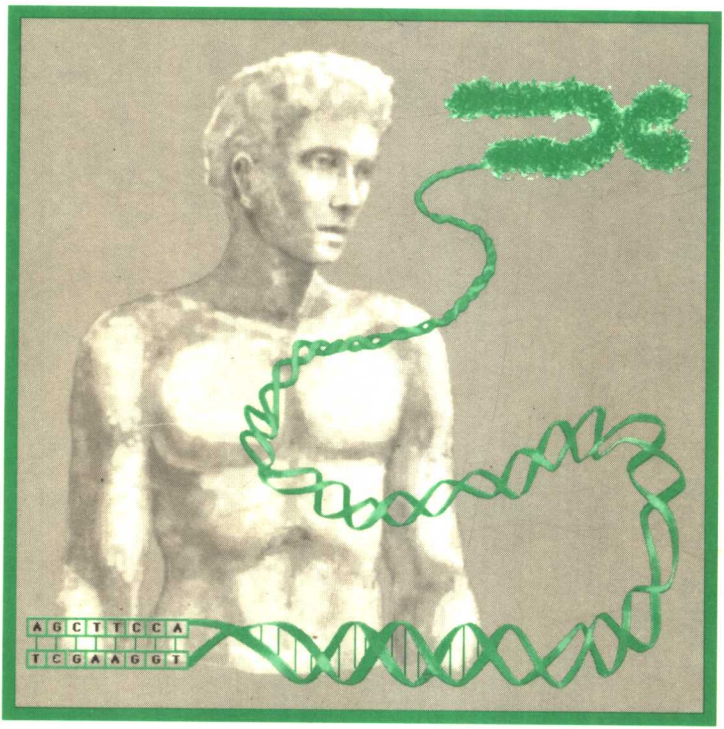
Medical Perspectives Series

Editors: Andrew P. Read & Terence Brown

The Human Genome

T. Strachan



BIOS SCIENTIFIC PUBLISHERS

THE HUMAN GENOME

T. Strachan
University Department of Medical Genetics, St Mary's Hospital, Hathersage Road, Manchester M13 0JH, U.K.

BIOS SCIENTIFIC PUBLISHERS

© BIOS Scientific Publishers Limited, 1992

All rights reserved. No part of this book may be reproduced or transmitted, in any form or by any means, without permission.

First published in the United Kingdom 1992 by
BIOS Scientific Publishers Limited,
St Thomas House, Becket Street, Oxford OX1 1SJ.

A CIP catalogue record for this book is available from the British Library.

ISBN 1 872 748 80 5

To my family and the memory of Hugh M. Strachan, 1947–1979

Typeset by Enset Photosetting Limited, Bath, U.K.
Printed by Information Press Ltd, Oxford, U.K.

PREFACE

The last few years have witnessed extraordinary advances in our understanding of the human genome. The genes which underlie many important inherited disorders such as cystic fibrosis and Duchenne muscular dystrophy have recently been isolated and studied, as have many genes which cause human cancers. These developments have led to improved diagnosis of genetic disease and to a greatly increased understanding of the molecular basis of single gene disorders.

Currently, much effort is being devoted to identifying genes implicated in the pathogenesis of common multifactorial disorders such as cancer, heart disease, mental illness, etc. In addition, gene technology has begun to be applied to devising new treatments for human disease. Against this background, the Human Genome project, one of the most ambitious scientific endeavors ever undertaken, has recently been initiated with the ultimate aim of isolating and characterizing each of the 50 000 to 100 000 genes in the human genome. The aim of this book is to provide non-specialist and specialist alike with a concise description of our current knowledge of the human genome and the ways in which it is influencing medical research and practice.

My thanks go to the series editors, Drs Andrew Read and Terry Brown for their helpful comments on the manuscript, and to my colleagues Paul Sinnott, Paul Sinclair, Carolyn Watson and Andrew Wallace for providing photos reproduced in Figures 3.8, 4.9, 4.10 and 6.8, respectively.

T. Strachan

ABBREVIATIONS

ADA	adenosine deaminase
APC	adenomatous polyposis coli
ARMS	amplification refractory mutation system
ARS	autonomously replicating sequence
ASO	allele-specific oligonucleotide
bp	base pair
CF	cystic fibrosis
CFTR	cystic fibrosis transmembrane regulator
cM	centimorgan
DGGE	denaturing gradient gel electrophoresis
DMD	Duchenne muscular dystrophy
FAP	familial adenomatous polyposis
FISH	fluorescence *in-situ* hybridization
HLA	human leukocyte antigen complex
HUGO	Human Genome Organization
Ig	immunoglobulin
kb	kilobase
LCR	locus control region
LDL	low density lipoprotein
LINE	long interspersed nuclear element
Mb	megabase
mRNA	messenger RNA
MHC	major histocompatibility complex
mt	mitochondrial
NF1	neurofibromatosis type I
PCR	polymerase chain reaction
PFGE	pulsed field gel electrophoresis
PIC	polymorphism information content
rDNA/RNA	ribosomal DNA/RNA
RFLP	restriction fragment length polymorphism
RSP	restriction site polymorphism
SINE	short interspersed nuclear element
SnRNA	small nuclear RNA
SSCP	single-strand conformation polymorphism
STS	sequence-tagged site
TIL	tumor infiltrating lymphocytes
TNF	tumor necrosis factor
tRNA	transfer RNA
VNTR	variable number of tandem repeats
YAC	yeast artificial chromosome

此为试读,需要完整PDF请访问: www.ertongbook.com

CONTENTS

1
ORGANIZATION AND EXPRESSION OF THE HUMAN GENOME

1.1 Structure of genomic DNA

The human genome consists of DNA molecules in the form of a double helix in which the two strands of the DNA duplex are held together by weak hydrogen bonds. Each strand has a linear backbone of residues of deoxyribose (a 5-carbon sugar) which are linked by covalent phosphodiester bonds. Covalently attached to carbon atom number 1' of each sugar residue is a nitrogenous base, either a pyrimidine (cytosine or thymine), or a purine (adenine or guanine; see *Figure 1.1*). A sugar with an attached base and phosphate group therefore constitutes the basic repeat unit of a DNA strand, a nucleotide. As the phosphodiester bonds link carbon atoms number 3' and number 5' of successive sugar residues, one end of each DNA strand, the so-called 5' end, will have a terminal sugar residue in which carbon atom number 5' is not linked to a neighboring sugar residue. The other end is defined as the 3' end because of a similar absence of phosphodiester bonding at carbon atom number 3' of the terminal sugar residue. The two strands of a DNA duplex always associate (anneal) in such a way that the 5'→3' direction of one DNA strand is the opposite to that of its partner.

Genetic information is encoded by the sequence of bases in the DNA strands. Hydrogen bonding occurs between laterally opposed bases, base pairs, of the two strands of a DNA duplex according to Watson–Crick rules: adenine (A) specifically binds to thymidine (T) and cytosine (C) specifically binds to guanine (G). Consequently, the two strands of a DNA duplex are said to be complementary (or to exhibit base complementarity) and the sequence of bases of one DNA strand can readily be determined if the DNA sequence of its complementary strand is already known. It is usual, therefore, to describe a DNA sequence by writing the sequence of bases of one strand only, and in the 5'→3' direction, which is the direction of synthesis of new DNA molecules during DNA replication and also of transcription when RNA molecules are synthesized using DNA as a template. However, when describing the sequence of a DNA region encompassing two neighboring bases (really a dinucleotide) on one DNA strand, it is usual to insert a 'p' to denote a connecting phosphodiester bond. For example, CpG means that a cytosine is covalently linked to a neighboring guanine on the same DNA strand, while a CG base pair means a cytosine on one DNA strand is hydrogen bonded to a guanine on the complementary strand.

In the process of DNA replication, the two DNA strands unwind and each strand directs the synthesis of a complementary DNA strand to generate two daughter DNA

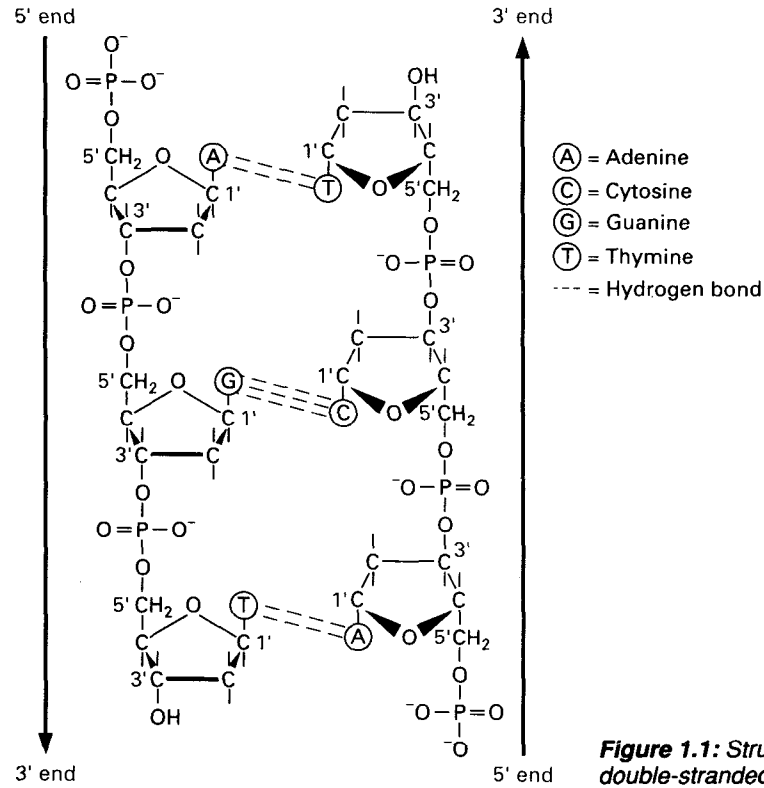


Figure 1.1: *Structure of double-stranded DNA.*

duplexes which are identical to the parent molecule. However, during gene expression, individual genes are transcribed from only one of the two available DNA strands. The strand from which the RNA is transcribed, and which is complementary in base sequence to the RNA molecule, is the so-called template strand (or anti-sense strand). The transcribed single-stranded RNA molecule is therefore a faithful copy of the other DNA strand of the DNA duplex, the sense strand, except that the sugar in the RNA molecule is ribose, and uracil (U) bases replace the original thymines. In the case of documented gene sequences it is customary to show only the DNA sequence of the sense strand. Orientation of sequences relative to a gene sequence is commonly dictated by the sense strand and by the direction of transcription. For example, the 5' end of a gene refers to the DNA sequence at the 5' end of the sense strand, and sequences upstream or downstream of a gene refer to sequences which flank the gene at the 5' or 3' ends of the sense strand respectively.

1.2 The nuclear and mitochondrial genomes

The genetic information in human cells is organized in the form of two genomes, a complex nuclear genome and a simple mitochondrial genome. The difference in complexity of the two genomes reflects the predominance of the nuclear genome in providing the great bulk of essential genetic information, most of which is ultimately decoded to specify polypeptide synthesis on cytoplasmic ribosomes. Although

f252065f

mitochondria possess their own ribosomes, the mitochondrial genome specifies only a very small portion of the specific mitochondrial functions; the bulk of the mitochondrial polypeptides are encoded by nuclear genes and are imported from the cytoplasm.

1.2.1 The nuclear genome

The nuclear DNA content of individual human cells is determined by the number of nuclei and the number of chromosomes in that cell. The specialized germ cells, eggs and sperm cells, are haploid cells in which there is a single copy of the nuclear genome with the DNA being distributed between 23 chromosomes, comprising 22 autosomes and a single sex chromosome, X or Y. Fusion of a normal egg cell and sperm cell at conception generates a diploid zygote with two genome copies (2C) and 46 chromosomes, consisting of 23 pairs of homologous chromosomes, that is, an homologous pair of each of the 22 autosomes and two sex chromosomes which may be completely homologous (XX), or partially homologous (XY).

Table 1.1: *DNA content of human chromosomes*[a]

Chromosome	Percentage of total length	Amount of DNA (Mb)	Chromosome	Percentage of total length	Amount of DNA (Mb)
1	8.3	250	13	3.6	110
2	7.9	240	14	3.5	105
3	6.4	190	15	3.3	100
4	6.1	180	16	2.8	85
5	5.8	175	17	2.7	80
6	5.5	165	18	2.5	75
7	5.1	155	19	2.3	70
8	4.5	135	20	2.1	65
9	4.4	130	21	1.8	55
10	4.4	130	22	1.9	60
11	4.4	130	X	4.7	140
12	4.1	120	Y	2.0	60

[a]The DNA content is given for chromosomes prior to entering the S (DNA replication) phase of cell division (see *Figure 1.3*); data abstracted from reference 1.

Subsequent mitotic DNA duplication and cell division events during development and growth result in the great majority of somatic cells containing a single diploid nucleus. Exceptions include examples of naturally polyploid cells in which there are additional rounds of chromosome duplication prior to cell division (e.g. certain liver cells normally each have 92 chromosomes and are tetraploid, i.e. have four copies of the haploid genome) and cells which are multinucleated (e.g. fully differentiated muscle cells) or which lack a nucleus (e.g. mature red blood cells). Additionally, although the number of chromosomes in a normal diploid somatic cell remains 46 until the anaphase stage of mitosis, the cell becomes effectively tetraploid following DNA duplication during the earlier S phase of the cell cycle as described below.

By and large, somatic cells carry the same genetic information as the zygote from which they arise. Except for sequences in the non-homologous regions of the X and Y chromosomes in males (see Section 2.2), diploid cells contain two copies of each individual nuclear gene (or DNA sequence) normally present in haploid cells. A pair of such homologous DNA sequences, which are located at identical positions on homologous chromosomes (i.e. at the same locus) are referred to as alleles. Most nuclear DNA sequences show Mendelian inheritance, whereby in diploid cells one allele is inherited from each parent. An individual is said to be homozygous or heterozygous at a specific locus if the two alleles at that locus are, respectively, identical or different in sequence. As Y chromosomes are transmitted exclusively by males, Y chromosome sequences are present as single copies in male diploid cells, as are X chromosome sequences.

The nuclear genome of a human haploid cell contains about 3×10^9 bp of DNA and an average size chromosome has approximately 1.3×10^8 bp (or 130 megabases) of DNA but can vary between approximately 50 Mb and 250 Mb (see *Table 1.1*). The DNA content of each chromosome is thought to consist of a single linear double-stranded DNA molecule which, if fully uncoiled, would be between 1.7 and 8.5 cm long. In the cell the structure of each chromosome is highly ordered [2] and compaction of the chromosomal DNA is achieved by complexing with various DNA-binding proteins. The most fundamental unit of packaging is the nucleosome which consists of a central core complex of eight basic histone proteins (two each of histones H2A, H2B, H3 and H4) around which a stretch of 146 bp of double-stranded DNA is coiled in 1.75 turns (*Figure 1.2*). Adjacent nucleosomes are connected by a short length of spacer DNA. The elementary fiber of linked nucleosomes is in turn coiled into a chromatin fiber of 30 nm diameter which can be resolved by electron microscopy.

At the metaphase stage of cell division the chromosomes become even more condensed and can be resolved by optical microscopy as structures which are over 1 μm wide and range in length from 2 μm (chromosome 21) to 10 μm (chromosome 1). At this stage the DNA in the chromosome arms, but not that in the centromere, has already duplicated in preparation for cell division. The metaphase chromosome consists of two laterally opposed chromatids which remain bound to each other at the centromere. Each chromatid consists of loops of chromatin fiber, containing approximately 30–90 kb of DNA per loop, which are attached to a central scaffold of non-histone acidic protein and the resulting complex is further compacted by coiling. As each chromatid contains double-stranded DNA (except for the centromere DNA), the genomic DNA in a human somatic cell from late S phase right up to the anaphase stage of mitosis is effectively tetraploid (4C), although the chromosome number is still 46 (*Figure 1.3*).

A variety of treatments cause chromosomes in dividing cells to appear as a series of alternating light and dark staining bands. In G-banding, for example, the chromosomes are subjected to controlled digestion with trypsin before staining with the DNA-binding chemical Giemsa which reveals alternating positively (dark G-bands) and negatively staining regions (pale G-bands). Bands are classified according to their relative location on the short arm (p) or long arm (q) of specific chromosomes. For example, 17p12 means sub-band 2 of band 1 on the short arm of chromosome 17. Further sub-division is also possible: 17q21 can be divided into 17q21.1, 17q21.2 and 17q21.3 (*Figure 1.2*). As Giemsa shows preferential binding to A + T rich DNA

此为试读,需要完整PDF请访问: www.ertongbook.com

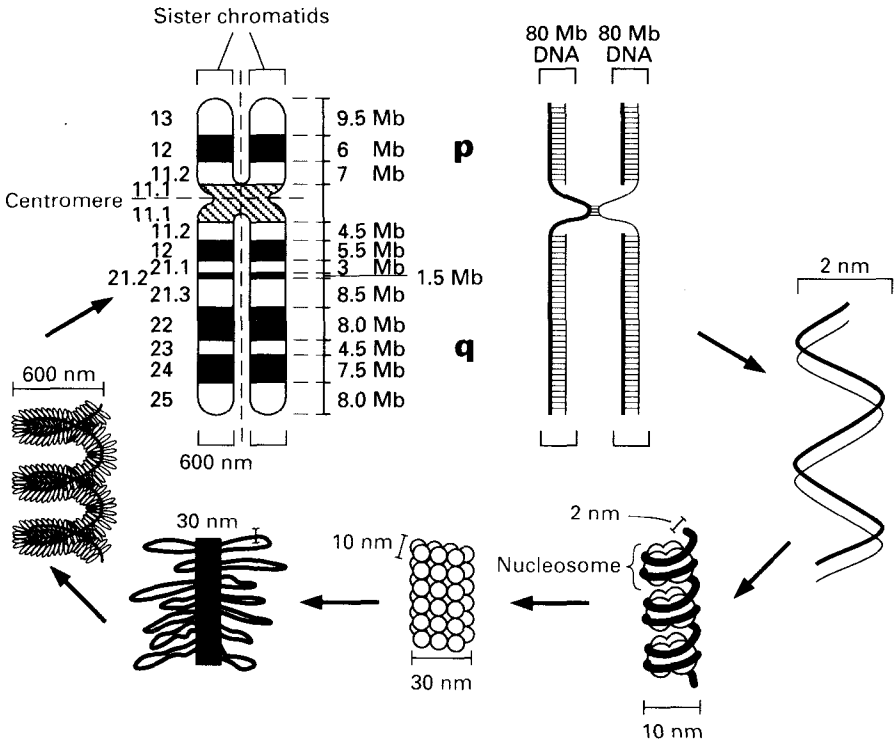


Figure 1.2: *From DNA duplex to metaphase chromosome (human chromosome 17, Giemsa-stained, 550 band preparation).*

sequences, the dark G-bands have been considered to be rich in A + T bases, while the pale G-bands are rich in G + C. Although, therefore, the average base composition of human genomic DNA is about 40% G + C, the alternating pale and light bands are thought to reflect the compartmentalization of the human genome into isochores, defined chromosomal regions in which the base composition of the DNA is comparatively homogeneous but which is variable between isochores [3]. The dark G-bands are thought to be deficient in genes (see below); during the cell cycle these regions condense early but replicate late. In contrast, pale G-bands represent regions which condense late, but replicate early; they are rich in G + C and in genes. Additionally, the two types of chromosomal regions differ in their predominant association with particular classes of interspersed repetitive DNA (see Section 1.8.4). At the resolution of approximately 550 bands per set of mitotic metaphase chromosomes (karyogram–see *Figure 1.4*), an average size band corresponds to approximately 6 Mb of DNA.

In the haploid nuclear genome the total number of genes is thought to be approximately 50 000–100 000. On this basis all nucleated cells have, on average, one gene per 30–60 kb, and about 2000–4000 per average chromosome. In a 400-band mitotic metaphase karyogram, one might anticipate about 100–200 genes on average per band. However, as noted above, average gene density is dependent on the base composition of the chromosomal region containing the gene and pale G-bands are relatively enriched in genes at the expense of dark G-bands.

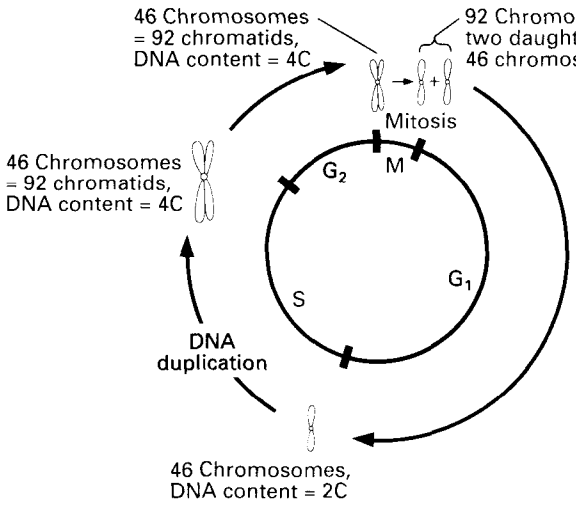


Figure 1.3: *Human chromosomal DNA content during the cell cycle.*

1.2.2 The mitochondrial genome

The mitochondrial genome is defined by a single type of circular double-stranded DNA molecule, 16 569 bp long, which has been completely sequenced [4]. 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 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, while males cannot transmit their mitochondria to subsequent generations.

Most human cells contain several hundred mitochondria. During mitotic cell division, the mitochondria of the dividing cell segregate in a purely random way to the two daughter cells. In each mitochondrion there are between about two and ten copies of the approximately 16.6 kb mitochondrial DNA molecule. Accordingly, although a single mitochondrial DNA molecule has only about 1/8000 as much DNA as an average sized chromosome, the total mitochondrial complement can account for up to about 0.5% of the DNA in a nucleated somatic cell. Although the mitochondrial DNA is principally double-stranded, a small section, the D loop, has a triple DNA strand structure due to the additional synthesis of a segment of mitochondrial DNA, 7S DNA (see *Figure 1.5*).

The 16.6 kb human mitochondrial DNA molecule contains 37 genes, 28 of which are encoded by the heavy (H) DNA strand, which is rich in guanines, and nine by the light (L) DNA strand (*Figure 1.5*). Of the 37 mitochondrial genes, 13 encode polypeptides which, together with the products of at least 50 nuclear genes, account for four of the five respiratory complexes, the multichain enzymes of oxidative phosphorylation, which are engaged in the production of ATP. The sub-units of the fifth respiratory complex, succinate–CoQ reductase, and all other mitochondrial proteins are encoded exclusively by the nuclear genome. The remaining 24 mitochondrial genes encode 22

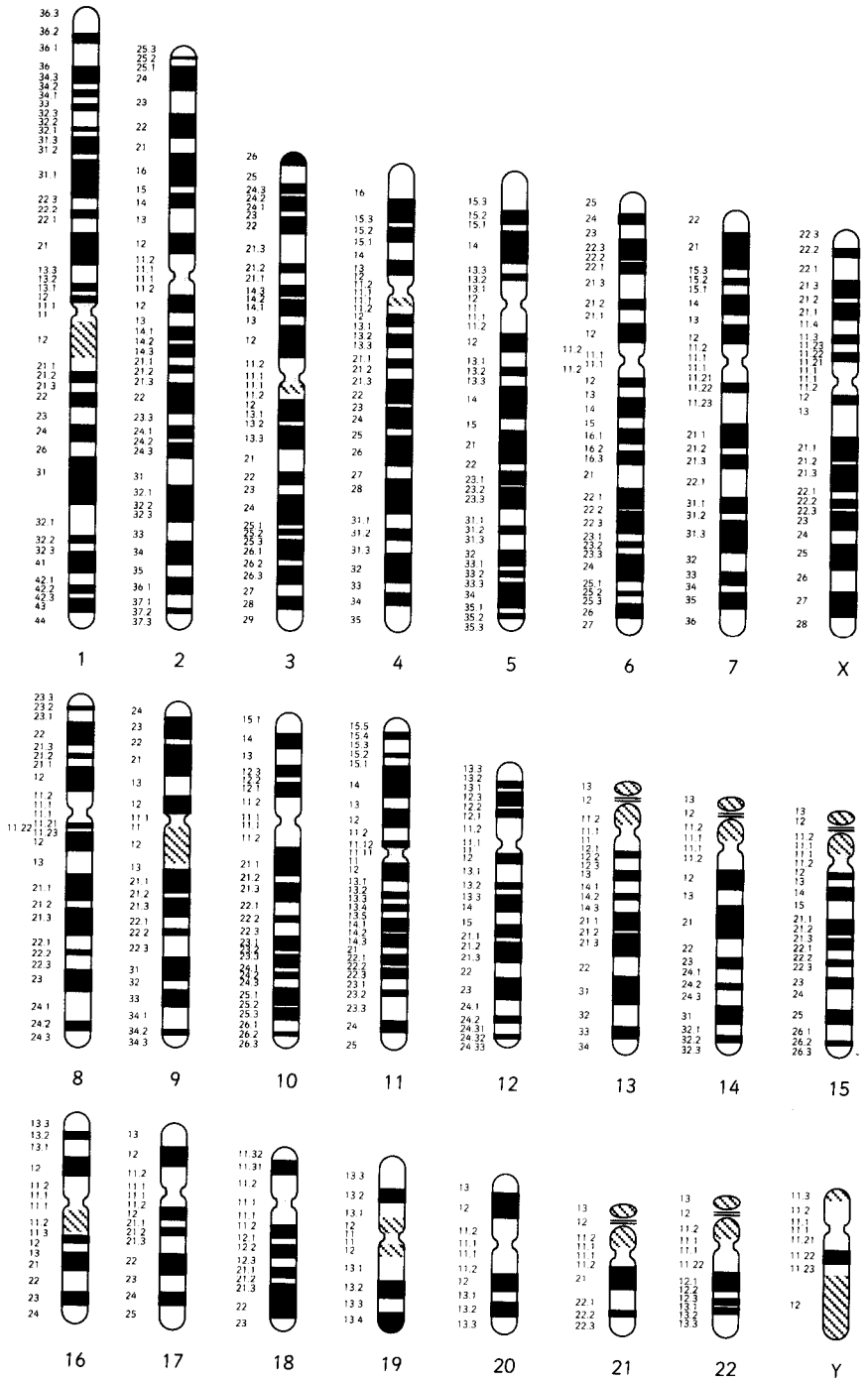


Figure 1.4: *Banding pattern of human chromosomes (G-banding, 550 band karyogram).*

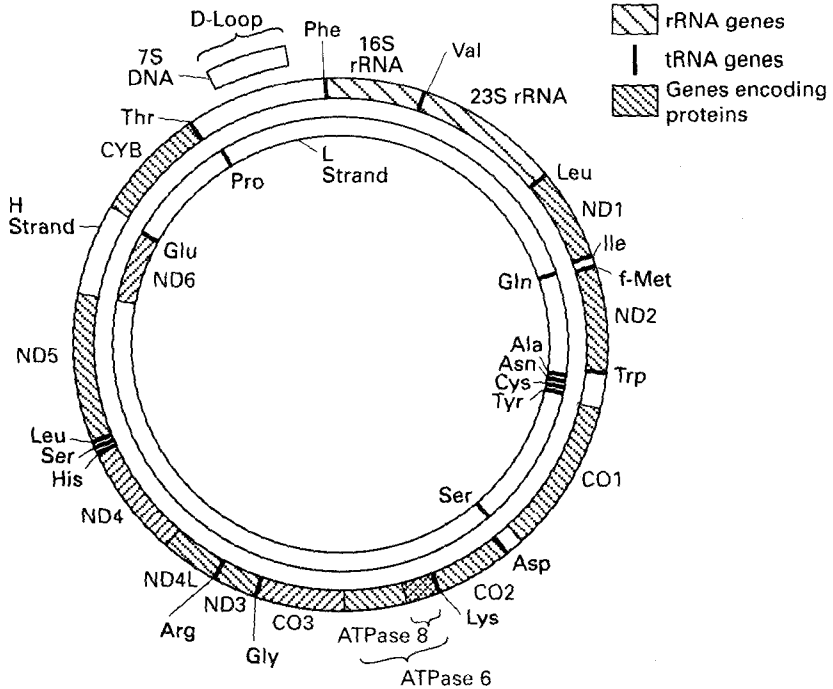


Figure 1.5: *Organization of human mitochondrial DNA. ND1–ND6: genes encoding NADH dehydrogenase sub-units 1–6. CO1–CO3: genes encoding cytochrome C oxidase sub-units 1–3. CYB–gene encoding cytochrome B.*

types of tRNA and two rRNA molecules which constitute part of the mitochondrial protein synthesis machinery; other components, for instance the aminoacyl tRNA synthetases, are encoded exclusively by nuclear genes.

Of the 22 types of tRNA encoded by mitochondria, eight can recognize families of four codons which differ only at the third base, and 14 recognize pairs of codons which are identical at the first two base positions and share either a purine or a pyrimidine at the third base. The remaining four codons, UAG, UAA, AGA, and AGG cannot be recognized by mitochondrial tRNA and act as stop codons (*see Table 1.2*).

Consequently, the genetic code employed to decipher mitochondrial-encoded mRNA on mitochondrial ribosomes differs from that used to decipher nuclear-encoded mRNA on cytoplasmic ribosomes. In addition, the mitochondrial and nuclear genomes differ in many other aspects of their organization and expression (*Table 1.3*).

1.3 Coding and non-coding DNA

Only a small fraction of the human genome (about 2–3%) is coding DNA, DNA sequences which directly specify a polypeptide or mature functional RNA product. The great majority of the genome is non-coding DNA (see *Figure 1.6*). At present, the bulk of extragenic DNA has no known function, either genetic or chromosomal, and accordingly has sometimes been considered to represent 'junk DNA'. Other extragenic non-coding DNA sequences have specific chromosomal functions.

Centromeres. Essential for ensuring proper disjunction of the chromosomes into daughter cells following cell division at meiosis and mitosis. Centromeric DNA predominantly comprises arrays of tandemly repeated DNA sequences, but the anticipated role of the latter in centromeric function has not been established (see Section 1.8.1).

Telomeres. These are required for ensuring complete replication of the DNA at the chromosome termini. In their absence, chromosomes become 'sticky' and will fuse with each other. In meiotic cells telomeres appear to be attached to the nuclear membrane and are the sites at which pairing of homologous chromosomes initiates. Telomeric DNA is composed of small arrays of tandemly repeated DNA (see Section 1.8.2).

Transcriptionally active DNA is generally marked by an altered chromatin structure which confers sensitivity to the enzyme DNase I, an endonuclease which nicks the individual strands of duplex DNA in a manner that is largely sequence-independent. Transcriptional activity is also often inversely correlated with the degree of methylation of cytosine residues. During DNA replication, transcriptionally active DNA is thought to be shielded from DNA methylases by transcription-associated protein factors. However, in other tissues where the same DNA is transcriptionally inactive, DNA methylases may gain access to the DNA and methylate it. The methylated DNA is subsequently bound by nuclear proteins such as MeCP-1, thereby limiting access to the DNA by transcription factors.

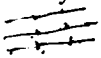


Table 1.2: *The genetic code in human cells*

AAA	Lys	ACA	Thr	AGA	$Arg^{N}/STOP^{M}$	AUA	Ile^{N}/Met^{M}
AAC	Asn	ACC	Thr	AGC	Ser	AUC	Ile
AAG	Lys	ACG	Thr	AGG	$Arg^{N}/STOP^{M}$	AUG	Met
AAU	Asn	ACU	Thr	AGU	Ser	AUU	Ile
CAA	Gln	CCA	Pro	CGA	Arg	CUA	Leu
CAC	His	CCC	Pro	CGC	Arg	CUC	Leu
CAG	Gln	CCG	Pro	CGG	Arg	CUG	Leu
CAU	His	CCU	Pro	CGU	Arg	CUU	Leu
GAA	Glu	GCA	Ala	GGA	Gly	GUA	Val
GAC	Asp	GCC	Ala	GGC	Gly	GUC	Val
GAG	Glu	GCG	Ala	GGG	Gly	GUG	Val
GAU	Asp	GCU	Ala	GGU	Gly	GUU	Val
UAA	STOP	UCA	Ser	UGA	$STOP^{N}/Trp^{M}$	UUA	Leu
UAC	Tyr	UCC	Ser	UGC	Cys	UUC	Phe
UAG	STOP	UCG	Ser	UGG	Trp	UUG	Leu
UAU	Tyr	UCU	Ser	UGU	Cys	UUU	Phe

^{N,M}Alternative interpretations of nuclear and mitochondrial codons.

Table 1.3: *The human nuclear and mitochondrial genomes*

	Nuclear genome	Mitochondrial genome
Size	3000 Mb	16.6 kb
Number of different DNA molecules	23 (in XX) or 24 (in XY) cells, all linear	1 circular DNA molecule
Total number of DNA molecules per cell	23 in haploid cells; 46 in diploid cells	Several thousand
Associated protein	Several classes of histone and non-histone protein	Largely free of protein
Number of genes	50 000–100 000	37
Gene density	1/30–1/60 kb	1/0.45 kb
Repetitive DNA	Large fraction – see *Figure 1.6*	Very little
Transcription	The great bulk of genes are transcribed individually	Continuous transcription of multiple genes
Introns	Found in most genes	Absent
Percentage of coding DNA	2–3%	Approximately 95%
Codon usage	See *Table 1.2*	See *Table 1.2*
Recombination	At least once for each set of homologous chromosomes at meiosis	None
Inheritance	Mendelian for sequences on X and autosomes; paternal for sequences on Y	Exclusively maternal

1.3.1 Clustering of coding DNA and of non-coding DNA in defined chromosomal regions

As the nuclear chromosomes uncoil following cell division certain chromosomal regions continue to remain condensed throughout the life cycle of the cell. They appear as dark-staining areas, termed heterochromatin, and have been presumed to be genetically inactive. In contrast, the bulk of the chromatin is euchromatin which

此为试读,需要完整PDF请访问: www.ertongbook.com