Organization of the Human Genome

9

KEY CONCEPTS

- The human genome is subdivided into a large nuclear genome with more than 26,000 genes, and a very small circular mitochondrial genome with only 37 genes. The nuclear genome is distributed between 24 linear DNA molecules, one for each of the 24 different types of human chromosome.

- Human genes are usually not discrete entities: their transcripts frequently overlap those from other genes, sometimes on both strands.

- Duplication of single genes, subchromosomal regions, or whole genomes has given rise to families of related genes.

- Genes are traditionally viewed as encoding RNA for the eventual synthesis of proteins, but many thousands of RNA genes make functional noncoding RNAs that can be involved in diverse functions.

- Noncoding RNAs often regulate the expression of specific target genes by base pairing with their RNA transcripts.

- Some copies of a functional gene come to acquire mutations that prevent their expression. These pseudogenes originate either by copying genomic DNA or by copying a processed RNA transcript into a cDNA sequence that reintegrates into the genome (retrotransposition).

- Occasionally, gene copies that originate by retrotransposition retain their function because of selection pressure. These are known as retrogenes.

- Transposons are sequences that move from one genomic location to another by a cut-and-paste or copy-and-paste mechanism. Retrotransposons make a cDNA copy of an RNA transcript that then integrates into a new genomic location.

- Very large arrays of high-copy-number tandem repeats, known as satellite DNA, are associated with highly condensed, transcriptionally inactive heterochromatin in human chromosomes.
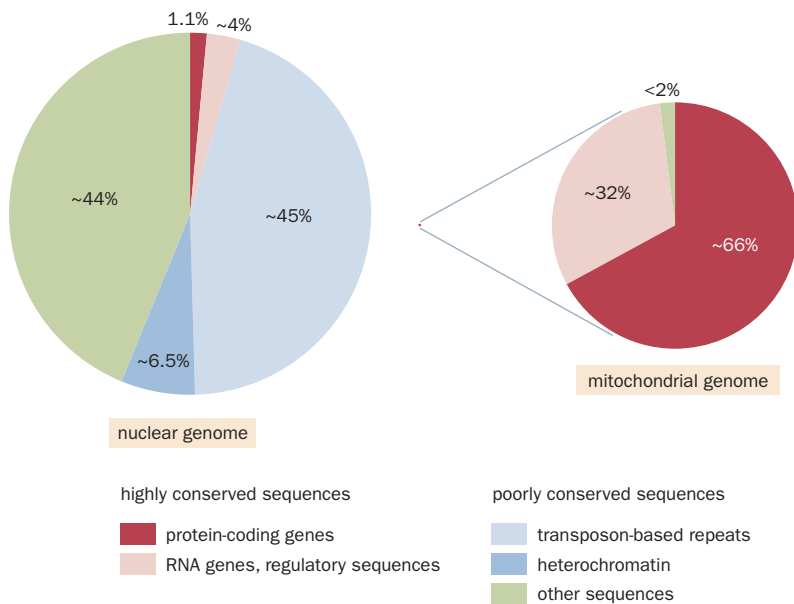
Figure 9.1 Sequence conservation and sequence classes in the human nuclear and mitochondrial genomes. To get an idea of the vast difference in scale between the nuclear (left) and mitochondrial (right) genomes, the tiny red dot in the center represents the equivalent of 25 mitochondrial DNA (mtDNA) genomes on the same scale as the single nuclear genome on the left. Note also the profound difference between the two genomes in the fractions of highly conserved DNA and also in the fraction of highly repetitive noncoding DNA.

The human genome comprises two parts: a complex *nuclear genome* with more than 26,000 genes, and a very simple *mitochondrial genome* with only 37 genes (**Figure 9.1**). The nuclear genome provides the great bulk of essential genetic information and is partitioned between either 23 or 24 different types of chromosomal DNA molecule (22 autosomes plus an X chromosome in females, and an additional Y chromosome in males).

Mitochondria possess their own genome—a single type of small circular DNA—encoding some of the components needed for mitochondrial protein synthesis on mitochondrial ribosomes. However, most mitochondrial proteins are encoded by nuclear genes and are synthesized on cytoplasmic ribosomes before being imported into the mitochondria.

As detailed in Chapter 10, sequence comparisons with other mammalian genomes and vertebrate genomes indicate that about 5% of the human genome has been strongly conserved during evolution and is presumably functionally important. Protein-coding DNA sequences account for just 1.1% of the genome. The other 4% or so of strongly conserved genome sequences consists of non-protein-coding DNA sequences, including genes whose final products are functionally important RNA molecules, and a variety of *cis*-acting sequences that regulate gene expression at DNA or RNA levels. Although sequences that make non-protein-coding RNA have not generally been so well conserved during evolution, some of the regulatory sequences are much more strongly conserved than protein-coding sequences.

Protein-coding sequences frequently belong to families of related sequences that may be organized into clusters on one or more chromosomes or be dispersed throughout the genome. Such families have arisen by gene duplication during evolution. The mechanisms giving rise to duplicated genes also give rise to nonfunctional gene-related sequences (*pseudogenes*).

One of the big surprises in the past few years has been the discovery that the human genome is transcribed to give tens of thousands of different *noncoding RNA* transcripts, including whole new classes of tiny regulatory RNAs not previously identified in the draft human genome sequences published in 2001. Although we are close to obtaining a definitive inventory of human protein-coding genes, our knowledge of RNA genes remains undeveloped. It is abundantly clear, however, that RNA is functionally much more versatile than we previously suspected. In addition to a rapidly increasing list of human RNA genes, we have also become aware of huge numbers of pseudogene copies of RNA genes.

A very large fraction of the human genome, and other complex genomes, is made up of highly repetitive noncoding DNA sequences. A sizeable component is organized in tandem head-to-tail repeats, but the majority consists of interspersed repeats that have been copied from RNA transcripts in the cell by reverse

transcriptase. There is a growing realization of the functional importance of such repeats.

In this chapter we primarily consider the *architecture* of the human genome. We outline the different classes of DNA sequence, describe briefly what their function is, and consider how they are organized in the human genome. In later chapters we describe other aspects of the human genome: how it compares with other genomes, and how evolution has shaped it (Chapter 10), DNA sequence variation and polymorphism (Chapter 13), and aspects of human gene expression (Chapter 11).

9.1 GENERAL ORGANIZATION OF THE HUMAN GENOME

The DNA sequence of the human mitochondrial genome was published in 1981, and a detailed understanding of how mitochondrial DNA (mtDNA) works has been built up since then. The more complex nuclear genome has been a much more formidable challenge. Comprehensive sequencing of the nuclear genome began in the latter part of the 1990s, and by 2004 essentially all of the euchromatic portion of the genome had been sequenced. Our knowledge of the nuclear genome remains fragmentary, however. As we see below, we still do not know how many genes there are in the nuclear genome, and recently obtained data are radically changing our perspective on how it is organized and expressed.

The mitochondrial genome is densely packed with genetic information

The human mitochondrial genome consists of a single type of circular double-stranded DNA that is 16.6 kilobases in length. The overall base composition is 44% (G+C), but the two mtDNA strands have significantly different base compositions: the heavy (H) strand is rich in guanines, but the light (L) strand is rich in cytosines. Cells typically contain thousands of copies of the double-stranded mtDNA molecule, but the number can vary considerably in different cell types.

During zygote formation, a sperm cell contributes its nuclear genome, but not its mitochondrial genome, to the egg cell. Consequently, the mitochondrial genome of the zygote is usually determined exclusively by that originally found in the unfertilized egg. The mitochondrial genome is therefore maternally inherited: males and females both inherit their mitochondria from their mother, but males do not transmit their mitochondria to subsequent generations. During mitotic cell division, the multiple mtDNA molecules in a dividing cell segregate in a purely random way to the two daughter cells.

Replication of mitochondrial DNA

The replication of both the H and L strands is unidirectional and starts at specific origins. Although the mitochondrial DNA is principally double-stranded, repeat synthesis of a small segment of the H-strand DNA produces a short third DNA strand called 7S DNA. The 7S DNA strand can base-pair with the L strand and displace the H strand, resulting in a triple-stranded structure (**Figure 9.2**). This region contains many of the mtDNA control sequences (including the major promoter regions) and so it is referred to as the *CR/D-loop region* (where CR denotes control region, and D-loop stands for displacement loop).

The origin of replication for the H strand lies in the CR/D-loop region, and that of the L strand is sandwiched between two tRNA genes (**Figure 9.3**). Only after about two-thirds of the daughter H strand has been synthesized (by using the L strand as a template and displacing the old H strand) does the origin for L-strand replication become exposed. Thereafter, replication of the L strand proceeds in the opposite direction, using the H strand as a template.

Mitochondrial genes and their transcription

The human mitochondrial genome contains 37 genes, 28 of which are encoded by the H strand and the other nine by the L strand (see Figure 9.3). Whereas nuclear genes often have their own dedicated promoters, the transcription of mitochondrial genes resembles that of bacterial genes. Transcription of mtDNA
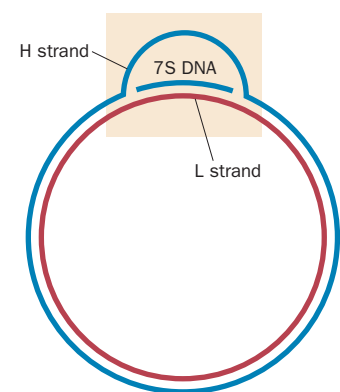

Figure 9.2 D-loop formation in mitochondrial DNA. The mitochondrial genome is not a simple double-stranded circular DNA. Repeat synthesis of a small segment of the H (heavy) strand results in a short third strand (7S DNA), which can base-pair with the L (light) strand and so displace the H strand to form a local triple-stranded structure that contains many important regulatory sequences and is known as the *CR/D-loop region* (shown by shading and in enlarged form for clarity).
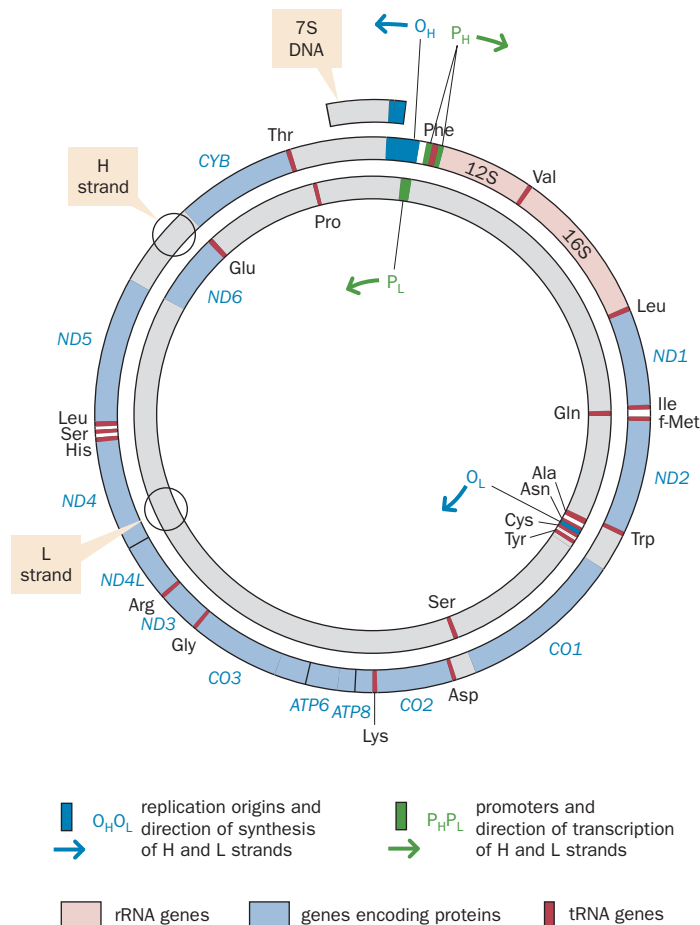
Figure 9.3 The organization of the human mitochondrial genome. The H strand is transcribed from two closely spaced promoter regions flanking the tRNA^{Phe} gene (grouped here as P_H); the L strand is transcribed from the P_L promoter in the opposite direction. In both cases, large primary transcripts are produced and cleaved to generate RNAs for individual genes. All genes lack introns and are closely clustered. The symbols for protein-coding genes are shown here without the prefix *MT-* that signifies mitochondrial gene. The genes that encode subunits 6 and 8 of the ATP synthase (*ATP6* and *ATP8*) are partly overlapping. Other polypeptide-encoding genes specify seven NADH dehydrogenase subunits (*ND4L* and *ND1*–*ND6*), three cytochrome *c* oxidase subunits (*CO1*–*CO3*), and cytochrome *b* (*CYB*). tRNA genes are represented with the name of the amino acid that they bind. The short 7S DNA strand is produced by repeat synthesis of a short segment of the H strand (see Figure 9.2). For further information, see the MITOMAP database at http://www.mitomap.org/.

starts from common promoters in the CR/D-loop region and continues round the circle (in opposing directions for the two different strands), to generate large multigenic transcripts. The mature RNAs are subsequently generated by cleavage of the multigenic transcripts.

Almost two-thirds (24 out of 37) of the mitochondrial genes specify a functional noncoding RNA as their final product. There are 22 tRNA genes, one for each of the 22 types of mitochondrial tRNA. In addition, two rRNA genes are dedicated to making 16S rRNA and 12S rRNA (components of the large and small subunits, respectively, of mitochondrial ribosomes). The remaining 13 genes encode polypeptides, which are synthesized on mitochondrial ribosomes. These 13 polypeptides form part of the mitochondrial respiratory complexes, the enzymes of oxidative phosphorylation that are engaged in the production of ATP. However, the great majority of the polypeptides that make up the mitochondrial oxidative phosphorylation system plus all other mitochondrial proteins are encoded by nuclear genes (**Table 9.1**). These proteins are translated on cytoplasmic ribosomes before being imported into the mitochondria.

Unlike its nuclear counterpart, the human mitochondrial genome is extremely compact: all 37 mitochondrial genes lack introns and are tightly packed (on average, there is one gene per 0.45 kb). The coding sequences of some genes (notably those encoding the sixth and eighth subunits of the mitochondrial ATP synthase) show some overlap (**Figure 9.4**) and, in most other cases, the coding sequences of neighboring genes are contiguous or separated by one or two noncoding bases. Some genes even lack termination codons; to overcome this deficiency, UAA codons have to be introduced at the post-transcriptional level (see Figure 9.4).

The mitochondrial genetic code

Prokaryotic genomes and the nuclear genomes of eukaryotes encode many hundreds to usually many thousands of different proteins. They are subject to a universal genetic code that is kept invariant: mutations that could potentially change

TABLE 9.1 THE LIMITED AUTONOMY OF THE MITOCHONDRIAL GENOME

Mitochondrial component	Encoded by	
	Mitochondrial genome	Nuclear genome
Components of oxidative phosphorylation system	**13 subunits**	**80 subunits**
I NADH dehydrogenase	7	42
II Succinate CoQ reductase	0	4
III Cytochrome b–c_1 complex	1	10
IV Cytochrome c oxidase complex	3	10
V ATP synthase complex	2	14
Components of protein synthesis apparatus	**24 RNAs**	**79 proteins**
rRNA	2	0
tRNA	22	0
Ribosomal proteins	0	79
Other mitochondrial proteins	**0**	**All[a]**

[a]Includes mitochondrial DNA and RNA polymerases plus numerous other enzymes, structural and transport proteins, etc.

the genetic code are likely to produce at least some critically misfunctional proteins and so are strongly selected against. However, the much smaller mitochondrial genomes make very few polypeptides. As a result, the mitochondrial genetic code has been able to drift by mutation to be slightly different from the universal genetic code.

In the mitochondrial genetic code there are 60 codons that specify amino acids, one fewer than in the nuclear genetic code. There are four stop codons: UAA and UAG (which also serve as stop codons in the nuclear genetic code) and AGA and AGG (which specify arginine in the nuclear genetic code; see Figure 1.25). The nuclear stop codon UGA encodes tryptophan in mitochondria, and AUA specifies methionine not isoleucine.

The mitochondrial genome specifies all the rRNA and tRNA molecules needed for synthesizing proteins on mitochondrial ribosomes, but it relies on nuclear-encoded genes to provide all other components, such as the protein components of mitochondrial ribosomes and aminoacyl tRNA synthetases. Because there are only 22 different types of human mitochondrial tRNA, individual tRNA molecules need to be able to interpret several different codons. This is possible because of *third-base wobble* in codon interpretation. Eight of the 22 tRNA molecules have anticodons that each recognize families of four codons differing only at the third base. The other 14 tRNAs recognize pairs of codons that are identical at the first two base positions and share either a purine or a pyrimidine at the third base. Between them, therefore, the 22 mitochondrial tRNA molecules can recognize a total of 60 codons [(8 × 4) + (14 × 2)].

Figure 9.4 The *MT-ATP6* and *MT-ATP8* genes are transcribed in different reading frames from overlapping segments of the mitochondrial H strand. *MT-ATP8* is transcribed from nucleotides 8366 to 8569, and *MT-ATP6* from 8527 to 9204. After transcription, the RNA encoding the ATP synthase 6 subunit is cleaved after position 9206 and polyadenylated, resulting in a C-terminal UAA codon where the first two nucleotides are derived ultimately from the TA at positions 9205–9206 and the third nucleotide is the first A of the poly(A) tail.
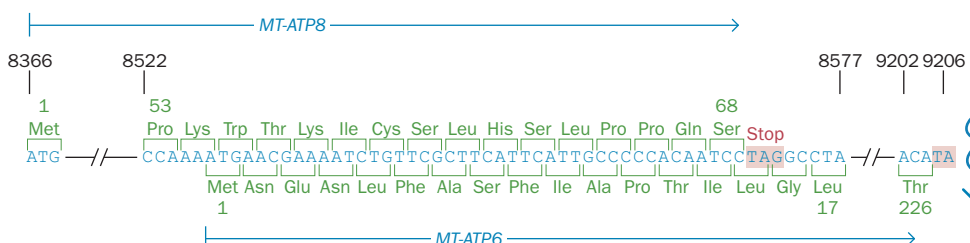
TABLE 9.2 THE HUMAN NUCLEAR AND MITOCHONDRIAL GENOMES		
	Nuclear genome	**Mitochondrial genome**
Size	3.1 Gb	16.6 kb
Number of different DNA molecules	23 (in XX cells) or 24 (in XY cells); all linear	one circular DNA molecule
Total number of DNA molecules per cell	varies according to ploidy; 46 in diploid cells	often several thousand copies (but copy number varies in different cell types)
Associated protein	several classes of histone and nonhistone protein	largely free of protein
Number of protein-coding genes	~21,000	13
Number of RNA genes	uncertain, but >6000	24
Gene density	~1/120 kb, but great uncertainty	1/0.45 kb
Repetitive DNA	more than 50% of genome; see Figure 9.1	very little
Transcription	genes are often independently transcribed	multigenic transcripts are produced from both the heavy and light strands
Introns	found in most genes	absent
Percentage of protein-coding DNA	~1.1%	~66%
Codon usage	61 amino acid codons plus three stop codons[a]	60 amino acid codons plus four stop codons[a]
Recombination	at least once for each pair of homologs at meiosis	not evident
Inheritance	Mendelian for X chromosome and autosomes; paternal for Y chromosome	exclusively maternal

[a]For details see Figure 1.25.

In addition to their differences in genetic capacity and different genetic codes, the mitochondrial and nuclear genomes differ in many other aspects of their organization and expression (**Table 9.2**).

The human nuclear genome consists of 24 widely different chromosomal DNA molecules

The human nuclear genome is 3.1 Gb (3100 Mb) in size. It is distributed between 24 different types of linear double-stranded DNA molecule, each of which has histones and nonhistone proteins bound to it, constituting a chromosome. There are 22 types of autosome and two sex chromosomes, X and Y. Human chromosomes can easily be differentiated by chromosome banding (see Figure 2.15), and have been classified into groups largely according to size and, to some extent, centromere position (see Table 2.3).

There is a single nuclear genome in sperm and egg cells and just two copies in most somatic cells, in contrast to the hundreds or even thousands of copies of the mitochondrial genome. Because the size of the nuclear genome is about 186,000 times the size of a mtDNA molecule, however, the nucleus of a human cell typically contains more than 99% of the DNA in the cell; the oocyte is a notable exception because it contains as many as 100,000 mtDNA molecules.

Not all of the human nuclear genome has been sequenced. The Human Genome Project focused primarily on sequencing *euchromatin*, the gene-rich, transcriptionally active regions of the nuclear genome that account for 2.9 Gb. The other 200 Mb is made up of permanently condensed and transcriptionally inactive (constitutive) heterochromatin. The heterochromatin is composed of long arrays of highly repetitive DNA that are very difficult to sequence accurately. For a similar reason, the long arrays of tandemly repeated transcription units encoding 28S, 18S, and 5.8S rRNA were also not sequenced.

The DNA of human chromosomes varies considerably in length and also in the proportions of underlying euchromatin and constitutive heterochromatin (**Table 9.3**). Each chromosome has some constitutive heterochromatin at the

TABLE 9.3 DNA CONTENT OF HUMAN CHROMOSOMES

Chromosome	Total DNA (Mb)	Euchromatin (Mb)	Heterochromatin (Mb)	Chromosome	Total DNA (Mb)	Euchromatin (Mb)	Heterochromatin (Mb)
1	249	224	19.5	13	115	96.3	17.2
2	243	240	2.9	14	107	88.3	17.2
3	198	197	1.5	15	103	82.1	18.3
4	191	188	3.0	16	90	79.0	10.0
5	181	178	0.3	17	81	78.7	7.5
6	171	168	2.3	18	78	74.6	1.4
7	159	156	4.6	19	59	60.8	0.3
8	146	143	2.2	20	63	60.6	1.8
9	141	120	18.0	21	48	34.2	11.6
10	136	133	2.5	22	51	35.1	14.3
11	135	131	4.8	X	155	151	3.0
12	134	131	4.3	Y	59	26.4	31.6

Chromosome sizes are taken from the ENSEMBL Human Map View (http://www.ensembl.org/Homo_sapiens/Location/Genome). Heterochromatin figures are estimates abstracted from International Human Genome Sequencing Consortium (2004) *Nature* 431, 931–945. The size of the total human genome is estimated to be about 3.1 Gb, with euchromatin accounting for close to 2.9 Gb and heterochromatin accounting for 200 Mb.

centromere. Certain chromosomes, notably 1, 9, 16, and 19, also have significant amounts of heterochromatin in the euchromatic region close to the centromere (*pericentromere*), and the acrocentric chromosomes each have two sizeable heterochromatic regions. But the most significant representation is in the Y chromosome, where most of the DNA is organized as heterochromatin.

The base composition of the euchromatic component of the human genome averages out at 41% (G+C), but there is considerable variation between chromosomes, from 38% (G+C) for chromosomes 4 and 13 up to 49% (for chromosome 19). It also varies considerably along the lengths of chromosomes. For example, the average (G+C) content on chromosome 17q is 50% for the distal 10.3 Mb but drops to 38% for the adjacent 3.9 Mb. There are regions of less than 300 kb with even wider swings, for example from 33.1% to 59.3% (G+C).

The proportion of some combinations of nucleotides can vary considerably. Like other vertebrate nuclear genomes, the human nuclear genome has a conspicuous shortage of the dinucleotide CpG. However, certain small regions of transcriptionally active DNA have the expected CpG density and, significantly, are unmethylated or hypomethylated (*CpG islands*; **Box 9.1**).

The human genome contains at least 26,000 genes, but the exact gene number is difficult to determine

Several years after the Human Genome Project delivered the first reference genome sequence, there is still very considerable uncertainty about the total human gene number. When the early analyses of the genome were reported in 2001, the gene catalog generated by the International Human Genome Sequencing Consortium was very much oriented toward protein-coding genes. Original estimates suggested more than 30,000 human protein-coding genes, most of which were gene predictions without any supportive experimental evidence. This number was an overestimate because of errors that were made in defining genes (see Box 8.5).

To validate gene predictions supportive evidence was sought, mostly by evolutionary comparisons. Comparison with other mammalian genomes, such as

---

BOX 9.1 ANIMAL DNA METHYLATION AND VERTEBRATE CpG ISLANDS

DNA methylation in multicellular animals often involves methylation of a proportion of cytosine residues, giving 5-methylcytosine (mC). In most animals (but not *Drosophila melanogaster*), the dinucleotide CpG is a common target for cytosine methylation by specific cytosine methyltransferases, forming mCpG (**Figure 1A**).

DNA methylation has important consequences for gene expression and allows particular gene expression patterns to be stably transmitted to daughter cells. It has also been implicated in systems of host defense against transposons. Vertebrates have the highest levels of 5-methylcytosine in the animal kingdom, and methylation is dispersed throughout vertebrate genomes. However, only a small percentage of cytosines are methylated (about 3% in human DNA, mostly as mCpG but with a small percentage as mCpNpG, where N is any nucleotide).

5-Methylcytosine is chemically unstable and is prone to deamination (see Figure 1A). Other deaminated bases produce derivatives that are identified as abnormal and are removed by the DNA repair machinery (e.g. unmethylated cytosine produces uracil when deaminated). However, 5-methyl cytosine is deaminated to give thymine, a natural base in DNA that is not recognized as being abnormal by cellular DNA repair systems. Over evolutionarily long periods, therefore, the number of CpG dinucleotides in vertebrate DNA has gradually fallen because of the slow but steady conversion of **C**pG to **T**pG (and to **C**pA on the complementary strand; Figure 1B).

Although the overall frequency of CpG in the vertebrate genome is low, there are small stretches of unmethylated or hypomethylated DNA that are characterized by having the *normal*, expected CpG frequency. Such islands of normal CpG density (**CpG islands**) are comparatively GC-rich (typically more than 50% GC) and extend over hundreds of nucleotides. CpG islands are gene markers because they are associated with transcriptionally active regions. Highly methylated DNA regions are prone to adopting a condensed chromatin conformation, but for actively transcribing DNA the chromatin needs to be in a more extended, open unmethylated conformation that allows various regulatory proteins to bind more readily to promoters and other gene control regions.
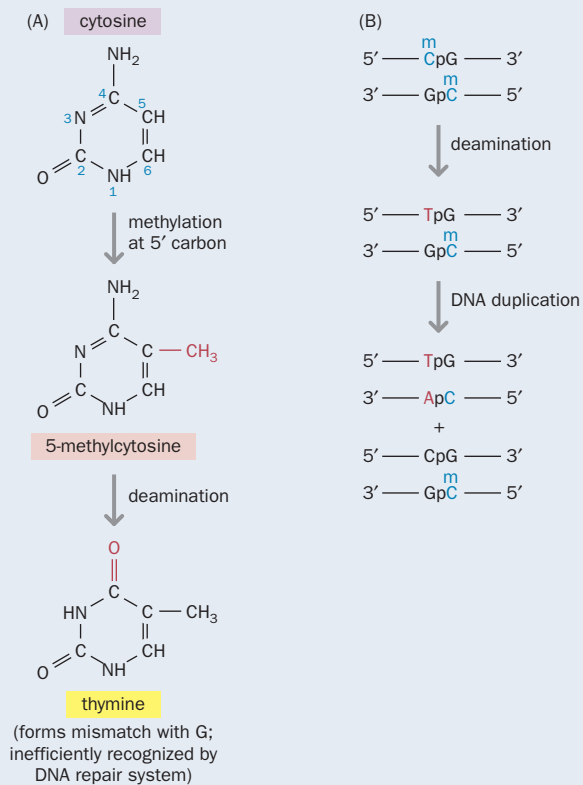

Figure 1 Instability of vertebrate CpG dinucleotides. (A) The cytosine in CpG dinucleotides is a target for methylation at the 5′ carbon atom. The resulting 5-methylcytosine is deaminated to give thymine (T), which is inefficiently recognized by the DNA repair system and so tends to persist (however, deamination of unmethylated cytosine gives uracil which is readily recognized by the DNA repair system). (B) The vertebrate CpG dinucleotide is gradually being replaced by TpG and CpA.

---

those of the mouse and the dog, failed to identify counterparts of many of the originally predicted human genes. By late 2009 the estimated number of human protein-coding genes appeared to be stabilizing somewhere around 20,000 to 21,000, but huge uncertainty remained about the number of human RNA genes. RNA genes are difficult to identify by using computer programs to analyze genome sequences: there are no open reading frames to screen for, and many RNA genes are very small and often not well conserved during evolution. There is also the problem of how to define an RNA gene. As we detail in Chapter 12, comprehensive analyses have recently suggested that the great majority of the genome—and probably at least 85% of nucleotides—is transcribed. It is currently unknown how much of the transcriptional activity is background noise and how much is functionally significant.

By mid-2009, evidence for at least 6000 human RNA genes had been obtained, including thousands of genes encoding long noncoding RNAs that are thought to be important in gene regulation. In addition, there is evidence for tens of thousands of different tiny human RNAs, but in many such cases quite large numbers of different tiny RNAs are obtained by the processing of single RNA transcripts. We look at noncoding RNAs in detail in Section 9.3.

The combination of about 20,000 protein-coding genes and at least 6000 RNA genes gives a total of at least 26,000 human genes. This remains a provisional total gene number; defining RNA genes is challenging and it will be some time before we obtain an accurate human gene number.

Human genes are unevenly distributed between and within chromosomes

Human genes are unevenly distributed on the nuclear DNA molecules. The constitutive heterochromatin regions are devoid of genes and, even within the euchromatic portion of the genome, gene density can vary substantially between chromosomal regions and also between whole chromosomes.

The first general insight into how genes are distributed across the human genome was obtained when purified CpG island fractions were hybridized to metaphase chromosomes. CpG islands have long been known to be strongly associated with genes (see Box 9.1). On this basis, it was concluded that gene density must be high in subtelomeric regions, and that some chromosomes (e.g. 19 and 22) are gene-rich whereas others (e.g. X and 18) are gene-poor (see Figure 8.17). The predictions of differential CpG island density and differential gene density were subsequently confirmed by analyzing the human genome sequence.

This difference in gene density can also be seen with Giemsa staining (G banding) of chromosomes. Regions with a low (G+C) content correlate with the darkest G bands, and those with a high (G+C) content with pale bands. GC-rich chromosomes (e.g. chromosome 19) and regions (e.g. pale G bands) are also comparatively rich in genes. For example, the gene-rich human leukocyte antigen (HLA) complex (180 protein-coding genes in a span of 4 Mb) is located within the pale 6p21.3 band. In striking contrast, the mammoth dystrophin gene extends over 2.4 Mb of DNA in a dark G band at Xp21.2 without evidence for any other protein-coding gene in this region.

Duplication of DNA segments has resulted in copy-number variation and gene families

Small genomes, such as those of bacteria and mitochondria, are typically tightly packed with genetic information that is presented in extremely economical forms. Large genomes, such as the nuclear genomes of eukaryotes, and especially vertebrate genomes, have the luxury of not being so constrained. Repetitive DNA is one striking feature of large genomes, in both abundance and importance.

Different types of DNA sequence can be repeated. Some are short noncoding sequences that are present in a few copies to millions of copies. These will be discussed further in Section 9.4. Many others are moderately long to large DNA sequences that often contain genes or parts of genes. Such duplicated sequences are prone to various genetic mechanisms that result in *copy-number variation* (*CNV*) in which the number of copies of specific moderately long sequences—often from many kilobases to several megabases long—varies between different haplotypes. Copy-number variation generates a type of *structural variation* that we consider more fully in Chapter 13, but we will consider some of the mechanisms below in the context of how genes become duplicated. It is clear, however, that CNV is quite extensive in the human genome. For example, when James Watson's genome was sequenced, 1.4% of the total sequencing data obtained did not map with the reference human genome sequence. As personal genome sequencing accelerates, new CNV regions are being identified with important implications for gene expression and disease.

Repeated duplication of a gene-containing sequence gives rise to a **gene family**. As we will see in Sections 9.2 and 9.3, many human genes are members of multigene families that can vary enormously in terms of copy number and distribution. They arise by one or more of a variety of different mechanisms that result in gene duplication. Gene families may also contain evolutionarily related sequences that may no longer function as working genes (*pseudogenes*).

Gene duplication mechanisms

Gene duplication has been a common event in the evolution of the large nuclear genomes found in complex eukaryotes. The resulting multigene families have from two to very many gene copies. The gene copies may be clustered together in one subchromosomal location or they may be dispersed over several chromosomal locations. Several different types of gene duplication can occur:

- *Tandem gene duplication* typically arises by crossover between unequally aligned chromatids, either on homologous chromosomes (*unequal crossover*) or on the same chromosome (*unequal sister chromatid exchange*). **Figure 9.5** shows the general mechanism. The repeated segment may be just a few kilobases long or may be quite large and contain from one to several genes. Two such repeats are said to be *direct repeats* if the head of one repeat joins the tail of its neighbor (→→) or *inverted repeats* if there is joining of heads (→←) or tails (←→). Over a long (evolutionary) time-scale the duplicated sequences can be separated on the same chromosome (by a DNA insertion or inversion) or become distributed on different chromosomes by translocations.

- *Duplicative transposition* describes the process by which a duplicated DNA copy integrates into a new subchromosomal location. Typically this involves *retrotransposition*: cellular reverse transcriptases make a cDNA copy of an RNA transcript, whereupon the cDNA copy integrates into a new chromosomal location. The same type of mechanism can often lead to defective gene copies and will be detailed in Section 9.2.

- *Gene duplication by ancestral cell fusion.* Aerobic eukaryotic cells are thought to have evolved through the endocytosis of a type of bacterial cell by a eukaryotic precursor cell. The current mitochondrial genome is thought to have derived from the bacterial cell's genome but is now a very small remnant because many of the original bacterial genes were subsequently excised and transferred to what is now the nuclear genome. As a result, the nuclear genome contains duplicated genes that encode cytoplasm-specific and mitochondrion-specific isoforms for certain enzymes and certain other key proteins.

- *Large-scale subgenomic duplications* can arise as a result of chromosome translocations. Euchromatic regions close to human centromeres and to telomeres (pericentromeric and subtelomeric regions, respectively) are comparatively unstable and are prone to recombination with other chromosomes. As a result, large segments of DNA containing multiple genes have been duplicated. Within the past 40 million years of primate evolution, this process has led to the duplication of about 400 large (several megabases long) DNA segments, accounting for more than 5% of the euchromatic genome. This type of duplication, known as **segmental duplication**, results in very high (often more than 95%) sequence identity between the DNA copies and can involve both intrachromosomal duplications and also interchromosomal duplications (**Figure 9.6**). Segmental duplications are important contributors to copy-number variation and to chromosomal rearrangements leading to disease and rapid gene innovation. We consider how they originate in Chapter 10.

- *Whole genome duplication.* It is now clear from comparative genomics studies that whole genome duplication has occurred at several times during evolution on a variety of different eukaryotic lineages. For example, there is compelling evidence that whole genome duplication occurred in the early evolution of chordates. This type of event could explain why vertebrates have four HOX clusters (see Figure 5.5). Whole genome duplication is detailed in Chapter 10 in the context of genome evolution.
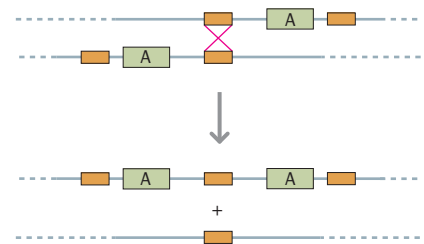

Figure 9.5 Tandem gene duplication. Crossover between misaligned chromatids can result in one chromosome with a tandem duplication of a sequence containing a gene (such as gene A shown here by a green box) and one in which the gene is lost. The mispairing of chromatids may be stabilized by closely related members of an interspersed repeated DNA family such as Alu repeats (as shown here by orange boxes). The crossover event leading to tandem gene duplication may result from unequal crossover (crossover between misaligned chromatids on homologous chromosomes) or unequal sister chromatid exchange (the analogous process by which sister chromatids are misaligned; see Figure 13.3 for an illustration).
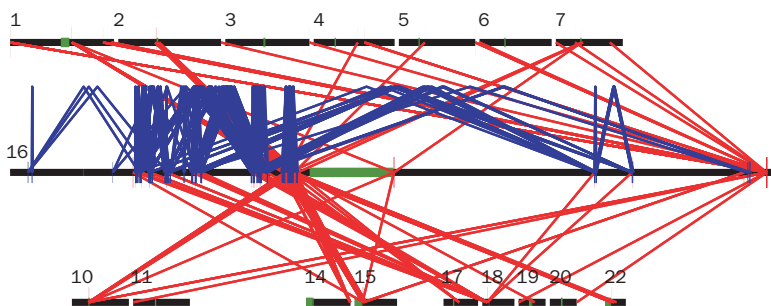

Figure 9.6 Segmental duplication. The horizontal bar in the center is a linear map of the DNA of human chromosome 16 (the central green segment represents heterochromatin). The black horizontal bars at the top and bottom represent linear maps of 16 other chromosomes containing large segments that are shared with chromosome 16, with red connecting lines marking the positions of homologous sequences. Intrachromosomal duplications are shown by blue chevrons (^) linking the positions of large duplicated sequences on chromosome 16. [From Martin J, Han C, Gordon LA et al. (2004) *Nature* 432, 988–994. With permission from Macmillan Publishers Ltd. For chromosomal coordinates and further information on segmental duplications in the human genome, dedicated segmental duplication databases can be accessed at http://humanparalogy.gs.washington.edu/ and http://projects.tcag.ca/humandup/]

9.2   PROTEIN-CODING GENES

For many years, molecular geneticists believed that the major functional end-point of DNA was protein. Studies of prokaryotic genomes supported this belief, partly because these genomes are rich in protein-coding DNA. It came as a surprise to find that the much larger genomes of complex eukaryotes have comparatively little protein-coding DNA. For example, protein-coding DNA sequences account for close to 90% of the *E. coli* genome but just 1.1% of the human genome.

Human protein-coding genes show enormous variation in size and internal organization

Size variation

Genes in simple organisms such as bacteria are comparatively similar in size and are usually very short (typically about 1 kb long). In complex eukaryotes, genes can show huge variation in size. Although there is generally a direct correlation between gene and product sizes, there are some striking anomalies. For example, the giant 2.4 Mb dystrophin gene is more than 50 times the size of the apolipoprotein B gene but the dystrophin protein has a linear length (total amino acid number) that is about 80% of that of apolipoprotein B (**Table 9.4**).

A small minority of human protein-coding genes lack introns and are generally small (see note to Table 9.4 for some examples). For those that do possess

TABLE 9.4 STRUCTURAL VARIATION IN SIZE AND ORGANIZATION OF HUMAN PROTEIN-CODING GENES						
Human protein	Size of protein (no. of amino acids)	Size of gene (kb)	No. of exons	Coding DNA (%)	Average size of exon (bp)	Average size of intron (bp)
SRY	204	0.9	1	94	850	–
β-Globin	146	1.6	3	38	150	490
p16	156	7.4	3	17	406	3064
Serum albumin	609	18	14	12	137	1100
Type VII collagen	2928	31	118	29	77	190
p53	393	39	10	6.0	236	3076
Complement C3	1641	41	29	8.6	122	900
Apolipoprotein B	4563	45	29	31	487	1103
Phenylalanine hydroxylase	452	90	26	3	96	3500
Factor VIII	2351	186	26	3	375	7100
Huntingtin	3144	189	67	8.0	201	2361
RB1 retinoblastoma protein	928	198	27	2.4	179	6668
CFTR (cystic fibrosis transmembrane receptor)	1480	250	27	2.4	227	9100
Titin	34,350	283	363	40	315	466
Utrophin	3433	567	74	2.2	168	7464
Dystrophin	3685	2400	79	0.6	180	30,770

Where isoforms are evident, the given figures represent the largest isoforms. As genes get larger, exon size remains fairly constant but intron sizes can become very large. Internal exons tend to be fairly uniform in size, but the terminal exon or some exons near the 3′ end can be many kilobases long; for example, exon 26 of the *APOB* gene is 7.5 kb long. Note the extraordinarily high exon content and comparatively small intron sizes in the genes encoding type VII collagen and titin. In addition to *SRY*, other single-exon protein-coding genes in the nuclear genome include retrogenes (see Table 9.8) and genes encoding other SOX proteins, interferons, histones, many G-protein-coupled receptors, heat shock proteins, many ribonucleases, and various neurotransmitter receptors and hormone receptors.

introns, there is an inverse correlation between gene size and fraction of coding DNA (see Table 9.4). This does not arise because exons in large genes are smaller than those in small genes. The average exon size in human genes is close to 300 bp, and exon size is comparatively independent of gene length. Instead, there is huge variation in intron lengths, and large genes tend to have very large introns (see Table 9.4). Transcription of long introns is, however, costly in time and energy; transcription of the 2.4 Mb dystrophin gene takes about 16 hours. Thus, very highly expressed genes often have short introns or no introns at all.

Repetitive sequences within coding DNA

Highly repetitive DNA sequences are often found within introns and flanking sequences of genes. They will be detailed in Section 9.4. In addition, repetitive DNA sequences are found to different extents in exons. Tandem repetition of very short oligonucleotide sequences (1–4 bp) is frequent and may simply reflect statistically expected frequencies for certain base compositions. Tandem repetition of sequences encoding known or assumed protein domains is also quite common, and it may be functionally advantageous by providing a more available biological target.

The sequence identities between the repeated protein domains are often quite low but can sometimes be high. Lipoprotein Lp (a), encoded by the *LPA* gene on chromosome 6q26, provides a classical example. It contains multiple tandemly repeated kringle domains, which are each about 114 amino acids long and form disulfide-bonded loops. The different kringle domains are often nearly identical in amino acid sequence. Even at the nucleotide sequence level the DNA repeats that encode the kringle domains show very high levels of sequence identity, making them prone to unequal crossover. As a result, the *LPA* gene is subject to length polymorphism, and the number of kringle domains in lipoprotein Lp (a) varies but is usually 15 or more.

Different proteins can be specified by overlapping transcription units

Overlapping genes and genes-within-genes

Simple genomes have high gene densities (roughly one per 0.5, 1, and 2 kb for the genomes of human mitochondria, *Escherichia coli*, and *Saccharomyces cerevisiae*, respectively) and often show examples of partly overlapping genes. Different reading frames may be used, sometimes from the same sense strand. In complex organisms, such as humans, genes are much bigger, and there is less clustering of protein-coding sequences (**Table 9.5**).

Gene density varies enormously from chromosome to chromosome and within different regions of the same chromosome. In chromosomal regions with high gene density, overlapping genes may be found; they are typically transcribed from opposing DNA strands. For example, the class III region of the HLA complex at 6p21.3 has an average gene density of about one gene per 15 kb and is known to contain several examples of partly overlapping genes (**Figure 9.7A**).

Whole genome analyses show that about 9% of human protein-coding genes overlap another such gene. More than 90% of the overlaps involve genes transcribed from opposing strands. Sometimes the overlaps are partial, but in other cases small protein-coding genes are located within the introns of larger genes. The *NF1* (neurofibromatosis type I) gene, for example, has three small internal genes transcribed from the opposite strand (Figure 9.7B).

Recent analyses have also shown that RNA genes can frequently overlap protein-coding genes. The positioning of RNA genes will be covered in Section 9.3.

Genes divergently transcribed or co-transcribed from a common promoter

Some protein-coding genes share a promoter. In many cases the 5′ ends of the two genes are often separated by just a few hundred nucleotides and the genes are transcribed in opposite directions from the common promoter. This type of bidirectional gene organization may provide for common regulation of the gene pair.

Alternatively, genes with a common promoter are transcribed in the same direction to produce multigenic transcripts that are then cleaved to produce a

TABLE 9.5 HUMAN GENOME AND HUMAN GENE STATISTICS

SIZE OF GENOME COMPONENTS	
Mitochondrial genome	16.6 kb
Nuclear genome	3.1 Gb[a]
Euchromatic component	2.9 Gb (~93%)
Highly conserved fraction	~150 Mb (~5%)
Protein-coding DNA sequences	~35 Mb (~1.1%)
Other highly conserved DNA	~115 Mb (~3.9%)
Segmentally duplicated DNA	~160 Mb (~5.5%)
Highly repetitive DNA	~1.6 Gb (~50%)
Constitutive heterochromatin	~ 200 Mb (~7%; Table 9.3)
Transposon-based repeats	~1.4 Gb (~45%; Table 9.12)
DNA per chromosome	48 Mb—249 Mb (Table 9.3)
GENE NUMBER	
Nuclear genome	> 26,000
Mitochondrial genome	37
Protein-coding genes	~ 20,000–21,000
RNA genes	> 6000 (exact figure not known)
Pseudogenes related to protein-coding genes	> 12,000
GENE DENSITY	
Nuclear genome	>1 per 120 kb (but considerable uncertainty)
Mitochondrial genome	1 per 0.45 kb
LENGTH OF PROTEIN-CODING GENES	
Average length	53.6 kb
Smallest	a few hundred base pairs long (several examples)
Largest	2.4 Mb (dystrophin)
EXON NUMBER IN PROTEIN-CODING GENES	
Average number of exons in one gene[b]	9.8
Largest number in one gene	363 (in the titin gene)
Smallest number in one gene	1 (no introns—see Tables 9.4 and 9.7 for example)
EXON SIZE IN PROTEIN-CODING GENES	
Average exon size	288 bp (but exons at 3′ end of genes tend to be large)
Smallest	< 10 bp (various; e.g. exon 3 of the troponin I gene *TNNI1* is just 4 bp long)
Largest	18.2 kb (exon 6 in *MUC16* isoform-201)
INTRON SIZE IN PROTEIN-CODING GENES	
Smallest	< 30 bp (various)
Largest	1.1 Mb (intron 5 in *KCNIP4*)
RNA SIZE	
Smallest noncoding RNA	< 20 nucleotides (e.g. many transcriptional start site-associated RNAs are 18 nucleotides)
Largest noncoding RNA	Several hundred thousand nucleotides; e.g. *UBE3A* antisense RNA is likely to be close to 1 Mb
Largest mRNA	> 103 kb (titin mRNA, NF-2A isoform)
POLYPEPTIDE SIZE	
Smallest	tens of amino acids (various neuropeptides)
Largest	34,350 amino acids (titin, NF-2A isoform)

[a]Note that the total size can vary between haplotypes because of copy-number variation. [b]For the longest isoform. Data were obtained largely from ENSEMBL release 55 datasets.
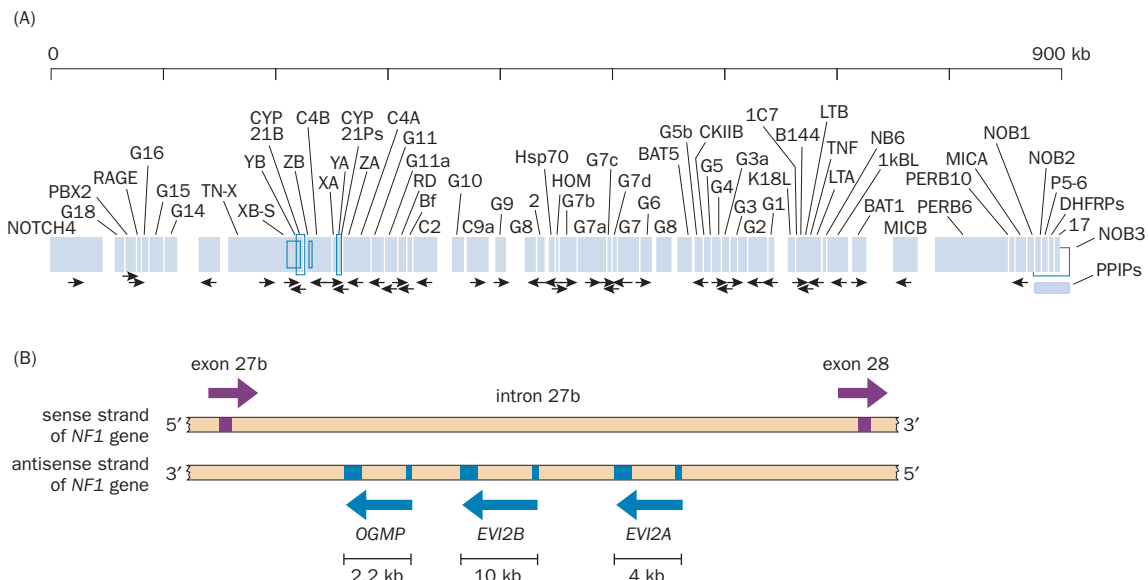
Figure 9.7 Overlapping genes and genes-within-genes. (A) Genes in the class III region of the HLA complex are tightly packed and overlapping in some cases. Arrows show the direction of transcription. (B) Intron 27b of the *NF1* (neurofibromatosis type I) gene is 60.5 kb long and contains three small internal genes, each with two exons, which are transcribed from the opposing strand. The internal genes (not drawn to scale) are *OGMP* (oligodendrocyte myelin glycoprotein) and *EVI2A* and *EVI2B* (human homologs of murine genes thought to be involved in leukemogenesis and located at ecotropic viral integration sites).

separate transcript for each gene. Such genes are said to form part of a *polycistronic* (= multigenic) transcription unit. Polycistronic transcription units are common in simple genomes such as those of bacteria and the mitochondrial genome (see Figure 9.3). Within the nuclear genome, some examples are known of different proteins being produced from a common transcription unit. Typically, they are produced by cleavage of a hybrid precursor protein that is translated from a common transcript. The A and B chains of insulin, which are intimately related functionally, are produced in this way (see Figure 1.26), as are the related peptide hormones somatostatin and neuronostatin. Sometimes, however, functionally distinct proteins are produced from a common protein precursor. The *UBA52* and *UBA80* genes, for example, both generate ubiquitin and an unrelated ribosomal protein (S27a and L40, respectively).

More recent analyses have shown that the long-standing idea that most human genes are independent transcription units is not true, and so the definition of a gene will need to be radically revised. Multigenic transcription is now known to be rather frequent in the human genome, and specific proteins and functional noncoding RNAs can be made by common RNA precursors. This will be explored further in Section 9.3.

Human protein-coding genes often belong to families of genes that may be clustered or dispersed on multiple chromosomes

Duplicated genes and duplicated coding sequence components are a common feature of animal genomes, especially large vertebrate genomes. As we will see in Chapter 10, gene duplication has been an important driver in the evolution of functional complexity and the origin of increasingly complex organisms. Genes that operate in the same or similar functional pathways but produce proteins with little evidence of sequence similarity are distantly related in evolution, and they tend to be dispersed at different chromosomal locations. Examples include genes encoding insulin (on chromosome 11p) and the insulin receptor (19p); ferritin heavy chain (11q) and ferritin light chain (22q); steroid 11-hydroxylase (8q) and steroid 21-hydroxylase (6p); and JAK1 (1p) and STAT1 (2q). However, genes that produce proteins with both structural and functional similarity are often organized in gene clusters.
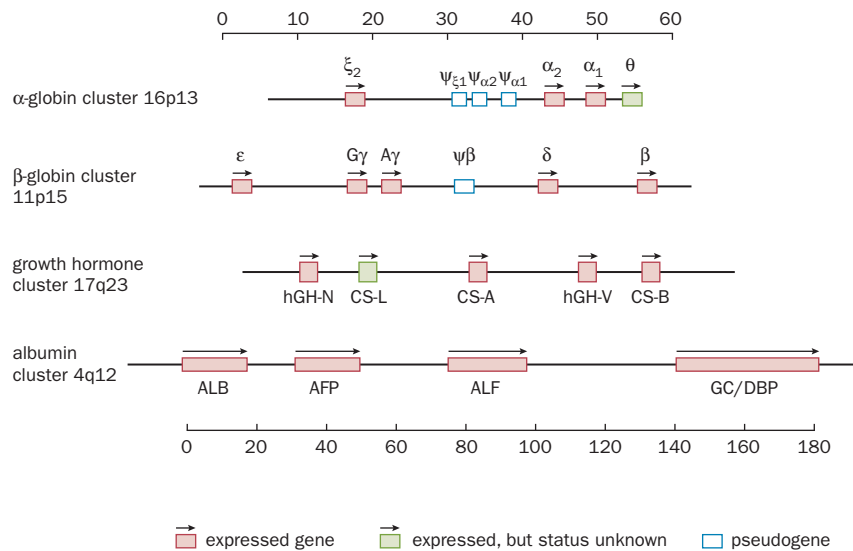
Different classes of human gene families can be recognized according to the degree of sequence similarity and structural similarity of their protein products. If two different genes make very similar protein products, they are most likely to have originated by an evolutionarily very recent gene duplication, most probably some kind of tandem gene duplication event, and they tend to be clustered together at a specific subchromosomal location. If they make proteins that are more distantly related in sequence, they most probably arose by a more ancient gene duplication. They may originally have been clustered together, but over long evolutionary time-scales the genes could have been separated by translocations or inversions, and they tend to be located at different chromosomal locations.

Some gene families are organized in multiple clusters. The β-, γ-, δ-, and ε-globin genes are located in a gene cluster on 11p and are more closely related to each other than they are to the genes in the α-globin gene cluster on 16p (**Figure 9.8**). The genes in the β-globin gene cluster on 11p originated by gene duplication events that were much more recent in evolution than the early gene duplication event that gave rise to ancestors of the α- and β-globin genes. An outstanding example of a gene family organized as multiple gene clusters is the olfactory receptor gene family. The genes encode a diverse repertoire of receptors that allow us to discriminate thousands of different odors; the genes are located in large clusters at multiple different chromosomal locations (**Table 9.6**).

Some gene families have individual gene copies at two or more chromosomal locations without gene clustering (see Table 9.6). The genes at the different locations are usually quite divergent in sequence unless gene duplication occurred relatively recently or there has been considerable selection pressure to maintain sequence conservation. The family members are expected to have originated from ancient gene duplications.

Different classes of gene family can be recognized according to the extent of sequence and structural similarity of the protein products

As listed below, various classes of gene family can be distinguished according to the level of sequence identity between the individual gene members.

- In gene families with closely related members, the genes have a high degree of sequence homology over most of the length of the gene or coding sequence. Examples include histone gene families (histones are strongly conserved, and subfamily members are virtually identical), and the α-globin and β-globin gene families.

TABLE 9.6 EXAMPLES OF CLUSTERED AND INTERSPERSED MULTIGENE FAMILIES			
Family	Copy no.	Organization	Chromosome location(s)
CLUSTERED GENE FAMILIES			
Growth hormone gene cluster	5	clustered within 67 kb; one pseudogene (Figure 9.8)	17q24
α-Globin gene cluster	7	clustered over ~50 kb (Figure 9.8)	16p13
Class I HLA heavy chain genes	~20	clustered over 2 Mb (Figure 9.10)	6p21
HOX genes	38	organized in four clusters (Figure 5.5)	2q31, 7p15, 12q13, 17q21
Histone gene family	61	modest-sized clusters at a few locations; two large clusters on chromosome 6	many
Olfactory receptor gene family	> 900	about 25 large clusters scattered throughout the genome	many
INTERSPERSED GENE FAMILIES			
Aldolase	5	three functional genes and two pseudogenes on five different chromosomes	many
PAX	9	all nine are functional genes	many
NF1 (neurofibromatosis type I)	> 12	one functional gene at 22q11; others are nonprocessed pseudogenes or gene fragments (Figure 9.11)	many, mostly pericentromeric
Ferritin heavy chain	20	one functional gene on chromosome 11; most are processed pseudogenes	many

- In gene families defined by a common protein domain, the members may have very low sequence homology but they possess certain sequences that specify one or more specific protein domains. Examples include the PAX gene family and SOX gene family (**Table 9.7**).

- Examples of gene families defined by functionally similar short protein motifs are families of genes that encode functionally related proteins with a DEAD box motif (Asp-Glu-Ala-Asp) or the WD repeat (**Figure 9.9**).

Some genes encode products that are functionally related in a general sense but show only very weak sequence homology over a large segment, without very significant conserved amino acid motifs. Nevertheless, there may be some evidence for common general structural features. Such genes can be grouped into an evolutionarily ancient **gene superfamily** with very many gene members. Because multiple different gene duplication events have occurred periodically during the long evolution of a gene superfamily, some of the gene members make proteins that are very divergent in sequence from those of some other family members, but genes resulting from more recent duplications are more readily seen to be related in sequence.

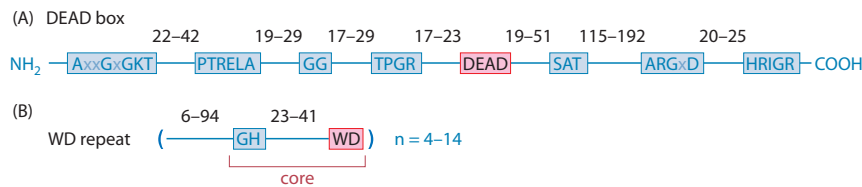| TABLE 9.7 EXAMPLES OF HUMAN GENES WITH SEQUENCE MOTIFS THAT ENCODE HIGHLY CONSERVED DOMAINS		
Gene family	Number of genes	Sequence motif/domain
Homeobox genes	38 *HOX* genes plus 197 orphan homeobox genes	homeobox specifies a homeodomain of ~60 amino acids; a wide variety of different subclasses have been defined
PAX genes	9	paired box encodes a paired domain of ~124 amino acids; *PAX* genes often also have a type of homeodomain known as a paired-type homeodomain
SOX genes	19	SRY-like HMG box which encodes a domain of 70–80 amino acids
TBX genes	14	T-Box encodes a domain of ~170 amino acids
Forkhead domain genes	50	the forkhead domain is ~110 amino acids long
POU domain genes	16	the POU domain is ~150 amino acids long

(A) DEAD box

22–42   19–29 17–29   17–23   19–51   115–192   20–25

NH₂ —[AxxGxGKT]—[PTRELA]—[GG]—[TPGR]—[DEAD]—[SAT]—[ARGxD]—[HRIGR]— COOH

(B)
6–94   23–41
WD repeat   (—[GH]—[WD])   n = 4–14
core

Two important examples of gene superfamilies are the Ig (immunoglobulin) and GPCR (G-protein-coupled receptor) superfamilies. Members of the Ig super-family all have globular domains resembling those found in immunoglobulins, and in addition to immunoglobulins they include a variety of cell surface proteins and soluble proteins involved in the recognition, binding, or adhesion processes of cells (see Figure 4.22 for some examples). The GPCR superfamily is very large, with at least 799 unique full-length members distributed throughout the human genome. All the GPCR proteins have a common structure of seven α-helix transmembrane segments, but they typically have low (less than 40%) sequence similarity to each other. They mediate ligand-induced cell signaling via interaction with intracellular G proteins, and most work as rhodopsin receptors.

Gene duplication events that give rise to multigene families also create pseudogenes and gene fragments

Gene families frequently have defective gene copies in addition to functional genes. A defective gene copy that contains at least multiple exons of a functional gene is known as a **pseudogene** (**Box 9.2**). Other defective gene copies may have only limited parts of the gene sequence, sometimes a single exon, and so are sometimes described as *gene fragments*.

Clustered gene families often have defective gene copies that have arisen by tandem duplication. These are examples of *nonprocessed pseudogenes*. Copying can be seen to have been performed at the level of genomic DNA because non-processed pseudogenes contain counterparts of both exons and introns and sometimes also of upstream promoter regions. However, even if the copy has sequences that correspond to the full length of the functional gene, closer examination will identify inappropriate termination codons in exons, aberrant splice junctions, and so on. Classical examples of nonprocessed pseudogenes are found in the α-globin and β-globin gene clusters (see Figure 9.8). Sometimes, smaller truncated gene copies and gene fragment copies are also evident, as in the class I HLA gene family (**Figure 9.10**).

A few gene families that are distributed at different chromosomal locations can also have nonprocessed pseudogene copies of a single functional gene. Certain types of subchromosomal region, notably pericentromeric and subtelomeric regions, are comparatively unstable. They are prone to recombination events that can result in duplicated gene segments (containing both exons and introns) being distributed to other chromosomal locations. The gene copies are typically defective because they lack some of the functional gene sequence. Two illustrative examples are sequences related to the *NF1* (neurofibromatosis type I) and the *PKD1* (adult polycystic kidney disease) genes.

The *NF1* gene is located at 17q11.2. Because of pericentromeric instability, multiple nonprocessed *NF1* pseudogene/gene fragment copies are distributed over seven different chromosomes, nine being located at pericentromeric regions (**Figure 9.11A**). The *PKD1* gene has over 46 exons spanning 50 kb and is located in a subtelomeric region at 16p13.3. Six nonprocessed *PKD1* pseudogenes have been generated by segmental duplications during primate evolution and have inserted into locations within a region that is about 13–16 Mb proximal to the *PKD1* gene, corresponding to part of band 16p13.1 (Figure 9.11B). The pseudogenes lack sequences at the 3′ end of the *PKD1* gene but have sequence counterparts of much of the genomic sequence spanning exons 1–32, showing 97.6% to 97.8% sequence identity to the PKD1 sequence.

Processed pseudogenes are defective copies of a gene that contain only exonic sequences and lack an intronic sequence or upstream promoter sequences. They arise by *retrotransposition*: cellular reverse transcriptases can use processed gene

BOX 9.2 THE ORIGINS, PREVALENCE, AND FUNCTIONALITY OF PSEUDOGENES

Pseudogenes are usually thought of as defective copies of a functional gene to which they show significant sequence homology. They typically arise by some kind of gene duplication event that produces two gene copies. Selection pressure to conserve gene function need only be imposed on one gene copy; the other copy can be allowed to mutate more freely (*genetic drift*) and can pick up inactivating mutations, producing a pseudogene. However, some sequences are referred to as pseudogenes even though they have not originated by DNA copying. For example, as we will see in Chapter 10, humans have rare *solitary pseudogenes* that are clearly orthologs of functional genes in the great apes and became defective after acquiring harmful mutations in the human lineage.

Different gene duplication mechanisms can give rise to multiple functional gene copies and defective pseudogenes. Either the genomic DNA sequence is copied, or a cDNA copy is made (after reverse transcription of a processed RNA transcript) that integrates into genomic DNA. For a protein-coding gene, copying at the genomic DNA level can result in duplication of the promoter and upstream regulatory sequences as well as of all exons and introns. A defective gene that derives from a copy of a genomic DNA sequence is known as a *nonprocessed pseudogene* (**Figure 1A**). Such pseudogenes usually arise by tandem duplication so that they are located close to functional gene counterparts (see Figures 9.8 and 9.10B for examples), but some are dispersed as a result of recombination (see Figure 9.11 for examples).

Copying at the cDNA level produces a gene copy that typically lacks introns, promoter elements, and upstream regulatory elements. Very occasionally, a processed gene copy can retain some function (a *retrogene*; see Table 9.7). However, because they lack important sequences needed for expression, most processed gene copies degenerate into *processed pseudogenes* (sometimes called *retrotransposed pseudogenes*; Figure 1B).

Prevalence and functionality of pseudogenes

Eukaryotic genomes typically have many pseudogenes. A long-standing rationale for their abundance is that gene duplication is evolutionarily advantageous. New functional gene variants can be created by gene duplication, and pseudogenes have long been viewed as unsuccessful by-products of the duplication mechanisms. Although some prokaryotic genomes seem to have many pseudogenes, pseudogenes are generally rare in prokaryotes because their genomes are generally designed to be compact.

The great majority of what are conventionally recognized as human pseudogenes are copies of protein-coding genes simply because it is relatively easy to identify them (by looking for frameshifting, splice site mutations, and so on). There are more than 8000 different processed pseudogene copies of protein-coding genes in the human genome, plus more than 4000 nonprocessed pseudogenes (see the pseudogene database at

http://www.pseudogene.org). Only about 10% of the 21,000 human protein-coding genes have at least one processed pseudogene, but highly expressed genes tend to have multiple processed pseudogenes. For example, the cytoplasmic ribosomal protein contains 95 functional genes encoding 79 different proteins (16 genes are duplicated) and 2090 processed pseudogenes.

RNA pseudogenes are often difficult to identify as pseudogenes (there is no reading frame to inspect, and RNA genes often lack introns). Nevertheless, pseudogene copies of many small RNAs are common (see the table below), notably if they are transcribed by RNA polymerase III (genes transcribed by RNA polymerase III often have internal promoters).

RNA family	Number of human genes	Number of related pseudogenes
U6 snRNA	49	~800
U7 snRNA	1	85
Y RNA	4	~1000

As will be described in Section 12.4, the Alu repeat, the most abundant sequence in the human genome, seems to have originated by the copying of 7SL RNA transcripts, and many other highly repeated interspersed DNA families in mammals are copies of tRNA. So, in a sense, RNA pseudogenes have come to be the most common sequence elements of mammalian genomes.

All the pseudogenes are located in the nuclear genome, but they do include defective copies of genes that reside in the mitochondrial genome (*mitochondrial pseudogenes*). The mitochondrial genome originated from a much larger bacterial genome and over a long evolutionary time-scale much of the DNA of the large precursor mitochondrial genome migrated in a series of independent integration events into what is now the nuclear genome. mtDNA pseudogenes now account for at least 0.016% of nuclear DNA (or about 30 times the content of the mitochondrial genome).

The functionality of pseudogenes has been an enduring debate, and different pseudogene classes have been envisaged. A significant number of pseudogenes (mostly processed pseudogenes) are transcribed, and antisense pseudogene transcripts may regulate parent genes. Pseudogenes have also been directly implicated in the production of endogenous siRNAs that regulate transposons, as described in Section 9.3. Finally, some pseudogene sequences may be co-opted for a different function. They have been described as *exapted pseudogenes*. An example is provided by the *XIST* gene. It makes a noncoding RNA that regulates X-chromosome inactivation, and two of its six exons are known to have originated from a pseudogene copy of a protein-coding gene.

Figure 1 Origins of nonprocessed and processed pseudogenes. (A) Copying of genomic DNA sequence containing gene A can produce duplicate copies of gene A. Strong selection pressure needs to be applied to one of the copies to maintain gene function (bold arrow), but the other copy can be allowed to mutate (dashed arrow). If it picks up inactivating mutations (red circles), a nonprocessed pseudogene (ψA) can arise. (B) A processed pseudogene arises after cellular reverse transcriptases convert a transcript of a gene into a cDNA that then is able to integrate back into the genome (see Figure 9.12 for details). The lack of important sequences such as a promoter usually results in an inactive gene copy.
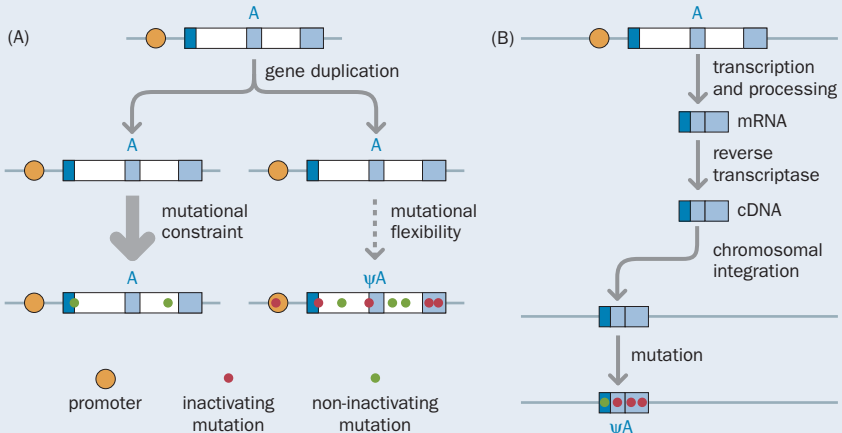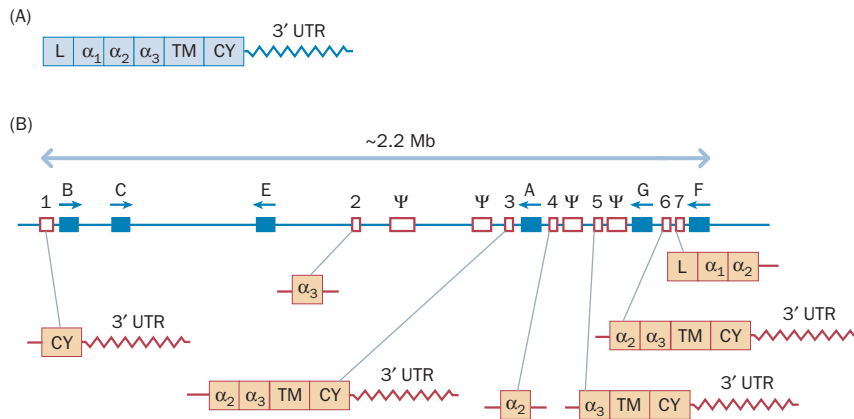
(A) Structure of a class I HLA heavy-chain mRNA. The full-length mRNA contains a polypeptide-encoding sequence with a leader sequence (L), three extracellular domains (α_1, α_2, and α_3), a transmembrane sequence (TM), a cytoplasmic tail (CY), and a 3′ untranslated region (3′ UTR). The three extracellular domains are each encoded essentially by a single exon. The very small 5′ UTR is not shown. (B) The class I HLA heavy chain gene cluster is located at 6p21.3 and comprises about 20 genes. They include six expressed genes (filled blue boxes), four full-length nonprocessed pseudogenes (long red open boxes labeled ψ), and a variety of partial gene copies (short open red boxes labeled 1–7). Some of the latter are truncated at the 5′ end (e.g. 1, 3, 5, and 6), some are truncated at the 3′ end (e.g. 7), and some contain single exons (e.g. 2 and 4).

transcripts such as mRNA to make cDNA that can then integrate into chromosomal DNA (**Figure 9.12**). Processed pseudogenes are common in interspersed gene families (see Table 9.5).

Processed pseudogenes lack a promoter sequence and so are typically not expressed. Sometimes, however, the cDNA copy integrates into a chromosomal DNA site that happens, by chance, to be adjacent to a promoter that can drive expression of the processed gene copy. Selection pressure may ensure that the processed gene copy continues to make a functional gene product, in which case it is described as a **retrogene**. A variety of intronless retrogenes are known to have testis-specific expression patterns and are typically autosomal homologs of an intron-containing X-linked gene (**Table 9.8**).

One rationale for retrogenes may be a critical requirement to overcome the lack of expression of certain X-linked sequences in the testis during male meiosis. During male meiosis, the paired X and Y chromosomes are converted to heterochromatin, forming the highly condensed and transcriptionally inactive **XY body**. Autosomal retrogenes can provide the continued synthesis in testis cells of certain crucially important products that are no longer synthesized by genes in the highly condensed XY body.
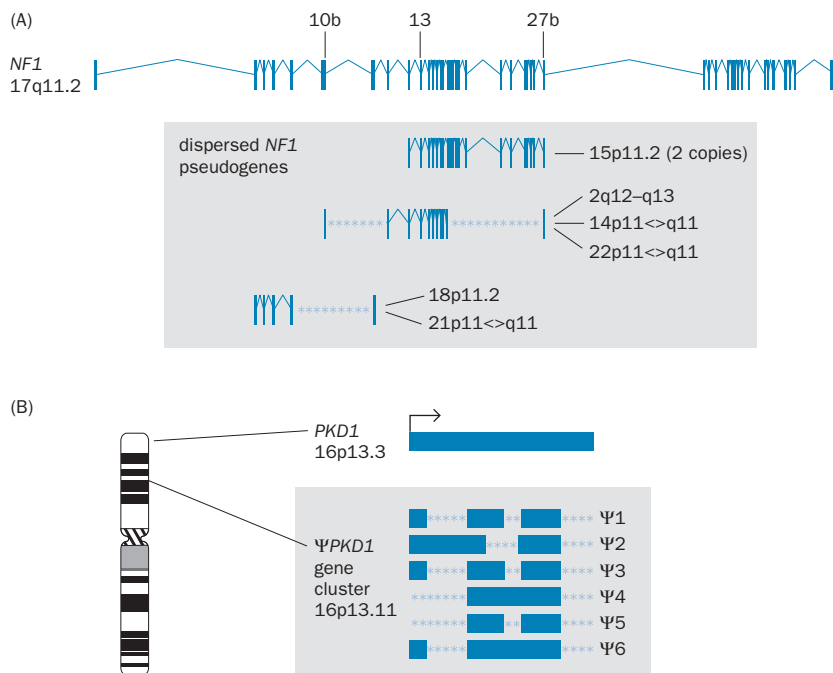

(A) The *NF1* neurofibromatosis type I gene is located close to the centromere of human chromosome 17. It spans 283 kb and has 58 exons. Exons are represented by thin vertical boxes; introns are shown by connecting chevrons (^). Highly homologous defective copies of the *NF1* gene are found at nine or more other genome locations, mostly in pericentromeric regions. Each copy has a portion of the full-length gene, with both exons and introns. Seven examples are shown here, such as two copies on 15p that have intact genomic sequences spanning exons 13 and 27b. Rearrangements have sometimes caused the deletion of exons and introns (shown by asterisks). (B) The 46 kb *PKD1* polycystic kidney disease gene is located close to the telomere of 16p and has over 40 exons. As a result of segmental duplication events during primate evolution, large components of this gene have been duplicated and six *PKD1* pseudogenes are located at 16p13.11, with large blocks of sequence (shown as blue boxes) copied from the *PKD1* gene (asterisks represent the absence of counterparts to *PKD1* sequences).
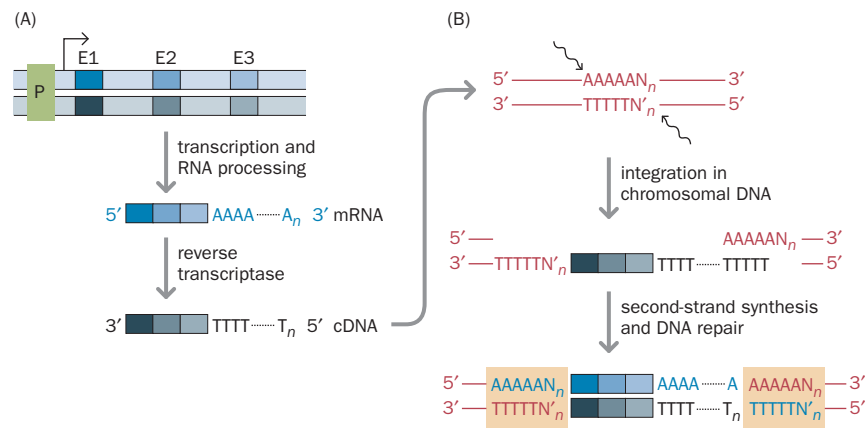
(A) In this example, a protein-coding gene with three exons (E1–E3) is transcribed from an upstream promoter (P), and introns are excised from the transcript to yield an mRNA. The mRNA can then be converted naturally into an antisense single-stranded cDNA by using cellular reverse transcriptase function (provided by LINE-1 repeats). (B) Integration of the cDNA is envisaged at staggered breaks (indicated by curly arrows) in A-rich sequences, but could be assisted by the LINE-1 endonuclease. If the A-rich sequence is included in a 5′ overhang, it could form a hybrid with the distal end of the poly(T) of the cDNA, facilitating second-strand synthesis. Because of the staggered breaks during integration, the inserted sequence will be flanked by short direct repeats (boxed sequences).

9.3   RNA GENES

Much of the attention paid to human genes has focused on protein-coding genes because they were long considered to be by far the functionally most important part of our genome. By comparison, genes whose final products are functional **noncoding RNA (ncRNA)** molecules have been so underappreciated that one of the two draft human genome sequences reported in 2001 contained no analyses at all of human RNA genes! RNA was seen to be important in very early evolution (**Box 9.3**) but its functions were imagined to have been very largely overtaken by DNA and proteins. In recent times, the vast majority of RNA molecules were imagined to serve as accessory molecules in the making of proteins.

The last few years have witnessed a revolution in our understanding of the importance of RNA and, although the number of protein-encoding genes has been steadily revised downward since draft human genome sequences were reported in 2001, the number of RNA genes is constantly being revised upward. The tiny mitochondrial genome was always considered to be exceptional because 65% (24 out of 37) of its genes are RNA genes. Now we are beginning to realize that the RNA transcribed from the nucleus is not so uniformly dedicated to protein synthesis as we once thought; instead, it shows great functional diversity.

What has changed our thinking? First, completely unsuspected classes of ncRNA have recently been discovered, including several prolific classes of tiny regulatory RNAs. Second, recent whole genome analyses using microarrays and high-throughput transcript sequencing have shown that at least 85% and possibly more than 90% of the human genome is transcribed. That is, more than 85% of the nucleotide positions in the euchromatic genome are represented in primary transcripts produced from at least one of the two DNA strands. Two other major surprises were the extent of multigenic transcription and the pervasiveness of bidirectional transcription. The recent data challenge the distinction between genes and intergenic space and have forced a radical rethink of the concept of a gene (**Box 9.4**).

TABLE 9.8 EXAMPLES OF HUMAN INTRONLESS RETROGENES AND THEIR PARENTAL INTRON-CONTAINING HOMOLOGS		
Retrogene	**Intron-containing homolog**	**Product**
GK2 at 4q13	*GK1* at Xp21	glycerol kinase
PDHA2 at 4q22	*PDHA1* at Xp22	pyruvate dehydrogenase
PGK2 at 6p12	*PGK* at Xq13	phosphoglycerate kinase
TAF1L at 9p13	*TAF1* at Xq13	TATA box binding protein associated factor, 250 kD
MYCL1 at 1p34	*MYCL2* at Xq22	homolog of v-Myc oncogene
GLUD1 at 10q23	*GLUD2* at Xq25	glutamate dehydrogenase
RBMXL at 9p13	*RBMX* at Xq26	RNA-binding protein important in brain development

BOX 9.3 THE RNA WORLD HYPOTHESIS

Proteins cannot self-replicate, and so many evolutionary geneticists consider that *autocatalytic* nucleic acids must have pre-dated proteins and were able to replicate without the help of proteins. The *RNA world hypothesis* developed from ideas proposed by Alexander Rich and Carl Woese in the 1960s. It imagines that RNA had a dual role in the earliest stages of life, acting as both the genetic material (with the capacity for self-replication) and also as effector molecules. Both roles are still evident today: some viruses have RNA genomes, and noncoding RNA molecules can work as effector molecules with catalytic activity. RNAse P, for example, is a ribozyme that can cleave substrate RNA without any requirement for protein, and certain types of intron are autocatalytic and able to splice themselves out of RNA transcripts without any help from proteins (see text). Another observation consistent with RNA's being the first nucleic acid is that deoxyribonucleotides are synthesized from ribonucleotides in cellular pathways.

As well as storing genetic information, RNA has been imagined to have been used subsequently to synthesize proteins from amino acids. Different RNAs, including rRNA and tRNA, are central in assisting polypeptide synthesis. Many ribosomal proteins can be deleted without affecting ribosome function, and the crucial peptidyl transferase activity—the enzyme that catalyzes the formation of peptide bonds—is a ribozyme. However, RNA has a rather rigid backbone and so is not very well suited as an effector molecule. Proteins are much more flexible and also offer more functional variety because the 20 amino acids can have widely different structures and offer more possible sequence combinations (a decapeptide provides 20^{10} or about 10^{13} different possible amino acid sequences, whereas a decanucleotide has 4^{10} or about 10^6 different possible sequences).

The replacement of RNA with DNA as an information storage molecule provided significant advantages. DNA is much more stable than RNA, and so is better suited for this task. Its sugar residues lack the 2′ OH group on ribose sugars that makes RNA prone to hydrolytic cleavage. Greater efficiency could be achieved by separating the storage and transmission of genetic information (DNA) from protein synthesis (RNA). All that was needed was the development of a reverse transcriptase so that DNA could be synthesized from deoxynucleotides by using an RNA template.

We have known for many decades that various ubiquitous ncRNA classes are essential for cell function. Until recently, however, we have largely been accustomed to thinking of ncRNA as not much more than a series of *accessories* that are needed to process genes to make proteins. Transfer RNAs are needed at the very end of the pathway, serving to decode the codons in mRNA and provide amino acids in the order they are needed for insertion into growing polypeptide chains. Ribosomal RNAs are essential components of the ribosomes, the complex ribonucleoprotein factories of protein synthesis.

Other ubiquitous ncRNAs were known to function higher up the pathway to ensure the correct processing of mRNA, rRNA, and tRNA precursors. Various small RNAs are components of complex ribonucleoproteins involved in different processing reactions, including splicing, cleavage of rRNA and tRNA precursors, and base modifications that are required for RNA maturation. Typically these RNAs work as *guide RNAs*, by base pairing with complementary sequences in the precursor RNA.

We have also long been aware of a few ncRNAs that have other functions, such as RNAs implicated in X-inactivation and imprinting, and the RNA component of the telomerase ribonucleoprotein needed for synthesis of the DNA of telomeres (see Figure 2.13). But these RNAs seemed to be quirky exceptions.

In the past decade or so, however, there has been a revolution in how we view RNA. Many thousands of different ncRNAs have recently been identified in animal cells. Many of them are developmentally regulated and have been shown to have crucial roles in a whole variety of different processes that occur in specialized tissues or specific stages of development. Several ncRNAs have already been implicated in cancer and genetic disease.

Now that the human genome is known to have close to 20,000 protein-coding genes, about the same as in the 1 mm nematode *Caenorhabditis elegans*, which has only about 1000 cells—the question now is whether RNA-based regulation is the key feature in explaining our complexity. Certainly, the complexity of RNA-based gene regulation increases markedly in complex organisms, as is described in Chapter 10. Maybe it is time to view the genome as more of an RNA machine than just a protein machine.

BOX 9.4 REVISING THE CONCEPT OF A GENE IN THE POST-GENOME ERA

Until genome-wide analyses were performed, a typical human protein-coding gene was imagined to be well defined and separated from its neighbors by identifiable intergenic spaces. The gene would typically be split into several exons. Directed from an upstream promoter, a primary transcript complementary to just one DNA strand would undergo splicing. The functionally unimportant (junk) intronic sequences would be discarded, allowing fusion of the important exonic sequences to make an mRNA. Expression would be regulated by nearby regulatory sequences typically located close to the promoter.

This neat and cosy image of a gene has been buffeted by a series of complications. It has long been known that some nuclear genes partly overlap others or are entirely embedded within much larger genes. Different products were known to be produced from a single gene by using alternative promoters, alternative splicing, and RNA editing. Very occasionally, sequences from different genes could be spliced together at the RNA level, sometimes even in the case of genes on different chromosomes (*trans-splicing*). Occasionally, natural antisense transcripts were observed that could be seen to regulate the expression of the sense transcript of a gene. Some genes were known to have regulatory elements located many hundreds of kilobases away, sometimes within another gene.

Despite the above complications, scientists were not quite ready to give up on the simple idea of a gene described in the first paragraph above, until whole genome analyses shattered previous misconceptions. The significant findings that forced a reappraisal of gene organization were:

- *Transcription is pervasive*. More than 85% of the euchromatic human genome is transcribed, and multigenic transcription is common, so that the distinction between genes and intergenic space is now much less apparent (**Figure 1**). About 70% of human genes are transcribed from both strands.
- *Coding DNA accounts for less than one-quarter of the highly conserved (and presumably functionally important) fraction of the genome*. This meant that there must be many more functionally important

noncoding DNA sequences than had been expected. And so it proved. An unexpectedly high number of conserved regulatory sequences and numerous novel noncoding RNAs (ncRNAs) have been, and continue to be, identified, often within or spanning introns of known protein-coding genes.

As a result of intensive analyses of mammalian transcriptomes, many thousands of transcripts of unknown function have been uncovered. Typically, ncRNA and protein-coding transcripts overlap, creating complicated patterns of transcription (**Figure 2**). In some cases, such as the imprinted *SNURF–SNRPN* transcript at 15q12, a single transcript contains a coding RNA plus a noncoding RNA that are separated by RNA cleavage.
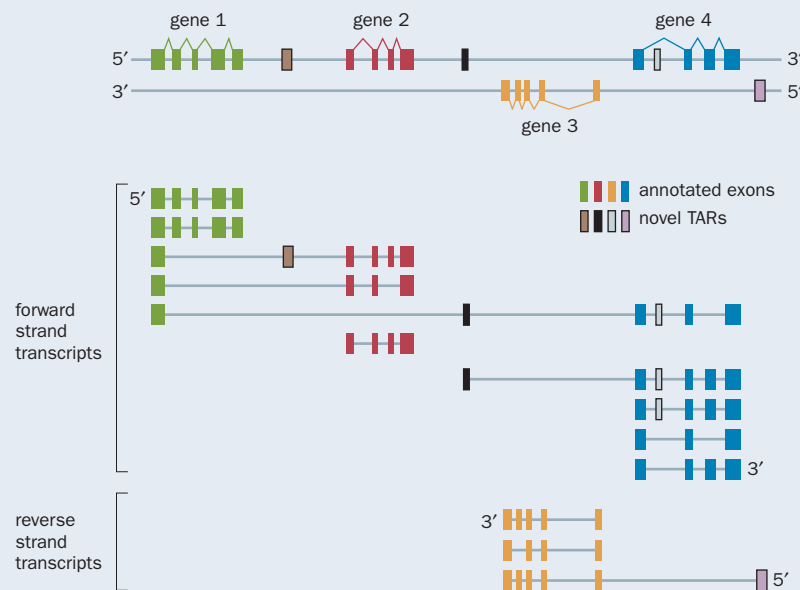

Figure 1 Blurring of gene boundaries at the transcript level. In the past, the four genes at the top would be expected to behave as discrete non-overlapping transcription units. As shown by recent analyses, the reality is more complicated. A variety of transcripts often links exons in neighboring genes. The transcripts frequently include sequences from previously unsuspected transcriptionally active regions (TARs). [From Gerstein MB, Bruce C, Rozowsky JS et al. (2007) *Genome Res.* 17, 669–681. With permission from Cold Spring Harbor Laboratory Press.]
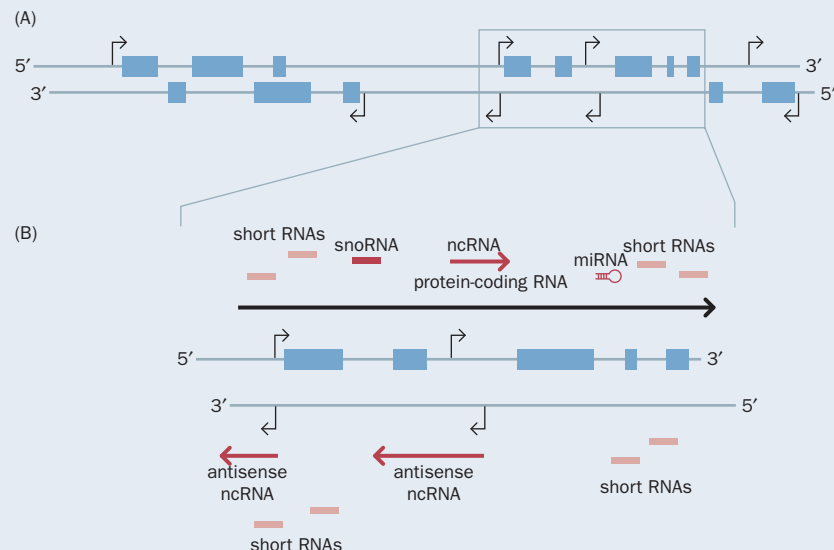
Figure 2 Extensive transcriptional complexity of human genes. (A) Human genes are frequently transcribed on both strands, as shown in this hypothetical gene cluster. (B) A single gene can have multiple transcriptional start sites (right-angled arrows) as well as many interleaved coding and noncoding transcripts. Exons are shown as blue boxes. Known short RNAs such as small nucleolar RNAs (snoRNAs) and microRNAs (miRNAs) can be processed from intronic sequences, and novel species of short RNAs that cluster around the beginning and end of genes have recently been discovered (see the text). [From Gingeras TR (2007) *Genome Res.* 17, 682–690. With permission from Cold Spring Harbor Laboratory Press.]
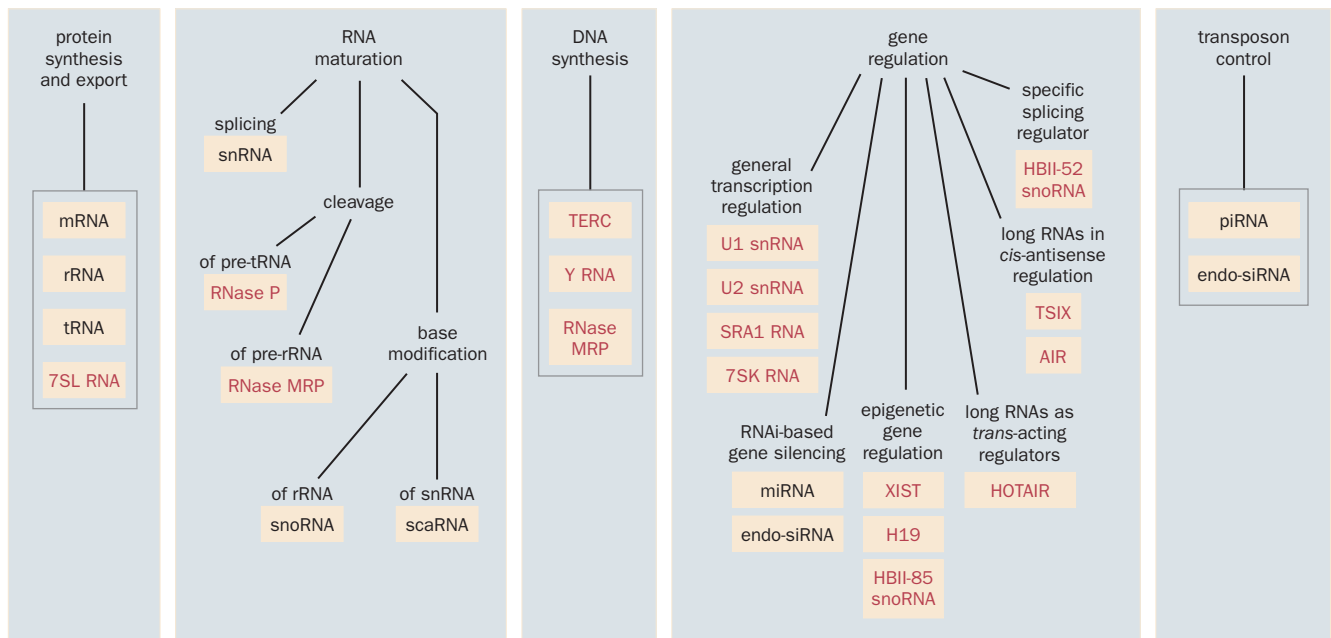
Figure 9.13 Functional diversity of RNA. Various ubiquitous RNAs function in housekeeping roles in cells, and in protein synthesis and protein export from cells (using 7SL RNA, the RNA component of the signal recognition particle). RNAs involved in RNA maturation include spliceosomal small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), small Cajal body RNAs (scaRNAs), and two RNA ribonucleases. Telomere DNA synthesis is performed by a ribonucleoprotein that consists of TERC (the telomerase RNA component) and a reverse transcriptase (see Figure 2.13). The Y RNA family are involved in chromosomal DNA replication, and RNase MRP has a crucial role in initiating mtDNA replication as well as in cleaving pre-rRNA precursors in the nucleolus. Diverse classes of RNA serve a regulatory function in gene expression. Although some do have general accessory roles in transcription, regulatory RNAs are often restricted in their expression to certain cell types and/or developmental stages. Three classes of tiny RNA use RNA interference (RNAi) pathways to act as regulators: Piwi protein-interacting RNAs (piRNAs) regulate the activity of transposons in germ-line cells; microRNAs (miRNAs) regulate the expression of target genes; and endogenous short interfering RNAs (endo-siRNAs) act as gene regulators and also regulate some types of transposon. Some members of the snoRNA family, such as HBII-85 snoRNA, are involved in epigenetic gene regulation. A large number of long RNAs are involved in regulating various genes, often at the transcriptional level. Some are known to be involved in epigenetic gene regulation in imprinting, X-inactivation, and so on. RNA families are shown in black type; individual RNAs are shown in red.

Figure 9.13 gives a modern perspective on the functional diversity of RNA. In this section we consider the functions and gene organizations of the different human RNA classes (**Table 9.9**). Numerous databases have been developed recently to document data on ncRNAs (**Table 9.10**).

More than 1000 human genes encode an rRNA or tRNA, mostly within large gene clusters

Ribosomal RNA genes

In addition to the two mitochondrial rRNA molecules (12S and 16S rRNA), there are four types of cytoplasmic rRNA, three associated with the large ribosome subunit (28S, 5.8S, and 5S rRNAs) and one with the small ribosome subunit (18S rRNA). The 5S rRNA genes occur in small gene clusters, the largest being a cluster of 16 genes on chromosome 1q42, close to the telomere. Only a few 5S rRNA genes have been validated as functional, and there are many dispersed pseudogenes.

The 28S, 5.8S, and 18S rRNAs are encoded by a single multigenic transcription unit (see Figure 1.22) that is tandemly repeated to form megabase-sized *ribosomal DNA* arrays (about 30–40 tandem repeats, or roughly 100 rRNA genes) on the short arms of each of the acrocentric human chromosomes 13, 14, 15, 21, and 22. We do not know the precise gene numbers because the ribosomal DNA arrays were excluded from the Human Genome Project, as a result of the technical difficulties in obtaining unambiguous ordering of overlapping DNA clones for long regions composed of very similar tandem repeats.

TABLE 9.9 MAJOR CLASSES OF HUMAN NONCODING RNA

RNA class	Subclass or evolutionary/ functional subfamily	No. of different types	Function	Gene organization, biogenesis, etc.
Ribosomal RNA (rRNA), ~120–5000 nucleotides	12S rRNA, 16S rRNA	1 of each	components of mitochondrial ribosomes	cleaved from multigenic transcripts produced by H strand of mtDNA (Figure 9.3)
	5S rRNA, 5.8S rRNA, 18S rRNA, 28S rRNA	1 of each	components of cytoplasmic ribosomes	5S rRNA is encoded by multiple genes in various gene clusters; 5.8S, 18S, and 28S rRNA are cleaved from multigenic transcripts (Figure 1.22); the multigenic 5.8S–18S–28S transcription units are tandemly repeated on each of 13p,14p,15p, 21p, and 22p (= rDNA clusters)
Transfer RNA (tRNA), ~70–80 nucleotides	mitochondrial family	22	decode mitochondrial mRNA to make 13 proteins on mitochondrial ribosomes	single-copy genes. tRNAs are cleaved from multigenic mtDNA transcripts (Figure 9.3)
	cytoplasmic family	49	decode mRNA produced by nuclear genes (Figure 9.13)	700 tRNA genes and pseudogenes dispersed at multiple chromosomal locations with some large gene clusters
Small nuclear RNA (snRNA), ~60–360 nucleotides	spliceosomal family with subclasses Sm and Lsm (Table 9.10)	9	U1, U2, U4, U5, and U6 snRNAs process standard GU–AG introns (Figure 1.19); U4atac, U6atac, U11, and U12 snRNAs process rare AU–AC introns	about 200 spliceosomal snRNA genes are found at multiple locations but there are moderately large clusters of U1 and U2 snRNA genes; most are transcribed by RNA pol II
	non-spliceosomal snRNAs	several	U7 snRNA: 3′ processing of histone mRNA; 7SK RNA: general transcription regulator; Y RNA family: involved in chromosomal DNA replication and regulators of cell proliferation	mostly single-copy functional genes
Small nucleolar RNA (snoRNA), ~60–300 nucleotides	C/D box class (Figure 9.15A)	246	maturation of rRNA, mostly nucleotide site-specific 2′-O-ribose methylations	usually within introns of protein-coding genes; multiple chromosomal locations, but some genes are found in multiple copies in gene clusters (such as the HBII-52 and HBII-85 clusters—Figure 11.22)
	H/ACA class (Figure 9.15B)	94	maturation of rRNA by modifying uridines at specific positions to give pseudouridine	
Small Cajal body RNA (scaRNA)		25	maturation of certain snRNA classes in Cajal bodies (coiled bodies) in the nucleus	usually within introns of protein-coding genes
RNA ribonucleases, ~260–320 nucleotides		2	RNase P cleaves pre-tRNA in nucleus + mitochondria; RNase MRP cleaves rRNA in nucleolus and is involved in initiating mtDNA replication	single-copy genes
Miscellaneous small cytoplasmic RNAs, ~80–500 nucleotides	BC200	1	neural RNA that regulates dendritic protein biosynthesis; originated from Alu repeat	1 gene, *BCYRN1*, at 2p16
	7SL RNA	3	component of the signal recognition particle (SRP) that mediates insertion of secretory proteins into the lumen of the endoplasmic reticulum	three closely related genes clustered on 14q22
	TERC (telomerase RNA component)	1	component of telomerase, the ribonucleoprotein that synthesizes telomeric DNA, using TERC as a template (Figure 2.13)	single-copy gene at 3q26
	Vault RNA	3	components of cytoplasmic vault RNPs that have been thought to function in drug resistance	*VAULTRC1, VAULTRC2*, and *VAULTRC3* are clustered at 5q31 and share ~84% sequence identity
	Y RNA	4	components of the 60 kD Ro ribonucleoprotein, an important target of humoral autoimmune responses	*RNY1, RNY3, RNY4*, and *RNY5* are clustered at 7q36

TABLE 9.9 (*cont.*) MAJOR CLASSES OF HUMAN NONCODING RNA

RNA class	Subclass or evolutionary/ functional subfamily	No. of different types	Function	Gene organization, biogenesis, etc.
MicroRNA (miRNA), ~22 nucleotides	> 70 families of related miRNAs	~1000	multiple important roles in gene regulation, notably in development, and implicated in some cancers	see Figure 9.17 for examples of genome organization, and Figure 9.16 for how they are synthesized
Piwi-binding RNA (piRNA), ~24–31 nucleotides	89 individual clusters	> 15,000	often derived from repeats; expressed only in germ-line cells, where they limit excess transposon activity	89 large clusters distributed across the genome; individual clusters span from 10 kb to 75 kb with an average of 170 piRNAs per cluster
Endogenous short interfering RNA (endo-siRNA), ~21–22 nucleotides	many	probably more than 10,000[a]	often derived from pseudogenes, inverted repeats, etc.; involved in gene regulation in somatic cells and may also be involved in regulating some types of transposon	clusters at many locations in the genome
Long noncoding regulatory RNA, often > 1 kb	many	> 3000	involved in regulating gene expression; some are involved in monoallelic expression (X-inactivation, imprinting), and/or as antisense regulators (Table 9.11)	usually individual gene copies; transcripts often undergo capping, splicing, and polyadenylation but antisense regulatory RNAs are typically long transcripts that do not undergo splicing

[a]Based on extrapolation of studies in mouse cells.

Transfer RNA genes

The 22 different mitochondrial tRNAs are made by 22 tRNA genes in mtDNA. The Genomic tRNA Database lists over 500 human tRNA genes that make a cytoplasmic tRNA with a defined anticodon specificity. The genes can be classified into 49 families on the basis of anticodon specificity (**Box 9.5**). There is only a rough correlation of human tRNA gene number with amino acid frequency. For example, 30 tRNA genes specify the comparatively rare amino acid cysteine (which accounts for 2.25% of all amino acids in human proteins), but only 21 tRNA genes specify the more abundant proline (which has a frequency of 6.10%).

Although the tRNA genes seem to be dispersed throughout the human genome, more than half of human tRNA genes (273 out of 516) reside on either chromosome 6 (with many clustered in a 4 Mb region at 6p2) or chromosome 1. In addition, 18 of the 30 Cys tRNAs are found in a 0.5 Mb stretch of chromosome 7.

TABLE 9.10 MAJOR NONCODING RNA DATABASES

Database	Description	URL
NONCODE	integrated database of all ncRNAs except rRNA and tRNA	http://www.noncode.org
Noncoding RNA database	sequences and functions of noncoding transcripts	http://biobases.ibch.poznan.pl/ncRNA/
RNAdb	comprehensive mammalian noncoding RNA database	http://research.imb.uq.edu.au/rnadb/
Rfam	noncoding RNA families and sequence alignments	http://rfam.sanger.ac.uk/
antiCODE	natural antisense transcripts database	http://www.anticode.org
sno/scaRNAbase	small nucleolar RNAs and small Cajal body-specific RNAs	http://gene.fudan.sh.cn/snoRNAbase.nsf
snoRNA-LBME-db	comprehensive human snoRNAs	http://www-snorna.biotoul.fr/
Genomic tRNA Database	comprehensive tRNA sequences	http://lowelab.ucsc.edu/GtRNAdb/
Compilation of tRNA sequences and sequences of tRNA genes	just as its name suggests	http://www.tRNA.uni-bayreuth.de
miRBase	miRNA sequences and target genes	http://microrna.sanger.ac.uk/
piRNAbank	empirically known sequences and other related information on piRNAs reported in various organisms, including human, mouse, rat, and *Drosophila*	http://pirnabank.ibab.ac.in/

BOX 9.5 ANTICODON SPECIFICITY OF EUKARYOTIC CYTOPLASMIC tRNAs

There is no one-to-one correspondence between codons in cytoplasmic mRNA and the tRNA anticodons that recognize them. The 64 possible codons are shown in **Figure 1**, alongside the (unmodified) anticodons. Horizontal lines join codon–anticodon pairs. Alternative codons that differ in having a C or a U at the third base position can be recognized by a single anticodon (*third-base wobble*, shown by chevrons). There are three rules for decoding cytoplasmic mRNA codons:

- *Codons in two-codon boxes.* Those codons ending with U/C that encode a different amino acid from those ending with A/G are known as two-codon boxes. Here the U/C wobble position is typically decoded by a G at the 5′ base position in the tRNA anticodon. For example, at the top left for Phe, there is no tRNA with an **A**AA anticodon to match the UU**U** codon, but the **G**AA anticodon can recognize both UU**U** and UU**C** codons in the mRNA (see Figure 1).
- *Non-glycine codons in four-codon boxes.* Four-codon boxes are those in which U, C, A, and G in the third, wobble, position all encode the same amino acid. Here the U/C wobble position is decoded by a chemically modified adenosine, known as inosine, at the 5′ position in the anticodon (blue shaded boxes; see Figure 11.31 for the structure of inosine). Inosine can base pair with A, C, or U. For example, at the bottom left, the GU**U** and GU**C** codons of the four-codon valine box are decoded by a tRNA with an anticodon of **A**AC, which is no doubt modified to **I**AC. The **I**AC anticodon can recognize each of GU**U**, GU**C**, and GU**A**. To avoid possible translational misreading, tRNAs with inosine at the 5′ base of the anticodon cannot be used in two-codon boxes.
- *Glycine codons.* The four-codon glycine box provides the one exception to the rule above: GGU and GGC codons are decoded by a **G**CC anticodon, rather than the expected **I**CC anticodon.

Thus, only 16 anticodons are required to decode the 32 codons ending in a U/C. The minimum set of anticodons is therefore 45 (64 minus 3 stop codons, minus 16). On this basis, one would predict a total of 45 different classes of human tRNA. However, despite the generality of third-base wobble, three pairs of codons ending in U/C are served by two anticodons each (see Figure 1), and so there are an extra three tRNA classes. In addition, a specialized tRNA carries an anticodon to the codon UGA (which normally functions as a stop codon). At high selenium concentrations, this tRNA will very occasionally decode UGA to insert the 21st amino acid, selenocysteine, in a select group of selenoproteins. Thus, there are 45 + 3 + 1 = 49 different classes of human tRNA, encoded by several hundred genes (see Figure 1).

Figure 1 Over 500 different human cytoplasmic tRNAs decode the 61 codons that specify the standard 20 amino acids. The relationships between the 64 possible codons (positioned next to amino acids on the left of the four major columns) and the corresponding anticodons (to the right of the four columns) are shown. The number next to each anticodon is the number of different human tRNAs that are documented in the Genomic tRNA Database (see Table 9.9) as carrying that anticodon. Note that 12 of the 61 anticodons that could recognize the codons that specify the standard 20 amino acids are not represented in the tRNAs (shown by dashes). This happens because of wobble at the third base position of most codons where the third base is U or C (exceptions are for codon pairs AU**U**/AU**C**, AA**U**/AA**C**, and UA**U**/UA**C**). The (3) indicated by an asterisk signifies that there are three different selenocysteine tRNAs with an anticodon that can recognize the codon UGA, which normally serves as a stop codon. The shaded adenines are most probably a modified form of adenine known as inosine, in which the amino group attached to carbon 6 is replaced by a C=O carbonyl group.

Dispersed gene families make various small nuclear RNAs that facilitate general gene expression

Various families of rather small RNA molecules (60–360 nucleotides long) are known to have a role in the nucleus in assisting general gene expression, mostly at the level of post-transcriptional processing. Initially, such RNAs were simply labeled as *small nuclear RNAs* (*snRNAs*) to distinguish them from pre-mRNA. Many of them were known to be uridine-rich and they were named accordingly (U2 snRNA, for example, does not honor a famous Irish rock band but simply indicates the second uridine-rich small nuclear RNA to be classified). Like rRNA,

TABLE 9.11 THE Sm AND Lsm CLASSES OF SPLICEOSOMAL snRNA

	Sm class[a]	Lsm class
Component of major spliceosome	U1, U2, U4, and U5 snRNAs	U6 snRNA
Component of minor spliceosome	U11, U12, U4atac, and U5 snRNAs	U6atac snRNA
Structure	see Figure 9.14A	see Figure 9.14B
Transcribed by	RNA polymerase II	RNA polymerase III
Bound core proteins	Sm proteins (SmB, SmD1, SmD2, SmD3, SmE, SmF, SmG)	seven Lsm proteins (LSM2–LSM8)
Location	synthesized in the nucleus and then exported to the cytoplasm, where they each associate with seven Sm proteins and undergo 5′ and 3′ end processing; then re-imported into the nucleus to undergo more RNA processing in Cajal bodies, before accumulating in speckles to perform spliceosomal function	never leave the nucleus; undergo maturation in the nucleolus and then accumulate in speckles to perform spliceosomal function
Site-specific nucleotide modification	performed by scaRNAs in the Cajal bodies of the nucleus	performed by snoRNAs in the nucleolus

[a]Although not a spliceosomal snRNA, U7 snRNA has a similar structure to the Sm class of snRNAs and five of its seven core proteins are identical to core proteins bound by the Sm snRNAs.

snRNA molecules bind various proteins and function as ribonucleoproteins (snRNPs).

Subsequently, various snRNAs, including some of the first to be classified, were found to be involved in post-transcriptional processing of rRNA precursors in the nucleolus; they were therefore re-classified as *small nucleolar RNAs* (*snoRNAs*), for example U3 and U8 snoRNAs. More recently, membership of the classes has been based on structural and functional classification.

A third group of small RNAs have been identified that resemble snoRNAs but are confined to Cajal bodies (also called *coiled bodies*), discrete nuclear structures in the nucleus that are closely associated with the maturation of snRNPs. They have been termed *small Cajal body RNAs* (*scaRNAs*). Hundreds of mostly dispersed human genes are devoted to making snRNA and snoRNA, and there are many hundreds of associated pseudogenes.
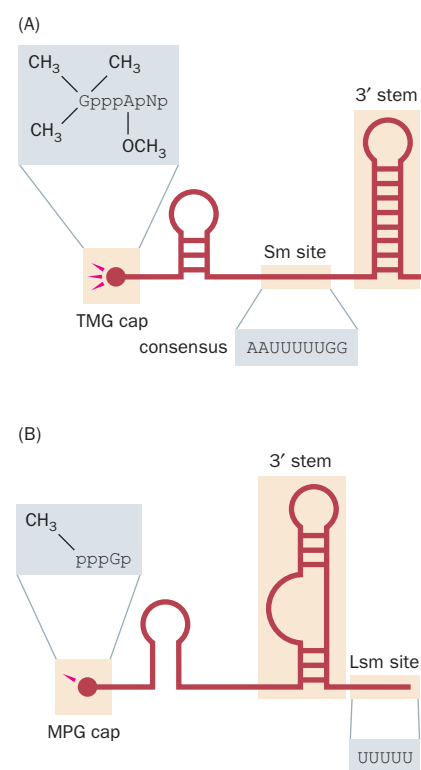
Spliceosomal small nuclear RNA (snRNA) genes

The nine human spliceosomal snRNAs vary in length from 106 to 186 nucleotides and bind a ring of seven core proteins. U1, U2, U4, U5, and U6 snRNAs operate within the major spliceosome to process conventional GU–AG introns (see Figure 1.19). U4atac, U6atac, U11, and U12 snRNAs form part of the minor spliceosome that excises rare AU–AC introns. Each of the spliceosomal snRNPs contains seven core proteins that are identical within a subclass and a unique set of snRNP-specific proteins. The Lsm subclass is made up of just U6 and U6atac snRNAs; the other spliceosomal snRNAs belong to the Sm subclass (**Table 9.11** and **Figure 9.14**).

More than 70 genes specify snRNAs used in the major spliceosome. They include 44 identified genes specifying U6 snRNA and 16 specifying U1 snRNA. There is some evidence for clustering. Multiple U2 snRNA genes are found at 17q21–q22 but the copy number varies; a cluster of about 30 U1 snRNA genes is located at 1p36.1.

Figure 9.14 Structures of Sm-type and Lsm-type spliceosomal snRNAs. (A) Sm-type snRNAs contain three important recognition elements: a 5′-trimethylguanosine (TMG) cap, an Sm-protein-binding site (Sm site), and a 3′ stem–loop structure. The Sm site and the 3′ stem elements are required for recognition by the survival motor neuron (SMN) complex for assembly into stable core ribonucleoproteins (RNPs). The consensus Sm site directs the assembly of a ring of the seven Sm core proteins (see Table 9.10). The TMG cap and the assembled Sm core proteins are required for recognition by the nuclear import machinery. (B) Lsm-type snRNAs contain a 5′-monomethylphosphate guanosine (MPG) cap and a 3′ stem, and terminate in a stretch of uridine residues (the Lsm site) that is bound by the seven Lsm core proteins.

Non-spliceosomal small nuclear RNA genes

Not all snRNAs within the nucleoplasm function as part of spliceosomes. Both U1 and U2 snRNAs also have non-spliceosomal functions. U1 snRNA is required to stimulate transcription by RNA polymerase II. U2 snRNA is known to stimulate transcriptional elongation by RNA polymerase II. Several other small nuclear RNAs with a non-spliceosomal function have been well studied. They tend to be single-copy genes but there are many associated pseudogenes. Three examples are given below.

- U7 snRNA is a 63-nucleotide snRNA that is dedicated to the specialized 3′ processing undergone by histone mRNA which, exceptionally, is not polyadenylated.

- 7SK RNA is a 331-nucleotide RNA that functions as a negative regulator of the RNA polymerase II elongation factor P-TEFb.

- The Y RNA family consists of three small RNAs (less than 100 nucleotides long) that are involved in chromosomal DNA replication and function as regulators of cell proliferation.

Small nucleolar RNA (snoRNA) genes

SnoRNAs are between 60 and 300 nucleotides long and were initially identified in the nucleolus, where they guide nucleotide modifications in rRNA at specific positions. They do this by forming short duplexes with a sequence of the rRNA that contains the target nucleotide. There are two large subfamilies. H/ACA snoRNAs guide site-specific pseudouridylations (uridine is isomerized to give pseudouridine at 95 different positions in the pre-rRNA). C/D box snoRNAs guide site-specific 2′-O-ribose methylations (there are 105–107 varieties of this methylation in rRNA). Single snoRNAs specify one, or at most two, such base modifications (**Figure 9.15**).

At least 340 human snoRNA genes have been found so far, but there may be many more because snoRNAs are surprisingly difficult to identify with the use of bioinformatic approaches. The vast majority are found within the introns of larger genes that are transcribed by RNA polymerase II. These snoRNAs are produced by processing of the intronic RNA, and so the regulation of their synthesis is coupled to that of the host gene. Many snoRNA genes are dispersed single-copy genes. Others occur in clusters. For example, the large imprinted *SNURF–SNRPN* transcription unit at 15q12 contains six different types of C/D box snoRNA gene, two of which are present in large gene clusters: one contains about 45 almost identical HBII-52 snoRNA genes and the other 29 HBII-85 snoRNA genes (see Figure 11.22).

Most snoRNAs are ubiquitously expressed, but some are tissue-specific. For example, the six types of snoRNA gene within the *SNURF–SNRPN* transcription unit are predominantly expressed in the brain from only the paternal chromo-

Figure 9.15 Structure and function of snoRNAs. (A) C/D box snoRNAs guide 2′-O-methylation modifications. The box C and D motifs and a short 5′, 3′-terminal stem formed by intrastrand base pairing (shown as a series of short horizontal red lines) constitute a kink-turn structural motif that is specifically recognized by the 15.5 kD snoRNP protein. The C′ and D′ boxes represent internal, frequently imperfect copies of the C and D boxes. C/D box snoRNAs and their substrate RNAs form a 10–21 bp double helix in which the target residue to be methylated (shown here by the letter m in a circle) is positioned exactly five nucleotides upstream of the D or D′ box. R represents purine. (B) H/ACA box snoRNAs guide the conversion of uridines to pseudouridine. These RNAs fold into a hairpin–hinge–hairpin–tail structure. One or both of the hairpins contains an internal loop, called the pseudouridylation pocket, that forms two short (3–10 bp) duplexes with nucleotides flanking the unpaired substrate uridine (ψ) located about 15 nucleotides from the H or ACA box of the snoRNA. Although each box C/D and H/ACA snoRNA could potentially direct two modification reactions, apart from a few exceptions, most snoRNAs possess only one functional 2′-O-methylation or pseudouridylation domain.
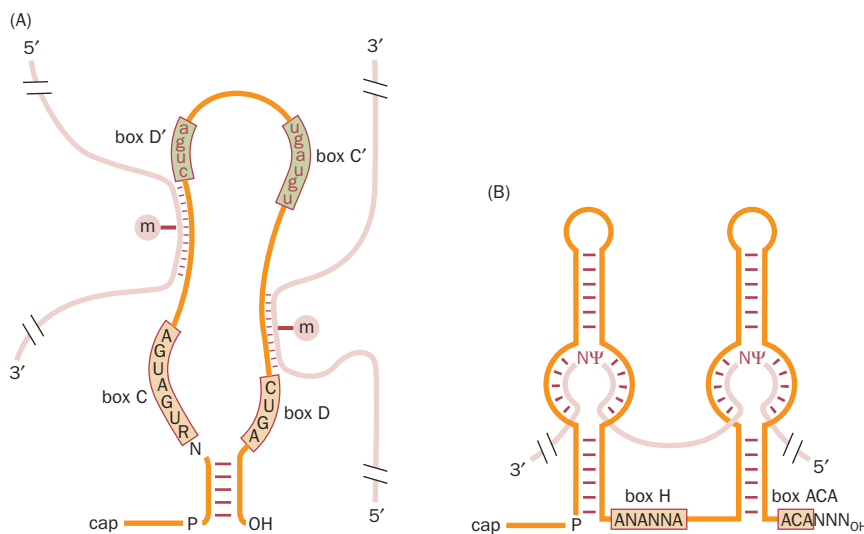
some 15. Nonstandard functions are known or expected for some snoRNA genes that do not have sequences complementary to rRNA sequences. For example, the HBII-52 snoRNA has an 18-nucleotide sequence that is perfectly complementary to a sequence within the *HTR2C* (serotonin receptor 2c) gene at Xp24, and regulates alternative splicing of this gene. The neighboring HBII-85 snoRNAs have recently been implicated in the pathogenesis of Prader–Willi syndrome (OMIM 176270).

Small Cajal body RNA genes

The scaRNAs resemble snoRNAs and perform a similar role in RNA maturation, but their targets are spliceosomal snRNAs and they perform site-specific modifications of spliceosomal snRNA precursors in the Cajal bodies of the nucleus. There are at least 25 human genes, each specifying one type of scaRNA. Like snoRNA genes, the scaRNA genes are typically located within the introns of genes transcribed by RNA polymerase II.

Close to 1000 different human microRNAs regulate complex sets of target genes by base pairing to the RNA transcripts

In addition to tRNA, we have known for some time about a variety of other moderately small (80–500-nucleotide) cytoplasmic RNAs. For example, the enzyme telomerase that synthesizes DNA at telomeres (see Figure 2.13) has both a protein component, TERT (telomerase reverse transcriptase), and also an RNA component, TERC, that is synthesized by a single-copy gene at 3q26.2 (see Table 9.8 for other examples). In the early 2000s it became clear that a novel family of tiny regulatory RNAs known as **microRNA** (**miRNA**) also operated in the cytoplasm.

MicroRNAs are only about 21–22 nucleotides long on average, and they were initially missed in analyses of the human genome. The first animal miRNAs to be reported were identified in model organisms such as the nematode (*C. elegans*) and the fruit fly (*Drosophila melanogaster*) by investigators studying phenomena relating to **RNA interference** (**Box 9.6**), a natural form of gene regulation that protects cells from harmful propagation of viruses and transposons.

Because many miRNAs are strongly conserved during evolution, vertebrate miRNAs were quickly identified and the first human miRNAs were reported in the early 2000s. miRNAs regulate the expression of selected sets of target genes by base pairing with their transcripts. Usually, the binding sites are in the 3′ untranslated region of target mRNA sequences, and the bound miRNA inhibits translation so as to down-regulate expression of the target gene.

Synthesis of miRNAs involves the cleavage of RNA precursors by nucleus-specific and cytoplasm-specific RNase III ribonucleases, nucleases that specifically bind to and cleave double-stranded RNAs. The primary transcript, the *pri-miRNA*, has closely positioned inverted repeats that base-pair to form a hairpin RNA that is initially cleaved from the primary transcript by a nuclear RNase III (known as Rnasen or Drosha) to make a short double-stranded pre-miRNA that is transported out of the nucleus (**Figure 9.16**). A cytoplasmic RNase III called dicer cleaves the pre-miRNA to generate a miRNA duplex with overhanging 3′ dinucleotides.

A specific RNA-induced silencing complex (RISC) that contains the endoribonuclease argonaute binds the miRNA duplex and acts so as to unwind the double-stranded miRNA. The argonaute protein then degrades one of the RNA strands (the *passenger strand*) to leave the mature single-stranded miRNA (known as the *guide strand*) bound to argonaute. The mature miRNP associates with RNA transcripts that have sequences complementary to the guide strand. The binding of miRNA to target transcript normally involves a significant number of base mismatches. As a result, a typical miRNA can silence the expression of hundreds of target genes in much the same way that a tissue-specific protein transcription factor can affect the expression of multiple target genes at the same time—see the targets section of the miRBase database listed in Table 9.9.

To identify more miRNA genes, new computational bioinformatics programs were developed to screen genome sequences. By mid-2009, more than 700 human miRNA genes had been identified and experimentally validated, but comparative genomics analyses indicate that the number of such genes is likely to increase. Some of the miRNA genes have their own individual promoters; others are part of

BOX 9.6 RNA INTERFERENCE AS A CELL DEFENSE MECHANISM

RNA interference (*RNAi*) is an evolutionarily ancient mechanism that is used in animals, plants, and even single-celled fungi to protect cells against viruses and transposable elements. Both viruses and active transposable elements can produce long double-stranded RNA, at least transiently during their life cycles. Long double-stranded RNA is not normally found in cells and, for many organisms, it triggers an RNA interference pathway. A cytoplasmic endoribonuclease called dicer cuts the long RNA into a series of short double-stranded RNA pieces known as **short interfering RNA** (**siRNA**). The siRNA produced is on average 21 bp long, but asymmetric cutting produces two-nucleotide overhangs at their 3′ ends (**Figure 1**).

The siRNA duplexes are bound by different complexes that contain an argonaute-type endoribonuclease (Ago) and some other proteins. Thereafter, the two RNA strands are unwound and one of the RNA strands is degraded by argonaute, leaving a single-stranded RNA bound to the argonaute complex. The argonaute complex is now activated; the single-stranded RNA will guide the argonaute complex to its target by base-pairing with complementary RNA sequences in the cells.

One type of argonaute complex is known as the *RNA-induced silencing complex* (*RISC*). In this case, after the single-stranded guide RNA binds to a complementary long single-stranded RNA, the argonaute enzyme will cleave the RNA, causing it to be degraded. Viral and transposon RNA can be inactivated in this way.

Another class of argonaute complex is the *RNA-induced transcriptional silencing* (*RITS*) complex. Here, the single-stranded RNA guide binds to complementary RNA transcripts as they emerge during transcription by RNA polymerase II. This allows the RITS complex to position itself on a specific part of the genome and then attract proteins such as histone methyltransferases (HMTs) and sometimes DNA methyltransferases (DNMTs), which covalently modify histones in the immediate region. This process eventually causes heterochromatin formation and spreading; in some cases, the RITS complex can induce DNA methylation. As a result, gene expression can be silenced over long periods to limit, for example, the activities of transposons.

Although mammalian cells have RNA interference pathways, the presence of double-stranded RNA triggers an interferon response that causes *nonspecific* gene silencing and cell death. This is described in Chapter 12 when we consider using RNA interference as an experimental tool to produce the specific silencing of pre-selected target genes. In such cases, artificially synthesized short double-stranded RNA is used to trigger RNAi-based gene silencing.
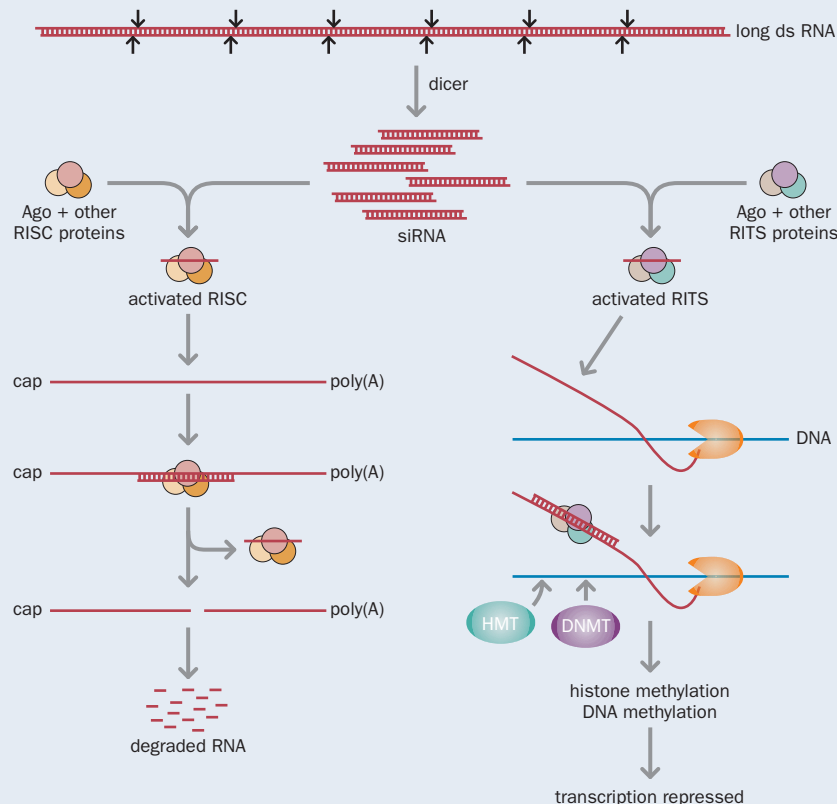
Figure 1 RNA interference. Long double-stranded (ds) RNA is cleaved by cytoplasmic dicer to give siRNA. siRNA duplexes are bound by argonaute complexes that unwind the duplex and degrade one strand to give an activated complex with a single RNA strand. By base pairing with complementary RNA sequences, the siRNA guides argonaute complexes to recognize target sequences. Activated RISC complexes cleave any RNA strand that is complementary to their bound siRNA. The cleaved RNA is rapidly degraded. Activated RITS complexes use their siRNA to bind to any newly synthesized complementary RNA and then attract proteins, such as histone methyltransferases (HMT) and sometimes DNA methyltransferases (DNMT), that can modify the chromatin to repress transcription.

a miRNA cluster and are cleaved from a common multi-miRNA transcription unit (**Figure 9.17A**). Another class of miRNA genes form part of a compound transcription unit that is dedicated to making other products in addition to miRNA, either another type of ncRNA (Figure 9.17B) or a protein (Figure 9.17C).

Many thousands of different piRNAs and endogenous siRNAs suppress transposition and regulate gene expression

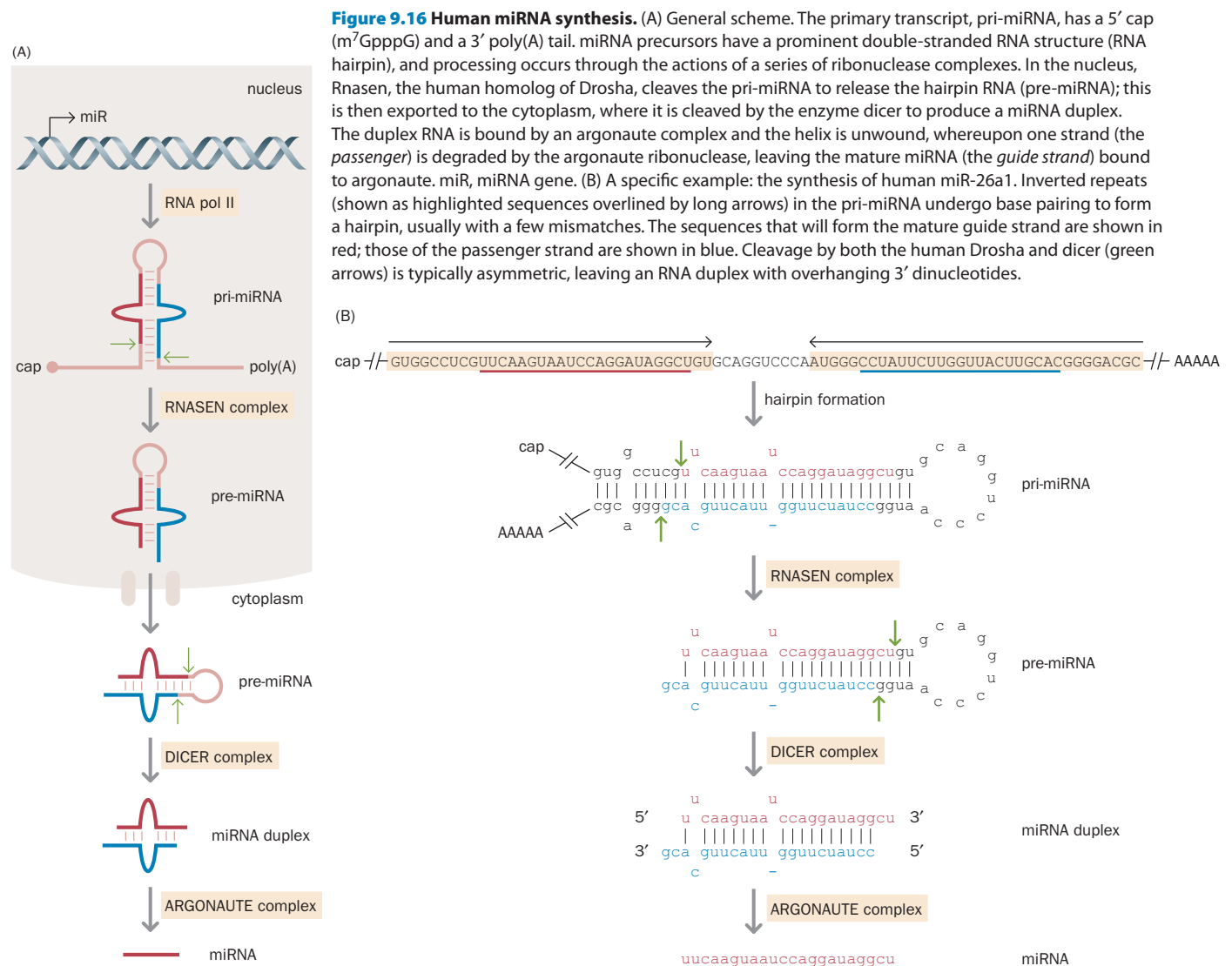The discovery of miRNAs was unexpected, but later it became clear that miRNAs represent a small component of what are a huge number of different tiny regulatory RNAs made in animal cells. In mammals, two additional classes of tiny regulatory RNA were first reported in 2006, and these are being intensively studied. Because huge numbers of different varieties of these RNAs are generated from multiple different locations in the genome, large-scale sequencing has been required to differentiate them.

Piwi-protein-interacting RNA

Piwi-protein-interacting RNAs (piRNAs) have been found in a wide variety of eukaryotes. They are expressed in germ-line cells in mammals and are typically 24–31 nucleotides long; they are thought to have a major role in limiting transposition by retrotransposons in mammalian germ-line cells, but they may also regulate gene expression in some organisms. Control of transposon activity is required because by integrating into new locations in the genome, active transposons can interfere with gene function, causing genetic diseases and cancer.

Figure 9.16 Human miRNA synthesis. (A) General scheme. The primary transcript, pri-miRNA, has a 5′ cap (m⁷GpppG) and a 3′ poly(A) tail. miRNA precursors have a prominent double-stranded RNA structure (RNA hairpin), and processing occurs through the actions of a series of ribonuclease complexes. In the nucleus, Rnasen, the human homolog of Drosha, cleaves the pri-miRNA to release the hairpin RNA (pre-miRNA); this is then exported to the cytoplasm, where it is cleaved by the enzyme dicer to produce a miRNA duplex. The duplex RNA is bound by an argonaute complex and the helix is unwound, whereupon one strand (the *passenger*) is degraded by the argonaute ribonuclease, leaving the mature miRNA (the *guide strand*) bound to argonaute. miR, miRNA gene. (B) A specific example: the synthesis of human miR-26a1. Inverted repeats (shown as highlighted sequences overlined by long arrows) in the pri-miRNA undergo base pairing to form a hairpin, usually with a few mismatches. The sequences that will form the mature guide strand are shown in red; those of the passenger strand are shown in blue. Cleavage by both the human Drosha and dicer (green arrows) is typically asymmetric, leaving an RNA duplex with overhanging 3′ dinucleotides.
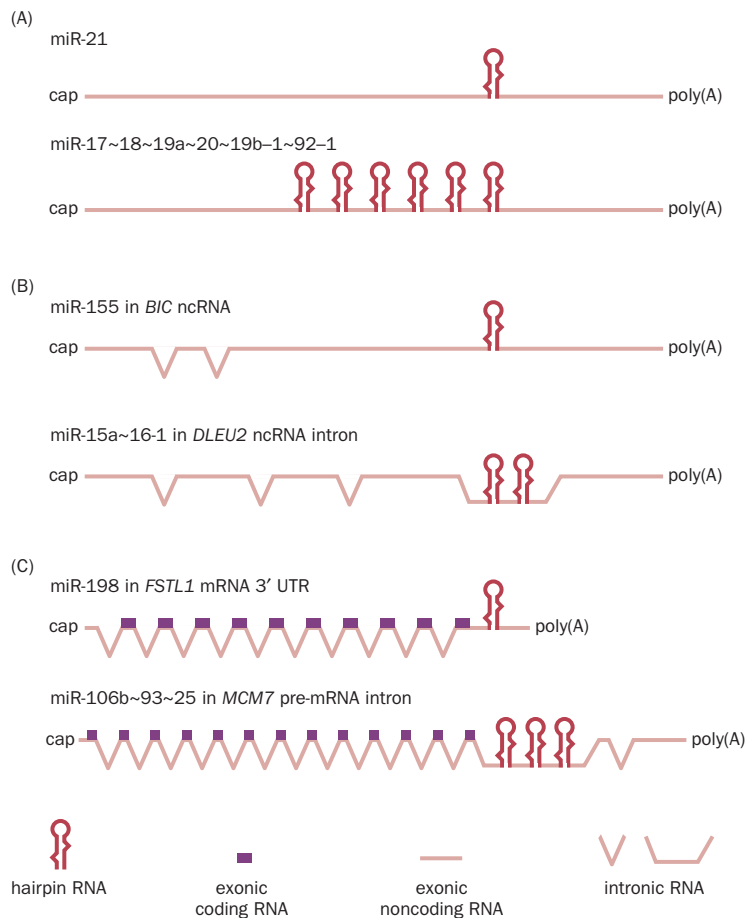
Figure 9.17 The structure of human pri-miRNAs. (A) Examples of transcripts that are used exclusively to make miRNAs: miR-21 is produced from a single hairpin within a dedicated primary transcript RNA; a single multigenic transcript with six hairpins that will eventually be cleaved to give six miRNAs, namely miR-17, miR-18, miR-19a, and so on. (B, C) Examples of miRNAs that are co-transcribed with a gene encoding either (B) a long noncoding RNA (ncRNA) or (C) a polypeptide. In each part, the upper example shows single miRNAs located within (B) an exon of an ncRNA (miR-155) and (C) in the 3′ untranslated region (UTR) within a terminal exon of an mRNA (miR-198). The lower examples show multiple miRNAs located within intronic sequences of (B) an ncRNA (miR-15a and miR-16-1) and (C) a pre-mRNA (miR-106b, miR-93, and miR-25). Cap, $m^7G(5′)ppp(5′)$ G. [Adapted from Du T & Zamore PD (2005) *Development* 132, 4645–4652. With permission from the Company of Biologists.]

More than 15,000 different human piRNAs have been identified and so the piRNA family is among the most diverse RNA family in human cells (see Table 9.8). The piRNAs map back to 89 genomic intervals of about 10–75 kb long (for more information see the piRNAbank database; Table 9.9). They are thought to be cleaved from large multigenic transcripts. In humans, the multi-piRNA transcripts contain sequences for up to many hundreds of different piRNAs.

piRNAs are thought to repress transposition by transposons through an RNA interference pathway, by association with piwi proteins, which are evolutionarily related to argonaute proteins (**Figure 9.18**).

Endogenous siRNAs

Long double-stranded RNA in mammalian cells triggers nonspecific gene silencing through interferon pathways, but transfection of exogenous synthetic siRNA duplexes or short hairpin RNAs induces RNAi-mediated silencing of specific genes with sequence elements in common with the exogenous RNA. As we will see in Chapter 12, this is an extremely important experimental tool that can give valuable information on the cellular functions of a gene. Very recently, it has become clear that human cells also naturally produce *endogenous siRNAs* (*endo-siRNAs*).

In mammals, the most comprehensive endo-siRNA analyses have been performed in mouse oocytes. Like piRNAs, endo-siRNAs are among the most varied RNA population in the cell (many tens of thousands of different endo-siRNAs have been identified in mouse oocytes). They arise as a result of the production of limited amounts of natural double-stranded RNA in the cell. One way in which this happens involves the occasional transcription of some pseudogenes (**Figure 9.19**).
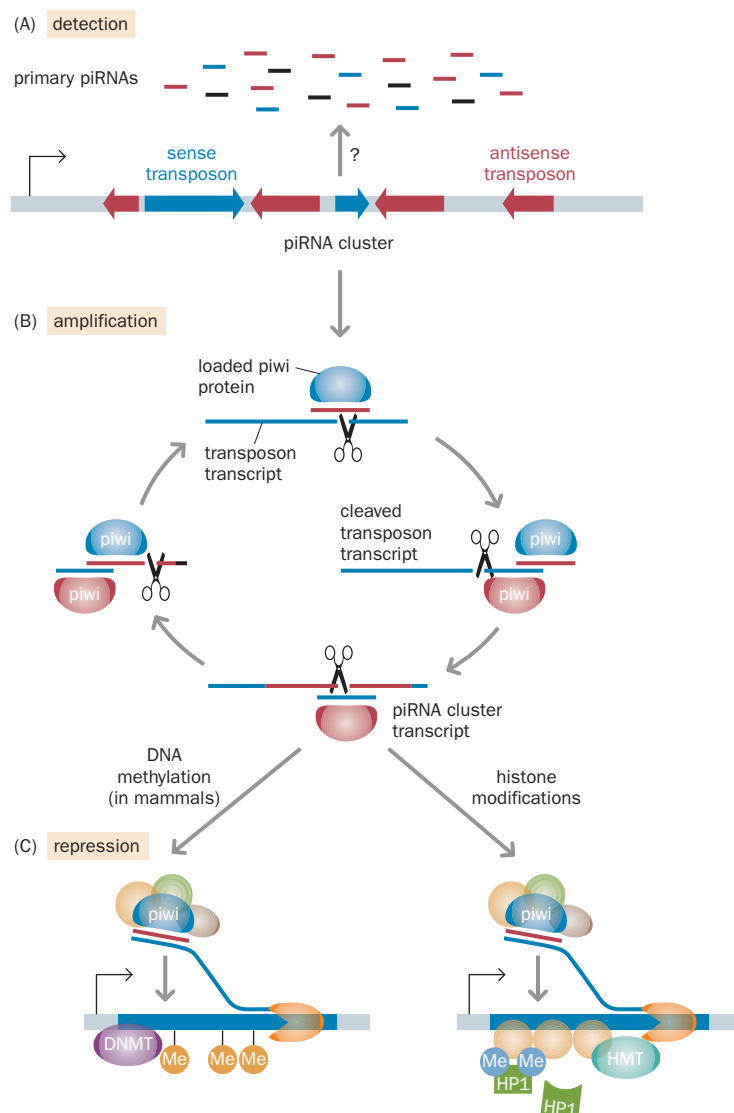
Figure 9.18 piRNA-based transposon silencing in animal cells. (A) Primary piRNAs (piwi-protein-interacting RNAs) are 24–31 nucleotides long and are processed from long RNA precursors transcribed from defined loci called piRNA clusters. Any transposon inserted in the reverse orientation in the piRNA cluster can give rise to antisense piRNAs (shown in red). (B) Antisense piRNAs are incorporated into a piwi protein and direct its slicer activity on sense transposon transcripts. The 3′ cleavage product is bound by another piwi protein and trimmed to piRNA size. This sense piRNA is, in turn, used to cleave piRNA cluster transcripts and to generate more antisense piRNAs. (C) Antisense piRNAs target the piwi complexes to cDNA for DNA methylation (left) and/or histone modification (right). DNMT, DNA methyltransferase; HMT, histone methyltransferase; HP1, heterochromatin protein 1. [From Girard A & Hannon GJ (2007) *Trends Cell Biol*. 18, 136–148. With permission from Elsevier.]

More than 3000 human genes synthesize a wide variety of medium-sized to large regulatory RNAs

Many thousands of different long ncRNAs, often many kilobases in length, are also thought to have regulatory roles in animal cells. They include antisense transcripts that usually do not undergo splicing and that can regulate overlapping sense transcripts, plus a wide variety of long mRNA-like ncRNAs that undergo capping, splicing, and polyadenylation but do not seem to encode any sizeable polypeptide, although some contain internal ncRNAs such as snoRNAs and piRNAs. The functions of the great majority of the mRNA-like ncRNAs are unknown. Some, however, are known to be tissue-specific and involved in gene regulation. Recently, in a systematic effort to identify long ncRNAs, 3300 different human long ncRNAs were identified as associating with chromatin-modifying complexes, thereby affecting gene expression.

Some long mRNA-like ncRNAs that are involved in epigenetic regulation have been extensively studied. The *XIST* gene encodes a long ncRNA that regulates X-chromosome inactivation, the process by which one of the two X chromosomes is randomly selected to be condensed in female mammals, with large regions becoming transcriptionally inactive. Many other long ncRNAs, such as the *H19* RNA, are implicated in repressing the transcription of either the paternal or maternal allele of many autosomal regions (*imprinting*). These mRNA-like ncRNAs are often regulated by genes that produce what can be very long antisense ncRNA transcripts that usually do not undergo splicing (**Table 9.12** gives examples).
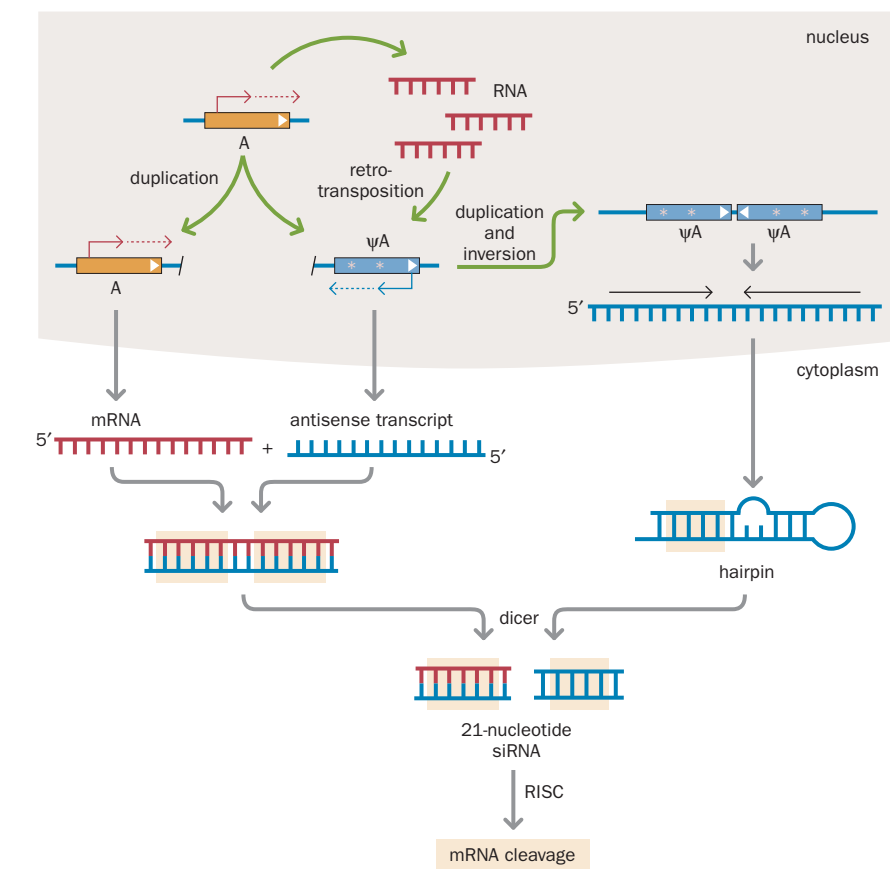
The pervasive involvement of long ncRNAs in regulating developmental processes is illustrated by a comprehensive (5 bp resolution) analysis of the transcriptional output at the four human *HOX* gene clusters. Although there are only 39 *HOX* genes, the transcriptional output of the *HOX* clusters was also found to include a total of 231 different long ncRNAs. Many of these are *cis*-acting regulators, but one of them, HOTAIR, was found to be a *trans*-acting regulator (see Table 9.12).

Some of the functional RNAs, such as XIST and AIR, have not been so well conserved during evolution. The fastest-evolving functional sequences in the human genome include components of primate-specific long ncRNAs that are strongly expressed in brain. We consider the evolutionary implications of such genes in Chapter 10.

TABLE 9.12 EXAMPLES OF LONG REGULATORY HUMAN RNAs				
RNA	**Size**	**Gene location**	**Gene organization**	**Function**
XIST	19.3 kb	Xq13	6 exons spanning 32 kb	regulator of X-chromosome inactivation
TSIX	37.0 kb	Xq13	1 exon	antisense regulator of XIST
H19	2.3 kb	11p15	5 exons spanning 2.67 kb	involved in imprinting at the 11p15 imprinted cluster associated with Beckwith–Wiedemann syndrome
KCNQTOT1 (= LIT1)	59.5 kb	11p15	1 exon	antisense regulator at the imprinted cluster at 11p15
PEG3	1.8 kb[a]	19q13	variable number of exons but up to 9 exons spanning 25 kb	maternally imprinted and known to function in tumor suppression by activating p53
HOTAIR	2.2 kb	12q13	6 exons spanning 6.3 kb	*trans*-acting gene regulator; although part of a regulatory region in the *HOX-C* cluster on 12q13, HOTAIR RNA represses transcription of a 40 kb region on the *HOX-D* cluster on chromosome 2q31

[a]Largest isoforms.

9.4 HIGHLY REPETITIVE DNA: HETEROCHROMATIN AND TRANSPOSON REPEATS

Genes contain some repetitive DNA sequences, including repetitive coding DNA. However, the majority of highly repetitive DNA sequences occur outside genes. Some of the sequences are present at certain subchromosomal regions as large arrays of tandem repeats. This type of DNA, known as heterochromatin, remains highly condensed throughout the cell cycle and does not generally contain genes.

Other highly repetitive DNA sequences are interspersed throughout the human genome and were derived by *duplicative transposition* (see Section 9.1). Sequences like this are sometimes described as *transposon repeats* and they account for more than 40% of the total DNA sequence in the human genome. In addition to residing in extragenic regions, they are often found in introns and untranslated sequences and sometimes even in coding sequences.

Constitutive heterochromatin is largely defined by long arrays of high-copy-number tandem DNA repeats

The DNA of constitutive heterochromatin accounts for 200 Mb or 6.5% of the human genome (see Table 9.3). It encompasses megabase regions at the centromeres and comparatively short lengths of DNA at the telomeres of all chromosomes. Most of the Y chromosome and most of the short arms of the acrocentric chromosomes (13, 14, 15, 21, and 22) consist of heterochromatin. In addition, there are very substantial heterochromatic regions close to the centromeres of certain chromosomes, notably chromomosomes 1, 9, 16, and 19.

The DNA of constitutive heterochromatin mostly consists of long arrays of high-copy-number tandemly repeated DNA sequences, known as *satellite DNA* (**Table 9.13**). Shorter arrays of tandem repeats are known as minisatellites and

TABLE 9.13 MAJOR CLASSES OF HIGH-COPY-NUMBER TANDEMLY REPEATED HUMAN DNA

Class[a]	Total array size unit	Size or sequence of repeat unit	Major chromosomal location(s)
Satellite DNA[b]	often hundreds of kilobases		associated with heterochromatin
α (alphoid DNA)		171 bp	centromeric heterochromatin of all chromosomes
β (*Sau*3A family)		68 bp	notably the centromeric heterochromatin of 1, 9, 13, 14, 15, 21, 22, and Y
Satellite 1		25–48 bp (AT-rich)	centromeric heterochromatin of most chromosomes and other heterochromatic regions
Satellite 2		diverged forms of ATTCC/GGAAT	most, possibly all, chromosomes
Satellite 3		ATTCC/GGAAT	13p, 14p, 15p, 21p, 22p, and heterochromatin on 1q, 9q, and Yq12
DYZ19		125 bp	~400 kb at Yq11
DYZ2		AT-rich	Yq12; higher periodicity of ~2470 bp
Minisatellite DNA	0.1–20 kb		at or close to telomeres of all chromosomes
Telomeric minisatellite		TTAGGG	all telomeres
Hypervariable minisatellites		9–64 bp	all chromosomes, associated with euchromatin, notably in sub-telomeric regions
Microsatellite DNA	< 100 bp	often 1–4 bp	widely dispersed throughout all chromosomes

[a]The distinction between satellite, minisatellite, and microsatellite is made on the basis of the total array length, not the size of the repeat unit.
[b]Satellite DNA arrays that consist of simple repeat units often have base compositions that are radically different from the average 41% G+C (and so could be isolated by buoyant density gradient centrifugation, when they would be differentiated from the main DNA and appear as *satellite bands*—hence the name).

microsatellites, respectively. Large tracts of heterochromatin are typically composed of a mosaic of different satellite DNA sequences that are occasionally interrupted by transposon repeats but are devoid of genes. Transposon repeats are also widely distributed in euchromatin and will be described below.

The vast majority of human heterochromatic DNA has not been sequenced, because of technical difficulties in obtaining unambiguous ordering of overlapping DNA clones. Thus, for example, only short representative components of centromeric DNA have been sequenced so far. However, Y-chromosome heterochromatin is an exception and has been well characterized. There are different satellite DNA organizations, and the repeated unit may be a very simple sequence (less than 10 nucleotides long) or a moderately complex one that can extend to over 100 nucleotides long; see Table 9.13.

At the sequence level, satellite DNA is often extremely poorly conserved between species. Its precise function remains unclear, although some human satellite DNAs are implicated in the function of centromeres whose DNA consists very largely of various families of satellite DNA.

The centromere is an epigenetically defined domain. Its function is independent of the underlying DNA sequence; instead, its function depends on its particular chromatin organization, which, once established, has to be stably maintained through multiple cell divisions. Of the various satellite DNA families associated with human centromeres, only the α-satellite is known to be present at all human centromeres, and its repeat units often contain a binding site for a specific centromere protein, CENPB. Cloned α-satellite arrays have been shown to seed *de novo* centromeres in human cells, indicating that α-satellite must have an important role in centromere function.

The specialized telomeric DNA consists of medium-sized arrays just a few kilobases long and constitutes a form of *minisatellite DNA*. Unlike satellite DNA, telomeric minisatellite DNA has been extraordinarily conserved during vertebrate evolution and has an integral role in telomere function. It consists of arrays of tandem repeats of the hexanucleotide TTAGGG that are synthesized by the telomerase ribonucleoprotein (see Figure 2.13).

Transposon-derived repeats make up more than 40% of the human genome and arose mostly through RNA intermediates

Almost all of the interspersed repetitive noncoding DNA in the human genome is derived from **transposons** (also called *transposable elements*), mobile DNA sequences that can migrate to different regions of the genome. Close to 45% of the genome can be recognized as belonging to this class, but much of the remaining unique DNA must also be derived from ancient transposon copies that have diverged extensively over long evolutionary time-scales.

In humans and other mammals there are four major classes of transposon repeat, but only a tiny minority of transposon repeats are actively transposing. According to the method of transposition, the repeats can be organized into two groups:

- *Retrotransposons* (also abbreviated to *retropons*). Here the copying mechanism resembles the way in which processed pseudogenes and retrogenes are generated (see Figure 9.12): reverse transcriptase converts an RNA transcript of the retrotransposon into a cDNA copy that then integrates into the genomic DNAs at a different location. Three major mammalian transposon classes use this copy-and-paste mechanism: long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and retrovirus-like elements containing long terminal repeats.

- *DNA transposons*. Members of this fourth class of transposon migrate directly without any copying of the sequence; the sequence is excised and then re-inserted elsewhere in the genome (a cut-and-paste mechanism).

Transposable elements that can transpose independently are described as *autonomous*; those that cannot are known as *nonautonomous* (**Figure 9.20**). Of the four classes of transposable element, LINEs and SINEs predominate; we describe them more fully below. The other two classes are briefly described here.
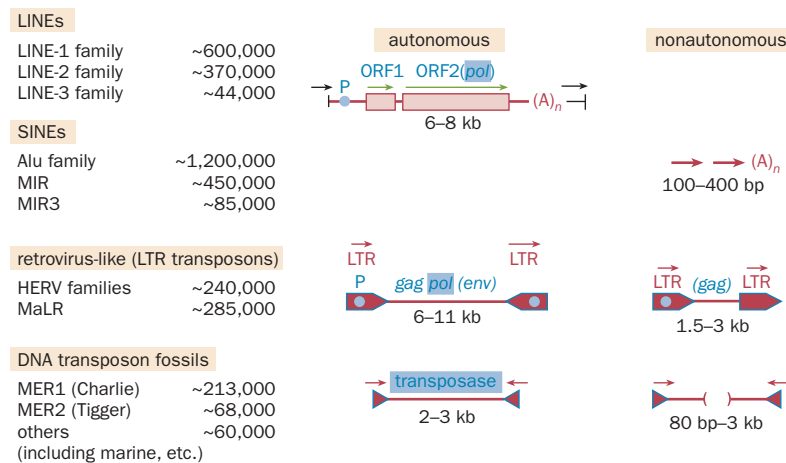
LINEs

LINE-1 family	~600,000
LINE-2 family	~370,000
LINE-3 family	~44,000

SINEs

Alu family	~1,200,000
MIR	~450,000
MIR3	~85,000

autonomous

ORF1 ORF2 (*pol*)
P
6–8 kb

nonautonomous

(A)$_n$
100–400 bp

retrovirus-like (LTR transposons)

| HERV families | ~240,000 |
| MaLR | ~285,000 |

LTR LTR
P *gag pol (env)*
6–11 kb

LTR (*gag*) LTR
1.5–3 kb

DNA transposon fossils

MER1 (Charlie)	~213,000
MER2 (Tigger)	~68,000
others	~60,000
(including marine, etc.)	

transposase
2–3 kb

80 bp–3 kb

Figure 9.20 Mammalian transposon families. Only a small proportion of members of any of the illustrated transposon families may be capable of transposing; many have lost such a capacity after acquiring inactivating mutations, and many are short truncated copies. Subclasses of the four main families are listed, along with sizes in base pairs. ORF, open reading frame. [Adapted from International Human Genome Sequencing Consortium (2001) *Nature* 409, 860–921. With permission from Macmillan Publishers Ltd.]

Human LTR transposons

LTR transposons include autonomous and nonautonomous retrovirus-like elements that are flanked by long terminal repeats (LTRs) containing necessary transcriptional regulatory elements. *Endogenous retroviral sequences* contain *gag* and *pol* genes, which encode a protease, reverse transcriptase, RNAse H, and integrase. They are thus able to transpose independently. There are three major classes of human endogenous retroviral sequence (HERV), with a cumulative copy number of about 240,000, accounting for a total of about 4.6% of the human genome (see Figure 9.20).

Very many HERVs are defective, and transposition has been extremely rare during the last several million years. However, the very small HERV-K group shows conservation of intact retroviral genes, and some members of the HERV-K10 subfamily have undergone transposition comparatively recently during evolution. Nonautonomous retrovirus-like elements lack the *pol* gene and often also the *gag* gene (the internal sequence having been lost by homologous recombination between the flanking LTRs). The MaLR family of such elements accounts for almost 4% or so of the genome.

Human DNA transposon fossils

DNA transposons have terminal inverted repeats and encode a transposase that regulates transposition. They account for close to 3% of the human genome and can be grouped into different classes that can be subdivided into many families with independent origins (see the Repbase database of repeat sequences at http://www.girinst.org/repbase/index.html). There are two major human families, MER1 and MER2, plus a variety of less frequent families (see Figure 9.20).

Virtually all the resident human DNA transposon sequences are no longer active; they are therefore transposon fossils. DNA transposons tend to have short lifespans within a species, unlike some of the other transposable elements such as LINEs. However, quite a few functional human genes seem to have originated from DNA transposons, notably genes encoding the RAG1 and RAG2 recombinases and the major centromere-binding protein CENPB.

A few human LINE-1 elements are active transposons and enable the transposition of other types of DNA sequence

LINEs (long interspersed nuclear elements) have been very successful transposons. They have a comparatively long evolutionary history, occurring in other mammals, including mice. As autonomous transposons, they can make all the products needed for retrotransposition, including the essential reverse transcriptase. Human LINEs consist of three distantly related families: LINE-1, LINE-2, and LINE-3, collectively comprising about 20% of the genome (see Figure 9.20). They are located primarily in euchromatic regions and are located preferentially in the dark AT-rich G bands (Giemsa-positive) of metaphase chromosomes.

Of the three human LINE families, LINE-1 (or L1) is the only family that continues to have actively transposing members. LINE-1 is the most important human transposable element and accounts for a higher fraction of genomic DNA (17%) than any other class of sequence in the genome.

Full-length LINE-1 elements are more than 6 kb long and encode two proteins: an RNA-binding protein and a protein with both endonuclease and reverse transcriptase activities (**Figure 9.21A**). Unusually, an internal promoter is located within the 5′ untranslated region. Full-length copies therefore bring with them their own promoter that can be used after integration in a permissive region of the genome. After translation, the LINE-1 RNA assembles with its own encoded proteins and moves to the nucleus.

To integrate into genomic DNA, the LINE-1 endonuclease cuts a DNA duplex on one strand, leaving a free 3′ OH group that serves as a primer for reverse transcription from the 3′ end of the LINE RNA. The endonuclease's preferred cleavage site is TTTT↓A; hence the preference for integrating into AT-rich regions. AT-rich DNA is comparatively gene-poor, and so because LINEs tend to integrate into AT-rich DNA they impose a lower mutational burden, making it easier for their host to accommodate them. During integration, the reverse transcription often fails to proceed to the 5′ end, resulting in truncated, nonfunctional insertions. Accordingly, most LINE-derived repeats are short, with an average size of 900 bp for all LINE-1 copies, and only about 1 in 100 copies are full length.

The LINE-1 machinery is responsible for most of the reverse transcription in the genome, allowing retrotransposition of the nonautonomous SINEs and also of copies of mRNA, giving rise to processed pseudogenes and retrogenes. Of the 6000 or so full-length LINE-1 sequences, about 80–100 are still capable of transposing, and they occasionally cause disease by disrupting gene function after insertion into an important conserved sequence.

Alu repeats are the most numerous human DNA elements and originated as copies of 7SL RNA

SINEs (short interspersed nuclear elements) are retrotransposons about 100–400 bp in length. They have been very successful in colonizing mammalian genomes, resulting in various interspersed DNA families, some with extremely high copy numbers. Unlike LINEs, SINEs do not encode any proteins and they cannot transpose independently. However, SINEs and LINEs share sequences at their 3′ end, and SINEs have been shown to be mobilized by neighboring LINEs. By parasitizing on the LINE element transposition machinery, SINEs can attain very high copy numbers.

The human Alu family is the most prominent SINE family in terms of copy number, and is the most abundant sequence in the human genome, occurring on average more than once every 3 kb. The full-length Alu repeat is about 280 bp long and consists of two tandem repeats, each about 120 bp in length followed by a short A_n/T_n sequence. The tandem repeats are asymmetric: one contains an internal 32 bp sequence that is lacking in the other (Figure 9.21B). Monomers,

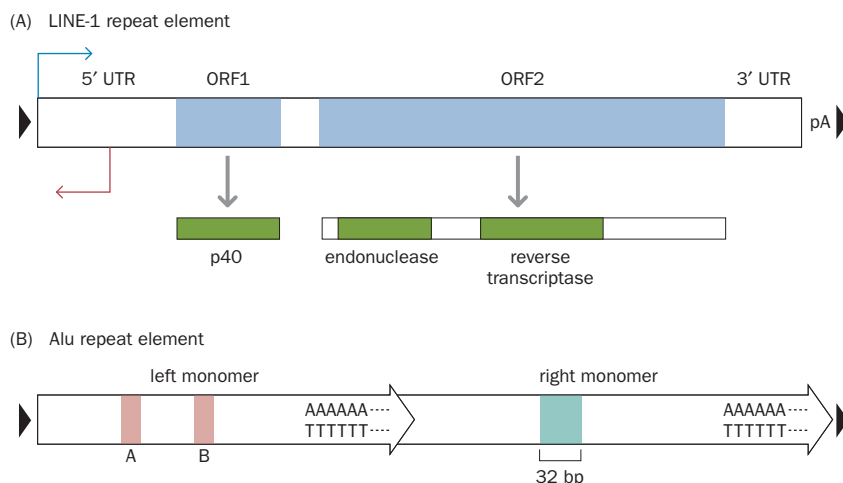(A) LINE-1 repeat element

(B) Alu repeat element

Figure 9.21 The human LINE-1 and Alu repeat elements. (A) The 6.1 kb LINE-1 element has two open reading frames: ORF1, a 1 kb open reading frame, encodes p40, an RNA-binding protein that has a nucleic acid chaperone activity; the 4 kb ORF2 specifies a protein with both endonuclease and reverse transcriptase activities. A bidirectional internal promoter lies within the 5′ untranslated region (UTR). At the other end, there is an A_n/T_n sequence, often described as the 3′ poly(A) tail (pA). The LINE-1 endonuclease cuts one strand of a DNA duplex, preferably within the sequence TTTT↓A, and the reverse transcriptase uses the released 3′-OH end to prime cDNA synthesis. New insertion sites are flanked by a small target site duplication of 2–20 bp (flanking black arrowheads). (B) An Alu dimer. The two monomers have similar sequences that terminate in an A_n/T_n sequence but differ in size because of the insertion of a 32 bp element within the larger repeat. Alu monomers also exist in the human genome, as do various truncated copies of both monomers and dimers.

containing only one of the two tandem repeats, and various truncated versions of dimers and monomers are also common, giving a genomewide average of 230 bp.

Whereas SINEs such as the MIR (mammalian-wide interspersed repeat) families are found in a wide range of mammals, the Alu family is of comparatively recent evolutionary origin and is found only in primates. However, Alu subfamilies of different evolutionary ages can be identified. In the past 5 million or so years since the divergence of humans and African apes, only about 5000 copies of the Alu repeat have undergone transposition; the most mobile Alu sequences are members of the Y and S subfamilies.

Like other mammalian SINEs, Alu repeats originated from cDNA copies of small RNAs transcribed by RNA polymerase III. Genes transcribed by RNA polymerase III often have internal promoters, and so cDNA copies of transcripts carry with them their own promoter sequences. Both the Alu repeat and, independently, the mouse B1 repeat originated from cDNA copies of 7SL RNA, the short RNA that is a component of the signal recognition particle, using a retrotransposition mechanism like that shown in Figure 9.12. Other SINEs, such as the mouse B2 repeat, are retrotransposed copies of tRNA sequences.

Alu repeats have a relatively high GC content and, although dispersed mainly throughout the euchromatic regions of the genome, are preferentially located in the GC-rich and gene-rich R chromosome bands, in striking contrast to the preferential location of LINEs in AT-rich DNA. However, when located within genes they are, like LINE-1 elements, confined to introns and the untranslated regions. Despite the tendency to be located in GC-rich DNA, newly transposing Alu repeats show a preference for AT-rich DNA, but progressively older Alu repeats show a progressively stronger bias toward GC-rich DNA.

The bias in the overall distribution of Alu repeats toward GC-rich and, accordingly, gene-rich regions must result from strong selection pressure. It suggests that Alu repeats are not just genome parasites but are making a useful contribution to cells containing them. Some Alu sequences are known to be actively transcribed and may have been recruited to a useful function. The *BCYRN1* gene, which encodes the BC200 neural cytoplasmic RNA, arose from an Alu monomer and is one of the few Alu sequences that are transcriptionally active under normal circumstances. In addition, the Alu repeat has recently been shown to act as a *trans*-acting transcriptional repressor during the cellular heat shock response.

CONCLUSION

In this chapter, we have looked at the architecture of the human genome. Each human cell contains many copies of a small, circular mitochondrial genome and just one copy of the much larger nuclear genome. Whereas the mitochondrial genome bears some similarities to the compact genomes of prokaryotes, the human nuclear genome is much more complex in its organization, with only 1.1% of the genome encoding proteins and 95% comprising nonconserved, and often highly repetitive, DNA sequences.

Sequencing of the human genome has revealed that, contrary to expectation, there are comparatively few protein-coding genes—about 20,000–21,000 according to the most recent estimates. These genes vary widely in size and internal organization, with the coding exons often separated by large introns, which often contain highly repetitive DNA sequences. The distribution of genes across the genome is uneven, with some functionally and structurally related genes found in clusters, suggesting that they arose by duplication of individual genes or larger segments of DNA. Pseudogenes can be formed when a gene is duplicated and then one of the pair accumulates deleterious mutations, preventing its expression. Other pseudogenes arise when an RNA transcript is reverse transcribed and the cDNA is re-inserted into the genome.

The biggest surprise of the post-genome era is the number and variety of non-protein-coding RNAs transcribed from the human genome. At least 85% of the euchromatic genome is now known to be transcribed. The familiar ncRNAs known to have a role in protein synthesis have been joined by others that have roles in gene regulation, including several prolific classes of tiny regulatory RNAs and thousands of different long ncRNAs. Our traditional view of the genome is being radically revised.

In Chapter 10 we describe how the human genome compares with other genomes, and how evolution has shaped it. Aspects of human gene expression are elaborated in Chapter 11. Within Chapter 13 we also consider human genome variation.

FURTHER READING

Human mitochondrial genome

Anderson S, Bankier AT, Barrell BG et al. (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.

Chen XJ & Butow RA (2005) The organization and inheritance of the mitochondrial genome. *Nat. Rev. Genet*. 6, 815–825.

Falkenberg M, Larsson NG & Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. *Annu. Rev. Biochem*. 76, 679–699.

MITOMAP: human mitochondrial genome database. http://www .mitomap.org

Wallace DC (2007) Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. *Annu. Rev. Biochem*. 76, 781–821.

Human nuclear genome

Clamp M, Fry B, Kamal M et al. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* 104, 19428–19433.

Ensembl human database. http://www.ensembl.org/Homo _sapiens/index.html

GeneCards human gene database. http://www.genecards.org

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945.

Nature Collections: Human Genome Supplement, 1 June 2006 issue. [A collation that includes papers analyzing the sequence of each chromosome plus reprints of the papers reporting the 2001 draft sequence and the 2004 finished euchromatic sequence, available electronically at http://www.nature.com/nature/supplements/collections /humangenome/]

NCBI Human Genome Resources. http://www.ncbi.nlm.nih.gov /projects/genome/guide/human/

UCSC Genome Browser, Human (*Homo sapiens*) Genome Browser Gateway. http://genome.ucsc.edu/cgi-bin/hgGateway

Organization of protein-coding genes

Adachi N & Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell* 109, 807–809.

Li YY, Yu H, Guo ZM et al. (2006) Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput. Biol*. 2, e74.

Sanna CR, Li W-H & Zhang L (2008) Overlapping genes in the human and mouse genomes. *BMC Genomics* 9, 169.

Soldà G, Suyama M, Pelucchi P et al. (2008) Non-random retention of protein-coding overlapping genes in Metazoa. *BMC Genomics* 9, 174.

Gene duplication, segmental duplication, and copy-number variation

Bailey JA, Gu Z, Clark RA et al. (2002) Recent segmental duplications in the human genome. *Science* 297, 1003–1007.

Conrad B & Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu. Rev. Genomics Hum. Genet*. 8, 17–35.

Kaessmann H, Vinckenbosch N & Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet*. 10, 19–31.

Linardopoulou EV, Williams EM, Fan Y et al. (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* 437, 94–100.

Redon R, Ishikawa S, Fitch KR et al. (2006) Global variation in copy number in the human genome. *Nature* 444, 444–454.

Tuzun E, Sharp AJ, Bailey JA et al. (2005) Fine-scale structural variation of the human genome. *Nat. Genet*. 37, 727–732.

The complexity of the mammalian transcriptome and the need to redefine genes in the post-genome sequencing era

Gerstein MB, Bruce C, Rozowsky JS et al. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res*. 17, 669–681.

Gingeras T (2007) Origin of phenotypes: genes and transcripts. *Genome Res*. 17, 682–690.

Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet*. 10, 833–844.

Kapranov P, Cheng J, Dike S et al. (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488.

General reviews of noncoding RNA

Amaral PP, Dinger ME, Mercer TR & Mattick JS (2008) The eukaryotic genome as an RNA machine. *Science* 319, 1787–1789.

Carninci P, Yasuda J & Hayashizaki Y (2008) Multifaceted mammalian transcriptome. *Curr. Opin. Cell Biol*. 20, 274–280.

Griffiths-Jones S (2007) Annotating non-coding RNA genes. *Annu. Rev. Genomics Hum. Genet*. 8, 279–298.

Marakova JA & Kramerov DA (2007) Non-coding RNAs. *Biochemistry (Moscow)* 72, 1161–1178.

Mattick JS (2009) The genetic signatures of noncoding RNAs. *PLoS Genet*. 5, e1000459.

Prasanth KV & Spector DL (2007) Eukaryotic regulatory RNAs: an answer to the genome complexity conundrum. *Genes Dev*. 21, 11–42.

Small nuclear RNA and small nucleolar RNAs

Kishore S & Stamm S (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* 311, 230–232.

Matera AG, Terns RM & Terns MP (2007) Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol*. 8, 209–220.

Sahoo T, del Gaudio D, German JR et al. (2008) Prader–Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nuclear RNA cluster. *Nat. Genet*. 40, 719–721.

MicroRNAs and ncRNAs as developmental regulators

Bushati N & Cohen SM (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol*. 23, 175–205.

Chang T-C & Mendell JT (2007) microRNAs in vertebrate physiology and disease. *Annu. Rev. Genomics Hum. Genet*. 8, 215–239.

Makeyev EV & Maniatis T (2008) Multilevel regulation of gene expression by microRNAs. *Science* 319, 1789–1790.

Rinn JL, Kertesz M, Wang JK et al. (2007) Functional demarcation of active and silent chromatin domains in human *HOX* loci by non-coding RNAs. *Cell* 129, 1311–1323.

Stefani G & Slack F (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol*. 9, 219–230.

piRNAs and endogenous siRNAs

Aravin AA, Sachidanandam R, Girard A et al. (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316, 744–747.

Girard A & Hannon GJ (2007) Conserved themes in small RNA-mediated transposon control. *Trends Cell Biol*. 18, 136–148.

Tam OH, Aravin AA, Stein P et al. (2008) Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* 453, 534–538.

Watanabe T, Totoki Y, Toyoda A et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 453, 539–543.

Antisense and long noncoding regulatory RNAs

He Y, Vogelstein B, Velculescu VE et al. (2008) The antisense transcriptomes of human cells. *Science* 322, 1855–1858.

Khalil AM, Guttman M, Huarte M et al. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* 106, 11667–11672.

Ponting CP, Oliver PL & Reik W (2009) Evolution and function of long noncoding RNAs. *Cell* 136, 629–641.

Wilusz JE, Sunwoo H & Spector DL (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev*. 23, 1494–1504.

Promoter- and termini-associated RNAs

Affymetrix/Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009) Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* 457, 1028–1042.

Pseudogenes and retrogenes

D'Errico L, Gadaleta G & Saccone C (2004) Pseudogenes in metazoa: origins and features. *Brief. Funct. Genomic. Proteomic*. 3, 157–167.

Duret L, Chureau C, Samain S et al. (2006) The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312, 1653–1655.

Sasidharan R & Gerstein M (2008) Protein fossils live on as RNA. *Nature* 453, 729–731.

Zhang D & Gerstein MB (2004) Large-scale analysis of pseudogenes in the human genome. *Curr. Opin. Genet. Dev*. 14, 328–335.

Zheng D & Gerstein MB (2007) The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet*. 23, 219–224.

Heterochromatin and transposon-based repeats

Choo KH, Vissel B, Nagy A et al. (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res*. 19, 1179–1182.

Faulkner GJ, Kimura Y, Daub CO et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet*. 41, 563–571.

Henikoff S, Ahmad K & Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293, 1098–1102.

Mariner PD, Walters RD, Espinoza CA et al. (2008) Human Alu RNA is a modular transacting repressor of mRNA transcription during heat shock. *Mol. Cell* 29, 499–509.

Mills RE, Bennett EA, Iskow RC & Devine SE (2007) Which transposable elements are active in the human genome? *Trends Genet*. 23, 183–191.

Muotri AR, Marchetto MCN, Coufal NG & Gage FH (2007) The necessary junk: new functions for transposable elements. *Hum. Mol. Genet*. 16, R159–R167.

Repbase: database of repeat sequences. http://www.girinst.org/repbase/index.html

Wicker T, Sabot F, Hua-Van A et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet*. 8, 973–983.

Yang N & Kazazian HH Jr (2006) L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat. Struct. Mol. Biol*. 13, 763–771.