CHAPTER

MOLECULAR BIOLOGY OF THE CELL

1

The aim of the first chapter is to provide a primer covering our understanding of the basic molecular and cellular processes of the cell, which inform a scientific understanding of hematologic diseases.

COMPARTMENTALIZATION OF THE CELL

A central evolutionary advance was the compartmentalization of cells, as shown in Fig. 1.1. The cell is bounded by a complex cell membrane that allows regulation of molecules into and out of the cell. Within the cytoplasm a number of different organelles perform key functions. For example, as described later in this chapter, mitochondria are critical for adenosine triphosphate (ATP) generation and heme biosynthesis. Proteins are translated from amino acids and undergo post-translational modification in the Golgi complex and rough endoplasmic reticulum. Depending on the cell type, there are specialized structures within the cytoplasm that allow the cell to perform its specialized role.

THE NUCLEUS

As we focus in on the nucleus, it is clear that it is also bounded by a specialized nuclear envelope and membrane (Fig. 1.2). Entry and exit out of the nucleus is regulated by nuclear pores. Within the nucleus, deoxyribonucleic acid (DNA) is tightly packaged by proteins and the DNA/protein complex is known as chromatin. Chromatin has different appearances under light or electron microscopes. When DNA is tightly packaged (and the genes more likely to be not expressed), it is known as heterochromatin. Under the light/electron microscope it appears darker. When DNA is less tightly packaged it is called euchromatin and is lighter in appearance. The other visible structure within the nucleus, in some cells, is the nucleolus, where ribosomal genes are transcribed and assembly of the ribosome takes place (as discussed later).

The DNA in the nucleus is distributed among 22 pairs of autosomal chromosomes (numbered 1–22, in order of size) and two sex chromosomes (Fig. 1.3A). When cells are in the metaphase phase of the cell cycle, chromosomes condense and can be visualized by a technique called karyotyping. Chromosomes are divided into two arms: a short arm, termed

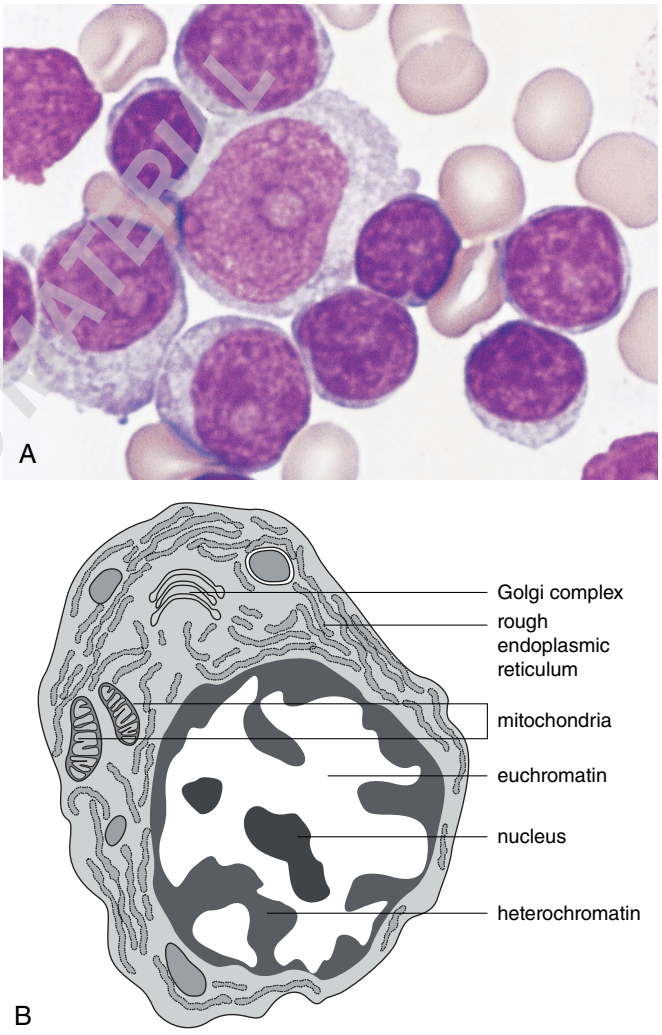


Fig. 1.1. A, Photomicrograph showing the morphology of many cells with prominent nucleoli, in this case, B cells. **B,** A schematic representation of the intracellular composition as visualized by electron microscopy. The nucleus is composed of euchromatin, which is less condensed, paler, and more transcriptionally active, and heterochromatin, which is more condensed, darker, and less transcriptionally active. In cytoplasm subcellular organelles including mitochondria, rough endoplasmic reticulum, and the Golgi complex are shown. The function of these organelles is discussed later. (Courtesy of Professor JV Melo.)

Color Atlas of Clinical Hematology: Molecular and Cellular Basis of Disease, Fifth Edition. A. Victor Hoffbrand, Paresh Vyas, Elías Campo, Torsten Haferlach, and Keith Gomez

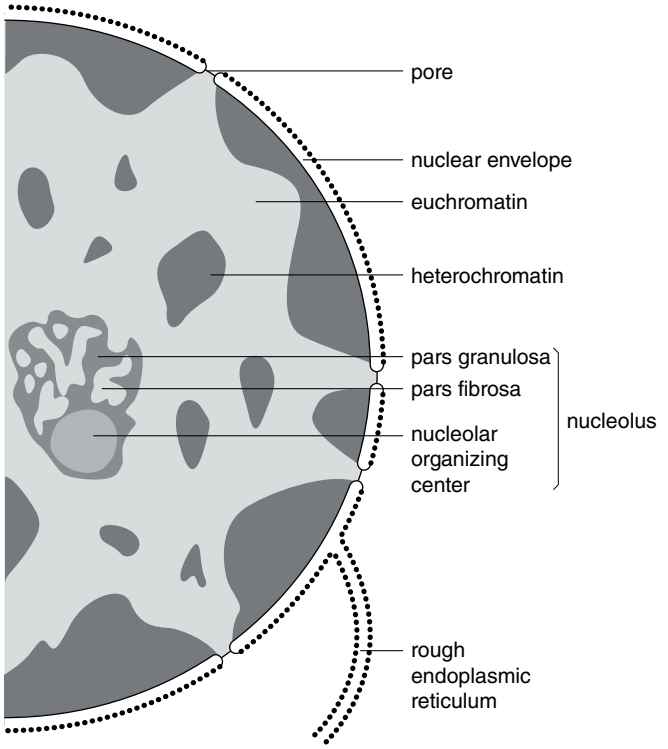


Fig. 1.2. Schematic representation of a portion of the nucleus. The nucleus is highly compartmentalized, containing specialized structures. The nucleolus is composed of a pars granulosa, a pars fibrosa, and a nucleolar organizing center and makes transfer RNA. The nucleus is bounded by a nuclear envelope that is lined by rough endoplasmic reticulum. There is controlled entry and exit into the nucleus via nuclear pores.

p, and a longer arm, q. The region where the chromosomes join is termed the centromere. Chromosomes are further subdivided into light and dark bands (depending on how they stain with the Giemsa dye) (Fig. 1.3B). When cells are not in metaphase, chromosomes are more diffusely spread through the nucleus. Most current evidence suggests that the chromosomes occupy discrete territories (chromosomal territories) within a nucleus (Fig. 1.3C). These territories need not be contiguous and can be shared with other chromosomes. However, there are still many aspects of how chromosomes are organized that remain unclear. For example, what constrains chromosomes to territories and how do territories affect gene regulation? Recent work suggests that within chromosomal territories chromatin exists in topologically associated domains (TADs) and that actively expressed genes along the chromosome and possibly even from different chromosomes may congregate in specialized structures where RNA is made from (transcribed) from genes. This process is called transcription and the specialized structures are known as transcription factories (see later).

The sequencing of the human genome was a landmark in biology. It allowed all the human genes arrayed along the chromosomes to be catalogued (Table 1.1). Genes are divided into protein-coding genes (of which there are ~21 000), genes that encode different types of RNA (e.g. ribosomal RNA, micro-RNAs, small nuclear RNA), and RNA moieties that are not translated into a functional protein or RNA (pseudogenes). The genome also dedicates sequence to other RNA species that do not make protein but that regulate either transcription or the production of protein from RNA (a process known as translation). These RNA sequences include micro-RNAs, long and

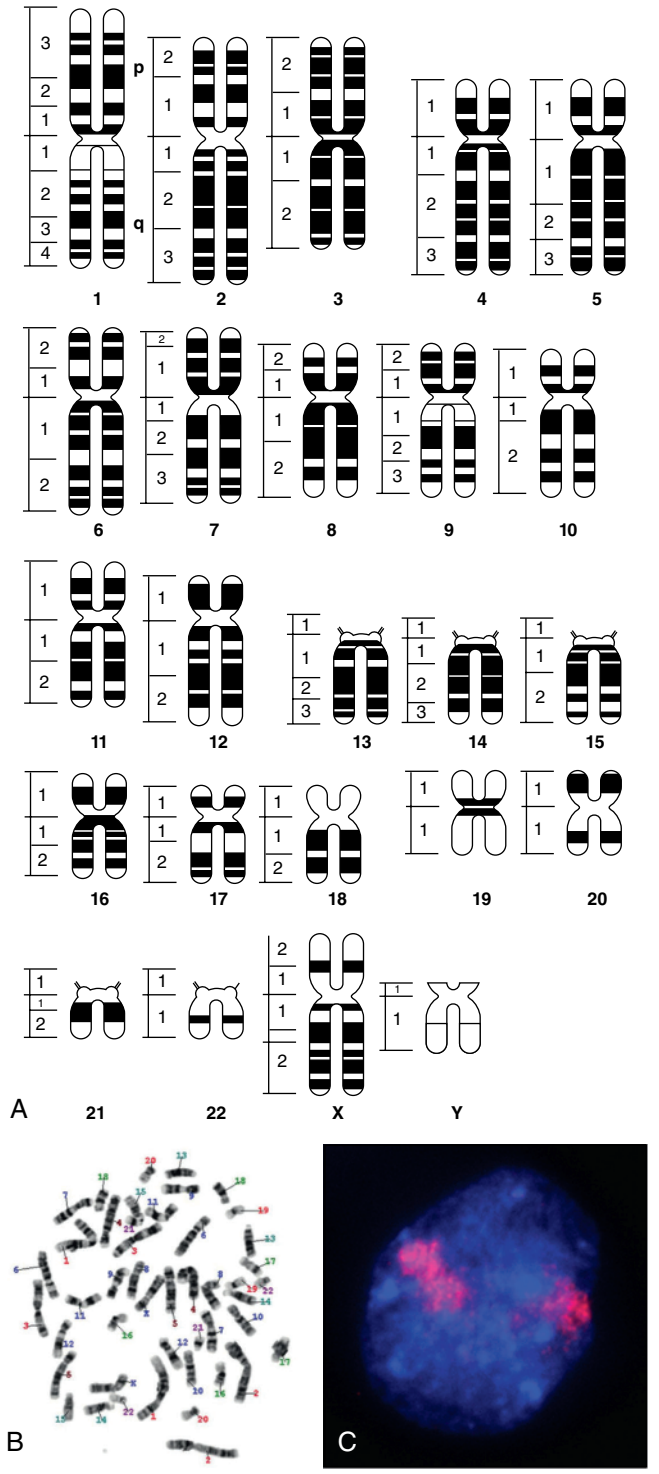


Fig. 1.3. A, DNA in the human nucleus is organized into 46 chromosomes. There are two copies of chromosomes 1–22 with two sex chromosomes (XX or XY). Each chromosome is divided into a short arm (p) and a long arm (q) and then subdivided into major numeric subsections. For example, the short arm of chromosome 1 (1p) has three subsections and the long arm (1q) has four subsections. **B,** The gross subdivision of chromosome can be visualized by Giemsa staining of chromosomes that have been subject to brief proteolytic cleavage. (B, Courtesy of Professor H Lodish.) **C,** Within an interphase nucleus chromosomes occupy discrete territories. The figure shows the territory occupied by chromosome 11 (red color) in a primary erythroblast. (C, Courtesy of Jo Green and Dr. Veronica Buckle.)

TABLE 1.1. ALL THE GENES AND OPEN READING FRAMES IN THE HUMAN GENOME HAVE BEEN CHARACTERIZED FROM THE SEQUENCING OF THE HUMAN GENOME. THIS TABLE SHOWS THE SIZE OF EACH CHROMOSOME (IN MEGABASES) AND NUMBER OF GENES AND PSEUDOGENES ON EACH CHROMOSOME.

Chromosome number	Size (Mb)	Gene	Pseudogene
1	248.96	5,078	1,372
2	242.19	3,862	1,166
3	198.3	2,971	887
4	190.22	2,441	799
5	181.54	2,578	766
6	170.81	3,000	876
7	159.35	2,774	896
8	145.14	2,152	661
9	138.4	2,262	702
10	133.8	2,174	631
11	135.09	2,920	835
12	133.28	2,521	680
13	114.36	1,381	477
14	107.04	2,055	583
15	101.99	1,814	555
16	90.34	1,920	451
17	83.26	2,432	541
18	80.37	988	295
19	58.62	2,481	514
20	64.44	1,349	329
21	46.71	756	202
22	50.82	1,172	348
X	156.04	2,158	859
Y	57.23	577	395
MT	0.016569	37	–

short noncoding RNAs. There are also sequences dedicated to regulating transcription of individual genes or banks of genes; these are called promoters and enhancers. This provides a primary description of our genetic makeup. The characterization of the human genome is still being refined as we understand more about how genes are organized and how transcriptional expression and protein translation is controlled.

Genes themselves are composed of DNA, which is made up of four nucleotides. Each nucleotide consists of a phosphate group linked by a phosphoester bond to a pentose sugar molecule (ribose) that lacks a hydroxyl group (thus it is deoxyribose), which is then attached to one of four heterocyclic carbon- and nitrogen-containing organic rings: adenine (A), cytosine (C), guanine (G) and thymidine (T). C and T are known as pyrimidines and A and G as purines. These are then linked together into polynucleotides via phosphoester bonds. As James Watson and Francis Crick correctly proposed, these are organized into two associated antiparallel polynucleotide strands that have a 5′ to 3′ direction and form a double helix. The strands are held in register by base-pairing between the two strands such that each A is paired with a T via two hydrogen bonds and each C with a G via three hydrogen bonds. Hydrophobic and van der Waals interactions combine with the thousands of hydrogen bonds to give the double helix great stability. In the common "B" form, the helix is right handed and makes a complete turn every 3.4 nm (about 10 base pairs) (Fig. 1.4A,B). The space between the strands creates a major and minor groove. In low humidity, DNA can adopt a more compact form with 11 base pairs per helical turn ("A" form) (Fig. 1.4C). Finally, short stretches of DNA composed of alternate purines and pyrimidines can form an alternate stacked Z structure.

GENE TRANSCRIPTION AND MESSENGER RNA TRANSLATION: THE PRODUCTION AND JOURNEY OF mRNA

A copy of the DNA of genes is transcribed into RNA by transcription in the nucleus. RNA is processed and transported into the cytoplasm. RNA corresponding to protein genes is then translated in the cytoplasm. Not surprisingly, these processes are very complex, affording opportunities for the cell to exquisitely regulate the complement of proteins made but also vulnerable to errors that lead to disease.

Genes are transcribed by one of three different RNA polymerases (RNA Pol I, II, and III). RNA Pol II transcribes most protein-coding genes. The remaining genes are transcribed by RNA Pol I and III. These include genes encoding ribosomal RNAs (makes ribosomes, see later), small nuclear RNAs (involved in processing RNA in a process called splicing, see later), and some transfer RNAs (involved in protein translation, see later). Pol I- and Pol II-transcribed genes will not be discussed in detail further in this book. However, it is important to remember that in a typical rapidly growing mammalian cell, ~80% of total RNA is ribosomal RNA and ~15% is transfer RNA.

When RNA is transcribed, a gene is said to be "expressed." Transcription of each gene begins at the 5′ end of the gene at its transcriptional start site (TSS) (Fig. 1.5). For any one gene the TSSs can either be single or multiple over several neighboring nucleotides. The DNA sequence 5′ of the gene helps to regulate transcription and is known as the promoter. This sequence works with other sequences (called regulatory sequences or *cis*-elements, see later) to provide finely tuned control over the amount of mRNA produced. In Chapter 9 the regulatory sequences involved in globin gene expression are described.

The body of the gene is segmented into exons separated by intervening sequences (introns). The exonic sequence is divided into protein-coding and noncoding sequences. RNA Pol II makes a RNA copy of the whole of the gene (primary transcript). This RNA species is then processed within the nucleus. As the nascent elongating primary transcript is produced a 5′ 7-methylguanine cap is added to the 5′ end to protect the RNA from enzymatic degradation. In addition, as nascent RNA transcript (heterogeneous RNA [hnRNA]) emerges from the RNA Pol II, it is sheathed in a large set of nuclear proteins in structures called heterogeneous ribonuclear particles (hnRNPs). hnRNP-associated proteins are important for transport of the RNA species and probably aid in the processing of RNA. Once the primary transcript is made, the 3′ end of the transcript is recognized by a protein complex that includes an enzyme called an endonuclease that cleaves the RNA transcript to produce a 3′

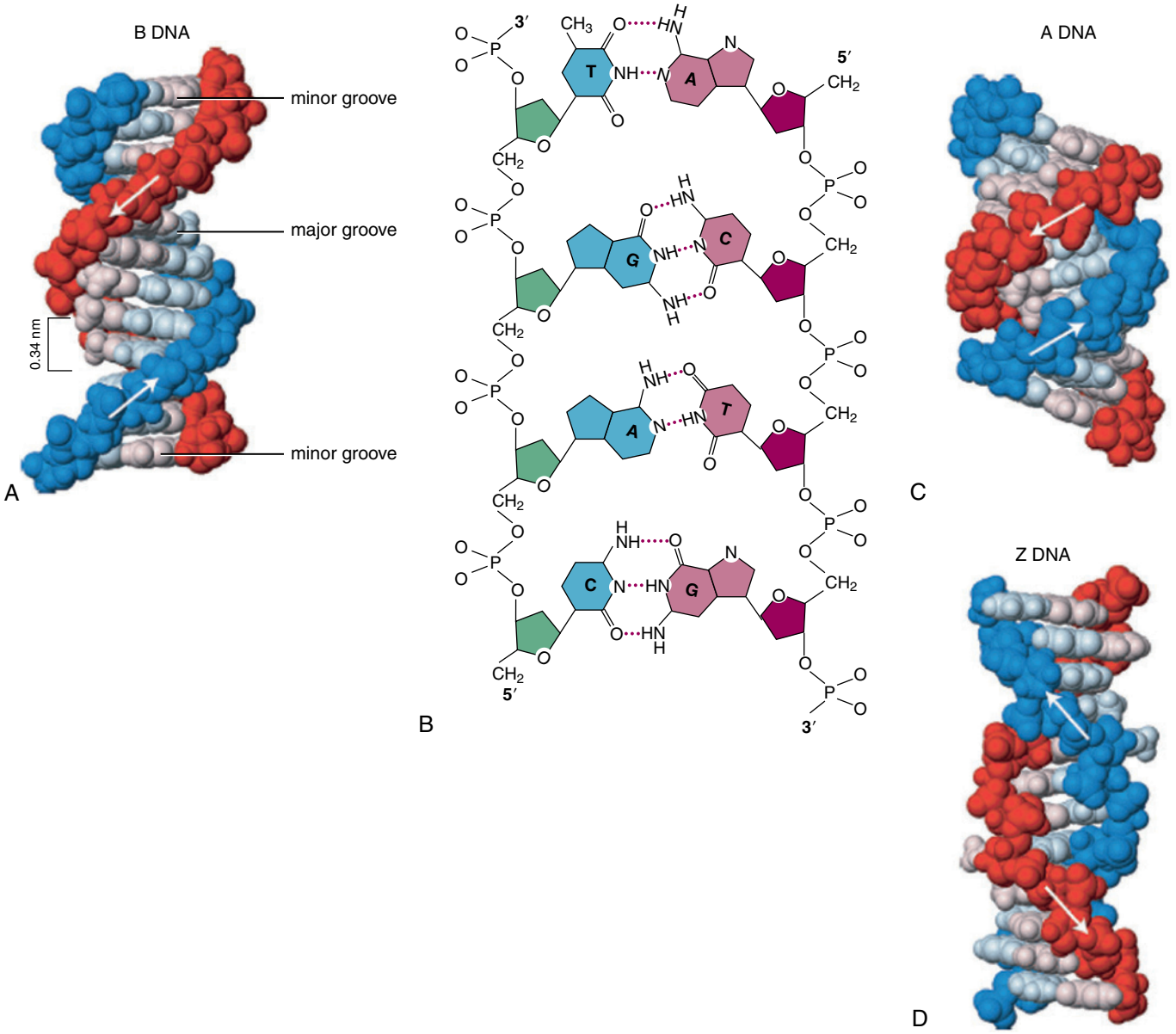


Fig. 1.4. Models of various structures adopted by DNA. **A,** Space-filling model of the "B" form of DNA. This is the common form of DNA, with a helical turn every 10 base pairs. The major and minor grooves are visible. **B,** Stick model shows that DNA is composed of a sugar phosphate backbone with the bases ("A," "C," "G," and "T") pointing inward (blue and light brown). **C,** More compact "A" form of DNA with 11 bases pairs per helical turn. **D,** "Z" form of DNA is a left-handed helix. (Courtesy of Professor H Lodish.)

end of the RNA. This then allows the enzyme polyadenylate polymerase to attach a homo-polymeric string of A residues to the 3′ end of the transcript called the poly-A tail. Increasing recent evidence suggests that the processes of transcription initiation, elongation by RNA Pol II, and 3′ end processing may be co-regulated.

The introns are then spliced out to form the mature mRNA species (a simplified version of this process is presented in Fig. 1.6). Splicing is an elaborate process that involves a large number of steps catalyzed by a splicing complex (or spliceosome), which contains small nuclear RNAs (snRNAs) and proteins, and that produces small nuclear ribonuclear particles (snRNPs). It is estimated that over 100 proteins are involved in splicing. At first approximation, this process is probably as complex as regulation of transcriptional initiation and translation. One reason why a cell invests this degree of effort into splicing is that it allows a cell to generate multiple different mRNA species from a single gene, contributing to the biological complexity that an organism can achieve from a limited gene set. However, the genes encoding proteins controlling splicing can acquire mutations and the mutant proteins may cause aberrant splicing, leading to hematologic diseases. One important part of splicing is that the two nucleotides that lie in the intron and mark the boundary of an exon–intron are almost always invariant (Fig. 1.6). Thus, the 5′ end of the intron is usually marked by the dinucleotide "GU" whereas the 3′ end has "AG."

Like hnRNA, mRNA is wrapped in chaperone proteins to form mRNPs (messenger ribonuclear particles) that are exported from the nucleus through a water-impermeable phospholipid bilayer, the nuclear envelope that is studded with proteins and pores (Figs. 1.7 and 1.8). The nuclear pore complex (NPC) is a large structure (~125 million Da), about 30 times the size of the ribosome. It is made of multiple copies of a large number (~100)

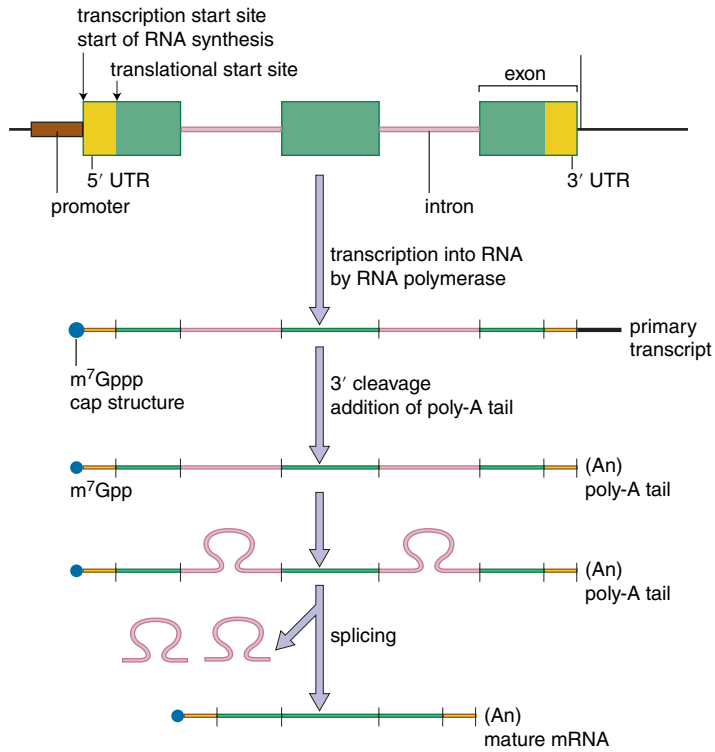


Fig. 1.5. Most genes encoding proteins are first transcribed into mature messenger RNA (mRNA) via multiple steps. *Top*, Genes are divided into exons (shown as boxes) separated by introns (shown as pink lines). Preceding the transcribed region is a promoter region (brown box) that helps regulate transcription timing and rate. A transcriptional initiation site marks the beginning of transcription. The beginning and end of the transcribed regions are usually not translated into protein and are known as the 5′ and 3′ untranslated regions (UTRs) (depicted as yellow boxes). Translated areas are shown as green boxes. The whole gene from the transcriptional start site is transcribed by RNA polymerase to make a primary transcript. This has a specialized cap structure at its 5′ end to protect the transcript from degradation. The 3′ end of the transcript is then cleaved and a tail of "A" nucleotide residues (known as a poly-A tail, "An") is added at the 3′ end of the transcript (to protect the end from degradation). Then the introns loop out (see later for details) and are spliced out to create the mRNA moiety.

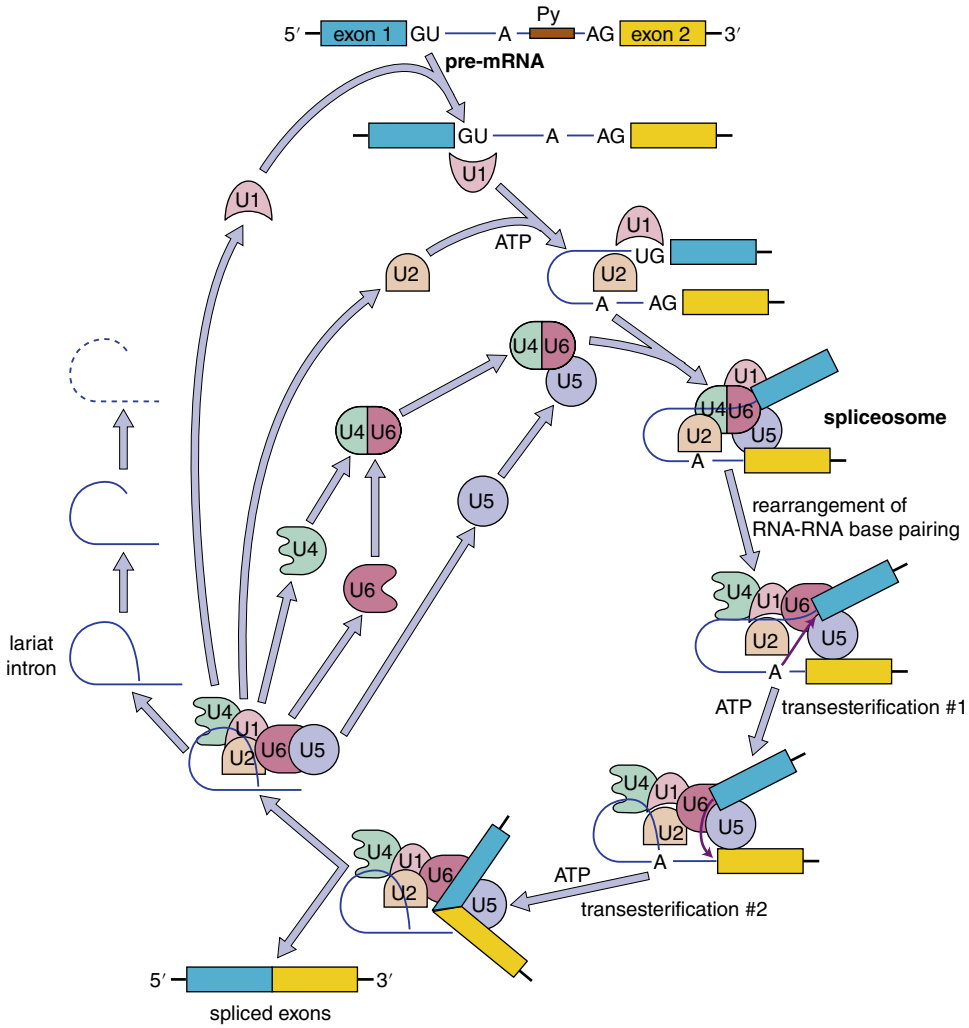


Fig. 1.6. Detail of splicing out of introns coordinated by splicing small nuclear ribonuclear protein particles (snRNPs) U1, U2, and U4–U6. U1 and U2 snRNPs associate with unspliced transcript in an ordered sequence at specific nucleotides ("GU") at the 5′ intron–exon boundary and a pyrimidine tract (Py) near an "A" nucleotide known as the branch point. U4–U6 then assemble, catalyzing an ATP-dependent rearrangement of RNA base-pairing structure. The snRNPs then catalyze two transesterification reactions that allow the exons to join. The intervening intron forms a lariat structure that is degraded. The snRNPs are recycled. Mutation in genes controlling splicing are recurrently detected in myeloid malignancies (Chapters 13 and 15).

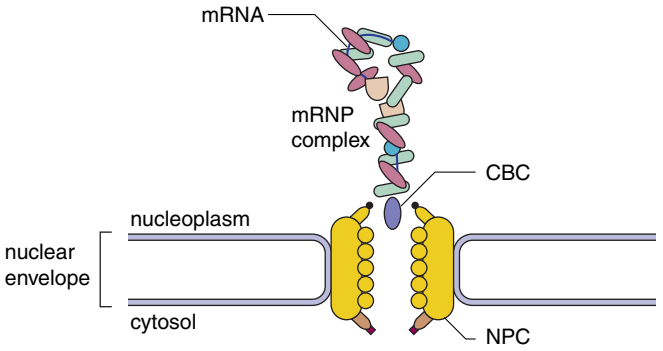


Fig. 1.7. Nascent primary transcripts and mRNAs are associated with nuclear proteins to form heterogeneous ribonuclear protein particles (hnRNPs). Some of these hnRNPs help transport mRNA out of the nucleus. The 5′ end of the mRNA–hnRNP complex (mRNP) associates with a cap-binding complex (CBC) that is exported through a specialized nuclear pore complex (NPC). Some of the hnRNPs remain in the nucleus and are recycled. The mRNA then interacts with cytosolic mRNP-binding proteins that escort the mRNP to ribosomes to be translated. mRNP export is an active, controlled, and coordinated process.

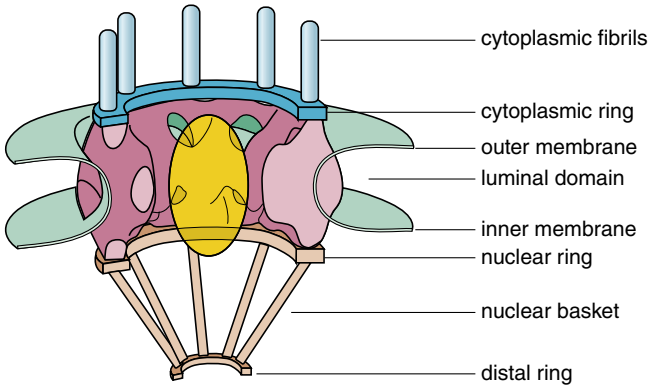


Fig. 1.8. Detailed schematic view of a eukaryotic nuclear pore complex. It is a highly ordered structure that is embedded in the nuclear and cytosolic membranes.

of proteins. It has a ring-basket structure. The ring points into the nucleus and filaments that form a basket. The structure is then embedded in the nuclear envelope. mRNPs are exported to the cytoplasm in a GTP-dependent process through the NPC. It is facilitated by a subset of RNP proteins that contain amino acid sequences functioning as nuclear export signals (NES). Similarly, proteins made in the cytosol have to be imported into the nucleus through NPCs. Such proteins often, but not always, have nuclear localization signal sequences (NLS).

Once in the cytoplasm, mRNA is covered by cytosolic proteins. The stability of mRNAs is variable and can be regulated. This can help determine the amount of mRNA available for protein synthesis.

Just as large macromolecular machines are required to make DNA and mRNA, proteins are translated from mRNA in a large structure called a ribosome. The details of translation are described in Fig. 1.9. The large and small ribosomal subunits, with the aid of specific translational initiation proteins (factors), locate the translational start site, which is usually the first codon (the three RNA nucleotides) for the amino acid methionine, in an ATP- and GTP-dependent process. As codons are composed of triplets of mRNA nucleotides, there are three different reading frames or ways in which mRNA triplets can be read by the ribosome. The frame that is selected is defined by the position of the start codon. Once engaged, the ribosome moves along the mRNA and sequentially adds amino acids to the growing peptide chain by recognizing sequential triplets of mRNA nucleotides (codons) (Fig. 1.9B). The amino acid added to the peptide is defined by the RNA codon selected and, as shown in Fig. 1.10, there is a code for how the different RNA codons specify particular amino acids. Of note, certain RNA codons specify "stop" signals (as well as a start signal, see above) that cause peptide chain termination. In eukaryotic cells (i.e. organisms with cells that have internal compartments, such as mammalian cells), multiple ribosomes commonly engage and concurrently translate a single mRNA to form a circular polysome to increase the efficiency of protein translation. As ribosomes finish translation at the 3′ end, the subunits quickly reassemble to reinitiate synthesis at the 5′ end of the mRNA.

Nascent peptide chains have to be properly folded, and amino acids modified and then either directed to the right cellular compartments or labeled for export. Proteins with a specific signal sequence direct the ribosome to the endoplasmic reticulum where protein synthesis is completed and peptides are directed to the Golgi complex and sorted for different destinations (the secretory pathway) (Figs. 1.11 and 1.12). In other cases, proteins complete synthesis in cytosolic ribosomes and are directed to other compartments (the nucleus, mitochondria, peroxisome). The transport of proteins depends on signal sequences (e.g. a nuclear localization signal) and interaction with specific receptor/transport proteins.

DNA MUTATIONS CAN ALTER PROTEIN SYNTHESIS BY A NUMBER OF MECHANISMS

Changes in a DNA sequence are called variants. Sometimes changing the DNA sequence can have deleterious consequences, in which case the variant is known as a mutation. DNA mutations occur at a variety of places in the gene locus and can cause aberrant mRNA and protein production (Fig. 1.13). For example, point nucleotide substitutions in the coding sequence (Fig. 1.13) can cause either an amino acid substitution (missense mutation) or introduce a stop codon (nonsense mutation). Deletions or additions of nucleotides (other than in multiples of three nucleotides) can cause an alteration in the reading frame (frameshift mutation). In addition to mutations in the coding sequence, mutations can also occur in the promoter (or other distal *cis*-regulatory elements) to alter transcription; in the invariant splice acceptor donor sites in the intron at the intron–exon boundary to affect splicing; or in sequences that control 3′ end processing (poly-A addition sites) and 5′ end processing (addition of the cap site) (Fig. 1.14). The whole spectrum of these mutations is exemplified in the germline in the β-globin gene in β-thalassemia (see Chapter 9) or, to a lesser extent, in acquired mutations in FMS-like tyrosine kinase 3 (*FLT3*) in acute myeloid leukemia (AML) (Fig. 1.15) (see Chapter 13).

TRANSCRIPTIONAL CONTROL OF GENE EXPRESSION

One major level of control of protein production is by regulating mRNA transcription. For any gene, expression is regulated by regulatory DNA sequences (*cis*-elements), proteins (transcription

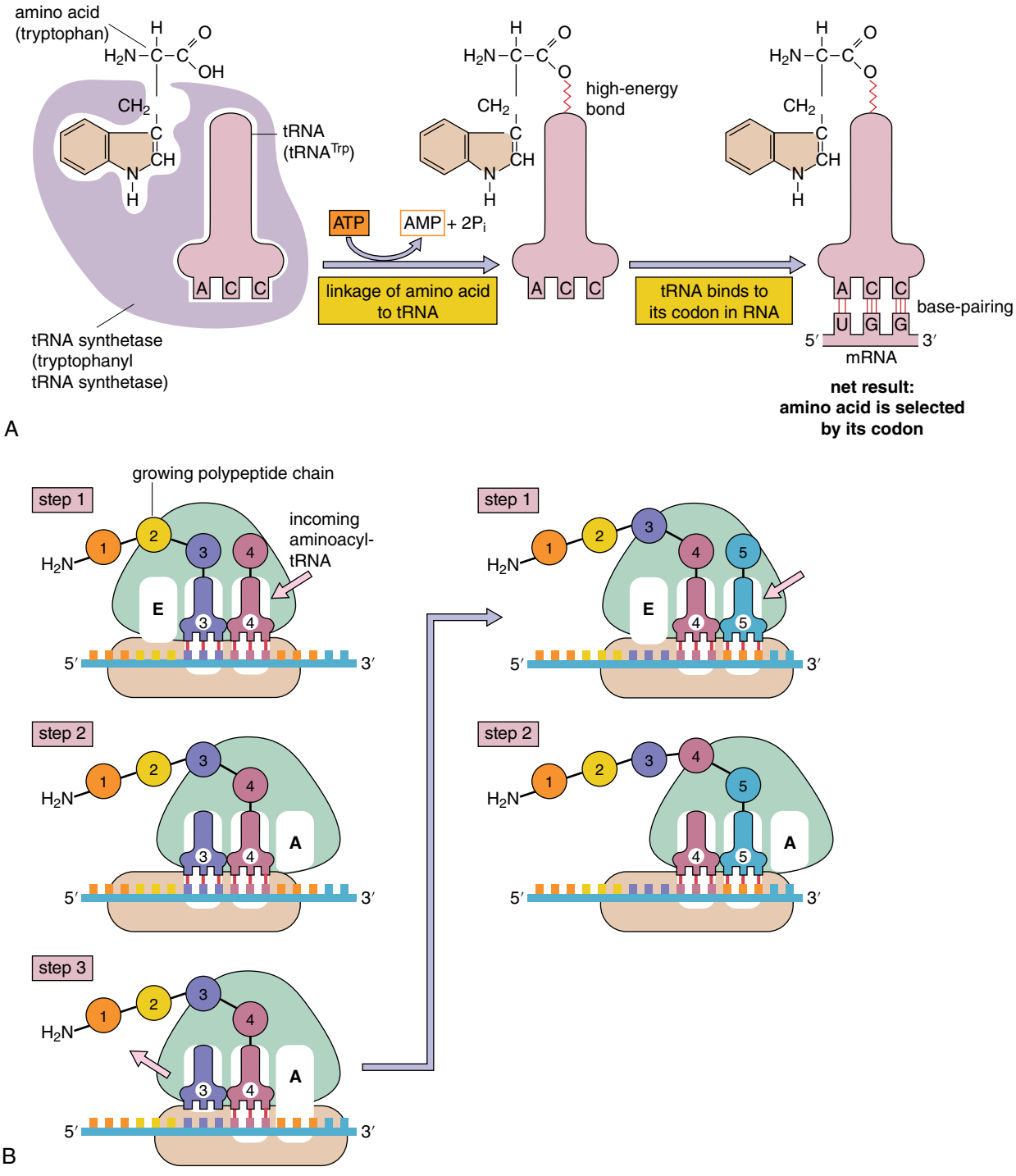


Fig. 1.9. Messenger RNA is translated into proteins. **A,** mRNAs docked in the ribosome interact with transfer RNAs (tRNAs) that bind amino acids. Different amino acids bind specific tRNA molecules. Here, the amino acid tryptophan is coupled to a specific $tRNA^{Trp}$ by an adapter aminoacyl-tRNA synthetase in an ATP-dependent reaction. The triple RNA nucleotide "ACC" sequence in the $tRNA^{Trp}$ (anticodon sequence) then base pairs to a triple RNA sequence "UGG" (codon) in mRNA. Thus the amino acid is selected by specific codon–anticodon recognition. **B,** To form an elongating protein polypeptide in a ribosome from mRNA, amino acids are added sequentially as specific tRNAs are recruited by codon–anticodon base-pairing. In this figure, in *step 1* on the left-hand side, amino acids 1, 2, and 3 have already been added. Amino acid is docked (via the specific tRNA) next to amino acid 3. In *step 2*, amino acid 4 is bound to amino acid 3 and a new tRNA-docking position ("A") is available for the next tRNA to dock. In *step 3*, mRNA moves along and is opposite position "A" and the tRNA that brought in amino acid 3 is ejected. These steps are then repeated on the right-hand side to allow docking of tRNA that brings in amino acid 5, which is added to the growing polypeptide chain.

factors and transcriptional cofactors) that regulate the transcription of the gene by binding either directly, or indirectly, to *cis*-element. Finally, as DNA is very tightly packed to accommodate it into the nucleus, *cis*-elements and the genes themselves have to be "unpacked" to allow the genes to be expressed. Regulation of gene expression through control of packing and unpacking of DNA is known as epigenetic regulation of gene expression. We discuss it in more detail later.

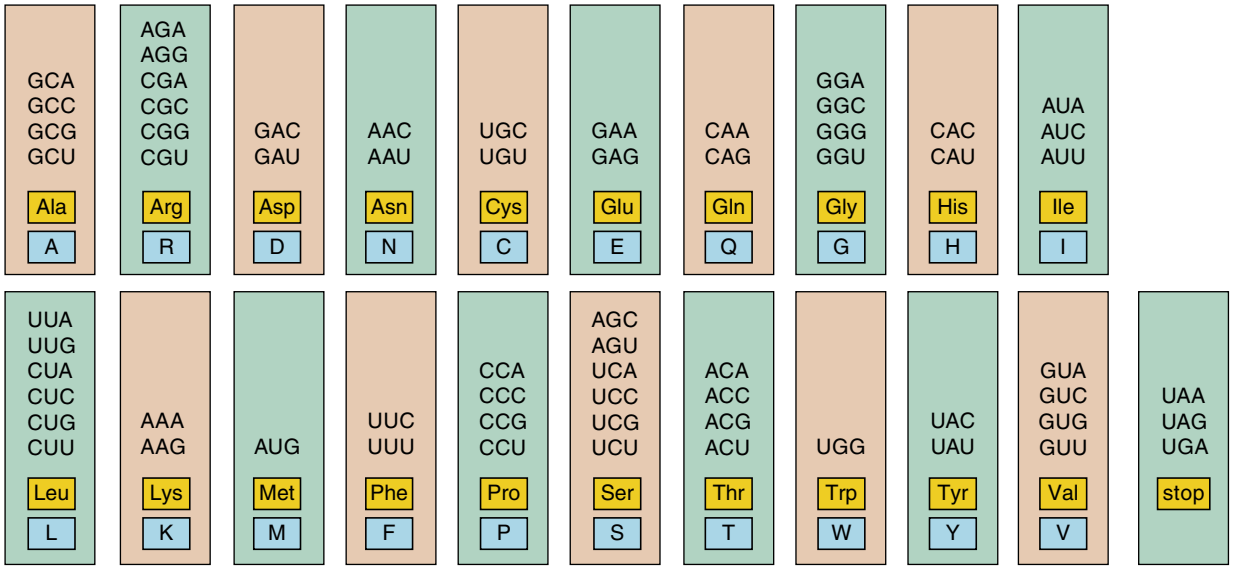


Fig. 1.10. Different RNA codons are used to code for amino acids. *Top left*, For example, the amino acid alanine (three-letter code Ala [yellow box] and one-letter code A [blue box]) is coded by the codons "GCA," "GCC," "GCG," and "GCU." The codons "UAA," "UAG," and "UGA" bring a halt to addition of amino acids and are known as stop codons.

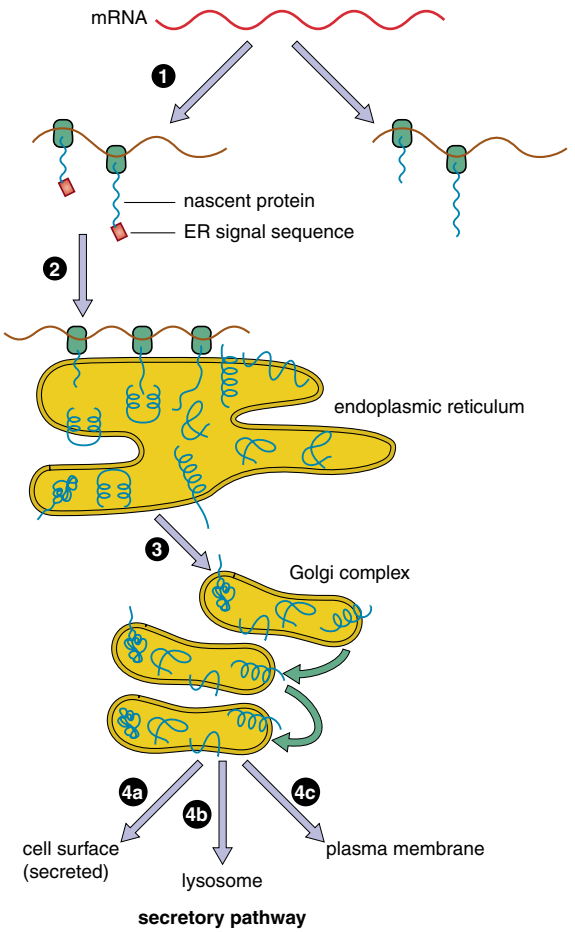


Fig. 1.11. Newly synthesized proteins are folded and post-translationally modified in the rough endoplasmic reticulum and the Golgi apparatus. *Step 1*, Ribosomes (green rectangles) synthesize polypeptide chains (blue lines) from an mRNA template (red line). Proteins with a signal sequence (pink square) are taken up by the rough endoplasmic reticulum, where translation is completed. Proteins without a signal sequence complete translation in the cytosol on free ribosomes. *Step 2*, In the rough endoplasmic reticulum the proteins are folded and post-translationally modified. *Step 3*, They are transferred to the Golgi via transport vesicles. *Step 4a,b,c*, The folded protein is then sorted for onward transport.

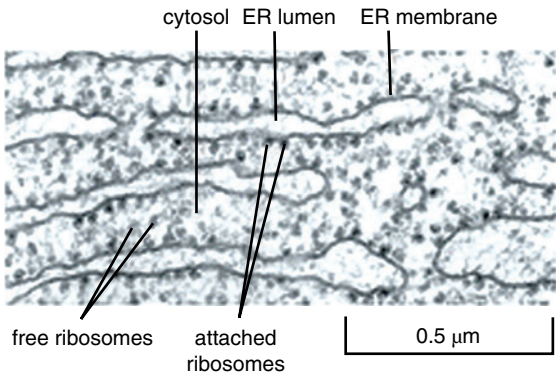


Fig. 1.12. Electron micrograph of ribosomes attached to the rough endoplasmic reticulum in a pancreatic cell.

CIS-ELEMENTS AND TRANSCRIPTION FACTORS

Figure 1.16 shows that *cis*-elements can be located near to the gene. For example, the promoter defines the location of transcription start site(s) and the directionality of transcription. The *cis*-elements can also be located at distance from the gene and can either promote (enhancer) or repress (silencer) transcription. The *cis*-elements are composed of multiple binding sites for transcription factors (TFs). Thus, the *cis*-element acts as a docking site for multiple DNA-binding TFs that in turn tether other TFs and cofactors that do not bind DNA.

Like many other proteins, TFs have a modular structure. For example, DNA-binding TFs such as MAX have a domain called the basic region (b) that binds DNA and an adjacent region, the helix-loop-helix (HLH) motif that interacts with other TFs and cofactors (Fig. 1.17). These domains allow TFs to be organized into families that share similar amino acid sequences and structural domains. Thus, MAX is part of the bHLH TF family. TFs in a common family often bind similar ~6–10 DNA base pair sequences (binding sites) and bind with similar protein partners.

Once TFs and cofactors bind *cis*-elements proximal and distal to the gene most, some but not all, evidence suggests that the *cis*-elements come together by looping out intervening DNA

(Figs. 1.18 and 1.19). Either before or after looping, the complex of TFs and cofactors recruits RNA Pol II containing preinitiation complex. RNA Pol II is then phosphorylated on a specific domain (the C-terminal domain) that allows RNA Pol II to disengage from the preinitiation complex and elongate the RNA chain as it proceeds along the gene.

Study of hematopoietic genes has contributed significantly to our understanding of eukaryotic gene expression. Examples include the α and β globin gene loci (Fig. 1.20). In the β-globin locus, multiple closely spaced distal *cis*-elements combine to form a specific type of enhancer known as a locus control region (LCR). TFs and cofactors are bound at the LCR and even more distal *cis*-elements are thought to bind together by looping out intervening DNA and promote transcription of different globin genes in a developmental stage-specific manner.

CHROMATIN AND EPIGENETIC CONTROL OF GENE EXPRESSION

DNA is highly packaged in a nucleus. For a TF to gain access to a short DNA sequence in a particular *cis*-element of an individual gene, the chromatin associated with that sequence has to be specifically unpacked (Figs. 1.21–1.23). Metaphase chromosomes are progressively unpackaged via intermediate states (Fig. 1.22) called chromosome territories that are poorly defined to transcriptional activation domains (see next paragraph) to a 30 nm chromatin fiber. These higher order structures are gradually being understood, as new, remarkable techniques, such as high-resolution electron microscopy and genome-wide molecular mapping

Wild-type sequences

amino acid	N-Phe	Arg	Trp	Ile	Ala	Asn-C
mRNA	5′-UUU	CGA	UGG	AUA	GCC	AAU-3′
DNA	3′-AAA	GCT	ACC	TAT	CGG	TTA-5′
	5′-TTT	CGA	TGG	ATA	GCC	AAT-3′

Missense

3′-AAT	GCT	ACC	TAT	CGG	TTA-5′
5′-TTA	CGA	TGG	ATA	GCC	AAT-3′
N-Leu	Arg	Trp	Ile	Ala	Asn-C

Nonsense

3′-AAA	GCT	ATC	TAT	CGG	TTA-5′
5′-TTT	CGA	TAG	ATA	GCC	AAT-3′
N-Phe	Arg	Stop			

Frameshift by addition

3′-AAA	GCT	ACC	ATA	TCG	GTT A-5′
5′-TTT	CGA	TGG	TAT	AGC	CAA T-3′
N-Phe	Arg	Trp	Tyr	Ser	Gln

Frameshift by deletion

GCTA				
CGAT				
3′-AAA	CGT	ATC	GGT	TA-5′
5′-TTT	GGA	TAG	CCA	AT-3′
N-Phe	Gly	Stop		

Fig. 1.13. Changes in coding parts of the DNA sequence of a gene can alter the protein produced. In this example, the wild-type (normal) protein and corresponding mRNA and DNA sequences are shown at the top. Point mutations that change the amino acid encoded in the protein are known as missense mutations. Here a "T" to "A" change in the DNA (reading 5′ to 3′) alters the protein sequence from phenylalanine (Phe) to leucine (Leu). On occasion the amino acid change can alter protein function. *Below*, The point change ("C" to "A") introduces a stop codon. This is a nonsense mutation. *Below*, If nucleotide/nucleotides is/are added (the nucleotide "T") or deleted ("CGAT"), this alters the reading frame (the order in which the triplets of nucleotides are read as codons) and alters protein sequence. This is a frameshift mutation. Nonsense and frameshift mutations usually have a more profound effect on protein sequence.

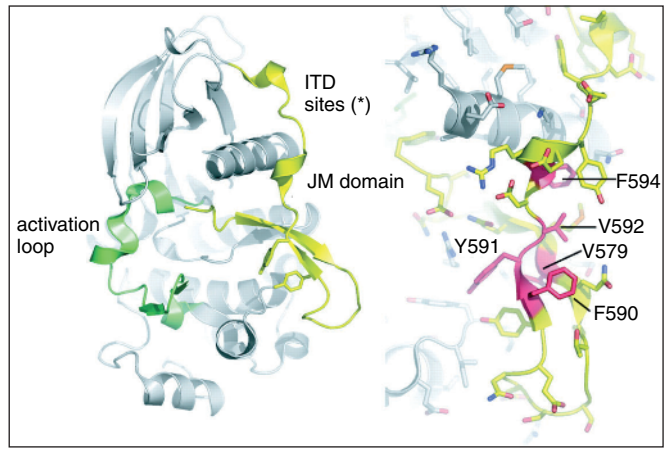


Fig. 1.15. Ribbon model of the crystal structure of the FMS-like tyrosine kinase 3 (FLT3) domain (with green activation loop and yellow juxtamembrane [JM] domain). The positions of internal tandem duplications (ITDs) in the JM domain, leading to FLT3 activation, are indicated. *Right*, Close view of the mutation sites in the JM domain (yellow). The structure is shown as a ribbon backbone, with side-chains as colored sticks.

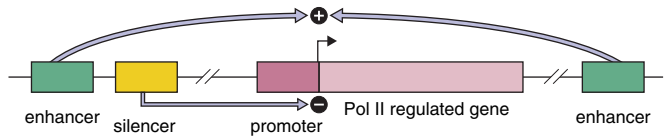


Fig. 1.16. Regulation of genes transcribed by RNA polymerase II is controlled by multiple DNA sequences (*cis*-elements). Here, the gene is regulated by proximal sequences—a promoter that abuts the transcriptional initiation site (arrow) and distal sequences that function to both enhance (enhancers) and repress (silencers) gene expression.

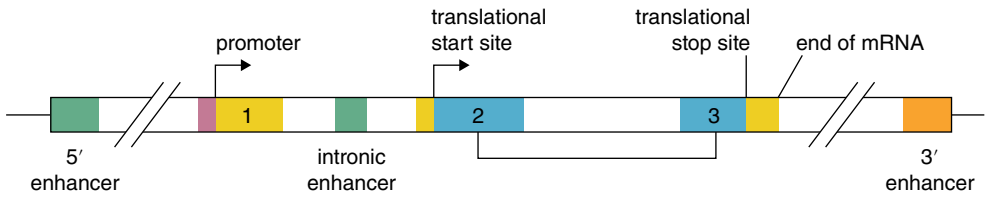


Fig. 1.14. Schematic representation of a gene locus. On the left is the 5′ end of the locus and an enhancer is shown as a green box, the promoter as a pink box, and the transcriptional start site as a arrow. This gene has three exons and the exons are shown as boxes. Noncoding portions of exon are in yellow and coding portions of exons are in blue. Exon 2 and 3 have both noncoding and coding regions. The translational start site is indicated. The gene also has a 3′ enhancer shown in orange.

of long-range chromatin interactions, are deployed to visualize these structures. The 30 nm chromatin fiber is composed of fibrils of DNA wrapped around nucleosomes (composed of histone octamer – two units each of H2A, H2B, H3, and H4) and finally to naked DNA. Histone octamers can include variant histones, of which there are a number, for example, H2A.Bbd, H2A.X, H3.1, H3.2, H3.3, and H3.X. H3.Y, that serve diverse roles in DNA replication and transcription of DNA into RNA.

TADs describe an intermediate level of chromatin packing where tens of kilobases to megabases, DNA, and chromatin

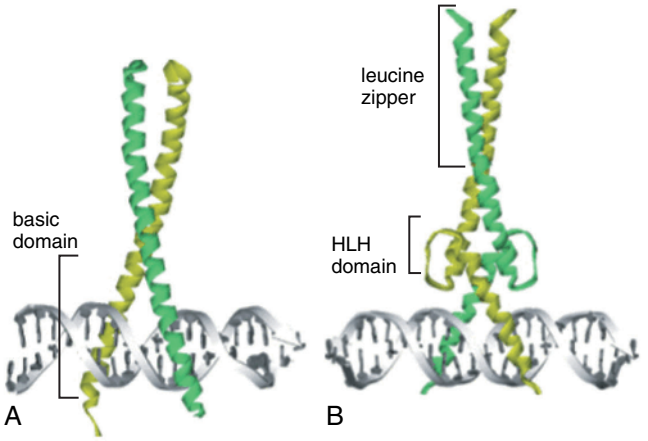


Fig. 1.17. A, Crystal structure of two transcription factors (green shades) shows the intimate contact between the basic residues (basic domain) and specific sequences in the major groove of DNA (white double helix). The two TFs then interact with each other via helical structures that are rich in leucine residues (leucine zipper domains). Thus these TFs contain bzip domains. **B,** Crystal structure of two TFs (green shades) that bind the major groove of DNA (white shade) via basic residues (basic domain). Immediately following this domain are helix-loop-helix domains (HLH) followed by leucine zipper regions. Therefore these TFs have bHLH domains with leucine zippers.

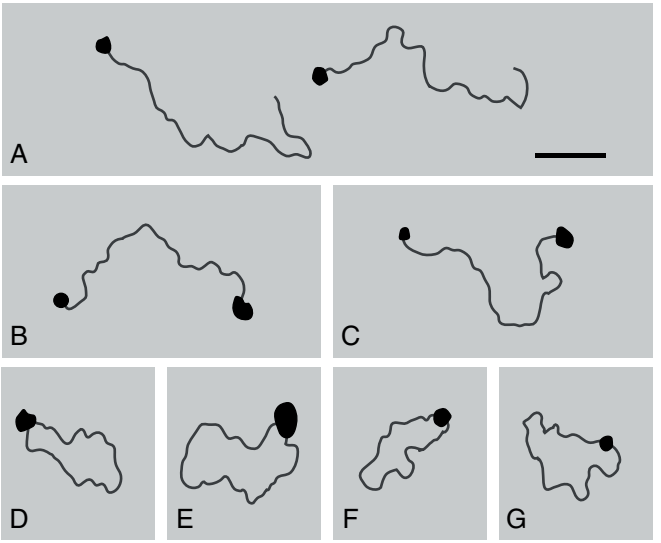


Fig. 1.19. Looping between *cis*-elements can be visualized. **A,** Schematic representation of an electron micrograph of the transcription factor Sp1 (shown as a black spot) bound to a cognate binding site at one end of DNA (irregular line). **B** and **C,** Two cognate DNA-binding sites were engineered at either end of the DNA fragment. **D–G,** Over time, Sp1 bound to the ends of the DNA self-associated and looped the intervening DNA out.

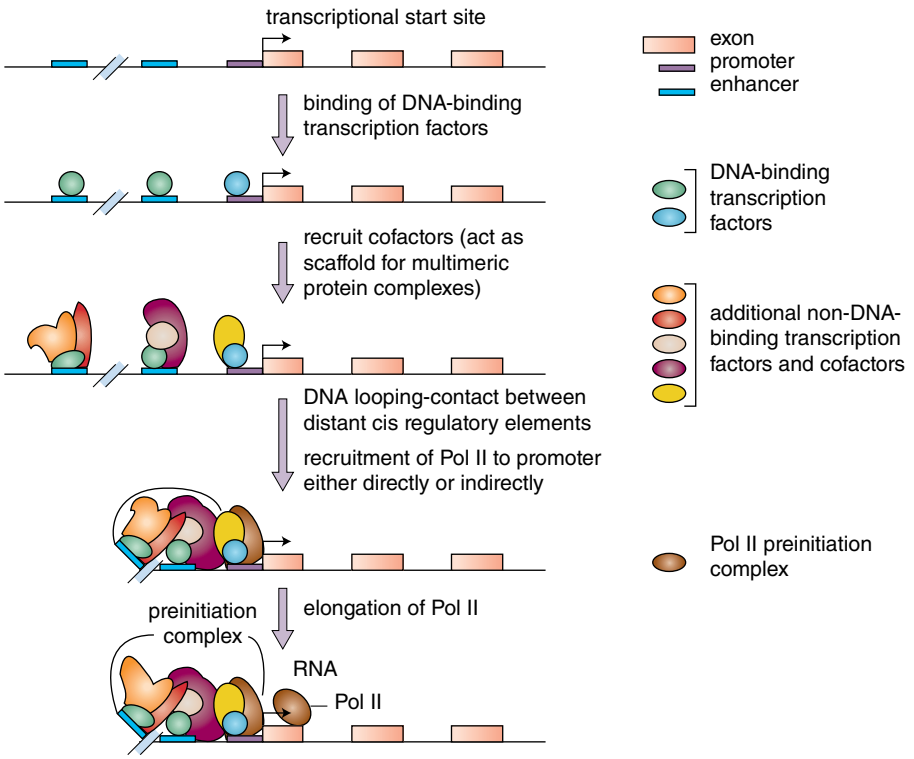


Fig. 1.18. Gene expression is controlled by transcription factors that bind *cis*-elements. DNA-binding TFs binding to *cis*-elements then recruit cofactors and other transcriptional regulators. It is also likely that DNA-binding TFs and non-DNA-binding TFs/cofactors may form preformed complexes that bind directly to DNA. Binding of RNA polymerase II, the preinitiation complex, and DNA looping between different *cis*-elements triggers transcription and elongation of polymerase II along the gene.

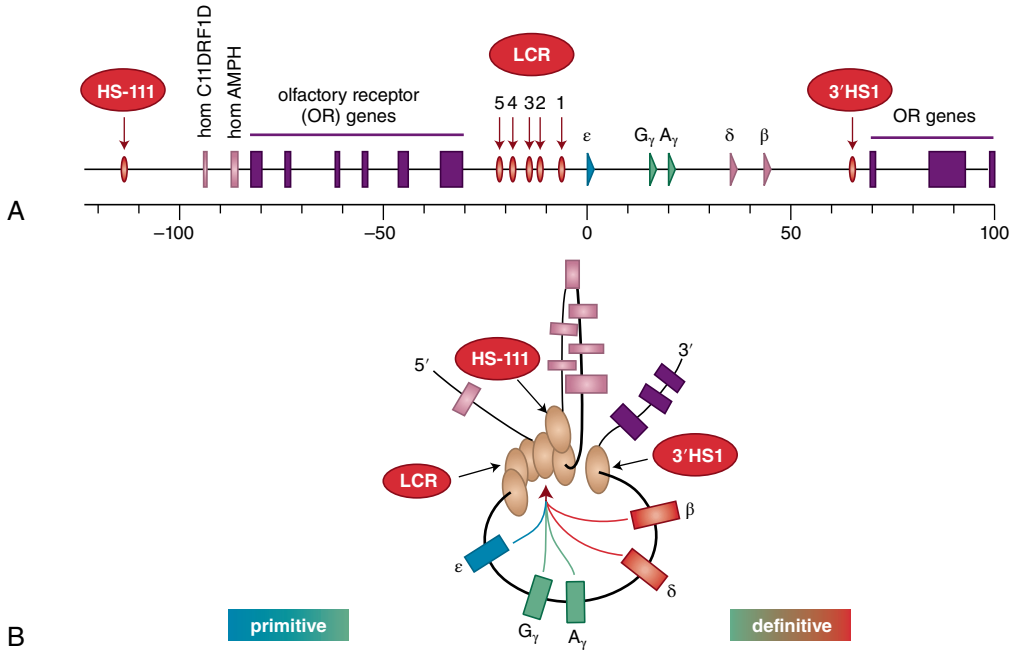
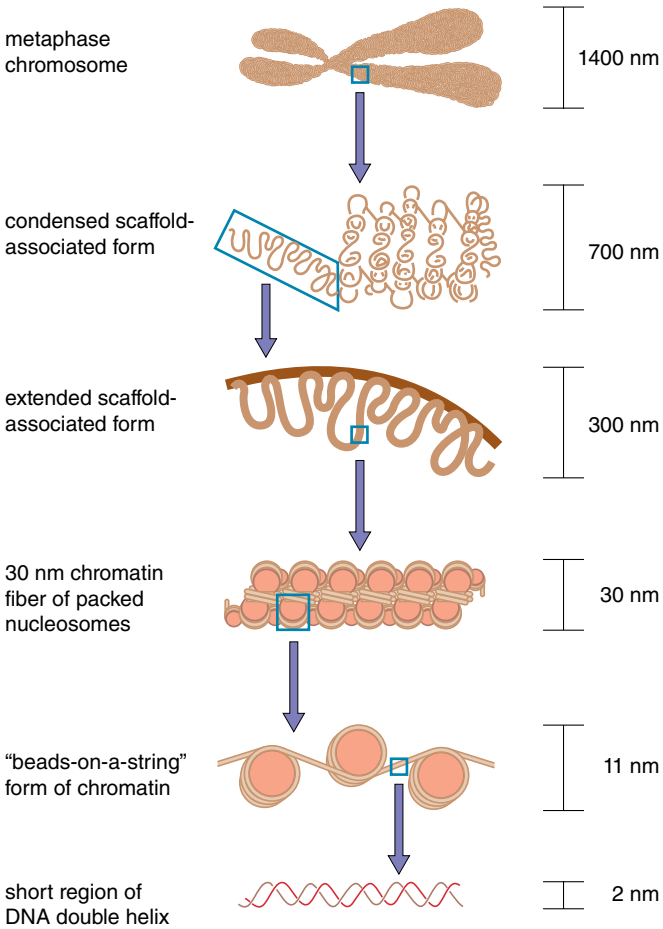


Fig. 1.20. The *cis*-elements that regulate gene expression can be distributed over a large area and can be complex. One of the best-studied gene loci is the human β-globin gene cluster on chromosome 11. **A,** There are five genes in the β-globin gene cluster (ε, Gγ, Aγ, δ, and β, depicted as arrowheads). At the 5′ of the gene cluster there are five *cis*-elements (numbered 1 to 5, marked by arrows and collectively termed the locus control region, LCR). Two more *cis*-elements are located 111 kilobases 5′ (HS-111) and approximately 65 kilobases 3′ (3′HS1). The whole β-globin domain (genes and *cis*-elements) is embedded in the midst of a bank of olfactory receptor (OR) genes (purple blocks) and two other genes (pink blocks). The scale below the locus is in kilobases. **B,** In a nucleus, current evidence supports a model where all the β-globin cluster *cis*-elements physically interact (even though widely dispersed) with each other and with a β-like globin gene. Embryonic ε globin is expressed first (in primitive red cells), then the fetal γ genes, and finally the adult δ- and β-globin genes (both in definitive red cells).



proteins are organized into spatial domains (Fig. 1.22C,D). Within each TAD there usually lie a number of genes and the DNA sequences that regulate those genes. The whole point of a TAD is that it segments chromatin such that all the DNA sequences capable of regulating genes usually function within the TAD they are located in (Fig. 1.22D–E). Thus, TAD boundaries help delimit the extent over which DNA sequences that regulate gene expression can function. TAD boundaries are marked by proteins called cohesins, and a protein called CTCF (CCCTC-binding factor). Thus, abnormalities in cohesin and CTCF have the potential to disrupt TAD boundaries and alter gene expression. Importantly, mutations in cohesins and CTCF are recurrently present in hematologic malignancies, especially myeloid blood cancers (AML and myelodysplastic syndromes).

Regulation of the selective packing or unpacking of chromatin affords another level at which control on gene expression can be exerted. Control of expression of specific

Fig. 1.21. Genes do not exist as naked DNA in the nucleus but are highly packaged. This figure shows that when metaphase chromosomes are unpackaged they are composed of fibrils of chromatin and may be associated with nuclear structures (one example that has been suggested is nuclear scaffolds that are often attached to nuclear membranes). When further unpacked, these chromatin fibrils are composed of 30 nm chromatin fibers, which in turn are composed of DNA wrapped around histone protein octamers called nucleosomes. When the 30 nm fiber is unwrapped the nucleosomes are linked by intervening DNA (like beads on a string). Nucleosomes can be temporarily shifted (remodeled) to expose naked DNA. The exact physical structure of higher order chromatin structure (30 nm fiber and higher orders of packaging) is unclear. Regulating the wrapping and unwrapping of DNA also affords a layer of regulation on controlling which genes are expressed (unwrapped) and which are not (wrapped).

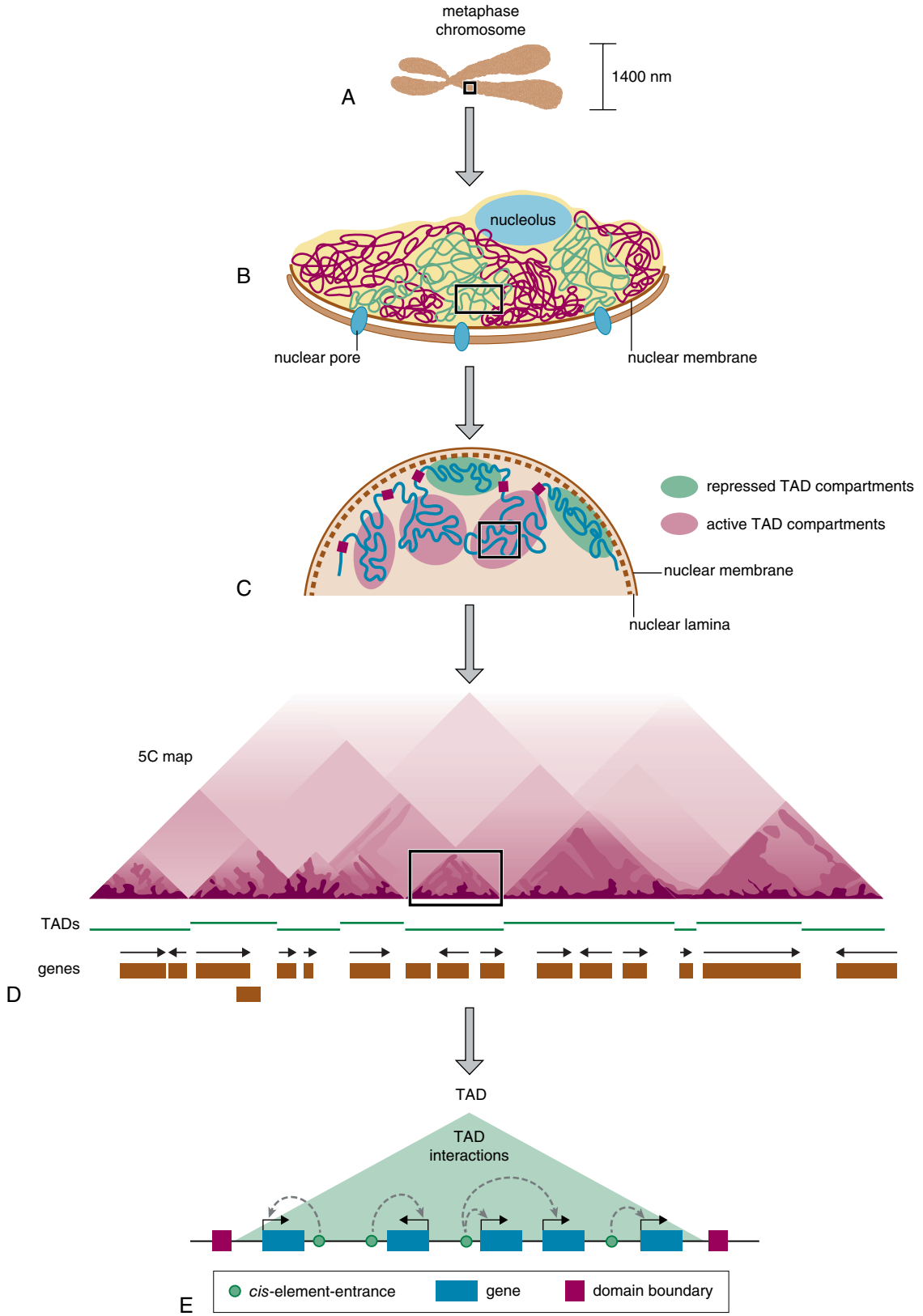


Fig. 1.22. Schematic representation of how DNA-encoding genes and the chromatin that coats genetic loci are packaged and unpackaged to allow selective gene expression. **A,** A metaphase chromosome and portion of the chromosome is examined in more detail in part B. **B,** Portion of chromosome composed of poorly defined chromatin fibers (irregular red and green lines that extend from the nuclear membrane into the nucleus) arranged into a number of chromosomal territories. The territories are shown in red and green. Genes encoding ribosomal RNAs are located in the nucleolus. Nuclear pores allow communication to the cytoplasm. **C,** Portion of chromosomal territory composed of a number of chromatin fibers that are arranged into topologically associated domains (TADs). Some are actively transcribed (pink) and some are repressed (green). Boundaries between TADs are shown as red squares. **D,** A number of TADs shown in greater detail by a technique called five-dimensional chromatin conformation capture (5C). This technique maps physical interactions within TADs and between TADs (shown as triangles). The stronger the interaction the darker the shade of color of the triangle. *Below*, Nine TADs of varying lengths are shown as lines. Under each TAD are a number of genes. Each gene is composed of exons (small vertical lines) and introns (space between lines with arrows). The arrows indicate the direction in which the gene is transcribed. **E,** Closer view of a TAD. The TAD boundaries are shown as red squares. There are five genes within the TAD (blue boxes). The arrows at the ends of each gene indicate the direction of transcription (5′ to 3′). The green circles indicate *cis*-elements that control transcriptional expression of the gene. Note, one *cis*-element may control expression of more than one gene.

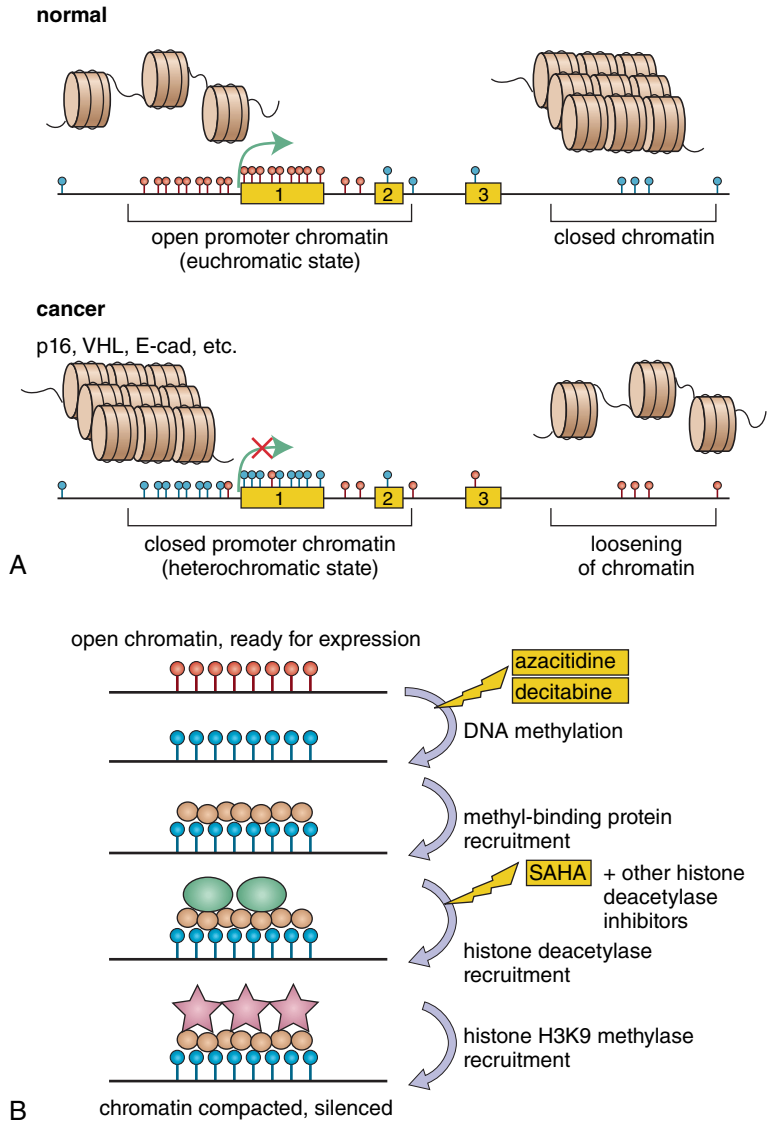


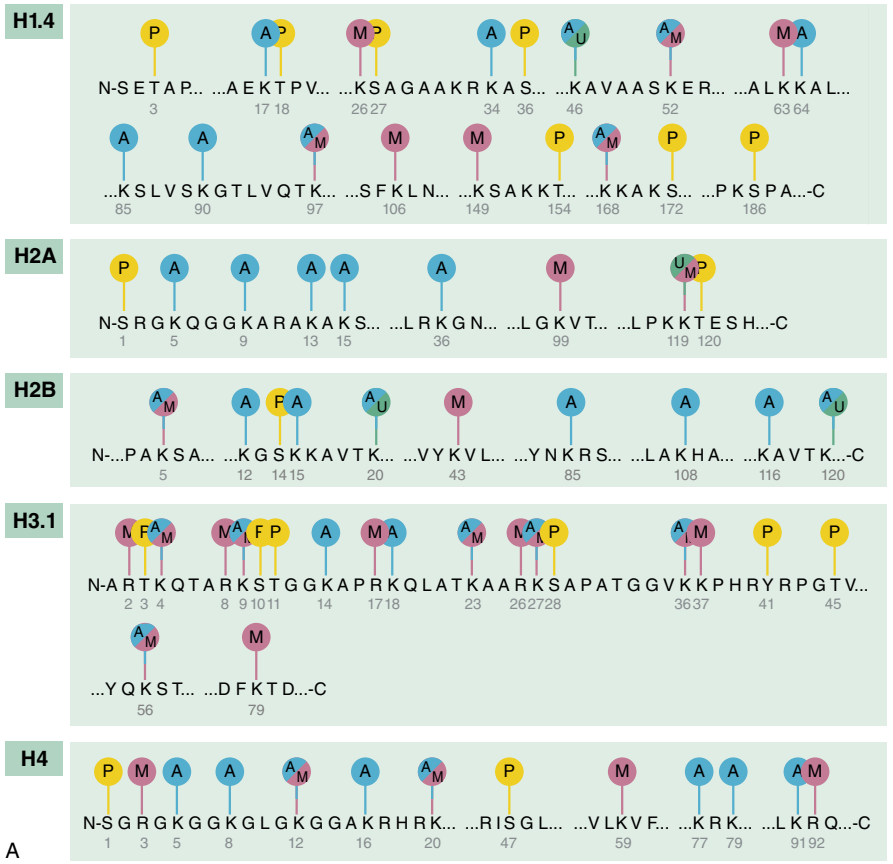
Fig. 1.23. A, Genes are in yellow boxes. Unwrapped genes (1) are in an open chromatin, euchromatic state. These genes have unmethylated CpG DNA residues (shown as pink lollipops) and are transcribed (green arrow). Genes that are not transcribed (2 and 3) often have methylated CpG residues (blue lollipops). Closed chromatin is shown as closely packaged nucleosomes. In cancer, gene expression is often inappropriate, and this is reflected in the changed chromatin structure. Thus genes such as p16, VHL, and E-cadherin (E-cad) are not transcribed. CpG residues are methylated and the chromatin is in a closed conformation. **B,** The transition from an open to a closed chromatin state occurs via multiple steps that can be targeted by drugs. *Top*, A gene locus is in an open chromatin conformation and the dinucleotide CpGs are not methylated (pink lollipops). DNA can then be methylated (blue lollipops). This can be inhibited by DNA methyltransferase inhibitors that are used in clinical practice (azacitidine and decitabine). Methyl CpG-binding protein (brown circles) is recruited to methylated DNA and this is thought to facilitate binding of histone deacetylases and histone H3K9 and H3K27 methylases. Acetylated histones are associated with open chromatin and facilitate transcription, whereas deacetylated histones and histone H3 methylated at residue lysine (K) 9 and K27 are associated with closed chromatin and transcriptional repression. A number of drugs in clinical use inhibit histone deacetylases. These include sodium valproate, SAHA, and MGCD0103.

genes by regulation of the state of chromatin is not encoded in DNA of the gene and thus is termed epigenetic regulation of gene expression. Huge advances have been made in our understanding of the epigenetic regulation and chromatin structure and its influence on gene expression both in normal cells and to a lesser extent in diseases such as some of the hematologic malignancies (Fig. 1.23).

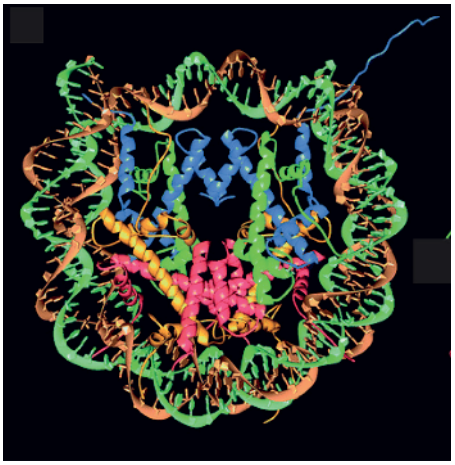
Regulation of the selective packing or unpacking of chromatin affords another level at which control on gene expression can be exerted. Control of expression of specific genes by regulation of the state of chromatin is not encoded in DNA of the gene and thus is termed epigenetic regulation of gene expression. Huge advances have been made in our understanding of the epigenetic regulation and chromatin structure and its influence on gene expression both in normal cells and to a lesser extent in diseases such as some of the hematologic malignancies (Fig. 1.23). This is exemplified in Fig. 1.23A, where the *cis*-elements that regulate expression of a key gene are shown in either an open chromatin conformation, allowing access to TFs and expression of the gene, or in a closed chromatin conformation (where gene expression is repressed). The normal chromatin state at key genes that control cell fate (e.g. growth, self-renewal, and differentiation) can be altered in a pathogenic manner in cancer.

This increased knowledge has led to the development of a new class of therapies for hematologic disease. To understand how these drugs may work in outline, it is helpful to consider epigenetic regulation of gene expression in a little more detail.

Control of packing is principally mediated by histones and methylation of DNA at the dinucleotide CpG (Fig. 1.23A,B). Histones can be post-translationally modified (acetylated, phosphorylated, methylated, or ubiquinated) at multiple residues by a large number of enzymes in a complex manner (Fig. 1.24). The complex and varying nature of the post-translational histone modifications (or histone marks) for any segment of chromatin is known as the "histone code" for that segment. Some of the histone marks cause the chromatin to be less tightly packed and activate gene expression. Consequently they are known as activating marks. Examples of activating marks include acetylation of histone H3 and H4, such as trimethylation of histone H3 at lysine residue 36 (me3H3K36) and H3 at lysine residue 4 (me3H3K4). Conversely, some marks are known as repressive marks as they cause the chromatin to become more tightly packed, repressing gene expression. Examples of repressive marks include trimethylation of histone H3 at lysine residue 9 (me3H3K9) and trimethylation of histone H3 at lysine residue 27 (me3H3K27). Both activating and repressive marks are laid down on chromatin regionally



A



B

Fig. 1.24. A, All histones are subject to post-transcriptional modifications, which mainly occur in histone tails. The amino acid sequence (as a single-letter code) is shown from histones H1.4, H2A, H2B, H3.1, and H4. The numbers below the single-letter code refer to the position of the amino acid. The main post-transcriptional modifications are depicted in this figure: A, acetylation (blue), M, methylation (red), P, phosphorylation (yellow), and U, ubiquitination (green). Some amino acid can be subject to more than one type of modification. Source: Portela A, Esteller M. *Nat Biotechnol* 2010;28:1057–1068. Reproduced with permission of Springer Nature. **B,** The view is down the DNA superhelix axis of a nucleosome core particle showing ribbon traces for the 146 base pair DNA phosphodiester backbones (green and brown) and eight histone protein main chains (blue: H3; green: H4; yellow: H2A; red: H2B. Source: Luger K, et al. *Nature* 1997;389:251–260. Reproduced with permission of Springer Nature.

in a very precise manner. The activating mark me3H3K4 is only seen at promoters, whereas me3H3K36 is present at enhancers. Similarly, the repressive mark me3H3K27 is seen at heterochromatin. Regional localization of histone marks allows identification of promoters and enhancers.

The histone-modifying enzymes are called "writers" as they "write" or impart a series of post-translational modification to histones. The histone marks imparted by writers allow chromatin to interact with a large number of chromatin-associated proteins called "readers." Finally, histone marks can be removed by other enzymes and these are known as "erasers." Reader, writer, and eraser proteins have modules (or domains) that "write," "read," and "erase" the histone code. All these proteins are grouped into families based on the modules that "write" or "read" or "erase" (Fig. 1.25). Recurrent mutations in the genes encoding these histone-modifying proteins are often seen in blood diseases.

Similarly, CpG dinucleotides can be methylated by DNA methyltransferases (DNMT) (Fig. 1.26). There are two broad classes of DNMTs: DNMT3A and DNMT3B are de novo methylases whereas DNMT1 maintains methylated DNA through DNA replication. Demethylation of DNA occurs by a very complex set of enzymatic reactions as shown in Fig. 1.26. These precise details of methylation are important as the proteins involved in DNA methylation and demethylation are often mutated in hematologic malignancies, pointing to the importance of this process for normal hematopoiesis.

Methylation of DNA is associated with more highly packaged DNA and repression of gene expression. Methylation of CpG residues promotes binding of methyl CpG-binding proteins (MeCPs). This in turn facilitates binding of writers and readers of repressive histone marks such as histone deacetylases (Fig. 1.23). Drugs that inhibit DNA methyltransferases (e.g. azacitidine and decitabine) and writers of repressive (e.g. vorinostat or pabinostat) would potentially reverse the repression of genes, especially those that promote differentiation of malignant cells (Fig. 1.23).

TRANSCRIPTION FACTORS, CONTROL OF GENE EXPRESSION, AND LINEAGE COMMITMENT

A cardinal event in the differentiation of a lineage-specific cell is the elaboration of a lineage-specific program of gene expression (Fig. 1.27). Expression of lineage-specific genes, as of all genes, is dependent on *cis*-elements and TFs (see earlier). These TFs can either be widely expressed or have a restricted pattern of expression. For example, there is a small of subset of TFs that are principally or exclusively expressed in blood cells. It is the action of these hematopoietic TFs that are critical in directing hematopoietic-specific gene expression. One such hematopoietic TF is GATA1. GATA1 is expressed in erythroid cells and megakaryocytes as well as eosinophils and mast cells. In all these cell

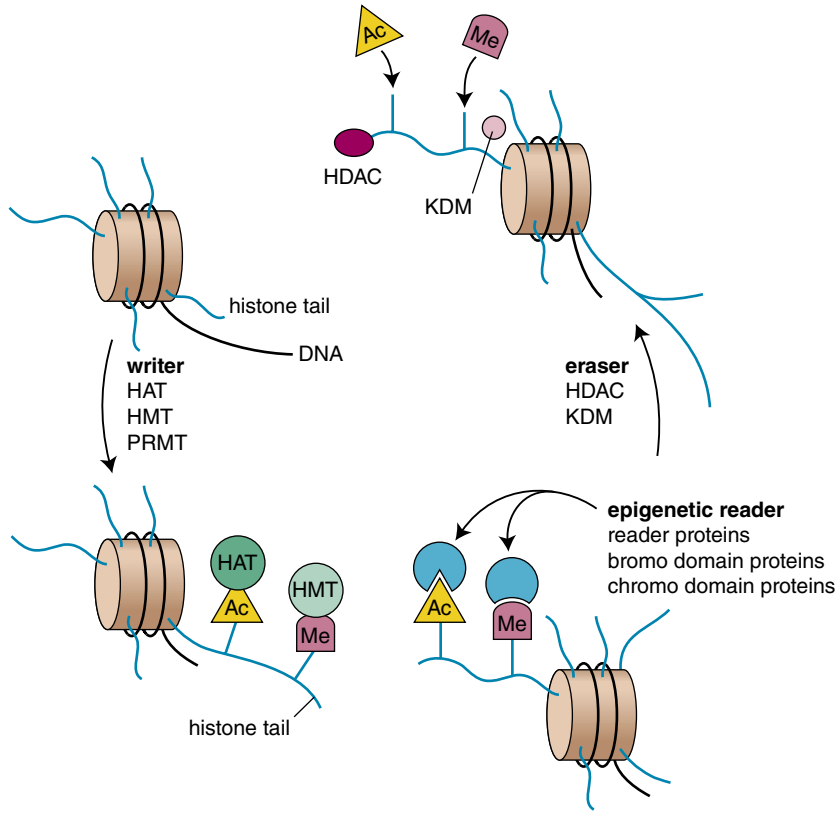


Fig. 1.25. Epigenetic regulation is a dynamic process. Proteins called epigenetic writers, such as histone acetyltransferases (HATs), histone methyltransferases (HMTs), and protein arginine methyltransferases (PRMTs), modify histones with post-translational modifications (known as epigenetic marks) on amino acid residues on histone tails. These post-translational modifications are recognized by proteins called epigenetic readers. They have amino acide domains called bromo domains and chromo domains that bind to these epigenetic marks. The post-translational modifications on histones are removed by proteins called epigenetic erasers, such as histone deacetylases (HDACs) and lysine demethylases (KDMs), that catalyze the removal of epigenetic marks. Addition and removal of these post-translational modifications of histone tails leads to the addition and/or removal of other marks in a highly complicated histone code. Together, histone modifications regulate various DNA-dependent processes, including transcription, DNA replication, and DNA repair. Source: Katrina J, et al. *Nat Rev Drug Discov* 2014;13:673–691. Reproduced with permission of Springer Nature.

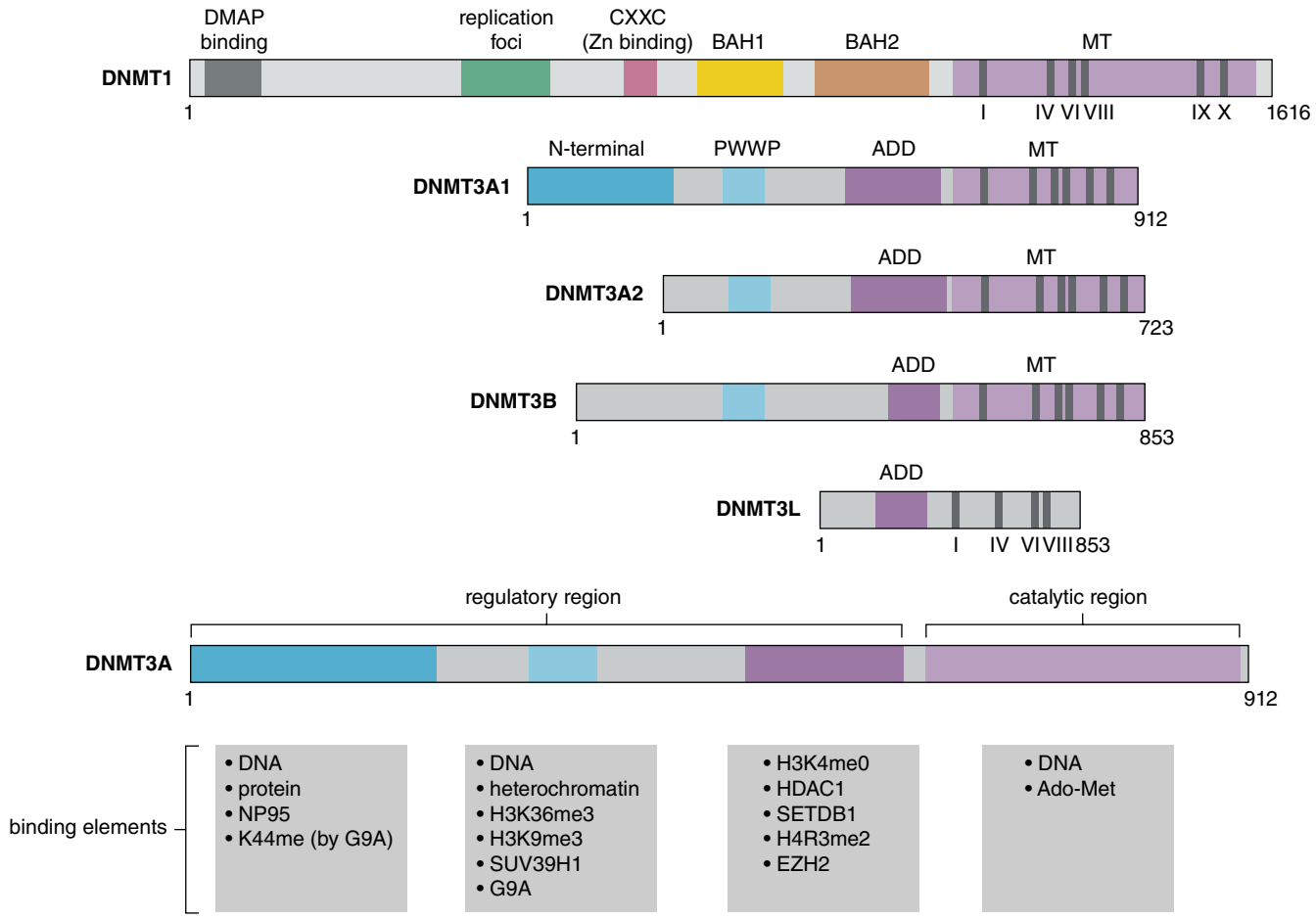


Fig. 1.26. DNA methyltransferase 1 (DNMT1), DNMT3A1, DNMT3A2, DNMT3B, DNMT3-like (DNMT3L), and major DNMT3A splice isoforms are depicted as boxes. Protein length is indicated as number of amino acids at the end of each protein. Each DNMT3A protein isoform has a modular structure, being composed of a mixture of protein domains. Each domain performs a different function. Domain abbreviations: ADD, ATRX-DNMT3-DNMT3L (related to the plant homology [PHD]-like domain of regulator ATRX); BAH, bromo adjacent homology domain; DMAP, DNMT1-associated protein; PWWP, Pro-Trp-Trp-Pro. MT is the catalytic methyltransferase domain, and I, IV, VI, IX, and X are motifs in the catalytic domain: motif I allows the binding of the methyl group donor AdoMet (*S*-adenosyl methionine). Motifs I and X are for cofactor binding and motifs VIII and IX are for DNA binding. The catalysis of DNA methylation occurs at the IV, VI, and VIII motifs. Source: Yang L, et al. *Nat Rev Cancer* 2015;15:152–165. Reproduced with permission of Springer Nature.

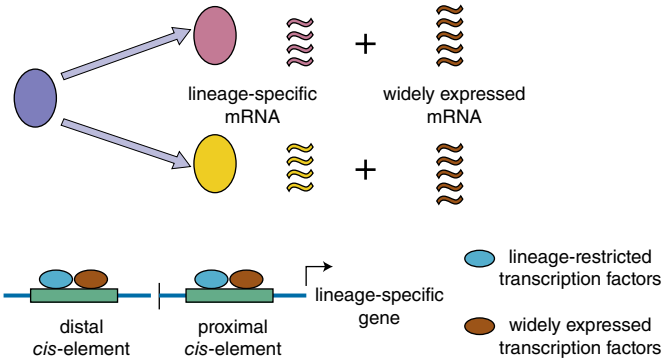


Fig. 1.27. A cardinal event in lineage specification is the expression of a program of lineage-specific gene expression. A multipotential cell (blue) differentiates into cells of two different lineages (pink and yellow). Although the phenotype of any cell, including cells of a specific lineage, is the sum of the lineage-specific and widely expressed RNAs, the specific phenotype of a lineage is a function of the lineage-specific RNAs. *Below*, Expression of a lineage-specific gene is controlled by DNA sequences (*cis*-elements) and widely expressed and lineage-restricted transcription factors. Thus, ultimately, one important element in regulating lineage specification is the complement of lineage-restricted TFs that are expressed and the transcriptional networks they control.

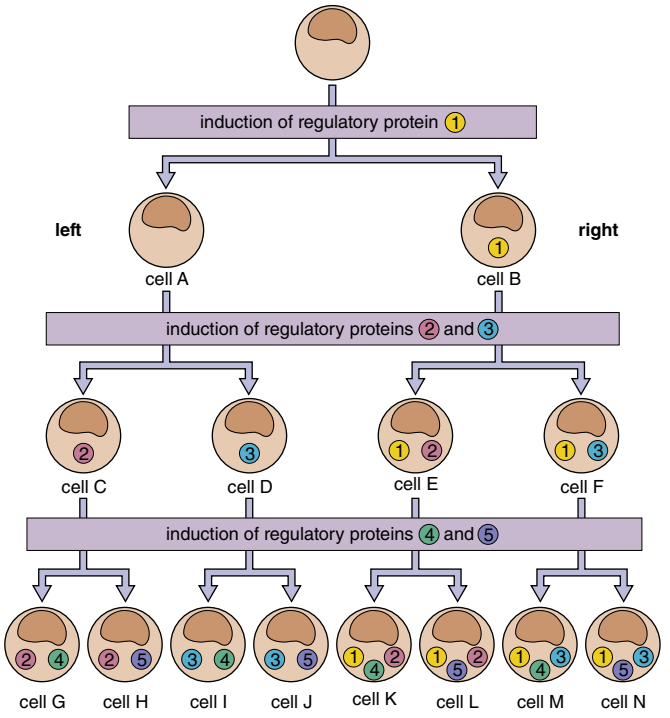


Fig. 1.28. This simplified diagram illustrates the principle that combinations of a limited number of critical regulatory proteins can generate different cell types (lineages) during development and adult life. Thus, in the embryonic life, the cell at the apex of the hierarchy and all its progeny divide asymmetrically such that the cell on the left always produces an even-numbered regulatory protein whereas the cell on the right produces an odd-numbered protein. The protein complement produced by a cell is then perpetuated by its progeny (cellular memory). In this simplistic scenario, five different regulatory proteins generate eight cell types (G to N). With continuation of this scheme, 10 000 cell types would be generated by only 25 different gene regulatory proteins.

types, it is critically required for expression of most of the genes associated with terminal maturation. Extrapolating this more broadly to hematopoiesis, accumulating evidence suggests that a small subset of critical hematopoietic TFs generate all blood cells by working in a combinatorial manner to direct lineage-specific programs of expression (Fig. 1.28). These TFs are not only crucial for normal blood cell programs but the genes encoding them are often a target of acquired mutation that leads to hematologic malignancy.

MICRO-RNAs

Over the last few decades a previously unrecognized class of RNAs, micro-RNAs (miRNAs), have been shown to play important biological roles in controlling expression of proteins in normal cells (Fig. 1.29). Micro-RNAs are encoded by RNA Pol II-transcribed genes to produce pre-miRNAs. These are processed initially in the nucleus and then in the cytoplasm. Here, mature 22 base pair miRNAs bind principally to the untranslated regions of mRNA transcripts leading either to mRNA degradation or repression of translation of mRNAs.

REGULATORY NONCODING RNAs

A relatively new class of RNA molecules have recently been shown to control gene expression and are collectively termed noncoding RNAs (ncRNAs). These include Piwi RNA (piRNAs) (Fig. 1.30A) and long noncoding RNAs (lncRNAs) (Fig. 1.30B). piRNAs are 26–31 nucleotides long and collectively are the most abundant ncRNAs in animal cells. Their roles are still being discerned in hematopoiesis and more generally in control of RNA and protein expression. To date, their clearest role is in epigenetic and translational silencing retrotransposons. The lncRNAs are the largest class of ncRNAs, usually between 200 and 400 nucleotides. It is estimated that there are 10 000–60 000 lncRNAs, a number far greater than protein-encoding mRNAs. The postulated roles of lncRNAs are set out in Fig. 1.30B. Some of these have been shown biochemically in cells and others *in vitro*.

Functional analysis of lncRNAs is in its infancy. It will be quite some time before we fully understand their physiological role and contribution to pathology. But it is likely that they will modulate a cell's response to its environment and in that regard contribute to the complexity of biological responses seen in hematopoiesis. This is likely become a very active area of research.

DNA REPLICATION AND TELOMERES

Every time a human cell divides, its 6 billion base pairs have to be faithfully replicated. This extraordinary task is accomplished daily in billions of cells, for the most part without deleterious consequence. When a cell enters the phase in cell cycle (S, synthesis phase, see later) where the genome is replicated, replication is initiated at multiple areas in the genome called replication foci. DNA replication then proceeds in a semiconservative manner, meaning that the two DNA strands in a double helix separate, are individually replicated, and the resulting two double helices segregate into daughter cells (Fig. 1.31). Thus, each daughter cell has a strand of newly synthesized DNA and a strand from the parental cell. Given that DNA polymerase, the enzyme that replicates DNA, does so in a 5′ to 3′ manner, only one strand is replicated continuously (the leading strand) whereas the other strand has to be replicated in short fragments (Okazaki fragments) (Fig. 1.31B,C). For the lagging strand a number of additional steps are required to ligate the discontinuous Okazaki

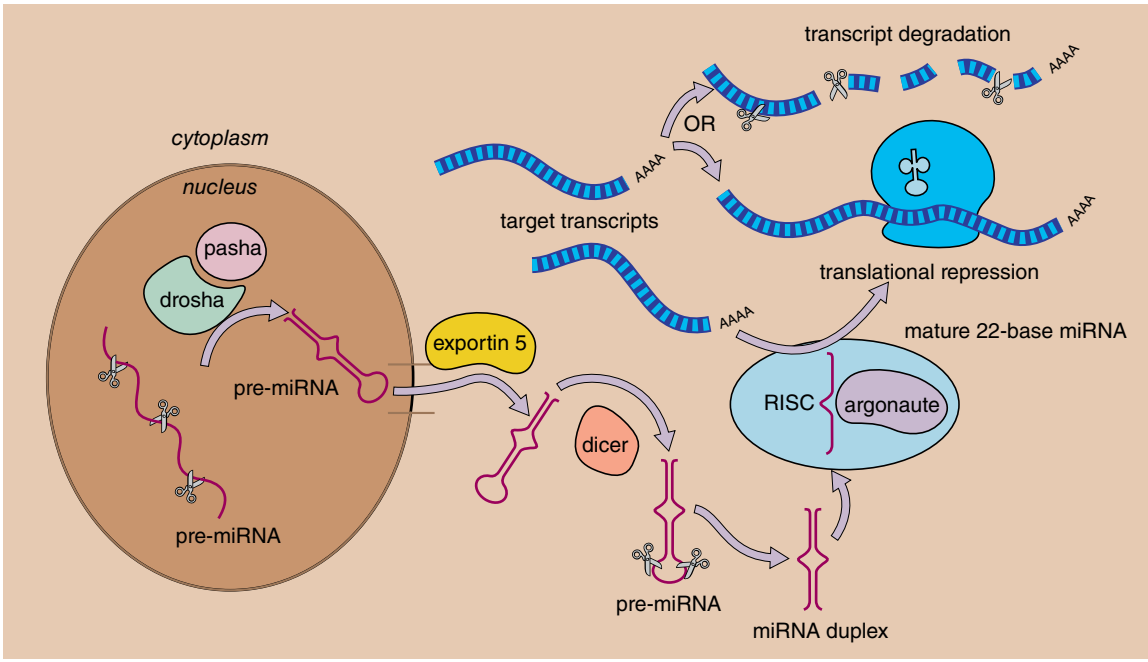


Fig. 1.29. An important class of molecules that regulates gene expression are micro-RNAs (miRNAS). A simplified diagram illustrates how miRNAs are produced and how they regulate miRNA and protein levels. Most miRNAs are expressed from RNA polymerase II-regulated genes as pre-miRNAs. This is cleaved by a nuclear RNase III Drosha and its partner protein DGCR8 Pasha into 50–80 base pair stem–loop pre-miRNAs. These pre-miRNAs are actively exported with the help of exportin-5 into the cytoplasm, where another nuclease, Dicer, excises them into 20–24 base pair mature miRNA duplexes. One of the two miRNA strands is then bound by the multiprotein complex called RNA-induced silencing complex (RISC), which contains, among other proteins, the protein argonaute (RISC/Argonaute). This strand can then repress gene expression by binding to mRNAs by partial sequence complementarity. Binding of mature miRNAs can either inhibit translation of mRNAs or target the mRNA for degradation.

fragments into a continuous DNA strand. DNA synthesized from multiple foci is then ligated together.

One special problem created by the semiconservative mode of DNA replication is the replication of the lagging DNA strand at the ends of chromosomes (Fig. 1.32). This is overcome by having repetitive sequences (called telomeres) at the ends of chromosomes that decrease with each round of replication. For cells that need to self-renew and maintain a high proliferative potential (e.g. germ and other stem cells), the enzyme telomerase can extend the repetitive sequences and compensate for loss at replication. Telomerase is a specialized ribonuclear protein complex composed of a RNA component called TERC that binds to the end of the leading strand and this is replicated by a specialized reverse transcriptase that is also a component of telomerase, called TERT. Loss of telomerase function leads to progressive telomere shortening and this can have catastrophic consequences for cell viability and can lead to transformation of the cell (Fig. 1.33). The degree of maintenance of telomeres determines the number of generations a cell can produce and is often increased above normal in malignant cells. This is discussed further in Chapter 12, where the importance of telomeres in human is elegantly demonstrated by acquired and germline mutations in the patients with dyskeratosis congenita and aplastic anemia.

MUTATIONS AND HOW THEY RESULT IN DISEASE

During DNA replication, errors in fidelity can lead to single base changes (Fig. 1.34A) that will create allele-specific changes in DNA sequence composition that are known as single-nucleotide polymorphism (SNPs) or single-nucleotide variants (SNVs).

If SNVs occur in coding sequence they can either be silent or result in mutation with a functional consequence (Fig. 1.34B). The single-nucleotide missense substitution of a T to G in the sixth codon of the β-globin gene that changes a glutamic acid to valine to produce a βs globin allele (a sickle β-globin allele) is an example of a pathogenic SNP. Due to selective pressure, multiple such functional missense mutations can be detected, even in a single gene such as β-globin or G6PD (Fig 1.34C) (see Chapter 8), producing a large of number of alleles associated with disease.

When SNVs occur in nongeneic parts of the genome they can either have no consequence or potentially alter the regulation of a key gene if the SNV occurs in *cis*-elements (promoters or enhancers) or creates a new *cis*-element, thereby altering the DNA binding of TFs, leading to altered gene expression (Fig. 1.34D). One example of this is mutations that promote expression of the TF SCL/TAL1 or c-myc by promoting its aberrant expression in T-cell acute lymphoblastic leukemia. More examples like this are now coming to light from SNV association studies.

DNA replication errors can also lead to translocation of chromosomes (Fig. 1.35). This is one of the most common karyotypic abnormalities in hematologic malignancy, which affects the expression of the gene either quantitatively (especially frequent in acute lymphoblastic leukemia) or qualitatively. Pathologically important translocations often lead to production of fusion transcripts where the genes at the sites of translocation produce proteins of altered function. For example, BCR-ABL in chronic myeloid leukemia, PML-RARA in acute promyelocytic leukemia (APML or AMLM3), or PBX-EHA in pre B-ALL (see Chapters 13 and 14). Translocation can also cause disease by altering gene expression with pathologic consequences by

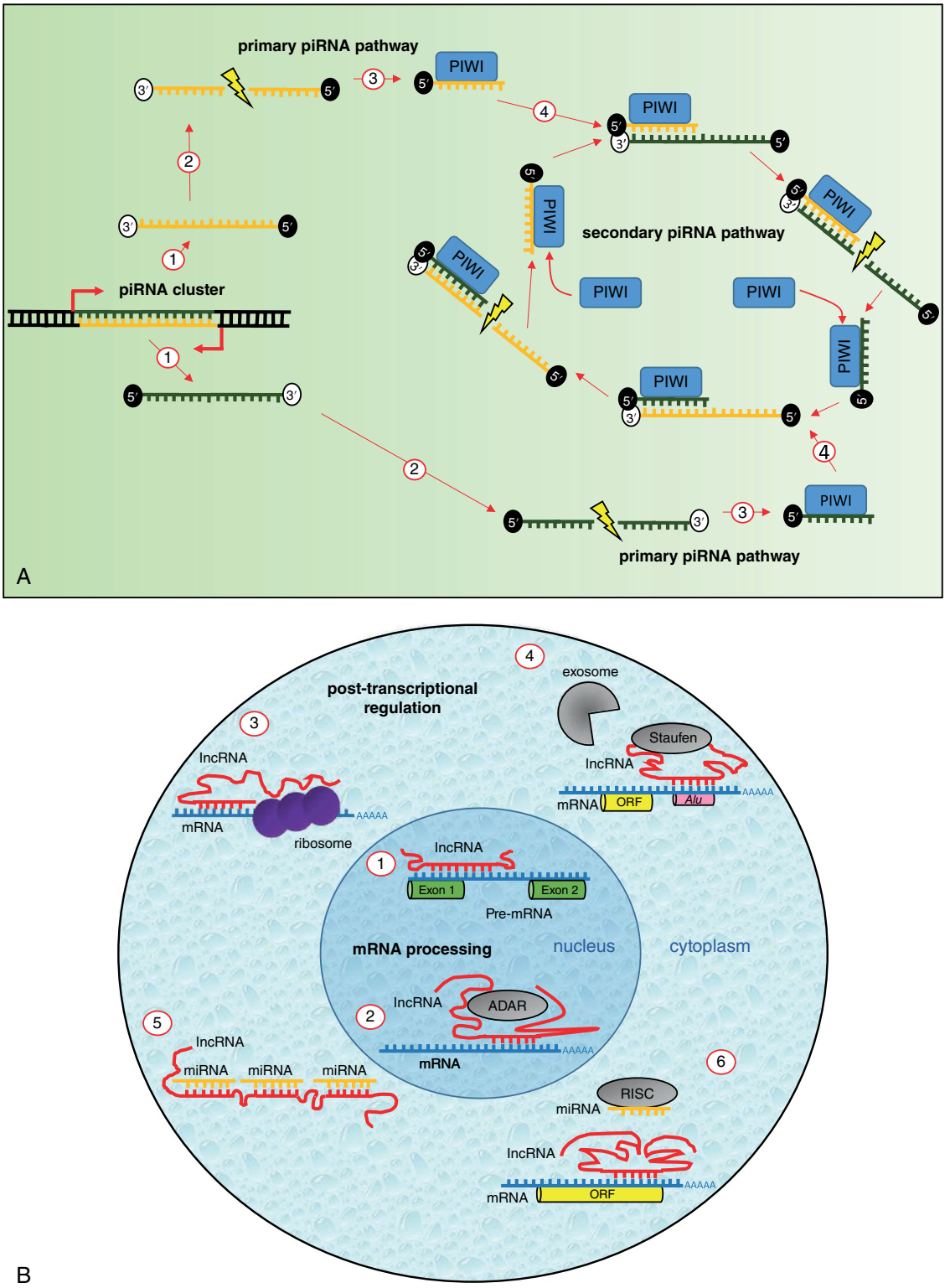


Fig 1.30. A, PIWI proteins and piRNAs regulate expression of genes and transposons at both transcriptional and post-transcriptional levels. *Step 1*, Sense and antisense piRNA precursor transcripts are transcribed from piRNA clusters in the nucleus. *Step 2*, piRNA precursor transcripts are exported to the cytoplasm and processed by the primary biogenesis pathway to generate mature sense piRNAs. *Step 3*, Mature piRNAs consisting of the 5′ end of the precursor then associate with PIWI proteins to enter the secondary piRNA pathway. *Step 4*, The PIWI:piRNA complexes associate with the complementary sequence in unprocessed precursor piRNA (or transposons and protein-coding transcripts) and mediate cleavage. The resulting cleaved 5′ end of the piRNA precursors is taken up by another PIWI protein and the precursor (or transposon or protein-coding transcript) is silenced. This process is known as the ping-pong cycle. PIWI–piRNA complexes interact with polysomes; mRNA cap-binding complex (CBC), P-body components, and piRNAs are mapped to the 3′ UTR of mRNAs. The PIWI–piRNA complexes can enter the nucleus and regulate gene transcription through epigenetic mechanisms including heterochromatin formation and DNA methylation. **B,** The roles of lncRNA in mRNA processing and post-transcriptional regulation. Within the nucleus, lncRNA modulates mRNA processing in one of two ways: (1) mRNA is bound at regions overlapping exon:intron boundaries. (2) lncRNA recruits mRNA-editing enzymes, such as adenosine deaminase (ADAR), to complementary mRNA sequences. In the cytoplasm, lncRNA regulates post-transcriptional events through at least four distinct mechanisms: (3) Recruitment of post-transcriptional machinery to mRNA due to possession of sequence-specific domains, such as SIN EB2 repeat elements that have affinity for ribosomes. (4) lncRNAs that contain Alu repeat elements associate with Alu elements in the 3′ UTR of mRNA, which recruits Staufen to induce mRNA degradation via RNA exosomes. (5) Linear or circular lncRNAs can serve as molecular sponges to sequester miRNAs from their target sequences. (6) lncRNAs can mask sequences in mRNA that would serve as targets for miRNAs bound to RNA-induced silencing complex (RISC). Source: Wilkes MC, et al. *Mol Genet Metab* 2017;122(3):28–38. Reproduced with permission of Elsevier.

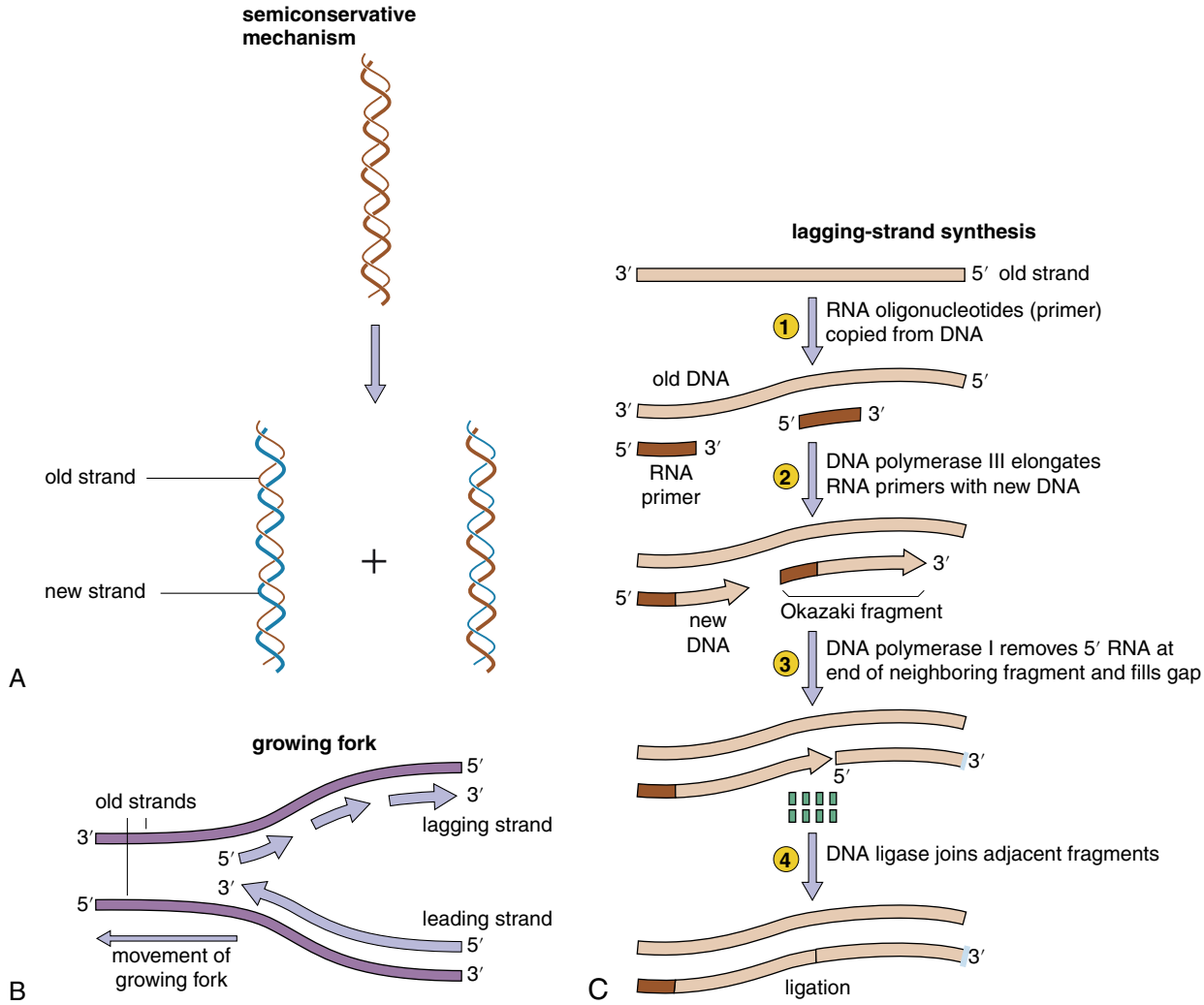


Fig. 1.31. DNA replication is semiconservative. **A,** The parental DNA is composed of two strands. During replication each parental strand acts as a template for replication, and thus progeny strands contain one new strand and one old strand. **B,** DNA replication is initiated at origins of replication, and as replication proceeds the parental strands separate and a replication fork forms. DNA is synthesized from a 5′ to 3′ direction, and this creates two types of strands of newly replicated DNA. One strand, called a leading strand, is continuously synthesized by DNA polymerase III by sequential addition of deoxyribonucleotides in the same direction as the movement of the replication fork (the bottom strand in this figure). In contrast, the other newly synthesized strand is made discontinuously (top strand) and is known as the lagging strand. **C,** Synthesis of the lagging strand requires multiple steps. First, multiple RNA oligonucleotides (primers) are synthesized from parental DNA strand templates. These serve to prime synthesis of fragments (called Okazaki fragments) of the new (lagging) strand. DNA polymerase I then removes the RNA primers, fills in the gaps with DNA, and finally DNA ligase joins adjacent DNA fragments.

altering the cell fate of hematopoietic stem/progenitor cells (Fig. 1.36). Control of cell fate is more extensively discussed in Chapters 2 and 3.

CELL CYCLE

One critical determinant of cell fate is whether cells enter the cell cycle. Although much is known about the molecular controls and cell biology of the cell cycle, we are still unearthing the complexity of how decisions are made on whether cell cycles are linked with external cues, the cellular history of the cell, and the cell compartment a cell is in (i.e. stem cell or progenitor cell).

The cell cycle is divided into phases (Fig. 1.37A and C). There is cyclic synthesis of DNA (S phase), a pause known as G_2, followed by chromosome condensation, nuclear envelope breakdown and chromosome segregation. This leads to chromosome decondensation, nuclear envelope reformation and cytokinesis terminating in separation of two daughter cells (mitosis or M phase). A critical set of proteins that control passage through cell cycle are the cyclins, the expression of which differs throughout the cell cycle by periodic changes in synthesis and degradation (Fig. 1.37B). They activate a set of protein kinases (CDKs) (Fig. 1.37B), which then phosphorylate various proteins. Phosphorylation of the retinoblastoma susceptibility gene product RB prevents RB from blocking the TFs (e.g. E2F), which is essential for transition from the G_1 to the S phase of the cell cycle. The cyclin/CDK complexes are regulated by inhibitors. Thus, cyclin D–CDK4 and cyclin D–CDK6 complexes are inhibited by a 16 kDa protein encoded by the *INK4a* gene and a 15 kDa protein encoded by the *INK4b* gene, inhibiting progress from mid to late G_1.

The *TP53* gene codes for a 53 kDa transcription control factor that mediates a block in the cell cycle at the G_1–S phase boundary).

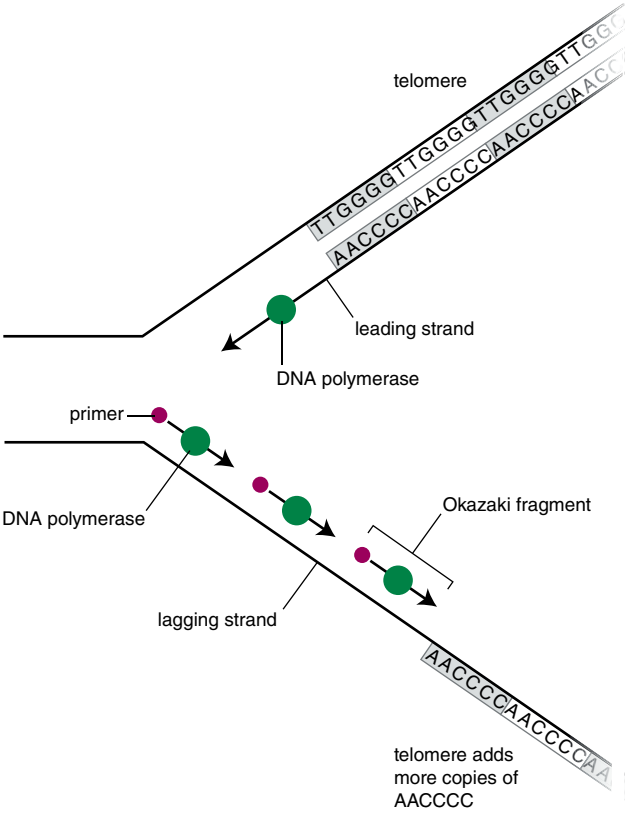


Fig. 1.32. Replication of the ends of chromosomes requires specialized structures called telomeres. Telomeres are DNA sequences composed of a motif that is repeated. In humans the repeat DNA motif is "TTGGGG" on one strand and "AACCCC" on the other. The leading strand is shown on the upper fork. The DNA polymerase (green circle) is replicating DNA. On the lagging strand a primer sequence allows short Okazaki fragments to be made. The telomere is then replicated on both strands by a specialized RNA–protein complex called telomerase.

This is mediated by a p21 cyclin CDK inhibitor, p21CIPI. Expression of p53 is induced by DNA damage that results from radiation or drugs. The cell is therefore held up in G_1 to allow the cell to repair the damage. If the damage is extensive, p53 induces apoptosis by increased expression of proapoptotic gene *BAX*. This is not the only example of how progression through cell cycle and response to external cues (in this case DNA damage) are linked with another cell fate choice option, namely apoptosis.

APOPTOSIS

Cell death can occur by necrosis or by a physiologically active mechanism (apoptosis, programmed cell death) (Fig. 1.38). Necrosis occurs in response to ischemia, chemical trauma, or hyperthermia. It affects many adjacent cells, and is characterized by cell swelling, with early loss of plasma membrane integrity and swelling of organelles and nucleus. There is usually an inflammatory infiltrate of phagocytic cells in response to spillage of cell contents into surrounding space.

Programmed cell death occurs by an active process that requires calcium ions. Nuclear condensation, nuclear fragmentation, and cytoplasmic vacuolation occur early, with later changes in the organelles and plasma membrane (Fig. 1.39). Apoptosis also involves digestion of cell DNA by an endonuclease to produce on a gel a ladder of regular bands 180 base pairs apart. Cleavage occurs by double-stranded breaks on linker regions between nucleosomes. The final part of the apoptosis pathway involves caspase enzymes. The executioner caspase 3 cleaves a restricted set of cellular proteins, including polyadenosine diphosphate–ribose polymerase, laminin, and gelsolin (Fig. 1.39). Caspase 3 is activated by caspase 9. This in turn is activated by the apoptotic protease 1 (APAF-1), which itself is activated by cytochrome *c*. Cytochrome *c* is released from mitochondria when the proapoptotic protein BAX is in excess and forms homodimers. Cells are protected from apoptosis by BCL-2, which binds to BAX and

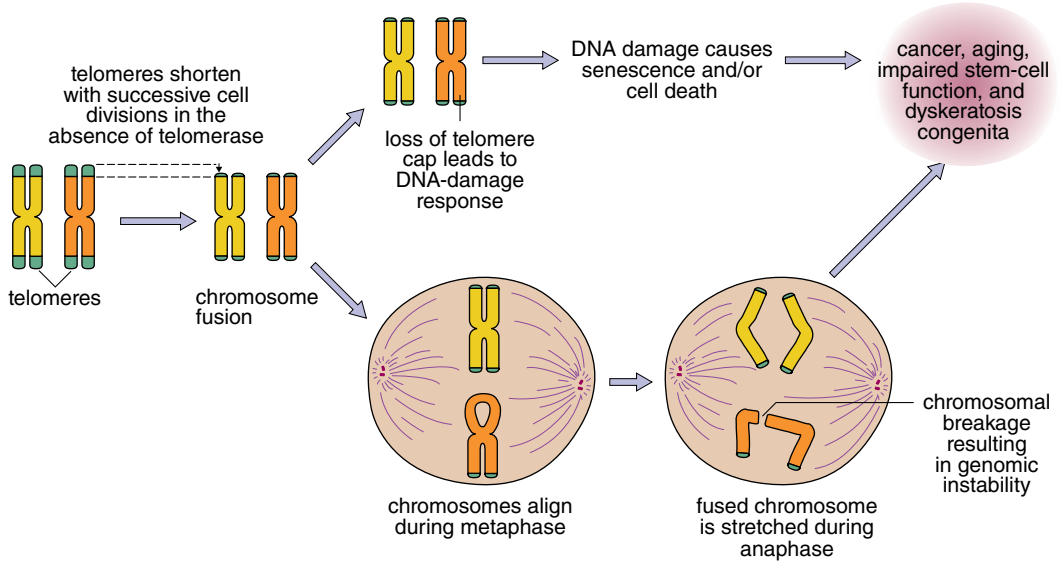


Fig. 1.33. The enzyme complex telomerase is essential for maintaining telomere length, and this is critical for chromosomal and cell viability. Lack of telomerase activity leads to telomere shortening with successive cell division. This can activate DNA damage responses and lead to cell senescence and cell death. Alternatively, telomere shortening can precipitate chromosomal fusion, and genomic instability. A specific condition associated with impaired telomerase activity is dyskeratosis congenita (see Chapter 12), in which children and young adults can have hematopoietic stem cell failure and leukemia. More generally, impaired telomere function has been implicated in many cancers and aging.

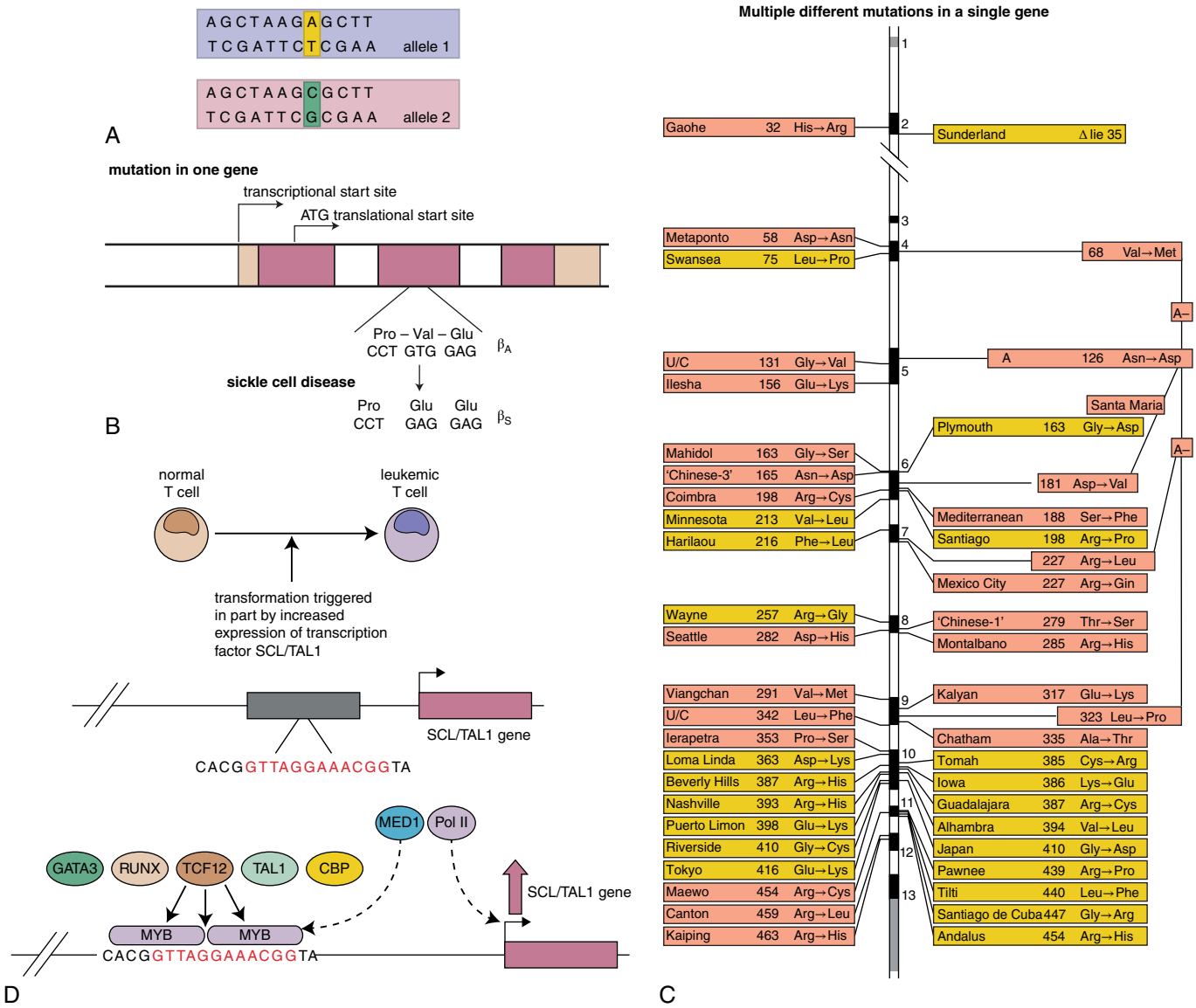


Fig 1.34. Errors in DNA replication or the maintenance of the methylated state of CpG residues can change DNA sequence. **A,** When the DNA sequence change occurs at single nucleotide it is known as a single-nucleotide polymorphism (SNP). In this example, the "A" residue on allele 1 is changed to a "C" residue on allele 2. **B,** When SNP occurs in a coding sequence it can alter the protein sequence. In this example, a "T" residue at the sixth codon in exon 2 of the human β-globin gene is changed to an "A" residue. This changes a valine (val) to glutamic acid (glu). This results in a β-sickle allele. In the homozygous state this causes sickle cell disease (see Chapter 9). **C,** Mutations can occur at multiple positions in the gene. In the glucose-6-phosphate dehydrogenase (*G6PD*) gene, many different mutations cause G6PD deficiency. These may cause drug sensitivity (pink) or more rarely chronic nonspherocytic hemolytic anemia (NSHA) (yellow). The exons are shown as black bands except exon 1, which is noncoding and shown in gray. Source: C, Adapted with permission from Vulliamy TJ et al. Molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends Genet* 1992;8:138–143. **D,** Nucleotide changes can also occur in DNA sequences that control RNA expression of a gene. In this example, the gene encoding the important blood transcription factor SCL/TAL1 is shown as a pink box on the left. 5′ of the gene a sequence "GTTAGGAAACGG" has been erroneously inserted (shown as pink letters) and this creates a new binding site for transcription factor MYB. MYB, in turn, attracts the additional transcription factors GATA3, RUNX1, TCF12, TAL1 itself, and CBP. This transcription factor complex works with the general transcription factor MED1 and RNA polymerase II (Pol II) to increase transcription of SCL/TAL1, which is oncogenic. This situation is seen in human T-cell acute lymphoblastic leukemia (see Chapter 14). Source: Vulliamy TJ, et al. *Trends Genet* 1992;8:138–143. Reproduced with permission of Elsevier.

thereby inhibits cytochrome *c* release and caspase activation. Apoptosis is promoted by BAD, which forms heterodimers with BCL-2. Apoptosis may also be stimulated by direct DNA damage (e.g. by radiation or drugs) or by withdrawal of a growth factor (e.g. interleukin 3 [IL-3]). When IL-3 is present it promotes cell survival in IL-3-responsive cells by stimulating protein kinase B to phosphorylate BAD, thus preventing its association with BCL-2.

ORGANELLES IN CELLS

MITOCHONDRIA

Mitochondria are complex organelles that are the main sites of ATP production (Fig. 1.40A,B) during aerobic metabolism, important for heme biosynthesis (see Chapter 5) and finally play a role in apoptosis. They are among the largest organelles and can comprise up to 25% of the cellular volume. They are

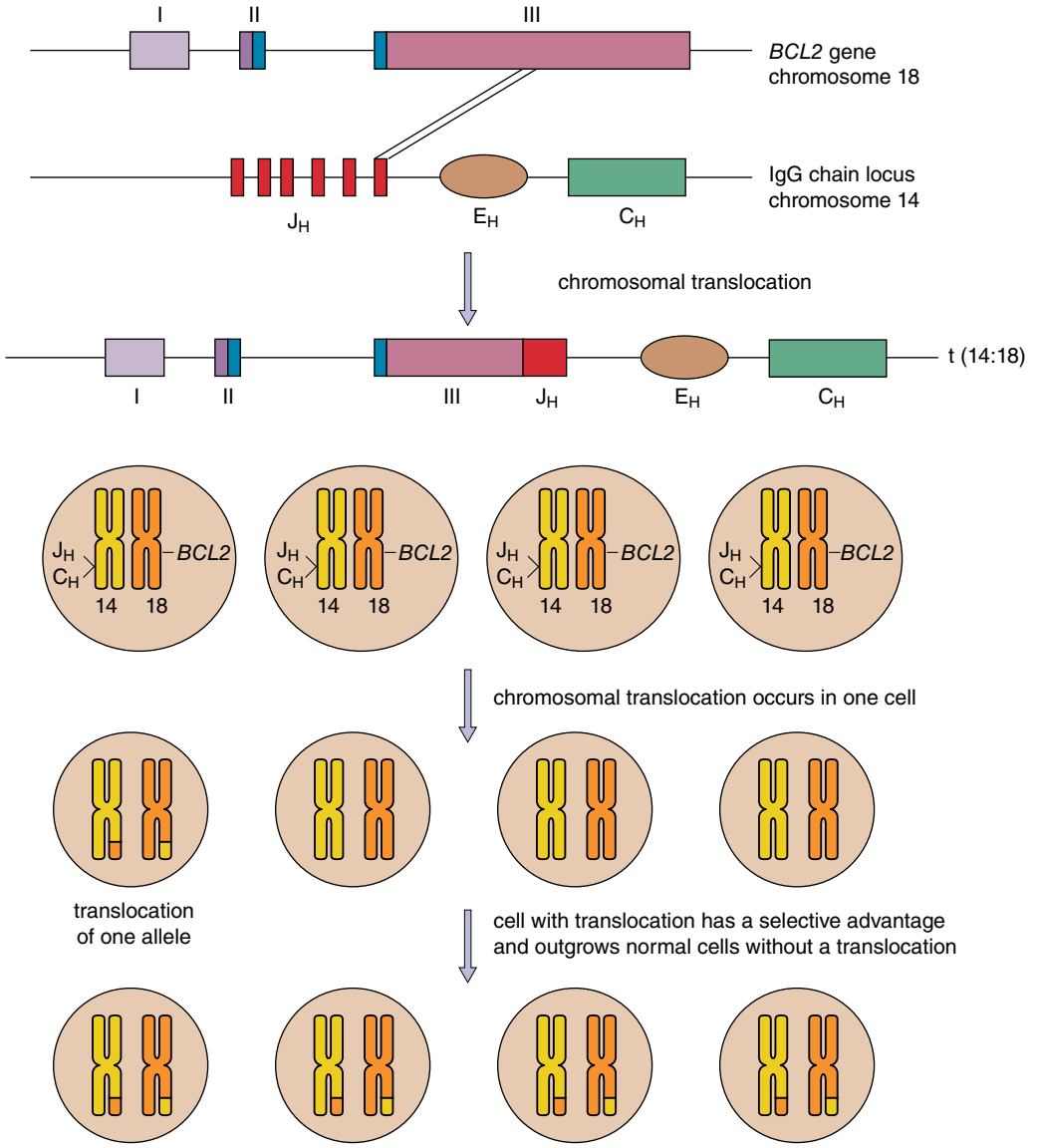


Fig. 1.35. Chromosomal translocation can lead to disease-causing mutation. In this example, the *BCL2* gene on chromosome 18 is involved in a translocation with the Ig heavy chain locus on chromosome 14. In this case the translocation breakpoints are inside the genes and give rise to a fusion transcript. Acquisition of this translocation provides a selective advantage to cells.

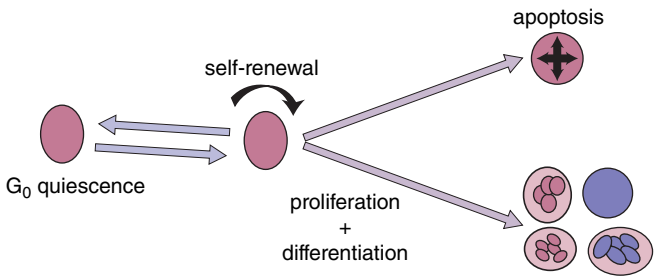


Fig. 1.36. Schematic showing the cell fate choices available to a hematopoietic stem/progenitor cell.

composed of two structurally and functionally distinct inner and outer membranes. The inner membrane has a large number of invaginations or cristae that protrude into the central space or matrix of the mitochondria. The complex biochemical pathways that produce ATP from aerobic metabolism of glucose (via pyruvate produced by the glycolytic pathway) can result in up to 34 molecules of ATP for every molecule of glucose. This involves phosphorylation and oxidation. Fatty acids can also be metabolized to CO_2 to generate ATP. Metabolism of pyruvate and fatty acids generates NADH and $FADH_2$ molecules that are oxidized to NAD^+ and FAD, and the resulting protons are pumped across the inner mitochondrial membrane. ATP generation is powered by the resulting proton-motive force. Thus, it is easy to see why mitochondria are required to provide energy for a cell.

LINK BETWEEN METABOLISM AND GENE EXPRESSION

It is being increasingly appreciated that cellular energetics is a vital consideration in the maintenance of tumor cell viability. This is as true for blood cancers as for other cancers. At least two broad mechanisms operate. First, by-products of

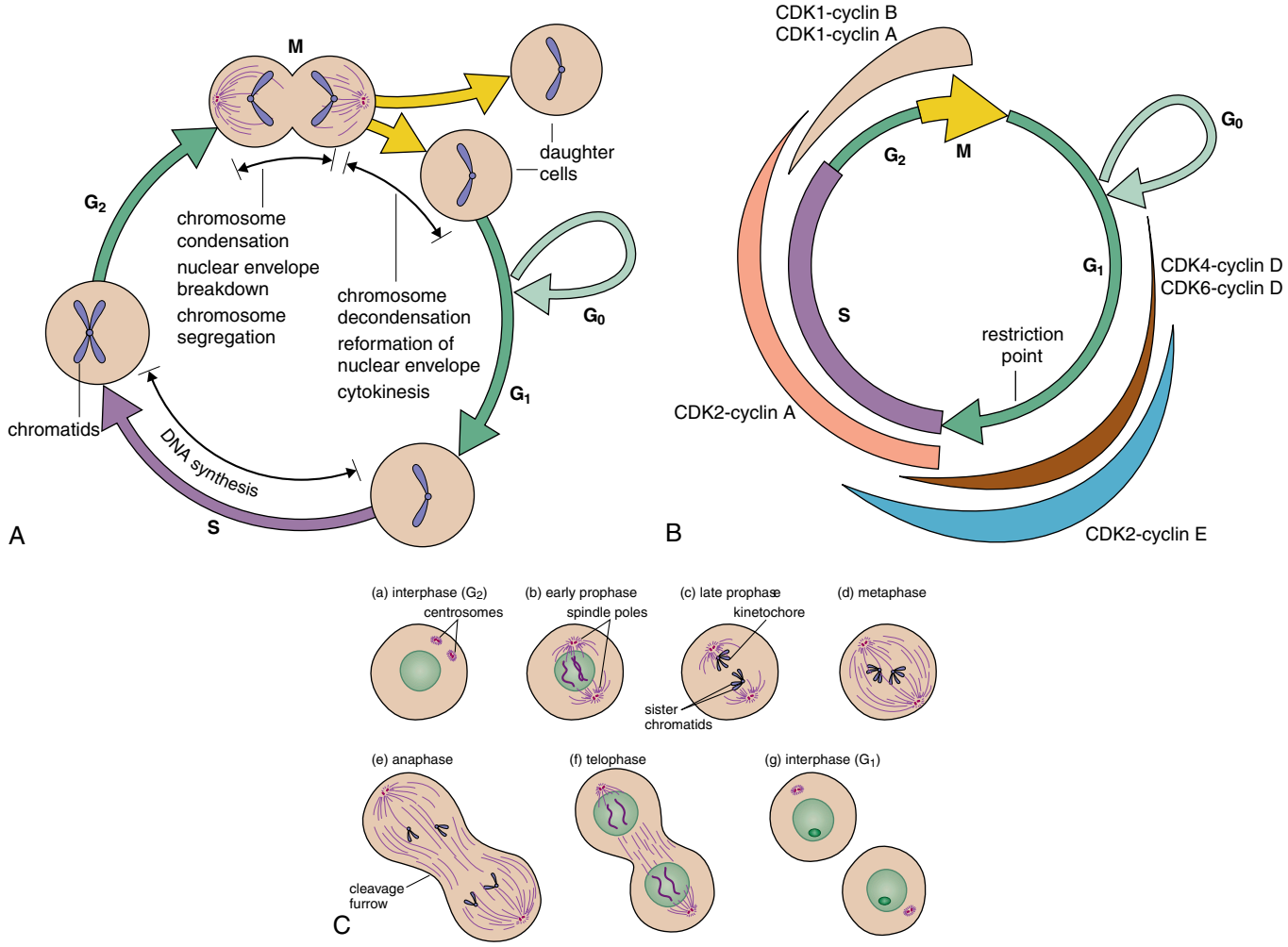


Fig. 1.37. A, Schematic representation of the cell cycle. Quiescent cells ("G_0") enter the cycle in "G_1" and progress to synthesize DNA in the "S" phase. In the G_2 phase cells have 4*n* DNA content and this leads to mitosis "M." The G_1, S, and G_2 phases collectively are known as interphase. Although chromosomes are condensed only in mitosis, they are shown here in condensed form throughout the cell cycle to emphasize the numbers of chromosomes at different stages in the cell cycle. **B,** Passage through the cell cycle is, in part, regulated by proteins that themselves cycle. These include the cycle-dependent kinases (CDKs) that are physically associated (i.e. are a complex) with proteins that regulate them, called cyclins. Different combinations of CDKs and cyclins are required to progress through different stages of the cell cycle. Thus CDK2–cyclin E complex is important in G_1. Critical regulatory points in the cell cycle are called restriction points, and the location of the G_1 restriction point is shown. **C,** During the cell cycle a key facet is the proper separation of sister chromatids (generated in "S" phase). (a) In late G_2, structures called centrioles are replicated. (b) In early prophase, chromosomes and associated centrioles move to cell poles. Now the chromosomes start to condense and can be seen as threads. The nuclear membrane begins to disaggregate. (c) In late prophase, chromosome condensation is complete. The chromosome centromeres are visible, and they progressively move to the pole and microtubular spindle fibers connect the centromeres to the poles of the cell. (d) In metaphase, chromosomes move to the cell equatorial plane. (e) In anaphase, the sister chromatids separate into independent chromosomes. Each centromere is connected to the pole by a spindle fiber and moves to the pole. Simultaneously, the cell elongates. Cytokinesis begins and cleavage furrows appear. (f) In telophase, new nuclear membranes are seen and chromosomal decondensation starts. The cells now reenter G_1 and interphase. (g) Interphase.

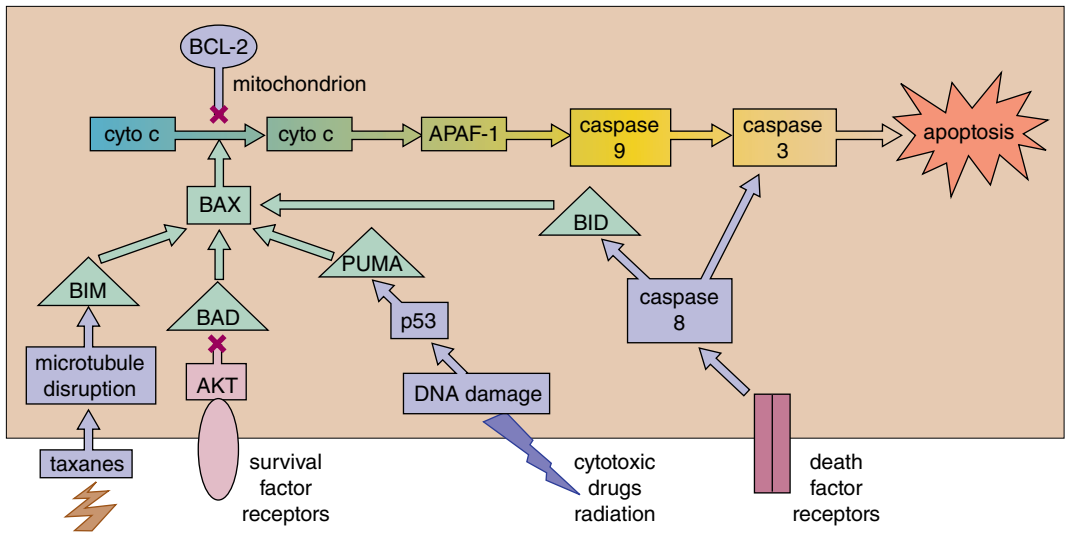


Fig. 1.38. This simplified diagram shows key elements of the pathways that lead to caspase 3 activation and apoptosis. A number of stimuli can activate the pathway. See text for details.

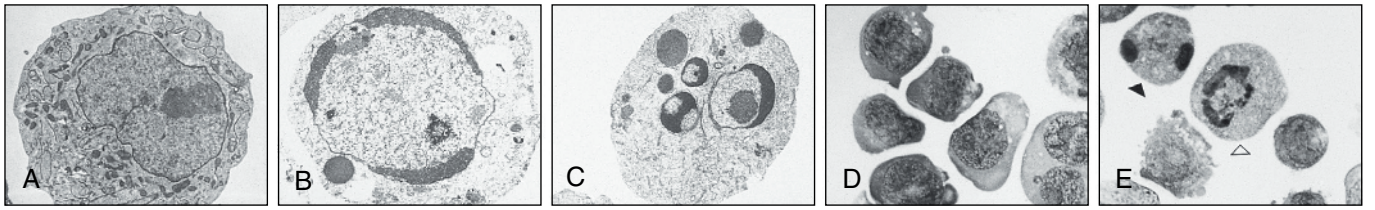


Fig. 1.39. Apoptosis: **(A–C)** electron microscopic and **(D, E)** light microscopic appearances. **A,** Normal K562 cell line. **B,** Early apoptotic cell showing chromatin condensation at the nuclear periphery. **C,** Later apoptotic cell showing both chromatin condensation and nuclear fragmentation. **D,** Normal K562 cell line. **E,** Early apoptotic cell (open arrowhead) with peripheral chromatin condensation and late apoptotic cell (solid arrowhead) with both chromatin condensation and nuclear fragmentation. Source: Riordan FA, et al. *Oncogene* 1998;16:1533–1542. Reproduced with permission from Springer Nature.

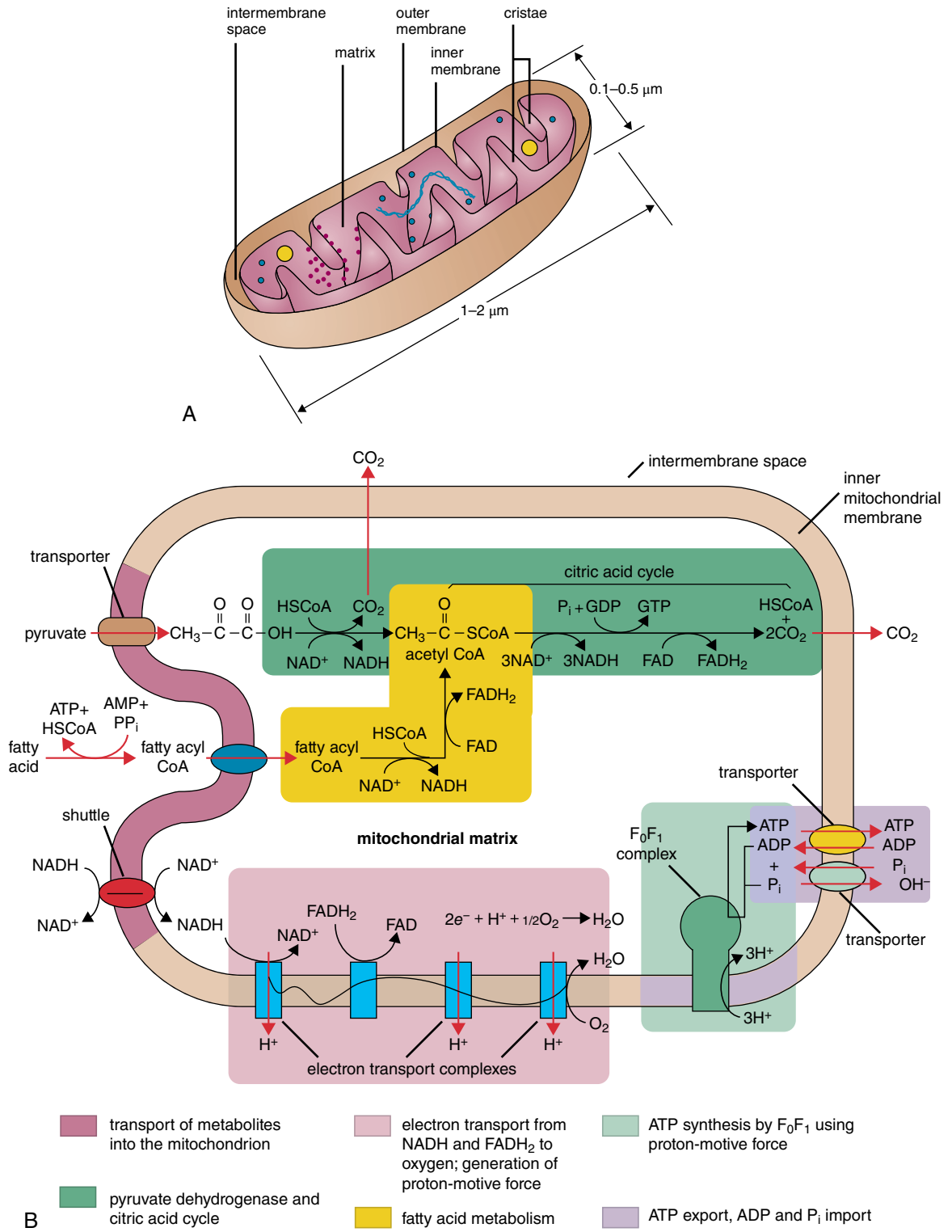


Fig. 1.40. A, Three-dimensional schematic diagram of a mitochondron cut longitudinally. The ATP-producing complexes (F_0F_1, red cell dots) are located on the inner membrane protruding inward. Mitochondrial DNA (blue), ribosomes (blue circles), and granules (yellow dots) are shown. **B,** Summary of aerobic oxidation of pyruvate illustrates some of the complexity of the biochemistry within the mitochondria.

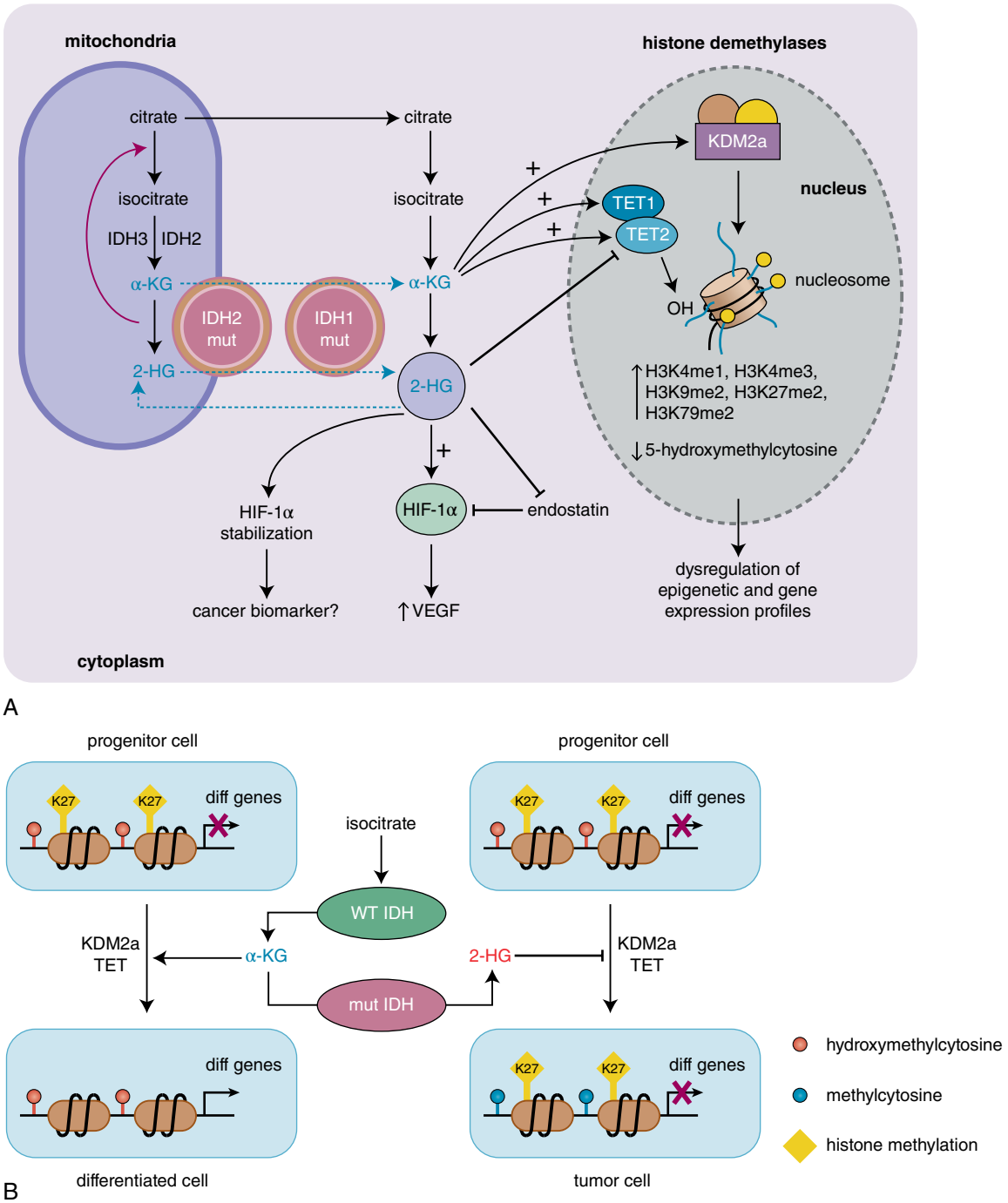


Fig 1.41. Cellular energetics and gene expression are linked. Here, we show one example of this. **A,** As part of the Krebs cycle, citrate is converted to isocitrate, which in turn is converted to α-ketoglutarate (α-KG) by the enzymes isocitrate dehydrogenase (IDH). There are three isoforms of IDH (IDH1, IDH2, and IDH3). IDH2 and IDH3 are located in the mitochrondria and IDH1 is cytoplasmic. α-KG is a required cofactor for the enzymes that modify histones (e.g. histone demethylases KDM2a) and demethylate DNA (TET1 and TET2). Mutations in IDH1 and IDH2 are found in myeloid cancer (see Chapter 14). Mutant IDH1 and IDH2 convert α-KG to the oncometabolite 2-hydroxyglutarate (2-HG). 2-HG inhibits histone demethylases and TET1 and TE2, thus antagonizes α-KG function. Furthermore, 2-HG promotes a cell's adaption to hypoxia (often seen in cancers) by promoting expression of HIF-1α. **B,** This deregulates histone and DNA methylation, alters gene expression, and promotes oncogenesis by stalling differentiation and promoting proliferation. A progenitor cell is shown on top on both left- and right-hand sides. Normally, α-KG would activate histone demethylase KDM2A and DNA TET demethylases. This removes methyl groups on histone H3K27 (which repress transcription) and keep cytosine residues as either cytosine or hydroxymethylcytosine. Both these epigenetic marks are permissive for transcription allowing expression of genes, for example those associated with differentiation. In contrast, 2-HG blocks TET and KDM2A enzymes, resulting in continued histone H3K27 methylation and cytosine residues being methylated. Both of these epigenetic marks repress expression of genes required for differentiation and lead to differentiation block, which a hallmark of blood cancers.

metabolism control gene expression (Fig. 1.41). An example here is that 2-ketoglutarate, an intermediate of the citric acid cycle, is an essential cofactor of enzymes that regulate DNA methylation and histone modification. Similarly amino acid metabolism provides the methyl donors for DNA and histone methylation. The proteins mediating the links between metabolism and gene expression are mutated in blood diseases (see Chapter 13).

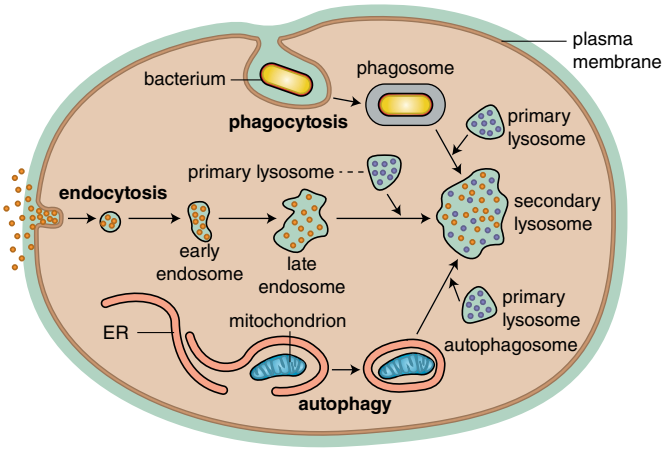


Fig. 1.42. Lysosomes degrade ingested extracellular particles (e.g. bacteria) and intracellular particles. Three main pathways deliver material to lysosomes: phagocytosis, endocytosis, and autophagy

REMOVAL OF CIRCULATING AND CELLULAR DEBRIS BY LYSOSOMES

Lysosomes are closed intracellular compartments composed of a single membrane that are responsible for degrading intracellular components no longer required for the cell's metabolism. Material is mostly taken up by a lysosome by three routes: phagocytosis, endocytosis, and autophagy (Fig. 1.42). Phagocytosis is when material is taken up into a membrane-bound phagosome. Endocytosis is the process by which small portions of the plasma membrane invaginate to form a small membrane-bound vesicle (endosomes). The endosomes is then combined with a primary lysosome to create a secondary lysosome. Secondary lysosomes can also form when primary lysosomes fuse with phagosomes. Finally, aged mitochondria are removed by a process known as autophagy, in which an autophagosome combines with a primary lysosome to make a secondary lysosome.

Lysosomes then release enzymes (termed acid hydrolases) that work at acid pH to denature the lysosome contents (e.g. proteins). There is an ATP-dependent pump that generates the acid pH and lysosomal enzymes work best in acid (pH 4.8) conditions and not in neutral cytosolic pH.

PROTEIN UBIQUITINATION

Ubiquitin (ubiquitous immunopoietic peptide) is a highly conserved 76-amino-acid 8.5 kDa peptide that is used to mark proteins for destruction. Ubiquinated proteins are targeted to the proteosome, which cleaves ubiquitin-tagged proteins in an ATP-dependent process to yield peptides and intact ubiquitin. Ubiquitin is added by a conjugating enzyme (ubiquitinating complex). First, ubiquitin is activated and bound to the enzyme E1 and then is transferred to the enzyme E2. E2 binds the ubiquitin ligases E3 (there are many different types of E3 ligase) (Fig. 1.43). The protein targeted for destruction is recognized by internal sequences. Successive conjugations of ubquitin moieties (at least four) usually to a lysine residue are required for proteosome targeting. The ubiquitin–proteosome pathway is a central process in controlling protein turnover in the cell. There are several hematologic diseases associated with this pathway, including forms of Fanconi anemia (mutations in genes for a large E3 ligase) and von Hippel–Lindau disease (mutations in genes for another E3 ligase). Furthermore, a widely used class of drugs in blood cancers (called ImiDs or CelMods) attach to ubiquitin E3 ligase to mark specific proteins for destruction (see Chapter 21).

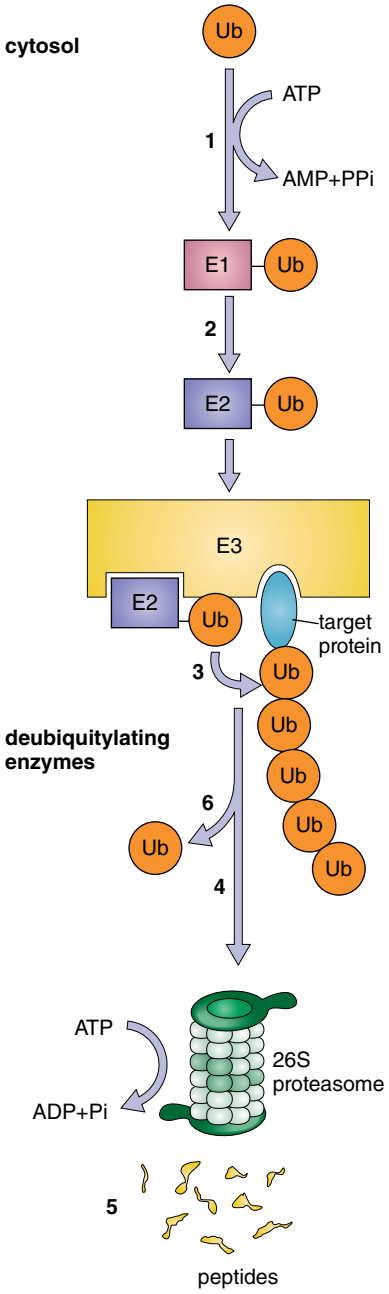


Fig. 1.43. Proteosomal degradation of cellular protein. Ubiquitin (Ub) is added to enzyme E1 (ubiquitin-activating enzyme) in an ATP-dependent process (1). Ubiquitin is transferred to protein E2 (ubiquitin–carrier protein) (2). This is then complexed to ubquitin ligase (3). E3 binds E2/ubiquitin and the target protein destined for destruction. This allows ubiquitin to be transferred to the polyubiquitin chain on the target protein. The polyubiquitinated protein is then proteolyzed in a 26S proteosome in an ATP-dependent process.