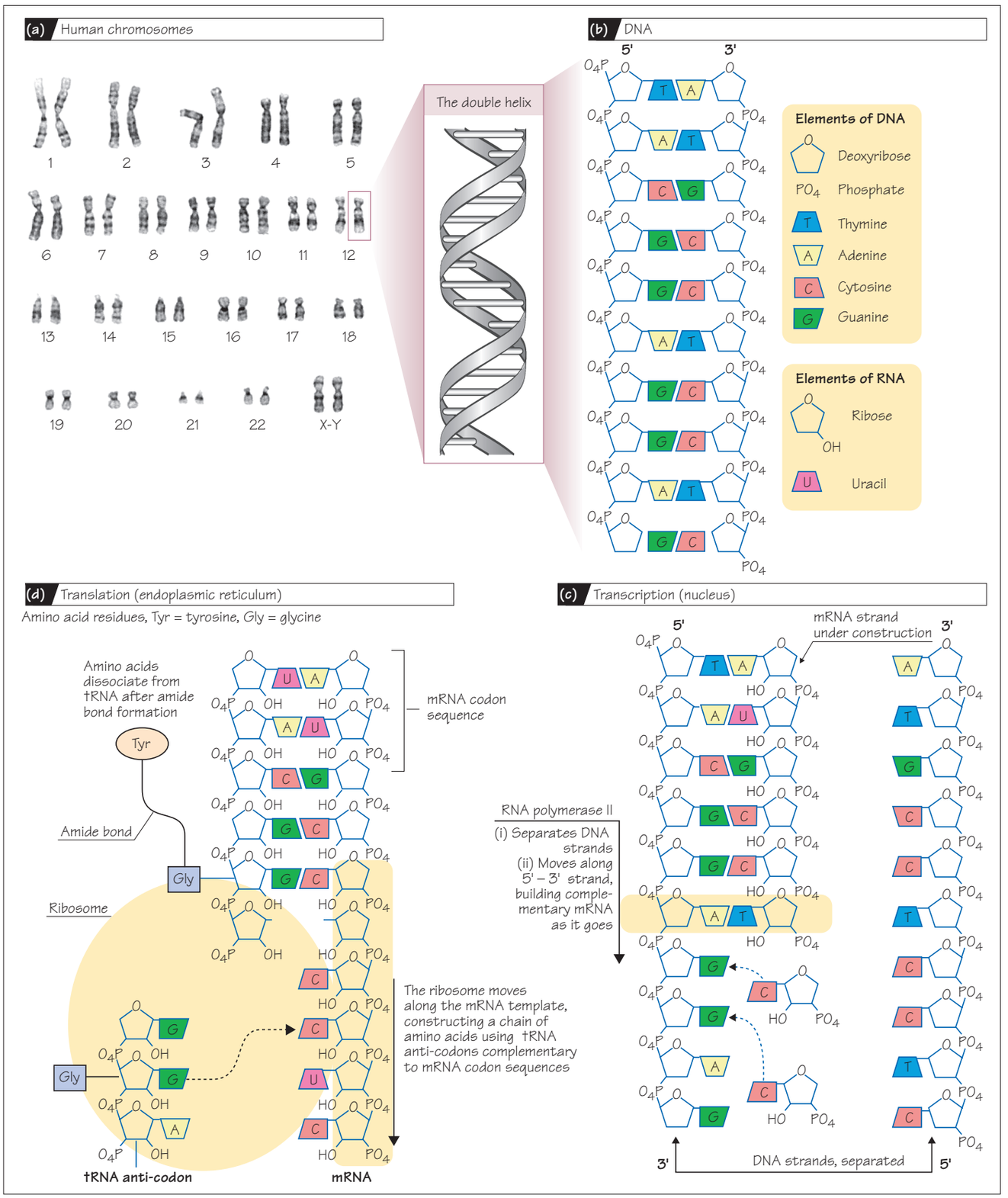
1 Physiology and the genome



DNA and genes

The genetic heritage of *Homo sapiens* amounts to some 30 000 genes packed into 44 somatic chromosomes arranged in pairs, plus the sex chromosomes X and Y (Fig. 1a). Multicellular organisms, such as ourselves, are made up of **eukaryotic** cells, which means that the genes are stored in the cell **nucleus** (Chapter 4). The genetic material comprises **deoxyribonucleic acid** (**DNA**; Fig. 1b), an alternating chain of phosphate and sugar (deoxyribose) residues, with **nucleotide bases** attached to each sugar moiety. This molecule self-organizes into a helix made up of two polarized, complementary strands that are arranged so that the phosphate terminal (**5′ terminus**) of one is opposite the sugar terminal (**3′ terminus**) of the other.

The two filaments of the helix are held together by hydrogen bonds between pairs of bases. The nucleotides found in DNA are **adenine** (**A**), **cytosine** (**C**), **guanine** (**G**) and **thymine** (**T**). An adenine on one strand of the helix will always be found opposite a thymine residue on the other, whereas cytosine is always found opposite guanine. The human genome comprises some 3.2×10^9 (3.2 billion) of these **base pairs**. The linear arrangement of bases forms the genetic code or **genome**. The working draft of the human genome that was announced in 2001 describes the sequences of base pairs in each chromosome. This information alone does not tell us where all the genes are located, as only some 1.5% of the genome is accounted for by translatable code. Each gene codes for a specific protein, and a large number of gene sequences have been identified by the process of **cloning**. The average length of these sequences in humans is roughly 27 000 base pairs, even though an average-sized protein requires a code of not much more than 1000 base pairs. The useful DNA code is contained within short sequences, known as **exons**, but these are intermingled with much longer sections, known as **introns**, that do not code for protein. In addition, each gene has a number of regulatory sequences that allow the gene to be switched on and off and indicate where the gene begins and ends. Not only does each gene contain large sections of non-coding DNA, but between working genes there are interposed long sequences of what appears to be non-useful, so-called 'mobile' material that has the ability to replicate itself for insertion at points along the DNA molecule.

Gene transcription and translation

The decoding of a gene into a protein is a two-stage process involving **transcription** of DNA into **messenger ribonucleic acid** (**mRNA**), followed by **translation** of mRNA into protein. In eukaryotic organisms, the DNA code is read by an enzyme known as **RNA polymerase II** which constructs sequences of mRNA from the DNA template. RNA polymerase II attaches to DNA with the help of a set of proteins, known as **general transcription factors**. These factors identify the starting sequence of a gene (the **promoter** region), unravel the chromosomal material (within which DNA is very tightly coiled) and unzip a portion of the paired helix to allow the polymerase to bind. Only one strand of the DNA is read at any one time, and transcription always begins at the 5′ end of a sequence. RNA polymerase II moves along the gene sequence base by base, creating an mRNA strand that is complementary to the original DNA (Fig. 1c). In RNA, the sugar residue in the backbone is ribose, and the base **uracil** (**U**) substitutes for thymine. Thus, mRNA is constructed with a uracil for each adenine on the DNA, a cytosine for a guanine, and so on, until the sequence indicating the end of the gene is reached. During the process, any RNA derived from introns is removed and transcription errors are corrected. Once complete, the mRNA strand moves from the nucleus to the **endoplasmic reticulum** (Chapter 4) where structures known as **ribosomes** convert the RNA message into protein. Within RNA, strings of three consecutive bases (**codons**) code for one of the 20 amino acids used in human proteins, or act as start or stop signals. **Transfer RNA** (**tRNA**) exists in three-base sequences, known as **anticodons**, which are complementary to the codons present in the mRNA (Fig. 1d). Each type of tRNA binds a specific amino acid so that, as the ribosome moves along the mRNA strand, a chain of amino acids is formed that mirrors the codon sequence of the RNA message.

What the genome tells us

The protein products of gene transcription organize themselves and other biological molecules into living organisms. The functions of these proteins and the assemblies they produce are at least as important as the genome itself in determining the nature and operational characteristics of organisms. The mouse genome is more than 90% homologous to that of humans, yet there are some striking differences between the two species, some of which must be explained by postgenomic processes. Although the sequencing of the genome is a great achievement, in isolation it tells us nothing about the functioning of the human organism. Such insight can only come from investigations on higher level systems, taking advantage of the new opportunities provided by our knowledge of the code. This is modern physiology.