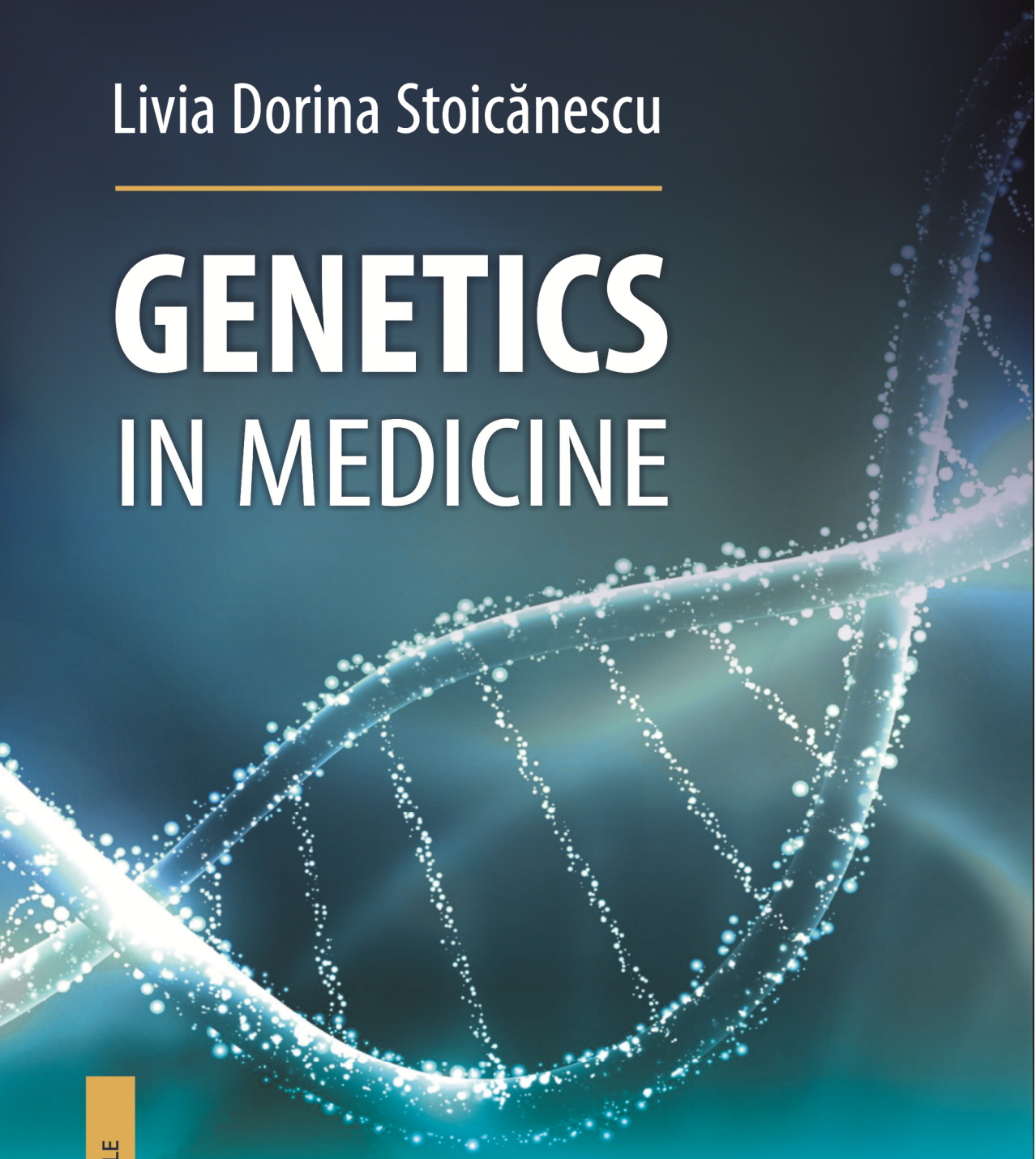
80 ani
UNIVERSITATEA
DE MEDICINĂ ȘI FARMACIE
„VICTOR BABEȘ" DIN TIMIȘOARA

Livia Dorina Stoicănescu

GENETICS IN MEDICINE



MANUALE

Editura „Victor Babeș"
Timișoara, 2025

Editura „Victor Babeş”
Piaţa Eftimie Murgu nr. 2, cam. 316, 300041 Timişoara
Tel./Fax 0256 495 210
e-mail: evb@umft.ro
www.umft.ro/ro/organizare-evb/

Director general: Prof. univ. dr. Sorin Ursoniu

Colecţia: MANUALE
Coordonatori colecţie: Prof. univ. dr. Codruţa Şoica
Prof. univ. dr. Daniel Lighezan

Referent ştiinţific: Prof. univ. dr. Alis Dema

ISBN 978-606-786-522-6

CONTENTS

Chapter I
ROLE OF GENETICS IN MEDICINE

Medical genetics is both a basic science and a clinical specialty. It is one of the most rapidly advancing fields of medicine and is now an integral part of all aspects of biomedical science.

Medical genetics is the science of human biologic variation as it relates to health and disease. Medical genetics can also be defined as the science of diagnosis, prevention and management of genetic disorders. The most important item of the clinical geneticist might be said to be the command of syndromology and dysmorphology. Genetic diseases include a large number of disorders. Either syndromology or dysmorphology can be determined by mutations, but also by nongenetic factors or by the interaction of both these factors.

The expansion and application of genetic knowledge have already had fruitful consequences for clinical medicine. Today, at least 1/3 (one third) of the sick children in pediatric hospitals has genetic disorders.

Human development depends on genetic and environmental factors. A person's genetic composition (genome) is established at conception. The genetic information is carried in the DNA of the chromosomes and mitochondria. Most diseases probably have some genetic component, the extent of which varies. DNA's capacity to replicate constitutes the basis of hereditary transmission. DNA also provides the genetic code, which determines cell development and metabolism by controlling RNA synthesis.

The sequence of the nucleotides that comprise DNA and RNA determines protein composition and thus its function.

Within medical genetics, there are many fields of interest such as, the study of chromosomes, their structure and the relations between chromosomes abnormalities and many disorders; on the other hand, the other major areas are the study of structure and function of individual genes (molecular and biochemical genetics), and the application to diagnosis; in addition to that, the population genetics, developmental genetics and immunogenetics have also medical relevance, especially in relation to the understanding and prevention of human disease.

In addition to clinical practice, medical genetics is applied to the care of patients in the fields of genetic counseling, population screening to identify persons at risk of developing or transmitting a genetic disorder, and prenatal diagnosis.

Genetic counseling, which combines the provision of risk information with a support function, is maturing into a new health profession. Population screening for genetic disease has become an important public health initiative. Prenatal diagnosis which makes use of many clinical and laboratory specialties in addition to genetics, is probably today the chief area in which genetics is applied to patient care.

Medical genetics is relevant to many clinical specialties. In adult medicine, it is recognized that many common disorders, such as coronary artery disease, hypertension, diabetes mellitus and many, many others diseases have important genetic components, and that preventive medicine could be more effective if it could be directed towards special genetically defined high risk groups.

The entire heredity of a person is named genotype; in other words, the genotype is the genetic constitution from the two genomes: one maternal, and the other one, paternal. The genome is the complete DNA sequence, containing the entire genetic information of a gamete. The haploid human genome consists of about 3 billion DNA base pairs that are distributed among 23 distinct chromosomes (22 autosomes and 1 sex chromosome).

The phenotype is the observable expression of a genotype as a morphological, biochemical or molecular trait. A phenotype may be abnormal (with variations) in all of the genetic disorders.

CLASSIFICATION OF GENETIC DISORDERS

In medical practice, the main significance of medical genetics is its role in the etiology of a large number of disorders. In fact, any disease is the result of the combined action of the genetic component and the environment, but in many disorders, the role of genetic component is larger, even completely.

Among disorders caused completely or partly by genetic factors, three main types are recognized:

1. Single gene disorders
2. Chromosome disorders
3. Multifactorial disorders

Single gene defects are caused by mutant genes. The mutation may be present on only one chromosome of a pair (with a normal allele -gene- on the homologous chromosome) or, on both chromosomes of the pair (both alleles of the pair). In either case, the cause is a single critical error in the genetic information.

Until today, more than 4000 disorders have been described and their impact in phenotype is significant. In several hundred of these diseases, the basic biochemical defects have been recognized and in many others, the affected gene has been already isolated and cloned. Most of affected genes are nuclear; a small number are mitochondria, such as genes for hereditary optic neuropathy, myoclonus epilepsy and encephalomyopathy. The trait of mitochondrially diseases is transmitted only by maternal inheritance. Mitochondrially transmitted diseases are recognized as a special type of single gene disorder.

In chromosome disorders, the defect is not due to a single mistake in the genetic material but to an excess or deficiency of the genes contained in whole chromosomes or chromosome segments.

Abnormalities of chromosomes may be either numerical or structural and may involve one or more autosomes, sex chromosomes, or both simultaneously. It is now known that chromosome disorders form a major category of genetic disease, accounting for a large proportion of all-reproductive wastage, congenital malformations, and mental retardation and as well are playing an important role in the pathogenesis of malignancy.

Multifactorial inheritance is responsible for several of developmental disorders resulting in congenital malformations and for many common disorders of adult life (with a late onset). It appears to be single error in the genetic information, but often a combination of a number of mutations that together can produce or predispose to a serious defect. Environmental factors may also be involved. Multifactorial disorders tend to recur in families (to have aggregation) but do not show the familial characteristic pedigree patterns of single gene traits.

Multifactorial diseases of adult life include: hypertension, diabetes mellitus, cancer, manic-depressive psychosis, schizophrenia, coronary artery disease and many others.

Short history of medical genetics

The main step of Medical Genetics was laid between 1865 when Mendel published his work about the nature of inheritance and 1956, when the correct human chromosome number was reported.

The terms dominant and recessive had been first introduced by Mendel (1865). His work and all his conclusions about the mechanisms of inheritance were not recognized at that time because the chromosomes, meiosis and its physical basis had not yet been discovered.

But, in 1900, mendelism was rediscovered by Hugo de Vries (1935), C. Correns (1933) and Tschermak (1935).

In 1902, Bateson made many contributions to genetics, including the introduction of the term " Genetics"and at the same time, he described the mechanisms of linkage. The true significance of linkage and the application of linkage and crossing-over to chromosome mapping were contributions of Thomas Hunt Morgan.

In the field of biochemical genetics, Garrod (1902), introduced the concept of inborn errors of metabolism. Alkaptonuria was the first of the disorders he investigated and supposed as an inheritance mendelian disorder. In 1958, some authors confirmed Garrod's predictions of an enzyme deficiency in alkaptonuria by demonstrating a deficiency of homogentisic acid oxidase in the patient's tissue. The gene for alkaptonuria has been mapped to the long arm of human chromosome 3 (Pollack, 1993).

In 1944, O. T. Avery, C. M. McLeod and collab., demonstrated that the hereditary information in bacteria is in DNA.

In 1956, J. H. Tijo and A. Levan developed new techniques for chromosome studies and improved their visibility. They determined the correct number of the human chromosomes.

In 1953, Watson and Crick suggested a model for the structure of DNA and established its complex structure and function. For this discovery, they were awarded the Nobel Prize in 1962.

During the past 40 years, techniques for the scientific study of human chromosomes have been developed and researchers began to explore the role of chromosome abnormalities in abnormal physical development and in mental retardation.

Since the mid-1970, with the aid of powerful new technologies for the manipulation and analysis of DNA, the field of molecular genetics has been transformed.

The discovery of specific restriction endonucleases made possible the isolation of small molecular fragments of DNA and short time after that, the cloning of DNA molecules has developed.

The first human genes, cloned through advanced methods, were made in 1977.

Around 1980, the method of somatic cell hybridization and the methods of molecular genetics contributed to the identification of the locus for many genes and provided the possibility for chromosome mapping.

The main methods for investigations in Medical Genetics are the following:

- Cytogenetics
- Molecular genetics

- Chromosome mapping
- Clinical genetics
- Clinical application of gene mapping
- Population genetics
- Genetic counseling

STUDY LEVELS OF HEREDITY

1. Molecular level: study of structure and function of genes (molecular and biochemical genetics) and the application to diagnosis
2. Cytogenetic level: study of chromosomes and the relationship between chromosome abnormalities and many disorders
3. Individual clinical level
4. Population level

Chapter II
DNA STRUCTURE, ORGANIZATION AND FUNCTIONALITY

Topology of DNA structure

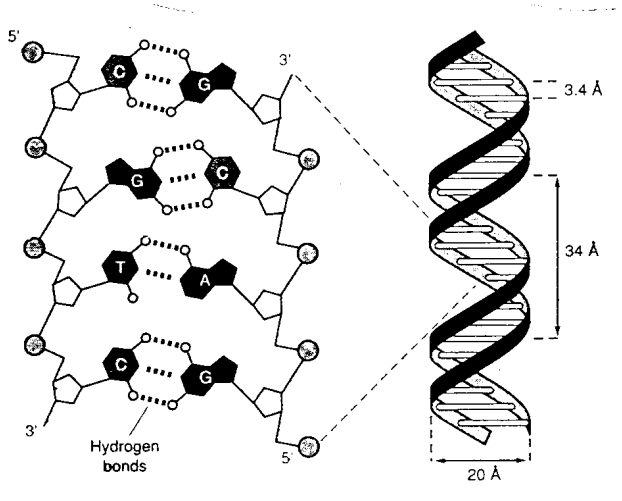
Genetic information is preserved by DNA structure. DNA is a macromolecule polymer of a linear array of deoxyribonucleotides, each of which consists of three components: a nitrogenous base, a sugar (deoxyribose) and phosphate. Sugar and phosphate groups link each base to adjacent bases on the same strand. The bases in DNA are either purines (adenine-A and guanine-G) or pyrimidines (cytosine-C and thymine-T), and together these nucleotides constitute the "four-letter alphabet" of DNA (of heredity) that is universal among organisms.

The DNA double helix is a two-stranded, spiraling structure that resembles a twisted ladder. Each strand is made of a sugar-phosphate backbone, with bases (adenine, thymine, cytosine, and guanine) attached to the sugar, the sequence of nucleotides. The two strands are held together by hydrogen bonds between complementary bases

In the Watson-Crick helical structure of double-stranded DNA first reported in 1953, pairing occurs between a purine base on one strand and a pyrimidine base on the opposite strand (G pairs with C, A pairs with T), thereby making each strand complementary to the other. It is the order of these bases that encodes the genetic information contained within DNA.

A typical DNA molecule consists of two polynucleotide chains, each containing several thousands to several million monomers (nucleotides). Each nucleotide in one chain is specifically linked by hydrogen bonds to a nucleotide in the other chain. It is said that the two chains are therefore complementary to each other.

The anatomical structure of DNA carries the chemical information that allows the exact transmission of genetic information from one cell to its daughter cells and from one generation to the next. At the same time, the primary structure of DNA specifies the amino acid sequences of the polypeptide chains of proteins.



The structure of DNA. Two-dimensional representation of the two complementary strands of DNA, showing the AT and GC base pairs. Note that the orientation of the two strands is antiparallel (left). The double-helix model of DNA, as proposed by Watson and Crick. The horizontal "rungs" represent the paired bases. The helix is said to be right handed because the strand going from lower left to upper right crosses over the opposite strand (right).

DNA polymorphism

There are three natural forms of DNA (A, B and Z). The origin of these different forms are related to the **conformation of the sugar** (*C2'-endo*/ *C3'-endo*) and the **orientation of the base relative to the sugar.**

Thus depending on base composition and physical conditions (Hydration/Salt-Content), DNA can assume several different **conformations (A, B, Z)**. Each conformation possesses specific parameters: diameter of the helix, number of bases per tour and distance between plan of bases.

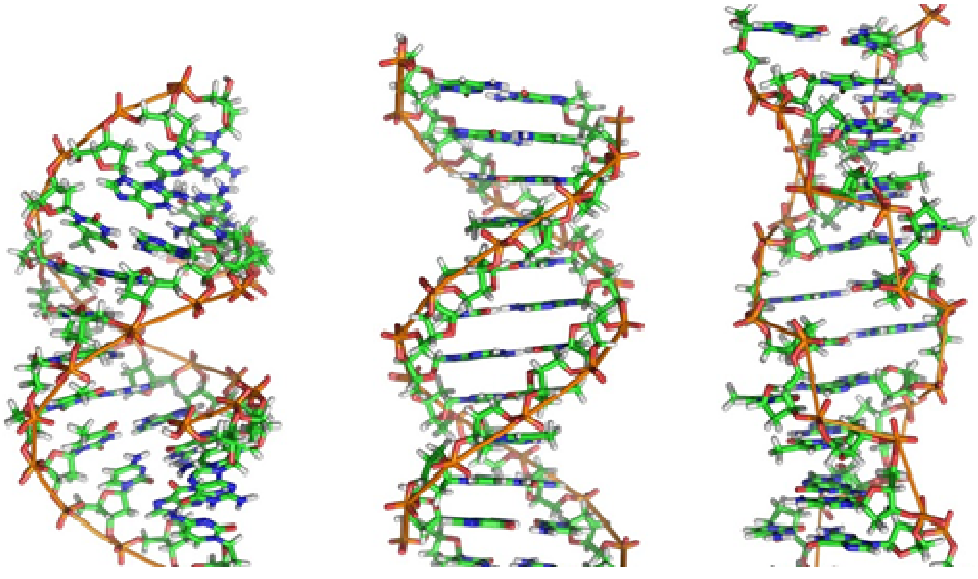
The **B-form** is the common natural form, prevailing under physiological conditions of low ionic strength and high degree of hydration. B-DNA arranges 10 nucleotides per helix tour. The plan of the bases is nearly perpendicular to the helix axis and the helix surface exhibits two prominent grooves (major and minor).

Besides the B form - described above, studies using X-ray diffraction methods, led to the discovery of another DNA form also, called *DNA-Z form* (with a zigzag course). The **Z-form** (Zigzag chain) is observed in DNA G-C rich regions. Z-DNA is longer, thinner and possesses an unusual left-handed helix (of 12 bases pairs/tour) with a single narrow deep groove. These Zigzag form mainly results from the alternation of purines and pyrimidines. In the Z form of DNA, a single deep minor groove replaces the major and minor grooves found in B DNA (represented by double helix). The Z DNA molecules are localized in certain limited regions of the chromosomes. Today, improved techniques indicate that DNA Z form provides recognition signals for certain important regulatory molecules.

The **A-form** is sometimes found in some parts of natural DNA in presence of high concentration of cations or at a lower degree of hydration (<65%). A-DNA possesses 11 nucleotides per tour and two grooves (a narrow deep major and a wide shallow minor).

The C-form and D-form are unusual forms with different base pairs. C-DNA is sometimes observed under 65% of hydration, while D-DNA is only found in artificial DNA.

The changes in the shape of DNA can affect its binding with proteins and may be involved in some regulation process during replication or transcription.



DNA forms: A-DNA, B-DNA and Z-DNA

(https://commons.wikimedia.org/wiki/File:A-DNA,_B-DNA_and_Z-DNA.png)

Comparison of A, B, and Z-DNA Feature	A Form	B-Form	Z-Form
Helix type or screw sense	Right-handed	Right-handed	Left-handed
Helix Diameter	~2.6 nm	~2.0 nm	~1.8 nm
Base pairs per turn of helix	11	10.4	12
Helix pitch or rise per turn of the helix	2.53 nm	~3.54 nm	4.56 nm
Helical twist or rotation per successive bp	33 degrees	36 degrees	60 degrees
Tilt of basepair from helix axis	19 degrees	6 degrees	9 degrees
Major groove	narrow and very deep	Wide and quite deep	Flat
Minor groove	Very broad and shallow	Narrow and quite deep	Very narrow and deep
Overall shapes of double helix	Broadest	Intermediate	Narrowest

Multistranded DNA structures have been described, studies concentrating on these structures due to their specific significant roles in vivo and potential use of therapeutics.

There is a DNA form, of three intertwined strands, called triplex DNA. It is known that a protooncogene (c-myc) has areas that form triplex DNA involved in the control of the quantity of protein made from this gene. A proto-oncogene is a normal gene that regulates cell growth and division, but can become an oncogene (a cancer-causing gene) if it is mutated.

Quadruplex DNA is found in the telomeres (the ends of the chromosomes), but also outside the telomeres, being associated with cell cycle progression. The same c-myc protooncogene may form quadruplex DNA.

Mitochondrial DNA

The *mitochondrial DNA* (mt DNA): although the vast majority of genes are located in the nucleus, a small but important subset resides in the cytoplasm, in the mitochondria. Mitochondrial genes exhibit exclusively maternal inheritance. All human cells have hundreds of mitochondria, each containing a number of copies of a small 9

Mitochondrial DNA is a circular molecule. The mitochondrial DNA molecules are only 0,03% of the length of the smallest nuclear chromosome. It contains 37 genes that are crucial for the cell's energy production processes, including genes for enzymes involved in oxidative phosphorylation, as well as ribosomal and transfer RNA. The mtDNA has been completely sequenced and is known to code for two types of ribosomal RNA, about 22 transfer RNAs and about 13 polypeptides involved in oxidative phosphorylation.

Applications:

- Ancestry tracing
- Forensics
- Studying human history
- Medical research

Mutations in mitochondrial genes have been demonstrated in several neuromuscular disorders, including the maternally transmitted hereditary optic neuropathy (Leber's syndrome), MELAS (Mitochondrial Encephalomyopathy, Lactic Acidosis, and Stroke-like episodes), Kearns-Sayre syndrome (KSS), and Leigh syndrome.

The mtDNA replicates within the mitochondria, unlike the nuclear DNA.

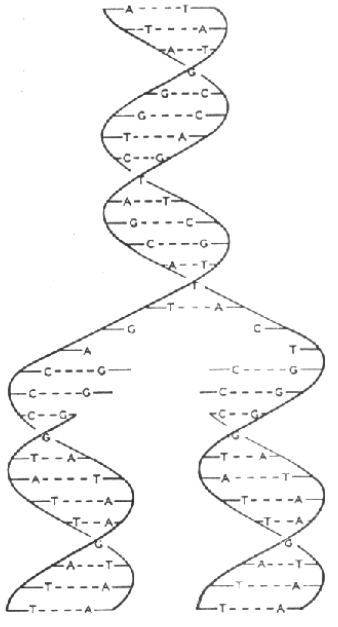
A unique feature of mtDNA is its maternal inheritance. The ovum is well supplied with mitochondria, but sperm contain few, and even those few do not persist in the offspring. The mother transmits her mtDNA to all her offspring.

Replication of DNA molecule

For cells to pass on their genetic material to their progeny, they must replicate their entire genome before cell division occurs. Replicating an entire genome is no easy matter since this process must be carried out efficiently in a defined period and with extreme fidelity.

Human cells carry out the process of DNA replication at a particular time in their life cycle; it takes place in the “S” phase towards the end of the interphase.

The anatomical structure of DNA is a linear double stranded structure with the two strands running in opposite polarity and held together by hydrogen bonding between A-T and G-C. During the process of DNA replication, the two strands separate, with each strand serving as a template for the synthesis of the two new daughter strands. This mode of replication is referred to as semiconservative replication. After the first round of DNA synthesis and cell division, each of the two daughter cells contains one old strand of DNA and one newly synthesized strand of DNA. Because of the specificity of the hydrogen bonding, where A pairs with T and G with C, the sequence of bases in the template strand dictates the sequence of bases in the newly synthesized strands. In this way, the nucleotide sequence of the DNA molecule is maintained in both daughter cells and is identical to the original parental molecule.



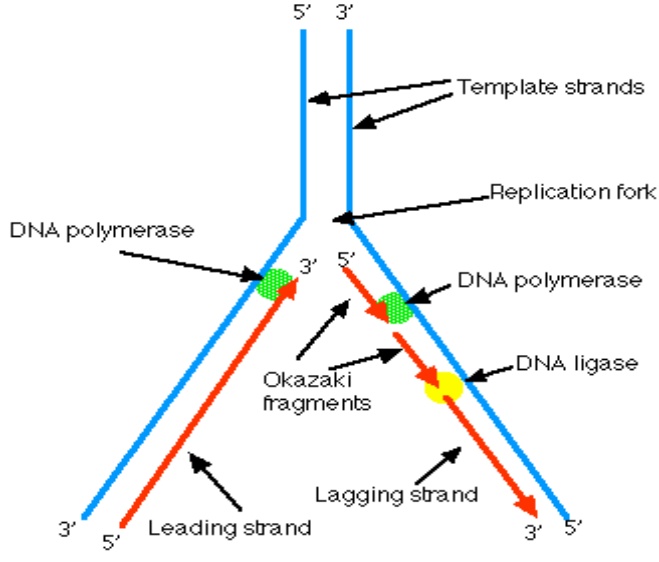
Replication of DNA double helix, resulting in two identical daughter molecules, each composed of one parental strand and one newly synthesized strand.

During replication several enzymes are active, which coordinate the process and repair the mistakes.

The replication process on each chromosome starts at specific positions, referred to as the origin of replication (replicons). These sites of initiation of DNA replication must have certain properties that signal the replicating system to start replication. As the two DNA strands separate and the bases are exposed, the enzyme *DNA polymerase* moves into position at the point where synthesis will begin.

The DNA polymerase then adds nucleotides in an exactly complementary manner, A to T and G to C. DNA polymerase catalyzes the formation of the hydrogen bonds between each arriving nucleotide and the nucleotides on the template strand. The template strand is always read in the 3' to 5' direction. The new DNA strand has to be synthesized in the 5' to 3'

direction. The DNA strands run antiparallel, thus one new strand will grow continuously as the point of replication will move along the template DNA. The other new strand will grow *discontiguously,* with short blocks of new DNA called *Okazaki fragments*. These sections are then joined by the action of the enzyme *DNA ligase,* which ligates the blocks together. As seen, once initiated, the replication complex proceeds bidirectionally, but is not synchronous throughout the chromosome.



Replication of DNA: leading and lagging strands

In DNA replication some proteins and several enzymes are involved, such as: ligase, endonuclease, exonuclease, DNA polymerase, helicase, etc.; their activity is necessary for replication and also to recognize the incorrect nucleotides, to remove them and to replace the proper complementary base.

Errors introduced during the normal process of DNA replication generate mutations. These errors can be recognized and repaired if DNA repair mechanisms are in good function.

DNA Damage Repair

Errors introduced during the normal process of DNA replication generate mutations. These errors can be recognized and repaired if DNA repair mechanisms are in good function.

Exonuclease activity is associated with DNA polymerases that are involved in DNA replication.

Recently, a new mismatch repair system has been described in all the cells from bacteria to human cells, known as methyl-directed mismatch repair. During the initial period of DNA replication, one strand (i.e., the template strand) is methylated, but the newly, synthesized DNA strand is not methylated. When a mispair is detected, correction specifically takes place on the nonmethylated (newly) DNA strand. This enables the repair system to correct the strand of DNA and prevents the mispair from leading to a mutation.

As expected from the important role that DNA replication and repair enzymes play in mutation surveillance and prevention, inherited defects that alter the function of enzymes for the correction of mispairs can lead to some autosomal recessive disorders such as: Xeroderma pigmentosum, ataxia telangiectasia, Fanconi anemia, Bloom syndrome, ADA deficiency and others.

In addition to replication errors, it is estimated that between 10,000 and 1 million nucleotides are damaged per human cell per day by spontaneous chemical processes such as depurination, demethylation, or deamination; by reaction with chemical mutagens (natural or otherwise) in the environment; and by exposure to ultraviolet or ionizing radiation. Some, but not all, of this damage is repaired. Even if the damage is recognized and removed, the repair machinery can create variants by introducing incorrect

bases. Thus, unlike replication-related DNA changes, which are usually corrected by proofreading mechanisms, nucleotide changes introduced by DNA damage and repair often lead to permanent variants.

The future of a variant is variable. Variants are most often repaired.

Restoration of a Damaged Site

Repair of Single-Strand Breaks: Breaks caused by mutagenic agents (X-rays or chemicals) can be quickly repaired by specific enzymes: ligase.

Removal and replacement of the damaged area

Mismatch repair: This process occurs after DNA replication in the manner of a "spelling checker". It is carried out by a group of proteins (nucleases) that can "scan" the DNA and detect incorrect or mismatched base pairs. The incorrect nucleotide is removed, then DNA polymerase operates again and restores the correct sequence.

Excision and then replacement of nucleotide groups: This mechanism intervenes in major DNA damage that creates blockages from replication to transcription, such as UV-induced dimers. The variant is recognized and the DNA strand with the variant is cut on both sides. The involved nucleotides are excised. DNA polymerase then completes the missing areas and ligase finally re-establishes the bonds.

Tolerance of a damaged area

Repair by recombination of daughter molecules: Not all DNA damage can be repaired immediately so some persist. If a replication eye contains alterations (e.g. a thymine dimer, replication blockages are normal. However, in eukaryotes, replication can be initiated at numerous points.

They can rejoin behind the dimer, leaving a sector of the molecule non-replicated. But this does not prevent cell division from being blocked. Recombination with a sister chromatid strand can lead to two "complete" sister molecules after insertion of the missing nucleotides. One of them still contains the dimer but the other is complete and not variant and division is also not blocked.

Chapter III

THE MOLECULAR PATHWAY OF GENE EXPRESSION

RNAs. Transcription. Translation. Genetic code. Central dogma of molecular biology.

Genetic code

Genetic information is contained in DNA molecules in the chromosomes within the cell nucleus. Each set of three bases constitutes a *codon (triplet)*, specific for a particular amino acid. Almost infinite variations are theoretically possible in the arrangement of the bases along a polynucleotide chain. At any position, there are four possibilities: A, T, C, G; thus there are 4n possible combinations in a sequence of n bases. For three bases, there are 43 =64 possible triplet (codons) combinations. These 64 codons constitute the genetic code.

The genetic code was deciphered through experiments in which synthetic polynucleotides were used. The first synthetic mRNA used was polyuracil (polyU), a sequence of nucleotides in which all the bases are uracil. Upon translation, the polyU mRNA directed the synthesis of a polypeptide chain composed exclusively of phenylalanines, which thus established that the triplet code for phenylalanine is UUU.

The other codons were then decoded in a similar manner. Because there are only 20 amino acids and 64 possible codons, most amino acids are

specified by more than one codon; hence the code is said to be degenerate. For instance, the base in the third position of the triplet can often be either purine (A or G) or pyrimidine (T or C) or, in some cases any of the four bases, without altering the code message. Leucine and arginine are each specified by six codons. Only methionine and tryptophan are each specified by a single, unique codon. Because the genetic code is degenerate, it is possible that a small change in the codon structure does not alter the amino acid sequence (silent mutation).

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

Codon table

Three of these codons are called stop (codons nonsense) codons because they designate the termination of translation of the mRNA at that point.

Other proprieties of genetic code: The code is *non-overlapping*, meaning that successive triplets are read in order and each nucleotide is part of only one codon. There are *no spaces or commas separating neighboring codons*, each nucleotide belongs to a codon. The genetic code is *unambiguous,* meaning that each codon specifies a particular amino acid and only that one. The code is nearly *universal*, meaning that almost all organisms use exactly the same genetic code.

Translation of a processed mRNA is always initiated at a codon specifying methionine. Methionine is therefore the first amino acid of each polypeptide chain, although it is usually removed before protein synthesizes is completed. The codon for methionine (the initiator codon, AUG from mRNA) establishes the reading frame of the mRNA. Genetic code refers to the codons in DNA molecules, also to the codons in mRNA.

Ribonucleic acids

Ribonucleic acid (RNA) is a single-stranded molecule involved in various cellular processes. The chemical structure of RNA is similar to that of DNA, except that each nucleotide in RNA has a ribose sugar component instead of a deoxyribose; in addition, uracil (U) instead of thymine is one of the pyrimidines of RNA. An additional difference between RNA and DNA is that RNA in most organisms exists as a single-stranded molecule, whereas DNA exists as a double-stranded helix (except transiently during DNA replication). It plays a critical role in gene expression, regulation, and protein synthesis. RNAs are central to the 'Central Dogma' of molecular biology: DNA → RNA → Protein.

Main Types of RNA are:

1. Messenger RNA (mRNA) – carries genetic information from DNA to ribosomes.

2. Ribosomal RNA (rRNA) – forms the structural and functional core of ribosomes.

3. Transfer RNA (tRNA) – delivers specific amino acids during protein synthesis.

These three are the major types involved in translation and gene expression.

Messenger RNA (mRNA) is synthesized during transcription from a DNA template. It contains codons that specify the amino acid sequence of a protein. Includes 5′ cap and 3′ poly(A) tail for stability and translation regulation.

Ribosomal RNA (rRNA) is a major component of ribosomes (along with proteins). rRNA catalyzes peptide bond formation (acts as a ribozyme). rRNAs are encoded by rDNA repeats and transcribed in the nucleolus. They are highly conserved across species, therefore are useful in evolutionary and phylogenetic studies.

Transfer RNA (tRNA) is an adaptor molecule linking mRNA codons with amino acids. It has a cloverleaf structure with an anticodon loop and amino acid attachment site. Each tRNA recognizes a specific codon via base pairing. Aminoacyl-tRNA synthetases ensure correct pairing between amino acids and tRNAs.

Non-Coding RNAs (ncRNAs)

Do not code for proteins but have crucial regulatory and structural roles.

Types include:

- microRNA (miRNA) – post-transcriptional regulation via mRNA silencing.
- small interfering RNA (siRNA) – mediates RNA interference (RNAi) for gene silencing.
- long non-coding RNA (lncRNA) – involved in chromatin remodeling, transcriptional regulation.
- small nuclear RNA (snRNA) – part of spliceosome for pre-mRNA splicing.
- small nucleolar RNA (snoRNA), with several functions within the nucleolus, such as: participating in making ribosomes, roles in the alternative splicing of pre-mRNA to different forms of mature mRNA. One small nucleolar RNA serves as the template for the synthesis of telomeres.
- XIST RNA is involved in inactivating one X chromosomes in females.

Clinical Relevance of RNA:

- RNA plays central roles in disease pathogenesis and therapy.
- Viral RNAs: many viruses (e.g., influenza, SARS-CoV-2) use RNA genomes.
- RNA-based therapeutics: mRNA vaccines, siRNA drugs (e.g., patisiran).
- Biomarkers: circulating miRNAs used in cancer diagnostics and prognosis.

Central dogma of molecular biology

The flow of information from gene to polypeptide is often called as ”the central dogma” of molecular biology and it states that this information moves in a single direction: DNA is transcribed into RNA, and then RNA is translated into a protein. This process is fundamental to gene expression. It is synthesized in the following formula:

transcription translation

DNA⇾ →mRNA ⇾ Polypeptide

replication

Ribonucleic acid molecules: mRNA (messenger RNA), tRNA (transfer RNA) and r-RNA (ribosomal RNA) are involved in gene expression.

The informational relationships among DNA, RNA and protein are circular: DNA directs the synthesis and sequence of RNA; RNA directs the synthesis and sequence of polypeptides (some specific proteins are involved in the synthesis and metabolism of DNA and RNA).

The most notable exception to the central dogma is **reverse transcription**, in which RNA encodes DNA. This process is used by retroviruses and transposons to parasitise the genomes of eukaryotes.

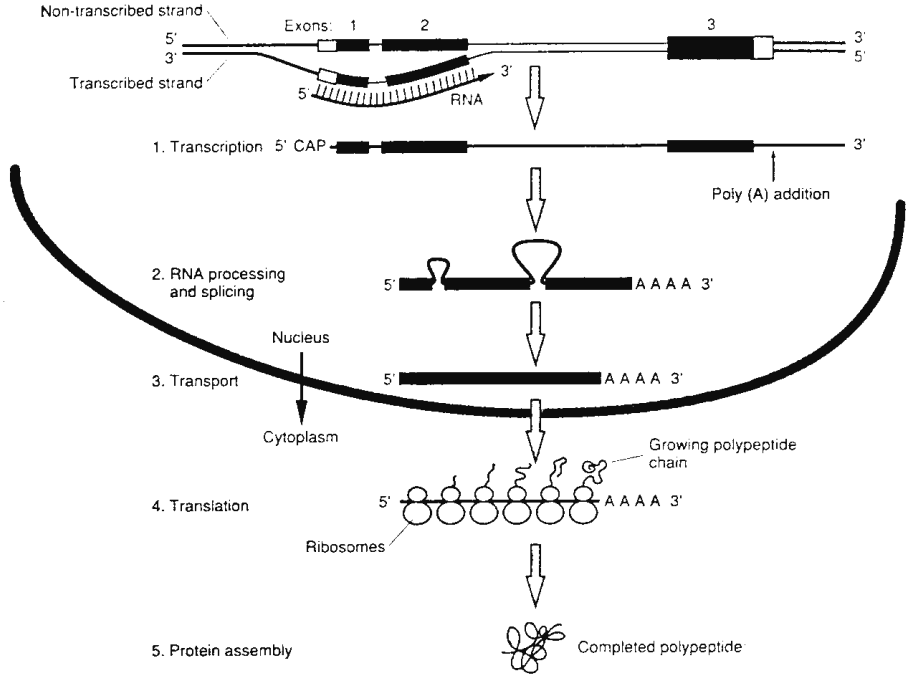
revers transcription *transcription* *translation*

RNA ⟶ DNA ⟶ mRNA ⟶ Polypeptide

Transcription

Translation

Genetic information is stored in DNA by means of codons (genetic code) in which the sequence of adjacent bases ultimately determines the sequence of amino acids in the encoded polypeptide. First mRNA is synthesized from DNA template through a process known as **transcription**.



Flow of information from DNA to RNA to protein for a hypothetical gene with three exons and two introns. Steps include transcription, RNA processing, RNA transport from the nucleus to the cytoplasm, and translation.

The RNA carrying the coded information in a form called messenger RNA (RNA) is then transported from the nucleus to the cytoplasm, where the RNA sequence is decoded, or translated to determine

the sequence of amino acids in the protein being synthesized. This process of **translation** occurs on ribosome's (they contain r-RNA) and involves a third type of RNA, transfer RNA (t-RNA), which provides the molecular link between the coded base sequence of the RNA and the amino acid sequence of the polypeptide.

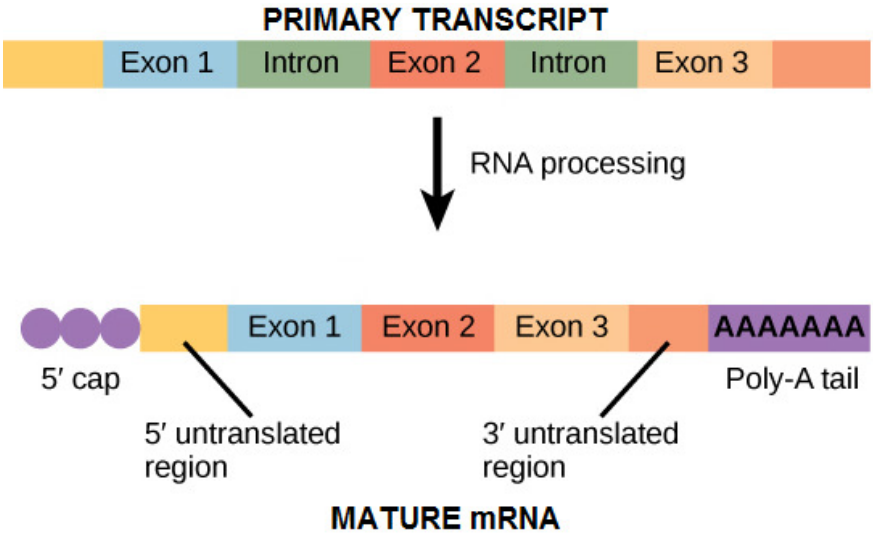
The nascent mRNA molecules must have its introns spliced out and its end modified before export into the cytoplasm as mature (final) mRNA. The transcription initiation site is a region of DNA that binds RNA polymerase (enzyme necessarily for mRNA synthesis). This region that initiates transcription is called promoter.

The transcription initiation site is a region of DNA that binds RNA polymerase (enzyme necessarily for mRNA synthesis). This region that initiates transcription is called **promoter**. Different protein transcription factors will bind to the promoter, usually on the 5′ end, of the gene to be transcribed. RNA polymerase will also bind to the complex of transcription factors and together will open the DNA double helix.

The RNA polymerase travels along the DNA strand in the $3' \rightarrow 5'$ direction and it assembles the ribonucleotides into a strand of RNA. Each ribonucleotide is inserted into the mRNA strand following the rules of complementarity. When transcription is complete, the primary transcript is released from the polymerase, which shortly thereafter, is released from the DNA strand.

RNA Processing

The nascent mRNA molecules must have its introns spliced out and its ends modified before export into the cytoplasm as mature (final) mRNA.



RNA processing

Steps in RNA processing:

Synthesis of the cap: a modified guanine will be attached to the 5' end of the pre-mRNA, protecting it from being degraded by enzymes that degrade RNA from the 5' end.

Splicing: Step-by-step removal of introns that are present in the primary transcript and splicing of the remaining exons.

The cutting of primary transcript in order to remove the introns must be done with great precision. If this process is not accurate and only one nucleotide is left over from an intron or one is removed from an exon, the reading frame of the gene will be changed, thus a mutation will occur.

Synthesis of the poly(A) tail: a stretch of nucleotides with adenine. This poly(A) tail is attached to the exposed 3' end. This completes the mRNA molecule, which is now ready for being exported to the ribosome.

The nascent precursor mRNA molecule is sometimes described as heterogeneous nuclear RNA which after splicing (in the nucleus) will form the mature mRNA. A critical role in the splicing of heterogeneous mRNA is assumed by small nuclear ribonucleoproteins.

After intron sequences are spliced out of the primary RNA transcript (heterogeneous RNA, the mature mRNA is transported from the nucleus to the cytoplasm, where it is translated into a polypeptide chain. In the cytoplasm, t-RNA molecules provide a bridge between mRNA and free amino acids. Adjacent groups of three nucleotides in mRNA (codons) each bind to complementary three nucleotide sequences in t-RNA (*anticodons).* Unlike most other nucleic acids, t-RNA molecules have a rigid tertiary structure. Each t-RNA (there are more than 40 types) has the *anticodon* located at one end and the amino acid - binding at the other end. A special role in binding of amino acid at the t-RNA has several aminoacyl-synthetases. These enzymes recognize different t-RNA and attach each t-RNA to the correct amino acid.

In the cytoplasm, t-RNA molecules provide a bridge between mRNA and free amino acids. Adjacent groups of three nucleotides in mRNA (codons) each bind to complementary three nucleotide sequences in t-RNA (*anticodons).* Unlike most other nucleic acids, t-RNA molecules have a rigid tertiary structure. Each t-RNA (there are more than 40 types) has the *anticodon* located at one end and the amino acid - binding at the other end. A special role in binding of amino acid at the t-RNA has several aminoacyl-synthetases. These enzymes recognize different t-RNA and attach each t-RNA to the correct amino acid.

Protein synthesis

The relationship between codon and amino acid sequence is referred to as the genetic code. The different tertiary structures of each t-RNA are specifically recognized by the proper t-RNA synthetase, ensuring the accuracy of the code. The translation process takes place between the codons of m-RNA and the anticodons of t-RNA. The last base in each codon is followed by the first base in the next and thus, the first codon in a mRNA molecule determines the reading frame for all subsequent codons. Translation ends when a stop codon is encountered in the same reading frame. The completed polypeptide is then released from the ribosomes, which becomes available to begin synthesis for another polypeptide.

The biochemistry of protein synthesis can be divided into stages of initiation, elongation and termination. All three processes occur on ribosomes, which contain proteins and much r-RNA. Active ribosomes are grouped together as up to 50 ribosomes per one m-RNA molecule, forming *polysomes.*

- *Initiation:* The small subunit of the ribosome binds to a site "upstream“ the start of the message. and proceeds downstream, until it encounters the start codon AUG. The large subunit of the ribosome will now be joined. The initiator tRNA will bind to the P site of the ribosome. In eukaryotes, initiator tRNA carries methionine (Met).
- *Elongation:* A tRNA that carries its amino acid able to base pair (as it has a complementary anti codon) with the next codon on the mRNA arrives at the A site. The ribosome will move one codon downstream, shifting the more recently arrived tRNA to the P site

and opening the A site for the arrival of a new tRNA, which will carry another aminoacid.

- *Termination:* Translation ends when the ribosome reaches one of the STOP codons (UAG, UAA, UGA).

After synthesis, many proteins undergo extensive posttranslational modifications. The polypeptide chain that is the primary translation product is folded and bound into a specific three-dimensional structure that is determined by the amino acid sequence itself. Two or more polypeptide chains, products of different genes, may combine to form a single protein. For example, two alfa-globin chains and two beta-globin chains combine to form a tetrameric $\alpha_2 \beta_2$ hemoglobin molecule.

The protein products may also be modified chemically by the addition of carbohydrates, by cleavage, by removing of specific sequences, becoming in this way active.

Exceptions of the relationship "one gene - one polypeptide"

Development of the techniques for the gene molecular structure analysis revealed some exceptions from "one gene-one polypeptide" concept.

One of the exceptions is: "a gene - more polypeptides" and is determined by different mechanisms at different levels of gene expression (transcription, post-transcription, translation, post-translational processing of the protein).

The mechanisms that lead to this exception:

- **alternative splicing**

As mentioned, the majority of eukaryotic genes are composed of exons and introns. For some genes, the usual splicing of pre-mRNAtakes place: the introns are excised and the exons are put together one after another, after translation resulting one single protein. For other genes, however, the splicing varies, depending on the tissue, developmental stage, or the physiological status of the cell.

One way is alternative splicing: the mechanism that generates several forms of mature mRNA from the same. Thus, several different proteins can be produced from the multiple mRNAs that are produced from the same gene. This is a regulatory mechanism by which variations in incorporating exons into mRNA lead to production of several related proteins, isoforms, or even completely different proteins.

- isoforms: the 4 forms of myelin basic protein
- different proteins: Calcitonin gene is processed differently in the thyroid and CNS. The pre-mRNA from this gene contains 6 exons; the calcitonin mRNA contains exons 1–4. In the CNS another mRNA is produced from this pre-mRNA by skipping exon 4, but including exons 1–3, 5, and 6, encoding a protein called CGRP (calcitonin gene related peptide).
- **alternative promoters:** information held by the same gene leads to formation of tissue-specific isoforms, with different properties, expressed in different ontogenetic periods and with different functions. The use of alternative promoters is common and is used to generate cell type specific mRNAs. These alternative promoters may be found even within introns of the gene. For example, the dystrophin gene that has more than 79 exons has at least eight different alternative promoters.

- **RNA Editing:** is a mechanism by which a programmed change of the primary transcript takes place, so mature mRNA will have a different sequence from the DNA strand that has been copied. Editing can take place for all types of RNA.
- **post-translational cleavage**: the mechanism occurs after synthesis of the polypeptide, which is cut into several smaller polypeptides. This phenomenon is encountered in the synthesis of certain hormones from a prohormone.

A further exception "**more genes - a polypeptide**" is evident in the case of polypeptides formed of different regions. For example, the light chains (L) and heavy (H) chains of immunoglobulins have a variable region (V) - which is the antigen recognition and attachment site and a constant region (C). These regions (V, C) are encoded by different genes located on the same chromosome. They rearrange (by somatic recombination) during B cell differentiation, forming a functionally complete sequence encoding the immunoglobulin chain.

Pleiotropy

Pleiotropy is characterized by multiple, diverse, phenotypic effects in multiple systems, organs, determined by a single gene (dominant) or a pair of genes (recessive).

Examples of diseases with pleiotropy: Marfan syndrome, neurofibromatosis, osteogenesis imperfecta (autosomal dominant); Bardet-Biedl syndrome, Fanconi anemia, Bloom syndrome, ataxia telangiectasia, cystic fibrosis (autosomal recessive).

Marfan syndrome is characterized by several types of clinical manifestations:

- Skeletal abnormalities (long limbs, long fingers (arachnodactyly), joint hyperlaxity (resulting in frequent luxations), deformations of the spine (kyphosis, scoliosis) or of sternum, etc.
- Ocular abnormalities: lens subluxation, myopia, etc.
- Cardiovascular Changes: aneurysms of the aorta, mitral valve prolapse

Marfan syndrome phenotypic changes are the result of mutations in one gene, mutations which cause defects in the primary structure of fibrillin, a major component of the connective tissue. Gene FBN1 is on chromosome 15q15. Multiple manifestations of the syndrome are the result of connective tissue alteration in different systems, thus there is a pathogenic link and a common mechanism. In such cases pleiotropy is said to be relational.

Another example is **Bardet-Biedl syndrome (BBS)**, an autosomal recessive ciliopathy. Clinical features are: obesity, polydactyly, hypogonadism, hearing loss, even deafness, retinopathy, mental retardation, etc. Diagnosis is based on clinical findings and can be confirmed by sequencing the genes known to be involved in the etiology of the disease and which are recorded in 80% of patients. BBS genes encode proteins that are involved in cilia biogenesis and function. Mutations in these genes cause ciliary defects responsible for the pleiotropic effects observed in BBS.

BBS has the estimated incidence of 1: 160 000 in populations from northern Europe and 1:13500 in some Arab populations. Molecular genetic testing is now available and today 18 genes that are associated with at least 18 clinical forms of Bardet-Biedl syndrome are known.

Chapter IV
GENE STRUCTURE AND ORGANIZATION

Genes – fundamental elements of heredity

Johannsen first applied the term gene to the hereditary determinant of a unit characteristic in 1911. The relation between gene and enzyme attained clear definition in the *one gene-one enzyme principal,* first succinctly stated by Beadle in 1945. This hypothesis has since been refined and extended to cover proteins that are not enzymes, as well as complex proteins of nonidentical polypeptide chains linked in various ways.

The functional unit of DNA that controls the structure of a single polypeptide chain is frequently called a *cistron.* The one gene-one enzyme principle was refined as the *one cistron-one polypeptide concept.*

A gene is a heritable unit of phenotypic variation; from a molecular standpoint, a gene is the linear collection of DNA sequences required to produce a functional RNA molecule, or a single transcription unit.

General structure of a typical human gene

Until the late 1970s, the gene was considered as a segment of DNA molecule containing the code for the amino acid sequence of a polypeptide chain and the regulatory sequences necessary for expression. This is now known to be an incomplete description of genes in the human genome (and

in most eukaryotic genomes). In fact, very few genes exist as continuous coding sequences; rather, the vast majority of genes are interrupted by one or more noncoding regions. These so-called intervening sequences or *introns,* are initially transcribed into RNA in the nucleus, but are not present in the mature mRNA in the cytoplasm and are thus not represented in the final protein product. Introns alternate with coding sequences, or *exons,* that ultimately encode the amino acid sequence of the product. Most genes usually contain several introns.

The usual linear order at a gene site is:

- regulatory element(s);
- promoter region (where the RNA polymerase complex binds);
- transcription start site (in 5' UTR) (including CAP site);
- ATG, translation initiation codon;
- exon(s) (variable number);
- introns (between exons, 5'GT and 3'AG, variable number);
- 3' UTR consisting of a translation stop codon (TAA, TGA, or TAG);
- AATAAA polyadenylation signal;
- the site for addition of poly A tail.

Although a few genes in the human genome have no introns, most genes contain at least one and usually several. For example the gene for a type of collagen is split into 52 separate exons. In many genes, the cumulative length of introns makes up a far grater proportion of gene's total length than do the exons. Whereas some genes are less than 1 kb in length, others, such as the factor VIII gene stretch on the for hundreds of kilo-bases. One exceptional gene, the dystrophin gene on the X-chromosome

(mutations of which lead to Duchenne muscular dystrophy) spans more than 2000 kb.

Individual exons correspond to structural and/or functional domains of proteins for which they code. The origin of intron / exon structure is thought to be extremely ancient and to predate the divergence of eukaryotes and prokaryotes. The prokaryotes have lost their introns during evolution, perhaps because of the strong selective pressure in these organisms to retain a small genome size.

Every gene occupies a specific locus along the chromosomes. A given locus can contain one of the alternative versions of a gene; the many versions of a gene in the population are called *alleles*. The same gene loci along the homologous chromosomes contain a pair of genes named alleles. When both alleles of homologous chromosomes are identical, it is said a *homozygote*; when the alleles are different it is a *heterozygote.*

The term homozygote and heterozygote refers to a person or a genotype.

Bidirectional organization of genes and overlapping genes: *g*enome size does not directly correlate with the complexity of the organism. Simple genomes have a high density of genes, which are partially overlapped. Many overlapping genes have been identified in the genome of prokaryotes. The genes of complex organisms are less crowded and the phenomenon of gene overlapping is much rarer. In humans the distance between neighboring genes is rarely 0 with partial overlapping of gene fragments. Although fewer overlapping genes, some have been identified in eukaryotes or mitochondria. For example, genes encoding HLA class III, located on chromosome 6p21.3 have a high density (about 1 gene / 15kb), resulting in gene overlapping. Bi-directional reading gene organization often occurs in genes encoding enzymes involved in DNA repair.

In large genomes genes are not evenly dispersed, but they form clusters of genes. For example out of the 144, or more than 300 known genes, in the human chromosome 21 or 22 respectively, 22% and 18% are located at about 1 kb from one another, while the average spacing distance between genes is ~ 85 kb. A study of 23,752 human genes revealed that more than 10% of the genes whose transcription start sites are separated by less than 1000 pairs of bases are transcribed bidirectionally.

George P. Rédei. Encyclopedia of Genetics, Genomics, Proteomics and Informatics, Springer, 2008, 207-208

Genes within genes: A particularly interesting type of overlapping is one in which a gene is completely included in another gene or in another's gene intron, although the coding sequences of these genes do not overlap. The human genome contains a relatively large proportion of genes located in the introns of other genes. It is estimated that there are 158 functional genes in intronic regions, but also 212 pseudogenes and three genes encoding small nucleolar RNA. These genes seem to be randomly distributed in all chromosomes and most encode proteins that are functionally different from those encoded by their host genes.

Examples of genes located in the introns of genes:

- gene for NF type I (there are three small internal genes transcribed in opposite directions, within intron 27 of the gene)
- gene for clotting factor VIII (in intron 22 of the gene there are two small internal genes, one read in the same direction as the host gene, the other read in the opposite direction)
- retinoblastoma gene (one internal gene)

Genes may be unique sequences or belong to a gene family

Gene family:

Gene families consist of structurally and functionally related genes with a common evolutionary origin. A set of genes in one genome all descended from the same ancestral gene. (a group of genes that has arisen by duplication of an ancestral gene). The genes in the family may or may not have diverged from each other.

For example, the Hb genes belong to one gene family created by gene duplication (making extra copies of a gene) and divergence (divergent changes in the copies of the gene).

Gene families may be close to one another in clusters or they may be dispersed, they may form a cluster on the same chromosome or they may be located on different chromosomes.

Examples include:

- collagen gene family comprises 28 members on different chromosomes
- globin genes familycomprises genes and pseudogenes from chromosomes 11 and 16
- olfactory receptor gene family comprises approximately 1,000 genes in over 25 chromosomal locations. The olfactory apparatus is able to recognize and distinguish thousands of diverse volatile chemicals. This function is mediated by a very large family of olfactory receptors encoded by approx. 1,000 genes, the majority of which are believed to be pseudogenes.
- human immunoglobulin gene family: more than 500 genes placed on different loci on the chromosomes,
- HLA gene family on chromosome 6: 17 genes (including several pseudogenes and gene fragments)

Housekeeping and tissue-specific genes

The genes that encode proteins involved in energy generation or nutrient transport of the cells are described as *housekeeping genes* and they may account for one-fifth (1/5) of the total number of human genes. The four-fifth (4/5) of genes that are not housekeeping are used only at specific times during development. The housekeeping genes lie in regions of DNA that are rich in G-C bases, stain lightly in metaphase chromosomes and replicate early, in the first half of ”S” phase. By contrast, tissue specific genes (non-housekeeping genes) lie in regions of DNA that are rich in A-T bases, stain darkly in metaphase chromosomes and replicate during the second half ” S” phase.

Pseudogenes. A region of DNA with many sequence elements like the other genes but which has no potential for transcription and does not code for a functional product is a *pseudogene.*

Pseudogenes may be part of a gene cluster or family. These gene duplicates are now evolutionary relics. They are related to functional genes but are no longer capable of being transcribed or translated. These pseudogenes cannot be transcribed or translated due to the accumulation of fatal errors such as a nonsense mutations or mutations in the promoter.

A gene superfamily consists of groups of genes that do not show a marked homology among each other. Nevertheless, they are related to each other by the occurrence of subdomains within the proteins encoded by them. Examples are genes encoding immunoglobulins, genes encoding some proteases, many enzymes and receptors, and many genes encoding cytokine receptors.

Mono-allelic expression

Usually both alleles are expressed at molecular level. However, some genes may show mono-allelic expression, (e.g. only maternal or only paternal allele will be expressed). This phenomenon is called allelic exclusion.

Mono-allelic expression may occur:

1. depending on the parental origin of the gene=genomic imprinting
2. independently, through the random repression of a certain allele, such as:

a) inactivation of an X chromosome in females

b) a process whereby only one immunoglobulin light chain and one heavy chain gene are transcribed in any one cell; the other genes are repressed.

c) unknown mechanisms

Multiple alleles

Every gene occupies a specific *locus* along the chromosomes. A given locus can contain one of the alternative versions of a gene; the many versions of a gene in the population are called *multiple alleles*.

Example of multiple alleles: t*he ABO system,* as there are more than two variants for a specific locus. There are four possible blood groups O, A, B and AB. The system is controlled by three allelic variants A, B and O. The genes are located on the long arm (q) of chromosome 9. A and B genes are dominant genes, and O gene is recessive. Dominant genes are expressed both in the homozygous state and in heterozygosity and recessive genes are expressed only in the homozygous state. If the homologous loci are occupied by different dominant genes these genes are said to be codominant.

Thus, the three allele will generate the following genotypes:

- AA or AO for blood type A
- BB or BO for blood type B
- AB for blood type AB
- OO for blood type O

Another example of multiple alleles, very important in the human genome, is the MHC (major histocompatibility complex), involved in the reactivity of the body and major defense systems. Studies describing CMH were rewarded with the Nobel Prize in 1980. Genes are located in the short arms (p) of chromosome 6 and are highly polymorphic. Several hundred alleles are known in the population (in HLA-A region are identified 250 haplotypes, in HLA-B 500 haplotypes and for DR locus over 300 alleles are known.

Chapter V
CONTROL MECHANISMS OF GENE EXPRESSION AND DIFFERENTIATION

Introduction

The controls that act on gene expression (the ability of a gene to produce a biologically active protein) are much more complex in eukaryotes than in prokaryotes. Control mechanisms of gene expression in eukariotes are complex, due to the following aspects:

- DNA molecule is bound to some proteins (histons)
- genes have a mosaic structure (exons, introns)
- mRNA synthesis takes place in two major steps: initially the primary transcript is formed, then processing of the primary transcript will produce the mature mRNA
- protein synthesis is carried out in the cytoplasm and the mRNA migrates from the nucleus to the cytoplasm.

Regulation of gene expression involves multiple factors that intervene in several stages and is accomplished through two main types of control:

- Long-term genetic control - also known as programmed control, is embedded in cell memory, hereditary and stable. It is usually irreversible, occurs during ontogenetic development and is related to cell differentiation.

- Short-term genetic control - also known as adaptive control, based on reversible molecular mechanisms, represented by changes in genes activity, leading to fluctuations in the intensity of DNA, RNA and proteins synthesis.

Genetic information is the same in all cells of a multicellular organism. From the zygote stage, the human body arises through repeated mitotic nuclear divisions. The first few mitotic divisions of the embryo produce cells that have the same genetic information, but then, due to the various control mechanisms, the cells diverge and give rise to different cell types. This differentiation (specialization) and its maintenance mainly depend on the cellular mechanisms that control formation of specific proteins.

Expression of genes in the human zygotes starts in the pre-implantation stages, between four to eight cells. This stage consists of two main phases: one from oocyte's meiosis II until the four cells embryo, when the mother's genes prevail and the other phase from 8 cells embryo until blastocyst, when mother's genes are gradually inactivated and zygote's genes become active, eg. specific genes for differentiation and/or organogenesis, genes involved in the development of multicellular organisms, cell adhesion, cell signaling, etc.

Control mechanisms determine the different expression of genes, resulting in cell specialization. Gene expression should be controlled on long term during cellular differentiation. Genes are continuously activated or inactivated in response to different signals from their internal and external environment. Obviously, only part of the total genetic information is expressed in a specialized cell.

! All body cells have the same genes. The cells of the body are not different because they contain different genes, but because the expression of these genes is different.

An estimate of the number of different mRNA sequences in human cells suggests that in a certain moment, a differentiated human cell expresses 5000-15000 genes out of about 25,000 genes. The large variations in size, shape, behavior and function of differentiated cells are determined by the expression of a collection of different genes in each cell type.

Once a cell has differentiated, it will generally remain differentiated and all its progeny cells will remain of the same type. Many differentiated cells, such as fibroblasts, smooth muscle cells or hepatocytes will divide many times during lifetime and will lead only to specialized cells identical to the parent cell (eg. hepatocytes will divide, forming only hepatocytes) .

Preserving the identity of the cell implies that changes in gene expression which cause the formation of differentiated cell have to be memorized and transmitted to daughter cells during subsequent cell divisions. How can cells accomplish this? Cells have several ways to ensure that their daughters will 'remember' what kind of cells they should be. One way is through a positive feedback loop, in which a key regulator activates transcription of its own gene, in addition to transcription of other cell-specific genes. Another way is by accurate propagation of condensed chromatin structure from mother cell to daughter cell, example: same inactive X chromosome is transmitted to all daughter cells. A third way is through DNA methylation.

In vertebrates DNA methylation occurs only at cytosine. This modification of cytosine generally inactivates genes by attracting proteins that block the expression of the genes. The patterns of DNA methylation of the cell are transmitted to the daughter cells by the action of an enzyme that copies the methylation pattern of the parent DNA strand to the - daughter DNA strand immediately after replication.
These mechanisms transmit information from mother cell to daughter cell without changing the actual nucleotide sequence of DNA, they are considered forms of ***epigenetic inheritance****.*

There are several control mechanisms which control cell differentiation and cell changes in different tissues. The genes, which direct the amino acid sequences in polypeptides, are called structural genes. Regulatory genes such as *the operator* and *the regulatory gene* coordinate their activity. The structural genes plus the operator and promoter form an *operon.*

In prokaryotes the operator is located adjacent to the structural genes. The promoter region to which RNA polymerase attaches is next to the operator. All these genes within an operon are transcribed on the same mRNA. The regulatory gene is located near the promoter. Jacob and Monod (awarded with Nobel prize in 1965 for discoveries on the genetic control of enzyme synthesis and viruses) were able to explain that the regulatory gene is responsible for the production of a repressor substance, which coordinates transcription. In fact, the repressor blocks the operator and thus, the structural genes. Transcription is thus effectively blocked by the interaction of the repressor with the operator site. This is the control of gene expression at the transcriptional level.

Gene Control in Eukaryotes

In eukaryotic cells, the control of gene expression takes place at several points:

1. Chromatin Structure: The physical structure of the DNA, as it exists compacted into chromatin, can affect the ability of transcriptional regulatory proteins (termed **transcription factors**) and RNA polymerases to find access to specific genes and to activate transcription from them. The presence of the histones and CpG methylation most affect accessibility of the chromatin to RNA polymerases and transcription factors. DNA methylation causes repression of transcription by inhibiting the binding of transcription factors to target sequences. Sequences that will be actively transcribed require demethylation. Methylation is a control process and it determines *when the gene will be expressed* (as in embryonic development, in which genes are switched on and off in a sequential manner), *the inactivation of one X chromosome in females* and in mammals *differential expression of certain genes depending on maternal or paternal origin* ("genomic imprinting"). Methylation of cytosine is very important and the methylation pattern together with the gene form the "**epigenome**".

2. Transcriptional Initiation: This is the most important mode for control of eukaryotic gene expression. Specific factors that exert control include the strength of **promoter elements** within the DNA sequences of a given gene, the presence or absence of **enhancer sequences** (which enhance the activity of RNA polymerase at a given promoter by binding specific transcription factors), and the interaction between multiple activator proteins and inhibitor proteins.

3. Transcript Processing and Modification: Eukaryotic mRNAs must be capped and polyadenylated, and the introns must be accurately

removed. Several genes have been identified that undergo tissue-specific patterns of alternative splicing, which generate biologically different proteins from the same gene.

mRNA editing: is a form of processing in which specific changes are produces. Thus, by modifying the mRNA, different length proteins will be produced. Example: mRNA for apolipoprotein B produces Apo-B100 in the liver, while in small intestine produces Apo-B48, the first 48 AA are the same as those of the Apo-B100. Apo-B48 occurs due to a change in a sense codon, which becomes stop codon and stops translation at AA number 48.

4. RNA Transport: A fully processed mRNA must leave the nucleus in order to be translated into protein.

5. Transcript Stability: Eukaryotic mRNAs can vary greatly in their stability. Certain unstable transcripts have sequences that are signals for rapid degradation.

6. Translational Initiation: Since many mRNAs have multiple methionine codons, the ability of ribosomes to recognize and initiate synthesis from the correct AUG codon can affect the expression of a gene product. Several examples have emerged demonstrating that some eukaryotic proteins initiate at non-AUG codons. This phenomenon has been known to occur in *E. coli* for quite some time, but only recently has it been observed in eukaryotic mRNAs.

Recently, it was discovered that noncoding RNA that is not directly involved in the production of a certain protein is far more widespread than previously believed and plays important roles in regulating gene expression. Such RNA is micro-RNA. In humans, for example, there are more than 400 different types of micro-RNA, which appear to regulate at least one third of

the total protein-coding genes. These RNAs control gene expression by base pairing with specific mRNAs, controlling their stability and translation.

Some of the proteins which processes micro-RNAs, also serve as a defense cellular mechanism: they destroy "foreign" RNA molecules, especially those that are double-stranded. Many viruses and transposable elements sometimes produce double-stranded RNA in their life cycles. This mechanism of degradation of target RNAs is called **RNA interference** (RNAi), helping to maintain control over these potentially dangerous "invaders".

In a practical sense, RNAi has become a powerful experimental tool that allows scientists to inactivate almost any gene in cell cultures. At the same time, using RNAi may represent a new approach to treat human diseases. Many diseases are due to inappropriate expression of genes, and the ability to inhibit these genes by this mechanism is very promising for future treatments.

7. Post-Translational Modification: Common modifications include glycosylation, acetylation, fatty acylation, disulfide bond formations, etc.

8. Protein Transport: In order for proteins to be biologically active following translation and processing, they must be transported to their site of action.

9. Control of Protein Stability: Many proteins are rapidly degraded, whereas others are highly stable. Specific amino acid sequences in some proteins have been shown to bring about rapid degradation.

Control of Eukaryotic Transcription Initiation

Transcription of the different classes of RNAs in eukaryotes is carried out by three different polymerases. RNA pol I synthesizes the rRNAs, except for the 5*S* species. RNA pol II synthesizes the mRNAs and some small nuclear RNAs (snRNAs) involved in RNA splicing. RNA pol III synthesizes the 5*S* rRNA and the tRNAs. The vast majority of eukaryotic RNAs are subjected to post-transcriptional processing.

The most complex controls observed in eukaryotic genes are those that regulate the expression of RNA pol II-transcribed genes, the mRNA genes. Almost all eukaryotic mRNA genes contain a basic structure consisting of coding exons and non-coding introns and basal promoters of two types and any number of different transcriptional regulatory domains (see diagrams below). The basal promoter elements are termed **CCAAT-boxes** (pronounced cat) and **TATA-boxes** because of their sequence motifs. The TATA-box resides 20 to 30 bases upstream of the transcriptional start site and is similar in sequence to the prokaryotic Pribnow-box (consensus TATAT/AAT/A, where T/A indicates that either base may be found at that position).

Numerous proteins identified as **TFIIA, B, C**, etc. (for **t**ranscription **f**actors regulating RNA pol **II**), have been observed to interact with the TATA-box. The CCAAT-box (consensus GGT/CCAATCT) resides 50 to 130 bases upstream of the transcriptional start site. The protein identified as **C/EBP** (for **C**CAAT-box/**E**nhancer **B**inding **P**rotein) binds to the CCAAT-box element.

There are many other regulatory sequences in mRNA genes, as well, that bind various transcription factors (see diagram below). Theses regulatory sequences are predominantly located upstream (5') of the

transcription initiation site, although some elements occur downstream (3') or even within the genes themselves. The number and type of regulatory elements to be found varies with each mRNA gene. Different combinations of transcription factors also can exert differential regulatory effects upon transcriptional initiation. The various cell types each express characteristic combinations of transcription factors; this is the major mechanism for cell-type specificity in the regulation of mRNA gene expression.

The genetic information is the same in all the cells of the body of a multicellular organism. From the zygote onward, a human organism arises through repeated nuclear divisions. The first few mitotic divisions of the embryo produce cells that are essentially the same (they have the same genetic information), but then, the cells diverge and give rise to different cell types. This differentiation (specialization) and its maintenance depend mainly on cell mechanisms that control the formation of specific protein products. Obviously, only a part of the total genetic information is expressed in any one cell.

There are several control mechanisms which control cell differentiation and cell change in different tissues. The genes, which direct the amino acid sequences in polypeptides, are called structural genes. Regulatory genes such as *the operator* and *the regulatory gene* coordinate their activity. The structural genes plus the operator and promoter compose an *operon.*

The operator is located adjacent to the structural genes. The promoter region to which RNA polymerase attaches to begin transcription is next to the operator. All these genes and sites within an operon are transcribed on the same mRNA. The regulatory gene is located near the promoter.

Jacob and Monod were able to explain that the regulatory gene is responsible for the production of a *repressor substance,* which coordinate the transcription. In fact, the repressor blocks the operator and thus the RNA polymerase that attaches to the promoter site cannot transcribe the information of the structural genes into messenger RNA. Transcription is thus effectively blocked by the interaction of the repressor with the operator site. This is the control of gene expression at a **transcriptional level**. Regulation of gene expression also occurs **before the translational level**, when the splicing process takes place in order to form the final m-RNA. The messenger RNA for each gene undergoes extensive processing to remove intronic sequences. This process can include alternative splicing (two or more different ways). This means that one gene could carry information for more than one kind of product. On the other hand, this process is seen in that a given gene though its transcript may be handled differently, depending on the cell types, is not actually expressed at all in most cells of the organism. Housekeeping genes are indeed expressed in all cells; however, only a small percentage of the genome is being expressed in any given cell type.

Regulation of gene expression also occurs at a **posttranslational level**. These processes can include the regulation of the export of mRNA from the nucleus into the cytoplasm, alternative splicing of transcripts, polyadenylation of transcripts, translation of final mRNA and, stability of mRNA. There are numerous examples where gene function is regulated through alternative splicing including control of immunoglobulin secretion, production of calcitonin versus calcitonin gene-related peptide and, modulation of the structure and function of other products.

The mechanisms of regulation of gene expression are also engaged in **the process of organism development during embryo-fetal time**. It is the special triumph of molecular biology to have replaced the phenomenology of embryology with molecular explanations of mechanisms. *Induction*, fields and gradients have given way to descriptions of the molecules engaged in fertilization, cell division and cell determination and differentiation. Molecular mechanisms of intercellular communication, migration, adhesion, association, morphogenesis and selective cell death have been discovered and lineages of cells have been traced to their final roles in the coordinated assembly of specialized cells in the tissues and organs of the embryo. All these processes are under strict genic control and while chance enters in, for example, in the choice of which cells to migrate in which direction and which to die off, there is little leeway for mistakes, which would result in distortions. So, embryogenesis is more nearly **programmed** than later fetal development.

A considerable **regulation of gene expression** is observed to be closely related to the gestational age of the fetus and largely independent of environmental factors. Several observations provide evidence for the preprogrammed gene expression. A very evident example is the developmental changes in globin gene expression, which will be summarized, in the following observations.

The pattern of Hb synthesis changes during development. In the very early embryo Hb synthesis is restricted to the yolk sac and the production of Hb Gower. At about 8 weeks of gestation, the fetal liver takes over, synthesizing predominantly Hb F and a small amount (less than 10%) of Hb A. Between about 18 weeks and birth, the liver is progressively replaced by bone marrow as the major site of red-cell production and this is

accompanied in the later stages of gestation with a reciprocal switch in production of Hb F and A, which continues, until by the end of the first year when Hb F production has dropped to less than 2 percent. This switch from Hb F to Hb A production means the reduction of gamma gene's activity in parallel with the increasing alpha gene's activity.

The above example shows the developmental regulation of globin genes during different developmental prenatal periods. In humans, the switch of globin gene activity seems to be closely related to the gestational age of the fetus depending of developmental specific period.

Chapter VI
ORGANIZATION OF GENOMIC DNA

The human genome is found mostly in the nucleus (called the nuclear genome), with approximately 26,000 genes, but there is also a very small quantity of circular mitochondrial DNA that forms the mitochondrial genome, which has only 37 genes. The organization of DNA in the human genome is complex.

The nuclear genome is distributed between 24 linear DNA molecules, one DNA molecule for each of the 24 different human chromosomes.

Molecular studies on the structure of the human genome revealed that a gene, in general, is composed of approximately 20,000-30,000 nitrogenous base pairs (longer genes have been identified, however, made up of hundreds of thousands of pairs of bases). Due to the fact that the genomic DNA consists of approximately 3 billion bases, it may form hundreds of thousands of genes. But in the human genome there are only about 26,000 structural genes, thus only a small part of the total amount of DNA is encoding proteins, the remainder will replicate, but will not be transcribed.

This part of the genome was called *junk DNA* or *extragenic DNA*. The possibility that a large proportion of the complex human genome to be non-functional has been discussed for a long time. It seems a paradox the

observation that the genome size does not correlate with the complexity of an organism and that closely related species may have very different genome sizes. It would seem an inefficient activity of the human genome in comparison with bacterial genome, which contains only coding sequences. Completion of sequencing the human genome was a major achievement in modern biology, but considerable challenge that remains is the identification and description of all genes structures and other functional elements. It is estimated that nearly 99% of the approximately 3.3 billion nucleotides that constitute the human genome does not encode proteins. More recently, it has been shown that many gene loci associated with specific normal traits, with certain human diseases or susceptibility to diseases are also outside the coding region.

These findings suggest that the human genome noncoding regions might have a wide range of important elements in terms of function, with various control functions, but maybe with other roles, too.

Types of organization of the nucleotide sequences in nuclear DNA:

The organization of DNA in the human genome is complex. There is:

- coding DNA and non-coding DNA
- unique DNA and repetitive DNA

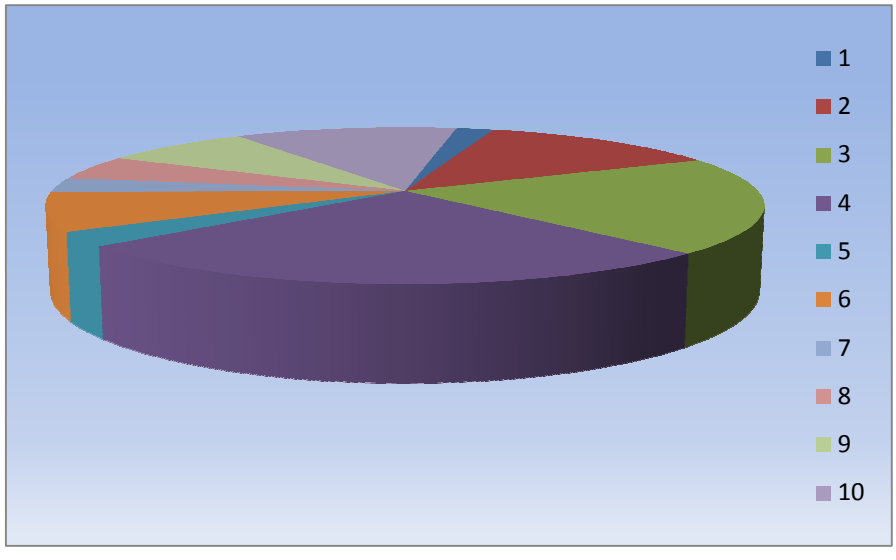
Coding DNA and non-coding DNA

Coding DNA:

- Is about 1.5-2% of the genome
- Encodes the proteins that underly the phenotypic traits

Non-coding DNA:

- Represents about 98% of the genome
- Is contained in:
  a. pseudogenes
  b. gene fragments
  c. introns
  d. telomeres
  e. sequences encoding the rRNA or tRNA
  f. regulatory sequences
  g. mobile elements: transposons and retrotransposons. Transposons are nucleotide sequences that change position from a genomic location to another, through a mechanism "cut and paste" or "copy and paste". Retrotransposons are complementary DNA (cDNA), complementary copies of RNA, which then integrate into a new genomic location.



Types of organization of the nuclear DNA nucleotide sequences

1. protein encoding genes; 2. SINEs; 3. LINEs; 4. introns; 5. transposons; 6. retrotransposon; 7. simple repetitive sequences; 8. duplicated segments; 9. various unique sequences; 10. heterochromatin

Unique DNA and repetitive DNA

Unique DNA

Single-copy or unique DNA is DNA whose nucleotide sequence is represented only once per haploid genome. It includes coding sequences for structural genes, which account for about 3% of the genome. The remainder is represented by intronic sequences or spacer DNA. The single-copies of DNA are interspersed with various repetitive DNA along the chromosomes. The great majority of genetic information for the synthesis of polypeptides is encoded by single-copies DNA.

Repetitive DNA

The remaining DNA contains two classes of *repetitive DNA*.

A nucleotide sequence of a certain length can be repeated several times in the DNA molecule of a cell. Approximately 50% of non-coding DNA sequences consist of repetitive DNA, which is represented by nucleotide sequences repeated more than 20 times per genome. Repetitive elements in the genome differ in terms of position, nucleotide sequence, size, number of copies and the presence or absence of coding regions within them. These repetitive elements can be:

• dispersed in the genome or

• arranged in tandem.

Tandem repeats represent adjacent repeats of a sequence consisting of two or more nucleotides. Ex: ATTCGATTCGATTCG = $(ATTCG)_3$.

Depending on the number of repeats in the human genome, these sequences are classified into:

- ***highly repetitive DNA***, about 10% of total genome, present at more than 10^6 copies per genome, sequences with variable lengths, in long tracts of up to 100 Mb

Depending on the size of these repeats we can can be distinguish: satellite DNA, minisatellite DNA and microsatellite DNA. Satellite DNA is transcriptionally inactive, as is most of minisatellite DNA, but microsatellite DNA can be found in coding regions.

- *Satellite DNA* is made up of a large number of tandem repeats of approximately 100 nucleotides. Satellite DNA may be present in heterochromatic regions, especially pericentromeric (having a role in attachment of chromatids to the spindle) and telomeric regions (involved in DNA replication at this level). Alpha-satellite DNA is a subtype of satellite DNA, consisting of repetitions of 171 base pairs, with a role in the kinetochore formation and normal cell division.
- *Minisatellite DNA* (variable number of tandem repeats-VNTRs) contains repeats of a number of 10-60 base pairs. These sequences are highly polymorphic and are located at or near the telomeres. Their role is not entirely known, are important for linkage analyses studies.

An important family of such minisatellite DNA is telomeric DNA, consisting of tandem repeats of six nucleotides, added by a specialized enzyme, telomerase. These repeats are responsible for telomeres function, as they protect chromosome ends and provide a mechanism for their replication. In a germ cell telomere length is about 15 kb, but an aging somatic cell telomeres are much shorter.

- *microsatellite DNA* (short tandem repeats-STRs) consists of units represented by 2-10 nucleotides repeated about 100,000 times / haploid genome. STRs instability is implicated as a progression factor in some cancers (eg. colorectal cancer). The number of repetitions in a given locus is highly variable between individuals of the same species.

Three nucleotides repeats are of great clinical importance due to their direct involvement in the determinism of the disease. Example: The three nucleotides CAG in Huntington's disease or CGG in fragile -X syndrome.

The number of repeats for a given minisatellite or microsatellite may differ between individuals. This feature is the basis of DNA fingerprinting.

- ***moderately repetitive DNA***, about 30% of the genome, present at between 10 - 10^5 copies per genome, found throughout the euchromatin, average 300bp in size

It includes genes for histones and rRNA. It may be involved in regulation of gene expression, supported by the fact that the single-copies of DNA are interspersed with various repetitive DNA along the chromosomes. This class includes interspersed repeats, which are repeated DNA sequences located at dispersed regions in a genome. They are also called mobile elements or transposons. A DNA sequence may be copied to a different location through DNA recombination and aftermany generations, such sequences could spread over various regions in the genome. In mammals, the most common mobile elements are *LINEs* (Long Interspersed Nuclear Elements) and *SINEs* (Short Interspersed Nuclear Elements).

The most common LINEs in humans is the L1 family. There are about 60,000 to 100,000 L1 elements in the human genome. Such elements inserted into the introns of the functional genes will decrease transcription of those genes. The level of gene expression will lower with the longer size of L1 element. It is estimated that humans have L1 elements in about 79% of their genes, perhaps they are a mechanism for establishing the baseline level of gene activity.

The most abundant SINEs in humans is the *Alu* family.

Transposons they cause diseases as their integrations into biologically important genes lead to insertional mutations. However, although a large number of TEs are transcriptionally active, only a small subset (<0.01%) are able to transpose, thus able to cause mutations. \

It is now known that insertion of L1 elements caused:

- haemophilia
- Duchenne muscular dystrophy
- sporadic breast and colon cancer.

Insertion of an Alu element in NF1 gene (chromosome 17) may lead to neurofibromatosis type 1.

Integrations have been also seen in oncogenes or in tumor suppressor genes, leading to different malignancies, but the exact mechanism is yet unclear.

The great majority of genetic information for the synthesis of polypeptides is encoded by single-copies DNA.

The repetitive DNA molecules have a complex role in regulatory genes.

The inheritance of the about 26,000 genes in each of us defines us as individuals. However, there are substantial variations between every individual's genes. These variations can either be represented by repeats of certain nucleotide sequences (repetitive DNA) or may be changes of a single nitrogenous base (called polymorphism if the frequency of> 1% and mutations if the frequency is <1%).

Polymorphism of a single nitrogenous base is called SNP (single nucleotide polymorphism, read snip) and is the most common form of genetic variation between individuals. The number of these small polymorphisms is extremely high. Their study is useful in medicine, may help predict response to administration of drugs, determine susceptibility to environmental factors or the risk of getting a particular disease. The differences between the genomes of different individuals are also called variants. Most of these variations are harmless, some are beneficial, while others cause different diseases or susceptibility to certain diseases. Understanding how genomic variants contribute to the disease may help improving diagnosis, treatment and identifying new methods of prevention.

A new project "1000 genomes project" identified more than 99% of existing variants in the human genome, occurring at a frequency of at least 1% in the studied populations (polymorphisms). They identified about 88 million sites in the human genome that vary between individuals, establishing a database available to researchers as a standard reference for how genome structure varies within populations and worldwide. This catalog presents a huge number of variations/polymorphisms and can now be used in studies of human biology and medicine, providing the basis for a new understanding of how inherited differences in the DNA may contribute to the risk of disease and to different responses to drugs. The study showed that the greatest genomic diversity is in African populations, in accordance with the evidence that humans originated in Africa and the African migrations have established other populations around the world.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 2015;526:68–2015;526:68–74

Mitochondrial genome organization

Mitochondria possess their own genome, represented by small circular DNA, encoding some of the components required for mitochondrial protein synthesis in mitochondrial ribosomes. However, most mitochondrial proteins are encoded by nuclear genes and are synthesized in the cytoplasm ribosomes before being imported into mitochondria.

- mitochondrial genes do not have promoter sequences.
- coding sequences forms 37 genes, which are encoding:
  - polypeptides (constituents of the oxidative phosphorylation system) (13 genes),
  - tRNA (22 genes)
  - rRNA or (2 genes).
- mtDNA genes do not have introns.
- Transmission is exclusively maternal.

CHROMOSOMAL BASIS OF INHERITANCE

The human genome functions as a complex, folded, three-dimensional chromatin polymer. Understanding how the human genome is folded and organized inside the cell nucleus is essential for understanding how genes are regulated in normal development and disregulated in diseases.

Organization of genetic material is different during interphase and during cell division. Chromosomes (during cell division) and chromatin (during interphase) are two morphological forms of the same genetic nuclear material.

Chromosomes are the vehicles of inheritance. Their behavior in cell division provides the mendeleian basis for laws of inheritance. In addition, their organization is important for the other major functions of DNA:

transcription and replication. Chromosomes contain virtually all the DNA of the cells, the only exception being the tiny amount of DNA in the mitochondria. Mitochondria are derived wholly from the ovum, and thus the few mitochondrial genes show strictly maternal inheritance. The chromosomes are derived equally from mother and father. Each ovum and sperm contains a complete set of 23 different chromosomes, the haploid number (n), while the diploid fertilized egg and virtually every cell of the organism arising from it has two compete sets (2n=46).

Each human diploid cell nucleus contains about 6x109 base pairs (bp) of DNA.

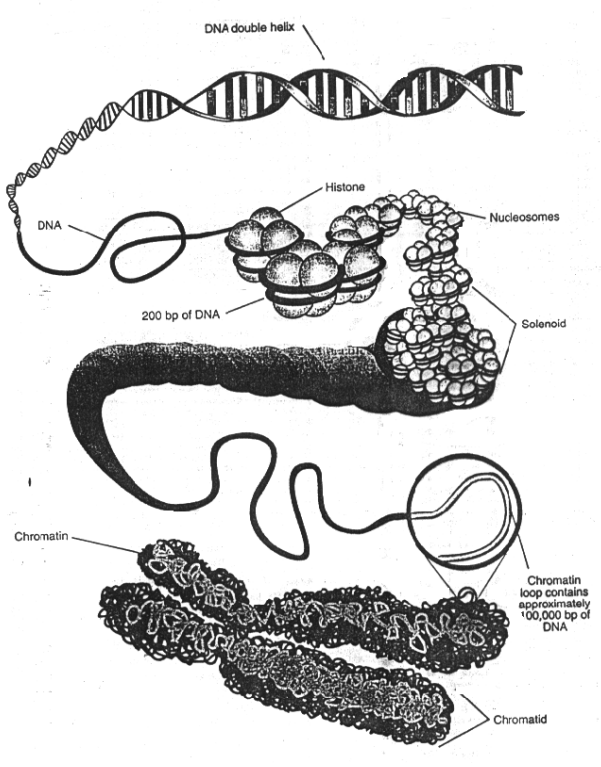
Structure of human chromosomes

The chromosomes are organized in a manner that facilitates the functions of its DNA throughout the cell cycle such as transcription and replication.

The DNA molecule of a chromosome exists as a complex with some of basic chromosomal proteins called histones and with a heterogeneous group of nonhistone proteins that are much less well characterized. This complex of DNA and protein is called *chromatin.* The DNA molecule is packaged in a hierarchic fashion of everlarger supercoils. The first level of supercoiling is the nucleosome, the basic structural unit of chromatin. Each nucleosome is a cylinder in which 140 base pairs (bp) of DNA are wrapped around a histone core containing two molecules (copies) of each of the four core histones: H2A, H2 B, H3 and H4; the histones constitute an octamer, around which a segment of DNA double helix winds (note that the DNA is on the outside the nucleosome). A connecting piece of DNA (about 60 bp long) lies between adjacent nucleosomes together with one molecule of histone H1 (internucleosomal region).

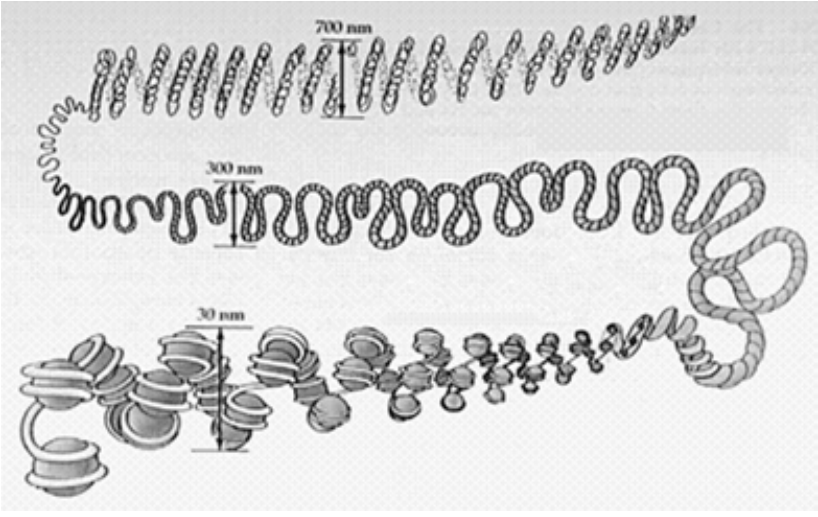
The next higher level of packaging is the secondary chromatin structure solenoid in which some nucleosomes are compacted (they appear under the electron microscope as a thick fiber).

Histone H1 blocks transcription, whereas nonhistone proteins permit access to RNA polymerase complexes and thus transcription.



Patterns of DNA coiling. DNA is wound around histones to form nucleosomes. These are then folded and finally organized into chromosomes

The solenoids are themselves packed into loops attached to some nonhistone protein (a matrix), forming the 300 nm fiber. The final level of packaging is characterized by the 700 nm fiber.



Patterns of DNA coiling. 30 nm, 300 nm, 700 nm fibers

During the cell cycle, the chromatin fibers go through orderly stages of condensation and decondensation. In the interphase nucleus, chromosomes and chromatin (they being the same material) are quite decondensed in relation to the highly condensed state of chromatin in metaphase. Nonetheless, even in interphase chromosomes, a part of DNA chromatin is condensed. Unlike the chromosomes seen in stained preparations under the microscope or in photographs, the chromosomes of living cells are fluid and dynamic structures. The loops of the chromatin fibers may be the beginning of the knoblike thickenings called chromomeres observable under the microscope in early prophase chromosomes, as mitosis begins. As chromosomes condense further, adjacent chromomeres will be compacted into larger ones. These clusters of chromomeres eventually become the dark-staining bands of G-banded prophase or metaphase chromosomes.

At prophase, chromosomes can show about 800-1000 bands after a special stained chromosome preparation (high-resolution banding). After

metaphase, as cells complete mitosis or meiosis, chromosomes decondense and return to their relaxed state as chromatin in the interphase nucleus, ready to begin the cycle again. This chromatin contains nonrepetitive DNA molecules and they are transcribed into RNA. They tend to replicate early in the DNA synthesis stage of the cell cycle. In contrast, the chromatin that stains darkly throughout the cell cycle, even in interphase is called heterochromatin. This chromosome region contains nontranscribed DNA and it replicates later in the "S" phase of the cell cycle.

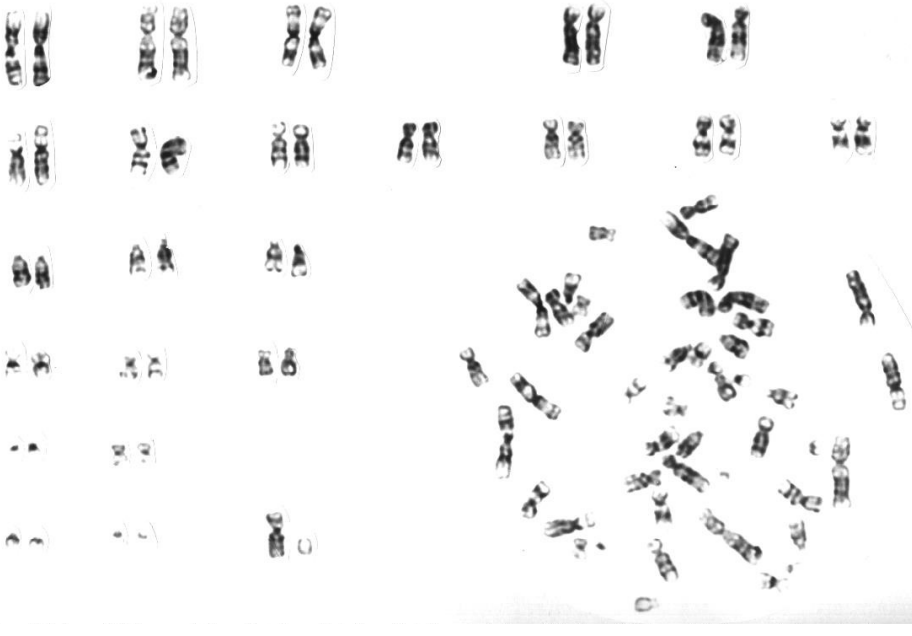
The heterochromatin which is composed of satellite DNA in regions such as centromere, acrocentric short arms and on the chromosomes 1"q", 9"q", 16"q", and Y"q", is known as constitutive heterochromatin; the inactive regions of X chromosome of the female cells contain facultative heterochromatin.

Many areas of heterochromatin are located close to chromosome centromere and to telomeres of acrocentric chromosomes. It may play a structural role in chromosome organization.

The chromosome bands (G, C, R, Q, etc.) reflect a very general functional organization, in which euchromatin alternates with heterochromatin.

The DNA in the G, C, Q bands (positive bands) and in the R bands (negative bands) is usually in a condensed state, reflecting its inactivity throughout most of the cell cycle.

Chromosomes are not visible during interphase, except for highly contracted heterochromatic regions, which are transcriptionally inactive that is involved in RNA synthesis. One example of such inactive DNA is sex chromatin, the Barr body (facultative heterochromatin, described later).

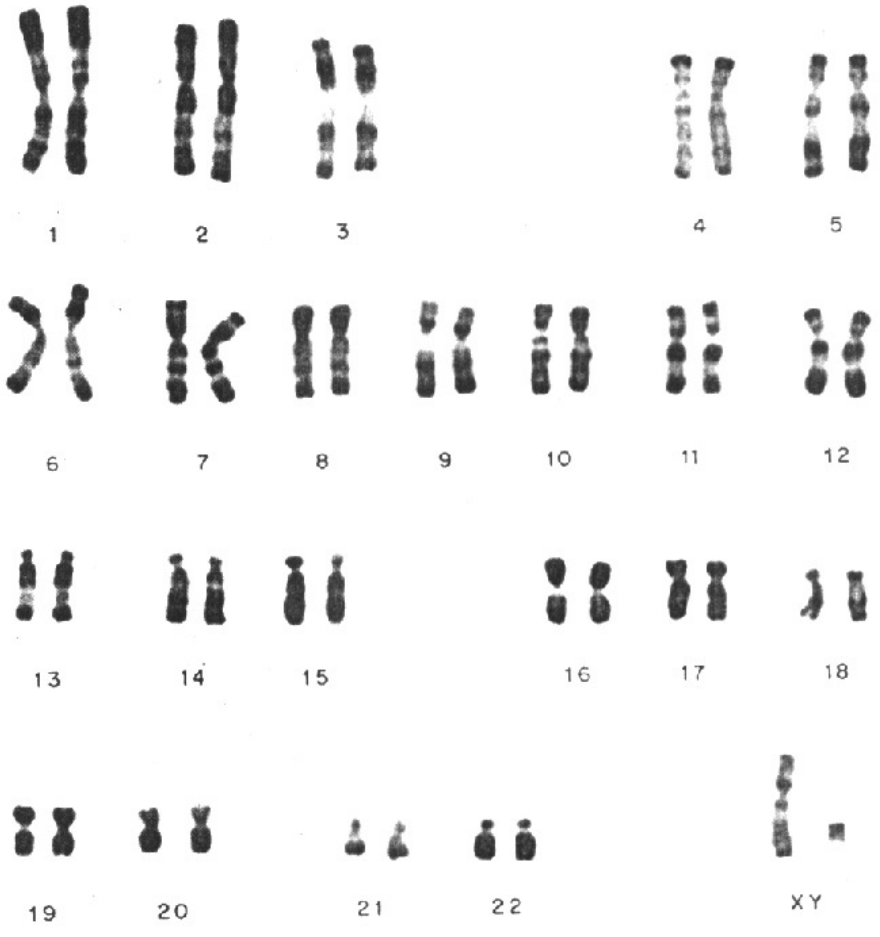


G-banded human male karyotype – from our collection

G-banded human female karyotype – from our collection

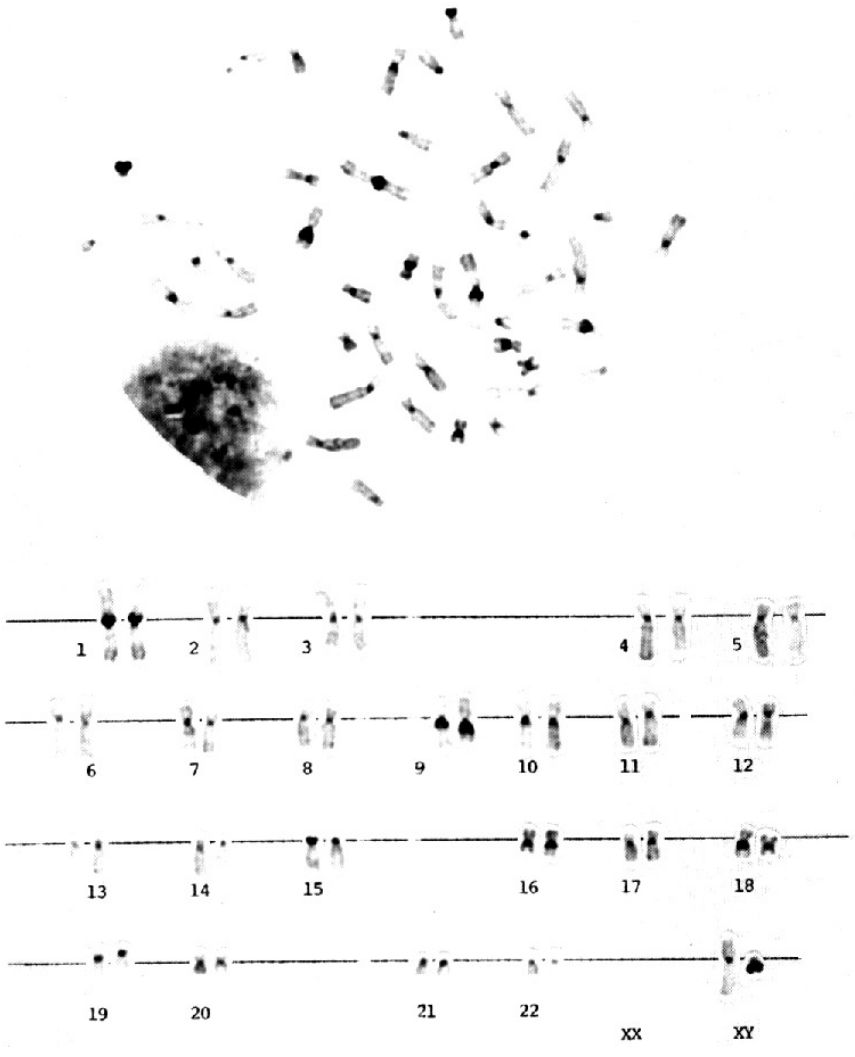
.

A male G-banded metaphase – from our collection.



An R-banded human male karyotype -from Genetics in Medicine, Thompson, 1991

Even though the chromosomes are not usually visible at interphase, they retain their continuity and can be visualized by special techniques. The most important of these is the premature chromosome condensation (PCR) technique, which can be brought about by fusing an interphase cell with a metaphase cell. As a result, the nuclear membranes of the chromosomes condense sufficiently to become visible, even though they may still be quite long and thin. This method permits detailed study of the haploid chromosomes in human sperm.



Metaphase and karyotype of human male showing C-banding (Thompson, 1991)

Chapter VII
MUTATIONS IN HUMAN DISEASES

Genetic Variability

Genetic variability represents the set of phenomena that ensure the appearance of differences between individuals of the same population and between different populations. These differences in genetic information may have no effect on the phenotype or may lead to the emergence of genetic diseases. Between these two extremes lie polymorphisms (variants with a frequency of 1% or more in a population).

Terminology

In some disciplines, the term "mutation" is used to indicate "a change/a variant," while in others it is used to indicate "a variant that causes disease." Similarly, the term "polymorphism" is used both to indicate "a change that does not cause disease" or "a change found at a frequency of 1% or more in a population." To avoid this confusion, we will use neutral terms such as "variant" and "change." The term "mutation" has developed a negative connotation. Who would want to carry a mutation and thus be a "mutant"?

Current guidelines from authoritative organizations recommend using only neutral terms such as "variant" and "change" (see Richards 2015, Genet. Med. 17:405–424).

In cancer genetics, the terms “mutation” and “mutational load” are often used. While changes in an individual’s genome compared to the reference genome should be called “variants,” additional changes (de novo somatic variants) in cancerous tissue are referred to as “mutations.” In such cases, “mutational load” and “tumor mutational burden” are acceptable terms, but only for variants proven not to be present in the individual’s germline.

Another commonly used term is “pathogenic variant.” A variant is not “pathogenic” by itself unless it can be causally linked to an observed phenotype in a patient, for example:

- causes disease when found in a male (X-linked recessive disorder),
- causes disease when combined with a modification of the other allele (autosomal recessive),
- causes disease when inherited from the father (imprinting disorder).

Each variant may have both a functional and a clinical classification (Houge 2022. Eur J Hum Genet 30, 150–159). A classification system with categories can be used:

Functional classification focuses on the impact on gene function

- affects function - pathogenic variant,
- probably affects function - likely pathogenic variant,
- unknown effect on function - variant of uncertain significance,
- probably does not affect function - likely benign variant,
- does not affect function - benign variant.

Clinical classification focuses on the genotype-phenotype relationship including:

- unknown clinical relevance (clinical VUS),
- gene-phenotype match,
- known risk factor, variant of interest
- pathogenic variant (depending on penetrance, it can be: low, moderate, high).

Across the spectrum of diversity from rare variants to more common polymorphisms, genetic variants arise in the context of fundamental processes of cell division such as DNA replication, DNA repair, DNA recombination, and chromosome recombination in mitosis or meiosis.

Genetic recombinations

Genetic recombinations are normal phenomena that occur during meiosis and fertilization and determine the appearance of new genetic combinations. These recombinations sum up natural exchanges of genetic material of variable sizes: between chromosomes, regions of chromosomes and genes.

Chromosomal recombination during meiosis

During meiosis I, 2 recombinatory phenomena occur: intra and interchromosomal.

Ø *Intrachromosomal recombination* occurs during prophase I and is the consequence of crossing-over. This type of recombination ensures the equal and reciprocal exchange of genetic material between homologous chromosomes. During prophase, homologous chromosomes align perfectly along the chromatids, forming the synapse. At this level, crossovers occur

between homologous chromatids. At the crossover points, chromosomal breaks occur followed by gene exchange between homologous chromosomes.

Due to intrachromosomal recombinations, at the end of prophase I, each chromosome will have segments of maternal and paternal origin. The number of intrachromosomal recombinations depends on the size of the chromosomes and sex, being 5-6 for large chromosomes and at least 1-2 for small chromosomes. The number of intrachromosomal recombinations is higher in oogenesis, being estimated at 70-75 recombinations per cell, compared to 40-45 recombinations per cell in spermatogenesis.

In pathological cases, imperfect alignment of homologous chromosomes will lead to unequal crossing-over, in which parts of non-homologous genes are exchanged, resulting in the synthesis of structurally and functionally modified proteins.

Ø *Interchromosomal recombination* occurs in anaphase of meiosis I. After chromosome disjunction and migration, a chromosome group consisting of a mixture of maternal and paternal chromosomes will arrive at each cell pole.

The two processes of meiotic recombination make gametes genetically different, and each chromosome of a gamete will contain genetic material of maternal origin mixed with genetic material of paternal origin.

Genomic recombination

Genomic recombination is achieved by the fusion of the maternal genome with the paternal genome during fertilization. In humans, marriages between genetically different people (unrelated) cause the gametes participating in fertilization to present different “hereditary dowry” among

themselves, so that the descendants of a couple are genetically different, thus ensuring heterogeneity, hereditary variability. Conversely, when the individuals of a couple are related (consanguineous), the degree of variability is reduced, increasing the risk of recurrence of recessive diseases (risk of meeting two heterozygotes with pathogenic variants in the same recessive gene).

Genome Diversity

Individuals vary greatly in a wide range of biological functions, driven in part by variation among their genomes.

Variation detected in a typical human genome (differences from the reference genome):

- ≈5-10 million SNPs (varies by population)
- 25,000-50,000 rare variants (private variants, previously observed in < 0.5% of tested individuals)
- ≈75 de novo variants not detected in the parental genome
- 3-7 new copy number variations (CNVs) involving ≈500 kb of DNA
- 200,000-500,000 indels (1-50 bp) (varies by population)
- 500-1000 deletions 1-45 kb, overlapping ≈200 genes
- ≈150 insertions/deletions without frameshifting
- ≈200-250 frameshifting variants
- 10,000-12,000 of synonymous SNPs
- 8,000-11,000 non-synonymous SNPs in 4,000-5,000 genes
- 175-500 rare non-synonymous variants
- 1 new non-synonymous variant
- ≈100 premature stop codons

• 40-50 splice site disruption variants

• 250-300 genes with probable loss-of-function variants

• ≈25 genes estimated to be completely inactivated

Diagnostic methods of genetic variability

Detailed study of the DNA molecule, identification and classification of genetic differences between individuals of the same species are possible through molecular genetics techniques:

- amplification techniques - polymerase chain reaction (PCR) - allow the multiplication of any DNA fragment in an unlimited number of copies;

- hybridization techniques - fluorescent in situ hybridization (FISH) - allow the comparison of different DNA samples, having various sources, and the identification of certain base pairs;

- sequencing techniques - sequencing of the human genome;

- microarray technique - allows the identification of gene variants as well as the activity of genes at the cellular level (gene expression).

Genetic variants and their impact

A genetic variant is defined as a permanent change in the structure and functionality of a part of the genetic material. Variants are the source of diversity between individuals, but they are also the cause of monogenic diseases and genetic predispositions in multifactorial diseases. The consequence of a mutation depends on its functional effect:

• it can be neutral,

• it can lead to an improvement in function (diversity, evolution) or

• to an impairment of a function (pathogenic effect).

Ø In a living cell, DNA is constantly exposed to various harmful factors that can lead to the appearance of variants. These factors are:

• of exogenous origin (radiation and genotoxic agents in the environment) mainly,

• of endogenous origin (free radicals),

• replication errors and

• accidental recombinations.

Ø The cell has its own repair "apparatus", which corrects most anomalies, but not all.

Genetic variants can be determined by exposure to various mutagenic factors, either external or internal. The latter represent the most important source of mutations, producing in particular errors during the replication of the DNA molecule, respectively the impossibility or inefficiency of the mechanisms for repairing these lesions. Statistical data show that during the life of an individual, 1017 cell divisions occur: approximately 2 × 1014 divisions are necessary to generate approximately 1014 adult cells, and the other mitoses will ensure the renewal of certain types of cells (especially epithelial ones). At each division, it is necessary to incorporate 6 × 109 new nucleotides, and the process of DNA replication must be carried out with great accuracy for the exact copying of hereditary information. Maintaining this high degree of fidelity is very difficult to achieve, so it is estimated that during the approximately 1017 mitoses during life, in general, each gene can undergo approximately 108-1010 mutations. In many cases, a variant that appears in a somatic cell can be lethal for that cell, so the variant will not propagate to other cells. However, in some cases, the variant can cause an inappropriate continuation of cell division, thus determining a malignant process.

Genetic variants can be:

Ø spontaneous - represent changes in the genetic material that occur during life, under the action of mutagenic factors in nature;

Ø induced (artificial), which are produced by physical and chemical mutagenic factors administered for a limited period of time, either for their therapeutic effects (x-ray therapy or chemotherapy in cancer), or in the laboratory, for research purposes.

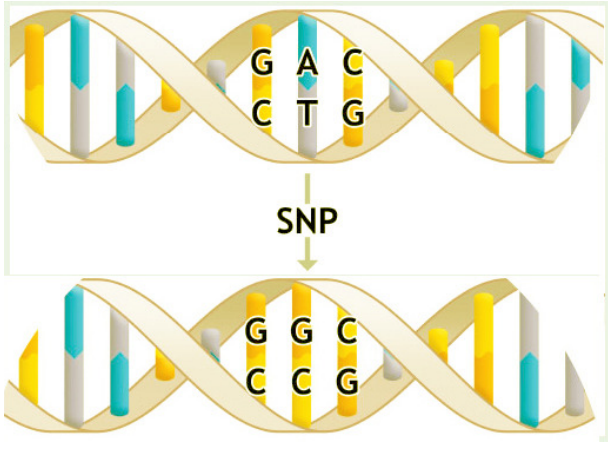
Genome polymorphism

The term "mutation" refers to any change in DNA sequence or changes that may affect genes and chromosomes. Before analyzing the harmful effects of variations in the genome, we must mention the importance of non-pathogenic variations in the genome that underlie diversity between individuals.

These non-pathogenic variations in the genome are called "polymorphisms". The notion of "polymorphism" is based on both non-pathogenic variations of the sequences, and on their frequency in the population (> 1%). Polymorphism is a mutation and can be present in a coding or non-coding region of a genome.

Several types of polymorphism have been described, eg:

- **SNPs** (single nucleotide polymorphisms): a polymorphism that occurs due to the replacement of a single nucleotide. These polymorphisms are very numerous and are distributed throughout the genome (approximately one SNP every 300 base pairs).



Example of a SNP

- **CNVs** (copy number variations): There is a variation in the number of copies of some large segments of the genome. Up to now, the variations have been identified in a proportion of about 15% for the human genome.
- **Polymorphic tandem repeats (VNTR-**variable number of tandem repeats): tandem repeats of varying lengths sequences (see repetitive DNA).

In general, variants can affect any part of the DNA molecule. Therefore, coding DNA and non-coding DNA are equally likely to undergo mutations. However, it is clear that the most important consequences on the phenotype are the variants of coding DNA.

Variants can be classified into three categories:

- genome variants
- chromosome variants
- gene variants

All three types of variant occur at appreciable frequencies and underline not only all genetic or heritable disease, but also many instances of cancer. In certain conditions, the variants can determine only “normal” variation (protein variation, genetic polymorphism).

Variants can occur in any cell, both in germline cells and in somatic cells. Only germline variants, however, can be perpetuated from one generation to the next and are thus the ones responsible for inherited disease. But this is not to say that somatic cell variants are not medically important. Somatic variant at the level of the genome, the chromosome, or the gene, which results in somatic mosaicism, is cause of phenotypic variation. Somatic variants in a number of genes can also give rise to a significant proportion of cancers.

The origin of variant

Genome and chromosome variants responsible for abnormalities of chromosome number involve the missegregation of a chromosome pair during germline cell division. The estimated rate of one missegregation event per 25 to 50 meiotic cell divisions derives from the observed incidence of chromosomally abnormal fetus and liveborn infants. Although these frequencies may seem high, genome and chromosome (structural) variants are rarely perpetuated because they are usually incompatible with survival or normal reproduction.

Gene variants include base-pair substitutions, insertions, deletions, duplication trinucleotide repeat sequences and RNA-splicing disorders. They can originate by either of two basic mechanisms: errors introduced during the normal process of DNA replication or base changes induced by mutagens.

The frequency of new variant

Most variants occur in somatic cells; depending on the nature of the variant, its location in the genome and the tissue involved, may or not lead to phenotypic variation.

A small number of the new variants occur in the germline cells, during any of the mitotic divisions that take place during spermatogenesis or oogenesis or during meiosis itself. These divisions that take place during spermatogenesis or oogenesis or during meiosis itself. These divisions differ between the sexes. In oogenesis, each haploid ovum has gone through an estimated 22 mitotic divisions and one meiotic division; these occur only in fetal life and cease by the time of birth. Spermatogenesis, on the other hand, involves a continuous series of cell divisions throughout life, resulting in a total of approximately 1 billion spermatogonia. These cells are the result of about 30 mitotic divisions from the embryo stage to the time of puberty and about 20-25 division cycles per year thereafter. Thus, each haploid sperm is the product of many hundreds of divisions, depending on age. Consequently, one might expect that the frequency of paternal new variants is age dependent and that certain kinds of new variant in genetic disease are more often of paternal, rather than maternal origin. In fact, this has been observed for some disorders, such as: neurofibromatosis, achondroplasia and hemophilia A. However, not all disorders appear to show this effect.

Because there are about 50 000 to 100 000 genes in the genome this suggests that at least 1 in 10 people is likely to have received a newly mutated gene from one or other parent.

The molecular bases of mutations

Many types of variant have been discovered with the application of molecular techniques and it is generally considered that variants cause genetic diseases. The variants in most genetic diseases are heterogeneous.

Types of variant in human genetic disease

1. Nucleotide substitutions (point variants)

Missense variants

Nonsense variants

RNA splicing variants

2. Deletions and insertions

Frameshift variants

Codon deletions and insertions

Gene deletions, duplications (often by unequal cross-over)

Insertion of repeated elements (interrupts coding sequence)

1. Nucleotide substitutions

A single nucleotide substitution (or point variant) in a DNA sequence can alter the code in a triplet of bases and cause the replacement of one amino acid by another in the gene product. Such variants are called missense variants because the alter the " sense" of the coding strand of the gene by specifying a different amino acid.

In many disorders, such as hemoglobinopathies, the vast majority of detected variants are missense variants. For example, sickle cell hemoglobin (HbS) is due to a change in only 1 of the 146 amino acids in beta globin, a substitution of valine for glutamic acid at the 6th position of the polypeptide. Other genetic disorders of hemoglobin (Hb) characterized by point variant are Hb C (due to a substitution of the same position of the

beta chain, the glutamic acid being replaced by lysine), HbM, Hb Kansas, Hb Hammersmith a. o.

Nucleotide changes that involve the substitution of one purine for the other or one pyrimidine for the other is called transitions. In contrast, the replacement of a purine for a pyrimidine (or vice versa) is called transversion. These variants cause one or the other type of substitution. A single base pair substitution with transition is well known in hemophilia A (an X-linked clotting disorder).

Another base substitutions can occur outside the coding sequences of a gene in untranslated region, e.g. hemophilia B (as hemophilia A, hemophilia B is an X-linked clotting disorder).

Chain termination variants (nonsense variant)

Normally, translation of mRNA ceases when a termination codon is reached. A variant that creates a termination codon can cause premature cessation of translation, whereas a variant that destroys a termination codon allows translation to continue until the next termination codon is reached. A variant that generates one of the three "stop" codons is called a nonsense variant.

Examples of nonsense variants are gene for neurofibromatosis (disorder of the nervous system), thalassemias (heterogeneous group of diseases of hemoglobin synthesis), hemoglobin Constant Spring (Hb C-S) and so forth.

RNA- splicing variants

The normal mechanism by which introns are excised from unprocessed RNA and exons spliced together to form a mature mRNA is

dependent on particular nucleotide sequences located at intron / exon (acceptor site) and exon/intron (donor site) boundaries. Variants that affect the required bases at either the splice donor or acceptor site interfere with normal RNA splicing at that site and often, abolish this process.

Examples of splice site variants are: some of molecular disorders as phenylketonuria, hemophilia B, lysosomal disorders, and enzyme defects, □-thalassemia, etc.

2. Deletions and insertions

The insertion or deletion of more base pairs can cause variants. Some deletions and insertions involve only a few nucleotides and can generally be detected only through molecular analysis involving nucleotide sequencing. In other cases, a substantial segment of a gene or an entire gene is deleted. Such variants can be detected using Southern blotting method; large insertions or duplications can be detected in the same manner. In rare instances, deletions are large enough to be visible at the cytogenetic level (on the chromosomes). These deletions are called microdeletion. Such deletions can be detected with high-resolution prometaphase banding. In many instances a microdeletion remove more than a single gene and causes a microdeletion syndrome.

In the case of deletions or insertions involving only a few base pairs, when the number of bases involved is not a multiple of three (is not an integral number of codons), such a variant in a coding sequence alters the reading frame of translation from the point of the variant on, which results in a different amino acid sequence of the encoded protein. These variants are called frameshift variants. An example of an insertion that causes a

frameshift is the most common variant in some of the lysosomal storage diseases.

Other small insertions or deletions do not cause a frameshift because the number of base pairs involved is a multiple of three. An example is a three base pair deletion observed in the mutant allele that causes cystic fibrosis (CF). This variant formed in nearly 70 percent of cystic fibrosis, causes synthesis of an abnormal gene product that is missing a single amino acid, the phenylalanine from position 508.

Large deletions and insertions have been described in numerous inherited disorders, being detected by Southern blotting (to be seen, methods of nucleic acid analysis). The observed frequency of such variants differs markedly among different genetic diseases; some disorders are characterized by a high frequency of deletions. For example, a total deletion is observed in more than 90 percent of cases of X-linked ichthyosis (genetic skin disorder) and in Duchenne muscular dystrophy (the same an X-linked disorder). Insertion of large amounts of DNA is less frequent cause of variant than is deletion. Large insertions have been described in sporadic cases of hemophilia A.

Expansion of trinucleotide repeats sequences

A novel mechanism of human gene variants that lead to hereditary disorders is the instability of certain trinucleotide repeats and their expansion in affected genes. A growing number of disorders, the majority of which involve the neuromuscular tissues have been found to be due to expansion of trinucleotide repeats (e.g., fragile X syndrome and Huntington disease). The trinucleotide repeat is polymorphic in the human populations. Rarely, however, the numbers of trinucleotide repeats are within a high-risk

category that is called prevariant. The normal polymorphic alleles contain between 10 and 50 triplets, the prevariant between 50 and 200, and the full variant of more than 200 triplets. The prevariant exhibits a high probability of further expansion (instability) into the disease-related alleles (full variant). Expansion of prevariants to full variants only occurs during female meiotic transmission.

Deletions and duplications caused by recombination

A frequent cause of variant, involves deletion or duplication mediated by recombination during the crossing-over. When two homologues pair of chromosome misalign during meiosis, the recombination will occur between mispaired genes and will lead to gene deletion or duplication. This mechanism known as *unequal crossing over* is responsible for gene deletions in several disorders.

One of the best examples of homologous unequal recombination is the case of Lepore gene (Hb Lepore), in which there is a hybrid gene between the delta and beta globin genes on chromosome 11. At the same manner, it is known a syndrome (gonadal dysgenesis) caused by an unequal mispairing of X and Y homologous sequences.

Mutagenic factors

A wide variety of physical agents as ionizing radiation have mutagenic effect. The corpuscular radiations, neutrons and protons, cause ionizations and are effective in producing variants. All the ionizing radiation can induce genetic change.

It is well known that even low dose exposures to radiation (natural exposures or radiographic exposures) contribute to the total population load

of variants, causing either somatic variants (cell cancer), or variants in germline cells and in fetus genome. Studies of leukemogenesis stemming from prenatal exposure to X-irradiation suggest the risk for an exposed fetus to develop leukemia. In the same way, it is reasonable to expect that radiation doses to produce risk for malformations for human offspring. Gonadal irradiation (germ line cell) at any time during life before conception may have some risk of a variant in the offspring. Besides the ability of ionizing radiation to induce gene variants, they are also effective in the production of structural chromosome changes (deletions, inversions, translocations, etc.).

For years has been discovered that several substances (mutagenic chemicals) can have mutagenic activity. They can produce pairing errors in the DNA molecules, alterations of the base distortions in the DNA and disorders of the mechanism of replication repair.

Several studies of RNA viruses and DNA viruses showed that they are involved in the development of cancer (contribute to tumorigenesis).

RNA tumor viruses can be divided into two classes-acute and chronic transforming viruses. The differences in the properties of the two classes of these mutagenic viruses can be traced to differences in their genetic content.

The chronic transforming viruses transform cells by random integration of DNA copy of the virus, termed the *provirus,* in to the host cell genome. The provirus can alter the genes in the region of the host chromosome where it integrates. If a proto-oncogene is contained in the region, provirus integration can alter the structure and/or expression of the proto-oncogene and thus contribute to tumorigenesis (proto-oncogenes are a cellular counterparts being an important regulators of many aspects of cell

growth). Several of the proto-oncogenes that are altered by chronic transforming viruses will be affected by somatic variants and contribute to tumorigenesis by inactivating or inhibiting the activities of tumor suppressor proteins.

Examples of human diseases resulting from different types of mutations:

Mutation	Disease
Single base-pair substitution: Missense mutation Nonsense mutation RNA splicing mutation Regulatory gene mutation	 Sickle-cell anemia, Hb-pathies C, M B-thalassemia, Hb Constant-Spring, Neurofibromatosis B-thalassemia, phenylketonuria, hemophilia B, lysososmal disorders hemophilia B
Frameshift mutations	Duchenne muscular dystrophy, beta-thalassemia, lysososmal storage diseases
Deletions: Small deletions Large deletion	 Cystic fibrosis (deletion of 1 codon) Duchenne muscular dystrophy, ichtyosis, hypercholesterolemia
Duplication	hemophilia A
Expansion of trinucleotides	X-fragile syndrome, Huntington disease

Mutation nomenclature examples

In order to describe mutational data, an official international nomenclature has been established (http://www.hgvs.org/mutnomen). The general rule consists in the description of the sequence variation location in relation to the gene coding sequence and the change that is present.

Example 1: c. 123A>G describes a substitution on cDNA, where A in 123 is replaced by G.

Example 2: a missense mutation in a gene: Exon 9: c.895G> T (p.Gly299Trp). It describes a substitution of guanine with a thymine in exon 9 of the gene, which leads to a replacement of tryptophan with the amino acid glycine at position 299 of the protein sequence.

References

1 Thompson & Thompson Genetics and Genomics in Medicine. 9th Edition (2023) Edited by Ronald Cohn, Stephen Scherer and Ada Hamosh .Publisher: Elsevier. ISBN: ISBN 978-0323547628

2 Emery's Elements of Medical Genetics and Genomics. (2021) Peter D Turnpenny. Edition. 16th, Elsevier Health Sciences, ISBN-13. 978-0702079665

3 Edward S. Tobias, Michael Connor, Malcolm Ferguson-Smith. Essential Medical Genetics, 6th Edition (2011). Ed. Wiley-Blackwell. ISBN: 978-1-405-16974-5

4 Jorde, L. B., Carey, J. C., & Bamshad, M. J. Medical Genetics and Genomics, 6th Edition (2020). Elsevier. ISBN: 9780323597371

5 D. Bourn. Diagnostic Genetic Testing. (2022). Ed. Springer Nature Switzerland AG,

6 Clinical Genetics and Genomics at a Glance. 1st Edition (2023), Edited by Lakhani N, Kulkarni K, Barwell J, Vasudevan P, Dorkins H. Publisher: Wiley-Blackwell. ISBN: ISBN 978-1-119-24095-2

7 Tom Strachan. Human Molecular Genetics: 5th Edition (2019); Publisher. Garland Science. ISBN: ISBN 978-0815345893

8 Andreescu NI, Trifa AP, Chiriță – Emandi A, Stoicănescu DL, Gug CR, Farcaș SS, Popa CA, Mihăilescu A, Simina IE, Gug MC. Curs de Genetică, Editura „Victor Babeș", 2025, ISBN 978-606-786-490-8.

9 D Stoicănescu. Medical Genetics. Ed. Eurostampa, 2016, ISBN 978-606-32-0326-8

10 https://www.who.int/health-topics/genomics#tab=tab_1

11 Hu Z, Suo Z, Liu W, Zhao B, Xing F, Zhang Y, Feng L. DNA conformational polymorphism for biosensing applications. Biosens Bioelectron. 2019 Apr 15;131:237-249. doi: 10.1016/j.bios.2019.02.019.

12 Zyner, K.G., Simeone, A., Flynn, S.M. et al. G-quadruplex DNA structures in human stem cells and differentiation. Nat Commun 13, 142 (2022). https://doi.org/10.1038/s41467-021-27719-1

13 Tao S, Hou Y, Diao L, Hu Y, Xu W, Xie S, Xiao Z. Long noncoding RNA study: Genome-wide approaches. Genes Dis. 2022 Nov 29;10(6):2491-2510. doi: 10.1016/j.gendis.2022.10.024.

14 Mattick JS, Amaral PP, Carninci P, Carpenter S, Chang HY, Chen LL, Chen R, Dean C, Dinger ME, Fitzgerald KA, Gingeras TR, Guttman M, Hirose T, Huarte M, Johnson R, Kanduri C, Kapranov P, Lawrence JB, Lee JT, Mendell JT, Mercer TR, Moore KJ, Nakagawa S, Rinn JL, Spector DL, Ulitsky I, Wan Y, Wilusz JE, Wu M. Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat Rev Mol Cell Biol. 2023 Jun;24(6):430-447. doi: 10.1038/s41580-022-00566-8.

15 Lodish H, Berk A, Kaiser C, Krieger M, Bretscher A, Ploegh H, Martin K, Yaffe M, Amon A. Molecular Cell Biology, Ninth Edition, 2021, Publisher Freeman WH. 978-1319208523

16 Chen LL, Kim VN. Small and long non-coding RNAs: Past, present, and future. Cell. 2024 Nov 14;187(23):6451-6485. doi: 10.1016/j.cell.2024.10.024.

17 Hebenstreit D, Karmakar P. Transcriptional bursting: from fundamentals to novel insights. Biochem Soc Trans. 2024 Aug 28;52(4):1695-1702. doi: 10.1042/BST20231286.

18 Huang, Y., Zhang, P., Wang, H., Chen, Y., Liu, T., & Luo, X. (2024). Genetic Code Expansion: Recent Developments and Emerging Applications. Chemical Reviews, 124(21), 11962–12005. https://doi.org/10.1021/acs.chemrev.4c00275

19 Li S, Vemuri C, Chen C. DNA topology: A central dynamic coordinator in chromatin regulation. Curr Opin Struct Biol. 2024 Aug;87:102868. doi: 10.1016/j.sbi.2024.102868.

20 Gromiha MM, Harini K. Protein-nucleic acid complexes: Docking and binding affinity. Curr Opin Struct Biol. 2025 Feb;90:102955. doi: 10.1016/j.sbi.2024.102955.

21 Kosztolányi G, Cassiman J-J. The medical geneticist as expert in the transgenerational and developmental aspects of diseases. European Journal of Human Genetics, 2010; 18:1075–1076

22 Jiménez-Morales S, Pérez-Amado CJ, Langley E, Hidalgo-Miranda A. Overview of mitochondrial germline variants and mutations in human disease: Focus on breast cancer (Review). Int J Oncol. 2018 Sep;53(3):923-936. doi: 10.3892/ijo.2018.4468.

23 https://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518/

24 https://www.genome.gov/Funded-Programs-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements

25 Covic M, Gorduza EV, Stefanescu D, Sandovici I: Genetica si genomica medicala. Ed a IV-a. 2024; Ed. Polirom, Iasi. ISBN, 978-973-469-793-9.

26 Eckersley-Maslin MA, Spector DL. Random monoallelic expression: regulating gene expression one allele at a time. Trends Genet. 2014 Jun;30(6):237-44. doi: 10.1016/j.tig.2014.03.003.

27 Lyu H, Li Y, Chen X, Liu Y, Liu E, Cheng X. Topologically associating domains of chromatin on single-cell Hi-C data: a survey of bioinformatic tools and applications in the light of artificial intelligence. Front Genet. 2025 Jul 1;16:1602234. doi: 10.3389/fgene.2025.

28 Fogg JM, Judge AK, Stricker E, Chan HL, Zechiedrich L. Supercoiling and looping promote DNA base accessibility and coordination among distant sites. Nat Commun. 2021 Sep 28;12(1):5683. doi: 10.1038/s41467-021-25936-2.

29 Haws SA, Simandi Z, Barnett RJ, Phillips-Cremins JE. 3D genome, on repeat: Higher-order folding principles of the heterochromatinized repetitive genome. Cell. 2022 Jul 21;185(15):2690-2707. doi: 10.1016/j.cell.2022.06.052.

30 Gilbert N. Marenduzzo D. Topological epigenetics: The biophysics of DNA supercoiling and its relation to transcription and genome instability. (2025) Current Opinion in Cell Biology, 92, February, 102448. https://doi.org/10.1016/j.ceb.2024.102448

31 Bickmore WA. The Spatial Organization of the Human Genome. Annu. Rev. Genomics Hum. Genet. 2013;14:67–84

32 Cardozo Gizzi AM. A Shift in Paradigms: Spatial Genomics Approaches to Reveal Single-Cell Principles of Genome Organization. Front Genet. 2021 Nov 19;12:780822. doi: 10.3389/fgene.2021.780822.

33 Roth GV, Gengaro IR, Qi LS. Precision epigenetic editing: Technological advances, enduring challenges, and therapeutic applications. Cell Chem Biol. 2024 Aug 6:S2451-9456(24)00309-X. doi: 10.1016/j.chembiol.2024.07.007.

34 Dai W, Qiao X, Fang Y, Guo R, Bai P, Liu S, Li T, Jiang Y, Wei S, Na Z, Xiao X, Li D. Epigenetics-targeted drugs: current paradigms and future challenges. Signal Transduct Target Ther. 2024 Nov 26;9(1):332. doi: 10.1038/s41392-024-02039-0.

35 Eckersley-Maslin MA, Spector DL. Random monoallelic expression: regulating gene expression one allele at a time. Trends Genet. 2014 Jun;30(6):237-44. doi: 10.1016/j.tig.2014.03.003.

36 Duttke S.H.C., Lacadie S.A., Ibrahim M.M., Glass C.K., et al. Human Promoters Are Intrinsically Directional. Molecular Cell, 2015, 57(4):674–684

37 Tena JJ, Santos-Pereira JM. Topologically Associating Domains and Regulatory Landscapes in Development, Evolution and Disease. Front Cell Dev Biol. 2021 Jul 6;9:702787. doi: 10.3389/fcell.2021.702787.

38 Capp JP. Interplay between genetic, epigenetic, and gene expression variability: Considering complexity in evolvability. Evol Appl. 2021 Feb 19;14(4):893-901. doi: 10.1111/eva.13204.

39 Shaye D, Liu CC, Tollefson TT. Cleft Lip and Palate: An Evidence-Based Review. Facial Plast Surg Clin North Am. 2015; 23(3):357-72

40 Leija-Salazar M, Piette C, Proukakis C. Review: Somatic mutations in neurodegeneration. Neuropathol Appl Neurobiol. 2018 Apr;44(3):267-285. doi: 10.1111/nan.12465.

41 http://www.csun.edu/~cmalone/pdf360/Ch15-2repairtanspose.pdf

42 http://www.genecards.org

43 https://www.orpha.net/

44 https://hgvs-nomenclature.org/stable/

45 Kent DG, Green AR. Order Matters: The Order of Somatic Mutations Influences Cancer Evolution. Cold Spring Harb Perspect Med. 2017 Apr 3;7(4):a027060. doi: 10.1101/cshperspect.a027060.

46 Ingle RG, M Elossaily G, Ansari MN, Makhijani S. Unlocking the potential: advancements and applications of gene therapy in severe disorders. Ann Med. 2025 Dec;57(1):2516697. doi: 10.1080/07853890.2025.2516697.

47 Maeder ML, Gersbach CA. Genome-editing Technologies for Gene and Cell Therapy. Mol Ther. 2016 Mar;24(3):430-46. doi: 10.1038/mt.2016.10.