for this discrepancy?

1.5 A woman is heterozygous for an X-linked trait that leads to expression of two different forms of an enzyme. The two forms are separable as two distinct bands when the enzyme protein is run through an electric field by electrophoresis. If you were to test cultured skin fibroblasts and isolate enzyme, what would you expect to see? If you could isolate single fibroblasts and grow them into colonies before extracting the enzyme and subjecting it to electrophoresis, what would you expect to see?