1

Genomics and Epigenetics

COPYRIGHTED MATERIAL

As early as the second half of the 19th Century, the work of Gregor Mendel followed by that of Hugo de Vries, Carl Correns and Erich von Tscherma led to the first postulate concerning the heredity of specific characteristics: due to the presence of a certain type of "particles" present within each organism and transmitted from one generation to another.

It was only at the beginning of the 20th Century that the term "gene" was introduced to designate the fundamental entity at the origin of heredity. In 1911, Thomas Hunt Morgan succeeded in showing that chromosomes, identified in the 1880s and 1890s, carry genes.

1.1. DNA, RNA and genetic code

Deoxyribonucleic acid (DNA) was isolated in the 1870s. But it was not until the 1940s that Oswald Avery and his team highlighted its role as a carrier of genetic information, hitherto generally attributed to proteins. Research was then increasingly oriented towards the question of its composition. In 1950, Erwin Chargaff discovered, by analyzing the DNA of different organisms, that the quantities of adenine (A) and thymine (T) were equivalent, as were those of guanine (G) and cytosine (C)[1]. These proportions made sense a few years later, when Rosalind Franklin was able to obtain X-ray diffraction images of DNA. It is from these clichés that in 1953, Francis Crick and James Watson created the famous double

1 We speak of nitrogenous bases to designate A, T, G and C.

helix model in which the nitrogenous bases are oriented towards the inside of the molecule and where the A bases of a single strand are associated with T, and the G with C in what we now call base pairs.

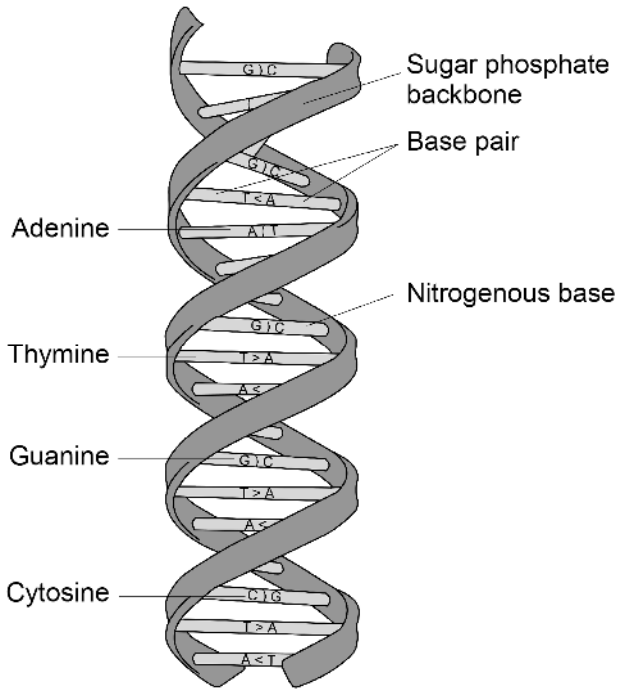

Figure 1.1. *DNA double helix (© MesserWoland, DNA structure and bases PL.svg, Wikimedia commons)*[2]

In the 1960s, the principles of the genetic code, correspondence between nucleotide triplets (codons) and amino acids (protein components), and of the genetic program were developed [KLU 12]. The basics of DNA functioning were elucidated. François Jacob and Jacques Monod succeeded in describing the phenomenon of transcription of a gene, and therefore of DNA, into a messenger ribonucleic acid (mRNA), a molecule that moves

2 License link: https://creativecommons.org/licenses/by-sa/3.0/legalcode.fr. The original work has been translated from Polish into French.

from the cell nucleus into the cytosol[3] before being translated into a protein. Each cell is thus genetically programmed to synthesize proteins with specific functions[4].

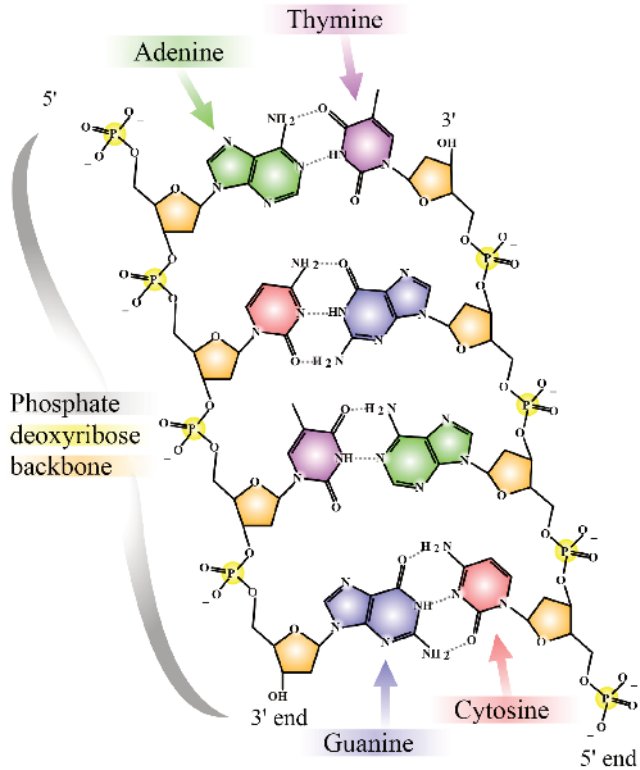

Figure 1.2. *The nitrogenous bases (A, G, C, T) are linked together by a phosphate deoxyribose backbone to form a strand of DNA. The double strand is formed by the hydrogen bonds between corresponding nitrogenous bases: A with T and C with G (© Madprime, Wikimedia commons)[5]. For a color version of the figures in this book see www.iste.co.uk/barbet/cancers.zip*

3 In this book focusing on cancer and cancer therapy, the cells of interest are eukaryotic cells that possess a nucleus containing DNA. The cytosol is the liquid present inside a cell, in which the different constituent structures of the cell (organelles) are bathed. The cytosol and organelles of the cell are called cytoplasm.

4 See video "What is genomics" on https://www.ontariogenomics.ca/about-us/what-is-genomics/.

5 License link: https://creativecommons.org/licenses/by-sa/3.0/deed.en.

NOTE.– Unlike nucleic acids, which consist of chains of ribose or deoxyribose groups and phosphate groups, proteins consist of chains of amino acids linked together by a characteristic bond, called a peptide bond, linking the carboxylate group of one to the amine group of the next. A protein is therefore a large polypeptide. Many smaller compounds, called peptides, also play various roles in biology.

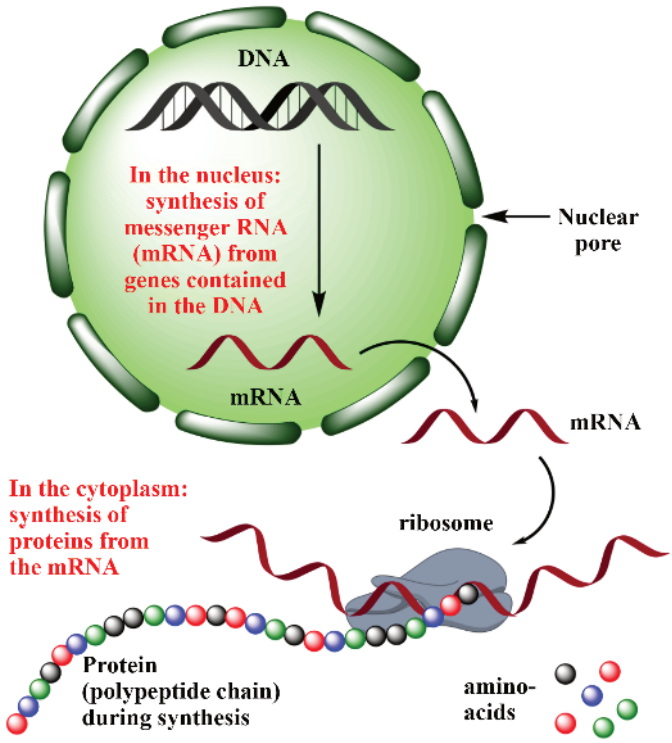

Figure 1.3. *Transcription and translation (© E. Jaspard, University of Angers)*

NOTE.– mRNA is a single-stranded copy of DNA that allows the transport of genetic information. In mRNA, as in the many other RNAs involved in cell life, the T-base (thymine) is replaced by the U-base (uracil).

The protein translation code is based on triplets of nucleotides. For example: the AUG triplet codes for the amino acid methionine, the GCU, GCC, GCA or GCG triplets, all code for the same amino acid, alanine; the

UAG, UAA or UGA triplets mark the end of the coding part of an RNA; they are "stop" codons.

These RNAs are most often single-stranded, but double-stranded RNAs also exist in living organisms, particularly in certain viruses, where they act as carriers of genetic information.

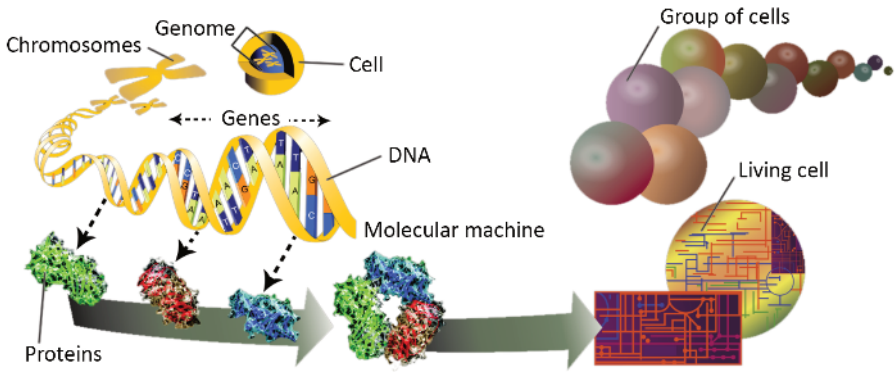

Figure 1.4. *Chromosomes, components of the genome present in cells, contain DNA. The coding genes of DNA are converted, via the transcription and translation processes, into proteins that perform a multitude of functions in the body (© PD-USGov-DOE)*

NOTE.– As early as 1838, the cell theory, in which "the cell is the structural and functional unit of plants and animals", was introduced by Matthias Jakob Schleiden. Thereafter, the nucleus is considered as the main organelle of the eukaryotic cell. In 1858, it was shown that every cell comes from a pre-existing cell. And the genetic continuity of the nucleus was established in 1878. DNA was isolated as early as 1869 without, of course, its structure being identified yet.

1.2. Sequencing and genomics

The very first sequence, a sequence of amino acids, of a polypeptide (small protein), insulin, was obtained in the 1950s by Frederick Sanger and his colleagues at Cambridge University. It was quickly followed by many

other examples. Since these sequences are necessarily "coded" in DNA, a growing interest developed in determining the sequence of nucleotides, the basic elements of genetic material, within the different DNA or RNA fragments. In 1972, Walter Fiers carried out the first sequencing of a gene. It is a gene of the MS2 bacteriophage[6] whose genetic material consists of RNA. In 1977, the genome of the same virus was sequenced in its entirety. A year later, the first sequencing of a DNA genome, that of the bacteriophage ΦX174, was performed through the work of Frederick Sanger and Walter Gilbert, carried out independently of each other. However, ΦX174 DNA is a single-stranded DNA and not a double helix. The Sanger method is based on selective enzymatic synthesis. It uses an enzyme complex called DNA polymerase which has the ability to copy DNA by successively adding nucleotides, in a complementary manner (A with T, C with G), to an already present DNA fragment (the primer). Specific nucleotides, called deoxynucleotides and fluorescently labeled, are added in small quantities. They interrupt the synthesis and produce fluorescent fragments that are separated by electrophoresis according to their size. The DNA sequence can then be read by a sequencer [HEA 16].

These discoveries herald the beginning of the era of genomics, a discipline that tends to study the functioning of an organism from its genome and no longer just on the scale of a single gene. However, it was not until 1995 that the first genome sequence of a living organism, the *Haemophilus influenzae* bacterium, was obtained. A year later, it was the turn of the first multicellular eukaryotic organism, the *Caenorhabditis elegans* worm. And in 2000, those of the fruit fly (*Drosophila melanogaster*) and of the arabette (*Arabidopsis thaliana*) were sequenced.

NOTE.– Viruses, whose genomes were sequenced earlier, are not strictly speaking "living beings", because they do not feed and cannot reproduce autonomously, but this is still a matter of debate. Bacteria are living beings in their own right, but they do not have a nucleus: they are prokaryotes. Yeasts and cells that make up multicellular organisms have a nucleus that contains the chromosomes. They are eukaryotes.

6 Bacteriophages are viruses that infect only bacteria. Safe for humans, they are easy to produce in large quantities.

For Piotr Slonimski, co-founder of the *Centre de génétique moléculaire du CNRS* in Gif-sur-Yvette:

> "If we consider it [genomics] as the understanding of the meaning of a genome, we will remember that the research was first done using experimental genetic techniques, that is, by identifying observable, phenotypic traits for which we knew a wild type and a mutant type. Conversely, where functional genetics started from a phenotype and went back to the genotype - to a disruption of the phenotype corresponds that of a gene - genomics looks in the genome for the elements that explain the difference between the wild phenotype and that of the mutant" [PIC 02].

The idea that diseases can be caused by genetic abnormalities appears early on. As a result, the possibility of replacing "defective" genes becomes possible. The American geneticist Edward Tatum mentioned this possibility as early as 1958:

> "With a more complete understanding of the functioning and regulation of gene activity in development and differentiation, these processes may be more efficiently controlled and regulated, not only to avoid structural or metabolic errors in the developing organism, but also to produce better organisms" [TAT 59].

We can already quite clearly see in this passage the ethical problems that the scientific community will soon be confronted with. From the 1970s onwards, controversies and media debates soon appeared, as the risk of a new eugenics came to arouse fear among the public. But Tatum does not only state possible goals, he also pursues them:

> "This might proceed in stages from the *in vitro* biosynthesis of better and more efficient enzymes, to the biosynthesis of the corresponding nucleic acid molecules, and to the introduction of these molecules into the genome of organisms, […]" [TAT 59].

These projections became reality less than two decades later.

Two other essential discoveries will allow the development of sequencing and genomics from the 1960s onwards. Enzymes capable of cutting the DNA of bacteriophage λ were found in the bacterium *Escherichia coli* and called restriction enzymes. Later, it was demonstrated that the cleavage of simian virus 40 (SV40) DNA by restriction enzymes affords specific fragments that can be separated by polyacrylamide gel electrophoresis. These discoveries are fundamental because they allow DNA to be manipulated and lead to the development of recombinant DNA technology.

In 1986, Kary Mullis added the second pillar of genomics with the polymerase chain reaction (PCR). This *in vitro* DNA amplification method is based on the Sanger sequencing technique. But the innovative aspect lies in a chain reaction induced by a repetition of temperature transition cycles. The first step is to mix double helix DNA, nucleotides, DNA polymerase and primers. At each cycle, the first heating (at 95°C) allows the DNA to be denatured (separate the two strands). It is therefore necessary to use a polymerase resistant to these high temperatures, such as those found in thermophilic bacteria. The temperature is then lowered (to 40–50°C) to induce the binding of the primers and then increased to allow for the synthesis of complementary strands by DNA polymerase intervention. This cycle is repeated many times. At the end of each cycle, the number of strands is doubled. For use for sequencing purposes, the introduction of dideoxynucleotides is necessary, and the fragments are separated by electrophoresis. Manipulations are considerably facilitated, and DNA sequencing becomes faster [ELF 07]. PCR was initially developed by Mullis to determine the nature of a nucleotide at a specific location in the DNA molecule. It is therefore particularly useful for diagnostic purposes, as genetic diseases are often caused by the substitution of a single nucleotide at a specific position on a gene. But the scope of this technique extends beyond medical borders. This is a major step forward in other areas, such as forensics, agri-food and paleogenetics.

The genetic origin of antibody diversity was demonstrated in the late 1970s, through the work of Susumu Tonegawa and Nobuchimi Hozumi showing the existence of recombination of immunoglobulin genes in somatic cells (all cells forming an organism, with the exception of germ cells responsible for reproduction).

Tonegawa first showed that the light chain of antibodies is encoded by several gene segments. These segments have the particularity of appearing in several copies in the same chromosome. It is thus shown that DNA undergoes recombination (it is cut and then reconnected) during cell differentiation. Subsequent research made it possible to identify the different gene segments encoding the variable and constant regions of the light and heavy antibody chains. Finally, we know that, in the germline genome, the variable region of the heavy chain is encoded by three types of gene segments (called V for variable, D for diversity and J for junction); that of the light chain by two (V and J). But, in the genome of B lymphocytes, these segments are united in the same region of a chromosome, following a recombination of DNA, called V(D)J, catalyzed by a set of enzymes that eliminates some of the genes. Since this recombination is random, a new assembly of the different gene segments V, D or J is possibly generated for each cell. Other random events, such as the removal or addition of nucleotides at the junction between genes, occur during the process. All this leads to the high diversity of antibodies.

The issue of T-cell antigen receptors (TCR, see Chapter 3) was initially more problematic: unlike antibodies, these receptors are not secreted into the blood. However, Tak Mak and Mark Davis were able to identify the coding genes as early as 1984 and extended the process explaining the genetic origin of antibody diversity to that of TCR.

Box 1.1. *Origin of antibody and T-cell receptor diversity [HOZ 76, TOY 84]*

In the 1970s, particularly by demonstrating the genetic origin of antibody diversity, it was also understood that the genome could be modified through natural mechanisms, that it was not immutable.

In 1986, the “Human Genome Project” (HGP), a project to sequence the entire human genome, was launched [LAT 10]. The stated objective was to be able to understand, prevent and treat genetic diseases and cancers. The ambition is considerable compared to what has been achieved upstream. The human genome is much larger than those previously sequenced. Huge investments were made to improve existing technologies. The end of raw

sequencing was announced in 2000. It will take a few years for the project to be really finalized. The number of human genes is approximately 25,000 for more than 3 billion base pairs. In comparison, Drosophila has about 13,000 genes and 165 million base pairs. Since proteins generally contain only a few hundred to a few thousand amino acids, each encoded by a triplet of nucleotides, the majority of DNA base pairs are not in coding parts of the genome.

Regardless of the influence of PCR, sequencing techniques continue to evolve. In the early 2000s, costs and times were still particularly high. To meet these constraints, high throughput sequencing appeared in 2005: methods such as pyrosequencing or real-time sequencing of a single molecule were developed and led to impressive improvements. The first complete sequencing of a human genome, completed in 2003, lasted 10 years and cost $3 billion US. Today, the price is less than $1000 US, and it takes a few days to get the same result.

What is related to genome sequencing is commonly reported as structural genomics. During the 1990s, more and more genomes were sequenced. A second branch then emerged: the so-called functional genomics. As the sequence of nucleotides within the DNA can be identified, it is logical to determine the associated biological functions. The goal is not, as is the case in genetics, to associate a function with a gene. The objective is the correlation between the genotype, as a whole, and the phenotype. In an oncological context, the aim is to associate cancer risks with the genome and its expression. Several approaches are possible.

1.3. Transcriptome and proteome

Transcriptomics is the analysis of the transcriptome, i.e. the analysis of all the mRNAs resulting from the transcription of the genome at a given time. A single genome can give rise to several transcriptomes. Not all genes are expressed by the different cells in the same organism, and expression levels are variable. In addition, the transcription of a gene varies over the life of the cell. The so-called serial analysis of gene expression (SAGE) methods, in 1995, and the “Microarrays”, in 1997, were developed and used

to quantify and analyze transcribed mRNAs using retrotranscription followed by amplification of complementary DNA by PCR[7] [BAL 11].

The organisms studied were, in the first instance, yeasts. More recently, another technology, called RNA-seq, has been developed that uses a high throughput sequencing method. One of the main objectives is to identify the levels of expression of the different genes and, by incidence, the networks of co-expression of genes. This concept, introduced in 1999, aims to determine which genes are associated, from a transcriptional and functional point of view.

In oncology, transcriptomics leads to a finer understanding of carcinogenesis, a process in which a normal cell is transformed into a tumor cell so that the issue can be addressed at the level of each patient or even each lesion.

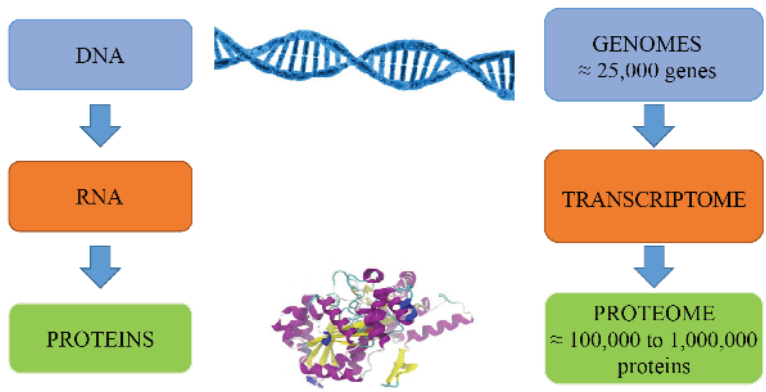

Figure 1.5. *Genome/transcriptome/proteome relationship*

The second approach is proteomics, i.e. the study of all the proteins of a cell or tissue at a given time (we speak of proteome) [INS 13]. As with the transcriptome, the proteome necessarily varies over time. Different proteomes can be derived from a single genome, and a gene can express different proteins depending on the cellular conditions to which it is

7 Howard Temin and Satoshi Mizutani on one side, David Baltimore on the other, discover the reverse transcriptase in the Rous sarcoma virus. This enzyme makes it possible to do the opposite of gene transcription: to synthesize DNA from RNA. This enzyme, combined with PCR in what is called RT-PCR, plays an essential role in cloning genes and quantifying RNA.

associated. The first to be identified appeared at the end of the 1990s and relate, as for transcriptomes, to yeasts. A technique called two-hybrid screening was used to detect protein–protein interactions. In the early 2000s, new labeling and purification methods were developed. The use of mass spectroscopy, which underwent a major evolution in the 1990s, has led to improved sample analysis.

The objective of proteomics is to obtain information about location, ligand binding sites, characteristics of proteins that interact with each other, etc. It becomes possible to compare data between healthy and diseased individuals or to identify the signaling pathways involved in the development of diseases. In addition, by monitoring the evolution of a group of proteins over time, particularly in the case of an abnormality, an expression profile can be developed to determine the protein mechanisms at the origin of the pathology [LEH 07]. In 2011, like the HGP, the Human Protein Project (HPP) was launched to identify all proteins encoded by genes in humans.

As early as 1916, Peyton Rous showed that cells can undergo malignant transformations after having been in contact with a virus. Later, the term oncogene appears. Until the early 1970s, it referred to the specific genetic material of the viruses that causes cell transformations. In 1972, researchers hypothesized that viruses or carcinogens (chemicals, radiations, etc.) can induce genetic mutations in normal cells, altering the functions of certain genes, called proto-oncogenes, gradually transforming them into cancer-inducing oncogenes. Finally, two possible processes were highlighted to explain the appearance of cancer. The first involves a retrovirus: having an RNA genome, the latter is retrotranscribed into DNA and can thus be integrated into the chromosomes of the normal cell. This induces a malignant transformation at the origin of the cancer. The second possible process involves carcinogens or spontaneous mutations, which alter the proto-oncogenes in the cell. This process is exacerbated by genetic predispositions, such as mutations in the ATM protein (ataxia telangiectasia mutated, a protein that repairs double-stranded DNA breaks), that cause ataxia telangiectasia syndrome and significantly increase the risk of lymphoma or leukemia. Proto-oncogenes are mostly genes regulating cell division, which explains why their modifications lead to a disorderly proliferation of cells.

Box 1.2. *Origin of tumors: oncogenes and proto-oncogenes [CHI 08]*

The technical and scientific advances we have mentioned are of paramount importance in oncology. Genomic, transcriptomic and proteomic analyses raise increasing interest from specialists. Sequencing an individual's genome is now possible. It makes it possible to characterize cancers based on genomic criteria, to identify mutations associated with the appearance of tumors, in order to develop targeted and adapted therapeutic strategies.

In 2008, the International Cancer Genome Consortium (ICGC), preceded by the Cancer Genome Atlas (TCGA) project, was launched. The objective was to study more than 25,000 tumor genomes and transcriptomes to produce a repertoire of genomic anomalies at the origin of the most common cancers in humans [LAT 10]. But another type of mapping also concentrates the attention of researchers, that of the epigenome.

1.4. Epigenetics[8], the missing link

For just over two decades, in addition to strictly genetic and genomic aspects, the influence of other factors has been increasingly highlighted and is now emerging as a fundamental criterion, not only for the understanding of oncological processes but also, more generally, for the understanding of species evolution in general. This is epigenetics [INS 15].

Michel Morange, professor of biology at the ENS, talks about a "concept that partially denies the 'fatality' of genes": other horizons are opening up, genetic mutations are no longer the only phenomena that can explain phenotypic changes from one generation to another. There are chemical modifications that do not alter the DNA sequence, the sequence of nucleotides, but only act during transcription, i.e. the conversion of DNA into RNA: it is the DNA reading process that is altered. Through epigenetic mechanisms, the "functioning" of genes is modified without changing their initial structure. Thus, the fact that a gene initially expressed may become inactive is partly explained in terms of epigenetics. And, although they are reversible and not hardcoded in DNA, epigenetic phenomena can be transmitted to the next generation. They can also be induced by environmental signals, such as smoking, stress or diet [DRO 19].

8 See "What is Epigenomics", available at https://www.youtube.com/watch?v=zcJPXISDxkM.

In fact, the discovery of the first epigenetic factors dates back to the late 1940s, through the work of American biologist Pamela Lewis, even before the double helix structure of DNA was elucidated. It highlighted the existence of protein complexes called polycomb and the genes associated with them. But it was only in the late 1970s, under the leadership of geneticist Edward Lewis, that their role as inhibitors of the expression of certain genes was demonstrated. In fact, in the case of polycomb, the genes concerned are first and foremost related to a regulatory function of the organizational plan of living organisms, of the arrangement of organs in relation to each other. Other proteins, trithorax, were quickly revealed as having an opposite effect to polycomb. However, it was not until the 1990s and 2000s, with the evolution of high throughput sequencing methods, in particular, that the mechanisms of action really began to be defined [FLE 18]. As Andrew Saurin, a researcher at the *Institut de biologie du développement de Marseille* (IBDM), explained:

> "By interfering in the field of genetics, biochemistry will make it possible to characterize these protein complexes at the molecular level and thus understand that they are involved in the epigenetic modifications of histone tails, those proteins around which DNA is wrapped to be compacted inside the cell nucleus" [FLE 18].

Histones, which are involved in transcription, can be seen as the bases for winding DNA. The epigenetic modifications they undergo (methylation, acetylation or others) have the effect of promoting or hindering this winding, thus making certain parts of the DNA accessible or inaccessible, certain genes active or inactive (their transcription), respectively. A loop composed of a part of DNA and a histone is called a nucleosome. Chromatin is the whole structure combining DNA and histones[9].

NOTE.– Although there are no real standards, histone methylation, as opposed to acetylation, often tends to promote DNA wrapping and, consequently, to inactivate certain genes.

9 Although the true role of chromatin and histones was only discovered in the 20th Century, their identification dates back to the end of the 19th Century.

Like histone modification, chromatin can also be affected and induce a similar process (activation or inactivation of genes) through a DNA methylation phenomenon that takes place at certain DNA bases. To be more precise, according to current knowledge, it is mainly cytosine that becomes methylated to 5-methyl-cytosine in the so-called promoter sequences[10]. Like histones, DNA methylation often results in the inactivation of genes: it can silence genes. This is, in fact, one of the most fundamental phenomena in the regulation of gene expression [KHO 18].

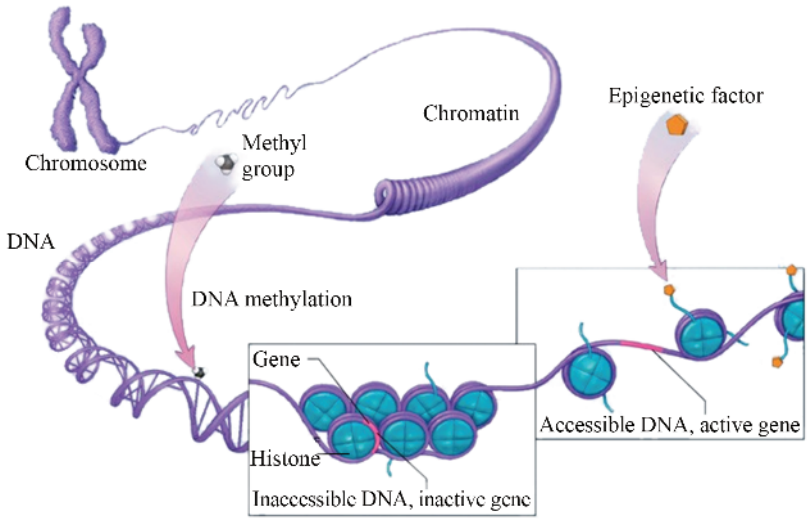

Figure 1.6. *Epigenetic mechanisms (© public domain)*[11]

Other molecules can also have an epigenetic impact. Non-coding RNAs, in particular, such as microRNAs, can have an effect on, for example, the elimination of methylation or the assembly of chromatin. It should also be noted that there are probably other effectors and mechanisms not yet known. In 2017, for example, researchers were able to identify the mode of action of certain proteins in the polycomb group by their ability to inhibit the

10 A promoter sequence is a small part of DNA close to a gene, by which transcription is initiated via DNA polymerase binding. It is therefore an essential element for transcription into RNA.

11 License link: https://creativecommons.org/publicdomain/zero/1.0/deed.en.

reactivity of polymerase, an enzyme that allows DNA to be transcribed into RNA; before specifying:

> "According to current scientific knowledge, it is estimated that this form of gene expression regulation in which polycomb is involved concerns nearly two thirds of the 20,000 genes in our genome" [FLE 18].

A year earlier, in 2016, the antitumor function of this same polycomb protein family was highlighted:

> "By specifically binding to hundreds of genes involved in tumor formation, such as genes that control cell proliferation, this protein combination is able to regulate their action, both temporally and spatially" [FLE 18].

More generally, it has been shown since the early 2000s that epigenetic abnormalities play a major role in tumor activity, either by activating oncogenes or by inhibiting tumor suppressor genes. For example, several types of cancers are associated with an overall reduction in methyl-cytosine levels in the genome compared to healthy tissues. On the other hand, it is sometimes observed that some tumor suppressor genes are silenced by *de novo* methylation of their promoter. Tumor activity can thus be explained not only in terms of genetic mutations but also by epigenetic modifications.

Although the understanding of epigenetic phenomena is not yet sufficiently advanced to ensure the effectiveness of treatments that play on these types of factors, innovative therapeutic strategies are being developed that could eventually lead to a renewal in the field of oncology. The reversibility of epigenetic modifications leaves room for hope. The use of inhibitors of DNA methylation, histone modification (methylation and acetylation, in particular) or other processes is thus possible, even if the molecules currently in existence do not have sufficient selectivity to be acceptable in terms of toxicity and, therefore, to allow therapeutic use. Nevertheless, the prospects for the coming years seem encouraging, according to Giacomo Cavalli, research director at the Institute of Human Genetics in Montpellier:

> "In the case of polycomb, the possibility of neutralizing the active site of the protein responsible for its negative action

> opens the way to new forms of therapeutic treatment. Molecules tested in clinical trials have produced encouraging results for the treatment of circulating tumors such as lymphomas and for certain solid tumors" [FLE 18].

Genetics, genomics and epigenomics thus play a major role in oncology. And, long before "epimedicines" began to be considered, the genetic aspect was already inseparable from the development of different anti-cancer strategies, first of all those of chemotherapy, which flourished during the second half of the 20th Century.