PART I MOLECULAR AND CELLULAR BASIS OF HEMATOLOGY

CHAPTER 1

ANATOMY AND PHYSIOLOGY OF THE GENE

Andrew J. Wagner, Nancy Berliner, and Edward J. Benz, Jr.

Normal blood cells have limited life spans; they must be replenished in precise numbers by a continuously renewing population of progenitor cells. Homeostasis of the blood requires that proliferation of these cells be efficient yet strictly constrained. Many distinctive types of mature blood cells must arise from these progenitors by a controlled process of commitment to, and execution of, complex programs of differentiation. Thus developing red blood cells must produce large quantities of hemoglobin but not the myeloperoxidase characteristic of granulocytes, the immunoglobulins characteristic of lymphocytes, or the fibrinogen receptors characteristic of platelets. Similarly, the maintenance of normal amounts of procoagulant and anticoagulant proteins in the circulation requires an exquisitely regulated production, destruction, and interaction of the components. Understanding the basic biologic principles underlying cell growth, differentiation, death, and the homeostasis of critical proteins requires a thorough knowledge of the structure and regulated expression of genes because the gene is now known to be the fundamental unit by which biologic information is stored, transmitted, and expressed in this regulated fashion.

Genes were originally characterized as mathematic units of inheritance. They are now known to consist of molecules of deoxyribonucleic acid (DNA). By virtue of their ability to store information in the form of nucleotide sequences, to transmit it by means of semiconservative replication to daughter cells during mitosis and meiosis, and to express it by directing the incorporation of amino acids into proteins, DNA molecules are the chemical transducers of genetic information flow. Efforts to understand the biochemical means by which this transduction is accomplished have given rise to the disciplines of molecular biology and molecular genetics.

THE GENETIC VIEW OF THE BIOSPHERE: THE CENTRAL DOGMA OF MOLECULAR BIOLOGY

The fundamental premise of the molecular biologist is that the magnificent diversity encountered in nature is ultimately governed by genes. The capacity of genes to exert this control is in turn determined by relatively simple stereochemical rules, first appreciated by Watson and Crick in the 1950s. These rules govern the types of interactions that can occur between two molecules of DNA or ribonucleic acid (RNA).

DNA and RNA are linear unbranched polymers consisting of four types of nucleotide subunits. Each nucleotide is distinguished from the others by a unique purine or pyrimidine "base" projecting from the chain. Proteins are linear unbranched polymers consisting of 21 types of amino acid subunits. Each amino acid is distinguished from the others by the chemical nature of its side chain, the moiety not involved in forming the peptide bond links of the chain. The properties of cells, tissues, and organisms depend largely on the aggregate structures, properties and biochemical activities of their proteins, and the interactions occurring among them. The central dogma of molecular biology states that genes control these properties by encoding the structures of proteins, controlling the timing and amount of their production, and coordinating their synthesis with that of other proteins. The information needed to achieve these ends is transmitted (expressed) from DNA and translated into proteins by a class of nucleic acid molecules called RNA. Genetic information thus flows in the direction DNA → RNA → protein. This central dogma provides, in principle, a universal approach for investigating the biologic properties and behavior of any given cell, tissue, or organism by study of the controlling genes. Methods permitting direct manipulation of DNA and RNA sequences should then be universally applicable to the study of all living entities. Indeed, the power of the methodologies of molecular genetics lie in the universality of their utility.

One exception to the central dogma of molecular biology that is especially relevant to hematologists is the storage of genetic information in RNA molecules in certain viruses, notably the retroviruses associated with T-cell leukemia and lymphoma, and the human immunodeficiency virus. When retroviruses enter the cell, the RNA genome (the term "genome" refers to the totality of DNA or RNA sequences encoding the genetic information of a cell, tissue, or organism) is copied into a DNA replica (cDNA). This is accomplished with RNA-dependent DNA polymerases, enzymes also called *reverse transcriptases*. This DNA representation of the viral genome is then expressed according to the pathway specified by the central dogma. Retroviruses thus represent a variation on the theme rather than a true exception to or violation of the dogma. There are also some RNA viruses (coronaviruses being the most universally known example) that carry an RNA-dependent RNA polymerase capable of replicating many copies of its own RNA genome. These messenger RNAs (mRNAs) then encode proteins essential to their life cycle.

THE ANATOMY AND PHYSIOLOGY OF THE GENE

DNA and RNA Structure

DNA molecules are extremely long, unbranched polymers of nucleotide subunits. Each nucleotide contains a sugar moiety called deoxyribose, a phosphate group attached to the 5′ carbon position, and a purine or pyrimidine base attached to the 1′ position (Fig. 1.1). The linkages in the chain are formed by phosphodiester bonds between the 5′ position of each sugar residue and the 3′ position of the adjacent residue in the chain (see Fig. 1.1). The sugar-phosphate links

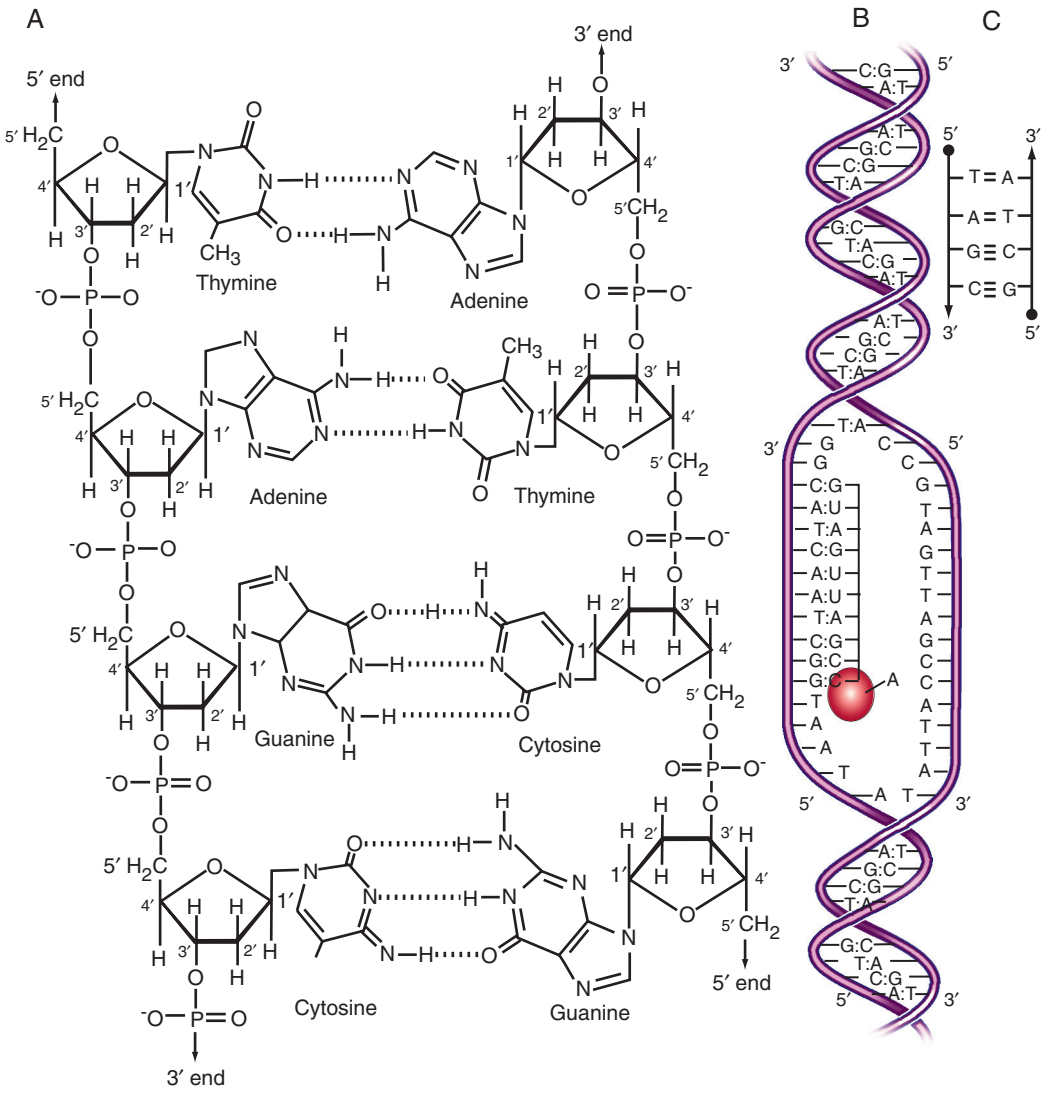

Figure 1.1 STRUCTURE, BASE PAIRING, POLARITY, AND TEMPLATE PROPERTIES OF DNA. (A) Structures of the four nitrogenous bases projecting from sugar phosphate backbones. The hydrogen bonds between them form base pairs holding complementary strands of DNA together. Note that A–T and T–A base pairs have only two hydrogen bonds, whereas C–G and G–C pairs have three. (B) The double helical structure of DNA results from base pairing of strands to form a double-stranded molecule with the backbones on the outside and the hydrogen-bonded bases stacked in the middle. Also shown schematically is the separation (unwinding) of a region of the helix by mRNA polymerase, which is shown using one of the strands as a template for the synthesis of an mRNA precursor molecule. Note that new bases added to the growing RNA strand obey the rules of Watson-Crick base pairing (see text). Uracil (U) in RNA replaces T in DNA and, like T, forms base pairs with A. (C) Diagram of the antiparallel nature of the strands, based on the stereochemical 3′ → 5′ polarity of the strands. The chemical differences between reading along the backbone in the 5′ → 3′ and 3′ → 5′ directions can be appreciated by reference to (A). *A*, Adenosine; *C*, cytosine; *G*, guanosine; *T*, thymine; *U*, uracil.

form the backbone of the polymer, from which the purine or pyrimidine bases project perpendicularly.

The haploid human genome consists of 23 long, double-stranded DNA molecules tightly complexed with histones and other nuclear proteins to form compact linear structures called *chromosomes*. The genome contains approximately 3 billion nucleotides; the individual chromosomes range from 50 to 200 million bases in length. By convention they are numbered from the longest (chromosome 1) to the shortest (chromosome 22), with the sex chromosomes getting the special designation X and Y. Females inherit the XX genotype and males, XY. The individual genes are aligned along each chromosome. The human genome contains about 2000 to 30,000 genes. Blood cells, like most somatic cells, are diploid. That is, each chromosome is present in two copies, so there are 46 chromosomes consisting of approximately 6 billion base pairs (bp) of DNA.

The four nucleotide bases in DNA are two purines (adenosine and guanosine) and two pyrimidines (thymine and cytosine). The basic chemical configuration of the other nucleic acid found in cells, RNA, is quite similar, except that the sugar is ribose (having a hydroxyl group attached to the 2′ carbon rather than the hydrogen found in deoxyribose) and the pyrimidine base uracil is used in place of thymine. The bases are commonly referred to by a shorthand notation: the letters A, C, G, T, and U are used to refer to adenosine, cytosine, guanosine, thymine, and uracil, respectively.

The ends of DNA and RNA strands are chemically distinct because of the 3′ → 5′ phosphodiester bond linkage that ties adjacent bases together (see Fig. 1.1). One end of the strand (the 3′ end) has an unlinked (free at the 3′ carbon) sugar position, and the other (the 5′ end) has a free 5′ position. There is thus a directionality (polarity) to the sequence of bases in a DNA strand: the same sequence of bases read in a 3′ → 5′ direction carries a different meaning than if read in a 5′ → 3′ direction. Cellular enzymes can thus distinguish one end of a nucleic acid from the other and one strand from its paired mate; most enzymes that "read" the DNA sequence tend to do so only in one direction (3′ → 5′ or 5′ → 3′ but not both). For instance, most nucleic acid–synthesizing enzymes read the template strand in 3′ → 5′ direction, thus adding new bases to the strand in a 5′ → 3′ direction.

Storage of Genetic Information in the Nucleotide Sequences of DNA

The ability of DNA molecules to store information resides in the sequence of nucleotide bases arrayed along the polymer chain. Under the physiologic conditions in living cells, DNA is thermodynamically most stable when two strands coil around each other to form a double-stranded helix. The strands are aligned in an "antiparallel" direction, having opposite 3′ → 5′ polarities (see Fig. 1.1). The DNA strands are held together by hydrogen bonds between the bases on one strand and the bases on the opposite (complementary) strand. The stereochemistry of these interactions allows bonds to form between the two strands only when adenine on one strand pairs with thymine at the same position of the opposite strand, or guanine with cytosine. These are the "Watson-Crick" rules of base pairing. Two strands joined together in compliance with these rules are said to have "complementary" base sequences. Similar rules apply to the formation of DNA-RNA or RNA-RNA double-stranded hybrids, except that A-U base pairs replace A-T pairs.

These thermodynamic rules imply that the sequence of bases along one DNA strand immediately dictates the sequence of bases that must be present along the complementary strand in the double helix. For example, whenever an A occurs along one strand, a T must be present at that exact position on the opposite strand; a G must always be paired with a C, a T with an A, and a C with a G.

Single-stranded nucleic acids can also fold back on themselves if two complementary sequences exist at different points along the molecule, thus forming "hairpin loops." Hairpin loop structures create secondary structures that affect the accessibility of sequences and the interaction of the molecule with proteins or other nucleic acids.

Transmission of Genetic Information to the Next Generation

Enzymes that replicate (polymerize) DNA and RNA molecules obey the base-pairing rules. By using an existing strand of DNA or RNA as the template, a new (daughter) strand is copied (transcribed) by reading processively along the base sequence of the template strand, adding to the growing strand at each position only that base that is complementary to the corresponding base in the template according to the Watson-Crick rules. Thus a DNA strand having the base sequence 5′-GGCTATG-3′ could be copied by DNA polymerase only into a daughter strand having the sequence 3′-CCGATAC-5′. Note that the sequence of the template strand provides all the information needed to predict the nucleotide sequence of the complementary daughter strand. Genetic information is thus stored in the form of base-paired nucleotide sequences.

If a double-stranded DNA molecule is separated into its two component strands and each strand is then used as a template to synthesize a new daughter strand, the product will be two double-stranded daughter DNA molecules, each identical to the original parent molecule. This semiconservative replication process is exactly what occurs during mitosis and meiosis as cell division proceeds (Fig. 1.2). The rules of Watson-Crick base pairing thus provide for the faithful transmission of exact copies of the cellular genome to subsequent generations.

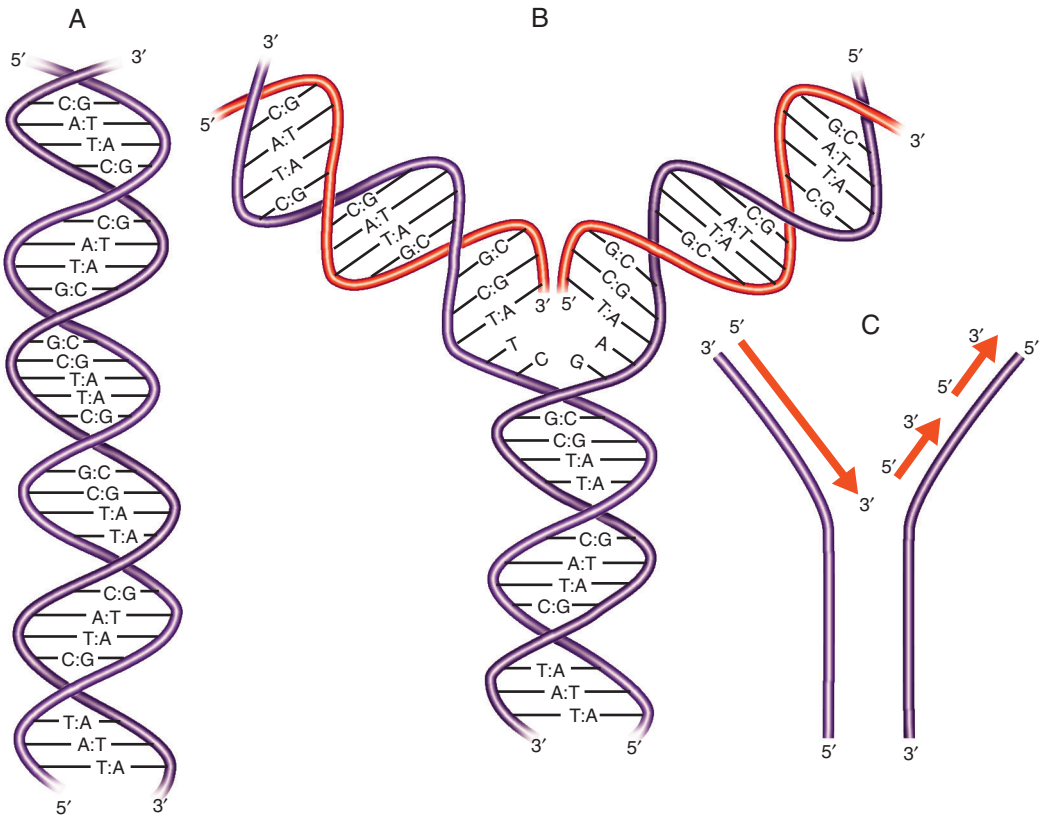

Figure 1.2 SEMICONSERVATIVE REPLICATION OF DNA. (A) The process by which the DNA molecule on the left is replicated into two daughter molecules, as occurs during cell division. Replication occurs by separation of the parent molecule into the single-stranded form at one end, reading of each of the daughter strands in the 3′ → 5′ direction by DNA polymerase, and addition of new bases to growing daughter strands in the 5′ → 3′ direction. (B) The replicated portions of the daughter molecules are identical to each other *(red)*. Each carries one of the two strands of the parent molecule, accounting for the term *semiconservative replication*. Note the presence of the replication fork, the point at which the parent DNA is being unwound. (C) The antiparallel nature of the DNA strands demands that replication proceed toward the fork in one direction and away from the fork in the other *(red)*. This means that replication is actually accomplished by reading of short stretches of DNA followed by ligation of the short daughter strand regions to form an intact daughter strand.

The Expression of Genetic Information Via Translation Into Proteins Using the Genetic Code

The information stored in the DNA base sequence of genes achieves its impact on the structure, function, and behavior of organisms by governing the structures, timing, and amounts of proteins and certain RNAs synthesized in the cells. The primary structure (i.e., the amino acid sequence) of each protein determines its three-dimensional conformation and therefore its properties (e.g., shape, enzymatic activity, ability to interact with other molecules, localization, and stability). In the aggregate, these proteins control cell structure and metabolism. The process by which DNA achieves its control of cells through protein synthesis is called *gene expression*.

An outline of the basic pathway of gene expression in eukaryotic cells is shown in Fig. 1.3. The DNA base sequence of the "minus," "anticoding" strand is first copied into an RNA molecule with a complementary base sequence, called *premessenger RNA* (pre-mRNA), by mRNA polymerase. Pre-mRNA thus has a base sequence identical to

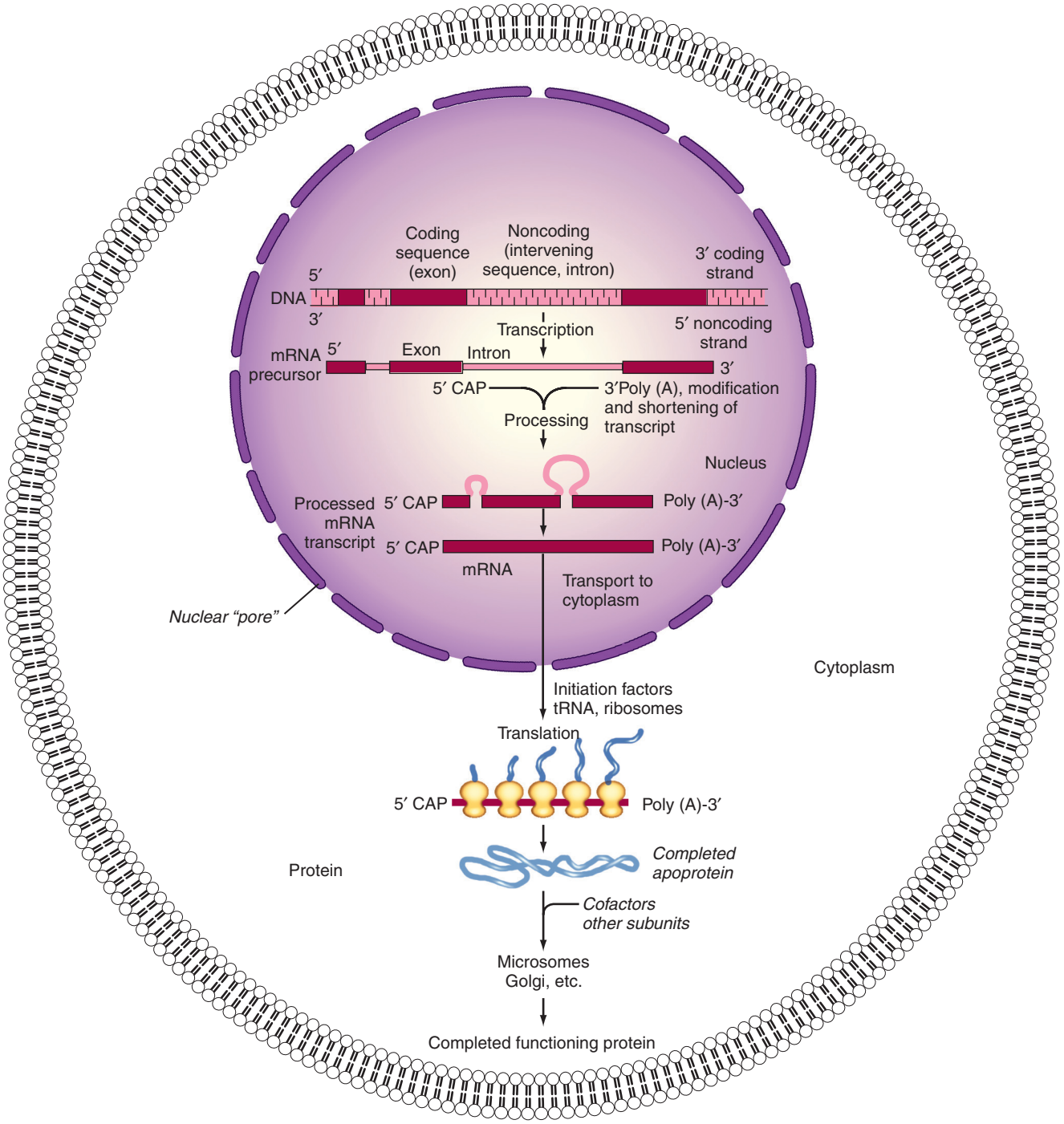

Figure 1.3 SYNTHESIS OF mRNA AND PROTEIN—THE PATHWAY OF GENE EXPRESSION. The diagram of the DNA gene shows the alternating array of exons *(red)* and introns *(shaded color)* typical of most eukaryotic genes. Transcription of the mRNA precursor, addition of the 5′-CAP and 3′-poly (A) tail, splicing and excision of introns, transport to the cytoplasm through the nuclear pores, translation into the amino acid sequence of the apoprotein, and posttranslational processing of the protein are described in the text. Translation proceeds from the initiator methionine codon near the 5′ end of the mRNA, with incorporation of the amino terminal end of the protein. As the mRNA is read in a 5′ → 3′ direction, the nascent polypeptide is assembled in an amino → carboxyl terminal direction.

the DNA "plus" or "coding" strand. Genes in eukaryotic species consist of tandem arrays of sequences encoding mature mRNA (exons) alternating with sequences (introns) present in the initial mRNA transcript (pre-mRNA) but absent from the mature mRNA. The entire gene is transcribed into the larger precursor, which is then further processed (spliced) in the nucleus. The introns are excised from the final mature mRNA molecule, which is then further processed, as discussed later, and exported to the cytoplasm to be decoded (translated) into the amino acid sequence of the protein by association with a biochemically complex group of ribonucleoprotein structures called *ribosomes*. Ribosomes contain two subunits: the 60 S subunit contains a single, large (28 S) ribosomal RNA (rRNA) molecule complexed with multiple proteins, and the 40 S subunit. The RNA component of the 40 S subunit is a smaller (18 S) rRNA.

Ribosomes read an mRNA sequence in a ticker tape fashion three bases at a time, inserting the appropriate amino acid encoded by each three-base code word or codon into the appropriate position of the growing protein chain. This process is called *mRNA translation*. The glossary used by cells to know which amino acids are encoded by each DNA codon is called the *genetic code* (Table 1.1). Each amino acid is encoded by a sequence of three successive bases. Because there are four code letters (A, C, G, and U) and because sequences read in the 5′ → 3′ direction have a different biologic meaning than sequences read in the 3′ → 5′ direction, there are 4^3, or 64, possible codons consisting of three bases.

There are 21 naturally occurring amino acids found in proteins. Thus more codons are available than amino acids to be encoded. As noted in Table 1.1, a consequence of this redundancy is that some amino acids are encoded by more than one codon. For example, six distinct codons can specify incorporation of arginine into a growing amino acid chain, four codons can specify valine, two can specify glutamic acid, and only one each methionine or tryptophan. However, in no case does a single codon encode more than one amino acid. Codons thus predict unambiguously the amino acid sequence they encode. In contrast, one cannot easily read backward from the amino acid sequence to decipher the *exact* encoding DNA sequence. These facts are summarized by saying that the code is degenerate but not ambiguous.

Some specialized codons serve as punctuation points during translation. The methionine codon (AUG), when surrounded by a consensus nucleotide sequence motif (the Kozak box) near the beginning (5′ end) of the mRNA, serves as the initiator codon signaling the first amino acid to be incorporated. All proteins initially begin with a methionine residue, but this is often removed later in the translational process. Three codons, UAG, UAA, and UGA, serve as translation terminators, signaling the end of translation.

The adaptor molecules mediating individual decoding events during mRNA translation are small (40 bases long) RNA molecules called *transfer RNAs* (tRNAs). When bound into a ribosome, each tRNA exposes a three-base segment within its sequence called the *anticodon*. These three bases attempt to pair with the three-base codon exposed on the mRNA. If the anticodon is complementary in sequence to the codon, a stable interaction among the mRNA, the ribosome, and the tRNA molecule results. Each tRNA also contains a separate region that is adapted for covalent binding to an amino acid. The enzymes that catalyze the binding of each amino acid are constrained in such a way that each tRNA species can bind only to a single amino acid. For example, tRNA molecules containing the anticodon 3′-AAA-5′, which is complementary to a 5′-UUU-3′ (phenylalanine) codon in mRNA, can be bound to or charged with only phenylalanine; tRNA containing the anticodon 3′-UAG-5′ can be charged with only isoleucine, and so forth.

tRNAs and their amino acyl tRNAs transduce nucleic acid information into the amino acid sequence that determines it physiologic properties. Ribosomes provide the structural matrix on which tRNA anticodons and mRNA codons become properly exposed and aligned in an orderly, linear, and sequential fashion. As each new codon is exposed, the appropriate charged tRNA species is bound. A peptide bond is then formed between the amino acid carried by this tRNA and the C-terminal residue on the existing nascent protein chain. The growing chain is transferred to the new tRNA in the process, so that it is held in place as the next tRNA is brought in. This cycle is repeated until completion of translation. The completed polypeptide is then transferred to other organelles for further processing (e.g., to the endoplasmic reticulum and the Golgi apparatus) or released into the cytosol for association with other subunits to form complex multimeric proteins (e.g., hemoglobin) and so forth, as discussed in Chapters 4 and 7.

TABLE 1.11 The Genetic Code[a] Messenger RNA Codons for the Amino Acids

Alanine	Arginine	Asparagine	Aspartic Acid	Cysteine
5′-GCU-3′	CGU	AAU	GAU	UGU
GCC	CGC	AAC	GAC	UGC
GCA	CGA			
GCG	AGA			
	AGG			

Glutamic Acid	Glutamine	Glycine	Histidine	Isoleucine
GAA	CAA	GGU	CAU	AUU
GAG	CAG	GGC	CAC	AUC
		GGA		AUA
		GGG		

Leucine	Lysine	Methionine	Phenylalanine	Proline[b]
UUA	AAA	AUG[c]	UUU	CCU
UUG	AAG		UUC	CCC
CUU				CCA
CUC				CCG
CUA				
CUG				

Serine	Threonine	Tryptophan	Tyrosine	Valine
UCU	ACU	UGG	UAU	GUU
UCC	ACC		UAC	GUC
UCA	ACA			GUA
UCG	ACG			GUG
AGU				
AGC				
Chain Termination[d]				
UAA				
UAG				
UGA				

[a]Note that most of the degeneracy in the code is in the third base position (e.g., lysine, AA [G or C]; asparagine, AA [C or U]; valine, GUN [where N is any base]).
[b]Hydroxyproline, the 21st amino acid, is generated by posttranslational modification of proline. It is almost exclusively confined to collagen subunits.
[c]AUG is also used as the chain-initiation codon when surrounded by the Kozak consensus sequence.
[d]The codons that signal the end of translation, also called nonsense or termination codons, are described by their nicknames *amber* (UAG), *ochre* (UAA), and *opal* (UGA).
A, Adenosine; *C*, cytosine; *G*, guanosine; *T*, thymine; *U*, uracil.

REGULATION OF GENE EXPRESSION

Virtually all cells of an organism receive a complete copy of the DNA genome inherited at the time of conception. The diversity of distinct cell types and tissues found in any complex organism is possible only

because different portions of the genome are selectively expressed or repressed in each cell type. Each cell must "know" which genes to express, how actively to express them, and when to express them. This biologic necessity has come to be known as *gene regulation* or *regulated gene expression*. Understanding gene regulation provides insight into how pluripotent stem cells determine that they will express the proper sets of genes in daughter progenitor cells that differentiate along each lineage. Major hematologic disorders (e.g., the leukemias and lymphomas), immunodeficiency states, and myeloproliferative syndromes result from derangements in the system of gene regulation. An understanding of the ways that genes are selected for expression thus remains one of the major frontiers of biology and medicine. Chapters 2, 4, and 6 offer a more thorough coverage of these topics. The following sections provide brief introductions.

Chromatin and the Epigenetic Regulation of Gene Expression

Only a small fraction of the 6 billion base pairs of DNA present in a diploid human cell codes for proteins or for the ribosomal, transfer, and spliceosome RNAs, even including the nearby DNA sequences (promoters, repressors, enhancers, silencers, and insulator sequences) that are needed to support regulated protein synthesis. As discussed later and in Chapter 4, many additional species of RNA molecules exhibiting important regulatory effects on gene expression have been and still are being discovered. Yet, less than 10% of the genome accounts for all DNA sequences having a known function in gene expression. The remainder is called "DNA dark matter." It is being intensively investigated, but its purpose and impact on homeostasis remain unknown. A major challenge for cells, then, is how to find the genes and how to identify and activate only those genes whose expression it needs for its vital functions. The field of study that has arisen to address these questions is called epigenetics. This section provides only a brief introduction to epigenetics; Chapter 2 offers a thorough review and documents the increasing importance of epigenetics to hematology.

Most of the DNA in living cells is inactivated by formation of a nucleoprotein complex called *chromatin*. The histone and nonhistone proteins in chromatin effectively sequester genes from enzymes needed for expression. The most tightly compacted chromatin regions are called *heterochromatin*. *Euchromatin*, less tightly packed, contains actively transcribed genes. Activation of a gene for expression (i.e., transcription) requires that it become less compacted and more accessible to the transcription apparatus. These processes involve both *cis*-acting and *trans*-acting factors. *Cis*-acting elements are regulatory DNA sequences within or flanking the genes. They are recognized by *trans*-acting factors, which are nuclear DNA–binding proteins needed for transcriptional regulation.

DNA sequence regions flanking genes are called *cis*-acting because they influence expression of nearby genes only on the same chromosome. These sequences do not usually encode mRNA or protein molecules. They alter the conformation of the gene within chromatin twisting or kinking the surrounding DNA in ways that facilitate or inhibit access to the factors that modulate transcription. When exogenous nucleases (DNAses) are added experimentally in small amounts to nuclei, these exposed regions are especially sensitive to their DNA-cutting action. Thus DNAse hypersensitive sites in chromatin have come to be useful as markers for regions in or near genes that are accessible for transcription (Chapter 2).

DNA methylation is an epigenetic structural feature that also marks differences between actively transcribed and inactive genes. Most eukaryotic DNA is heavily methylated; that is, the DNA is modified by the addition of a methyl group to the 5 position of the cytosine pyrimidine ring (5-methyl-C). In general, heavily methylated genes are inactive; active genes are relatively hypomethylated, especially in the 5′ and 3′ flanking regions containing the promoter and other regulatory elements (see "Enhancers, Promoters, and Silencers"). These flanking regions frequently include DNA sequences with a high content of Cs and Gs (CpG islands). Hypomethylated CpG islands serve as markers of actively transcribed genes. For example, a search for undermethylated CpG islands on chromosome 7 facilitated the search for the gene for cystic fibrosis.

DNA methylation is facilitated by DNA methyltransferases (DMTs). DNA replication incorporates unmethylated nucleotides into each nascent strand, thus leading to demethylated DNA. For cytosines to become methylated, the methyltransferases must act after each round of replication. After an initial wave of demethylation early in embryonic development, regulatory elements are methylated during various stages of development and differentiation (Chapter 2). Aberrant DNA methylation also occurs as an early step during tumorigenesis, leading to silencing of tumor suppressor genes and of genes related to differentiation. This finding has led to induction of DNA demethylation as a target in cancer therapy. Indeed, 5-azacytidine, a cytidine analog that inhibits DMT, and the related compound decitabine, are approved by the US Food and Drug Administration (FDA) for use in myelodysplastic syndromes, and their use in cases of other malignancies is being investigated.

The mechanisms by which particular regions of DNA are targeted for methylation are under intense investigation. It is becoming increasingly apparent that this modification begets further alterations in chromatin proteins that in turn influence gene expression.

The "opening" of chromatin is necessary but not sufficient for genes to be expressed. The sequences within the now-accessible regions of DNA that are intended for transcription, and no others, must be identified and configured for binding by the intranuclear factors and mRNA polymerase that will execute the transcription program. This is accomplished by the presence of sequences embedded near or within the gene that are recognized by specific proteins that activate or inactivate transcription depending on which stimulatory or inhibitory proteins the sequences attract. These are discussed in the next section.

The major protein components of chromatin are histones, which are a small, highly basic protein family that binds tightly to the acidic residues in DNA. Histones can be acetylated, reducing their affinity for DNA, or methylated, which stabilizes their binding. Histone acetylation, phosphorylation, and methylation of the N-terminal tail are the focus of intense study for their potential roles in opening or closing access to regions of DNA for expression. For example, acetylation of histone lysine residues (catalyzed by histone acetyltransferases) is associated with transcriptional activation. Conversely, histone deacetylation (catalyzed by histone deacetylase) leads to gene silencing. Histone deacetylases are recruited to areas of DNA methylation by DMT and by methyl–DNA-binding proteins, thus linking DNA methylation to histone deacetylation. Drugs inhibiting these enzymes have been demonstrated to be active anticancer agents and continue to be the focus of ongoing studies. The regulation of histone acetylation and deacetylation appears to be linked to gene expression, but the roles of histone phosphorylation and methylation are less well understood. Current research suggests that in addition to gene regulation, histone modifications contribute to the "epigenetic code" and are thus a means by which information regarding chromatin structure is passed to daughter cells after DNA replication occurs.

Regulatory Sequence Motifs in or Near Genes: Enhancers, Promoters, and Silencers

Several types of *cis*-active DNA sequence elements have been defined according to the presumed consequences of their interaction with nuclear proteins (see Fig. 1.5). *Promoters* are found just upstream (to the 5′ side) of the start of mRNA transcription (the CAP). mRNA polymerases appear to bind first to the promoter region and thereby gain access to the structural gene sequences downstream. Promoters thus serve a dual function of being binding sites for mRNA polymerase and marking for the polymerase the downstream point at which transcription should start.

Enhancers are more complicated DNA sequence elements. Enhancers can lie on either side of a gene or even within the gene. Enhancers are bound by enhancer binding proteins, thereby stimulating expression of genes nearby. The domain of influence of enhancers

(i.e., the number of genes to either side whose expression is stimulated) varies. Some enhancers influence only the adjacent gene; others seem to mark the boundaries of large multigene clusters (gene domains) whose coordinated expression is appropriate to a particular tissue type or a particular time. For example, the very high levels of globin gene expression in erythroid cells depend on the function of an enhancer that seems to activate the entire gene cluster and is thus called a *locus-activating region* (see Fig. 1.5). The nuclear factors interacting with enhancers are probably induced into synthesis or activation as part of the process of differentiation. Chromosomal rearrangements that place a gene that is usually tightly regulated under the control of a highly active enhancer can lead to overexpression of that gene. This commonly occurs in Burkitt lymphoma, for example, in which the MYC proto-oncogene is juxtaposed and dysregulated by an immunoglobulin enhancer.

Silencer sequences serve a function that is the obverse of enhancers. When bound by the appropriate nuclear proteins, silencer sequences cause repression of gene expression. Some evidence indicates that the same sequence elements can act as enhancers or silencers under different conditions, presumably by being bound by different sets of proteins having opposite effects on transcription. *Insulators* are sequence domains that mark the "boundaries" of multigene clusters, thereby preventing activation of one set of genes from "leaking" into nearby genes. The concerted actions of enhancers, silencers, and insulators delineate the specific DNA sequences to be transcribed or prevented from transcription within an opened region of chromatin.

One way that activation of transcription of a genomic DNA segment is accomplished is by a "looping" out phenomenon whereby some DNA binding proteins first bind to each end of a potentially expressed segment of open chromatin; those proteins then bind to one other, pulling the ends together and forming a looped-out segment of chromatin. Additional factors then bind to enhancers, silences, promotors, and enhancers, thereby demarcating those parts meant for transcription or silencing. Loops, in other words, may be a secondary structure that identifies areas primed for transcription (see Fig. 2.1).

Transcription Factors

Transcription factors are nuclear proteins that exhibit gene-specific DNA binding. Considerable information is now available about these nuclear proteins and their biochemical properties, but their physiologic behavior remains incompletely understood. Common structural features have become apparent. Most transcription factors have DNA-binding domains sharing homologous structural motifs (cytosine-rich regions called zinc fingers, leucine-rich regions called leucine zippers, and so on), but other regions appear to be unique. Some factors recognize specific DNA sequence motifs within promoters, enhancers, silencers, or insulators and bind directly to them, whereas others bind to these factors, forming complexes that promote or inhibit transcription. Many factors implicated in the regulation of growth, differentiation, and development (e.g., homeobox genes, proto-oncogenes, antioncogenes) appear to be DNA-binding proteins and may be involved in the steps needed for activation of a gene within chromatin. These factors are discussed in more detail in several other chapters (see Chapters 2, 4, and 6); when mutated, many are involved in the pathogenesis of blood dyscrasias, such as c-myc and c-myb.

Regulation at the Level of Pre-mRNA and mRNA Metabolism

In eukaryotic cells, mRNA is initially synthesized in the nucleus (see Figs. 1.3 and 1.4). Before the initial transcript becomes suitable for translation in the cytoplasm, mRNA processing and transport occur by a complex series of events including excision of the portions of the mRNA corresponding to the introns of the gene (mRNA splicing), modification of the 5′ and 3′ ends of the mRNA to render them more stable and translatable, and transport to the cytoplasm. Moreover, the amount of any particular mRNA moiety in both prokaryotic and eukaryotic cells is governed not only by the composite rate of mRNA synthesis (transcription, processing, and transport) but also by its degradation by cytoplasmic ribonucleases (RNA degradation). Many mRNA species of special importance in hematology (e.g., mRNAs for growth factors and their receptors, proto-oncogene mRNAs, acute phase reactants) are exquisitely regulated by control of their stability (half-life) in the cytoplasm.

Posttranscriptional mRNA metabolism is complex. Only a few relevant aspects are considered in this section. Chapter 4 provides more detail.

Pre-mRNA Splicing

The initial transcript of eukaryotic genes contains several subregions (see Fig. 1.4). Most striking is the tandem alignment of exons and introns. Precise excision of intron sequences and ligation of exons is critical for production of mature mRNA. This process is called mRNA splicing, and it occurs on complexes of small nuclear RNAs and proteins called snRNPs; the term *spliceosome* is also used to

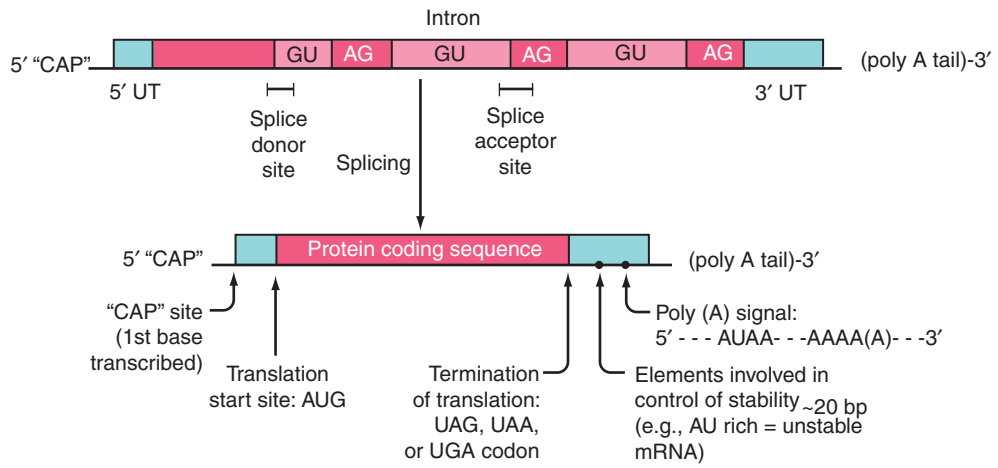

Figure 1.4 ANATOMY OF THE PRODUCTS OF THE STRUCTURAL GENE (mRNA PRECURSOR AND mRNA). This schematic shows the configuration of the critical anatomic elements of an mRNA precursor, which represents the primary copy of the structural portion of the gene. The sequences GU and AG indicate, respectively, the invariant dinucleotides present in the donor and acceptor sites at which introns are spliced out of the precursor. Not shown are the less stringently conserved consensus sequences that must precede and succeed each of these sites for a short distance.

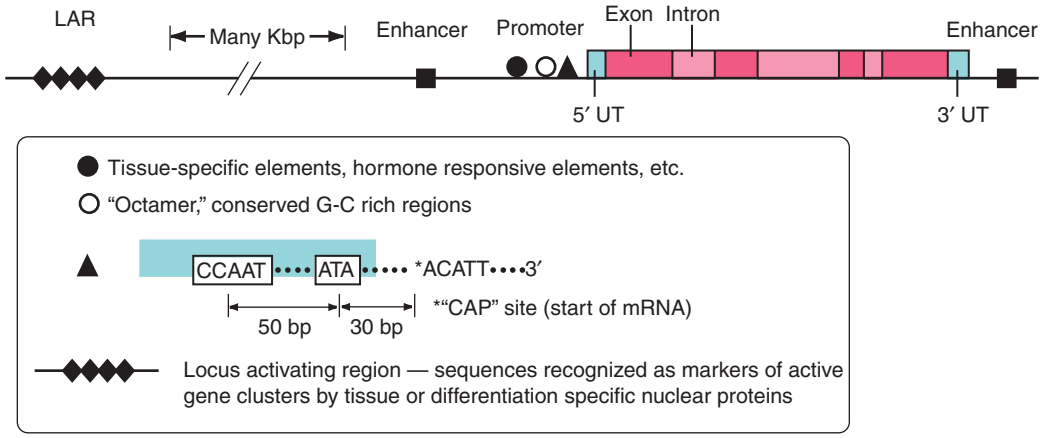

Figure 1.5 REGULATORY ELEMENTS FLANKING THE STRUCTURAL GENE. (*For more information refer to suggested readings from Jones B; Kumar A, et al; Waddington S, et al.)

describe the intranuclear organelle that mediates mRNA splicing reactions. The biochemical mechanism for splicing is complex. A consensus sequence, which includes the dinucleotide GU, is recognized as the donor site at the 5′ end of the intron (5′ end refers to the polarity of the mRNA strand coding for protein); a second consensus sequence ending in the dinucleotide AG is recognized as the acceptor site, which marks the distal end of the intron (see Figs. 1.4 and 1.5). The spliceosome recognizes the donor and acceptor and forms an intermediate lariat structure that provides for both excision of the intron and proper alignment of the cut ends of the two exons for ligation in precise register.

mRNA splicing has proven to be an important mechanism for greatly increasing the versatility and diversity of expression of a single gene. Several different mRNA and protein products can arise from a single gene by selective inclusion or exclusion of individual exons from the mature mRNA products. This phenomenon is called *alternative mRNA splicing*. It permits a single gene to code for multiple mRNA and protein products with related but distinct structures and functions. The mechanisms by which individual exons are selected or rejected are complex and highly context-specific, varying among different cell types, differentiation stages, and physiologic states. Chapter 4 provides additional details. For present purposes, it is sufficient to note that important physiologic changes in cells can be regulated by altering the patterns of mRNA splicing products arising from single genes.

Many inherited hematologic diseases arise from mutations that derange mRNA splicing. For example, some of the most common forms of the thalassemia syndromes and hemophilias (see Chapters 41 and 134) arise by mutations that alter normal splicing signals or create splicing signals where they normally do not exist (activation of cryptic splice sites). Conversely, mutations altering key protein factors that modulate alternative splicing pathways are known to contribute to the pathogenesis of bone marrow dyscrasias (see Chapters 59, 61, and 66).

Modification of the Ends of the mRNA Molecule

Most eukaryotic mRNA species are polyadenylated at their 3′ ends. Polyadenylation results in the addition of stretches of 100 to 150 "A" residues at the 3′ end. Such an addition is often called the *poly-A tail* and is of variable length. Polyadenylation facilitates rapid early cleavage of the unwanted 3′ sequences from the transcript and is also important for stability or transport of the mRNA out of the nucleus. Signals near the 3′ extremity of the mature mRNA mark positions at which polyadenylation occurs. The consensus signal is AUAAA (see Fig. 1.4). Mutations in the poly-A signal sequence have been shown to cause thalassemia (see Chapter 41).

At the 5′ end of the mRNA, a complex oligonucleotide having unusual phosphodiester bonds is added. This structure contains the nucleotide 7-methyl-guanosine and is called *CAP* (see Fig. 1.4). The 5′-CAP enhances both mRNA stability and the ability of the mRNA to interact with protein translation factors and ribosomes.

5′ and 3′ Untranslated Sequences Within mRNAs That Modulate Stability and Translatability

Most mature mRNAs contain sequence motifs at the 5′and 3′ ends of the molecule extending beyond the initiator and terminator codons that mark the beginning and the end of the sequences actually translated into proteins (see Figs. 1.4 and 1.5). These so-called 5′ and 3′ untranslated regions (5′ UTRs and 3′ UTRs) influence both mRNA stability and the efficiency with which mRNA species can be translated. For example, if the 3′ UTR of a very stable mRNA (e.g., globin mRNA) is swapped with the 3′ UTR of a highly unstable mRNA (e.g., the c-myc gene), the c-myc mRNA becomes more stable. Conversely, attachment of the 3′ UTR of c-myc to a globin molecule renders it unstable. Instability is often associated with repeated sequences rich in A and U in the 3′ UTR (see Fig. 1.4). The UTRs in mRNAs coding for proteins involved in iron metabolism mediate altered mRNA stability or translatability by binding iron-laden proteins and thus govern iron storage and turnover (see Chapter 36).

Transport of mRNA From Nucleus to Cytoplasm: mRNP Particles

An additional potential step for regulation or disruption of mRNA metabolism occurs during the transport from nucleus to cytoplasm. mRNA transport is an active, energy-consuming process (Chapter 4). Moreover, at least some mRNAs appear to enter the cytoplasm in the form of complexes bound to proteins (mRNPs). mRNPs may regulate stability of the mRNAs and their access to translational apparatus. Some evidence indicates that certain mRNPs are present in the cytoplasm but are not translated (masked message) until proper physiologic signals are received.

Regulation of mRNA Processing and Stability

As mentioned earlier, cells can regulate the relative amounts of different protein isoforms arising from a given gene by altering the relative amounts of an mRNA precursor that are spliced along one pathway or another (alternative mRNA splicing). Many striking examples of this type of regulation are known—for example, the ability of B lymphocytes to make both immunoglobulin M (IgM) and IgD at the same developmental stage, changes in the particular

isoforms of cytoskeletal proteins produced during red blood cell differentiation, and a switch from one isoform of the *c-myb* proto-oncogene product to another during red blood cell differentiation. Abnormalities of mRNA splicing due to mutations at the splice sites can lead to defective protein synthesis, as can occur in β-globin pre-mRNA, leading to some forms of β-thalassemia. The effect of controlling the pathway of mRNA processing used in a cell is to include or exclude portions of the mRNA sequence. These portions encode peptide sequences that influence the ultimate physiologic behavior of the protein, or the RNA sequences that alter stability or translatability.

The importance of the control of mRNA stability for gene regulation is being increasingly appreciated. The steady-state level of any given mRNA species ultimately depends on the balance between the rate of its production (transcription and mRNA processing) and its destruction. One means by which stability is regulated is the inherent structure of the mRNA sequence, especially the 3′ and 5′ UTRs. As already noted, these sequences appear to affect mRNA secondary structure, recognition by nucleases, or both. Different mRNAs thus have inherently longer or shorter half-lives, almost regardless of the cell type in which they are expressed. Some mRNAs tend to be highly unstable. In response to appropriate physiologic needs, they can thus be produced quickly and removed from the cell quickly when a need for them no longer exists. In contrast, globin mRNA is inherently quite stable, with a half-life measured in the range of 15 to 50 hours. This is appropriate for the need of reticulocytes to continue to synthesize globin for 24 to 48 hours after the ability to synthesize new mRNA has been lost by the terminally mature erythroblasts.

The stability of mRNA can also be altered in response to changes in the intracellular milieu. This phenomenon usually involves nucleases capable of destroying one or more broad classes of mRNA defined on the basis of their 3′ or 5′ UTR sequences. Thus, for example, histone mRNAs are destabilized after the S-phase of the cell cycle is complete. Presumably this occurs because histone synthesis is no longer needed. Induction of cell activation, mitogenesis, or terminal differentiation events often results in the induction of nucleases that destabilize specific subsets of mRNAs. Selective stabilization of mRNAs probably also occurs; for example, α-globin mRNA is stabilized by the protective binding of a specific stabilizing protein to a nuclease target sequence in its 3′ UTR.

Another critical mechanism that ensures the efficiency and fidelity of gene expression is nonsense-mediated decay (NMD). NMD has evolved to deal with the fact that common classes of mutations (either germ line or somatic, and including point mutations, "frame shifts" due to small deletions or insertions, and mutations causing mis-splicing; see Chapters 3 and 4) result in the creation of a premature translation termination codon in the translation reading frame (also stop codons or nonsense mutations). Nonsense codons can also be created by transcription or processing errors occurring during expression of normal genes. Indeed, as many as 5% to 30% of mature mRNA transcripts may carry nonsense codons in some cells under certain conditions. These mRNAs can be translated only into fragments of the intended protein and are thus physiologically useless. This impairs the efficiency of gene expression, expending the considerable energy required for even partial translation while serving no functional purpose. Moreover, those fragments fold abnormally and can trigger stress responses such as the unfolded protein response (Chapter 4) that can trigger other undesired cellular reactions. These fragments can also contain some of the functional domains of the intended complete protein. These can interact deleteriously with other cellular components, deranging cellular homeostasis.

NMD addresses these issues by recognizing nonsense codons and destroying the affected mRNA, thus avoiding its translation. The process exists across evolution from yeast to mammals. It is mediated by complex protein and RNA components functioning and supporting at least two recognition and destruction pathways. It is becoming clear that the integrity of these pathways is likely relevant to multiple disease states, including neoplasia.

Regulation at the Level of mRNA Translation

The amount of a given protein accumulating in a cell depends not only on the amount of the mRNA present but also on the rate at which it is translated into the protein and the stability of the protein. Translational efficiency depends in part on the structural features of any given mRNA, including polyadenylation, secondary structure of the 5′ and 3′ UTRs, and presence of the 5′ cap. The amounts and state of activation of protein factors needed for translation are also crucial. The secondary structure of the mRNA, particularly in the 5′ UTR, greatly influences the intrinsic translatability of an mRNA molecule by constraining the access of translation factors and ribosomes to the translation initiation signal in the mRNA. Secondary structures along the coding sequence of the mRNA may also have some impact on the rate of elongation of the peptide.

Changes in capping, polyadenylation, and translation factor efficiency affect the overall rate of protein synthesis within each cell. These effects tend to be global rather than specific to a particular gene product. However, these effects influence the relative amounts of different proteins made. mRNAs whose structures inherently lend themselves to more efficient translation tend to compete better for rate-limiting components of the translational apparatus, but mRNAs that are inherently less translatable tend to be translated less efficiently in the face of limited access to other translational components. For example, the translation factor eIF-4 tends to be produced in higher amounts when cells encounter transforming or mitogenic events. This causes an increase in overall rates of protein synthesis but also leads to a selective increase in the synthesis of some proteins that were underproduced before mitogenesis because they competed less well when the supply of active eIF-4 was limiting. It is also now being increasingly recognized that several classes of low-molecular-weight RNAs (micro-RNAs [miRNAs]) can have profound effects on the output of proteins from individual mRNAs or related groups of mRNAs by recognizing specific sequences in them and thereby altering stability or translatability.

Translational regulation of individual mRNA species is critical for some events important to blood cell homeostasis. For example, as discussed in Chapter 36, the amount of iron entering a cell is an exquisite regulator of the rate of ferritin mRNA translation. An mRNA sequence called the *iron response element* is recognized by a specific mRNA-binding protein but only when the protein lacks iron. mRNA bound to the protein is translationally inactive. As iron accumulates in the cell, the protein becomes iron bound and loses its affinity for the mRNA, resulting in translation into apoferritin molecules that bind the iron.

Tubulin synthesis involves coordinated regulation of translation and mRNA stability. Tubulin regulates the stability of its own mRNA by a feedback loop. As tubulin concentrations rise in the cell, it interacts with its own mRNA through the intermediary of an mRNA-binding protein. This results in the formation of an mRNA-protein complex and nucleolytic cleavage of the mRNA. The mRNA is destroyed, and further tubulin production is halted.

Heterogeneity of rRNAs and tRNAs

The 18 S and 28 S rRNAs, the many ribosomal proteins needed to assemble a ribosome, and tRNAs are encoded by many genes and are actually quite heterogeneous. The heterogeneity also varies among cell types and under varied cellular states such as the nutritional stress found in cancer cells. These variations appear to create significant alterations in the translatability of specific mRNAs. These effects can be blunted or accentuated by the tendency of different ribosome classes to favor or disfavor certain patterns of codon use. Disease states have been associated with mutations in these proteins and RNAs (ribosomeopathies), and manipulation of this complexity for therapeutic purposes is under intense investigation.

These few examples of posttranscriptional regulation emphasize that cells tend to use every step in the complex pathway of gene expression as points at which exquisite control over the amounts of a particular protein or RNA species can be regulated. In other chapters, additional levels of regulation are described (e.g., regulation of the production,

stability, activity, localization, and access to other cellular components of the proteins that are present in a cell [see Chapters 6 and 7]).

Roles of Small Interfering RNAs, Micro RNAs, Short Hairpin RNAs, and Long Noncoding RNAs in Regulating Gene Expression

Cells were once thought to possess only three basic classes of RNA molecules: mRNA, rRNAs (5 S, 18 S, and 28 S), and tRNA. Moreover, the physiologic capacity of these RNA species was thought to be only informational, their nucleic acid sequences serving as codons, anticodons, or binding sites for ribosomal proteins, splicing and translation factors, mRNA transport factors, etc. Two fundamental discoveries have profoundly changed our view of the biologic role of RNAs. First was the recognition that some RNA molecules have catalytic activity that sustain key steps in gene expression such as pre-mRNA splicing. In cells, these activities are often carried out within ribonucleic acid (RNP) complexes. The second was the discovery that cells contain a potpourri of small RNA species in both the nucleus and the cytoplasm. Collectively these RNA moieties provide another layer of complex posttranscriptional mechanisms modulating gene expression. Some of these small RNAs might modulate transcription and processing as well.

One such process is carried out *by small interfering RNAs (siRNAs)*: short, double-stranded fragments of RNA containing 21 to 23 bp (Fig. 1.6). The process is triggered by perfectly complementary double-stranded RNA, which is cleaved by Dicer, a member of the RNase III family, into siRNA fragments. These small fragments of double-stranded RNA are unwound by a helicase in the RNA-induced silencing complex (RISC). The antisense strand anneals to mRNA transcripts in a sequence-specific manner and in doing so brings the endonuclease activity within the RISC to the targeted transcript. An RNA-dependent RNA polymerase in the RISC may then create new siRNAs to processively degrade the mRNA, ultimately leading to complete degradation of the mRNA transcript and abrogation of protein expression.

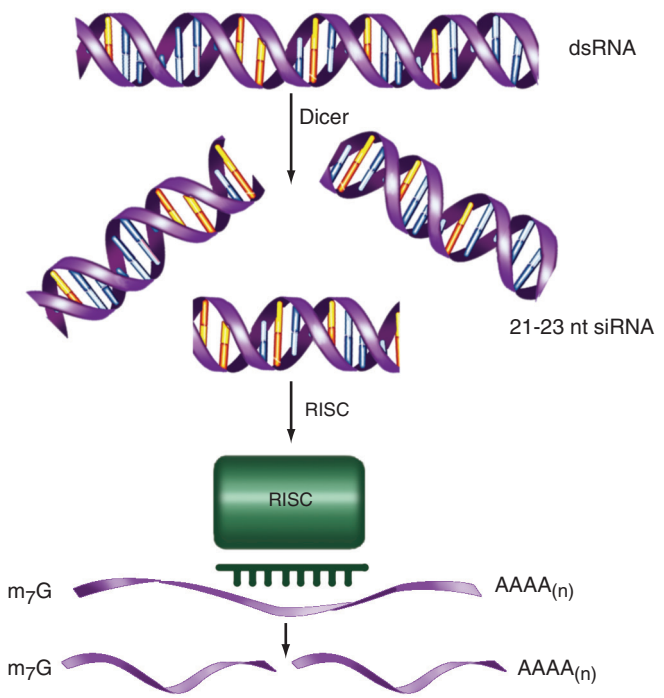

Figure 1.6 mRNA DEGRADATION BY siRNA. dsRNA is digested into 21- to 23-bp (base pair) small interfering RNAs by the Dicer RNase. These RNA fragments are unwound by RISC and bring the endonucleolytic activity of RISC to mRNA transcripts in a sequence-specific manner, leading to degradation of the mRNA. *dsRNA*, double-stranded RNA; *RISC*, RNA-induced silencing complex; *siRNA*, small interfering RNA.

Although this endogenous process likely evolved to destroy invading viral RNA, the use of siRNA has become a commonly used tool for evaluation of gene function. Sequence-specific synthetic siRNA may be directly introduced into cells or introduced via gene transfection methods and targeted to an mRNA of a gene of interest. The siRNA will lead to degradation of the mRNA transcript and accordingly prevent new protein translation. This technique is a relatively simple, efficient, and inexpensive means to investigate cellular phenotypes after directed elimination of expression of a single gene. Experimentally, engineered short hairpin RNAs (shRNAs) are used extensively to degrade or block the translation of a gene's mRNA product in a highly specific fashion, thus allowing one to target or "knock down" the expression of any gene or collection of genes at will and allowing assessment of a cell's behavior in the absence of expression of the targeted genes.

miRNAs, or MIRs, are 22-nucleotide small RNAs encoded by the cellular genome that alter mRNA stability and protein translation. These genes are transcribed by RNA polymerase II and capped and polyadenylated similar to other RNA polymerase II transcripts. The precursor transcript of approximately 70 nucleotides is cleaved into mature miRNAs by the enzymes Drosha and Dicer. One strand of the resulting duplex forms a complex with the RISC that together binds the target mRNA with imperfect complementarity. Through mechanisms that are still incompletely understood, miRNA suppresses gene expression, likely either through inhibition of protein translation or through destabilization of mRNA. miRNAs appear to have essential roles in development and differentiation and are aberrantly regulated in many types of cancer cells. The identification of miRNA sequences, their regulation, and their target genes are areas of intense study.

Other classes of small RNA molecules, such as circular or ringed RNAs and glycosylated RNAs, are under active study. Discussion of these is beyond the scope of this chapter. Moreover, a class of extraordinarily long RNA transcripts (long noncoding RNA [lncRNA]) has been known to exist for decades, but its functions are just beginning to be uncovered. lncRNA may be support an important mechanism for "opening" large domains of chromatin to access by mRNA polymerase (RNA polymerase II), transcription factors, enhancer- and silencer-binding proteins, etc., so the genes within that domain can be expressed. This might also provide clues into the role played by DNA "dark matter" in gene regulation, if the signals for the production, start points, and end points of lncRNAs are encoded in the regions "opened" by lncRNA transcription.

Some Illustrative Structural Features of the Genome Relevant to Hematology

Structural genes are separated from one another by as few as 1 to 5 kilobases or as many as several thousand kilobases of DNA. Almost nothing is known about the reason for the erratic clustering and spacing of genes along chromosomes. It is clear that intergenic DNA contains a variegated landscape of structural features that provide useful tools to localize genes, identify individual human beings as unique from every other human being (DNA fingerprinting), and diagnose human diseases by linkage. Only a brief introduction is provided here.

Polymorphism and Single Nucleotide Polymorphisms

The genomic landscape of each of our genomes is dotted with scattered sequence differences that distinguish us from any other living creature. These are a consequence of the nonzero error rate of base copying during normal DNA replication; under normal circumstances it is

approximately $1/10^6$. In other words, one of 1 million bases of DNA will be miscopied (mutated) during each round of DNA replication. A set of enzymes called *DNA proofreading enzymes* corrects most of these mutations so that the rate of mutation following a normal cell division is closer to $1/10^9$. When these enzymes are themselves altered by mutation, the rates of mutation (and therefore the odds of neoplastic transformation) increase considerably. If these mutations occur in bases critical to the structure or function of a protein or gene, altered function, disease, or a lethal condition can result. Most pathologic mutations tend not to be preserved throughout many generations because of their unfavorable phenotypes. Exceptions, such as the hemoglobinopathies, occur when the heterozygous state for these mutations confers selective advantage in the face of unusual environmental conditions, such as malaria epidemics. These "adaptive" mutations drive the dynamic change in the genome with time (evolution).

Because these copying errors occur randomly most will occur in either the vast stretches of intergenic DNA or the "silent" bases of gene DNA, such as the degenerate third bases of codons. They thus do not pathologically alter the function of the gene or its products. These clinically harmless mutations are called *DNA polymorphisms*. DNA polymorphisms can be regarded in exactly the same way as other types of polymorphisms that have been widely recognized for years (e.g., eye and hair color, blood groups). They are variations in the population that occur without apparent clinical impact. Each of us differs from other humans in the precise number and type of DNA polymorphisms that we possess. Most polymorphisms represent single-nucleotide changes and are called single-nucleotide polymorphisms (SNPs).

DNA polymorphisms breed true. In other words, if an individual's DNA contains a G 1200 bases upstream from the α-globin gene, instead of the C most commonly found in the population, that G will be transmitted to that individual's offspring. Note that if one had a means for distinguishing the G at that position from a C, one would have a linked marker for that individual's α-globin gene. Before the completion of the human genome project, only limited regions of the genome could be analyzed by direct DNA sequencing and SNPs were detectable only if they altered the recognition site for one or more restriction endonucleases, enzymes that cut DNA only at sites possessing a specific recognition sequence (Fig. 1.7). SNPs not altering such sites were not readily detectable. Contemporary DNA sequencing methods now allow for routine comprehensive cataloging of SNPs in a population or individual. However, the principles of choosing the right comparison populations and of the "breeding true" through generations remain important principles in interpreting the results.

The importance of polymorphic variations in each is that they can be used to identify individuals uniquely and to compare two individuals at the genomic level. For example, the severity of sickle cell anemia varies greatly, even within families, even though the disease is always caused by a specific point mutation in the β-globin gene. This suggests that the products of other genes exert a modifying effect on clinical phenotype. By scanning the genomes of many sickle cell patients of varying severity, an SNP was identified in less severely affected individuals near the Bcl11a gene, which was then shown to participate in the perinatal shutdown of fetal hemoglobin synthesis. Less severely affected individuals turned out to express a less active variant of Bcl11a. Similarly, the pattern of variations in the polymorphisms strung along the HLA gene cluster on chromosome 6 (i.e., the "haplotype") can be measured to compare the HLA "match" between two individuals and assess the compatibility of a potential bone marrow donor and recipient. The term haploidentical transplant is derived form a donor-recipient pair who have matching HLA cluster haplotypes.

Repeated Sequence Motifs

A related important feature of the DNA landscape is the existence of highly repeated DNA sequence motifs. A DNA sequence is said to be repeated if it or a sequence very similar (homologous) to it occurs more than once in a genome. Some multicopy genes, such as the histone genes and the rRNA genes, are repeated DNA sequences. However, most repeated DNA occurs outside genes, or within introns. Indeed, 30% to 45% of the human genome appears to consist of repeated DNA sequences.

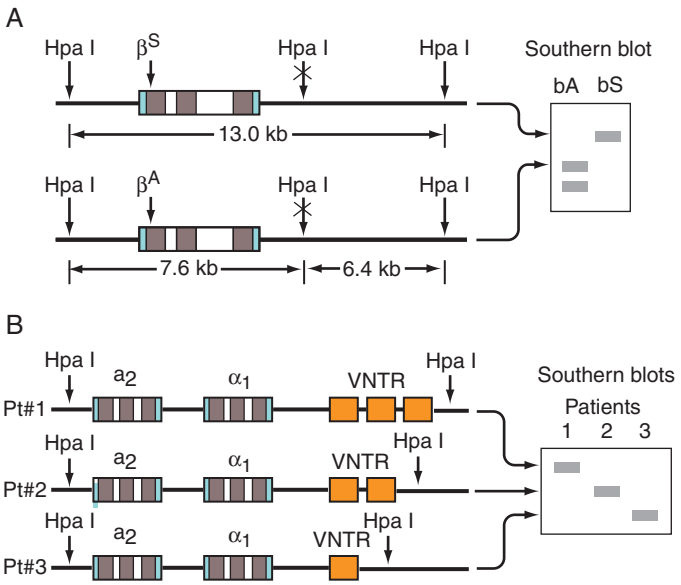

Figure 1.7 TWO USEFUL FORMS OF SEQUENCE VARIATION AMONG THE GENOMES OF NORMAL INDIVIDUALS. (A) Presence of a DNA sequence polymorphism that falls within a restriction endonuclease site, thus altering the pattern of restriction endonuclease digests obtained from this region of DNA on Southern blot analysis. (Readers not familiar with Southern blot analysis should return to examine this figure after reading later sections of this chapter.) (B) A variable-number tandem repeat (VNTR) region (defined and discussed in the text). Note that individuals can vary from one to another in many ways according to how many repeated units of the VNTR are located on their genomes, but restriction fragment length polymorphism differences are in effect all-or-none differences, allowing for only two variables (restriction site presence or absence).

The function of repeated sequences remains unknown, but their presence has inspired useful strategies for detecting and characterizing individual genomes. For example, a pattern of short repeated DNA sequences, characterized by the presence of flanking sites recognized by the restriction endonuclease Alu-1 (called "Alu repeats") occurs approximately 300,000 times in a human genome. These sequences are not present in the mouse genome. If one wishes to infect mouse cells with human DNA and then identify the human DNA sequences in the infected mouse cells, one simply probes for the presence of Alu repeats. The Alu repeat thus serves as a signature of human DNA.

Classes of highly repeated DNA sequences (tandem repeats) have proven to be useful for distinguishing genomes of each human individual. These short DNA sequences, usually less than a few hundred bases long, tend to occur in clusters, with the number of repeats varying among individuals (see Fig. 1.6). Alleles of a given gene can therefore be associated with a variable number of tandem repeats (VNTRs) in different individuals or populations. For example, there is a VNTR near the insulin gene. In some individuals or populations, it is present in only a few tandem copies, but in others, it is present in many more. When the population as a whole is examined, there is a wide degree of variability from individual to individual as to the number of these repeats residing near the insulin gene. It can readily be imagined that if probes were available to detect a dozen or so distinct VNTR regions, each human individual would differ from virtually all others with respect to the aggregate pattern of these VNTRs. Indeed, it can be shown mathematically that the probability of any two human beings sharing exactly the same pattern of VNTRs is exceedingly small if approximately 10 to 12 different VNTR elements are mapped

for each person. A technique called *DNA fingerprinting* that is based on VNTR analysis has become widely publicized because of its forensic applications.

There are many other classes of repeated sequences in human DNA. For example, human DNA has been invaded many times in its history by retroviruses. Retroviruses tend to integrate into human DNA and then "jump out" of the genome when they are reactivated, to complete their life cycle. The proviral genomes often carry with them nearby bits of the genomic DNA in which they sat. If the retrovirus infects the DNA of another individual at another site, it will insert this genomic bit. Through many cycles of infection, the virus will act as a transposon, scattering its attached sequence throughout the genome. These types of sequences are called *long interspersed elements*. They represent footprints of ancient viral infections.

MOLECULAR GENETIC METHODOLOGIES ALLOWING THE ISOLATION, ANALYSIS, AND MANIPULATION OF GENES

The application of molecular genetics to the understanding, diagnosis, treatment, and prevention of hematologic diseases became possible in limited ways during the 1970s and 1980s, when a variety of experimental methods, both biochemical and genetic, made it possible to isolate any desired DNA fragment from chromosomes, or from DNA copies of cellular RNA (cDNAs). These methodologies, such as "Southern" blotting analysis of DNA, "Northern" blotting of RNA, and initial DNA sequencing techniques, although elegant, were laborious and required sophisticated personnel and equipment. They are now largely of historical interest, although still useful for some purposes. Four methodologies that made widespread routine use of DNA- and RNA-based disease-oriented research, diagnostics, and therapeutics feasible are the polymerase chain reaction (PCR), gene cloning, high-throughput DNA sequencing, and gene transfer techniques. The latter allows one to insert of the genetic material of choice into almost any desired cells, tissues, or organisms. All of these capabilities have been greatly enhanced by advances in computational methods, computerization, and automation. These four merit a brief introductory discussion because they are alluded to in many chapters in this book.

The Polymerase Chain Reaction

The development of the PCR revolutionized DNA-based strategies for diagnosis and treatment. It permits the detection, synthesis, and isolation of specific genes and allows one to discriminate among the alleles of a gene differing by as little as one base. It requires only readily available equipment and basic technical skills. A specimen consisting of only minute amounts of material will suffice; in most circumstances, no special preparation of the tissue is necessary. PCR made direct genetic and genomic analyses readily accessible to clinical, epidemiologic, and forensic laboratories. This single advance fueled quantum increases in the use of direct gene analysis for diagnosis of human diseases. Indeed, PCR analysis combined with direct DNA sequencing technologies have largely supplanted older strategies, such as restriction enzyme mapping and DNA/RNA blotting strategies for many research and diagnostic applications, although these older methods remain useful for some niche applications. PCR coupled with now-routinely available gene cloning methodologies allows one to synthesize in microgram quantities naturally occurring or engineered genes at will. These can then readily be inserted into cells, tissues, or organisms where they will be expressed and their physiologic or pathologic effects investigated. Similarly, industrial scale production of novel therapeutics based on the PCR-designed DNA itself or its expressed RNA or protein products is now routine. Hematopoietic growth factors and monoclonal antibody therapeutics are just two examples of widely used hematologic therapies that depended on these strategies.

PCR is based on the prerequisites for copying an existing DNA strand by DNA polymerase: an existing denatured strand of DNA to be used as the template and primers. Primers are short oligonucleotides, 12 to 100 bases in length, having a base sequence complementary to the desired region of the existing DNA strand. Oligonucleotide primers are now easily designed and produced using biochemical techniques developed in the 1970s and 1980s. The primer allows the polymerase to "know" where to begin copying. If the base sequence of the DNA of the gene under study is known (see DNA sequencing), two synthetic oligonucleotides complementary to sequences flanking the region of interest can be prepared. If these are the only oligonucleotides present in the reaction mixture, then the DNA polymerase can copy only daughter strands of DNA downstream from those oligonucleotides. In other words, it can copy only that gene. Recall that DNA is double stranded, that the strands are held together by the rules of Watson-Crick base pairing, and that they are aligned in antiparallel fashion. This implies that the effect of incorporation of both oligonucleotides into the reaction mix will be to synthesize two daughter strands of DNA, one originating upstream of the gene and the other originating downstream. The net effect is synthesis of only the DNA between the two primers, thus doubling only the DNA containing the region of interest. If the DNA is now heat denatured and then cooled again, allowing hybridization of the daughter strands to the primers, and the polymerization is repeated, then the region of DNA through the gene of interest is doubled again. Thus two cycles of denaturation, annealing, and elongation result in a selective quadrupling of the gene of interest. The cycle can be repeated 30 to 50 times, resulting in a selective and geometric amplification of the sequence of interest to the order of 2^{30} to 2^{50} times. The result is a millionfold or higher selective amplification of the gene of interest, yielding microgram quantities of that DNA sequence.

PCR achieved practical utility when DNA polymerases from thermophilic bacteria were discovered; when synthetic oligonucleotides of any desired sequence could be produced efficiently, reproducibly, and cheaply by automated instrumentation; and when DNA thermocycling machines were developed. Thermophilic bacteria live in hot springs and other exceedingly warm environments, and their DNA polymerases can tolerate 100°C (212°F) incubations without substantial loss of activity. The advantage of these thermostable polymerases is that they retain activity in a reaction mix that is repeatedly heated to the high temperature needed to denature the DNA strands into the single-stranded form. Microprocessor-driven DNA thermocycler machines can be programmed to increase temperatures to 95°C to 100°C (203°F to 212°F) (denaturation), to cool the mix to 50°C (122°F) rapidly (a temperature that favors oligonucleotide annealing), and then to raise the temperature to 70°C to 75°C (158°F to 167°F) (the temperature for optimal activity of the thermophilic DNA polymerases). In a reaction containing the test specimen, the thermophilic polymerase, a sufficient supply of primers to support the amplification, and the chemical components needed to sustain the multiple rounds of copying (e.g., nucleotide triphosphate precursors, reaction buffer, an adenosine triphosphate [ATP]-generating system to support the endothermic polymerase reaction), the thermocycler can conduct many cycles of denaturation, annealing, and polymerization in a completely automated fashion. The gene of interest can thus be amplified more than a millionfold in a matter of a few hours. The DNA product is readily identified and isolated by routine agarose gel electrophoresis. The DNA can then be analyzed by restriction endonuclease, digestion, hybridization to specific probes, sequencing, further amplification by cloning, and so forth.

Reverse transcriptases (RNA-dependent DNA polymerases) derived from retroviruses greatly extend the utility of PCR. By copying all the RNAs into their cDNAs, reverse transcriptase allows RNA sequences in a specimen to be amplified much like DNA sequences. This procedure, called reverse transcription (RT)-PCR, inserts a reverse transcriptase step into the beginning of the procedure, which then proceeds exactly like PCR. RT-PCR permits one to amplify all of the mRNAs expressed in a cell for high-throughput nucleotide sequence analysis, to detect just one or a few mRNAs to analyze their

expression patterns, or to clone them (see later) to isolate their encoding genes.

High-Throughput DNA and RNA Sequencing

Knowing the nucleotide base sequence of a gene, its RNA products, its flanking regulatory elements, and its variation in a disease state is essential to understanding its normal or pathologic behavior. Techniques for sequencing (i.e., deciphering the nucleotide base sequence) DNA that emerged in the 1970s were valuable but limited. Only short stretches of a few hundred bases could be read during a single "run." The methods required the use of radioactive tracers, sophisticated electrophoretic steps, and/or toxic chemicals. Nonetheless, the coding sequences of many genes relevant to hematologic disorders were obtained in this way. Fortunately, the human genome project inspired major technologic innovations (e.g., in the application of physicochemical and chromatographic principles to nucleic acid chemistry, the development of novel nonradioactive tracers, and the creation of software and firmware that allowed one to assemble the sequences of multiple independent sequencing "runs" of shorter fragments into a coherent sequence of the whole length of a gene). Sequencing of millions of nucleotides in a single sitting became feasible.

Modern sequencing techniques are commonly described as high-throughput sequencing or "next-gen" (i.e., next-generation) sequencing. Their efficiency and cost-effectiveness are such that whole genome sequences can now be gotten from a clinical specimen within a few days for a direct cost of less than a thousand dollars. The profound effect that these advances have had on the practical utility of DNA analysis in medicine is evident in the routine application of high-throughput sequencing to tumor specimens to identify therapeutic targets or infer prognostic information or the many thousands of SARS-CoV-2 genomes sequenced every day to track variants.

Next-gen sequencing has inspired the discipline of genomics, which attempts to understand the anatomy and functioning of any gene in the context of all of the DNA in the entire genome of a cell. Indeed, the technology has advanced to the point that one can sequence the genome of a single cell. Similarly, one can obtain the sequences of all of the mRNAs expressed in a specimen or even a single cell (the transcriptome) by first copying the cellular RNA into cDNA. This is called RNA sequencing or RNAseq.

Chapter 3 discusses genomics and the uses of sequencing technologies in hematology in greater detail.

Gene Cloning

PCR allows one to generate microgram amounts of pure DNA fragments up to a few kilobases in length. Most genes are considerably longer than that. To study their function or pathology, one needs to isolate the entire gene and its flanking sequences and insert it into cells for expression. Moreover, for any applications, such as manufacturing DNA reagents for diagnostic kits, the capability to generate much larger amounts is desirable. Gene cloning, or recombinant DNA technology, is a collection of methods that meets these goals. Basically, an amplified PCR fragment, or a mixture of all of the DNA fragments from a cell up to megabase lengths (1 megabase =1 million bp) generated by sonication or limited nuclease digestion, is modified at the ends with oligonucleotide "adaptors" that allow them to be ligated into a "vector." In this context, a vector is an engineered microbial DNA element that can be inserted into a host cell, where it will coexist with the host genome and be able to be expressed. The most common vectors are viral genomes that were engineered to retain infectiousness but have had their pathogenic properties removed from their genomes.

If the "recombinant" genome has been placed in a bacteriophage genome and exposed to an excess of host bacterial cells, each cell acquires a single recombinant molecule. When cultured at low density on petri plates, each colony that grows out is a clone derived from a single transfected bacterium that in turn contains and expresses a single recombinant molecule. Many screening techniques have been devised by which one can identify and purify the clone(s) containing the desired DNA fragment among the thousands of clones on the plates. The clone can then be grown in bulk culture to generate large amounts of that DNA fragment for analysis, used as a diagnostic or experimental probe, or refined for use as a therapeutic, for transfer into cells, tissue, or whole organisms for studies of its biologic function. "Gene cloning" is thus named for the fact that the method allows one to capture, purify, and mass produce any single desired DNA fragment (e.g., a whole gene) in a single bacterial clone. This clone can also be preserved in a manner that sustains viability and be used repeatedly to generate additional DNA. Much of our contemporary molecular understanding of hematologic pathobiology has been gleaned by application of gene cloning approaches. Important therapeutics, such as erythropoietin, granulocyte-macrophage colony-stimulating factor (GM-CSF), monoclonal antibody therapeutics, CAR-T cells, and many more, are derived from recombinant DNA molecule purified by gene cloning methods.

Extensions and variations of techniques of gene cloning into bacteria have made possible the cloning of genes into cells of a wide variety of species, including human tissue culture cells. This adds great versatility to the methodology for expressing large quantities of the RNAs or proteins encoded by the cloned genes with all the appropriate posttranslation modifications present in their natural state.

Use of Transgenic and Knockout/Knockin Organisms to Model Gene Function

Recombinant DNA technology has resulted in the identification of many disease-related genes. To advance the understanding of the disease related to a previously unknown gene, the function of the protein encoded by that gene must be verified or identified, and the way changes in the gene's expression influence the disease phenotype must be characterized. Analysis of the role of these genes and their encoded proteins was made possible by the development of recombinant DNA technology that allowed the production of mice that are genetically altered at the cloned locus. Mice can be produced that express an exogenous gene and thereby provide an in vivo model of its function. Linearized DNA is injected into a fertilized mouse oocyte pronucleus and reimplanted in a pseudopregnant mouse. The resultant transgenic mice can then be analyzed for the phenotype induced by the injected transgene. Placing the gene under the control of a strong promoter that stimulates expression of the exogenous gene in all tissues allows the assessment of the effect of widespread overexpression of the gene. Alternatively, placing the gene under the control of a regulatory sequence that can function only in certain tissues (a tissue-specific promoter) elucidates the function of that gene in a particular tissue or cell type. A third approach is to study control elements of the gene by testing their capacity to drive expression of a "marker" gene that can be detected by chemical, immunologic, or functional means. For example, the promoter region of a gene of interest can be joined to the cDNA encoding green jellyfish protein and activity of the gene assessed in various tissues of the resultant transgenic mouse by fluorescence microscopy. Use of such a reporter gene demonstrates the normal distribution and timing of expression of the gene from which the promoter elements are derived. Transgenic mice contain exogenous genes that insert randomly into the genome of the recipient. Expression can thus depend as much on the location of the insertion as it does on the properties of the injected DNA.

In contrast, any defined genetic locus can be specifically altered by targeted recombination between the locus and a plasmid carrying an altered version of that gene (Fig. 1.8). If a plasmid contains that altered gene with enough flanking DNA identical to that of the normal gene locus, homologous recombination can occur, and the altered gene in the plasmid will replace the gene in the recipient cell. Using a mutation that inactivates the gene allows the production of a null mutation, in which the function of that gene is completely lost. To induce such a mutation, the plasmid is introduced into an embryonic stem cell, and the rare cells that undergo homologous recombination

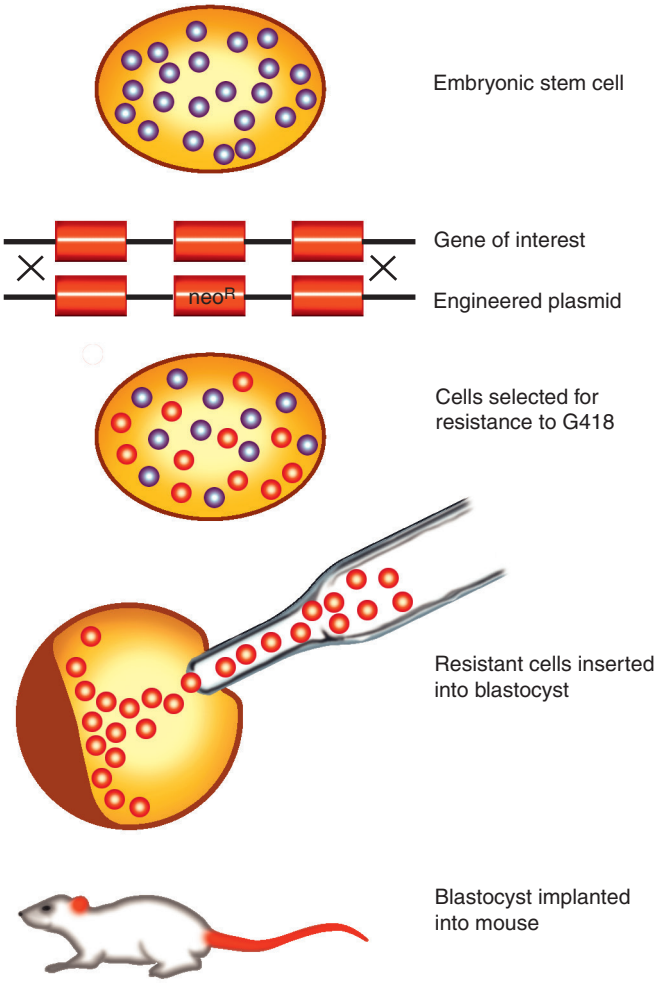

Figure 1.8 GENE "KNOCKOUT" BY HOMOLOGOUS RECOMBINATION. A plasmid containing genomic DNA homologous to the gene of interest is engineered to contain a selectable marker positioned so as to disrupt expression of the native gene. The DNA is introduced into embryonic stem cells, and cells resistant to the selectable marker are isolated and injected into a mouse blastocyst, which is then implanted into a mouse. Offspring mice that contain the knockout construct in their germ cells are then propagated, yielding mice with heterozygous or homozygous inactivation of the gene of interest.

are selected. The "knockout" embryonic stem cell is then introduced into the blastocyst of a developing embryo. The resultant animals are chimeric; only a fraction of the cells in the animal contain the targeted gene. If the new gene is introduced into some of the germline cells of the chimeric mouse, then some of the offspring of that mouse will carry the mutation as a gene in all of their cells. These heterozygous mice can be further bred to produce mice homozygous for the null allele.

Knockout mice reveal the function of the targeted gene by the phenotype induced by its absence. Methods for "knocking in" a gene have been developed to allow one to assess the functional consequences of replacing the function of the knocked-out gene with a modified version of that gene or an alternative gene with a related function. Genetically altered mice have been essential for discerning the biologic and pathologic roles of large numbers of genes implicated in the pathogenesis of human disease. These methods were originally developed in mice, but they have been extended to many animal species. The methods are now refined enough to generate recombinant organisms in which multiple endogenous genes are replaced by human genes, generating model organisms "humanized" for certain key functions (e.g., hemoglobin synthesis in mouse erythroid cells, certain immune functions); these humanized models are proving useful for preclinical testing of novel therapeutics.

DNA- AND RNA-BASED THERAPEUTICS

Gene Therapy and Gene Editing

The application of gene therapy to genetic hematologic disorders is an appealing idea. In some cases, this would involve isolating hematopoietic stem cells from patients with diseases with defined genetic lesions, inserting normal genes into those cells, and reintroducing the genetically engineered stem cells back into the patient. A few candidate diseases for such therapy include sickle cell disease, thalassemia, hemophilia, and adenosine deaminase–deficient severe combined immunodeficiency. The technology for separating hematopoietic stem cells and for performing gene transfer into those cells has advanced rapidly, and clinical trials are actively testing the applicability of these techniques. Indeed, the use of this "ex vivo" approach has led to the approval in Europe of a therapeutic gene for β-thalassemia. In other cases, such as treatments for hemophilia, the therapeutic gene is injected directly into a target tissue or infused. In both cases the gene must be packaged in a vector, usually a virus engineered to infect a particular cell type and to have lost any potential to cause a viral disease pathogenic. Presently, there are only few (but increasing, such as severe combined immunodeficiency syndromes, Wiskott–Aldrich disease, and thalassemia) proven therapeutic successes from gene therapy.

Progress in this field continues rapidly and is likely to accelerate as a consequence of the development of "gene editing" technologies (see Chapter 5). Among these, "CRISPR" is the most prominent current example. It is based on the discovery of enzyme systems used by microorganisms to excise foreign DNA sequences (e.g., integrated viral genomes) from the host genome. These systems can be adapted to insert, replace, or delete, in principle, any desired DNA sequence at its naturally occurring position in the host genome. For example, one could excise the mutation causing sickle cell anemia and replace it with the normal DNA sequence in the β-globin gene of a patient's hematopoietic stem cells and then reintroduce them into the patient's bone marrow without introducing any foreign DNA. This exciting technology is in clinical trials for a number of hematologic conditions, including hemoglobinopathies.

RNA Therapeutics

The recognition that abnormal expression of oncogenes plays a role in malignancy has stimulated attempts to suppress oncogene expression to reverse the neoplastic phenotype. One early attempt blocking mRNA expression is with antisense oligonucleotides. These are single-stranded DNA sequences 17 to 20 bases long, having a sequence complementary to the transcription or translation start of the mRNA. These relatively small molecules can be engineered with modified nucleotides that resist nucleotide destruction and freely enter the cell, where they complex to the targeted mRNA by Watson-Crick base pairing. Alternatively, one can use a modified gene therapy approach by transfecting the cells with a DNA segment encoding the antisense RNA. The binding of the oligonucleotide may directly block translation and clearly enhances the rate of mRNA degradation, thus downregulating the expression of the desired gene. The discovery, mentioned earlier, of naturally occurring small inhibitory RNAs has stimulated the development of RNA therapeutics that have largely superseded the original antisense approach.

RNA therapeutics is a burgeoning field of early drug development. Synthetic small hairpin RNAs containing modified nucleotides that stabilize them in the circulation and tissue spaces can be readily manufactured and engineered to contain any desired nucleotide sequences needed to identify and bind to only the targeted gene or RNA gene product, form metabolically active complexes with other intracellular

RNAs or proteins, and thereby achieve the desired therapeutic effect. RNA therapeutics are promising to be extremely versatile. In addition to binding to the target mRNA to block its translation and enhance its destruction, engineered shRNAs have been successfully designed to interact with the translational apparatus to "read through" or "skip over" nonsense codons, permitting completion of translation of the mutated protein, and to interact with the pre-mRNA splicing apparatus to alter the pattern of alternative mRNA splicing of the desired pre-mRNA in a physiologically favorable way. The latter strategy has been elegantly deployed to develop an FDA-approved therapy for spinal muscular atrophy. Using more conventional gene therapy methods to employ an shRNA targeting the binding of Bcl11a to its erythroid specific enhancer, thereby blocking the postnatal shutdown of fetal hemoglobin, is also being tested in clinical trials for treating sickle cell anemia and β-thalassemia.

FUTURE DIRECTIONS

The elegance of recombinant DNA technology and its successor technologies of genomics, epigenomics, proteomics, genetic therapies, gene editing, and RNA therapeutics resides in the capacity they confer on investigators to examine each gene as a discrete physical entity that can be purified, reduced to its basic building blocks for decoding of its primary structure, analyzed for its patterns of expression, and perturbed by alterations in sequence or molecular environment so that the effects of changes in each region of the gene can be assessed. Purified genes can be deliberately modified or mutated to create novel genes not available in nature. These provide the potential to generate useful new biologic entities, such as modified live virus or purified peptide vaccines, modified proteins customized for specific therapeutic purposes, and altered combinations of regulatory and structural genes that allow for the assumption of new functions by specific gene systems.

The most important impact of the genetic approach to the analysis of biologic phenomena is the most indirect. Diligent and repeated application of the methods outlined in this chapter to the study of many genes from diverse groups of organisms is beginning to reveal the basic strategies used by nature for the regulation of cell and tissue behavior. As our knowledge of these rules of regulation grows, our ability to understand, detect, and correct pathologic phenomena will increase substantially. So too will the complexity of ethical and policy issues about what comprises the appropriate and inappropriate uses of technologies capable of altering the nature of what it means to be human. For all of these reasons, it is incumbent on students of hematology to be as conversant with this discipline.

SUGGESTED READINGS

Bentley D. The mRNA assembly line: transcription and processing machines in the same factory. *Curr Opin Cell Biol*. 2002;14:336.

Collins FS, Doudna JA, Lander ES, Routimi CN. Human molecular genetics and genomics—important advances and exciting possibilities. *N Engl J Med*. 2021;384:1–4.

Dykxhoorn DM, Novina CD, Sharp PA. Killing the messenger: short RNAs that silence gene expression. *Nat Rev Mol Cell Biol*. 2003;4:457.

Fischle W, Wang Y, Allis CD. Histone and chromatin cross-talk. *Curr Opin Cell Biol*.. 2003;15:172.

Grewal SI, Moazed D. Heterochromatin and epigenetic control of gene expression. *Science*. 2003;301:798.

Jones B. Layers of gene regulation. *Nat Rev Genet*. 2015;16:128–129.

Jongbloed JDH, Lekanne Deprez RH, Vatta M. Introduction to molecular genetics. In: Baars HF, Doevendans PAFM, Houweling A, van Tintelen J, eds. *Clinical Cardiogenetics*. Cham: Springer; 2016.

Kloosterman WP, Plasterk RHA. The diverse functions of microRNAs in animal development and disease. *Dev Cell*. 2006;11:441.

Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci*. 2006;31:89.

Kumar A, Garg S, Garg N. Regulation of gene expression: RNA regulation. In: Meyers RA, ed. *Synthetic Biology,* Vol. 1. Weinheim: Wiley-VCH Verlag; 2014:61–121.

Lee TI, Young RA. Transcription of eukaryotic protein-coding genes. *Ann Rev Genet*. 2000;34:77.

Tefferi A, Wieben ED, Dewald GW, et al. Primer on medical genomics, part II: background principles and methods in molecular genetics. *Mayo Clin Proc*. 2002;77:785.

Waddington S, Privolizzi R, Karda R, et al. A broad overview and review of CRISPR-CAS technology and stem cells. *Curr Stem Cell Rep*. 2016;2:9–20.

Wilusz CJ, Wormington M, Peltz SW. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol*. 2001;2:237.