BIOLOGY SERIES

Trajectories of Genetics

Bernard Dujon
Georges Pelletier

iSTE
WILEY

Trajectories of Genetics

Series Editor
Marie-Christine Maurel

Trajectories of Genetics

Bernard Dujon
Georges Pelletier

ISTE
WILEY

First published 2020 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2019956925

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-533-6

Contents

Introduction

The applications of genetics are invading our daily lives. Whether it is for prenatal diagnosis, agronomy or forensic science, we appreciate the accuracy of genetic methods at the same time as we fear their power. What we are able to accomplish today was unimaginable not long ago. We are entering the era of personalized medicine and genetic therapeutics at the same time as we are leaving the empiricism that has prevailed until now for the description and exploitation of the biosphere. And progress is accelerating. But what is it all about? How can we understand what is happening if we do not have clear notions about the fundamental principles of the living world? Principles that have only been revealed slowly to scientists, whose investigations have often followed complex trajectories, are not very explicit to non-specialists.

When we talk about heredity, common language gives the appearance of simplicity. We hear, for example, that this child with light eyes inherited the **genes** from his/her grandmother, who also has light eyes. That's implied, everyone in the family knows it. This simple sentence hides the complexity of the gene. On the one hand, the gene is correctly perceived as this somewhat mysterious element transmitted from generation to generation, even skipping generations, because it's the grandmother we're talking about. On the other hand, the gene is here confused with the observable trait referred to as a ***phenotype****. Here, the eyes are light, but in the population there is a variety of shades, not just light eyes as opposed to dark eyes. Is there a gene for every shade? This seems very difficult to imagine. And then genetics also tells us that this child with blue eyes inherited exactly as many genes from his grandmother with light eyes as from each of his other grandparents who all had dark eyes. So, what is a gene? How does it work?

It was long believed that blood was the carrier of hereditary traits, hence the term "consanguineous" is usually used to designate filiations. We now know that it is ***DNA**** (deoxyribonucleic acid), a macromolecule present in all our cells and that there are tens of thousands of billions of copies of it in an adult human being, all derived from the unique molecules present in the egg that gave birth to him/her. With the progress of genetics, the acronym DNA has become commonplace. It has become a common term for the essence of a personality or even a thing. We hear about the DNA of a company or the DNA of a sports club, which (literally) makes no sense. DNA only exists in living organisms. Moreover, even for personalities, invoking the DNA of an artist to signify his/her talent or the DNA of a child passionate about horseback riding to justify his/her taste for horses brings nothing more to the understanding of the causes than talking about blood; except the appearance of being educated, poorly educated. Because the transition from blood to DNA has been accompanied by a considerable conceptual revolution that is totally ignored here: the particulate nature of the determinants of hereditary ***traits****. Genetics was born when the old vision of mixing "parental fluids" which was invoked to explain the heredity of the characters of the descendants was replaced by the notion of "particles" of heredity. While these imprecise fluids gave the illusion of continuity, each particle, which would later be called a gene, became a separate entity independent of the other particles with which it mixes in discrete proportions, quantifiable by the experimenter. This was clearly understood by Gregor Mendel as early as 1865. But then what is the relationship between the gene and DNA?

I.1. The multiple facets of the gene

A simple answer is not to be expected. The gene is a concept, but a concept that can be manipulated in a test tube! The following chapters describe the evolution of ideas as the nature, organization and functioning of the genetic material became clearer. Where are we at today? The gene remains counterintuitive, because of three main barriers or obstacles to its understanding: one operational, another temporal and the third essential.

The first obstacle is that the gene is both a DNA segment and a functional entity. In other words, its very nature changes depending on how

it is examined, a bit like elementary particles in quantum physics, which can be both particle of matter and wave. However, if DNA is a molecule that is perfectly defined at the chemical level (see Chapter 2) and can be manipulated *in vitro*, the notion of function remains an abstraction specific to biology that applies at different scales from the molecule – an ***enzyme**** for example – to the entire body – eye color for example. And if DNA can be broken down into its elementary components, the functions cannot. This makes the gene a primordial unit, a "particle", but composed of elements that can be separated from one another.

The second obstacle is that the gene is both the element that crosses generations and the element that acts on each of them. At each generation, the functional aspect prevails. Between generations, it is the quality of the transmission of the informational content that matters. This duality is based on two different types of molecules and two different processes. The transmission of the informational content is based on DNA and its duplication (referred to as ***replication****) before each cell division. The functional aspect is based on the copying of DNA (known as ***transcription****) into ***RNA molecules****, for ribonucleic acid, the other category of nucleic acids (see Chapter 2).

This leads us to the third obstacle, because the intergenerational permanence is non-physical, in the form of information – but it is carried by molecules and contained in the organization of their elements. And it is important to understand that at the material level it is not the gene itself that is transmitted from generation to generation, but copies of it, with inevitable potential for error, that is, ***mutation****. The gene, an element of permanence, thus automatically becomes an element of variation. Consequently, over the successive generations of a lineage, and therefore within populations of a species, the same[1] gene will acquire multiple forms, known as ***alleles****. This very important notion makes genetics "the science that uses variation to study permanence", as Philippe L'Héritier, one of the pioneers of this discipline in France, put it very well.

In summary, one could say that the gene is the information necessary to perform a biological function, written in a ***nucleic acid**** and transmitted

1 We will not enter here into the insoluble philosophical problem of knowing to what extent a whole remains the same when its parts change.

from generation to generation with a certain degree of imperfection. But where do the phenotypic traits come from then, if the light-eyed child has inherited as many genes from each of his grandparents, regardless of their eye color?

I.2. Genotype and phenotype: the reality of genetic determinism

We are entering into phenomena that were only slowly understood through a century and a half of research. Let us immediately eliminate the simplistic vocabulary used on the Internet or by some media that talks about genes for anything as soon as humans are involved. There is a jumble of genes for violence, depression, wanderlust, gluttony, generous buttocks, cowardice, resilience, rebellion, slimness, obesity, crime, schizophrenia, homosexuality, mathematical intelligence, etc. Curiously, there is no gene for stupidity that does characterize this list. This is because genetic determinism is anything but simple and direct, as this misleading vocabulary suggests.

First, because each gene can take many forms, it is the alleles that must be considered. Second, because many living organisms – including humans – are ***diploid****, it is the relationship between two alleles that counts: when the effect of one dominates that of the other, we speak of ***dominance**** of the first and ***recessivity**** of the second, but it is also possible that the effects of one and the other are added in variable proportions. Finally, because biological functions are highly intertwined, it is the interaction between alleles of different and sometimes many genes that comes into play. For example, more than 400 different genes are involved in determining the size of human adults, each gene being represented in the population by many alleles associated by pairs in each individual. It is therefore not surprising that the size of the individuals in populations forms an apparent continuum. Is it always so complex? No. Cystic fibrosis is caused by the dysfunction of a chloride ion transporter of the pulmonary epithelium which is the product of a precisely defined gene, located on our ***chromosome**** 7. We now know about 2,000 alleles of this gene that may be linked to the severity of the syndrome. In this case, the alleles of this single gene explain almost all observed phenotypes. The same is true for a complex developmental phenotype, such as the transformation of antennae into legs in the *Drosophila* fly, following the mutation of a single gene.

In fact, the vast majority of phenotypic traits depend simultaneously on genes with major effects – the easiest to identify – and genes with minor effects, whose lists are generally not complete. The power of genetic analysis is highly dependent on the organisms to which it applies, their degree of fertility and the genetic polymorphism existing within natural populations. In many cases in humans, the mutations in all the genes already identified do not explain all individuals with a particular phenotypic trait. It is said that there is a lack of ***heritability**** of this trait. Filling this gap is an important objective of current human genetics.

Before going any further, it is necessary to reconsider what are phenotypic traits. A plant placed on poor, arid soil will generally grow less well than its genetic twin (its cutting, for example) placed on rich, suitably irrigated soil. Similarly, an undernourished animal will suffer growth difficulties. These banal findings are important in practice, but of no interest to us here. The comparison of phenotypes between individuals has a genetic meaning only when these individuals are placed under comparable external conditions. But even under these conditions, the genetic determinism of the phenotype is not direct. In reality, genetically identical individuals (members of an **inbred line** or a ***clone**** or simply identical twins) placed under identical conditions will never be strictly identical phenotypically. There is always, for each phenotypic trait considered, an individual variation around the average value. Within a genetically homogeneous population, this variation generates a statistical distribution specific to this trait. Depending on the trait considered, the variance of the distribution will be more or less large. It is therefore essential to understand that what is transmissible to the offspring is the form and parametric values (mean, variance) of these distributions, not the precise phenotype of each individual. It was this fundamental observation that led Wilhelm Johannsen to define ***genotype**** and ***phenotype**** as early as 1903 (see Chapter 1). When we talk about genetic determinism, we must therefore keep in mind that the genotype, an element of intergenerational permanence, does not directly define the phenotype of each individual, but defines the statistical distribution of the phenotypes of all individuals who carry this same genotype, placed under the same external conditions.

Today, we know how to fully define an individual's genotype by fully sequencing his or her ***genome****. For a diploid, the two alleles of each gene must obviously be individually sequenced, which has only recently become

possible. However, except in specific cases of prenatal diagnosis, for example, it is still difficult to predict the resulting phenotype, as we are far from knowing all the interactions that can exist between the alleles of all genes. Genetic counselling therefore remains probabilistic.

There is also a time dimension, because the functional state of a gene depends not only on its allelic form (and possibly other genes of the same genome), but also on its functional state during the previous generation. This is called ***epigenetic**** effects. The molecular mechanisms responsible for them will be discussed in the following chapters. At this stage, it is sufficient to recall that epigenetic phenomena are themselves genetically determined and, for some of them, not only by the genotype of the individual concerned, but also by those of previous generations.

I.3. The products of genes

One would naively imagine that genes have a wide variety of products, given the number and diversity of observable phenotypic traits. It is exactly the opposite. There is only one type of direct gene product: RNA. Native RNA molecules copy the information from the DNA of the genes but, depending on their nucleotide ***sequences****, they then engage in very different functional pathways. It is therefore actually RNAs and not the genes themselves that are responsible for all the functions of living cells. But these RNA molecules have a limited lifetime and, with few exceptions, are not passed on to the next generation because they do not replicate. They therefore do not ensure intergenerational permanence. This is not the case with the RNA molecules of some ***viruses**** that replicate, invade cells and can be transmitted to offspring.

Among the functional pathways in which RNA molecules engage, one leads to the synthesis of another category of macromolecules made up of ***amino acids**** linked to one another in specific orders: the ***proteins****. Protein synthesis is a complex chemical process that involves several types of RNA molecules acting together in a coordinated manner. During this process, the information made of the succession of ***nucleotides**** in an RNA molecule, known as the ***messenger RNA**** sequence, is translated into an information made of the succession of amino acids in a protein, that is another sequence.

Since nucleotide sequences offer infinite combinations, the theoretical diversity of proteins on which phenotypic traits ultimately depend is unlimited. This explains how, through a single mechanism, genes can be at the origin of all phenotypic traits. But how is achieved this passage of information from RNA to proteins, known as the ***translation**** process?

This is where the ***genetic code**** comes in. This term is often used inappropriately by the public, leading to a total misunderstanding of genetics. The genetic code is not the informational content of a living organism (its genome) that is unique to it and differs between species and even between members of the same species. It is a universal deciphering code, common to all living organisms currently known. The genetic code, like all decryption codes, is a set of simple rules that establish the correspondence between messages made up of elements of different natures. It is somewhat like a message made up of a series of dots and dashes that can be deciphered in Latin script using Morse code. Here, the genetic code establishes the correspondence between a succession of nucleotides in a nucleic acid, a messenger RNA molecule, and a succession of amino acids in a protein that is its translation. The genetic code does not determine the nature of the synthesized protein. This depends only on the information carried by the messenger RNA, which in turn comes, through a sometimes complex path, from the information carried by the gene. To say, as we too often hear, that rewriting a genome modifies the genetic code of a cell is therefore a complete misinterpretation, which amounts to saying that the Morse code has been changed. Fortunately, this is not the case, because, as any telegraphic communication would become impossible, the cell would immediately become unable to decipher its own genome and disappear!

For reasons that will be explained in this book, the genetic code establishes the correspondence between sequences of three nucleotides (there are $4^3 = 64$ different ones) with the 20 main amino acids that make up proteins. It should be noted that the universality[2] of the genetic code in the known living world demonstrates the common ancestral origin of the protein synthesis mechanism used by modern living organisms. It is a mechanism that is practically fixed in terms of evolution, which is extremely rare in the living world. This universality allows the transfer of genes between different contemporary organisms. A phenomenon called ***transgenesis****, which, long

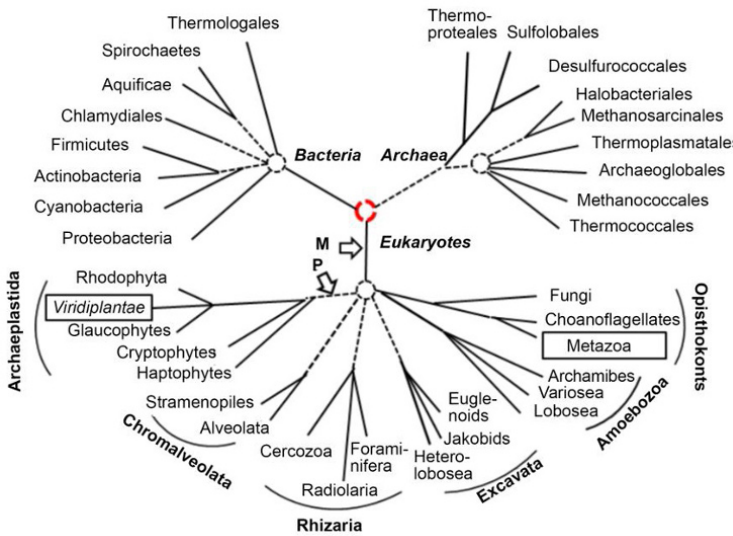
2 In reality, this is almost universal, as minor changes are observed in some species.

before its artificial use, played a critical role in natural biological evolution. In its absence we would not be here to talk about it.

But then, if all living organisms use the same genetic code and can sometimes even exchange genes, can we talk about human genes, cat, grasshopper, wheat, parrots or fish genes? And what is a species?

I.4. Unity and diversity of the living world

Historically, the notion of living species preceded that of biological evolution. Originally, species were defined as sets of individuals with common characteristics that were sufficiently distinctive from those of other sets that, therefore, constituted other species. If these characteristics changed over time, then new species could appear, hence the idea of transformism and evolution. For organisms that reproduce sexually, it is the viability and fertility of hybrids that indicates if both parents are members of the same species or of two different species, as if there were reproductive barriers between species. How then to reconcile these observations? Genetics was born in this historical context and very quickly sought to rationalize the bases of speciation by the idea of incompatibility between alleles of different genes, assuming that the same genes should be present in different allelic forms in different species. As a result, there are no longer any human genes, cat, grasshopper, wheat genes, etc., but only different forms of the same genes after divergence of their sequences due to mutations that have accumulated over successive generations from their common ancestors. Obviously, as the number of generations increases, the differences make it more and more difficult to recognize their common ancestry. But below this limit, it will be possible to reconstruct their probable relationships by comparing their sequences. The idea of trying to reconstruct genealogical links between species is not new. Antoine Nicolas Duchesne gave a first concrete illustration of this idea in his *Histoire naturelle des fraisiers* in 1766, concerning the strawberry species that he had studied and cross-bred. This idea of species genealogy, taken up mainly by Jean-Baptiste de Lamarck and Charles Darwin, was developed by their immediate successors, including Ernst Haekel, who proposed the first tree of life as we knew it then. With the sequencing of genomes, the search for evolutionary filiations has recently taken on a considerable scale, to the point that our ideas on the different branches of the evolutionary tree of the living world have become much clearer, sometimes modifying older trees (see Box I.1).



The tree is constructed according to commonly accepted phylogenetic principles based mainly on ribosomal RNA sequences, a molecule common to all living cellular organisms. Viruses are therefore excluded. Only the main branches are represented with the names of the corresponding groups. Many branches have very large numbers of known species, sometimes millions, and it is not known how many species remain to be discovered, probably a lot. Solid lines indicate the relationships established. The dotted lines indicate uncertainties. The small arrows place the primary endosymbiosis events of bacterial lines supposed to be at the origin of the formation of mitochondria (**M**) and chloroplasts (**P**) of eukaryotic cells. Mitochondria come from the Proteobacteria lineage while chloroplasts come from the Cyanobacteria lineage. It is obviously important to understand that these are not the current bacteria, but their very distant ancestors, who have now disappeared.

From the hypothetical origin (placed in the center, red), we distinguish three major domains: bacteria, archaea and eukaryotes. The names of the main subgroups of eukaryotes are shown in bold. It can be seen that on this global scale, all animals on the one hand and all plants and green algae on the other hand each represent only one terminal branch (*Metazoa* and *Viridiplantae*, respectively). The vast majority of the diversity of the living world is represented by microorganisms. This tree is based on the initial work of Carl Woese and his colleagues (Woese *et al.* 1990) with many modifications to accommodate subsequent results.

Box I.1. *Simplified tree of life. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

As early as 1937, Édouard Chatton differentiated between organisms whose cells have a ***nucleus**** containing the chromosomes, which are thus isolated from the cytoplasm, and others whose cells are not compartmentalized, leaving the chromosomes directly in the cytoplasm. The first, called ***eukaryotes****, include animals, plants, fungi, algae, but also a very wide variety of unicellular microorganisms that are generally not well known to the public. Eukaryotic cells also contain other ***organelles**** that carry their own genes: ***mitochondria**** and ***chloroplasts****. The latter, historically called ***prokaryotes****[3] because their greater simplicity could suggest that they were more primitive, are almost exclusively unicellular microorganisms initially confused under the name of ***bacteria****. The first sequence comparisons were to confirm the uniqueness of eukaryotes, but to distinguish two groups among the prokaryotes, one of which, now called ***archae****[4], had been much less studied than the other, which retained the name **bacteria**. The living world would therefore be made up of three main subdivisions, whose order of separation remains a subject of discussion. Archae, bacteria and eukaryotes are themselves subdivided into many subgroups. Eukaryotes contain at least five, perhaps eight, subgroups of which animals and plants represent only two particular lines among a much larger number of single-celled microorganism lines. This classification ignores the world of **viruses**, because they are not living cells, but molecular particles that must infect living cells to multiply. We know of viruses that infect archae, other bacteria, other eukaryotes, each with more or less broad host specificities. We are far from knowing all the existing viruses and it is likely that new ones are constantly being formed. The latest results of ocean exploration suggest that viruses alone represent the largest mass of the living world.

The contributions of the three groups of cellular organisms and viruses to the development of genetics were very different. As we will see, genetics was born from the study of heredity in plants and initially developed with the study of insects, fungi and few other organisms, all eukaryotes. It then became molecular, with the study of bacteria and their viruses, called ***bacteriophages****, which paved the way for fine structure analysis of the gene and provided the first tools that would later be used for genome analysis.

3 The exact term would have been “acaryotes”.

4 For historical reasons, the archaea were initially referred to as “archaebacteria”, while the other prokaryotic group was referred to as “eubacteria”. Despite this designation, there is no evidence that archaea are more primitive than bacteria.

Genomic tools now make it possible to address all organisms and their viruses, but there remains a large bias in our present knowledge between the different branches of the living world, hence the possibility for important new discoveries through the study of the lines that have remained practically unexplored. We will see some examples of this. This bias has never called into question the universality of the established principles, but it questions their completeness.

I.5. Permanent changes: lessons from genomes

The spectacular advances in genomics illustrate this question. It teaches us that the species differ not only from each other in the allelic forms of the genes they share in common, but also in the presence, absence or multiplication of certain genes. The same applies to members of the same species, albeit to a lower scale. Two individuals randomly selected from the human species differ not only in the allelic forms of their genes, but also in the exact number of their genes. Here, the difference remains small in numerical terms – a few units at most – but not necessarily in functional terms. For other organisms, in particular microorganisms, the number of genes common to all members of the species may be significantly less than the number of genes of each individual, itself much less than the estimated total number of genes of all members of the species.

Gene gains or losses that explain these variations are based on mechanisms that are only now beginning to be better understood. The loss of genes from generation to generation seems more or less constant, limited only by the possible deleterious effects that may result. These losses are numerically compensated by gains that can have two sources: acquisition by horizontal transfer from other contemporary organisms or *de novo* gene formation by mutations from existing sequences. These events remain rare at the individual level, but they play a considerable role in the evolution of populations. As a result, every genome is inherently imperfect, an instant image of entangled dynamic processes whose effects are mitigated by the resilience of their products. This is far from the simple and naive genetic determinism too often imagined where any modification of a genome would be prohibited.

I.6. The future of genetics: hopes and fears

It is rare to keep in mind the progress of past centuries. Their benefits are taken for granted. Their misdeeds are known or forgotten. The remarkable advances of the 20th Century in terms of agricultural production, animal husbandry, human health and domestication of micro-organisms had a broad genetic basis. Who remembers? And since they have been completed, what is left to be done? Current advances in genetics seem to raise as many questions among the public as they raise hope among specialists. When a genetic anomaly is responsible for a disease, what could be more natural than to look for ways to repair it, especially since it is disabling or greatly reduces life expectancy? This idea is not a new one; it emerged about 60 years ago, with the discoveries of the structure of DNA and the genetic code. Does the knowledge of genomes allow us to anticipate the appearance of illnesses? Does transgenesis make it possible to treat them? We have just shown the possibility of transforming lymphocytes into drugs against malignant cells. Can living organisms be created from genomes obtained by chemical synthesis of DNA? Today with microorganisms, tomorrow with animals or plants? Life is built on the chemistry of nucleic acids and proteins. Were there any other possible choices? Is a ***xenobiology**** possible, that is, a new biology separate from our living world because it uses other atoms and molecular structures?

Doesn't genetics give us the means to act ethically? This is not a new question. **William Bateson**, one of the founding fathers of genetics, expressed himself in this way in his speech at the closing banquet of the 4th Genetics Conference that was organized in Paris:

> What will become of genetics? We are at the beginning, and when you consider the depths and heights it can reach, we are truly dizzy. Genetics gives the human race a power that could never be predicted and that is extremely dangerous [...] perhaps – and I don't think I'm being ridiculous in saying this – will we have the power in a century to regulate the fate of the human race, and the types we don't want will not be born. I am not sure that a government with this power will not abuse it.

That was 1911. A century has now passed. Eugenicist abuses led to nothing but tragedies, as their foundations were so wrong. They were inspired more by Darwin than by Mendel. But the prophecy was fulfilled. We have the power to influence the fate of the human race, to deliberately

modify ecosystems, to artificially manufacture micro-organisms – including pathogens – and, why not, soon to bring extinct species back to life. How can we apply this in a useful way? This is the first question that makes sense, the other being how to preserve the curiosity of the unknown.

I.7. Origin of the book and content of the chapters

This book will illustrate some of the current trajectories of genetics, with an emphasis on the concepts on which they are based. It is derived from a symposium (Comptes Rendus Biologies 2016) held at the *Académie des sciences* in Paris in September 2016 on the occasion of the 150th anniversary of the publication of Mendel's work, *Versuche über Pflanzen-Hybriden*. This conference brought together French researchers – a country that has long been reluctant to recognize this discipline for its true value – and foreign researchers. This book will address the fundamental concepts and issues without which it is impossible to understand the development of current research. It will attempt to show that beyond the immediate benefits, it is intellectual curiosity that gives genetics its full impetus for the future. The novelty and power of the dogmas established by successive discoveries of genetics during the 1950s and 1960s could suggest that it had completed its work, while genes had not yet been chemically isolated. Subsequent developments in recombinant DNA and genomics, which some considered either dangerous or superfluous, would instead open up unsuspected areas of research that now shed light on all branches of biology, from the most fundamental to the most applied. The book will illustrate the transition between these two periods, the first three chapters dealing with the fundamentals, the next three chapters on genomics and the last three on current applications and expected developments.

We would like to thank Jean-Yves Chapron and Éric Postaire for their wise advices and Jean Weissenbach for his careful review of the manuscript.

I.8. References

Comptes Rendus Biologies (2016). Trajectories of genetics, 150 years after Mendel. *Comptes Rendus Biologies*, 339 (7/8), 223–336.

Woese, C.R., Kandlert, O., Wheelis, M.L. (1990). Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87, 4576–4579.

1

Following Ariadne's Thread from Genetics to DNA

> In order to make natural history a true science, we must focus ourselves to research that can reveal, not the individual and particular aspect of one animal or another, but the general process by which nature reproduces and preserves itself.

So wrote **Pierre Louis Moreau de Maupertuis** in 1752 (Moreau de Maupertuis 1752) more than a century before the beginning of genetics.

The natural sciences have long been interested in describing the diversity of species before considering the mystery of this commonplace feature of life that makes individuals of one species generate other individuals of the same species. While genetics met the recommendation of P. Maupertuis, the path taken by scientists to develop this science was, as we will see, more like a labyrinth than a Roman road!

1.1. The birth of genetics

In 2016, genetics turned 110 years old (Gayon 2016). This term was first introduced publicly by Bateson at the Third Conference on Plant Hybridization held in London in 1906[1], exactly 40 years after Mendel's

1 This conference was later renamed the *Third International Congress of Genetics* and included in an uninterrupted series of international conferences since 1899. They have been held every five years since 1948. The latest, the 22nd in the series, was held in Foz de Iguaçu, Brazil in 2018.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

publication (Mendel 1866), which had been poorly distributed and very generally misunderstood because it represented a methodological and conceptual break with everything that had existed before. However, in the century preceding Mendel's work, botanists and horticulturists observed many offspring of plant crosses. **Thomas A. Knight**, at the end of the 18th Century in England, crossed pea varieties differing in seed and leaf colors and observed their offspring, but without having the idea of counting the different types obtained. In France, by hybridizing melons differing by several characters, **Augustin Sageret** in the 1820s had the first intuition concerning the discontinuous nature of heredity, in opposition to the vision of heredity by mixing, like two fluids, the dominant idea of the time. He wrote:

> It appeared to me that, in general, the similarity of the hybrid to its two ancestors does not consist in an intimate fusion of the various characters specific to each, but rather in an equal or unequal distribution of these same characters.

At the beginning of his memoir, Mendel (see Box 1.1) justifies the literally "monastic" work that he had done during the eight preceding years by the desire "to follow up the developments of hybrid progenies" in ornamental plants. The first part of the memoir on pea hybrids (*Pisum sativum*) is a real experimental demonstration that ends with a kind of theorem:

> Hybrids produce ovular and pollen cells that correspond in *equal number* to all constant forms resulting from the combination of traits brought together by fertilization [which produced this hybrid].

In other words, the stated rule is simple: a hybrid that has received from one parent the form "A" and from the other the form "a" of a given character, will produce ***gametes**** "A" and "a" in equal numbers, A and a being mutually exclusive in these gametes, hence their name "allelomorphs[2]". This law is currently known as the "law of purity of gametes" or "Mendel's First Law". He continued:

2 Here we are talking about ***alleles****, see the glossary.

> This proposal provides a sufficient explanation of the diversity of forms in the descendants of hybrids as well as of the numerical relationships we observe between them.

Indeed, this simple equality explains the different types of plants and their proportions in the progeny of the pea hybrids that he had produced either by self-fertilization or by back-crossing with each of their parents[3]. Mendel systematically found the same proportions for the seven differential traits he followed, for which there are two contrasting states and only two, such as two colors (yellow/green) or two seed shapes (smooth/wrinkled), two pod shapes (uniform/strained), two plant sizes (high/dwarf), etc.

Gregor Johann Mendel was born on July 20, 1822, on the family farm in Heinzendorf in what is now the Czech Republic. He received elementary education in the village school where the teacher encouraged his parents to make him continue his studies in high school, which he did excellently from 1834 to 1840. In 1838, the family situation became more precarious after a serious accident prevented his father from working. However, Mendel continued his studies at the Institute of Philosophy in Olomouc for two years, then entered the monastery of Saint Thomas in Brno (Order of Saint Augustine) as a monk in the hope of becoming a teacher by completing his training at the monastery's expense. He was admitted to the novitiate in 1843, chosing Gregor as a first name, before being ordained priest in 1847, appointed parish priest in 1848 and assistant professor in 1849. In 1851, he went to study mathematics and physics at the Vienna Institute of Physics (Christian Doppler) and deepen his knowledge of entomology, paleontology, botany and plant physiology. Upon his return to Brno in 1854, he began his experiments on plant hybrids while teaching natural sciences and physics. The city of Brno and its monastery, then ruled by Abbot Napp, offered a particularly rich intellectual environment. In particular, the monastery was engaged in reflections on heredity, with objectives of application to sheep breeding and to the orchards of its domains.

Mendel had been interested in gardening and flowers since childhood. For example, he produced a variety of fuchsia that bears his name, and an original variety of peas from his hybridizations. He made crosses of pear, apple and cherry trees for the monastery's orchards and even obtained a medal for his stone fruit varieties. He bred white and gray mice and crossed them to follow the heredity of coat color, a task he could not pursue within the monastery. Throughout his life, he was fascinated by bees, practicing beekeeping and their selection. He had a strong reputation in the country as a meteorologist, taking an

3 By convention, the hybrid generation is called F1 and its offspring by self-fertilization is called F2.

interest in sunspots and taking precise and regular measurements until the day before his death. He had the opportunity to travel, going to Paris and London in 1862 (without being able to meet Darwin, absent at the time), to Germany, to the Alps, and to visit the Pope in Rome.

In 1868, he was appointed abbot at the head of the monastery, and recognized as an excellent teacher, an esteemed botanist and a highly valued citizen in the city of Brno. Part of his activity consisted in managing and inspecting the different dependences of the monastery. He had disputes with the government that was trying to tax monastic property. He was a member of many learned societies, curator of the *Institut des sourds et muets* (an institute for the deaf and mute) and, towards the end of his life, President of the Moravian Mortgage Bank. He died on January 6, 1884, as a result of kidney disease. His funeral was attended by a large crowd, paying tribute to a man highly appreciated by his fellow citizens, but unaware of his scientific contribution to biology. His successor at the head of the monastery burned his archives.

Often cited by other scientists as early as 1867, his demonstrations were generally misunderstood until the early 20th Century.

Box 1.1. *Gregor Johann Mendel (1822–1884)*

We now know that this rule corresponds to the mechanism of this particular cell division called ***meiosis**** that halves the number of **chromosomes** to form gametes (see Chapter 3). In Mendel's time, the existence of chromosomes was unknown. They were described in 1875 by **Eduard Strasburger** (Strasburger 1875) and named in 1888 by **Heinrich-Wilhelm Waldeyer-Hartz** (Waldeyer 1888). **Walther Flemming** (Flemming 1879) described their movement and distribution between the two daughter cells during ***mitosis**** in 1879, but it was **Edouard Van Beneden** who, in the *Parascaris equorum* nematode, first described meiosis in 1883 (Van Beneden 1883). The individuality and continuity of chromosomes during development were demonstrated by **Theodor Boveri** between 1887 and 1902, and it was in 1903 that **Walter S. Sutton** (Sutton 1903), explicitly linked the distribution of chromosomes during meiosis in a grasshopper (*Brachystola magna*) to Mendel's rule: chromosomes are distributed in pairs (like pairs of stockings!) before meiosis, and a gamete retrieves one from each pair. For a given pair, there will therefore be as many gametes possessing one member as gametes possessing the other, exactly like Mendel's A and a factors. The gametes, in turn, after ***fertilization****, will reproduce an organism with the initial number of

chromosomes. This is how the **chromosomal theory of heredity** developed, offering the first materialization on which relied, as it was said at the time, the "power to transmit some particular traits of the parents to the progenies, in addition to the characteristics of the species".

Defined at the time by **Wilhelm Johannsen** as "the science of the propagation of life", or "the science of the fixed elements that compose organisms", genetics was to be officially born in 1900 after the rediscovery of Mendel's work by three botanists, **Hugo de Vries, Carl Correns** and **Erich Von Tschermak**, working independently on different plant species (Campbell 1980). What became known as "Mendelism" was confirmed in mice, for coat characteristics, by **Lucien Cuénot** as early as 1902 (Cuénot 1902), and in humans two years later, for some cases of polydactylism, by **Charles B. Davenport** (Davenport 1904). It should be noted that the hereditary transmission of this trait in some families had already been described by P. Maupertuis in the 18th Century!

1.2. The foundations of a new science

Very quickly, during the first decade of the 20th Century, new concepts were developed, with the introduction of a new vocabulary. Bateson coined the terms ***homozygous**** to reflect the fact that an individual has received the same "element" from both parents (AA or aa using Mendel's symbolism), or ***heterozygous**** in the opposite case (Aa or aA). We call **alleles** the different forms taken by the same element (A or a, but also A1, A2... a1, a2, etc.), being well aware that only two forms at most can coexist in the same organism and only one in each of its gametes. We owe to de Vries the notion of **mutation**, which he introduced in 1886 in studies on the appearance of new forms in the oenothera (evening primrose) that he called "mutant forms". From 1901 onwards, he developed this new concept in relation to the work of Mendel in a famous book, including an evolutionary perspective, entitled *The Theory of Mutations* (de Vries 1901–1903). The term "mutation" refers to the sudden appearance, without apparent cause, of new characteristics of an organism that become heritable. The phenomenon had long been known to horticulturists, when an abnormal shoot (a sport) spontaneously appears on a tree giving, for example, flowers or fruits of a different color, without them drawing any conclusions. We will come back to this in the following chapters.

In contrast with many variations that can be found in the forms of plants, the ***mutants**** of de Vries, which were larger oenothera known for this reason as "gigas", had the particularity that the novel trait persisted in the progeny by self-fertilization[4]. He therefore saw the mutation as a mechanism that could explain the appearance of new species. It was not until 1907 that **Anne M. Lutz** showed that the gigas mutants were quite special, because they were plants whose chromosome numbers had spontaneously doubled (Lutz 1907). They had become ***polyploid**** and this feature was preserved in their offspring. The term ***ploidy**** refers to the number of chromosome sets of a cell or organism, so we have the ***haploid**** series (one set), ***diploid**** (two sets), triploid, tetraploid, etc. It should be noted that after doubling or quadrupling the number of their chromosomes, plants usually show only a size increase, but retain their fertility, while triplications (or odd ploidies of higher ranks) induce sterility phenomena. Sterility is sometimes useful. For example, if bananas are edible, it is because the domesticated banana trees are triploids, rendering their fruits seedless. The same applies to oysters without milt.

The idea that chromosomes show in perfect pairs had to be completed when, with the progress of cytological observation techniques, it became apparent that there was a pair of two different chromosomes in males of insects but not in females (Wilson 1905). This pair therefore seemed to determine sex (see Chapter 3). This was the first time that a characteristic could be directly assigned to a given chromosome, validating the chromosome theory of heredity[5]. It was only later that ***cytogenetics**** acquired techniques to describe chromosome pairs as karyotypes* representative of a given species. In humans there are 23 pairs of chromosomes (including the sex chromosome pair), in barley 7, etc. But it is for the identification of accidental alterations of individual ***karyotypes**** that cytogenetics has become especially useful. Thus, the presence of a supplementary chromosome in cell cultures of Down syndrome children was discovered as early as 1956, published three years later (Lejeune *et al.* 1959) and named *"**trisomy**** 21" to indicate the presence of three copies of

4 This is a critical point. It differentiates mutations that correspond to a genetic, and therefore heritable, change from accidental phenotypic variations that can affect a particular individual during his or her development.

5 Subsequent research aimed at finding differences in chromosomal shapes that could explain the human races as defined at the time will be ignored; all claims being only gross technical errors.

chromosome 21 (the penultimate smallest chromosome) instead of two (an example of ***aneuploidy****).

1.3. Gene, locus and genetic maps

Where did the idea of a gene come from? The distinction between *genotype* and *phenotype* introduced by Johannsen in 1903, mentioned in the Introduction, is fundamental because it marked an essential turning point in the evolution of ideas. Previously, these notions were implicitly confused for centuries, opening the door to beliefs about the heritability of characters acquired through the effect of the environment, with **Georges-Louis Buffon, Jean-Baptiste Lamarck** and even **Charles Darwin**, a belief that lasted until the middle of the 20th Century with Lysenkoism. In this vision, what is transmitted to the offspring by the parents are, with slight differences, "germs of organs" secreted by the organs themselves, the resemblance of the descendant to one or the other parent resulting from the dosage of these hypothetical germs from one or the other parents. Darwin, in his "pangenesis" theory, pushed this idea to the elementary structures of the organism, imagining gemmules coming from each cell that would accumulate in the gametes.

It is in this context that de Vries postulated in 1889 the existence of hereditary units, or "pangenes", which would no longer be emanations of somatic cells, but materialistic determinants of hereditary traits that do not leave the cell (de Vries 1889). For him, "the nucleus of the cell is the pool of hereditary characters". He made an analogy with the development of other sciences:

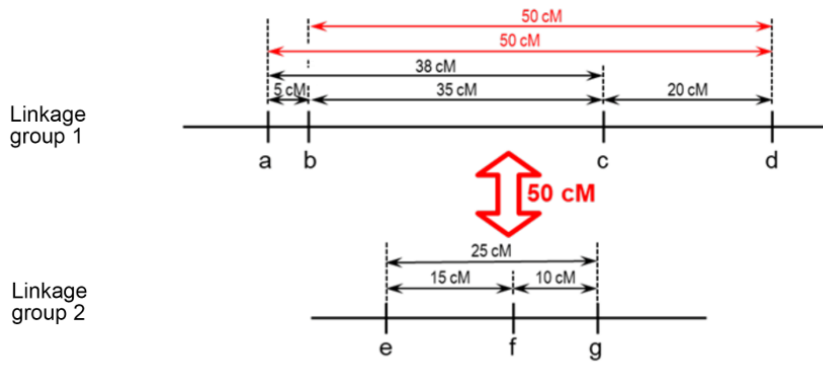
> Just as physics and chemistry can be reduced to the study of molecules and atoms, so biological sciences must study these hereditary units in order to seek in their combinations the explanation of the manifestations of the living world.

Taking these ideas up in the light of Mendel's work, Johannsen introduced in 1909 the simpler term "gene" to designate the fixed factors of Mendelian heredity, the "hereditary particles". He wanted a short term totally free of any speculative hypothesis and coined this word, taken from the Greek *gennao* meaning "to generate" (Wanscher 1975).

It remained to be understood the nature of these genes, because when Mendel analyzed the simultaneous transmission of several different traits, he observed that they behaved independently of each other. This gave rise to Mendel's second law, generally known as the "law of independence of characters". Was there not a contradiction with the chromosomal theory of heredity if we assumed that two elements (now genes) that determine two traits were born by the same chromosome? Shouldn't they, contrary to what Mendel observed, always be transmitted together since they were physically linked? As early as 1906, while studying the simultaneous transmission of two traits in peas, Bateson and his colleagues (Bateson *et al.* 1906) observed, in the offspring of a purple-flowered (P) and long-pollen (L) line crossed with a red-flowered (p) and round-pollen (l) line, too many plants maintaining the parental associations **P-L** or **p-l** compared to the proportions provided by Mendel (9/16 and 1/16, respectively, for the second generation). They introduced the notion of genetic "coupling" to reflect this trend, erroneously interpreting it as a preferential selection of gametes reproducing the parents' constitution.

A few years later, **Thomas H. Morgan**, noting the same phenomenon for the mutations of the fruit fly (*Drosophila melanogaster*), clarified the phenomenon known as ***genetic linkage****. He hypothesized that the genes responsible for the observed traits, eye color (red or white) and wing size (normal or short), were lying on the same chromosome, explaining their tendency to be transmitted together, but insisted that it is only a tendency depending on their relative distance on this chromosome. The closer the distance between the two positions (called ***loci****), the stronger the tendency. In other words, during gametogenesis, the two chromosomes of the same pair must exchange parts, otherwise there would be no flies of recombinant genetic constitution, different from the parents. These exchanges, called ***cross-over****, provide the gametes with recombinant chromosomes in which the two loci have broken their initial linkage, replaced by a new linkage (Morgan *et al*. 1915). But the cross-overs remained hypothetical, their existence was only revealed by the formation of recombinants between alleles of the two genetically linked loci. It was not until 1931, when **Barbara McClintock** and her colleague studied corn meiosis under a microscope, that they were able to observe the reality of this phenomenon of exchange between chromosomal arms (Creighton and McClintock 1931), a phenomenon confirmed a few years later by the observation of Drosophila meiosis (Stern 1936).

In the meantime, Morgan's team had established the first "genetic maps" of Drosophila chromosomes using the ***recombination**** frequency between loci as a measure of the genetic distance between them on this chromosome (see Box 1.2). A low frequency of recombination between two loci indicates a strong genetic link, that is, a short genetic distance between them. A high frequency of recombination indicates a weak genetic link, that is a long genetic distance between them. When the distance between the loci becomes too great, the genetic linkage disappears, that is, although carried by the same chromosome, two loci may be genetically independent of each other. This is what happened by chance to Mendel for the two traits relating to seeds, located on chromosome 1 of pea which he studied in detail, explaining his law of independence of characters (Blixt 1975).



Genetic maps are linear (or sometimes circular) diagrams that represent the relative positions of loci (here *a*, *b*, *c*... *g*) on a **genetic linkage group** and the distances between them, measured in **centimorgans (cM)**. These maps are established by counting the recombinants in the progeny of crosses between parents whose alleles differ at the loci considered.

One centimorgan represents 1% of recombined chromosomes in the progeny. It should be noted that, depending on the living organisms studied, the relationship between recombined chromosomes and observable recombinants in the progeny of a cross differs according to their mode of reproduction and the ploidy of the observable phase (see Chapter 3).

Two loci are **genetically linked to each** other when their distance is less than 50 cM. They are **genetically independent** at 50 cM, the value obtained for loci that are either carried by separate chromosomes or sufficiently distant from each other on the same chromosome. Short genetic distances add up (e.g. *e*, *f* and *g*).

When their values increase, the sums dwindle more and more as we approach 50 cM (e.g. $ac < ab + bc$, $bd < bc + cd$, $ad < ab + bc + cd$). The additivity of distances may not be sufficient to determine the order of loci (e.g. the order *a-b-c* could become *b-a-c* if the genetic distances of *ac* and *bc* become close to each other).

A linkage group is made up of loci **genetically linked to one another step by step**. For example, in linkage group 1, locus *d* is independent of loci *a* and *b* (red), but is linked to locus *c* which is linked to loci *a* and *b*. Linkage groups 1 and 2 can either belong to two different chromosomes or be sufficiently distant from each other on the same chromosome that all combinations between their loci are 50 cM apart.

The scale of genetic maps varies according to their application. They may concern a gene (fine maps), a group of genes, whole chromosomes or whole genomes. The relationship between genetic distances (in cM) and physical distances (in DNA nucleotides) varies greatly between organisms and is not constant along the same chromosome (hot or cold spots of recombination). The genetic distances within a same genome may vary according to its genetic content or the sex of individuals (for example, in humans, the female meiosis produces twice as many recombinants as the male meiosis).

Box 1.2. *Genetic Maps. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

With this method, which was revolutionary, genes were placed one after the other along the chromosomes, long before their very nature was understood. In reality, these were not really genes, but loci, each identified by the presence of at least two alleles conferring a recognizable character to the organism. The different loci were aligned along lines each representing a chromosome or a chromosome fragment, and graduated in centiMorgans (cM). One cM corresponds to the genetic distance between two loci, generating 1% of recombinant combinations of alleles in the gametes. The genetic maps thus established, often designated “factorial maps”, should not be confused with the physical maps of DNA and genomes (see Chapters 4 and 5).

1.4. Mutagenesis, first ideas on the material nature of the gene

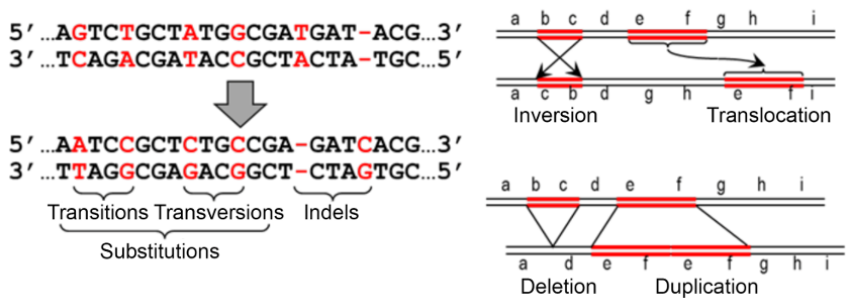
It was the discovery of radioactivity that provided the first indications on the material nature of genes. As early as 1908, **Stuart Gager** exposed pollen

grains or plant ovules to radium radiation and discovered that these treatments gave rise to mutant plants. Radioactivity (here gamma radiation) therefore had the power to alter the genetic material, the nature of which remained mysterious except that it became clear that it was made of atoms. The practical aspect of the phenomenon was important: interesting mutants could be collected from smaller numbers of plants than under normal conditions. But the mechanisms of this ***mutagenesis**** remained to be elucidated.

It was in 1927 that things became clearer, with the work on the *Datura stramonium* (Gager and Blakeslee 1927) and on the fruit fly (Muller 1927). Irradiation of anthers or floral ovaries with increasing doses of gamma rays in the former case, or irradiation of whole animals with increasing doses of X-rays in the latter, increased the chances of appearance of mutations in genes as well as in the chromosomes of the offspring. The mutations in the genes followed perfectly the rules of Mendelian ***segregation**** in progenies, but their nature eluded the power of analysis. In contrast, chromosomal mutations became clear: they were rearrangements of chromosome fragments suggesting that breaks had been caused by radiations and that broken fragments had been abnormally re-joined. The rates of these rearrangements were very high (17% on average for experiments on plants). In some cases, mutants carrying these rearrangements showed reduced fertility when crossed with their normal parent. Among these rearrangements, microscopic observations and subsequent genetic analyses carried out mainly on the fruit fly distinguished ***inversions****, ***deletions****, ***duplications**** and ***translocations**** (see Box 1.3).

The discovery of induced mutagenesis played a crucial role in the development of fundamental genetics, as well as in its applications with, for example, the production of new plant lines useful for agronomy or simply ornamental. Today, more than 3,200 mutations produced mainly by irradiation and selected for their interest can be found among cultivated plants, from fruit trees to field crop species (wheat, barley, rice, rape, etc.), and ornamental plants. In addition to the initial use of radiations (gamma, X, UV, etc.), chemical mutagenesis was added as the molecular nature of genes became better understood (see Chapter 2). A long list of molecules (mustard gas, sodium azide, ethyl methane sulfonate, nitrous acid, di-methyl sulphate, etc.), are demonstrated as ***mutagens**** and used for specific purposes under safe conditions in research laboratories. Standard tests have been developed

to measure the potential mutagenicity of new molecules used for food or in the environment. The molecular nature of the mutations produced depends on the precise physical or chemical nature of the mutagen, but, unlike modern DNA-based mutagenesis methods (see Chapter 6), physical or chemical mutagens act randomly throughout the genome and not on a specific genetic target.



A distinction is made between **point** and **structural** mutations. Mutations result from spontaneous or mutagen-induced errors in DNA replication or repair processes, causing daughter molecules to differ from the parental molecule. Conventional mutagens can act by chemically altering nucleotides (nitrous acid, nitrosoguanidine), forming covalent chemical bonds between neighboring or paired nucleotides (UV rays) or including other macromolecules (formaldehyde), intercalating between nucleotides (acridines) or by cleaving DNA chains on one or two strands (ionizing agents, X-rays). It should be noted that the normal metabolic reactions of the cell produce various oxidized molecules called ROS (Reactive Oxygen Species) which have a mutagenic power. Whether natural or artificial, physical and chemical mutagens act **randomly** throughout the genome. In contrast, nucleic acids (DNA and RNA) and proteins capable of specifically recognizing DNA sequences can cause **targeted mutations** (directed mutagenesis and genome editing).

Point mutations consist of nucleotide replacements (substitutions) or ***additions**** or ***subtractions**** of a small number of nucleotides (indels), examples of which are shown on the short fragment of double-stranded DNA sequence. Substitutions that maintain the relative purine-pyrimidine positions between the two DNA strands are called ***transitions****, those that reverse these relative positions are called ***transversions****. For each pair of nucleotides in the DNA, there is therefore one transition and two possible transversions (non-exhaustive illustration).

Structural mutations involve DNA segments of varying sizes, that can reach thousands or millions of pairs of nucleotides. Depending on their type, they can keep (inversions, translocations) or not keep (deletions, duplications) the number of gene copies. All structural mutations generate new DNA sequences at their junctions. Several structural mutations can be entangled as a result of the same mutational event.

Box 1.3. *Mutations*

1.5. First ideas on gene products

Long before the material nature of genes was understood, the question obviously arose as to how they could confer a particular character to the organism. How did the genes act? The problem did not seem simple. However, as early as 1902, **Archibald Garrod**, who was studying a metabolic anomaly in humans (alkaptonuria), had found that it was transmitted as a recessive Mendelian trait between generations (Garrod 1902). He concluded that hereditary units were the cause of the chemical reactions that take place in the body and, as early as 1909, extended this idea to several other metabolic diseases of genetic origin.

This hypothesis of a one-to-one relationship between a gene and a product was confirmed experimentally about 30 years later. In a series of experiments with imaginal eye disc grafts between *Drosophila melanogaster* larvae of different mutant lines for eye color, **George Beadle** and **Boris Ephrussi** concluded that the normally reddish-brown color of the fly's eyes required the synthesis of three diffusible substances, which they called ca+, v+, cn+ (Beadle and Ephrussi 1936, 1937). Some mutants do not produce ca+, others not v+, others not cn+. The results of cross-grafting between the mutants indicated that the three substances are formed by the sequential series: → ca+ → v+ → cn+, the transition from one substance to the next representing one or several chemical reactions each involving a gene identified by its mutants.

Pursuing the same idea of a direct relationship between genes, metabolic reaction and enzyme, **G. Beadle** and **Edward Tatum** (Beadle and Tatum 1945) systematically produced ***auxotrophic**** mutants in the fungus *Neurospora crassa*, that is mutants unable to develop on the usual growth medium without the addition of a particular metabolite (***amino acid****,

vitamin, etc.) that they become unable to synthesize because of the mutations. Among these, some mutants deficient in the biosynthesis of nicotinic acid (vitamin B3) proved particularly interesting, because their hybrids (diploids formed by crossing haploid mutants) were able to grow without addition of nicotinic acid. They had become **prototrophic**. There had therefore been functional complementation between the mutations, indicating that these mutations, although conferring the same auxotrophy, did not affect the same gene. Each gene therefore participates in a particular stage of the metabolic pathway leading to the synthesis of this vitamin by directing the synthesis of one of the enzymes. Thus was born an essential concept of genetics summarized by the expression "one gene – one enzyme". We now know that this concept applies to many genes. For example, Mendel's wrinkled seed peas result from a mutation affecting the enzyme that connects sugars to starch polymers, resulting in the appearance of mature seeds when dried. Note, however, that while the enzymes of metabolism are proteins, not all proteins are enzymes and it is therefore preferable to say more generally "one gene – one protein", a central idea that led to the discovery of the genetic code. It is only much later that we understood the importance of other genes, those whose products are RNAs that do not code for proteins (see Chapters 2 and 5).

Let's recapitulate. After half a century of research, it was understood that genes were functional entities that direct the synthesis of products (proteins). Their nature was still unknown, but they were locatable along the chromosomes (loci), they mutated and were transmitted to the offspring according to Mendel's principles only slightly modified by Morgan. Nothing was expected to disturb this beautiful Swiss watchmaking! And yet, several observations seemed to contradict the laws of hereditary transmission established by Mendel, the first of which dating back to the very beginning of genetics. Was there, in addition, a non-Mendelian genetics whose principles remained to be discovered?

1.6. The order of things and the elements of disorder

An important discordant note came from the genetic study of the coloring of maize grains. In some varieties, these characteristics had intriguing behaviors. They seemed unstable, with the ears and even the grains themselves showing colored areas and others not, as if mutations appeared at

high frequency during the plant's development. Did that require attention? That's what Barbara McClintock did from the 1940s onwards. In addition to their abnormally high frequency, these mutations were also reversible with a high frequency, that is once they disappeared, the coloring could reappear and sometimes simultaneously with the mutation of another characteristic (such as the quality of the grain starch, for example). Worse, when attempting to map these mutations on factorial maps by crossing, according to the now classical method, no specific loci were found, as if there were genetic elements capable of moving along the chromosomes or even jumping from one chromosome to another. This is the hypothesis McClintock made, by postulating that these mysterious transposable elements were likely to affect the functioning of genes by activating them, inactivating them or even inducing chromosomal instabilities (McClintock 1950). The idea overturned many of the foundations of genetics, and it took all of the tenacity and foresight of McClintock to gain acceptance by beginning to elucidate the mechanisms responsible for these phenomena and, in particular, by showing the existence of distinct families of transposable elements, each with active autonomous elements and inactive mutated ones. Autonomous elements are able to move alone in the genomes, but they are also able to mobilize the mutated elements of their own family by providing them with the necessary machinery for their movement. This discovery later proved to be fundamental to understanding the evolution of genomes (see Chapter 6). Our own genome, like that of many animals or plants, is made up for the most part of transposable elements and their mutated traces.

In addition, from the very beginning of genetics, there were other observations that were incompatible with the Mendelian heredity. Correns himself, one of the three who re-discovered Mendel's laws, reported as early as 1904 that female plants of the summer savory (*Satureja hortense*) pollinated by hermaphrodites of the same species produced a 100% female offspring, thus showing a strictly maternal heredity. It is now known that this is an example of male cytoplasmic sterility that is quite common in plants. Five years later, the same author reported another case of maternal heredity. Yellow *Mirabilis jalapa* plants (a symptom of chlorophyll deficiency), pollinated by normally green plants, produced an entirely yellow progeny, while the reciprocal crossing gave an entirely green progeny. At the same time, **Erwin Baur** observed that the hybrids between a green *Pelargonium zonale* plant and a white branch of a white/green chimeric plant were randomly distributed between white, green or chimeric white/green plants. Far from Mendelian regularity! It was therefore suspected, as early as those

years, that hereditary factors existed in the cellular cytoplasm in addition to nuclear chromosomes, and Baur hypothesized that these cytoplasmic genetic factors were localized in the plastids (Hagemann 2000). Other similar cases, ranging from male sterilities to chlorophyll deficiency (see Chapter 3) were then described in many plant species (Hagemann 2010). In either case, it was found that some nuclear genotypes abolished these cytoplasmic effects, restoring male fertility (first case in maize) or chlorophyll function (oenothera). Phenomena including cytoplasmic heredity should therefore be considered more broadly as the result of interactions between nuclear and cytoplasmic factors.

The phenomena of non-Mendelian heredity are not limited to plants. In 1949, Ephrussi discovered a new type of mutation in baker's yeast, induced by acriflavin, which, in crosses with normal yeasts, segregated during the mitotic cell divisions (diploids) following the formation of ***zygotes**** (Ephrussi and Hottinguer 1950). These totally irreversible mutations were called "petites colonies" because, on suitable growth media, they greatly reduced the size of yeast colonies. During subsequent meiosis, normal diploid cells never showed any evidence of segregation of the mutant alleles. The mutation did not therefore concern the chromosomes. Again, this was a cytoplasmic heredity whose rules seemed much less clear than Mendel's laws. Other observations at the same time with mutants of the unicellular alga *Chlamydomonas reinhardtii* led to the same ideas, which had allowed Ephrussi to say, as a joke, that there were two kinds of genetics: nuclear genetics and unclear genetics. Since petite colony mutations led to deficiency of mitochondrial respiratory chain components (which explained their slower growth), the idea emerged that cytoplasmic genetic factors were localized in the mitochondria (see Chapter 3).

1.7. Dissecting the invisible: allelism, cistron and the locus again

As we have seen previously, the location of a gene along a chromosome deduced from genetic recombination tests is referred to as locus. But the relationship between locus and gene is not simple, because, based on an operational definition, the locus covers variable dimensions that depend on the sensitivity of the test, that is in practice on the number of individuals that can be studied in the offspring of crosses. A locus may therefore cover one gene (or even several neighboring genes) or, on the contrary, represent only a part of it. Mutations will be considered part of the same locus so long as no

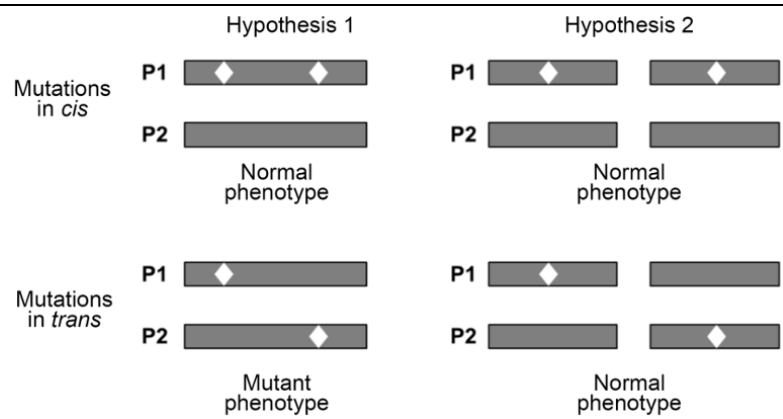
recombinants are observed between them. It is for this very reason that bacteriophages (bacterial viruses) have been essential to our understanding of the fine structure of the gene.

Bacteriophages were discovered in the years 1915–1917 by **Frederick Twort** and **Félix d'Hérelle**, who noticed that infectious elements for bacteria were able to pass through the porcelain filters used at the time to sterilize broths containing bacteria. They were therefore much smaller than bacteria, suggesting their viral nature. But it was not until the 1940s that they began to play a role in genetic research, when their growth, causing lysis of infected bacteria[6], began to be characterized on a quantitative basis (Ellis and Delbrück 1939). The limited phenotypic panoply offered by bacteriophages (in practice the mutations concerned the host spectrum and the shape of the lysis plaques) was largely compensated by their reproductive efficiency and the ease with which populations of considerable size, several orders of magnitude larger than what could be done with other organisms, could be handled. The discovery of genetic recombination in T2 and T4 bacteriophages by **Alfred Hershey** and his colleagues propelled these organisms to the forefront of the then emerging molecular genetics and allowed **Seymour Benzer**, a few years later, to elucidate the fine structure of the gene. With bacteriophages, the gene then becomes a set of loci, that is a set of mutations that can be separated from each other by recombination. At the extreme (see Chapter 2), the locus became a single nucleotide, because, with bacteriophages, two adjacent nucleotides along the DNA molecule could be separated by genetic recombination.

This analytical power led to a new fundamental notion of genetics, that of the ***cistron****. At this level of resolution, the locus was no longer able to define the limits of a gene with respect to its neighbors, because recombination ignores these limits. The interest of the cistron is that it defines the gene as a functional unit in the absence of precise knowledge of its exact nature. The word "cistron" is a neologism derived from the notions of ***cis**** and ***trans****, which define the two possible topologies between two mutated alleles in a cross. Defining a cistron simply requires the existence of

6 In practice, bacteria are cultured on the surface of an agar medium contained in a petri dish. They thus form a compact lawn on which each deposited bacteriophage will cause neighboring bacteria to lyse by its stepwise proximal infection and multiplication. The result will have the visible appearance of a more or less regular hole in the bacterial lawn, free of bacteria, called "lysis plaque".

recessive mutations relative to the normal allele and the production of hybrids by crossing, so that the normal function of the gene can appear. When two such mutations are brought in *trans* in the cross, two opposite results are obtained depending on whether the normal function of the gene reappears – we speak of "functional complementation", already mentioned – or not. In the first case, the two mutations are said to belong to two distinct cistrons, in the second case, they belong to the same cistron (see Box 1.4).



This is a functional test to determine whether two independent mutations at different ***loci**** (because they can recombine with each other) affect the same gene (hypothesis 1) or two distinct genes (hypothesis 2). Mutations are symbolized by white diamonds and genes by grey rectangles.

The test consists of comparing the phenotypes of hybrids resulting from two crosses that differ only by the parental origin of the two mutations. In the first case (*cis* position or coupling), both mutations are brought by the same parent (here P1). In the second case (*trans* position or repulsion), they are each brought by a different parent (P1 and P2).

In the *cis* position, the phenotype of the hybrid will be normal in both cases. This cross is used to verify that the mutations are recessive (an essential criterion for this test). In the *trans* position, the phenotype of the hybrid will be mutant in the first hypothesis (because there is no non-mutated copy of the gene), but normal in the second (because there is still one non-mutated copy of each gene). In the latter case, we speak of functional complementation between the two mutations.

We call **cistron** the set of recessive mutations which, brought in *trans* relative to one another, are incapable of functional complementation between

them. The cistron is therefore a functional definition of the gene. This test was initially developed with the fungus *Neurospora crassa* and then used in many organisms, including the bacteriophage T4. As an example, a child will be affected by a syndrome if the gametes of his/her two parents bring harmful recessive mutations in the same cistron, not in the case in which the mutations concern different cistrons.

Today, the *cis-trans* test has given way to other methods, but the term "cistron" has been retained to differentiate *monocistronic* transcripts, which correspond to a single gene, from *polycistronic* transcripts, which correspond to several successive genes along a chromosome.

Box 1.4. *Functional cis-trans test and cistrons*

The cistron is therefore the set of all recessive mutations which, brought in *trans* in hybrids, are incapable of functional complementation between them. It is therefore still an operational definition with a varying degree of resolution. With bacteriophages, the resolution is maximum. Finally, the relationship between cistron and locus, which is complex by nature, will be all the more precise – and therefore the gene better defined – if the number of mutations studied and the number of descendants analyzed are large. We understand that genetics is the art of balancing the desirable and the achievable.

1.8. The DNA trail

Let's summarize again our vision of genetics at this stage. Genes are functional entities, **cistrons**, independent of one another, and each producing a protein (or RNA). Genes are located along the chromosomes, each at a specific place, the **locus**. As a result, they follow the chromosomes during cell divisions (mitosis and meiosis), resulting results in their regular distribution among daughter cells and thus in the subsequent generations, as Mendel had already understood. But not all genes are on nuclear chromosomes, so there are genetic traits that do not follow Mendelian segregation over generations. In addition, there are other genetic elements that, unlike genes, move as they please from locus to locus, and can interfere with the functioning of genes. Finally, genes can mutate, forming different alleles, spontaneously or under the action of physical or chemical agents. So, if genes are of material nature, what are they made of?

It has been known since **Friedrich Miescher's** analyses in 1869 that chromosomes consist of proteins (basically charged) and a strong acid, originally called nuclein (now deoxyribonucleic acid, DNA), whose invariant chemical properties did not match the expected diversity of genes because the phenotypic traits they control are so varied (Dham 2005). Would genes therefore have to be sought among proteins whose biochemical properties are very variable? This simplistic reasoning did not withstand rigorous experimentation. But this experimentation did not directly derive from genetics. Indeed, it will not have escaped the reader's attention that genetics has developed with hybridization experiments carried out mainly on plants, insects, fungi, sometimes mice, but almost nothing else.

However, the bulk of the diversity of the living world is represented by microorganisms and genetics had practically ignored them. Except for experiments that gave strange results, such as **Frederick Griffiths'** early experiments (Griffiths 1928) on pneumococci that led to the discovery of bacterial ***transformation****, a phenomenon that would later play an essential role in the development of genetic engineering (see Chapter 4). He discovered that by simultaneously injecting mice with harmless R-form pneumococci (R for rough colonies) and virulent S-form pneumococci (S for smooth colonies) killed by heat treatment, the mice died from pneumonia following the proliferation of S-form pneumococci. The dead S-type bacteria had therefore transferred something that had transformed the living R-type bacteria into S-type. So this something had a genetic power. It was designated "transforming factor". It was not until 1944 that this experiment was repeated by **Oswald Avery's** group, this time by treating the transforming factor with protein-specific degradation enzymes (proteases) or DNA degradation enzymes (deoxyribonucleases). After multiple controls, the result was clear, Griffiths' transforming factor was DNA, and nothing else (Avery *et al.* 1944). So the genes were made of DNA. In the same year, **Erwin Schrödinger** published his now historic book, *What is Life?* (Schrödinger 1944), in which, applying the laws of physics to what was known about genetics, he was able to predict that the maximum size of the gene (which he still believed was made of proteins) was between one and a few million atoms. The result is perfectly relevant if we put it back on the DNA: one million atoms corresponds to about 15,000 base pairs, which is not far from an "average gene" of an animal or plant.

1.9. Important ideas to remember

– Genes are the elementary **units** of heredity that are transmitted from generation to generation according to precise rules.

– In a population, there are multiple forms of the same gene, called **alleles**, originating from **mutations**.

– The functional genetic unit is called **cistron**. Formally, the cistron is defined by the set of mutations that are incapable of functional complementation between them.

– Genes are located along the chromosomes at fixed locations, called **loci**.

– The different loci can be placed relative to one another on **genetic maps** by counting the recombinants following crosses between individuals carrying different alleles at these loci.

– There are mobile genetic elements, called **transposable elements** that spontaneously change their location in chromosomes.

– There are **mutations** called "structural", which alter the order or number of the different genes along chromosomes without directly changing them.

1.10. References

Avery, O., MacLeod, C.M., McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medicine*, 79, 137–158.

Bateson, W., Saunders, E.R., Punnett, R.C. (1906). Further experiments on inheritance in sweet peas and stocks: preliminary account. *Proceedings of the Royal Society B*, 77(517), 236–238.

Beadle, G.W., Ephrussi, B. (1936). The differentiation of eye pigments in Drosophila as studied by transplantation. *Genetics*, 21, 225–247.

Beadle, G.W., Ephrussi, B. (1937). Development of eye colors in Drosophila: diffusible substances and their interrelations. *Genetics*, 22, 76–86.

Beadle, G.W., Tatum, E.L. (1945). Neurospora. II. Methods of producing and detecting mutations concerned with nutritional requirements. *American Journal of Botany*, 32, 678–686.

Blixt, S. (1975). Why didn't Gregor Mendel find linkage? *Nature*, 256, 206.

Campbell, M. (1980). Did de Vries discover the law of segregation independently? *Annals of Science*, 37(6), 639–655.

Creighton, H., McClintock, B. (1931). A correlation of cytological and genetical crossing-over in *Zea mays*. *Proceedings of the National Academy of Sciences*, 17, 492–497.

Cuénot, L. (1902). La loi de Mendel et l'hérédité de la pigmentation du pelage chez les souris. *Archives de zoologie expérimentale et générale*, 10, 27–30.

Davenport, C.B. (1904). Wonder horses and Mendelism. *Science, New Series*, 19(473), 151–153.

Dham, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278, 274–288.

Ellis, E.L., Delbrück, M. (1939). The growth of bacteriophage. *The Journal of General Physiology*, 22, 365–384.

Ephrussi, B., Hottinguer, H. (1950). Direct demonstration of the mutagenic action of euflavine on baker's yeast. *Nature*, 166, 956.

Flemming, W. (1879). Beiträge zur Kenntniss der Zelle und ihrer Lebenserscheinungen. *Archiv für mikroskopische Anatomie*, 16, 302–436.

Gager, C.S., Blakeslee, A.F. (1927). Chromosome and gene mutations in *Datura* following exposure to radium rays. *Proceedings of the National Academy of Sciences*, 13, 75–79.

Garrod, A.E. (1902). The incidence of alkaptonuria: A study in chemical individuality. *The Lancet*, 2, 1616–1620.

Gayon, J. (2016). From Mendel to epigenetics: History of genetics. *Comptes Rendus Biologies*, 339, 225–230.

Griffiths, F. (1928). The significance of pneumococcal types. *Journal of Hygiene,* 27, 113–159.

Hagemann, R. (2000). Erwin Baur or Carl Correns: Who created the theory of plastid inheritance? *Journal of Heredity*, 91, 435–440.

Hagemann, R. (2010). The foundation of extranuclear inheritance: Plastid and mitochondrial genetics. *Molecular Genetics and Genomics*, 283, 199–209.

Lejeune, J., Gauthier, M., Turpin, R. (1959). Les chromosomes humains en culture de tissus. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 248, 602–603.

Lutz, A.M. (1907). A preliminary note on the chromosomes of *Oenothera Lamarckiana* and one of its mutants. *Oxford Academy GigasScience*, 26, 151–152.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36, 344–355.

Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 4, 3–47.

Moreau de Maupertuis, P.L. (1752); *Lettre sur le progrès des sciences*. George Conrad Walther, Berlin.

Morgan, H.T., Sturtevant, A.H., Muller, H.J., Bridges, C.B. (1915). *The Mechanism of Mendelian Heredity*. Henry Holt & Company, New York.

Muller, H.J. (1927). Artificial transmutation of the gene. *Science*, 66, 84–87.

Schrödinger, E. (1944). *What is Life?* Trinity College, Dublin.

Stern, C. (1936). Somatic Crossing Over and Segregation in *Drosophila melanogaster*. *Genetics*, 21, 625–730.

Strasburger, E. (1875). *Über Zellbildung und Zelltheilung im Pflanzenreiche*. Hermann Dabis, Iéna.

Sutton, W.S. (1903). The chromosomes in heredity. *The Biological Bulletin*, 4, 231–251.

Van Beneden, E. (1883). Recherches sur la maturation de l'œuf et la fécondation. *Archives de biologie*, 4, 265–640.

de Vries, H. (1889). *Intracelluläre Pangenesis*. Fischer, Iéna.

de Vries, H. (1901–1903). *Die Mutationstheorie*. Von Veit and Co., Leipzig.

Waldeyer, W. (1888). Über Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen. *Archiv für mikroskopische Anatomie*, 32, 1–122.

Wanscher, J.H. (1975). The history of Wilhelm Johannsen's genetical terms and concepts from the period 1903 to 1926. *Centaurus*, 19(2), 125–147.

Wilson, E.B. (1905). The chromosomes in relation to the determination of sex in insects. *Science*, 22, 500–502.

2

The Molecular Nature of Genes and Their Products

2.1. DNA and its replication

How can we understand that a chemically monotonous molecule like DNA is the molecular material for genes whose diversity seems to be infinite? The question seemed difficult. But biochemistry and crystallography have provided essential aids to genetics. In 1950, a chemist named **Erwin Chargaff** analyzed the proportions of the four nitrogenous bases used in the composition of DNA – adenine (A), thymine (T), guanine (G) and cytosine (C) – and observed a perfect chemical equality between A and T on the one hand and between G and C on the other hand (Chargaff *et al.* 1950). This was true for all the organisms studied, while the ratio (A+T)/(G+C) varied widely from one species to another. This compositional variation offered a possible explanation for the role of DNA as a carrier of genetic information, even if we were still far from understanding how.

It was Chargaff's rule, combined with the analysis of X-ray diffraction images of DNA crystals obtained by **Rosalind Franklin** and **Maurice Wilkins**, that led to the discovery of the "double helix" structure of DNA by **James Watson** and **Francis Crick** only three years later (Watson and Crick 1953a). In this model, two strands composed of molecules of ***deoxyribose**** (5 carbon sugar) chemically linked together by phosphate groups and each carrying one of the four nitrogenous bases, form a right-handed double helix (see Box 2.2). The two strands are joined together, because each base of one strand is linked by hydrogen bonds to the complementary base of the other strand, A with T and C with G, explaining Chargaff's rule.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

Such a molecule has two emerging properties, fundamental for genetics.

The first is to convey information represented by the succession of bases along the strands, now called the sequence. The amount of information thus carried is truly vertiginous, because the number of sequence combinations is 4^n, if n is the number of bases. For only 15 bases (one and a half turns of the DNA double helix), there are already about one billion combinations, for 30, one billion billion, and so on. For a gene of the size estimated by E. Schrödinger, the number greatly exceed the number of elementary particles that must exist in the whole universe. Life is therefore based on a considerable excess of information theoretically available.

The second property of DNA results from the complementarity of the bases between the two strands. This complementarity provides an immediate explanation for the duplication of genes needed before each cell division, as immediately understood by Watson and Crick. They hypothesized that if the two strands separate by breaking the hydrogen bonds that are low in energy, each of them can serve as a matrix for the synthesis of its complementary strand, producing two molecules that are identical to each other and identical to the original molecule (Watson and Crick 1953b). It was not yet known how this synthesis could be achieved, but biochemistry elucidated this problem, fully confirming the initial hypothesis, with the discovery of DNA polymerases by **Arthur Kornberg**.

2.2. Permanence and alteration of DNA, mutations

The idea of perpetuating genes by successive doublings of DNA molecules also explained another fundamental principle of genetics, the appearance of mutations. At the thermodynamic level, since the complementarity of the bases is ensured by the formation of hydrogen bonds, the process of synthesizing a new DNA strand cannot be free of a certain probability of error. Consequently, with a low frequency that depends on the physical-chemical conditions in which DNA replication takes place in the cell, the newly synthesized strands will not always be perfectly complementary to their matrix strands. DNA replication is a complex biochemical mechanism involving many enzymes and it is not necessary to detail it, except to say that it directly affects the error rate. However, it should be noted that the two daughter molecules resulting from DNA replication each carry an old strand, that of the mother molecule that served

as a matrix, and a new strand, the one that has just been synthesized. The replication errors will therefore affect this last strand. This allows living cells to correct these errors by using the old strand as a model to repair the new strand when there are marks to distinguish the two strands (post-replicative methylation of some bases). The molecular mechanisms involved in these sophisticated processes have emerged during evolution. They reduce mutation rates, but still have their limitations and unrepaired replication errors will result in point mutations in the next generation (when the new DNA strand will in turn serve as a matrix for the synthesis of its complementary strand).

The point mutation rate varies according to the organisms, their genetic heritage (more or less effective DNA repair mechanisms), but also according to living conditions (abundance or limitation of nutrients, presence of toxic elements, etc.) and the age of individuals. For example, the mutation rate of yeast under normal growth conditions is 3×10^{-10} per nucleotide per cell generation. In other words, each nucleotide has a chance of about one in three billion of being mutated with each DNA replication. This seems very little, but the yeast genome has 13 million nucleotides. At each cell division, one yeast in 200 will therefore carry a mutation. And in a volume of beer limited to the size of a glass, there are about 30 million mutated yeasts (out of a total population of about one billion cells), which is 5,000 times more than the number of genes in this yeast. The same calculations apply to humans and plants. With mutation rates comparable to those of yeast, about 40 new mutations are expected to appear in the genome of each newborn compared to the previous generation, which is exactly what is observed by full genome sequencing (see Chapter 4). Comparable numbers are found in each seed in plants.

Point mutations are not limited to DNA sequence changes resulting from replication errors. Without replicating, DNA molecules can undergo spontaneous chemical alterations (deamination, depurination) or induced alterations by exposure to chemical mutagens, electromagnetic or radioactive radiations, as already mentioned in the previous chapter, or simply by the action of normal cellular metabolism (oxidative radicals for example). If not repaired, these alterations will lead to mutations if the DNA molecules concerned give rise to offspring. By comparing the DNA sequences of a daughter cell with its parent, point mutations fall into two categories: nucleotide ***substitutions****, that is the replacement of one nucleotide by another, and insertions/deletions or ***indels**** affecting a single nucleotide or a

very small number of adjacent nucleotides. Substitutions and indels are responsible for the genome sequence polymorphism observed in all natural populations (see Chapter 5), with indels generally being less frequent than substitutions (by about an order of magnitude). Depending on their nature and their precise location in genomes, these mutations will have very different genetic consequences. In some cases, the substitution of a single specific nucleotide may result in immediate cell death due to the absence of the product of an essential gene. These mutations are obviously never found in natural populations. In other cases, mutations may affect the activity of a gene (by activating or inactivating it) without modifying its products or, on the contrary, modify its products more or less significantly. Finally, in the most frequent cases, mutations will not change anything at all; they will only increase the genetic polymorphism of populations. These are called neutral mutations to differentiate them from those with positive or negative phenotypic effects.

As discussed in Chapter 6 and illustrated in Box 1.3, mutations are not limited to the DNA sequence changes already discussed. In addition to point mutations, there is a whole range of so-called structural mutations, because they affect entire segments of chromosomes by rearranging them in different ways relative to each other or by altering the number of their copies without modifying the DNA sequences of these segments (except at junctions). Copy number alterations may concern particular genes or whole chromosomes. Copper resistance in yeast or trisomy 21 in humans are respective examples.

2.3. Protein synthesis and the central dogma of molecular biology

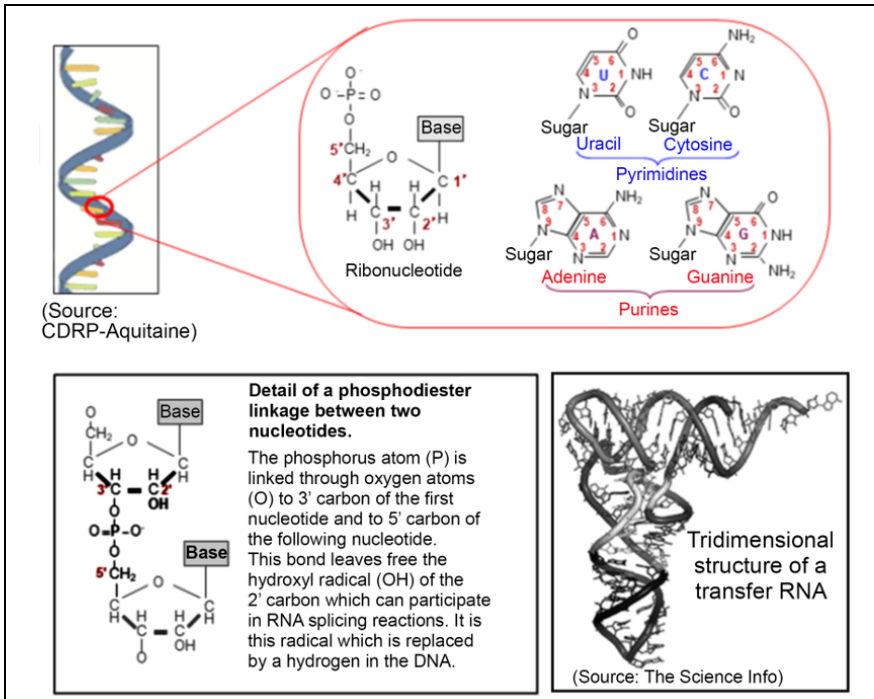
Let's leave the propagation of genes for a moment and return to their function. Before the discovery of the role of DNA, genetics had taught us that a gene corresponds to a protein and, although the reality is more complex as we will see later, this relationship whose mechanism remained mysterious played a central role in the emergence of molecular genetics. It was known that proteins were made up of amino acids: 20 molecular species in total, common to all living beings (22 in reality, if we take into account selenocysteine and pyrrolysine, two rare amino acids used by only a few organisms). The first major discovery was to understand that these amino acids are linked to each other by covalent chemical bonds called **peptide bonds**. A chain of amino acids forms a ***polypetide****. Each protein,

depending on its nature, consists of a single polypeptide or several. Hemoglobin in vertebrate blood, for example, consists of a total of four polypeptides, two of each kind, while myoglobin, its equivalent in muscles, consists of only one polypeptide.

It was by studying a relatively simple protein secreted by the pancreas, insulin, that **Frederick Sanger** and his colleagues first succeeded in determining the order of amino acids in a polypeptidic chain, its ***sequence**** (see Chapter 4). Insulin consists of two short polypeptides, an A chain and a B chain. The sequence of the B chain was described in 1951 (Sanger and Tuppy 1951), that of the A chain two years later (Sanger and Thompson 1953). Each chain has its own amino acid sequence. It should be noted that the theoretical combinatorics of amino acid sequences is even more gigantic than that of DNA, because it is a power of the number 20. There are more than ten trillion different sequences of only 10 amino acids long and, most natural polypeptide chains are generally made of hundreds of amino acids. Each polypeptide being a specific amino acid sequence, it was tempting to hypothesize that this sequence was related to the sequence of nucleotides in the corresponding gene. But how? Everything remained to be discovered about the mechanisms involved and we were very far from being able to determine the nucleotide sequences of DNA.

Several arguments suggested the existence of intermediate actors between DNA and proteins. First, protein synthesis does not occur in the nucleus of eukaryotic cells where DNA is found, but in their cytoplasm. There, the nascent proteins are associated with structures then called "microsomes", composed of proteins and RNA (for this reason, microsomes became ***ribosomes****). Therefore, some molecules must serve as intermediaries between the DNA of the nucleus and the site of protein synthesis in the cytoplasm. RNAs were good candidates for carrying genetic information, because some viruses such as tobacco mosaic virus, were known whose infectious particles are composed of a protein enveloping an RNA molecule, suggesting that this RNA was the genetic material of the virus. Like DNA, RNAs are polymers of phosphorylated sugars carrying nitrogen bases, but the sugar is ribose and uracil replaces thymine[1] (see Boxes 2.1 and 2.2). RNAs do not follow Chargaff's rule, they are molecules made up of a single strand. As many RNAs are unstable molecules, their study was very difficult at that time. We were limited to cellular extracts.

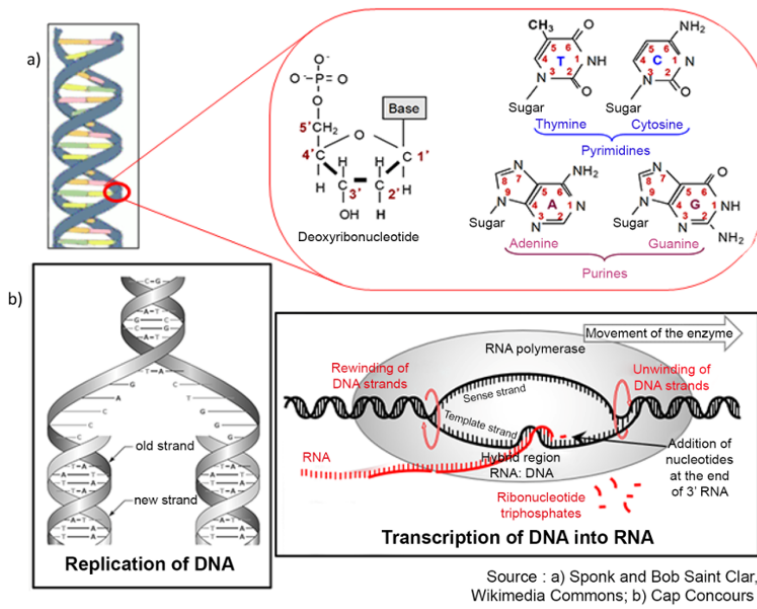
1 Thymine is a 5-methyl uracil.



An **RNA** molecule (top left) consists of a chain of polymerized nucleotides. Each **nucleotide** is formed by a sugar, the **ribose** carrying a phosphate group on its 5′ carbon and one of the four nitrogenous bases (purine or pyrimidine) on its 1′ carbon (bubble at the top right). Polymerization is carried out by covalent linkage of the phosphate group with the hydroxyl (OH) radical carried by the 3′ carbon of the previous nucleotide (**phosphodiester bond**, box above left). The RNA chain is therefore polarized from its 5′ end to its 3′ end. The succession of nucleotides along this chain constitutes the RNA **sequence**.

RNA molecules can fold back on themselves (or combine with other RNA molecules) to form helices held together by hydrogen bonds between complementary bases (A with U, G with C or U) when the sequences allow it. Other chemical interactions (Hoogsteen bonds that involve the face of purines opposite to ribose, for example) also contribute to the three-dimensional structure of RNA molecules (example of a **transfer RNA**, box above, right). Note that some nucleotides of some RNA molecules may undergo various specific chemical modifications after RNA synthesis.

Box 2.1. *Nucleic acids 1: RNA. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*



Source : a) Sponk and Bob Saint Clar, Wikimedia Commons; b) Cap Concours

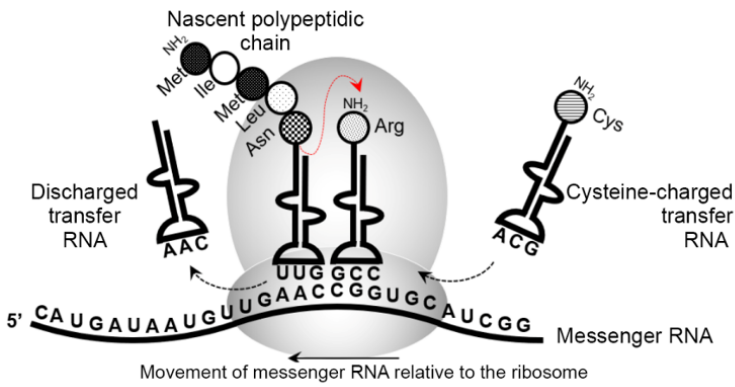
The **DNA** double helix (top left, here in its B form) consists of two chains of nucleotides polymerized in reverse orientation with respect to each other and maintained together by hydrogen bonds between the complementary bases (A with T, G with C). Each **nucleotide** is formed by a sugar bearing a phosphate group on its 5′ carbon and one of the four nitrogenous bases (purine or pyrimidine) on its 1′ carbon as for RNA, but the sugar here is **deoxyribose**. The succession of nucleotides along the chains is ensured by phosphodiester bonds as for RNA. The double helix has a diameter of 2 nanometers. Its length can reach centimeters.

Note the presence of hydrogen instead of hydroxyl group on the 2′ carbon of deoxyribose, which makes DNA molecules chemically less reactive than RNA molecules. It is one of the secrets of the origin of the genetic material of presently living organisms. Also note that uracil is replaced by 5-methyl-uracil, called thymine.

The sequence of one of the chains fully determines that of the other by the complementarity of the bases: A and T form two hydrogen bonds, G and C form three hydrogen bonds, which allows DNA **replication** (box above, left). Similarly, the sequence of one of the DNA chains determines that of a nascent RNA chain during the **transcription** process (rectangle above, right).

Box 2.2. *Nucleic acids 2: DNA. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

It was the chemists **Mahlon Hoagland** and **Paul Zamecnik** who, with their colleagues, were the first to reconstitute protein synthesis *in vitro* from rat liver extracts (Hoagland *et al*. 1958). They followed the synthesis by incorporating carbon-14-labeled amino acids into the protein fraction of these extracts. To function, this *in vitro* system had to include the 20 proteinogenic amino acids, microsomes, ATP (adenosine triphosphate, the molecule that provides energy) and a complex fraction, but known to be DNA-free. In particular, this fraction contained low molecular weight RNAs that, strangely enough, became chemically bound to the labeled amino acids. These RNAs were essential for protein synthesis to take place.



The figure shows a ribosome (gray solid) in the process of protein synthesis. **Ribosomes** are made up of two subunits (of different mass) each composed of a large RNA molecule (thousands of nucleotides) and specific proteins. Depending on the organisms, one (bacteria) or two (eukaryotes) additional small molecules of RNA (hundreds of nucleotides) exist in the large subunit. The **messenger RNA** molecule is symbolized by the thick line on which is represented the succession of nucleotides (A, C, G and U). **Amino acids** (patterned circles) are chemically bound to **transfer RNA** molecules (thick wavy lines). There are generally about 40 species of transfer RNA (their number varies slightly from one organism to another). Each transfer RNA species carries only one type of amino acid, but two (sometimes three) different transfer RNA species can carry the same type of amino acid, following the genetic code degeneracy rules (see Box 2.5).

Protein synthesis is a sequential dynamic process. When a **triplet** of messenger RNA **nucleotides** is correctly placed in the small ribosome subunit, the amino acid-charged transfer RNAs penetrate one by one into the **accepting site** of the large subunit from which they immediately exit (molecular agitation is

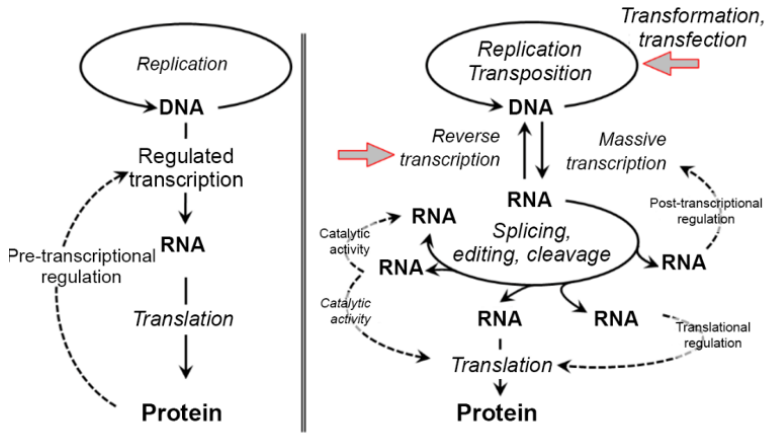
intense at normal life temperatures), unless the three nucleotides of their **anticodon** are complementary to the three nucleotides of the messenger RNA **codon**. In this case, a new **peptide** bond will be formed by transfer (red dotted arrow) of the nascent peptide chain that was linked to the previous transfer RNA (previous codon) on the amino group (NH_2) of the amino acid carried by the new transfer RNA. The previous transfer RNA, now unbound to an amino acid, is expelled and replaced by the new transfer RNA linked to the nascent peptide chain, freeing the ribosome acceptor site for the next codon and allowing the cycle to start again.

This complex mechanism is guided and catalyzed by a precise chemistry of RNA molecules accompanied by various specialized proteins. The specificity of charge of each species of transfer RNA for a given amino acid is ensured by enzymes called "tRNA-synthetases", or which there are 20 (one per amino acid).

Box 2.3. *Protein synthesis. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

At the same time, for theoretical reasons, Crick hypothesized that small RNA molecules should act as adapters for each of the twenty amino acids used in protein composition, each adapter being specific for an amino acid and capable of recognizing a nucleic acid sequence by complementarity of bases. It was imagined that specific enzymes were catalyzing the chemical binding of the adapter to the amino acid. The two approaches therefore converged towards the existence of a set of RNA molecules called "solubles" which will then be called ***transfer RNA**** (see Box 2.3). In reality, the soluble RNA fractions also contained many other molecules unrelated to protein synthesis.

From this stage, it became possible to propose a general scheme for the transfer of information from genes to proteins through RNA intermediaries. This scheme became the central dogma of molecular biology (see Box 2.4). In this scheme, the DNA was copied indefinitely by replication and was not susceptible to receiving any external information. On the other hand, DNA transmitted its information to RNA molecules which, through a then unknown mechanism, subsequently transmitted it to proteins. These RNA molecules were neither transfer RNAs nor ribosomal RNAs because their constant compositions are incompatible with the synthesis of thousands of different proteins in an organism. What were they and how was their synthesis from DNA decided upon?



In its initial version (left), the **central dogma** of molecular biology represented the unidirectional and direct transfer of genetic information (solid arrows) from DNA to proteins via RNA as a simple intermediate. RNA production through DNA transcription was considered to be regulated by specific proteins acting on operators associated with the promoters (dotted arrow). We imagined a precise correspondence between RNA molecules and genes. No genetic information was supposed to go from RNA to DNA.

In its current version (right), DNA is transcribed into RNA in a global way (with some possible regulations, however), but without a precise correspondence with genes. Once synthesized, most RNA molecules are quickly destroyed. Some undergo or catalyze various transformations (cleavage, splicing, edition) that lead them to a wide variety of functions (catalytic, informational or regulatory, dotted arrows).

Only a small fraction of the various RNA molecules produced will become messenger RNAs for protein synthesis. The transfer of genetic information from DNA to gene products is therefore no longer as direct as originally imagined, but it remains entirely **genetically determined**, because RNA transformations depend directly on their sequences, even when they involve the interaction of several distinct RNA molecules.

DNA can receive and incorporate new genetic information (red arrows) from RNA molecules inside the organism (reverse transcription) or outside the organism (virus, transfection) or from foreign DNA fragments (horizontal transfer, transformation). This **randomly** acquired information will then become hereditary.

Box 2.4. *The central dogma of molecular biology revisited*

By using bacterial conjugation[2], **François Jacob** and **Jacques Monod** solved part of the mystery. They found that, after introduction of a particular gene into a bacterium that did not have it, the synthesis of the corresponding protein began almost instantly. This was incompatible with the stability of the transfer RNA and ribosomal RNA molecules and suggested the existence of another type of shorter-lived RNA. These hypothetical molecules were named ***messenger RNA****, because they were supposed to bring the genetic information from DNA to ribosomes. In the same year 1961, **François Gros** and his colleagues, in Watson's laboratory, detected this RNA *in vivo* in cultures of the *Escherichia coli* bacterium (Gros *et al*. 1961). The pattern was thus clarified: the DNA of the gene is transcribed into a messenger RNA molecule, which introduces itself into the ribosomes where charged transfer RNAs are successively recruited in an order defined by the messenger RNA sequence, delivering amino acids which become chemically linked to each other in their order of arrival. But what correspondence exists between the nucleotide sequence of the messenger RNA (supposedly identical to that of the gene) and the amino acid sequence of the new synthesized polypeptide? This is where the genetic code comes in.

2.4. The genetic code: how to read the genetic message

As we have already mentioned in the introduction, it is essential to distinguish between two very different concepts, often confused in common language: that of **genetic code** and that of **genetic message**. The genetic message is the set of all instructions that each living cell (or organism) receives from its ancestors and transmits to its descendants, if any. These instructions are interpreted by complex mechanisms that involve various RNA molecules. The genetic code, on the other hand, is a set of extremely precise rules of interpretation that only concern certain instructions, those carried by a particular class of RNA molecules, the messenger RNAs, when they are involved in protein synthesis. The same rules apply to all messenger RNAs regardless of the gene from which they originate and, therefore, regardless of the message they carry. These are rules for translating a

2 Bacteria do not have sexual reproduction like eukaryotes. Their reproduction is strictly clonal. But some of them, under the action of particular plasmids, can inject part of their chromosome into another bacterial cell of the same species. This is called bacterial conjugation or, sometimes, parasexuality. The bacterium that receives the foreign DNA fragment is called a merodiploid.

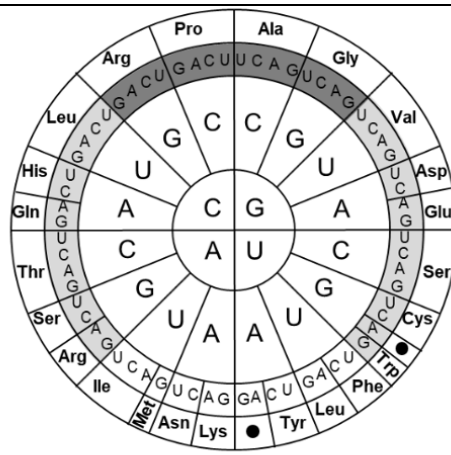
message, not the message itself. But what are these rules? Contrary to what many anticipated to be an enormously complex problem, decoding the genetic code was a matter of only five years, from 1961 to 1966, mainly carried out in the laboratories of **Gobind Khorana, Marshall Nirenberg** and **Robert Holley**, three American biochemists who developed complementary methods.

The ideas were guided by purely genetic work by Crick, **Sydney Brenner** and their colleagues, published in 1961 (Crick *et al.* 1961). They analyzed specific mutations of an *E. coli* bacteriophage (called T4) obtained by the action of a chemical mutagen that, by interposing itself between the nucleotides of the DNA, induces indel mutations by adding or subtracting a nucleotide. A large number of such mutations had been isolated and mapped in the same cistron of the bacteriophage (see Chapter 1), all mutants being unable to infect a particular strain of *E. coli* due to the lack of a necessary protein (while remaining able to infect another strain, otherwise the experiment would obviously not have been feasible). By associating different mutations between them the spectacular result occurred. Three mutations, provided they were of the same sign (addition or subtraction) and sufficiently close to each other, gave the bacteriophage a normal phenotype. In other words, a sequence shift of one or two nucleotides leads to the mutant phenotype, but a shift of three nucleotides allows the reading of the messenger RNA to "fall back into place" or recover. Protein synthesis must therefore be carried out by triplets of nucleotides. On a theoretical level, this result made sense. Nucleotide doublets only allow $4 \times 4 = 16$ possibilities, which is not enough for an unambiguous code since there are 20 proteinogenic amino acids. With triplets, combinatorics gives 64 possibilities, more than necessary. It should be noted that, as early as the 1950s, **George Gamow** had already reached this same conclusion. But how are triplets and amino-acids linked together?

Biochemistry answered this question. Nirenberg's method consisted of chemically producing synthetic polyribonucleotides, placing them in the presence of protein synthesis extracts *in vitro* with one of the twenty amino acids labeled with a radioactive atom, and then observing the presence of radioactivity in the synthesized protein fraction. But the synthesis of sufficiently long polyribonucleotides was difficult. The simplest to achieve were polyuridine (UUUUU...) or polyadenosine (AAAAA...) or mixtures of two nucleotides. Not all triplets could be examined in this way. An

alternative method, developed by Khorana, consisted of using short synthetic polyribonucleotides to directly capture the corresponding transfer RNAs (by an ultrafiltration system in the presence of cell extracts). By repeating the experiment 20 times for each triplet with a different radioactive amino acid each time, one could deduce which one corresponded to the triplet. It is by combining these two methods that the meaning of each of the triplets was established in 1966 (Nirenberg *et al*. 1966).

Let us pause for a moment on the properties of this code, because they inform us about the fundamental mechanisms of the living world, derived from its origin (see Box 2.5). The genetic code is *unambiguous*, that is, each of the 64 possible triplets, known as ***codons****, has a unique meaning. In the original version of the genetic code, the one that has proven to be the most widely used by present living organisms and is therefore called the "universal code", 61 codons correspond to amino acids. They are called sense codons. The other three (UAA, UAG and UGA), called non-sense or stop codons, signal the end of peptide chain synthesis. The excess of sense codons over the 20 proteinogenic amino acids (21 or 22 taking into account selenocysteine and pyrrolysine) is solved by the fact that several different codons may correspond to the same amino acid. We are talking about "synonymous" codons. There are, for example, two synonymous codons for cysteine, four for valine and even up to six for leucine, arginine or serine. The genetic code therefore seems to be *degenerate*. In reality, only the third nucleotide of codons is degenerate among the synonymous codon families. One possible hypothesis is that at the origin of life, only the first two nucleotides of each codon defined their meaning, the third one intervening only as a kind of spacer. According to this hypothesis, families of four codons meaning the same amino acid could have been the first to be stabilized. The influence of the third nucleotide in codon discrimination would only have developed as new transfer RNA species appeared with their particular chemical modifications. Unambiguity and degeneration combined mean that a given nucleotide sequence will always correspond to a single amino acid sequence in a polypeptide, while the opposite is not true. The same amino acid sequence may result from several (often many) different nucleotide sequences. This property is fundamental for the evolution of genomes (see Chapter 5).



Ala: L-Alanine
Arg: L-Arginine
Asn : L-Asparagine
Asp : L-Aspartic acid
Cys: L-Cysteine
Gln: L-Glutamine
Glu: L-Glutamic acid
Gly : Glycine
His: L-Histidine
Ile: L-Isoleucine
Leu: L-Leucine
Lys: L-Lysine
Met: L-Methionine
Phe: L-Phenylalanine
Pro: L-Proline
Ser: L-Serine
Thr : L-Threonine
Trp: L-Tryptophan
Tyr : L-Tyrosine
Val : L-Valine

The genetic code **matches** each of the 64 possible nucleotide triplets in messenger RNA (**codons**) to one of the 20 **amino acids** used in protein composition (list in box). Note that only the ***L-enantiomer**** is used.

Shown is the most recent representation of the genetic code (Grosjean and Westhof 2016) emphasizing the thermodynamic stability of **codon-anticodon pairings** (most stable dark gray, least stable white, intermediate light gray). The 64 codons are read from the center (first base) to the periphery of the circle (second, then third base of each codon). The corresponding amino acids are indicated in the outer circle according to their three-letter abbreviation. For example, the AUG codon corresponds to a methionine (Met), the GAA and GAG codons correspond to glutamic acid (Glu). The black dots indicate the **stop codons** (stopping protein synthesis), also called **non-sense** codons. There are three stop codons in the universal genetic code, UAA, UAG and UGA.

Shown is the so-called "universal" code used by almost all presently living organisms. Occasional variations exist in some phylums. For example, the UGA codon corresponds to a tryptophan in some bacterial species and in mitochondria. In some organisms or tissues, two other amino acids are used in the composition of some proteins, pyrrolysine and selenocysteine. They are coded by the codons UAG and UGA respectively.

Box 2.5. *The genetic code*

An essential point of the genetic code is that, in messenger RNAs, codons follow one another without overlap or intervals between them (except in very specific cases), thus creating a ***reading frame**** of base three. The absence of physical delimitation between the successive triplets means that the same sequence of messenger RNA can, in theory, be translated in three different ways depending on the **reading phase** used. In reality, the protein synthesis machinery uses particular signals in the messenger RNA sequence to start translation, which defines the phase to be used. It should be noted that it is generally easy to predict this phase from the sequence of a messenger RNA, because this is the one where there is a long ***open reading frame**** (ORF), that is the interval between two successive stop codons. In the other two phases, except in special cases, the intervals between successive stop codons are short (they should be on average 21 codons long for a random sequence of unbiased nucleotide composition, since there are three stop codons out of a total of 64).

The correspondence between codon and amino acid supposes that each codon is specifically recognized by the transfer RNA that carries this amino acid during protein synthesis. Transfer RNA molecules have a three-dimensional structure made up of single-stranded "loops" separated by double-stranded "stems" formed by the pairing of short complementary sequences within the RNA molecule. One of these loops contains the ***anticodon****, that is the three complementary bases of the codon, which will ensure its pairing during protein synthesis (see Box 2.3). Codon-anticodon pairings are defined by RNA thermodynamics and it would take too long to go into detail. It is sufficient to know that there are about 40 distinct species of transfer RNA in each living cell[3], each being specific to a single amino acid, but able to recognize several synonymous codons (because the pairing with the third nucleotide of the codons follows certain rules known as "wobble"). Therefore, there are most often two or three species of transfer RNA for the same amino acid. The set of transfer RNA species varies somewhat between organisms, in relation to the uneven use of the different synonymous codons. It is also these variations that are at the origin of the few modifications of the genetic code observed in certain organisms (yeasts, ciliates, diplomonads, etc.) and, more generally, in mitochondria and chloroplasts. Most of these changes concern non-sense codons that become sense codons (sometimes the opposite), but there are also rare cases of changes in the meaning of sense codons. Despite their limited nature,

3 Mitochondria and chloroplasts each have, in addition, their own set of transfer RNA.

changes in the genetic code significantly reduce the interchangeability of genes between cell compartments (nuclei, mitochondria, chloroplasts) or between organisms that differ in their code (horizontal transfers, see Chapter 6).

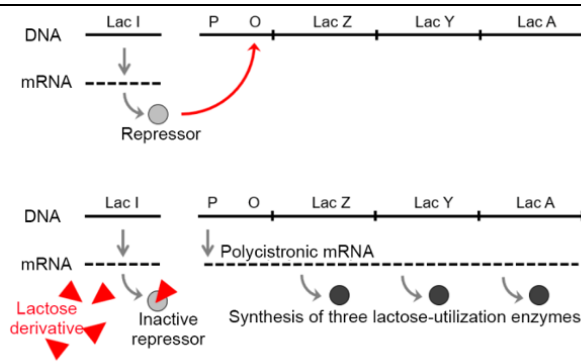
Finally, it is essential to recall that the genetic code applies exclusively to messenger RNAs, and not to genes, as some unfortunate common expressions tend to suggest. In fact, as will be seen later, many genes do not have messenger RNA among their products, which consist only of other classes of RNA molecules. Moreover, messenger RNAs are generally not the direct products of the genes that produce them, but distant derivatives of the primary transcripts after various stages of maturation. Before going into the details of these phenomena of crucial importance for the understanding of genetics, a brief historical detour is necessary.

2.5. First paradigm of gene expression: the bacterial lactose operon

The central dogma of molecular biology postulated that DNA was transcribed into RNA by following the rules of base pairing: A with U, G with C, T with A and C with G, but said nothing about the molecular mechanisms of this transcription and its regulation. However, as early as the 1950s, it was observed that bacteria growing on a culture medium containing both glucose and lactose first use glucose and only start using lactose when glucose is exhausted. They are therefore able to change their range of active enzymes according to the available sugars, which may seem "economical", but leaves intact the mystery of the mechanisms involved. Since enzymes are the products of genes, there had to be at least one regulatory step between the genes and their products. Similarly, the harmonious development of a multicellular organism is impossible to imagine without some coordination of the expression of its genes.

By studying the assimilation of lactose in a bacterium of the human intestinal flora, *E. coli*, F. Jacob, J. Monod and their colleagues made a decisive conceptual advance (Jacob *et al*. 1960; Jacob and Monod 1961). They distinguished two types of genes: those that encode enzymes, called "structural genes", and those (actually only one at that time) that regulate the expression of the former, called "regulatory genes". The product of the regulatory gene was not easy to guess. For some time, the authors imagined

that these could be RNA molecules that, through a process yet to be discovered, would interfere with the transcription of the structural genes.



A set of three genes encoding proteins (Lac Z, Lac Y and Lac A, scales not respected) and carried by the same locus of the *E. coli* chromosome is transcribed into RNA from the same P **promoter**, forming an **operon**. Another gene Lac I, corresponding to another chromosomal locus, leads to the synthesis of a protein with a high affinity for a short DNA sequence, called **operator** O, associated with the promoter. The binding of this protein to the operator (top) prohibits the transcription of the operon, hence its name of **repressor**. In the presence of lactose (actually a metabolic derivative of lactose, red triangles), the repressor is inactivated by a **conformational change**, which eliminates its **affinity for** the operator (bottom), thus releasing the transcription of the operon and leading to the synthesis of the lactose utilization enzymes.

This model, here oversimplified, had a considerable impact at the origin of molecular biology, because it made it possible for the first time to imagine how the expression of different genes could be co-regulated, paving the way for regulatory networks, even if reality later turned out to be a little more complex. This model also clearly illustrated for the first time the fundamental difference between **cis-acting mutations** (those that affect elements of the genome, as here the operator) and **trans-acting mutations** (those that affect gene products, as here the repressor). The distinction between cis- and trans-acting mutations by functional complementation tests between mutants is one of the fundamental bases of genetics.

There are other bacterial operons that work with positive regulations (transcriptional ***activators****) instead of the negative regulation (repressor) illustrated here by the lactose operon.

Box 2.6. *Eschericia coli lactose operon. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

It should be noted how modern this assumption is compared (in spite of it being not true in this instance) to everything we know today (see section 2.9). But other mutants, and what was known at the time about the cycle of the temperate[4] bacteriophage called λ, led to the idea that the regulatory gene, called *Lac I*, had a protein as a product. This protein prevents the transcription of the structural genes in absence of lactose, hence its name as a ***repressor****. In the presence of lactose, the repressor is inactivated and the transcription of structural genes takes place, resulting in the synthesis of the enzymes necessary for the use of lactose. But there are three of them, and their synthesis is coordinated: beta galactosidase, produced by the *lacZ* gene, which hydrolyzes lactose into galactose and glucose; permease, produced by the *lacY* gene, which allows lactose to enter the cell; and transacetylase, produced by the *lacA* gene (see Box 2.6). These three genes are closely linked to each other on the bacterial chromosome. They form three contiguous cistrons, hence the idea that the coordination of their expression is done by the common transcription of a polycistronic RNA from a single site called a ***promoter****, itself adjacent to a hypothetical new site, called ***operator****, allowing the action of the repressor.

In terms of the logic of isolated mutants, this operator has the following properties: it is adjacent to the group of structural genes whose expression it controls and it is sensitive to the presence of the repressor. Unlike the mutations in the *Lac I* gene that could be supplemented in *trans*, those of the operator only acted in *cis*, that is, the presence of another operator (provided by a recombinant sexual ***episome****) did not change the phenotype of the bacteria. Beyond the lactose ***operon****, it should be noted that this distinction between cis-acting and trans-acting mutations is one of the most fundamental bases of genetics. It makes it possible to differentiate between those parts of the genome that give rise to a product (therefore diffusible in trans) and those that are important only at the level of the genome itself and can therefore only act in *cis*.

It is now known that the functioning of the lactose operon of *E. coli* is a little more complex than the original model anticipated. In addition, there are variants: other bacterial operons work with positive regulations (activators),

4 A temperate bacteriophage can integrate its DNA into that of the host bacterium and remain silent for many generations of bacteria. In the presence of an inductive event, such a bacterium will produce a population of bacteriophages, as if it had just been infected, whereas the infection goes back to its very distant ancestor.

unlike the negative regulation (repressor) of the lactose operon, or sometimes with several promoters and operators. It is also known that the molecular mechanisms of gene expression are very different in the case of eukaryotic cells. But the value of the lactose operon was that it provided the first conceptual basis for gene expression by showing how *trans-active* elements (from any part of a genome) act on *cis-active* elements directly associated with genes of interest. These bases are universal to the whole living world. Trans-active elements allow a living cell to respond to the environmental variations encountered. In simple cases, such as that of the lactose operon, the elements themselves are the targets of effectors external to the cell (such as lactose here) that induce changes in molecular conformation (***allostery****). But more generally, trans-active elements respond to cascades of intracellular reactions in which a membrane receptor receives a signal from an external effector and transmits it to the elements of transcriptional regulation of the cell by a complex succession of chemical modifications of intermediary molecules. Here, we are talking about signaling routes. In all cases, however, it remains a mode of gene expression regulation at the transcriptional level, that is a mechanism in which the main role is played by the RNA synthesis activity. However, as we will see, RNA had many other surprises in store for us.

2.6. Reverse transcription and retrogenes

The first surprise was in 1970. As we began to understand that neither transposons nor organelle genetics contradicted Mendelian genetics, but brought new elements to it, a discovery changed the central dogma of molecular biology: RNA could serve as a template for synthesizing DNA. In other words, RNAs that were believed to be only intermediaries between genes and their products could also be the source of an inherited message transmitted to the offspring. This fundamental discovery came from the study of avian oncogenic viruses, now classified as ***retroviruses****. These viruses contain in their capsid a molecule of RNA and not DNA, hence their initial designation as *oncornaviruses*. When they infect cells, these viruses bring with them a particular protein, which uses their RNA as a template to synthesize a complementary strand of DNA, which then serves itself as a template to synthesize the other strand of DNA. Finally, the infected cell produces a double-stranded DNA that is a copy of the virus' RNA. Under the action of enzymes produced by the virus (because the viral RNA also serves as messenger RNA), the double-stranded DNA molecule is then

integrated into the genome of the host cell into a new chromosome locus. From then on, the viral genome will be transcribed into RNA molecules and replicated in the continuity of chromosomes during cell divisions, like any other gene of the host. The transcripts will serve as messenger RNA for the synthesis of viral proteins in which they will be encapsulated, thus completing the cycle by producing new viruses but leaving a copy of the viral genome in the genome of the infected cell which, if it survives, will transmit it to its offspring.

This mechanism explained that these viruses tend to produce cancers because, by integrating their genome into that of the host, they are perfect mutagens. Most importantly, it was discovered that among the proteins produced by these viruses, there was a new enzyme, a DNA polymerase as known for the replication of chromosomal DNA, but which uses RNA as template, which normal DNA polymerases don't do. In biochemical terms, it is an RNA-dependent DNA polymerase or, more simply, a **reverse transcriptase**. For a while, the phenomenon seemed to be restricted to viruses. No cellular genes were suspected to have an RNA ancestor. But, since the biochemical mechanism existed, it would have been surprising if it had not been used. It took a few more years to demonstrate this, but today it is clear that the formation of genes or gene fragments from RNA is an important mechanism in the evolution of genomes (see Chapter 6). Therefore, an accidental acquisition (such as, here, the infection by a virus) can become hereditary. We will come back to this later, because we still lack elements to fully understand this fundamental notion of genetics.

2.7. Exons, introns and splicing: the first complexity of RNA life

The reality of RNAs has not been quickly revealed, as many of these molecules have a very short lifespan, which has long hidden their immense diversity. They have only become truly accessible with the most modern methods of genomic analysis, combined with the use of mutants in certain model organisms such as yeasts. Biochemically, DNA transcription is now a very well described mechanism. It is a continuous stepwise phenomenon, carried out by a complex molecular machinery made up of many proteins, including an RNA polymerase[5]. Transcription includes an initiation step,

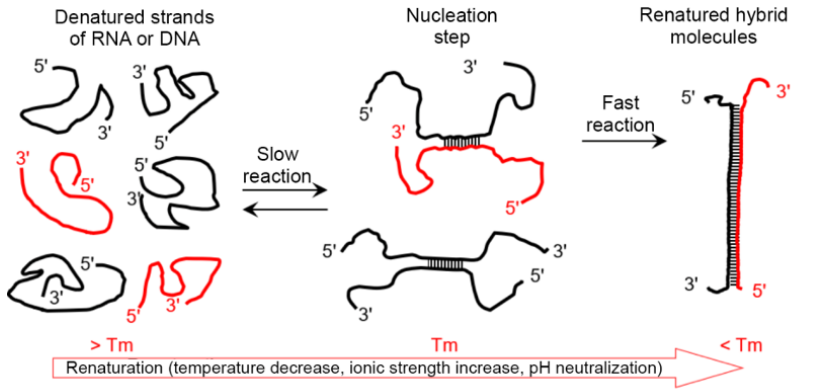
5. In eukaryotes, there are three different types of RNA polymerases, each specialized in the transcription of a gene category, those whose products are ribosomal RNAs (polymerase I), small stable RNAs (polymerase III), messenger RNAs or short-lived RNAs (polymerase II).

during which the interaction of this machinery with DNA defines the starting point along the DNA sequence by locally separating the two strands. It is followed by an elongation step (i.e. the progression of the synthesis of an RNA molecule with a sequence complementary to that of one of two DNA strands) which may be sensitive to various "signals" encountered along the template sequence. Finally, it ends with a specialized DNA sequence (***terminator****) or with an accidental interruption of the elongation following the signals encountered. The details of these mechanisms are beyond the scope of this book. What is important to remember is that while RNAs are the **only direct products** of genes, transcription does not always respect their limits; it can start and end at many sites upstream and downstream of genes, or even within genes. The relationships between primary transcripts and genes are therefore much more complex than previously thought.

In eukaryotes, almost all chromosome sequences are transcribed on one or the other strands of DNA used as template and, quite often, on both at the same time. There are preferential initiation and termination regions along these sequences – referred to respectively as "promoters" and "terminators" – but these do not systematically correspond to the intervals between genes. The result is an enormous diversity of primary transcripts, with extremely variable lifespans (usually very short), partially overlapping with each other and potentially covering all or part of one or more genes. It is the combination of the initiation rates at different promoters, the encounter of specific signals along the transcribed sequences and the rapid degradation by ***exosomes**** that regulates gene expression and makes the majority of final transcripts monocistronic. The preponderance of post-transcriptional phenomena, which will be discussed below, has very significant consequences, because RNA molecules engage in complex biochemical reactions that sometimes modify them enormously.

The first of these reactions concerns the ***splicing**** of ***introns****. Introns had not been predicted by genetics. Their discovery in 1977 was a surprise (Gilbert 1978). By hybridizing adenovirus messenger RNAs with the corresponding DNA fragments and examining the hybrid molecules under an electron microscope, it appeared that regions within the gene were absent from the messenger RNAs, forming characteristic DNA loops between the RNA-DNA hybrid segments (principles of nucleic acid hybridization are explained in Box 2.7). The messenger RNA was therefore not the complete copy of the gene, but only that of discontinuous regions of the gene. It was quickly realized that the same was true for many genes of eukaryotic

organisms. The genes were therefore fragmented, since non-contiguous sequences at the DNA level were attached together in the same messenger RNA by eliminating the interval between them[6].



Hydrogen bonds between **complementary bases** in nucleic acid helices are broken when temperature or pH increases and ionic strength decreases, destroying the three-dimensional structures of the RNA strands and separating the two strands from the DNA molecules. The resulting "denatured" molecules are polymers that each retain their own nucleotide sequence. **Tm** (temperature of melting) is the physical-chemical conditions at which 50% of the molecules are denatured. In addition to the physical-chemical parameters, the Tm of a nucleic acid molecule depends on its base composition and length.

Placed under conditions beyond the Tm, all molecules are **denatured**. Below Tm, hydrogen bonds seek to reform between fragments of molecules with complementary sequences. This reaction occurs in two stages, **nucleation** and **elongation**. The first corresponds to the **random collisions** between molecules that allow the formation of "nuclei" of some paired bases in complementary sequences. This reaction is reversible when close to the Tm. The second is an elongation of the base pairings on either side of the "nuclei" **if the sequences remain complementary** (example on top). It does not take place if the sequences are not complementary (example at bottom). By playing with the experimental conditions, it is therefore possible to identify or to isolate from complex mixtures the nucleic acid molecules (RNA or DNA) whose sequences are complementary to those of a chosen **probe**. This probe can be a natural or synthetic DNA or RNA fragment, labeled (radioactively or chemically) or bound to a solid material (column, plate).

6 It should be noted that this does not contradict the notions of cistron and locus and that is why classical genetics had not suspected the phenomenon.

Molecular hybridization of nucleic acids has played a significant role in gene cloning, directed mutagenesis, genome mapping (see Chapters 2 and 4) and continues to play an essential role in modern molecular genetics. By playing with the experimental conditions and the many chemical modifications available, it is possible to search only for sequences that are strictly complementary to the chosen probe or, on the contrary, partially complementary (similar but not identical sequences).

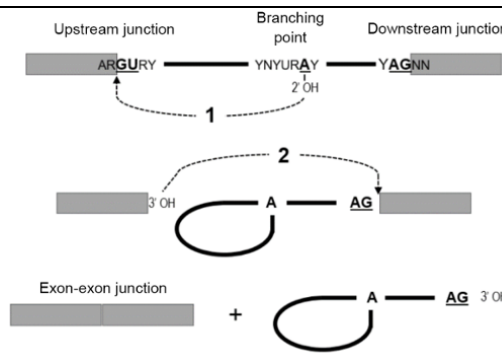
Box 2.7. *Molecular hybridization of nucleic acids. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

We will ignore here the various investigations that immediately followed this unexpected discovery to move directly to current knowledge. The fragmented genes give rise to RNA molecules called "primary transcripts", whose sequences are entirely collinear with those of DNA. But these RNA molecules very quickly engage in splicing mechanisms during which internal segments are eliminated, while flanking segments are spliced together. It is essential to understand that all these reactions occur exclusively at the level of RNA molecules[7], the genes themselves are never involved. The sequences eliminated during the splicing of the RNA received the name ***intron**** (for internal region), the remaining ***exon**** (expressed region). The corresponding regions of the genes are generally designated accordingly, although splicing does not exist with DNA.

RNA splicing is now very well understood at the biochemical level. It consists of two coupled ***transesterification**** reactions that create a new phosphodiester bond between two nucleotides that were not previously contiguous in the initial sequence (see Box 2.8). Chemically, the new bond is not distinguishable from the other phosphodiester bonds which form polyribonucleotide chains. But genetically, a **new sequence** is created by this new junction. The gene product is therefore not identical to the gene, but it is also not genetically undetermined, because the transesterification sites are strictly defined by the RNA sequences. The genetic message transmitted

7 For chemical reasons, it cannot be otherwise. Splicing reactions are transesterifications that require the presence of the hydroxyl group in the 2' position of the ribose. It is precisely this group that is replaced by a hydrogen atom in the deoxyribose of DNA (see Boxes 2.1 and 2.2). Introns and exons therefore only functionally exist in RNA molecules, never at the DNA level. It is only their image, made up of the corresponding sequence segments, that is reflected in the genes.

from generation to generation is not directly the one used at each generation, but it carries in itself the instructions for its usage.



The **exons** are symbolized by the gray rectangles, the **intron** by the thick line. The characteristic sequences are indicated (R: purine, Y: pyrimidine, N: any nucleotide). The splicing mechanism of spliceosomal and group II introns consists of two successive ***transesterifications**** (dotted lines noted 1 and 2) between particular sequences of the RNA molecule.

The first reaction forms a covalent bond between the 5′ phosphate group of the first nucleotide of the intron (always a G) and the hydroxyl group on the 2′ position of a specific nucleotide (always an A) of the branching point located inside the intron, thus releasing the 3′ hydroxyl group of the last nucleotide of the upstream exon (often a purine). At this intermediate stage, the downstream exon is attached to the intron, which shows a **lariat** figure.

The second reaction forms a covalent bond between the 5′ phosphate group of the first nucleotide of the downstream exon and the hydroxyl group of the last nucleotide of the upstream exon. The exon-exon junction is therefore chemically indistinguishable from any other nucleotide junctions of the RNA molecule (however, some organisms "mark" this junction by binding a specific protein complex). This reaction releases the intron in the form of a lariat which, depending on the case, will be degraded (most often) or involved in other reactions.

In the case of group II introns, these reactions are catalyzed by the intron itself, which behaves like a **ribozyme** (enzyme made of RNA). In the case of spliceosomal introns, these reactions occur within **spliceosomes**, which are complexes of specialized proteins and RNAs that recognize the upstream junction and branching point and trigger the catalysis of transesterification. Splicing of the other categories of introns (group I, transferRNA, etc.) follows slightly different mechanisms, but always involves two successive transesterification steps.

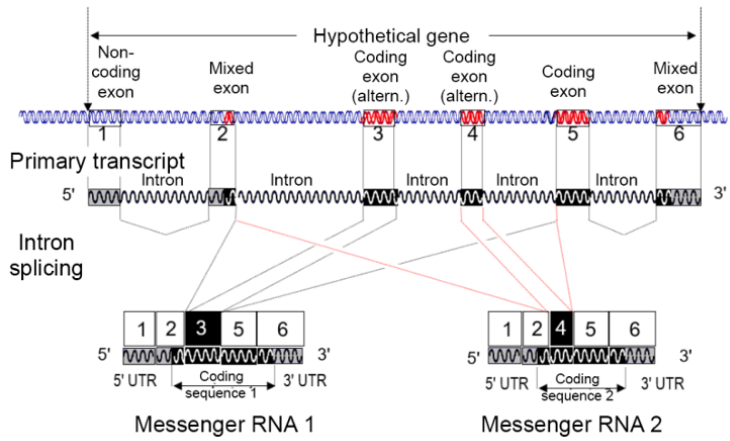
Box 2.8. *RNA splicing*

It is now known that there are several categories of introns that differ in the details of RNA splicing mechanisms. The introns of the precursors of transfer RNA molecules are eliminated by proteinic enzymes that cleave and ligate the RNA chain. The splicing of messenger RNA precursors in eukaryotes involves macromolecular complexes – called ***spliceosomes**** – made of specialized RNA (snRNA) and proteins. Finally, there are two groups of introns (present mainly in mitochondrial and chloroplastic genes but also in bacteria, archaea and some viruses), which themselves catalyze RNA splicing reactions without the help of other molecules. These are examples of catalytic RNAs or ***ribozymes****. In all cases, splicing uses the information written in the RNA sequences. In the case of spliceosomal introns, the sequences of the exon-intron junctions and of the branching point internal to the intron are recognized by the sequences of the spliceosomal snRNA to define the sites of the transesterification reactions. In the case of autocatalytic introns, the three-dimensional structures of the RNA molecules define the catalytic sites and their substrates.

While rare in bacteria and archaea, introns are extremely numerous in eukaryotes, where they can be found in practically all categories of transcripts produced by the nucleus and by cellular organelles containing DNA (mitochondria, chloroplasts and ***nucleomorphs**** when they exist), regardless of the nature of the final product of these genes: RNA or protein. Introns can be small in size (this is often the case with fungi and many unicellular eukaryotes), but they are often large in size, much larger on average, than exons (in mammals for example). In humans, there are, on average four or five introns per protein-coding gene. Some genes have many more and as a result, can extend over very long DNA sequences. For example, our *DMD* gene, involved in muscle contraction, covers more than 2.2 million nucleotides, more than many bacterial genomes.

Contrary to popular belief, the organization of genes into exons and introns is completely independent of protein synthesis. First, it concerns all categories of genes and not only those encoding proteins. Second, the fact that mature messenger RNAs (ready for translation) consist only of exons (by definition) does not mean that all exons are included in the translated sequence or that introns cannot be translated. In reality, the situation is much more complex. On the one hand, mature messenger RNAs consist of the assembly of exons, some of which are **coding** (included in the translated sequence), others **non-coding** (sequences upstream and downstream of the translated part) and others **mixed** (at junctions). These elements have their

equivalent positions in the corresponding genes. On the other hand, introns can themselves be translated because they contain a specific coding sequence (often the case in autocatalytic introns) or because of alternative splicing (see Box 2.9). Finally, it is not uncommon for introns of some genes to contain other genes within themselves (e.g. for small stable RNAs such as small nucleolar RNAs (snoRNAs) involved in the cutting of ribosomal RNA precursors) or mobile genetic elements or their remains. It is therefore often complex mosaic assemblages of intertwined genetic elements that are found in the DNA of eukaryotes. This entanglement is reminiscent of the fragmentation of files in computer memories. Obviously, to work, this requires keeping the information necessary for the correct reassembly of the parts. The genetic message transmitted from generation to generation is therefore not only the information necessary for the synthesis of gene products, but also the protocol for using this information (see section 2.9).



A hypothetical **eukaryotic** gene (top) is transcribed into **primary** RNA molecules (center) that engage in **splicing** reactions (to simplify, it is assumed here that there is only one type of primary transcript, which is rarely the case). In this theoretical scheme, exons 1 and 2 on the one hand and 5 and 6 on the other hand are always linked together. For them, splicing is not an alternative. On the other hand, exon 2 can be linked either to exon 3 or to exon 4 (themselves linked to exon 5 respectively) thus producing two different **mature** RNA molecules. This is called **alternative splicing** (black or red dotted line). In the case of messenger RNA molecules (shown here), the coding sequences of the different RNA molecules produced from the same gene, and therefore the synthesized proteins, may be partially different.

It should be noted that in any messenger RNA, the coding sequence is preceded by a sequence called the 5′ UTR and followed by a sequence called the 3′ UTR. Depending on the position of the introns, there will therefore be **coding exons** (fully included in the coding sequence), **non-coding exons** (fully excluded from the coding sequence) and **mixed exons** (overlapping the beginning or end of the coding sequence). The position of the different exons and introns can be reported on the DNA sequence level if the sequences of the messenger RNAs are available. Alternative splicing complicates this representation. Here, six exons (boxes) are defined with the fragments of sequences that can be used as coding sequence in red.

It should be noted that eukaryotic genes are often more complex, as other genetic elements, oriented with one or the other strand of DNA, may find their place in the sequences defined here as introns. These include cases of mobile elements (or their remains), genes for ***non-coding RNA****, or interfering RNA, repeated sequences or cis-active genetic elements such as ***enhancers**** of the gene or its neighbors.

Box 2.9. *Principle of correspondence between a gene and its transcripts. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

2.8. Sequence editing: the second complexity of RNA life

In addition to splicing reactions that reorganize sequence fragments or cleavages of polyribonucleotide chains that participate in the maturation of some transcripts (e.g. ribosomal RNAs), RNA molecules are also subject to other chemical modifications. Some, consisting of the addition of various chemical radicals to specific nucleotides (e.g. in transfer RNA molecules) do not modify the initial nucleotide sequence, but others do. The sequence of the RNA produced is thus no longer an exact copy of the DNA sequence. ***RNA editing**** is the process(es) by which additions, subtractions or precise modifications of nucleotides occur in RNA molecules.

The first example of RNA editing was discovered by chance in 1986: the messenger RNA of the *COXII* mitochondrial gene of a trypanosome contained four nucleotides absent from the corresponding DNA (Benne *et al.* 1986). This was disturbing. Where did the information come from to precisely modify this sequence in this way? We now know that such a phenomenon is not specific to this gene and this organism, but is widespread throughout the living world. RNA molecules undergo specific editings of their sequences, which may be limited to one or a few nucleotides or be massive, to the point of affecting half of the nucleotides.

This phenomenon is the result of several molecular mechanisms whose description would exceed the scope of this book. In the simplest version, one nucleotide can be changed to another by simple chemical modification *in situ*. For example, a C can be changed to a U by simple deamination, as is the case for the messenger RNA for apolipoprotein B in humans. In other cases, precise numbers of nucleotides (often U) are added (or sometimes removed) at specific sites of an RNA molecule, after precise cuttings and ligations of the nucleotidic chain. In both cases, an information is required to choose the target site and the type of modification made because, like in the cases of intron splicing, the editing phenomena do not occur randomly but rather follow a "pre-determined" plan producing functional RNAs.

The origin of this information is now clear. It is written in the sequence of the RNA to be edited or in those of other RNA molecules with which it interacts. Limited editings occur on sites determined by the local three-dimensional structure of the RNAs to be edited. The nucleotide sequence defines the structure which, in turn, defines the changes to be made to the primary sequence. By contrast, massive additions/subtractions of nucleotides involve the action of ***editosomes****, that is, ribonucleoprotein complexes containing small RNA molecules, called guide RNA or gRNA. These guide RNAs are themselves transcripts of other loci in the genome. By their sequence, they provide the editosome with the necessary information about the nature and number of nucleotides to be added or removed at each precise position of the RNA sequence to be edited. Long cascades of reactions can only occur in a defined order, because the products of one reaction are the substrates for the next reaction. In conclusion, in addition to the genetic information used to synthesize gene products, there is a protocol for using this information, which is none other else than part of the genetic information itself.

2.9. RNA interference and epigenetics

This duality has significant consequences, because the phenotype of an organism or a cell depends not only on the allele it carries, but also on the state of activity in which it is found: simplified, this is active or inactive (i.e. expressed or not). Now, in many cases, this state of activity may itself be inherited in a more or less stable fashion. This additional dimension of genetics gives rise to a wide variety of phenomena now grouped under the term "epigenetics", after having remained mysterious for a long time. This

designation actually encompasses a wide variety of mechanisms, including chemical modifications of DNA (methylation) or of proteins in chromatin (acetylation, methylation, phosphorylation), modifications which interfere with transcription as well as with interactions between RNA molecules that modulate their activity or lead to their degradation.

The story began very early. As far back as 1915, Bateson himself observed that "rogue" peas crossed with normal peas only produced rogue[8] peas in subsequent generations (Bateson and Pellew 1915). The same phenomenon was generalized in 1956 under the name "paramutation" by **Royal Brink**, who studied corn (Brink 1956). Something was therefore superimposed on Mendel's laws in the progeny of these plants. Independently, in 1920, it was observed that the symptoms of a plant infected with a virus[9] gradually fade as the plant grows: the young leaves appeared healthy and their infection by the same virus did not lead to any new symptoms, as if they were "immunized". Molecular methods have now confirmed the absence of viral particles in such tissues and also shown that it is not possible to infect them with a different virus if part of the sequence of its genome is common with that of the genome of the first virus (Ratcliff *et al*. 1997). Immunity, therefore, seemed to depend on the presence of small RNA sequences. What would happen then if a plant were infected with a recombinant virus containing a fragment of a messenger RNA of a plant gene? Can we "immunize" a plant against one of its own genes? Surprisingly, the answer is yes. Such viruses inactivate the expression of the corresponding gene, causing the same phenotype as a mutation that would inactivate that gene (Baulcombe 1999). This phenomenon, called virus-induced gene silencing or VIGS, has become a common tool for the genetic study of plants. But on which molecular mechanism is it based?

Other unexplained phenomena of gene inactivation had been observed over the years in different organisms. In the early 1990s, when researchers tried to increase the color intensity of petunia flowers by increasing, through transgenesis, the number of copies of the chalcone synthase gene (the enzyme acting at the beginning of the biosynthetic pathway of anthocyanins, the molecules responsible for the coloration of flowers in this species), they obtained the opposite result to the one expected (Napoli *et al*. 1990). Plants

8 These plants do not have the normal appearance of their ancestral varieties but show a complex of syndromes with, simultaneously, all organs of reduced size.

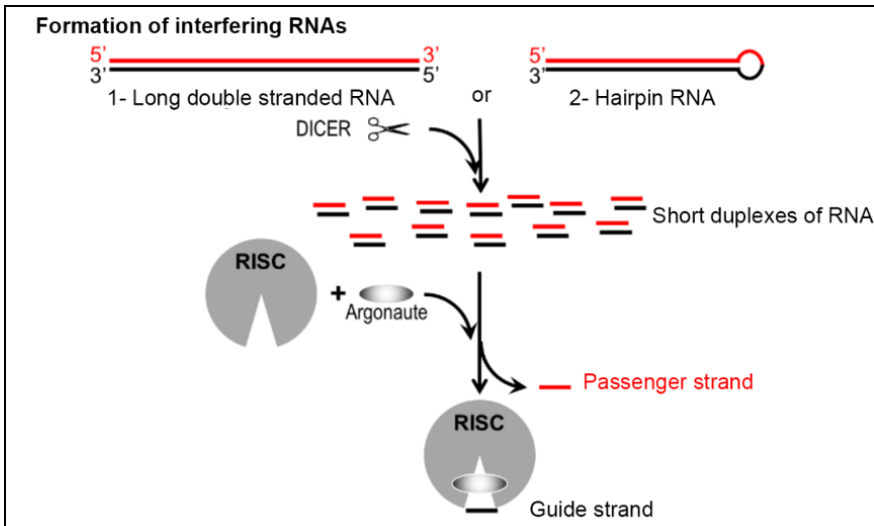
9 Most plant viruses have genomes made up of RNA and not DNA.

that carried several copies of the gene had flowers entirely white or with faint colored markings on a white background. It was later understood that the original gene and the copy introduced by transgenesis were both inactivated by the specific degradation of their messenger RNAs. The phenomenon, common to the other genes later studied in plants, was called post-transcriptional gene silencing or PTGS. Even more surprisingly, after grafting, the transgenic rootstock overexpressing a given gene was able to inactivate the corresponding gene of the non-transgenic graft (Palauqui *et al.* 1997). The inactivation signal of genes by degradation of their messenger RNA was therefore able to spread from cell to cell.

Similar phenomena of inactivation of genes in supernumerary copies were observed in filamentous fungi (molds) such as *Ascobolus immersus* or *Neurospora crassa* (quelling phenomenon, Romano and Macino 1992) and in the nematode *Caenorhabditis elegans* (Fire *et al.* 1998; Chang *et al.* 2012). It is now known that in *A. immersus*, the phenomenon results from a particular mechanism of methylation of repeated DNA sequences that occurs before meiosis, inactivating genes in supernumerary copies. In the other two species, on the contrary, as in the case of plant PTGS, the phenomenon results from a much more general mechanism, now called RNA ***interference**** (or RNAi), because it involves new RNA molecules that had previously escaped analysis. Originally perceived as a defense mechanism against invasive genetic elements such as viruses or transposons, RNA interference actually regulates a large number of cellular and developmental processes based on molecular mechanisms that are highly conserved in eukaryotes.

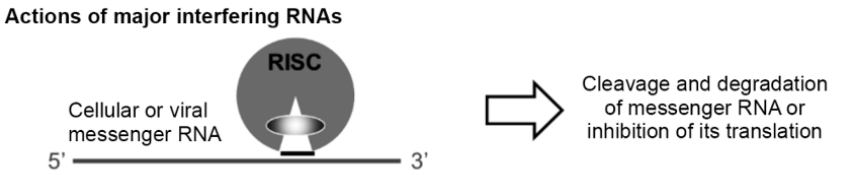
It was with *C. elegans* that we understood that RNA interference depended on specific small RNA molecules, called "interfering RNA" (Parrish *et al.* 2000). There are currently three main families, known as siRNA ("si" for short interfering), miRNA ("mi" for micro) and piRNA ("pi" for piwi-interacting). The latter family is more generally involved in the control of transposon activity, but not only this. These molecules are found in all studied eukaryotic organisms that have retained an RNAi mechanism, including of course plants (Hamilton and Baulcombe 1999), which have played a major role in elucidating the molecular mechanisms involved (see Box 2.10). It is now known that small duplexes of RNA (21–22 or 25–27 base pairs, depending on the family) are produced after specific cleavage of double-stranded RNA molecules by a nuclease called Dicer. These duplexes are then loaded onto a macromolecular complex called RISC

where, following the action of a protein called Argonaute, the two strands are separated from each other, allowing one of them (the "guide") to load the RISC complex onto a messenger RNA showing sequence complementarity with the guide RNA, hence providing the signal for a specific action. In cases of interference by siRNAs, derived from transcripts of the target gene itself, the messenger RNA is cleaved, leading to its eventual destruction. Interference by miRNAs involves their interaction with the untranslated part of the messenger RNAs, preventing their translation by ribosomes and also leading to their final destruction.



Several molecular mechanisms lead to the formation of double-stranded RNA molecules in linear form (replication of viral RNA, convergent transcription of the same chromosome segment on both DNA strands) or in hairpin form (transcription of a palindromic DNA sequence).

In all organisms possessing an **RNA interference machinery**, these molecules are the target of a ribonuclease of the RNAse III family, called **Dicer**, which cleaves them in short RNA duplexes of 21–22 nucleotides or 25–27 nucleotides as appropriate. These short duplexes are then loaded onto a macromolecular complex called **RISC** (for RNA induced silencing complex) where, with the help of a protein called **Argonaute**, the two strands are separated from each other, allowing one of them (guide) to carry the complex onto the messenger RNA molecules.



The sequence complementarity between the guide and the messenger RNA induces **cleavage** and then degradation of this messenger RNA (case of **siRNAs**) or inhibits its translation by ribosomes (case of **miRNAs**). There are other families of interfering RNAs.

Box 2.10. *RNA interference. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

Unlike siRNAs, miRNAs are derived from transcripts of distinct loci from the targeted messenger RNA. miRNA interference therefore represents adaptive mechanisms for regulating gene expression that play an important role in the developmental processes of multicellular organisms. By contrast, interference by siRNAs and piRNAs is more specifically oriented towards the inactivation of supernumerary copies of genes, and the resistance against viruses or the proliferation of transposable elements in genomes. The existence of RNA interference also provides us with a powerful tool to inactivate the expression of specific genes without rewriting genomes (see Chapter 6). Simply providing an entire cell or organism with a DNA fragment carrying a small palindromic sequence is sufficient to inactivate the expression of the gene targeted by that sequence.

2.10. Important ideas to remember

– The molecular structure of **DNA** that carries the genetic information in the form of the base sequence along a strand explains its identical **replication** by the chemical complementarity of the bases between the two strands (hydrogen bonds).

– DNA molecules serve as templates for the synthesis of **RNA** molecules by a mechanism called **transcription**, which uses the chemical complementarity of the bases. The process known as "gene expression" is regulated at the pre- and/or post-transcriptional level.

– Gene products are **RNA** molecules. There are many different families of RNA molecules differing by structures and functions. **Proteins** are the products of messenger RNAs.

– Newly synthesized RNA molecules undergo or catalyze various **maturation** stages that modify them more or less profoundly before reaching their final active forms. These are different chemical reactions, all precise and genetically determined.

– RNA molecules originating from different genes can interact specifically with one another. Some interactions can lead to **epigenetic** phenomena.

– The specific genetic information carried by each messenger RNA is **translated** according to the common rules of the **genetic code** during protein synthesis by the **ribosomes**.

– Translation involves another class of RNA, called **transfer RNA**, in which each species is covalently linked to a specific amino acid.

– Different genetic elements can be entangled in each other at the DNA level. Gene **fragmentation** (resembling that of computer files) is solved at the level of RNA molecules.

2.11. References

Bateson, W., Pellew, C. (1915). On the genetics of "rogues" among culinary peas (Pisum sativum). *Journal of Genetics*, 5, 13–36.

Baulcombe, D.C. (1999). Fast forward genetics based on virus-induced gene silencing. *Current Opinion in Plant Biology*, 2, 109–113.

Benne, R., Van den Burg, J., Brakenhoff, J.P., Sloof, P., Van Boom, J.H., Tromp, M.C. (1986). Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, 46, 819–826.

Brink, R.A. (1956). A genetic change associated with the R locus in maize which is directed and potentially reversible. *Genetics*, 41, 872–889.

Chang, S.S., Zhang, Z., Liu, Y. (2012). RNA interference pathways in fungi: mechanisms and functions. *Annual Review of Microbiology*, 66, 305–323.

Chargaff, E., Zamenhof, S., Green, C. (1950). Composition of human desoxypentose nucleic acid. *Nature*, 165, 756–757.

Crick, F.H., Barnett, L., Brenner, S., Watts-Tobin, R.J. (1961). General nature of the genetic code for proteins. *Nature*, 192, 1227–1232.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391, 806–811.

Gilbert, W. (1978). Why genes in pieces? *Nature*, 271, 501.

Gros, F. *et al*. (1961). Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*. *Nature*, 190, 581–585.

Grosjean, H., Westhof, E. (2016). An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Research*, 44, 8020–8040.

Hamilton, A.J., Baulcombe, D.C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286, 950–952.

Hoagland, M.B., Stephenson, M.L., Scott, J.F., Hecht, L.I., Zamecnik, P.C. (1958). A soluble ribonucleic acid intermediate in protein synthesis. *Journal of Biological Chemistry*, 231, 241–257.

Jacob, F., Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3, 318–356.

Jacob, F., Perrin, D., Sanchez, C., Monod, J. (1960). Opéron : un groupe de gènes avec l'expression coordonnée par un opérateur. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 250, 1727–1729.

Napoli, C., Lemieux, C., Jorgensen, R. (1990). Introduction of a chimeric chalcone synthase gene into *Petunia* results in reversible co-suppression of homologous genes in trans. *The Plant Cell*, 2, 279–289.

Nirenberg, M.W., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., Anderson, F. (1966). The RNA code and protein synthesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 31, 11–24.

Palauqui, J.C., Elmayan, T., Pollien, J.M., Vaucheret, H. (1997). Systemic acquired silencing: Transgene-specific post-transcriptional silencing is transmitted by grafting from silenced stocks to non-silenced scions. *The EMBO Journal*, 16, 4738–4745.

Parrish, S., Fleenor, J., Xu, S., Mello, C., Fire, A. (2000). Functional anatomy of a dsRNA trigger: differential requirement for the two trigger strands in RNA interference. *Molecular Cell*, 6, 1077–1087.

Ratcliff, F., Harrison, B.D., Baulcombe, D.C. (1997). A similarity between viral defense and gene silencing in plants. *Science*, 276, 1558–1560.

Romano, N., Macino, G. (1992). Quelling: transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences. *Molecular Microbiology*, 6, 3343–3353.

Sanger, F., Thompson, E. (1953). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. II. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 53, 353–374.

Sanger, F., Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. II. The investigation of peptides from enzymic hydrolysates. *Biochemical Journal*, 49, 463–490.

Watson, J.D., Crick, F.H. (1953a). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171, 737–738.

Watson, J.D., Crick, F.H. (1953b). Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171, 964–967.

3

Chromosomes and Reproduction

In living organisms, DNA molecules are included in structures that allow their transmission from generation to generation as well as exchanges with other DNA molecules. This organization is relatively simple in bacteria and archaea, where we already speak of chromosomes, although neither their biochemical constitution[1] nor their mode of hereditary transmission are comparable to those of eukaryotic nuclear chromosomes, which have a more complex structure and whose transmission is associated with gamete formation and sexual reproduction. In eukaryotes, DNA molecules specific to cellular organelles such as mitochondria and plastids are added, whose hereditary transmission is different. In prokaryotes, which do not have subcellular compartmentalization, genomes consist of all the DNA molecules present in the cytoplasm. Most of the time, these DNA molecules are circular with sizes varying between a few tens of thousands and a few million pairs of nucleotides. There may be one such molecule or more, depending on the species and sometimes even on individuals of the same species. The designation "chromosome" is reserved for DNA molecules that are consistently found in all individuals of the same species, while ***plasmids**** or **episomes** are used to designate those whose presence is inconstant.

3.1. The "true" chromosomes

As mentioned in Chapter 1, it is to Sutton that we owe what has been called the "chromosome theory of heredity", when in 1902, he was the first to make the connection between the Mendelian segregation of traits and the

1 Prokaryotic chromosomes are obviously composed of DNA, but this one is associated with proteins distinct from histones.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

distribution of condensed chromosomes among the daughter cells of meioses observed under the microscope. In addition, during mitosis and meiosis, each chromosome appeared duplicated in two copies, called the ***chromatids****, which remained attached to each other at a particular point, called ***centromere****, before separating between the two daughter cells. Chromosomes were discovered by W. Flemming as early as 1882, without anyone understanding what they were. Their name, which means "colored body" in Greek, referred to small sticks that could be stained by the techniques of the time and that appeared in the nuclei during cell divisions and mysteriously disappeared between these phases. It is now known that these chromosomes, or more precisely the ***chromatin**** of which they are made (see next paragraph), condense (becoming visible) and relax (disappearing) during the cell cycles of eukaryotic cells. The condensation factor (shortening, thickening) exceeds a thousand times. But what are these chromosomes made of?

It's now clear: DNA and special proteins, called **histones**. In interphases (apart from mitoses and meioses), when the chromosomes are decondensed, there is only one DNA molecule per chromosome. These molecules are therefore extremely long, several tens of centimeters in humans for example, while their diameter is only 2 nanometers (see Chapter 2). The length of DNA molecules is therefore several hundred thousand times larger than the diameter of the cell nuclei in which they are located. To achieve such a feat, the DNA is regularly wrapped around the histones, forming ***nucleosomes****. The nucleosomes, strung together, form the entire chromosome, like a pearl necklace where the thread would be wrapped around each pearl instead of passing through it.

Each nucleosome consists of a core, made up of eight proteins (2 histones of each type: H2A, H2B, H3 and H4) surrounded by about two turns of DNA. These structures are separated from each other by a short segment of DNA associated or not with another type of histone (H1 or its derivatives). Approximately 150 base-pairs of DNA, or 15 helix turns, are wrapped around each core, while the interval between two successive cores requires approximately 80 base pairs of DNA (8 helix turns). The same gene, especially if it is complex, will therefore be spread over several or even many successive nucleosomes. When the chromosomes condense, the successive nucleosomes are themselves wound into higher-order helices. The histone sequences hardly vary between organisms however evolutionarily

distant from each other, as if a functional optimum impossible to exceed had been reached here.

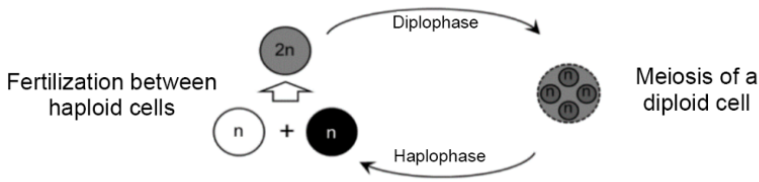
Depending on the living species, the number of chromosomes is extremely variable. At the extremes, in the Australian ant *Myrmecia pilosula*, the female is diploid, with 2 chromosomes, and the male is haploid with only one! The *Ophioglossum reticulatum* fern has been listed at 1,440! But in general, the number of chromosomes is adapted to the size of the genomes and the molecular mechanisms that ensure their correct segregation during cell divisions. Thus, all mosquitoes have 6 chromosomes (3 pairs), the diploid baker's yeast has 32 chromosomes (16 pairs), earthworms 36, cats and lions 38, rabbits and dolphins 44, humans and bats 46, sheep and burgundy snails 54, cows and goats 60, wolves and dogs 78, carps 100, etc. By convention, n designs the number of chromosome pairs, for example in humans n=23. We now know how these numbers change over time through duplications that increase them and through fusions that decrease them. In plants and other organisms such as fungi, the total number of chromosomes is often the result of more or less ancient hybridizations. For example, in *Poaceae* such as barley, wheat, rye and oats, the number of chromosomes is a multiple of 7, from 14 to 42. In sugar cane, it exceeds 100, following the hybridization of the species *Saccharum officinarum* ($2n = 80$ after duplications from an ancestor with $2n = 20$) with the species *S. spontaneum* ($2n = 40$ to 128 after duplications from an ancestor with $2n = 16$). Like genome size (see Chapter 4), the number of chromosomes therefore provides no direct information on the complexity or the way of life of the organisms in question.

3.2. Sexual reproduction and alternating generations

The fact that our own reproduction is biparental, like that of many other animals or plants, drew geneticists' attention to sexual reproduction very early on. However, this is only a specific phenomenon to eukaryotic organisms (even, not all of them), but its historical importance for our understanding of genetic mechanisms has been considerable and its complexity still raises many theoretical and practical problems. It deserves attention before examining other modes of reproduction in which essentially clonal lineages, which characterize mainly micro-organisms, dominate. But what is meant by sexual reproduction?

In a broad sense, sexual reproduction, which should rather be referred to as "gametic reproduction", involves the fusion of two gametes (fertilization)

to form an "egg" or zygote with a double number (diploid) of chromosomes. Depending on the organisms, the zygote itself or the diploid cells derived from it will carry out a particular division, called "meiosis", which ensures a return to the gametic (haploid) number of chromosomes. This alternation between a haploid phase (between meiosis and fertilization) and a diploid phase (between fertilization and meiosis) constitutes the cycle of sexual reproduction that is specific to eukaryotes, the two phases being of very variable relative importance from one group of organisms to another (see Box 3.1).



The sexual reproduction cycle of all eukaryotes always involves a **fertilization** step between two haploid **gametes** (*n* chromosomes), producing a diploid **zygote** (2*n* chromosomes), and a meiosis step in which a diploid cell produces four haploid cells (depending on the cases, they may all be viable or not). A complete cycle therefore always contains an alternation between haploid and diploid cells, but the relative importance of the two states varies greatly between organisms.

In many multicellular organisms, the cycle is **diplobiontic**, that is, the life of the organism consists only in the diploid phase, the haploid phase being limited to the gametes themselves which are produced immediately after the meiosis of a diploid cell of the germinative cell line. This is the case in the human species.

In other organisms, often unicellular, the cycle is **haplobiontic**, that is, the life of the organism consists of the haploid phase, the diploid phase being limited to zygotes that immediately enter meiosis themselves. This is the case, for example, for certain yeasts.

In other cases, the cycle is **haplo-diplobiontic**, that is, there are haploid organisms that produce gametes and diploid organisms with cells that will enter meiosis. These organisms may or may not be morphologically similar. They can be separated from each other or, on the contrary, associated in the form of organs of the same individual. In jellyfish or ferns, for example, haploid and diploid organisms are separated. We are talking about **alternate generations**. In mosses, the diploid phase grows at the peak of the leafy haploid phase.

In the haplo-diplobiontic cycle, haplophase and diplophase are not necessarily of equal importance and many cycles called "diplobiontic" actually fall into this category. In flowering plants, for example, the female haploid phase is derived from one of the four products of a meiosis which, after three successive cell divisions, will form the embryo sac. The male haploid phase consists of the pollen grain that germinates on the stigma of the flowers by emitting a pollen tube carrying two gametes that will ensure two fertilizations of two cells from the embryo sac. One will fertilize the egg cell (called oosphere) to give the diploid embryo and the other a two-nuclei cell to produce the triploid endosperm that will nourish this embryo.

Box 3.1. *Reproduction cycles in eukaryotes*

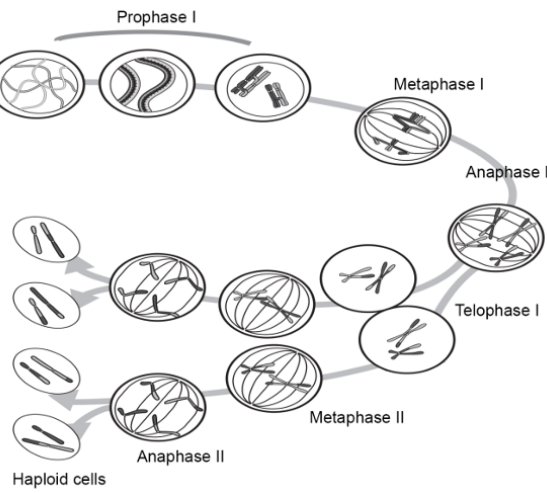
In species considered to illustrate a more primitive mode of reproduction, the zygote enters directly into meiosis and the visible form of the species is therefore haploid. This type of cycle, known as ***haplobiontic****, is observed for example in certain algae, mosses or fungi. In other, more complex eukaryotes, such as seed-bearing plants and most animals (but also fungi and some unicellular eukaryotes), the diploid phase predominates, the haploid phase being reduced to gametophytes (plants) or gametes (animals). The cycle is called ***diplobiontic****. Finally, there are organisms whose cycle is **haplo-diplobiontic**, that is the two phases, haploid and diploid, are of similar importance. This is the case, for example, in hepatica, certain red algae or animals such as jellyfish.

3.3. Meiosis

In all cases, meiosis must produce haploid cells that contain a complete set of chromosomes. But it does not reconstitute parental sets of chromosomes. On the contrary, it ensures a random redistribution of these parental sets by retaining, for each daughter cell, only one chromosome from each pair. To do this, meiosis links two successive cellular divisions following a single DNA replication (see Box 3.2). Starting from a diploid cell, it therefore produces four haploid cells[2]. We call these ***tetrads****.

2 Not all four daughter cells are necessarily viable. In female mammalian meiosis, for example, one of the four becomes the female gamete (called "ovule"), the others (called "polar globules") degenerate during their formation. The same is true for the formation of the female gametophyte in flowering plants and gymnosperms. When all four products are viable, as is often the case, for example, in fungi, the tetrads formed facilitate appropriate genetic

Meiosis is therefore in reality the entanglement of two successive modified mitoses. The first, called ***reductional****, does not end with the usual formation of two actual daughter cells ready to resume a normal cell cycle[3]. On the contrary, these cells enter directly the next G2 phase, ignoring the de-condensation of chromosomes, the DNA synthesis phase and the re-condensation of chromosomes, and lead to a second cell division, called ***equational****. Meiosis is a remarkably well-preserved mechanism where the main stages of meiotic divisions are similar in all species studied, suggesting that it must have already existed in the common ancestor of eukaryotes.



Meiosis is an uninterrupted succession of two cell divisions following a single DNA replication cycle. The ***prophase**** of the first division is the longest and most decisive step of meiosis, where five stages called **leptotene, zygotene, pachytene, diplotene** and **diakinesis** are distinguished on morphological criteria.

The leptotene stage corresponds to the **condensation** of chromosomes with chromatin loops attached to an axial proteinic element; recombination starts; the ends of the chromatids (**telomeres**) are connected to the nuclear membrane,

analyses, which have played a crucial role in our understanding of the mechanisms of meiotic functions (and dysfunctions).

3 A cell cycle has three phases after mitosis: a G1 phase in which the total number of chromosomal DNA molecules is *n* (haploid) or 2*n* (diploid), etc, followed by a DNA replication phase, called S, itself followed by a G2 phase, during which each chromosome is duplicated into two chromatids bringing the total number of chromosomal DNA molecules to 2 x *n* (haploid) or 2 x 2*n* (diploid), etc., *n* being the haploid number of chromosomes in a given species.

forming a structure called a "bouquet". At the *zygotene* stage, polymerization between the two axial elements of two homologous chromosomes takes place, forming a protein complex (**synaptonemal complex**) from the sites where recombination progresses, leading to the **pairing** of these chromosomes. At the *pachytene* stage, the chromosomes are fully linked by the synaptonemal complex and recombination is completed. At the *diplotene* stage, the complex depolymerizes, the homologous chromosomes are only retained together by the exchange sites (cross-over) between chromatids, called ***chiasmas****. At the *diakinesis* stage, the chromosomes are located near the nuclear envelope.

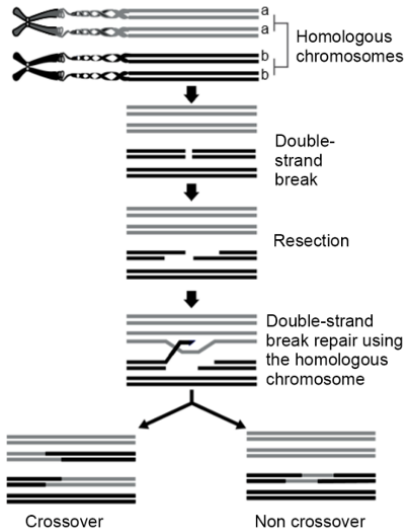
The next steps in the first meiotic division are ***metaphase I****, during which the nuclear envelope disappears (except in some phylums, such as fungi) and sister chromatids are linked to the cellular spindle by protein complexes located at the **centromeres**, while remaining attached to one another; ***anaphase I****, during which homologous chromosomes (each composed of two sister chromatids) migrate to the two opposite poles of the cell and **telophase I** during which the nuclear envelope is reformed and the chromosomes partially de-condensate before the second division begins simultaneously in each of the two nuclei (already isolated in two cells in some species).

The second meiotic division enters directly into ***metaphase II****, during which each sister chromatid is positioned on the spindle in opposite orientation to each other, followed by an ***anaphase II****, during which the cohesion of the centromeres of the sister chromatids disappears and each migrates to a cellular pole leading, finally, to 4 haploid cells. In phylums where all 4 cells are viable, we speak of tetrads.

Box 3.2. *The phases of meiosis*

During meiosis, the balanced distribution of parental chromosomes between haploid daughter cells is based on the formation of characteristic structures, called ***bivalents**** before the reductional division. These bivalents associate in pairs the chromosomes to be separated (after their condensation). Each pair includes two duplicated homologous chromosomes, each consisting of two ***chromatids****, themselves formed by a double DNA helix. These bivalents are the result of a succession of steps: the first shortly after the replication phase consisting of multiple cuts of the two strands (double-strand breaks) along the four DNA molecules (for each pair of chromosomes) while the condensed chromatids align with each other via specialized proteins (De Massy 2013). These programmed cuts are carried out by a meiosis-specific protein complex in which one of the partners, SPO11, acts like a molecular scissor. These cuts are transient. In 95 to 99% of cases, the strands are welded again identically (after equilibration of DNA super-coiling). But in a few

cases (usually at least once per pair of chromosomes) a connection is established between the two homologous chromatids, resulting in a reciprocal exchange between them. These exchanges or ***crossovers**** are essential for the formation of bivalents and therefore for the proper distribution of homologous chromosomes between daughter cells at reductional division. In mutants that no longer have *SPO11* activity and therefore no longer form DNA double-strand breaks and crossovers, division is disrupted, leading either to cell death (***apoptosis****) or to an unequal distribution of chromosomes in daughter cells. In both cases, there is sterility.



The figure symbolizes two **homologous chromosomes** (gray and black) each duplicated into two **chromatids** at the time of meiosis. A **double-stranded break** in the DNA can, after resection, lead to a potential **crossover**. In the case of crossover (left), there is a reciprocal exchange between two homologous chromatids.

In the vast majority of cases (95 to 99%) double-strand cuts do not lead to a crossover (right). What happens then? If the homologous chromosomes are different in the region of the cleavage site, one of the four DNA molecules will not exactly recover its original sequence, as revealed by DNA sequencing methods: traces of this cleavage will be found locally through the presence of a unilateral copy of genetic information (genetic **conversion**), as shown in the figure above (right).

Box 3.3. *Crossovers and non-crossovers*

Where do these molecular actors which make the cuts to DNA come from? For well-studied species, which represent only a small part of the tree of life, researchers first had difficulty answering this question because the proteins involved are very poorly conserved. It is with the studies of archaea (that do not make meiosis) that the problem started to be clarified. A topoisomerase, called TopoVI B, had been discovered there, involved in the release of the superhelicity of DNA molecules during replication and transcription. Imagine a long string, several hundred million times longer than its thickness, wrapped on small drums and from which you would have to separate the two strands, without making knots! Regular cuts are necessary, as long as they can be repaired: this is what topoisomerases do. It was then hypothesized that, in Eukaryotes, the complex that cuts DNA in meiosis would look like these topoisomerases. It was still necessary to explain how the same complex could be composed of proteins with such variable sequences in different taxonomic groups, associated with the highly conserved SPO11 protein. Finally, it is through the study of the three-dimensional structures of these proteins that it has been possible to make connections between the meiotic protein complex and the topoisomerase of archaea (Robert *et al.* 2016; Vrielinck *et al.* 2016). Despite large sequence differences, the predicted structures are remarkably well conserved. However, the functions are not identical, because the cut is ephemeral in archaea and repaired directly by the topoisomerase complex, whereas in meiosis it is only repaired by the interaction with other DNA molecules and with the help of other proteins. This similarity of the actors suggests that the nuclear genome of eukaryotes would originate from the archaea.

Do crossovers have any other purpose than to facilitate the halving of the number of chromosomes to produce gametes? It is striking to note that the transmission of the hereditary message by gametes, which one would imagine to be faithful, paradoxically begins with a multitude of DNA cuts, prone to inducing mutations! While crossovers are quantitatively restricted compared to the number of initial DNA cuts, they are the ones that generate new allelic associations (see Box 3.3). But they are also the ones that break up pre-existing associations. We are therefore faced with a dilemma, the interest of novelty in creating new associations, but the risk of destroying pre-existing favorable associations. It is therefore not surprising that the "moderate usage" of crossovers is genetically controlled. First, crossovers are not uniformly distributed along the chromosomes. There is often a preference for regions of the chromosomal arms far from the centromeres. Second, there are mutations, therefore genes and their products, which

modify the number of crossovers, or change their position. Finally, the frequency of crossovers varies with the stress caused by external conditions. Chromatid elongation, DNA methylation, chromatin chemical modifications and RNA splicing are all factors involved in this modulation. Especially in plants, given their fixation in soil, this plasticity can have adaptive consequences by regulating the genetic diversity of gametes and therefore of offspring. Finally, it should be noted that, despite the generality of the meiotic process, some species show particularities for certain stages. For example, the male Drosophila or the female silkworm pair their chromosomes without this resulting in any crossover.

A number of anomalies on which we will not expand can occur during meiosis. For example, a crossover may involve two different loci of the homologous chromatids, resulting in an unequal exchange between chromosomes, one undergoing a loss (deletion), the other a gain (duplication) of the fragment between the two loci. The duplicated copies of genes may evolve differently: forming a possible source of gene family. After fertilization, the lost region may reveal a genetic anomaly present on the corresponding region of the homologous chromosome: this is called ***hemizygosity****. Other anomalies concern the distribution of chromosomes during meiosis, resulting in gametes with chromosomal assortments that are either incomplete, and often lethal, or excessive, as in the case of certain trisomies (chromosomes 18 or 21) in humans.

3.4. Genetic determinism of sex

3.4.1. *From gametes to sex*

In a haplobiontic cycle, the same haploid organism can generate the two gametes that will carry out fertilization[4]. The resulting diploid cells will therefore be completely homozygous and, during the next meiosis, all haploid products will be genetically identical to each other and to the parental haploid organism. The sexual reproduction cycle is therefore carried out without any genetic exchange. It is the equivalent of a clonal reproduction interspersed with fertilizations and meioses, but without genetic consequences. Some fungi reproduce in this way. On the opposite,

4 The two gametes can be differentiated: we speak of “anisogamy”, or on the contrary undifferentiated, “isogamy”.

fertilization may only be possible between gametes from two different haploid organisms because of self-incompatibility. The resulting diploid cells will therefore have a certain probability of being heterozygous for some (or many) loci. This probability depends on the genetic differences between the two haploid parents. During the next meiosis, recombinants will be able to appear among the haploid products, with different genotypes from each of their two parents.

There are very varied situations for the production of male and female gametes in the case of the diplobiontic cycle. Some diploid organisms are ***hermaphrodites****, with male and female organs coexisting in the same individual, either simultaneously (snails, most flowering plants) or successively (fish). In other cases, these organs are carried by different individuals, whose sex, male or female, is determined genetically (sex chromosomes), or by environmental influences (temperature for example). It should be noted that sexual dimorphism leads to two types of meioses for the same organism, those that produce male gametes and those that produce female gametes. Their modalities (cross-over rate, frequency of chromosomal accidents, etc.) may be different, as is the case in the human species.

3.4.2. *Sex determinism in animals*

The mechanisms that determine whether a gamete is male or female, whether the organism that produces it is itself of one sex or another, are extremely diverse in eukaryotes and change rapidly on the evolutionary scale, but remain essentially genetic. In the simplest case, two homologous chromosomes carry a determinant of either one sex or the other. These chromosomes are referred to as sex chromosomes, and can be so differentiated from each other that they become morphologically different, like the X and Y chromosomes of mammals. As early as 1891, **Hermann Henking** (Henking 1891) noticed the presence of a chromosome that was not involved in meiosis in the male of the hemipterous insect *Pyrrhocoris apterus*, commonly called a “firebug”. This accessory chromosome later came to be designated by the letter X. In 1905, **Netti Stevens** (Stevens 1905), in the mealworm beetle *Tenebrio molitor*, discovered the smaller, male sex-linked Y chromosome and in the same year **Edmund Beecher Wilson** (Wilson 1905) obtained similar results in several types of hemiptera.

These discoveries, later extended to mammals and birds, supported the then emerging chromosomal theory of heredity. In mammals, the female sex is ***homogametic**** XX and the male sex is ***heterogametic**** XY with the Y chromosome necessary for male differentiation. In birds, the system is reversed and another pair of chromosomes differentiates into a Z chromosome and a W chromosome, so that males are homogametic ZZ and females heterogametic WZ. In Drosophila, the female sex is determined by the presence of two X chromosomes, so it is homogametic XX, and the male sex is determined by the presence of a single X chromosome, the Y chromosome being neutral, so that we have males which are XY and even XO where the Y chromosome is absent. But in other insects, such as lepidoptera, in some houseflies and crickets, females are heterogametic WZ, W carrying the female determinant and males are homogametic ZZ. Note the case of bees, where a particular locus determines sex, the females being heterozygous at this locus, the males being homozygous or more generally hemizygous, because they come from unfertilized eggs and are therefore haploid. Finally, in cold-blooded animals, a continuum of genetic and environmental mechanisms determines sex, for example in turtles or crocodiles, where it depends on the incubation temperature of the egg. So where is the logic (Bachtrog *et al.* 2014)?

Functionally, it was in 1990 (Sinclair *et al.* 1990) that the *SRY* gene carried by the human Y chromosome was identified. It is responsible in all mammals for testicular differentiation from undetermined gonads. This gene produces a transcription factor that activates, like a domino effect, other genes responsible for this differentiation, including that of the genital organs, and for suppressing the female differentiation pathway. The system is old, since it has remained stable since the origin of mammals. In birds, it is another transcription factor, produced by the *DMRT1* gene, carried by the Z chromosome, that determines sex by its dosage: two copies in males, only one in females. Genes from the same family are involved in the sexual differentiation of many organisms such as insects (Drosophila) or nematodes (*Caenorhabditis elegans*). It is important to note that in birds, the application of estrogens to ZZ eggs produces females. This hormonal susceptibility is found in other oviparous animals, such as reptiles. There is often also a susceptibility to temperature during embryonic development, even when well-differentiated chromosomes of type ZZ/ZW or XX/XY are present. Higher temperatures lead to females in some species and males in others. Global warming could affect the reproductive capacities of such populations.

Fish, from rays and sharks to the fugu, correspond to 400 million years of evolution and represent half of the *ca.* 60,000 described vertebrate species. How then have sex determinants of fishes evolved? As before, we can find a differentiation of sex chromosomes, but without continuity between similar species. For example, depending on the species of tilapia or stickleback, males may be either heterogametic of the type XY or homogametic of the type ZZ. Genes that regulate sexual identity evolve very rapidly. In medakas, the gene that initiates male development in XX/XY systems is variable within the same genus: *Dmy* in *Orysias latipes* and *O. curvinotus*, *Sox3* in *O. dancena*, *gsdf* in *O. luzonensis.* The latter gene, which is a member of the TGFβ growth factor family, has the same role in sablefish, *Anoplopoma fimbria*, belonging to a very different group of fishes. The sequencing of genomes shows that American pikes have lost the determinant they still shared with those of Europe only 220,000 years ago and have replaced it with another one that remains to be discovered: on the scale of evolution, this is a very short time frame!

Gene duplication and neo-functionalization (see Chapter 6) are the creative mechanisms of these sex-determination genes, but without long-term evolutionary continuity and in perpetual renewal. In this set of species, the same determinants may be found on different chromosomes. These chromosomes may differ from their homologs by restricting meiotic crossovers and losing the functionality of genes located in the chromosomal regions deprived of exchanges or, on the contrary, by retaining all their "evolutionary faculties", but losing the determinants in question. In addition, many of the nine orders of teleosteans show natural sequential ***hermaphroditism****, with alternating male and female phases in their adult lives. These alternations are caused, via the neuroendocrine system, by external stimuli such as temperature or the presence of congeners.

All this gives the impression of a great mess: simultaneous or successive hermaphroditism, separate sexes, variable genetic or environmental determinations do not show obvious evolutionary coherence (Herpin and Schartl 2015). In the face of such a variety, modern techniques of genomic analysis and editing will be able to shed light on these questions by more comprehensively identifying the determinants of sex in these groups of species. Moreover, shouldn't we also explore other evolutionary ***phyla**** to look for invariant elements?

3.4.3. *Sex determinism of brown algae*

Brown algae (Pheophyceae) are a group of multi-cellular eukaryotic organisms that have evolved independently of animals and plants for more than a billion years. They belong to the large group of Chromalveolata. It is the phylum that shows the greatest diversity of types of reproductive cycles and sex determination systems among eukaryotes. The rockweeds on our coasts, *Fucus vesiculosus* for example, have a diplobiontic reproductive cycle with differentiated male and female diploid individuals that produce gametes, respectively male and female, after two different meioses. ***Anisogamy**** is strong: the female gamete or "oosphere" is large, equivalent to an animal egg, while the male gamete is a flagellated spermatozoon. Fertilization takes place in the marine environment.

In another brown alga, *Ectocarpus siliculosus*, diploid individuals (called "sporophytes") produce U type (female) or V type (male) haploid spores during meiosis, which develop into morphologically very similar haploid gametophytes, but showing strong differentiation in the activity of many genes[5]. These gametophytes produce, respectively, the gametes U or V, which are quite similar, but play complementary roles during fertilization. In brown algae, ***isogamy**** (male and female gametes of the same size and shape) therefore coexists with anisogamy (female gametes larger than male gametes) and ***oogamy**** (spherical female gametes, voluminous and full of reserves, and male gametes in the form of a flagellate spermatozoon). It is not known what determines the size of the gametes, especially since the evolution among the different orders of Phaeophyceae can be in the oogamy to isogamy direction, as well as in the opposite direction. So what about the sex chromosomes in these organisms?

Whether in the genus *Fucus* or *Ectocarpus*, the genetic determinants of sex are found on a pair of chromosomes that have the properties of sex chromosomes as seen in animals: low gene density, accumulation of repeated DNA sequences and absence of meiotic cross-over in the sex-determining region (even if the latter is small as in *Ectocarpus*) (Ahmed *et al.* 2014). Beyond this region, there is an accumulation of genes

5 The homologous chromosomes, U and V, differ from each other by a central region containing more than a dozen genes. These are not different alleles, but totally different genes. We are talking about "idiomorphs".

differently expressed in the diploid phase (***the sporophyte****) compared to the haploid phase (***the gametophyte****). This differentiation between the U and V chromosomes has been established for a hundred million years. By experimentally producing UUV or UUUV gametophytes that happen to be male, it is shown that the V chromosome carries a determinant of male function, which is dominant over female function. A gene encoding a DNA binding protein, present in the region of the V chromosome that determines the male sex, is a good candidate because it is similar to the *DMRT1* gene, whose alleles are known to determine the sex of nematodes, insects, fishes and amphibians. It is also similar to analogous genes involved in sex determination in vertebrates, fungi or the green algae *Volvox*. Thus, under the versatility of physiological and anatomical modalities, some molecular constants appear at the genome level. But are the sex chromosomes really necessary?

This question comes immediately after the observation that can be made in the brown algae group as well as in the fish group: what is behind this diversity and volatility of sex determinism on the evolutionary scale? As we have seen, homologous sex chromosomes are distinguished from each other either by the necessary and sufficient presence of a single allele – the simplest case – or by the entire region surrounding the locus of the determinant, or by their entire length in which case one chromosome gradually stops genetic exchanges with its former homolog and evolves independently. It seems that the progressive differentiation pathway intrinsically contains a risk of disappearance of the chromosome that carries the sexual determinant as the function of its other genes is lost. This is the fate that seems to be that of the Y chromosome of mammals at the scale of evolution! The consequence is that all the genes carried by the X chromosome, numerous and significant in humans, are present in only one copy in males and therefore that the recessive alleles will be differently expressed between the two sexes. The other thing that the diversity of sex determinism illustrates is the tendency to escape "evolutionary senescence" by changing the determinant or changing the chromosome. A way of "staying young" in some way!

3.5. Clonal reproduction and its derivatives

The importance of sexual reproduction and the great diversity of its operational modes should not hide the existence of another mode of

reproduction in the whole living world, much more fundamental and general: clonal reproduction. A living cell, whether eukaryotic or prokaryotic, gives birth to two living cells without the help of any partner. And these cells will continue the process, giving birth to a ***clone**** that is a set of cells derived from the same ancestor cell by asexual multiplication. All cells of the same clone are therefore genetically identical to each other and to the ancestral cell of the clone, with the exception of mutations that can occur at each cell division. Clones can grow exponentially if, at each division, the two daughter cells retain the ability to divide or, on the contrary, linearly if only one of the two daughter cells retains this ability (this is the case for stem cells in differentiated tissues).

Many microorganisms have a clonal mode of reproduction that may be exclusive or, on the contrary, associated to some degree with sexual reproduction. In bacteria, the chromosome(s), and plasmid(s) if any, are distributed between the two daughter cells after replication without the benefit of the complex system of eukaryotic mitoses. There are other mechanisms to promote the equitable distribution of chromosomes between the two daughter cells. For plasmids with a high number of copies, a random distribution may be sufficient to ensure a high probability that both daughter cells will receive at least one copy. However, losses are observed in one or the other daughter cell, and more frequently when the number of copies is low. For eukaryotic cells, mitoses ensure an almost perfect distribution of sister chromatid pairs (from the duplication of a chromosome) between the two daughter cells. If plasmids exist, they will usually be treated randomly as in bacteria. The same applies to mitochondrial and chloroplastic genomes, which are generally composed of large enough number of copies to limit their random loss (see section 3.6).

Clonal reproduction also concerns certain multicellular organisms where it may be exclusive or associated with sexual reproduction. Algae and fungi use it frequently. Basal shoots, layering and bulbs are examples of this in plants. Some animals (e.g. coelenterates) also favor this mode of reproduction when conditions are favorable. Finally, it should not be forgotten that the development of multicellular organisms (even those whose reproduction is exclusively sexual) involves the clonal multiplication of cells from the zygote (if diplobiontic) or the spore (if haplobiontic). In humans, there are more than 40 mitotic cell generations between the zygote at the origin of the adult and its germ cells that will give birth to the future zygotes

of the next generation. To be precise, human reproduction is therefore sexual only once every 40 cell generations! This is the case for all multicellular organisms.

Clonal reproduction has the advantage of being more efficient than sexual reproduction, and is therefore used preferentially by fast-growing microorganisms and when conditions are favorable. The yeast *S. cerevisiae*, for example, undergoes about 100,000 mitotic cycles for a single sexual cycle. Some other yeast species have completely lost their sexuality. However, clonal reproduction has the disadvantage that any loss of a genetic element becomes irreversible. In most cases of evolutionary successful lineages, it is therefore observed that some degree of horizontal genetic exchanges exist, either between members of the same or related species (sexual reproduction or hybridization) or from totally different organisms living in the same environment (horizontal transfers).

3.6. The genetics of organelles

As we saw briefly in Chapter 1, the organellar mode of inheritance has long been defined negatively with respect to chromosomes. It was referred to as "non-chromosomal heredity" and even "non-Mendelian heredity", restricting the term "Mendelian" to the meiotic segregation of eukaryotic chromosomes. In reality, genes carried by mitochondrial and plastid DNA share all the properties common to all genes and are therefore perfectly Mendelian, but they have their own rules of hereditary transmission which are often significantly different from one organism to another. In addition, the number of genes contained in organelle genomes varies greatly among groups of organisms. The mitochondrial genome carries about 100 genes in *Reclinomonas americana*, a unicellular eukaryote belonging to the major group *Excavata*, compared to only five in *Plasmodium falciparum*, the agent of malaria, a member of the *Chromalveolata*. Similarly, the plastid genome can have up to 250 genes in red algae, compared to less than 20 in other *Chromalveolata* such as Dinoflagellates. In almost all cases, however, these genomes retain the genes for ribosomal RNAs and transfer RNAs that provide organelles with their autonomous protein synthesis within the eukaryotic cell.

3.6.1. *In unicellular eukaryotes*

Baker's yeast has played a major role in elucidating the rules of mitochondrial heredity from the moment that point mutations could be isolated instead of the massive genomic alterations of the previous "petite colonies" mutations. Some of these mutations conferred respiratory deficiency after inactivation of mitochondrial genes, others conferred resistance to drugs inhibiting the products of these same genes. These last mutations allowed classical genetic analyses by the quantitative examination of progenies after the crossing between pure parental lineages. It was found that mitochondrial heredity in yeast is biparental and that genetic recombination is omnipresent. In addition, it was found that parental and recombinant genotypes segregated very quickly from each other in diploid clones derived from zygotes. ***Heteroplasmic**** cells (containing simultaneously several mitochondrial genotypes) gave rise to mixtures of ***homoplasmic**** cells, of one type or another, in accordance with the idea that the various mitochondrial DNA molecules were more or less randomly distributed among the daughter cells during mitoses. The rapid formation of homoplasmic lineages during clonal multiplication also suggested that a limited number of mitochondrial DNA molecules give rise to the mitochondrial genome of each daughter cell. Accordingly, the meiosis of a homoplasmic cell produces four haploid products that are genetically identical to each other and identical to the mother cell in terms of the mitochondrial genome (Westermann 2014).

3.6.2. *In humans and animals*

The human mitochondrial genome (like all mammals) is a compact circular DNA molecule; genes touch each other (and may even overlap a little) and all are essential to cellular respiratory function. As in yeasts, some mutations are large deletions affecting several genes, others are point mutations. These mitochondrial DNA mutations can be responsible for severe pathologies in humans, most often affecting nerve and/or muscle cells. Since mitochondrial heredity is uniparental and maternal in most metazoans (with remarkable exceptions, such as mussels and a few other bivalves in which it is biparental), recombinants with paternal alleles are only exceptionally found. Sick children have inherited their mitochondrial genome from heteroplasmic mothers who are themselves generally healthy. The proportions of molecules carrying the mutated alleles may change

during early embryonic development and, beyond certain thresholds, trigger syndromes called "mitochondrial cytopathies". The severity of these syndromes varies according to the proportion of mutated mtDNA molecules, and of course also depending on the type of mutations (Rossignol *et al.* 1999).

3.6.3. *In plants*

In the vast majority of plant species, the genomes of cellular organelles are inherited exclusively from one of the two parents during sexual reproduction: the heredity is uniparental. In general, in mosses or ferns, plastids and mitochondria are inherited from the female gamete. In gymnosperms, if the mitochondrial genome comes from the female gamete, the plastid genome is inherited from the male gamete. In angiosperms, both organellar genomes are generally inherited from the mother. However, there are notable exceptions. The melon (*Cucumis melo*) inherits its mitochondria from the male gamete and its plastids from the female gamete. In alfalfa, if the mitochondria originate from the female gamete, the plastids are in mixture in the zygote, coming from both parents, with those of the male gamete predominating later in development. In pelargonium, there is no exclusion, paternal and maternal plastids and mitochondria are found in the embryo: this is a case of biparental heredity.

This situation has made it possible to identify genetic recombination between the mitochondrial genomes of both parents (Apitz *et al*. 2013). Such recombination had been achieved in the past by forcing two types of organelles to find themselves in the same cell after somatic cell fusions (Belliard *et al*. 1979). Genetic exchanges are then possible thanks to the continuous fusions and fissions of mitochondria that bring mitochondrial DNA molecules into contact with each other. Obtaining cells derived from recombination between plastid genomes requires the application of a high selection pressure. This is achieved, for example, by isolating two plastid mutants resistant to two different antibiotics, forcing their coexistence by somatic fusion and selecting *in vitro* double resistant cells by culture in the presence of both drugs (Medgyesy *et al*. 1985).

There are several mechanisms responsible for uniparental heredity: the physical exclusion of organelles when the differentiation of the male gamete "crowds them out", or active, genetically controlled degradation processes,

which eliminate, for example, the plastids of male gametes in potatoes or the plastids of female origin in the embryo of gymnosperms.

During the development of the plant, the genomes of the organelles replicate and are transmitted at each cell division without risk of being lost, because they are in numerous copies in a cell. When a mutation that is not harmful occurs in an organelle genome, a genetically different sub-population will gradually form and cohabit with the original population. But this heteroplasmy is transient, because during development, a stochastic purification process follows, the duration of which depends on the average number of organelle genomes per cell. The final result will consist of two sets of cells, one that exclusively carries the mutation and the other that does not. This process is particularly visible in a plant where a chloroplastic mutation appears, which will lead to white areas on branches or leaves.

3.7. Important ideas to remember

– Each DNA molecule, wrapped around particular proteins, forms a **chromatid** which, in eukaryotic cell nuclei, alternates between a relaxed structure in interphase of the cell cycle and a condensed structure during cell division. A nuclear chromosome consists of a single chromatid before DNA replication (G1 phase) and two chromatids between DNA replication and cell division (G2 phase).

– The sexual reproduction of eukaryotic organisms proceeds by an alternation of two phases, of varying importance depending on the case: the **haplophase** which starts after meiosis and leads to the formation of gametes, the **diplophase** which starts after fertilization between two gametes and continues until the next meiosis. Within each phase, cell reproduction is **clonal**.

– Meiosis allows recombinations (crossovers) between homologous chromosomes and leads to the re-assortment of parental alleles in gametes. Meiosis and fertilization ensure the **genetic shuffling** within populations.

– **Sex** determinism shows an astonishing evolutionary variability whose genetic components differ according to the categories of organisms.

– Depending on the organisms, reproduction is clonal or sexual in an exclusive or non-exclusive way.

– The heredity of mitochondrial and plastidial genomes is biparental or uniparental, **heteroplasmic** cells produce clones composed of different **homoplasmic** cells.

3.8. References

Ahmed, S., Cock, J.M., Pessia, E., Luthringer, R., Cormier, A., Robuchon, M., Sterck, L., Peters, A.F., Dittami, S.M., Corre, E., Valero, M., Aury, J.M., Roze, D., Van de Peer, Y., Bothwell, J., Marais, G.A., Coelho, S.M. (2014). A haploid system of sex determination in the brown alga. *Ectocarpus sp*. *Current Biology*, 24, 1945–1957.

Apitz, J., Weihe, A., Pohlheim, F., Börner, T. (2013). Biparental inheritance of organelles in *Pelargonium*: Evidence for intergenomic recombination of mitochondrial DNA. *Planta*, 237, 509–515.

Bachtrog, D., Mank, J.E., Peichel, C.L., Kirkpatrick, M., Otto, S.P., Ashman, T.-L., Hahn, M.W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J.C. (2014). Sex determination: Why so many ways of doing it? *PLOS Biology*, 12(7).

Belliard, G., Vedel, F., Pelletier, G. (1979). Mitochondrial recombination in cytoplasmic hybrids of *Nicotiana tabacum* by protoplast fusion. *Nature*, 281, 401–403.

De Massy, B. (2013). Initiation of meiotic recombination: How and where? Conservation and specificities among eukaryotes. *Annual Review of Genetics*, 47, 563–599.

Henking, H. (1891). Untersuchungen fiber die ersten Entwicklung – vorgfinge in den Eieren der Insekten. *Z. Wiss. Zool.*, 51, 685–736.

Herpin, A., Schartl, M. (2015). Plasticity of gene-regulatory networks controlling sex determination: Of masters, slaves, usual suspects, newcomers and usurpators. *EMBO Reports*, 16, 1260–1274.

Medgyesy, P., Fejes, E., Maliga, P. (1985). Interspecific chloroplast recombination in a *Nicotiana* somatic hybrid. *Proceedings of the National Academy of Sciences*, 82, 6960–6964.

Robert, T., Nore, A., Brun, C., Maffre, C., Crimi, B., Bourbon, H.M., de Massy, B. (2016). The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science*, 351, 943–949.

Rossignol, R., Malgat, M., Mazat, J.-P., Letellier, T. (1999). Threshold effect and tissue specificity implication for mitochondrial cytopathies. *Journal of Biological Chemistry*, 274(47), 33426–33432.

Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L., Smith, M.J., Foster, J.W., Frischauf, A.M., Lovell-Badge, R., Goodfellow, P.N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, 346, 240–244.

Stevens, N.M. (1905). *Studies in spermatogenesis with especial refrence to the "accessory chromosome"*, 36. Carnegie Institution of Washington, Washington D.C.

Vrielinck, N., Chambon, A., Vezon, D., Pereira, L., Chelysheva, L., De Muyt, A., Mézard, C., Mayer, C., Grelon, M. (2016). A DNA topoisomerase VI-like complex initiates meiotic recombination. *Science*, 351, 939–943.

Westermann, B. (2014). Mitochondrial inheritance in yeast. *Biochimica et Biophysica Acta*, 1837, 1039–1046.

Wilson, E.B. (1905). The chromosomes in relation to the determination of sex in insects. *Science*, 22, 500–502.

4

From Genetic Engineering to Genomics

In the early 1970s, the central dogma of molecular biology supplemented by the genetic code and by the mechanisms of transcriptional regulation in bacteria could give the impression to some that genetics had achieved its goal and that there was nothing important left to discover. In reality, no gene had yet been examined at the molecular level and the discoveries that followed show how false this impression was. A huge field of investigation opened up in which surprises were not going to be rare. Some of these have already been mentioned in Chapter 2. This chapter attempts to put them back in historical order by emphasizing the technical and conceptual changes that have transformed genetics since the early 1970s.

4.1. Restriction of DNA

The first discovery that gave genetics its molecular dimension was that of type II restriction ***endonucleases**** from bacteria, the first tools that allowed us to cut DNA molecules at specific sites defined by the sequence (Smith and Nathans 1973). It's a strange name, but let us go back a few years. A curious phenomenon, called "host restriction", had been discovered by studying the infection of *E. coli* by bacteriophages (Arber and Linn 1969). In some experiments, the bacteriophages resulting from an infection appeared to retain the "memory" of the bacterial strain that had produced them, in such a way that they became unable to infect other strains. This was surprising because, by reversing the order of the infected strains, there was no symmetry, the produced bacteriophages were able to infect all strains.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

After multiple experiments, it became clear that bacterial strains could "modify" the DNA of the produced bacteriophages and destroy the DNA of infecting bacteriophages if it was not properly "modified". Some bacterial strains did both (they were called r^+m^+), others modified without destroying (r^-m^+), and still others did nothing (r^-m^-). This ***restriction-modification**** system lead to the discovery of two categories of enzymes. The first (methylases) modify DNA by methylating[1] it at a particular site (a short sequence). The second (endonucleases) cut the DNA if it is not methylated at these sites, but ignore it if it is[2]. The restriction/modification sites in a DNA molecule correspond to specific sequences of a few nucleotides (typically 4 to 6). In the first restriction-modification system discovered, called type I, the cleavage of DNA by the endonuclease occurs at variable distances from the site that should have been methylated, producing fragments of variable sizes, dispersed around an average, therefore difficult to manipulate.

Subsequent discoveries of other restriction-modification systems, known as type II, brought decisive changes because, in them, the endonucleases recognize the same sites as the corresponding methylases and cleave DNA precisely at these sites or in their immediate vicinity. By applying these endonucleases *in vitro* on purified DNA molecules of a particular organism, DNA fragments of defined sizes were obtained, precisely limited by the presence of their restriction sites in the sequence. Such fragments may be separated from each other by electrophoresis. It was quickly recognized that the diversity of bacterial species offers a wide variety of methylases and endonucleases with different sites. By purifying many such endonucleases (nicknamed "restriction enzymes" in jargon), we became capable in the early 1970s, for the first time in history, of precisely cutting DNA molecules at specific sites, predetermined by their sequence (which we did not yet know). This paved the way for their manipulation *in vitro*.

1 Depending on the methylase, methylation occurs on carbon 6 or adenine (the most common case in bacteria) or carbon 5 of cytosine. But methylation only occurs on short, precise sequences, because methylases are associated with other proteins that ensure sequence recognition. In Type I restriction-modification systems, methylases, endonucleases and specificity proteins are part of a single multi-protein complex. In Type II restriction-modification systems, methylases and endonucleases act separately, each being capable of ensuring the specificity of DNA sequence recognition.

2 This is the general case; the opposite also exists.

But genomes are large (see Chapter 5) and the molecular weight of DNA is high[3]. The purification of a particular restriction fragment directly from the total DNA of a genome using standard biochemical methods can therefore only result in quantities of molecules far too small to be usable by the methods then available. The above discoveries would therefore have had only a limited impact without the arrival of recombinant DNA.

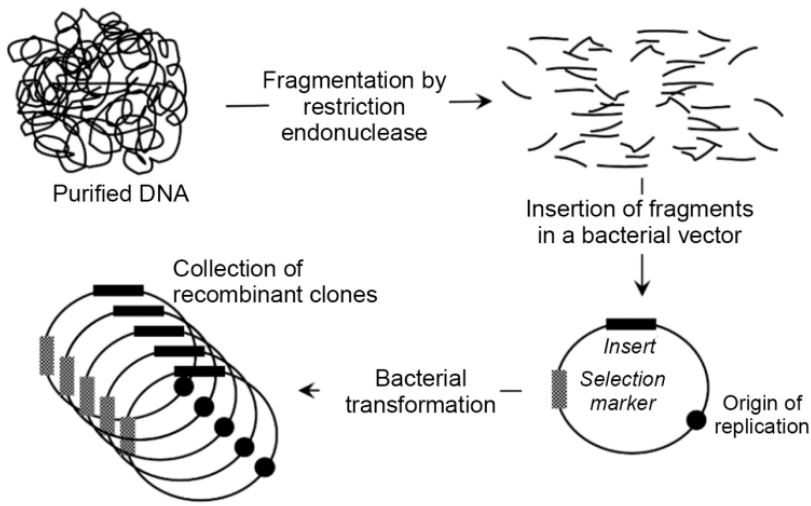
4.2. Recombinant DNA and the birth of genetic engineering

The purification of other bacterial enzymes, called "DNA-ligases", opened up this new era. Using co-factors that provide them with energy, these enzymes are able to weld the cleaved ends of DNA molecules and thus form recombinant molecules *in vitro*. The first successful reaction of this type was carried out in **Paul Berg's** laboratory in 1972 (Jackson *et al.* 1972). Fragments of DNA from the λ bacteriophage and from the galactose operon of *E. coli* were ligated with the DNA of the SV40 virus, which infects mammalian cells. The opposite, fragments of SV40 DNA ligated into the λ bacteriophage, was obviously possible and from this stage, taking into account the knowledge on bacteriophages, it became possible to integrate in *vitro* any DNA molecule of one's choice into the DNA of a bacteriophage which, during its natural multiplication in *E. coli*, would multiply the foreign DNA with it. The notion of artificial recombinant DNA was born. Bacteriophages were the first ***vectors****. Placed on suitable culture media, the infected bacteria provided the machinery, precursors and energy necessary for the propagation of recombinant bacteriophages, which thus formed clones. The notion of gene cloning was born.

Natural bacterial plasmids quickly became other good vectors, tending to quickly replace bacteriophages; later, artificial vectors were constructed from them to facilitate the manipulations (see Box 4.1). By ensuring that each transformed bacteria received only one recombinant DNA molecule and allowing it to multiply on an appropriate culture medium, bacterial clones were obtained, each containing multiple copies of the DNA fragment carried by the vector. The gene cloning operation therefore simultaneously allows the purification and the amplification of the relevant DNA fragment. From an entire genome, libraries of all DNA fragments of that genome cut at

3 For example, a double-stranded DNA molecule of 1,500 base pairs has a molecular weight of one million daltons, i.e. one microgram of this DNA (an amount usually manipulated) represents only one picomole (one thousandth of a billionth of a mole).

specific sites could be built, each bacterial clone amplifying a particular fragment. With molecular probes (see Chapter 2, Box 2.7) or appropriate experimental methods (there were many of them), it was then possible to search for fragments carrying the gene that one wished to study or manipulate. The success or failure of the operation was often decisive for the pursuit of young researchers' careers, as there was still so much to discover in a field that was finally opening up to exploration: the possibility of manipulating genes *in vitro*.



The discovery of type II bacterial **restriction endonucleases** has paved the way for DNA fragmentation between specific sites. However, from an entire genome, the fragments are generally far too numerous and each of them in too small quantities to be manipulated *in vitro*. The development of methods for the production of **recombinant DNA** *in vitro* has solved the dual problem of purifying each fragment and amplifying it, giving rise to **genetic engineering.**

As a result, it became possible to build recombinant DNA **libraries**, representing all or part of a genome to be studied, and to have access *in vitro* to the genes or their fragments in order to study or modify them according to specific needs. A great many discoveries followed.

Subsequent discoveries of endonucleases with very high specificity for DNA sequence recognition have paved the way for genome manipulation *in vivo*.

Box 4.1. *Recombinant DNA, nucleases and gene cloning*

With these techniques, molecular exploration of genes began, and the number of genes studied in different organisms increased rapidly, while analytical methods were being refined (electrophoresis of DNA fragments, radioactive labeling, molecular hybridizations, etc.) and gene transfer from one organism to another became possible (in practice, it was only the transfer of genes into bacteria). Arguably the most important fundamental discovery made at that time was that of ***introns**** (see Chapter 2). Genes still had a lot to teach us at the molecular level and the research in this area was accelerating. However, they would probably not have succeeded if another technological revolution had not come at the right time, that of DNA sequencing.

4.3. Sequencing of biological macromolecules

DNA sequencing would not have existed without the recombinant DNA methods on which it depends entirely, but it would probably not have existed either without the prior knowledge on sequencing the other two categories of biological macromolecules carrying information: RNA and proteins. Another brief retrospective look is useful. As soon as it was understood that DNA was a long polymer of nucleotides – in the early 1950s – and then that the genetic code was deciphered– in the 1960s – it became clear that the secrets of heredity lay in the sequences, that is in the order of succession of the elementary components of macromolecules: nucleotides for nucleic acids and amino acids for proteins (Sanger 2001). But determining these sequences, even very short ones, was still an inaccessible dream with the tools then available. It was proteins, not DNA, that became the first sequenced macromolecules, because the physical-chemical diversity of their amino acids (20 in total) was better suited to the biochemical methods available in the early 1960s. But, of course, proteins do not have genetic continuity and, since the genetic code is degenerate (see Chapter 2), it is impossible to deduce the sequence of a gene from that of its protein product among the many possible combinations[4]. The need to sequence nucleic acids therefore remained intact.

4 It should be noted, however, that a short peptide sequence gives rise to a sufficiently limited number of combinations to allow the synthesis of the corresponding oligonucleotides. This technique played a major role during the gene sequencing period.

Contrary to what current techniques might suggest, it was RNA molecules, not DNA molecules, that became the first sequenced nucleic acids. Again, because these molecules could be purified by biochemical methods and because enzymes – **ribonucleases** – were available to cut them at specific sites. It should be recalled that the discovery of DNA restriction endonucleases dates back only to the 1970s (see section 4.1). The first fully sequenced RNA molecule was the yeast alanine transfer RNA, completed in 1965 (Holley *et al.* 1965), followed two years later by *E. coli* 5S RNA (Brownlee *et al*. 1967), and five years later by the capsid gene (RNA) of the MS2 bacteriophage (Jou *et al.* 1971). Each time it was a considerable amount of work, involving a number of specific tricks that were difficult to generalize. But this time, we could expect to have the exact and unique image of the corresponding genes – in theory at least, because at that time the phenomena of RNA editing and splicing were unknown (see Chapter 2). In addition, the first two sequenced RNA molecules were from non-coding RNA genes and the third was one of the few examples of genomes consisting of RNA instead of DNA. Note, however, that the complete sequencing of the MS2 bacteriophage genome, made entirely of RNA, was the first sequenced genome and provided us with the direct confirmation of the accuracy of the genetic code (Fiers *et al.* 1975).

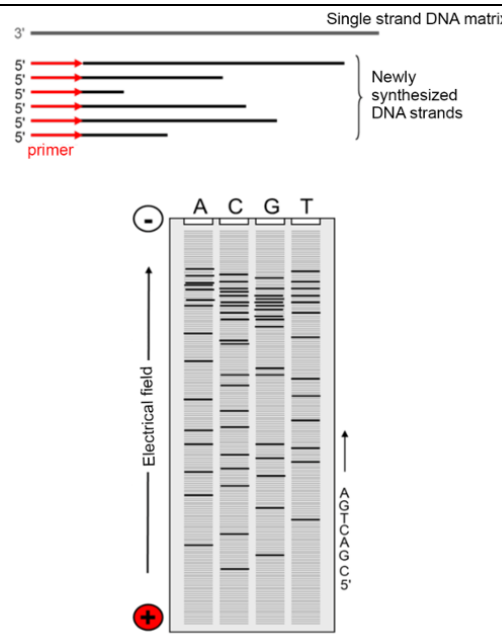
It was in 1977 that the sequencing of DNA molecules really began. Some assays, had been done before, but using very laborious techniques. By 1968, part of the cohesive ends of the lambda bacteriophage had been deciphered (Wu and Kaiser 1968). Ten nucleotides out of the twelve, obtained following a considerable amount of work that still involved specific tricks that could not be generalized. From 1977 onwards, two general methods, one chemical (Maxam and Gilbert 1977), the other enzymatic (Sanger *et al.* 1977), finally enabled us to read continuous sequence fragments of a few hundred nucleotides from purified DNA molecules. A first version of the enzymatic method had been published in 1975 (Sanger and Coulson 1975), thanks to which the DNA sequence of bacteriophage Φx174 was fully deciphered in 1977.

The chemical method was based on the partial degradation of DNA fragments, previously purified and labeled with a radioactive phosphorus atom at one end. Four different chemical reactions, each specific to a type of nucleotides, were performed. The enzymatic method was based on the *in vitro* synthesis of copies of DNA molecules. Four different reactions were

also performed, each leading to the partial interruption of DNA synthesis at the level of one of the four nucleotides. In both cases, the populations of radioactive molecules generated were analyzed by sufficiently resolving electrophoreses in order to separate molecules that differed in size from each other by only one nucleotide (see Box 4.2). By placing the four reactions side-by-side and navigating from one path to the other, the nucleotidic sequence of the DNA fragment studied could be read. In practice, this reading could extend to a few hundred nucleotides, no more. But the fragments of sequences obtained could then be assembled on paper, using their partial overlaps as a guide, to derive the gene sequences. We were still a long way from the sequencing of large genomes of entire organisms that began nearly 20 years later, but then started a period of sequencing genes of particular interest that has proven very productive in discoveries. The genes – or their fragments – were obtained by cloning in artificial bacterial vectors. They were identified by a variety of methods, including the possibility of synthesizing complementary DNA from messenger RNA molecules using reverse transcriptases *in vitro*. These complementary DNA molecules were labeled with a radioactive element and used as molecular hybridization probes to search for corresponding genes in recombinant clone libraries.

Conceptually, DNA sequencing was a real revolution. Instead of simply identifying genes from their mutations or products, as genetics had always done since its inception, it was now possible to identify them directly from their sequences and deduce those of their products. Instead of looking for mutations from phenotypic screens, the desired mutations could be made from the sequence and then examined for their consequences. We speak here of **directed mutagenesis** and **reverse genetics**. The development of DNA synthesis in the early 1980s considerably accelerated this powerful strategy.

With DNA sequences, new genomic elements were quickly discovered that were not accessible to conventional genetics because their mutation does not directly produce phenotypic changes. This is the case for ***pseudogenes****, nucleotide repetitions, traces of mobile genetic elements or viral sequences, mitochondrial or chloroplastic DNA segments (NUMT or NUPT) inserted in the nuclear genome, etc. (see Chapter 5). All these discoveries motivated the complete sequencing of genomes. But to achieve this, much strategic and technical progress was still needed.



The Sanger method consists of **synthesizing *in vitro*** complementary DNA strands from a short **primer** paired to a **single strand** DNA **template** to be sequenced. The primer is generally a synthetic **oligonucleotide**. The synthesis is carried out by a DNA polymerase which successively incorporates one by one the nucleotides complementary to the matrix strand. Four reactions are carried out in parallel with, in each case, a chemically modified nucleotide, the incorporation of which results in stopping the chain **synthesis** at the level of an A, a G, a C or a T. The low concentration of this **terminator** leads to a low probability of stopping the synthesis at each corresponding position, so as to obtain a family of molecules of partially nested sequences and different lengths which will then be separated by **electrophoresis** (see the diagram). The synthesized molecules carry a radioactive or fluorescent **marker** that allows them to be visualized after electrophoresis. The sequence can therefore be read by examining the molecules in the four channels in increasing size order, each one corresponding to a nucleotide (from bottom to top of the diagram). Many technical improvements have been made over the years, while maintaining the general principles of the method. The main one was the introduction of fluorescent nucleotides, which allowed automatic readings of the sequences in the form of **fluorograms**, each color corresponding to a nucleotide, as shown in www.iste.co.uk/dujon/genetics.zip.

Box 4.2. *DNA sequencing using the Sanger method. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

4.4. The beginnings of genomics: the very first genome sequences

Apart from the MS2 bacteriophage, whose genome of 3,569 RNA nucleotides was fully deciphered in 1975 in **Walter Fiers'** laboratory in Belgium (see section 4.3), it was another *E. coli* bacteriophage, ΦX 174, that became the first DNA genome ever sequenced. This was in 1977. This sequence, published by F. Sanger's laboratory in England, included only 5,375 nucleotides, but it paved the way for a development that has continued until today.

Special mention should be made for the sequencing of two mitochondrial genomes by F. Sanger's laboratory, first that of humans (16,569 nucleotides, published in 1981), then that of cattle (16,338 nucleotides, published in 1982), because, in addition to being the longest genomes at the time, they played a very important role in the development of new concepts in genomics (Anderson *et al.* 1981). Indeed, of the 13 protein-coding genes predicted from the sequence of the human mitochondrial genome, 8 looked like nothing previously known. Today, we speak of "orphan" genes to describe this situation. They are found in all new genomes sequenced to date. The other five genes were identifiable by their similarity to genes in the mitochondrial yeast genome, which had just been sequenced and whose function had been determined by extensive genetic work carried out in the 1970s. This gave rise to the notion of functional identification by sequence similarity, a method which, without replacing direct experimental demonstrations, considerably accelerates research on one organism by benefiting from the work done on other organisms. This is one of the hallmarks of modern genomics.

Subsequently, seven of the eight orphans, then known as URFs (unknown reading frames), were found to encode the mitochondrial proteins of the respiratory chain complex I, an essential function in most organisms, but absent in yeast. Thus, the idea was born that beyond the sequence divergences between homologous genes, it was the presence/absence of genes which is important in genome diversity. As a corollary, the identification of gene functions by sequence similarity, now called genome ***annotation****, depends on the knowledge already available in other biological systems. Beyond these small genomes themselves, it was the heuristic power of comparison between genomes that was clearly appearing as an essential complement to the experimental methods of molecular biology. An

intellectual schism was burgeoning. By comparing the mitochondrial genomes of humans and cattle, we could directly see their evolution on the corresponding time scale. Note in passing that it had also been discovered that, contrary to what had previously been imagined, the genetic code was not universal (the same observation had been made at the same time with yeast mitochondrial genes). The differences were limited in number, but of great functional importance (the UGA codon was not a stop, but coded tryptophan, the AUA codon coded methionine and not isoleucine). Other differences were to be discovered later.

Finally, we started to measure the importance of computer methods to exploit the sequences. At that time, however, sequence databases were only in their infancy, the results of gene sequencing work, not of genomes. But sequence comparison algorithms were beginning to develop. They would play a considerable role, with each new sequenced gene being immediately compared to all the sequences already available. In return, the sequence of this new gene was increasing the content of the databases in a virtuous circle, in principle, although (as already mentioned) focusing on what is already known.

4.5. The trigger

In the early 1980s, very few laboratories thought of sequencing entire genomes, with the exception of viruses, addressed by increasing size. The focus was on cloning and sequencing specific genes, previously identified by their function. To many, genes were more interesting for their products than for themselves. Of course, their sequences gave us the sequences of their products more efficiently than by sequencing the proteins themselves. Moreover, the gene sequences gave us the nature of the mutations that altered them and the elements of their immediate environment suspected to control their expression. But we remained in the classical intellectual schema of Mendelian genetics. The aim was to identify the genetic determinants of biological mechanisms considered to be of interest for study, without much concern about whether genomes would not carry other secrets in themselves.

Of course, we knew that the genomes were too large, because the amount of DNA present in the cells was always much higher than expected when considering the numbers of protein species estimated in these same cells. It was considered as a paradox. Even counting the introns, we stayed very far from the account. The idea of “junk DNA” then germinated: DNA without a

defined function and therefore certainly useless. Furthermore, since the techniques remained laborious and expensive, sequencing entire chromosomes seemed to many, not only out of reach, but above all futile.

It was in this context that in the mid-1980s, some enlightened minds came up with the idea of sequencing the human genome! The project seemed grandiose, comparable to the conquest of the moon or the construction of the Great Wall of China. It was a challenge because we didn't know how to do it well, hence there was no shortage of criticisms. The first problem encountered was the cost. Even with only one dollar to sequence a nucleotide, as W. Gilbert predicted (which was far below reality in laboratories at the time), more than $3 billion needed to be spent. Were there not other more sensible projects for this price? Then, the technical feasibility was a problem. DNA fragments a few hundred nucleotides long could be sequenced. But how to assemble these fragments into a single sequence of billions of nucleotides? Maps were not available to indicate their order and orientation. Building them required not only considerable work, but also a rare degree of expertise. In addition, some tools that later proved indispensable, such as *BAC* (Bacterial Artificial Chromosomes) vectors, had not yet been invented. And, who would do the work? If it were to be distributed, who would coordinate it and how? The idea was to build very large sequencing centers to centralize the work. But then existing laboratories would be excluded: the very ones where the scientific knowledge, technological know-how and heuristic potential were located. And who would work in these centers?

After many discussions and scattered attempts, and as technology and knowledge progressed rapidly, a global program, called the Human Genome Project and coordinated by the NIH in the United States, was launched in 1990. It marked the real beginning of deciphering the genetic heritage of *Homo sapiens*. This program was officially declared complete in 2004 (see section 4.7). During that time, many unforeseen events would change the order of things, and the human genome would not be the first sequenced genome.

4.6. The impact of the first real genomes

Before 1990, ideas had already evolved considerably and it became clear that the genomes of micro-organisms, in addition to their intrinsic interest,

could serve as effective stepping stones to test the strategies needed to sequence the human genome. This idea had not imposed itself, though, and it took several other enlightened minds to bring it to life. Historically, two bacteria (1995), shortly followed by yeast (1996) and a few other bacteria and archaea (1996 and 1997) were the first genomes of autonomously living organisms to be sequenced.

The first two bacteria were two moderate human pathogens, *Haemophilus influenzae* (ENT infections and non-viral influenza syndromes) and *Mycoplasma genitalium* (inflammation of the urogenital tract). In addition to the small sizes of their genomes, the first criterion for their choice (1.8 and 0.6 million nucleotides, respectively), it was the sequencing method that proved the most important result. While in all laboratories, genomes were sequenced from cloned fragments, precisely ordered relative to each other along genomic maps, **Craig Venter** and his colleagues at TIGR[5] undertook to randomly sequence the clones and then assemble the fragments of sequences obtained by computer using partial overlaps (Fleischmann *et al.* 1995). We now speak of random sequencing (Whole-genome shotgun or WGS) as opposed to ordered sequencing. It was a revolution that required having solved several difficulties, in particular, developing algorithms capable of assembling thousands (and soon millions) of sequence fragments in a reasonable time, taking into account repeated sequences in genomes and the intrinsic imperfection of the sequence of each fragment (reading errors). Given the importance of the issue, increasingly powerful algorithms were subsequently developed. Then it was necessary to find strategies to fill the gaps, because you can never cover the entire surface of a target by shooting at random – you would need an infinite number of shots.

This remains one of the main problems of the WGS strategy, now universally adopted for technical reasons (see section 4.8). The sequences of genomes deposited in the databases are therefore only very rarely complete. This is a quantitative problem that must be taken into account, aggravated by the fact that gaps are generally not random due to the very structure of genomes and cloning biases (see Chapter 5). Several percent of the human genome, for example, have never been assembled in the reference sequence.

5 TIGR: The Institute for Genomic Research, founded in 1992 by J. Craig Venter and based in Rockville, Maryland.

Baker's yeast, *Saccharomyces cerevisiae*, a unicellular fungus widely used as an experimental model in genetic laboratories (in addition to being the microorganism of alcoholic fermentation) was the third living organism ever sequenced (Goffeau *et al*. 1996). For the first time, it was a eukaryotic cell, like humans. Its genome consists of 16 chromosomes, totaling more than 13 million nucleotides located in the nucleus, plus mitochondrial DNA. This sequence was determined completely from map-ordered clones in a vast cooperative program coordinated by **André Goffeau** that involved many laboratories around the world, mainly in Europe. The sequence of the first chromosome was completed in 1992; it was a historic moment, because it offered us the first complete image of a eukaryotic chromosome. The sequences of three other chromosomes were to follow as early as 1994, when the distribution of human chromosomes among the global sequencing centers of the Human Genome Project had not yet been decided. The complete sequencing of the yeast genome was completed in 1996. With nearly 6,000 protein-coding genes predicted, of which only less than a thousand had already been discovered after several decades of genetic research, this sequence showed us the heuristic power of genomics at the same time as it made us discover an unexpected phenomenon, the large number of genes duplicated into families in a genome otherwise very compact. In the case of yeast, part of this phenomenon was later explained by the legacy of a complete duplication of the genome (itself resulting from hybridization between species), but the presence of paralogs is a general law of genomics. Genes have a strong tendency to duplicate and get lost over successive generations, and this is one of the major mechanisms of evolution.

Among the bacteria that followed, as early as 1997, were *Bacillus subtilis*, treated exactly like yeast, and *Escherichia coli*, two main experimental models, but also important human pathogens such as *Helicobacter pylori*, responsible for gastric ulcers and cancers, *Mycobacterium tuberculosis*, the tuberculosis agent or *Rickettsia prowazekii*, the typhus agent. The utilitarian nature of bacterial genomics appeared: we could go beyond model experimental systems to analyze the real living world, as it exists in nature. Certainly, the usefulness of model systems had not disappeared; quite the contrary, they became references against which natural organisms could be compared. But we could already see the possibility of sequencing natural populations.

The same trend developed for multicellular eukaryotes, only a little later, because their genomes are much larger. They generally measure hundreds or thousands of megabases (Mb). Two laboratory model organisms historically became the first fully sequenced multicellular eukaryotes, the nematode, *Caenorhabditis elegans* (97 Mb, published in 1998) (Science 1998) and the Brassicaceae, *Arabidopsis thaliana* (115 Mb, published in 2000) (Nature 2000). Both were sequenced after initial complete genome mapping, as for yeast and the Human Genome Project. On the other hand, the shotgun method was tested on the fly genome, *Drosophila melanogaster*, in 2000 (Adams *et al.* 2000). With a total size of 160 Mb containing repeated sequences, the result obtained was not complete (only the relaxed, "euchromatinian" part of the genome – about 70% of the total – was covered), but it paved the way for further developments. From this period onwards, the genomic sequences of various organisms have multiplied using the shotgun method more and more often, so that very many sequences are in reality just drafts with varying numbers of gaps. It is not necessarily an insurmountable problem if you remember it when using them. These incomplete sequences are useful for global comparative analyses, but should be considered with caution for more specific questions.

4.7. The human genome

While these pioneering works offered us the first images of the composition and organization of genomes and allowed us to refine experimental sequencing methods and computer methods of data analysis, the organization of human genome sequencing project remained difficult. The scope of the project suggested that the task should be distributed among several partners, but the lack of physical or genetic maps of sufficient resolution made this distribution risky. Human DNA libraries in conventional *E. coli* vectors would have required mapping a considerable number of recombinant clones to cover entire chromosomes. Cloning in yeast vectors (artificial chromosomes or YACs) could have solved the problem given the large size of the inserts, but their handling remained difficult for several reasons (instability, DNA purification, etc.). Two new features came to change the situation. On the one hand, a sufficiently detailed genetic map of the human genome was established in 1996 with molecular markers (Dib *et al.* 1996). On the other hand, a new type of bacterial cloning vector was developed by similarity to artificial yeast chromosomes. They were called BAC (for bacterial artificial chromosomes).

With sufficiently large inserts and greater ease of handling, they have made it possible to build collections of recombinant clones covering the entire human genome, which could be anchored on the genetic map. The sequence of the human genome was therefore assembled chromosome by chromosome as for yeast, *Caenorhabditis* and *Arabidopsis* in a coordinated international program that progressed steadily in the late 1990s.

But in the meantime, Venter and his colleagues had undertaken a competing project as part of a private initiative aimed at the total sequencing of the human genome using the shotgun method that had been tested and validated with Drosophila. The two programs did not use the same DNA and donors remained anonymous. The Human Genome Project had chosen to sequence a heterogeneous assortment of clones from different DNA libraries of a few individual donors. The genetic polymorphism (which was already estimated at the time to be in the order of one nucleotide over a thousand varying between two independent ***haplotypes****) was therefore ignored. The sequences obtained were only artificial chimeras between genomes of different individuals. Faced with the intensification of the competition and the importance of the issue, the two competing projects were officially declared complete in February 2001 and the results were published in parallel (Nature 2001; Venter *et al*. 2001). In reality, these were only draft sequences. The assembled fragments of sequences (contigs) were very short compared to the size of the chromosomes. There were more than 150,000 gaps left to fill, covering more than 10% of the euchromatinian sequence. It was only in 2004 that the international consortium of the Human Genome Project produced a second version of the human genome, in which only 341 gaps remained to be filled and which, with 2.85 billion nucleotides assembled, covered 99% of the euchromatinian part of the human genome (Nature 2004). With the heterochromatinian part, the sequence of which is still practically unknown today, the overall size of the human genome is 3.08 billion nucleotides. It was this latter version that served as a reference for further work.

The historical importance of this achievement should not lead us to forget what a reference sequence is: common data to which everyone refers for subsequent work, not a purpose, but a beginning. It is therefore not surprising that the sequencing of the human genome did not immediately have an impact in biomedical terms, as some people imagined. Genetics only begins when there is variation. Everything remained to be done to exploit this sequence. In the case of the human genome, we didn't even know how

many genes to expect. Before sequencing, fanciful estimates (100 to 150,000 genes) had been announced by commercial companies wishing to patent short sequences. Wagers had even been placed on this number. The first reasonable estimate was to come in 2000 from the comparison with a fish genome, *Tetraodon negroviridis* (Roest-Crollius *et al.* 2000). The human species needs to be self satisfied with a maximum of 28 to 34,000 genes encoding proteins; barely more than a modest nematode or even a cruciferous vegetable. The reality is closer to 20,000 genes encoding proteins. Together, they represent less than 2% of our genome, while more than 75% of the sequences are transcribed into RNA (Djebali *et al.* 2012).

4.8. New methods of genome sequencing and the current state of genomics

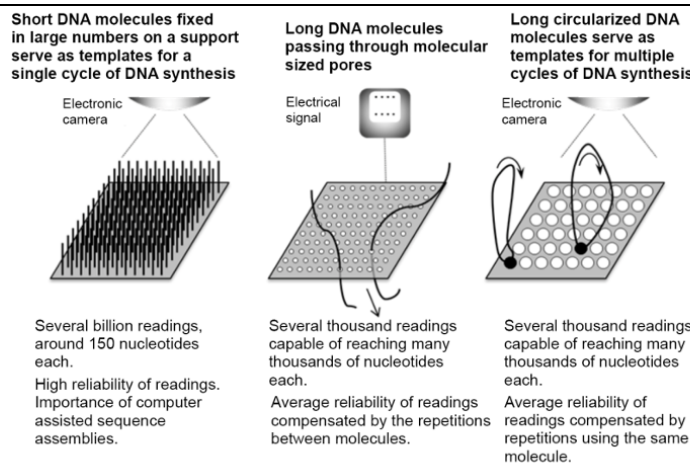
Until the mid-2000s, all DNA sequencing work was carried out using F. Sanger's enzymatic sequencing method, although, over the years, the addition of significant modifications increased its effectiveness. Among these, the replacement of radioactive labeling by fluorescent labeling paved the way for the development of the first automated processes. Several other improvements were also introduced, including fluorescent labeling of chain elongation terminators instead of primers, which made it possible to perform only one sequencing reaction per DNA molecule instead of four.

In the mid-2000s, DNA sequencing methods made a huge technological leap forward with the advent of the so-called NGS[6], which, through the application of new physical methods for image analysis of molecules attached to a surface, increased the number of simultaneously sequenced DNA molecules by several thousands, then millions of times, driving down costs in similar proportions.

Today, the latest equipment using these technologies is capable of sequencing several hundred million molecules in a single step, each one on more than a hundred nucleotides, producing a total of sequences several dozen times the size of the human genome (see Box 4.3). These technologies, based on very many short sequences that the computer must then assemble, are now complemented by completely different technologies capable of reading very long sequences on a smaller number of molecules. These long readings facilitate the assembly of the partially repeated

6 NGS: Next generation sequencing.

sequences that are very common in genomes. They also allow exploration of single DNA molecules in heterogeneous populations.



Today, the techniques of DNA sequencing known as "next generation sequencing" (**NGS**) are based on two complementary strategies:

– the parallelism of the reactions allowing to considerably increase the global producing capacities, but remaining with short individual readings (hundreds of nucleotides);

– the lengthening of individual readings by applications of novel techniques, but quantitatively reducing the parallelism of the reactions.

The first strategy, called solid-state sequencing, produces the bulk of the sequences in volume, given its efficiency and the reliability of the readings. For example, only one of the most commonly used devices in current usage is capable of sequencing several thousand human genomes per year or the equivalent in complex populations of organisms or ecosystems (***metagenome****). The main limitation of this strategy is the **computer assembly** of sequences, which becomes difficult with repeated sequences in genomes.

The second strategy is becoming increasingly important because it makes it possible to overcome the problem of assembling repeated sequences in genomes. In addition, it allows us to determine **individual variations** between DNA molecules. Its main limitation is the lower reliability of individual readings and the lower volume of sequences produced.

Box 4.3. *New DNA sequencing techniques*

By changing scale thanks to the panoply of these new devices, genomics has now changed in nature. Far from the recent years devoted to establishing reference sequences of well-chosen species including humans, the current power of low-cost DNA sequencing makes it possible to address the real genomes, that is those carried by the variety of individuals in populations. The time is probably not far away when all human beings will have their genome fully sequenced at birth (or probably even before). Some countries have already launched programs to sequence their entire populations. Discoveries can only accelerate, as we can already see, because genetics is the science that uses variation. The reference human genome is only a reference; the variations from this reference are the informative part. The same is true for natural populations as well as for farm animals or cultivated plants, for which several thousand sequences are already available. The quantitative depth of sequencing also allows direct access to complex natural populations without the need to first separate their constituents. This allows us to discover what we did not yet know, as exemplified by recent results on ***microbiotes**** or on the exploration of oceans (see Chapters 8 and 9). Conversely, we are also able to sequence the genome of a single cell. All fields of biology and its applications are being impacted by the developments in genomics.

4.9. Important ideas to remember

– The purification of enzymes capable of cutting DNA molecules at specific sites or of ligating them appropriately has opened up a new era of ***in vitro* genetics**. Genes (and soon genomes) have moved from conceptual objects to chemically manipulatable molecules.

– As soon as it was understood that nucleic acid and protein molecules were polymers of chemical elements in a limited number of species (4 nucleotides or 20 amino acids), it became clear that all the secrets of heredity lay in the **sequences** of these elements.

– DNA sequencing provides access to the most intimate details of the molecular structure of genes and entire **genomes**. But genetic analysis can only begin when there is a **variation**. More than from the reference sequences themselves, it is therefore from the comparison of sequences (mutants, genes and their products, individuals from the same natural population or even distinct species) that the maximum conclusions can be drawn.

– Current **DNA sequencing** technologies enable their application in the fields of individual medical diagnosis, genetic counseling, livestock and agriculture as well as in the exploration and monitoring of natural and/or microbial populations.

4.10. References

Adams, M.D. *et al*. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, 287, 2185–2195.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R., Young, I.G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290, 457–465.

Arber, W., Linn, S. (1969). DNA modification and restriction. *Annual Review of Biochemistry*, 38, 467–500.

Brownlee, G.G., Sanger, F., Barrell, B.G. (1967). Nucleotide sequence of 5S-ribosomal RNA from *Escherichia coli. Nature*, 215, 735–736.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, 380, 152–154.

Djebali, S. *et al*. (2012). Landscape of transcription in human cells. *Nature*, 489, 101–108.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., Van den Berghe, A., Volckaert, G., Ysebaert, M. (1975). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260, 500–507.

Fleischmann, R.D. *et al*. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, 269, 496–512.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G. (1996). Life with 6000 genes. *Science*, 274, 562–567.

Holley, R.W., Everett, G.A., Madison, J.T., Zamir, A. (1965). Nucleotide sequence in the yeast alanine transfer ribonucleic acid. *Journal of Biological Chemistry*, 240, 2122–2128.

Jackson, D.A., Symons, R.H., Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 69, 2904–2909.

Jou, W., Haegeman, G., Fiers, W. (1971). Studies on the bacteriophage MS2. Nucleotide fragments from the coat protein cistron. *FEBS Letters*, 13, 105–109.

Maxam, A.M., Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, 74, 560–564.

Nature (2000). The *Arabidopsis* genome initiative 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796–845.

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

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

Roest-Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier, P., Quétier, F., Saurin, W., Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genetics*, 25, 235–238.

Sanger, F. (2001). The early days of DNA sequences. *Nature Medicine*, 7, 267–268.

Sanger, F., Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94, 441–448.

Sanger, F., Nicklen, S., Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74, 5463–5467.

Science (1998). *C. elegans* sequencing consortium 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, 282, 2012–2018.

Smith, H.O., Nathans, D. (1973). A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. *Journal of Molecular Biology*, 81, 419–423.

Venter, J.C. *et al*. (2001). The sequence of the human genome. *Science*, 291, 1304–1351.

Wu, R., Kaiser, A.D. (1968). Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *Journal of Molecular Biology*, 35, 523–537.

5

Uniqueness and Polymorphism of Genomes

The word "genome" is not new. It was introduced in 1920 by **Hans Winkler**, a Professor of Botany at the University of Hamburg, to designate the genetic entirety of an organism reduced to its haploid equivalent (Winckler 1920). At that time, we were talking about genes, identified by their mutations. We were far from imagining the content of genomes as we know them today. But we already had the idea that this genetic entirety was finite – therefore accessible to exhaustive knowledge – and had a permanent character.

What is new, since a few decades ago, is that a genome is represented by a succession of nucleotides – the so-called "sequence" – that can be fully elucidated by DNA sequencing. These sequences are four-letter texts, with A, C, G, and T, each symbolizing one of the four nitrogenous bases of DNA: adenine, cytosine, guanine and thymine[1]. They are so long – the amount of information they carry is so great – that they can only be made accessible to the human mind by computers, while their deciphering – the sequencing of DNA – is done by specialized equipment (see Chapter 4).

1 In reality, the sequences recorded in computers use a broader alphabet (15 letters) in order to take into account the ambiguities of sequencing and the natural polymorphism of the sequences. In addition to the four primary letters A, C, G and T, the following are added: R for purine meaning A or G; Y for pyrimidine meaning C or T; S for strong meaning C or G (3 hydrogen bonds); W for weak meaning A or T (2 hydrogen bonds); K for keto meaning G or T (bearing a ketone radical); M for amino meaning A or C (carrying an amino radical); B for non-A meaning C, G or T; D for non-C meaning A, G or T; H for non-G meaning A, C or T; V for non-T meaning A, C or G and, finally, N for any one of the four primary letters.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

More than in the use of modern technologies – obviously essential – it is the **completeness** of the data that marks the limit between genomics and classical genetics. While the latter proceeds by reductionism by first identifying the elements associated with a phenomenon and then reconstructing their interactions by reasoning and experiments, genomics begins with a complete inventory of all existing elements, without an initial hypothesis, in order to provide an exhaustive basis for subsequent work to study them.

This is a major epistemological change that has troubled and still troubles many biologists, because it reverses fundamental reasoning and, in doing so, involves new heuristics. But the whole exceeds the sum of its parts and, in just two decades, genomics has profoundly influenced all of biology by revealing phenomena that remained inaccessible to ancient methods. And even if we still have a lot to learn today to correctly interpret the multiple biological dimensions included in DNA sequences, it has become totally impossible to ignore them. Even epigenetic phenomena are of genetic origin. So what do genomes tell us?

5.1. The immensity of nucleic acid sequences

Each genome has a finite size, measurable by the total number of nucleotides that make up its sequence. Kilobases (kb), megabases (Mb) or gigabases (Gb) are used to designate, respectively, one thousand, one million or one billion nucleotides in a sequence. These sizes vary enormously from one organism to another. The genome of the yeast *S. cerevisiae*, for example, is about 13 Mb and that of humans 3 Gb, which is 230 times more. However, it is not the largest genome. The genomes of small flowering plants such as *Paris japonica* (Melanthiaceae), or *Fritillaria sp*., the fritillaries (closely related Liliaceae family), are 50 or 30 times larger than the human genome. The smallest genomes are found in the world of viruses, but they are not autonomous organisms. Some bacteria or archaea capable of independent living have genomes in the order of 1 Mb. It seems to be a minimum. There are genomes that are smaller in size, but belong to parasites or organisms living in symbiosis. The largest genomes currently known are those of some amoebas, with more than 650 Gb, whereas they are unicellular eukaryotes in the same way as yeasts. Nearly six orders of magnitude separate the genomes of bacteria from those of amoebas, two microorganisms. In groups of multicellular organisms such as arthropods or angiosperms, for example,

genome size variation reaches three orders of magnitude. Overall, therefore, the size of genomes is not directly related to the apparent complexity of organisms, which will be explained in the following paragraphs. And it so happens that the genomes of related species, not always easily distinguishable from each other, differ in size by more than one order of magnitude.

In a mathematical sense, the size of a genome corresponds to a quantity of information[2]. As already seen in Chapter 2, a sequence of only 100 nucleotides, much smaller than most genes, offers 10^{60} combinations – a really "astronomical" number. In genetic terms, the possible variations are therefore infinite. Simply try to calculate the number of possible combinations of sequences of one million nucleotides, the size of the smallest known genomes of autonomous organisms. Your computer will refuse to do so. In other words, the living world we are currently observing, like all that may have existed during geological times and all that will exist in the future, can never be anything other than a tiny part of what is possible.

Obviously, not all combinations of nucleotide sequences have functional genetic significance. It is a fundamental question that remains open to try to estimate the number of elementary sequence patterns existing in the living world and, from there, to imagine how many other living worlds could exist if evolution had been different from what it has been during the history of the earth.

5.2. Components of genomes and their replication

Depending on the category of organisms, the genome corresponds to one or more DNA molecules[3] present in one or more cellular compartments. In bacteria and archaea, which have only one cell compartment, it is usually a single DNA molecule, usually circular, called the "chromosome", to which optional episomes can be added. However, some species have several

2 The fact that DNA molecules are double-stranded (except in particular cases of bacteriophages) does not double the volume of information they carry, because the two strands are strictly complementary to each other. The sequence of one of the two strands fully determines the sequence of the other and, as a result, only one strand is represented in databases.

3 In some viruses, the genome consists of one or more RNA molecules that multiply directly without ever passing through a DNA intermediary.

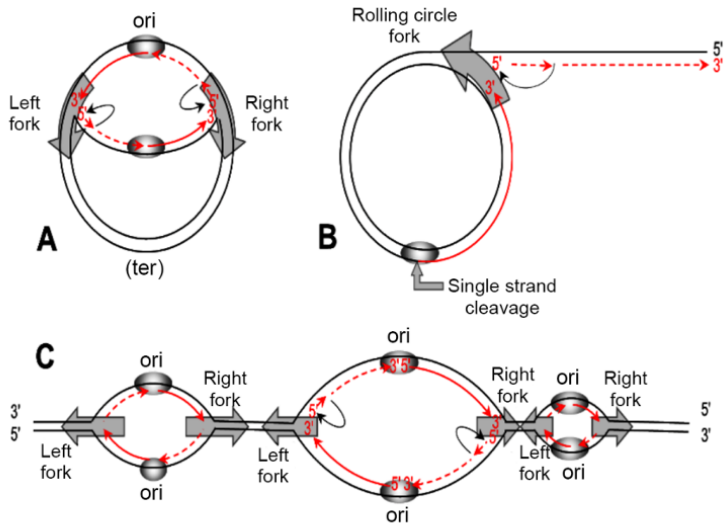
chromosomes. In eukaryotes, whose cells are formed by several compartments, the genome is always composed of several DNA molecules: one per nuclear chromosome, plus those present in mitochondria, chloroplasts (where they exist) and, possibly, nucleomorphs (old nuclei observed in certain lineages showing traces of recent ***endosymbioses****).

It should be noted that, to define its size, the genome is reduced to its haploid equivalent, that is the two copies of a pair of homologous chromosomes in a diploid are counted only once. The same applies to the multiple copies that are generally found in organelles (mitochondria and chloroplasts).

While these copies correspond to the same sequence in the case of homozygotes, the same cannot be said for heterozygotes, which is the most frequent situation in sexually reproducing organisms. In this case, it is useful to distinguish the two haplotypes, that is, the individual sequences of each of the two homologs, which is not always easy for technical reasons of DNA sequencing (only the most recent methods allow it). Similarly, some genomes, especially but not only in plants, are actually descendants of interspecific hybrids or total duplications, followed by gene loss and rearrangements (see Chapter 6). When we are able to identify the phenomenon, we speak of ***subgenomes**** to distinguish the parts of the genome from the different ancestral lines. This situation is seen fairly often in plants. For example, cotton, *Gossypium hirsutum*, is the result of spontaneous hybridization between an African species, *G. herbaceum*, and a Mexican species, *G. raimondi*, about one to two million years ago.

Genome replication depends on the structure of DNA molecules (see Box 5.1). In the case of large circular molecules such as bacterial chromosomes, replication generally starts from a single origin and progresses in a bidirectional divergent manner until the two forks meet, which can occur randomly, or more generally, at the level of a terminator. The entire chromosome corresponds to a single ***replicon****. In the case of the very long linear molecules of eukaryotic chromosomes, replication is initiated from several origins spaced from each other along the sequence and progresses in a bidirectional divergent manner until converging forks from two neighboring replicons meet. During the S phase of the cell cycle, some replication origins are activated early, others later. Not all origins are necessarily activated at each replication cycle. Activation is random, so the

same DNA segment can be replicated successively by forks originated from different origins along the same molecule.



DNA replication is carried out by a succession of **phosphoester** bonds formed between the hydroxyl radical carried by the 3' carbon of the last nucleotide of the growing chain and the phosphate group of a new nucleotide (see Chapter 2, Box 2.2). The elongation of a chain is therefore done exclusively by its 3' end. Since the two complementary chains of DNA are **antiparallel**, one of the two newly synthesized chains can progress continuously as the **replication fork** advances (continuous red strand), while the other must progress discontinuously (dotted red strand). The synthesis of this second chain is done by re-initiating new fragments (black arrows) from short primers made of RNA as the replication fork advances. The fragments are then ligated together.

According to the organisms, chromosome topologies lead to three distinct modes of replication. For circular DNA molecules (**A**, cases of bacterial and archaeal chromosomes and some eukaryotic episomes), two replication forks progress in opposite orientation from the same replication origin (ori), forming a **structure called theta**. The replication ends with the meeting of the two forks. For some circular DNA molecules (**B**, cases of some viruses or organelle genomes), replication is initiated from a 3' end (produced by strand cleavage or by recombination), which serves as a primer for the synthesis of the continuous strand, thus creating a replication fork whose progression can be infinite around the circle (**rolling circle** mode). For linear DNA molecules (**C**, cases of

eukaryotic chromosomes and some organelle genomes), replication is initiated from internal replication origins (ori) that can be multiple along the molecules. The replication ends with the meeting of converging forks originated from two distinct origins.

Box 5.1. *Modes of replication of DNA molecules*

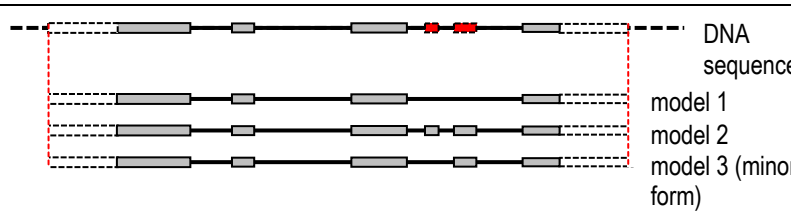
The genomes of organelles in eukaryotic cells show a great diversity of sizes and structures depending on the organisms (Smith and Keeling 2015). In the human species, as in many metazoans, the mitochondrial genome is made of small circular DNA molecules (16.5 kb in humans) in multiple copy number (normally identical) in each cell. In plants, the genomes of plastids and mitochondria are generally much larger and made of long linear DNA molecules bearing several repeated copies in tandem. Genomes therefore appear as circular, while the DNA molecules making them are linear. Sizes can vary considerably. The mitochondrial genome is 200 kb in turnips, 2,400 kb in melons and 11,000 kb in *Silene conica*. Plastid genomes, which average about 140 kb in flowering plants, can reach 1,000 kb in the green alga *Acetabularia acetabulum* or be reduced to 30 kb in Dinoflagellates. In yeasts, where the functions of the mitochondrial genome have been most extensively studied, there is a size variation of nearly one order of magnitude in mitochondrial genomes (from less than 20 kb to more than 100 kb) with no significant change in gene content. Genome size depends on the irregular presence of long mobile introns encoding proteins (see Chapter 2) and long intergenic sequences. They are also circular genomes (with a few exceptions) carried by long linear DNA molecules, suggesting a rolling-circle mode of replication. In some cases (green algae, protists, fungi or even animals), mitochondrial genomes are carried by linear DNA molecules with ***telomeres**** at the ends and can be reduced in size down to 13 kb (some green algae or yeast) or even 6 kb in a protozoan parasite of red blood cells (fatal for cattle herds) called *Babesia*. Finally, there are cases where mitochondrial or plastid genomes are fragmented between dozens of different small DNA molecules, each carrying sequences that will only become usable after assembly and massive editing of their RNA transcripts, as discussed in Chapter 9.

Unlike nuclear chromosomes, whose copy numbers remain fixed during the development of organisms (with rare exceptions in polyploid cells), copy numbers of organelle genomes vary between cells and, sometimes, environmental conditions. Thus, in *Arabidopsis thaliana*, we go from about

40 copies per chloroplast in a ***meristem**** cell to 600 when the leaf measures 0.5 mm, while when the leaf has reached maturity, the plastid DNA is in the form of very small molecules and some chloroplasts no longer even contain DNA at all. For mitochondria, it is difficult to talk about the number of DNA molecules per mitochondria, because they form a dynamic network, but the total number of copies of the mitochondrial genome per cell also varies in significant proportions around one hundred for yeasts, several hundred for mammals and several thousand for plants.

5.3. A little perspective on the content of genomes

Finally, a genome will therefore be represented by uninterrupted sequences of letters, each corresponding to a DNA molecule. Each sequence is therefore comparable to a single word, but this word has a considerable length. On the scale of this text, a small bacterial genome of one Mb is a word that would spread over about 250 pages! How can we read it? This is an operation called "genome annotation" (see Box 5.2), which consists of assigning a genetic meaning to each sequence segment. This is a difficult operation, because genetic elements are defined by their function, not by their structure. There are no material limits between them at the sequence level and they can even be entangled in each other (see Chapter 2). Moreover, in addition to genes, genomes contain many other types of genetic elements discovered later and which, in most cases, numerically exceed the genes (Lynch 2007).



An example of annotation of a fragment of a eukaryotic genome (region delimited by red dotted lines in the diagram), here the grapevine *Vitis vinifera* (Denoeud *et al.* 2008). The coding exons are in gray, the non-coding exons in white surrounded by dotted lines symbolizing uncertainty on their boundaries, the introns are symbolized by solid lines, the sequences outside the gene by dotted lines. Depending on the alternative splicing of introns, three gene models can be proposed. The sequencing of RNAs indicates their relative importance. The first model has 4 exons, the second 6 exons and the last 5 exons. The

annotation of the sequence (top line) must therefore take into account the fact that the same sequence segment (red boxes) can correspond simultaneously to introns and exons.

Identifying the genetic elements present in genomes from DNA sequences consists of defining the limits of the different types of functional elements that can be interpreted: protein coding genes, non-coding RNA genes, mobile elements and their traces, pseudogenes, cis-active elements of chromosomes (replication origins, centromeres, telomeres) and, for each of them, predicting if possible the exact functional nature of the element.

This operation is tricky for several reasons. First, the boundaries of genetically functional elements are not always unique. In general, several transcripts cover the same locus, with different origin and termination points. Second, many genetically functional elements are mosaics of exons and introns that can vary greatly in size. Very small exons (sometimes up to a single nucleotide) are very difficult to identify without the **RNA sequences** themselves. Large introns often host other genetic elements within them (traces of mobile elements, non-coding RNA genes). Finally, the assignment of a putative function to a new genetic element identified by its DNA sequence is based on **comparisons** with already known sequences that serve as references, giving a historical dimension to the process. Ideally, these references are related organisms on which experiments could be carried out in order to describe as extensively as possible the functions of the homologous genetic elements considered. In practice, this is not always possible and annotation by similarity of DNA sequences leads to a series of hypothetical, sometimes contradictory proposals.

With the considerable development of genome and RNA sequencing, the enrichment of **databases** and the arrival of new algorithmic principles better adapted to different types of organisms, genome annotation has made immense progress and many existing sequences are now being re-annotated before being used.

Box 5.2. *Genome annotation. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

To fully understand this situation, let us forget for a moment the history of our discoveries and imagine that a scientist from a distant galaxy arrives on Earth without preconceived ideas. He would find huge areas covered with trees or other types of vegetation[4] and a large number of bipeds very similar to each other and concentrated mainly in small areas of the planet with fewer

4 Hoping that he doesn't wait too long to visit!

trees, but which illuminate at night with light visible from space. And then even larger areas populated by a wide variety of aquatic organisms[5].

Of course, he would immediately sequence the DNA of these bipeds and that of a few trees and even perhaps a few marine animals, because in his galaxy, they have long understood the secrets of life. And what would he conclude? In all cases, genomes are mainly made up of nucleotide sequences repeated in a more or less imperfect way, among which are, in a dispersed way, short sequences that are not repeated very often or not at all, but which show a certain base-three periodicity (our visitor does not need to know the terrestrial genetic code to see this, the signal can be detected by mathematical processing of the sequence).

Logically, on this quantitative basis alone, he would deduce:

– that the more or less degenerate repeated sequences constitute the essential part of the genomes of the organisms on our planet and are therefore probably their basis, their essence and, perhaps, their origin;

– that these genomes must have captured, during their evolution, a small quantity of coding sequences (because of the periodicity) of common ancestral origin (because of phylogenetic traces that our visitor cannot ignore);

– that these terrestrial genomes have been able to preserve and multiply these coding sequences probably because they give them distinctive secondary characteristics (our visitor obviously immediately understood that trees, bipeds and fishes are not identical, but that they only differ in terms of details!).

None of this should surprise him, it's just like at home, because there are probably no other ways to build genomes. If he extends his stay with us, our visitor will have the opportunity to see that there are also simple, often microscopic organisms, with smaller genomes showing fewer repeated sequences (which increases the relative proportion of coding sequences, but without significantly changing their total number). He would logically conclude that it is the repeating sequences, more than the coding sequences, that best correspond to the complexity of the organisms.

5 Same remark.

We are far from our classic textbooks of biology! But we are getting closer to what genomics has revealed to us in less than two decades about the origin and evolution of genomes. Genes represent only a (small) part of the genomes. The rest consists of elements that are more or less abundant and more or less well evolutionarily conserved, long regarded as accessories ("junk DNA"), because their individual alteration by mutation remains, most often, without any immediate phenotypic effect, but which are in fact the main driving forces of genome evolution. We cannot understand genomes without taking into account the temporal dimension.

5.4. Traces of the past and driving forces for the future

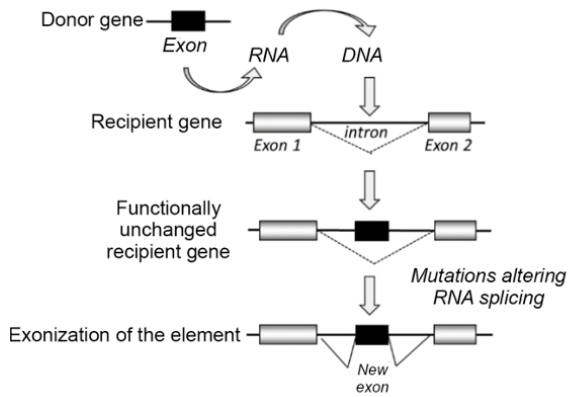
The most obvious traces of the past are pseudogenes. These are DNA sequences that resemble those of known active genes (from the same or another organism), but have a number of alterations that prevent their expression or render their product inactive. Pseudogenes are found in almost all organisms whose genomes have been studied. Our own genome has even more pseudogenes than active protein-coding genes. Where do they come from and, if they are useless, why don't they disappear?

It is now clear that pseudogenes have two origins. The first is simply the result of mutations. Often (but not only) after gene duplication, mutations can alter a gene without having a major deleterious effect, so that they are maintained in populations. They can even sometimes have a beneficial effect. In later generations, these pseudogenes will slowly fade away at the rate of neutral mutations (see below), their presence in the genome being due only to the difference in kinetics between their appearance and their disappearance. We are touching here on the fact that genomes are in reality only snapshots within continuous changes over successive generations, and are not optimized structures.

The second origin of pseudogenes is new copies of genes (or gene fragments) inserted elsewhere in the genome and showing traces of their passage through an RNA intermediate (e.g. loss of introns, polyadenylation)[6]. These are therefore more or less complete **retrocopies** of genes. It should be noted that this mechanism has a very strong power of evolutionary

6 This process of reintroducing sequence segments copied from RNA molecules into genomes is extremely important to remember in order to fully understand genetics, as this mechanism has the potential to create new heritable traits for the subsequent generations.

innovation, because, in addition to pseudogenes, retrocopies of RNA sequences can also merge with existing genes to form new genes (see Box 5.3). They are called ***retrogenes****. It is estimated that a very large number of our own genes have inherited such events, which have occurred in our ancestral lineage at various stages since the origin of vertebrates. The formation of DNA from RNA templates requires the activity of reverse transcriptases (see Chapter 2), which are generally produced by mobile elements, thus revealing their crucial importance in the evolution of genomes.



In exon shuffling, one or more RNA exons (solid box) produced from a donor gene can be retro-transcribed to DNA under the action of **reverse transcriptases** produced by retroviruses or class I transposable elements. Insertion of this DNA into the genome within an intron (solid line) of another gene will generally not produce any immediate phenotypic effect, as the splicing of the transcribed RNA (dotted lines) will remain unchanged. A minimal number of mutations are then sufficient to modify the splicing of the RNA and make the additional element to an exon of the recipient gene.

From the moment when it was understood that DNA can be synthesized from RNA molecules as templates (see ***retroviruses**** and ***retrotransposons****) and that RNA molecules themselves undergo or even catalyze splicing reactions, it became clear that RNA can be involved in the creation of new genes by reassembling pre-existing elements. This mechanism (originally hypothetical) came to be called exon-shuffling.

The first demonstration of its reality was provided in 1983 by a Drosophila retrogene, named *Jingwei*, which is expressed in the insect's testicles and brain and which comes from the insertion of accreted exons from an alcohol-

dehydrogenase gene into an intron of another gene, named *yande,* itself originated from the duplication of an ancestral gene called *yellow-emperor*. Since then, many cases of genes formed by exon shuffling have been discovered in different organisms. Almost a quarter of the protein-coding genes in our own genome bear the traces of this mechanism – a figure that is sufficient to demonstrate the importance of the action of transposable elements in the evolution of eukaryotic genomes.

Box 5.3. *Exon shuffling*

By their presence and that of their remains, mobile genetic elements explain a large part of the imperfectly repeated sequences observed by our extra-terrestrial friend. But it is their activity that is important, because they are both destructive and constructive. There are two main classes of mobile elements, defined by their mechanism of propagation (Wicker *et al*. 2007). All correspond to DNA sequences, generally smaller than 10 kb in size, that encode functions allowing them to move in a genome either after leaving their original locus (cut and paste mechanism) or, more frequently, by making a copy that will be inserted into another locus (copy and paste mechanism). The activity of mobile elements can create mutations by inactivating genes (white grapes, wrinkled peas, white eyes in Drosophila, golden cauliflower producing beta-carotene, some hemophilias in humans, etc.), but also generates a large part of the genetic innovation on which evolution is based.

Class I elements or **retrotransposons*** are short DNA sequences (a few kb) included in the continuity of the genome that carries them and therefore normally transcribed into RNA (specific regulations may exist). There are several families, some of which are similar to retroviruses (such as AIDS), but are not infectious. The originality is that their RNAs serve both as messenger RNA for the synthesis of a reverse transcriptase (and other enzymes) and as template for the synthesis of a complementary DNA strand by this reverse transcriptase. This strand will itself serve as a template for the synthesis of the second DNA strand and the double stranded DNA thus formed will then be integrated into the continuity of the cellular genome by the action of enzymes (integrases), themselves translated from the retrotransposon's messenger RNA. The result is the creation of a new copy of the mobile element at an ectopic site of the host genome, thus contributing to the formation of dispersed repeated sequences. It should be noted that this process can be repeated several times successively, in an exponential

manner, so that the propagation of mobile elements would be an explosive process if it were not regulated. Of course, retrotransposons are also subject to mutational degradation at the rate of neutral mutations, which explains the large number of altered copies found in all genomes in addition to active copies. For example, our own genome has several tens of thousands of copies of intact or altered Class I mobile elements belonging to several families. The very great diversity of abundance and nature of mobile elements in the different genomes reflects their irregular propagation, like short invasive waves interspersed with long periods of decline.

The mobile elements initially discovered in corn (see Chapter 1) belong to Class II, which also includes many other families of elements. These are ***transposons**** *stricto sensu*, whose propagation in genomes does not include RNA intermediates, unlike Class I elements. They are also short DNA sequences (a few kb) included in the continuity of genomes and transcribed into RNA molecules which, often after removal of introns by splicing, will become messenger RNAs for the synthesis of proteins that will ensure the propagation of these elements by their action at the DNA level. Depending on the family, this propagation is done according to "cut and paste" or "copy and paste" mechanisms. While the first type of mechanism has the potential to fragment chromosomes (which initially drew attention to corn), the second has a potentially invasive character as for Class I elements. The different families of Class II elements and their remnants therefore also contribute to the partially repeated sequences of genomes. Unlike Class I elements, which are restricted to eukaryotes, Class II elements are found in the entire living world.

The different families of mobile elements show unexpected distributions in the different categories of organisms. They do not respect (or very little) the ***phylogeny**** of the species. Similar elements can be found in plants and insects, for example, while related species can show different sets of elements. In some genomes, all active mobile elements may have completely disappeared, while their altered traces reflect recent activity in the ancestral lineage. It is interesting to note that mobile elements can play a role in speciation by creating a process called *hybrid* ***dysgenesis****, as observed in Drosophila. In other words, two individuals belonging to the same species, but one of whom carries active copies of certain mobile elements in its genome and the other not, can no longer produce fertile offspring together, because, after fertilization, the active copies of the first parent will massively mobilize the inactive copies of the other parent and destroy the genome. This

process therefore leads to the genetic isolation of individuals and subsequently to the formation of two species.

Despite their importance, mobile elements do not explain all the imperfect repeated sequences observed by our extra-terrestrial visitor, because there is an intrinsic tendency for chromosome segments to accidentally duplicate (sometimes even multiply a large number of times) at each cell generation (each DNA replication). Thus, in most genomes, even the smallest ones, there are more or less perfect duplications of segments varying in size. They are called **segmental duplications**. The mechanisms responsible for their formation are complex and it is not useful to detail them here, except to say that they are themselves genetically controlled, that is that there are genes involved in the synthesis of the molecular machineries involved and therefore mutations that can alter these processes. The largest segmental duplications can cover hundreds of thousands of nucleotides and contain several genes. They can be formed in tandem on the same chromosome or dispersed on different chromosomes. They are then transmitted to successive generations with slowly diverging sequences owing to the mutations that accumulate at each generation. In total, segmental duplications cover about 5% of our own genome, only counting those greater than a thousand nucleotides with more than 90% sequence identity. On average, each individual human genome shows about 150 segmental duplications, but the polymorphism within the population is such that several thousand sites of segmental duplications have been mapped across the human genome by comparing only a few hundred individuals. Segmental duplications are an important component of the structural mutations occurring in genomes that alter the number of copies of genes or fragments. They are responsible for the copy number variations (CNVs) observed at population levels. By sequencing the genomes of several hundred parent-child trios, it has recently been calculated that the losses (by deletion) or gain (by duplication) of DNA sequences at each generation cover several hundred kb of the human genome. This is not much in relation to genome size (some ten thousandths), but it is a lot over successive generations, because most of the time, fortunately, these alterations have no morbid effect – they only create diversity. Depending on the genes affected, however, some may have severe phenotypic effects (see Chapter 8).

Transposable elements are a source of innovation in genomes. For example, it is a particular recombination mechanism that, in jawed vertebrates including humans, creates the great diversity of T cell receptors

and immunoglobulins by combining fragments of different genes: V (for variables), D (diversity) and J (junction). This mechanism involves recombination enzymes encoded by the *RAG1* and *RAG2* genes, which are evolutionarily derived from genes encoding transposition enzymes in transposons (Zhang *et al*. 2019). This is an example of "molecular domestication", in which the original activity of transposition was lost and a new recombination activity acquired. It has been a crucial event in the evolution of the immune system of vertebrates allowing them to recognize the immense variety of foreign antigens, especially those of pathogenic organisms, parasites, bacteria and viruses.

5.5. Genes in genomes

These traces of the permanent mutational dynamics of genomes vary in their importance, depending on the organism. Generally abundant in eukaryotes, they are much rarer in bacterial genomes in which only genes remain, which, of course, cannot all disappear! And while genome sizes vary considerably among organisms as discussed in section 5.1, the number of genes, at least those whose end products are proteins, remain within a narrow range (about one order of magnitude)[7]. Thus, the human genome, with fewer than 23,000 protein-coding genes, is only four times that of yeast (~5,800 genes), and is half the size of that of a paramecium. The smallest genomes of autonomous bacterial species have about a thousand protein-coding genes. The total number of genes in a genome is therefore not directly related to the complexity of the organism (see Box 5.4 below).

So much attention is focused on genes because, of all the genome's elements, they have the most immediate impact on the organism's phenotype. When we look for the genetic determinants of phenotypic traits, as classical genetics does, we essentially come across genes. The other elements of genomes are very polymorphic and their mutational alteration is most often undetectable at the phenotypic level, which explains why they had long escaped genetics and why they may have been considered as accessories (junk DNA), because their importance is only revealed at the evolutionary scale. But while classical genetics played a crucial role in our understanding of genes, genomics proved to have some surprises still in store for us.

7 We are still far from knowing the numbers of genes whose end products are non-coding RNAs because the lifespan and functions of these molecules are so diverse (see Chapter 2).

Phyla and species	Genome (in Mb)	Number of genes	Coding part
Bacteria			
Mycoplasma genitalium	0.58	515	89%
Haemophilus influenzae	1.8	1,643	85%
Escherichia coli	5.1	4,380	78%
Archaeplastida			
Arabidopsis thaliana (arabidopsis)	120	27,000	33%
Triticum aestivum (wheat)	14,500	14,400	1.1%
Chondrus crispus (red algae)	105	9,600	14%
Chromalveolata			
Paramecium tetraurelia (paramecium)	72	39,600	82%
Plasmodium falciparum (malaria agent)	23	5,300	35%
Opisthokonts			
Saccharomyces cerevisiae (yeast)	13.1	5,860	70%
Caenorhabditis elegans (roundworm)	102	21,000	31%
Strongylocentrotus purpuratus (sea urchin)	814	23,300	43%
Homo sapiens (man)	3,000	22,300	1.1%
Mus musculus (mouse)	2,700	23,000	1.3%

Genome size varies by five orders of magnitude, from bacteria with a few mega-bases (Mb, million base pairs) to certain plants (Fritillaria) with 124,000 Mb. The number of protein-coding genes varies much less (less than two orders of magnitude).

The above examples give approximate values of genome size of some species, the number of protein coding sequences and their total proportion (in %) in the genome. For humans, mice or wheat, coding sequences represent just over one percent. It should also be noted that a single-celled organism such as a paramecium may have more genes than humans or mice. The names of the large phyla (bold) refer to Box I.1, “Simplified tree of life”, in the Introduction.

Box 5.4. *Composition of some typical genomes*

First, in all genomes, there seem to be too many genes – that is, sequences reveal the existence of multigenic families that traditional systematic mutagenesis approaches did not always suspect. Part of the

phenomenon corresponds to the fact that some genes are multiplied in order to ensure adequate levels of synthesis of their products. This is particularly the case for non-coding RNAs such as ribosomal RNAs, whose genes are generally amplified into a large number of tandem repeat units in a small number of chromosomal loci, or transfer RNAs, whose gene copies are, on the contrary, almost always dispersed in genomes. Examples of this type are rarer for protein-coding genes. In the latter case, on the contrary, the families of genes are derived from ancestral duplications that have produced copies whose sequences have then diverged over generations. This is the classic model of gene evolution by duplication-divergence originally proposed by **Susumu Ohno** in 1970 (Ohno 1970). But, with regard to one of the conceptual bases of genetics, derived from the Darwinian idea of natural selection, which prohibits two strictly iso-functional genes from cohabiting permanently in the same genome because one of them should be inactivated by random mutations without having phenotypic consequences, genomics shows us that redundancy is possible and even frequent. The reason is that each genome is only a snapshot between two temporal dynamics, that of the formation of gene copies (***paralogs****) and that of their loss. And as soon as the sequences diverge between the paralogs, which depends on a third dynamic, functional divergences can appear, which tend to stabilize multigene families. In a simple genome such as that of the yeast *S. cerevisiae*, more than 40% of the protein-coding genes are members of multigenic families. Some families may contain a very large number of gene copies corresponding to functional specializations, as is the case, for example, with olfactory receptor genes in mammals (several hundred genes). Note that this family has a significant number of pseudogenes in humans in direct relationship with our olfactory weakness compared to other mammals such as dogs.

Second, many genes are not essential. In cases where experimental inactivation of all genes has been systematically undertaken, such as in yeast[8], it has been surprising to discover that the proportion of genes essential to life remains low (about 20%). Most genes can therefore be lost one by one. In addition, while some gene deletions confer a measurable phenotypic change, in the majority of cases (more than 66% in yeast) they

8 These are systematic deletions of genes one by one, made in different strains, haploids or diploids, of different origin. Such collections are available for several yeast species. In other organisms, genes can be inactivated one by one by RNA interference without the need to remove them from chromosomes by deletion.

only produce a minor phenotypic effect or an effect perceptible only under very specific conditions. A few of them have no effect at all. This phenomenon may concern both single copy genes in genomes (***singletons****) and those in multigenic families, although it is somewhat more common in the latter, as expected. The fact that only a minority of genes are essential to life and that many of them have only a secondary role was also a great surprise of genomics. In retrospect, it explains that many functional genes have escaped the conventional systematic mutational screens used in experimental model systems designed to identify all the genes involved in a defined function. Of course, genetics had long predicted the existence of neutral mutations, but it had clearly underestimated their numerical significance. This significance has an immediate impact on the genetic structures of populations, because, in reality, genes are much more easily lost than previously thought. But before examining these consequences, a review of genetic determinism is useful.

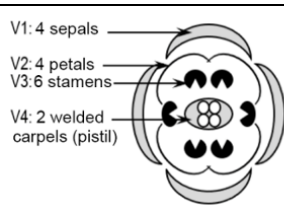
5.6. Genes and genetic determinism

The relationship between gene and phenotype is rarely as direct as in the case of Mendel's experiments. The complexity of this relationship has dominated genetics since its origins and has been reflected in the emergence of sometimes abstruse terms such as "epistasis", "penetrance", "expressivity" or "heritability", which deserve some attention, as they have become more important since we are able to sequence individual genomes. Outside the experimental framework on the genetically pure lines of model organisms, the problem of the relationship between gene and phenotype is complicated by the fact that the different individuals in natural populations are not genetically identical to each other, even if they are similar. The relationship between the gene under consideration and the distribution of the corresponding phenotypic trait thus becomes subject to the influence of genome polymorphism, that is the multiplicity of alleles of all other genes (sometimes called "genetic background"). This interference between different genes, called ***epistasis****, is extremely variable from one trait to another. At one of the extremes are the so-called ***monogenic**** traits, as were the smooth or wrinkled characters of the seeds of Mendel's peas. At the other extreme are the so-called ***polygenic**** traits, such as the size of individuals or skin color in the human species. In between, there is a continuous gradation of traits that depend simultaneously on major and minor genes. This is the most general case.

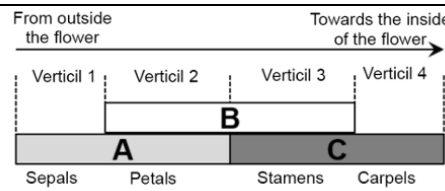
One of the consequences of the genetic polymorphism in populations is that the same trait may appear to be monogenic or polygenic depending on an individual's sexual partner. For example, resistance to cycloheximide in *S. cerevisiae* yeast is monogenic in about 70% of natural isolates (it depends only on mutations in the *PDR1* gene), but it becomes digenic (*PDR1* and *PDR5*) or even plurigenic in the remaining isolates. If we apply the above concepts to each of the genes involved in a polygenic trait, we see that we will often find ourselves in situations that are difficult to analyze for populations that do not allow experimental crosses, as is the case in humans. This is one of the major difficulties of human genetics (see Chapter 8), which requires large cohorts and high-density genotyping (or, to be even more precise, the integral sequencing of the genomes of all individuals in the cohorts) in order to identify the genes responsible for diseases.

This is obviously different in species where the isolation of mutants and experimental crosses reveal the full power of traditional genetics to identify the determinants of complex phenotypes. Take the example of floral development (see Box 5.5). *A priori*, a developmental process must involve the coordination of several genes. However, in this case, the isolation of mutants in *Arabidopsis thaliana* provided a surprisingly simple answer, with only four major genes involved, A, B, C and E. How can it be explained that a complex developmental process can be based on such a small number of genes? The answer here is that these genes determine the synthesis of transcription factors[9], that is proteins that activate or inhibit the transcription of other (many) genes that are directly involved in the development of particular flower parts. The four transcription factors are associated in tetramers that bind to DNA, so that the expression of A + E leads to sepals, A + B + E to petals, B + C + E to stamens, and C + E to carpels. These genes also inhibit the genes involved in the development of the leaf in floral parts, confirming the intuition of **Johan von Goethe**, who, as early as 1790 in his *Metamorphosis of Plants*, had understood that floral organs were modified leaves. Is it always so simple? Often yes. For example, Drosophila mutants are known to transform antennas into legs or to change the number of thoracic segments by altering only one gene at a time.

9 These are families of proteins preserved throughout the living world, of which there are members in humans, for example. The products of the B and C genes are MADS box proteins (acronym for MCM1 from yeast (*S. cerevisiae*), AGAMOUS from *A. thaliana*, DEFICIENS from snapdragon (*Antirrhinum majus*), and SRF from humans (*Homo sapiens*)).



Floral diagram of *Arabidopsis thaliana*



The interaction of three genes, A, B and C determines the differentiation of the four verticils

What determines the identity and position of each organ of a flower? It is a logical task! Multicellular organisms develop by building different tissues and organs. Everything looks as if an immutable construction plan were drawn up by a project manager. Thus, a flower develops (from the outside in) four different organs organized into four verticils: sepals (se), petals (pe), stamens (st) and carpels (ca). The normal formula is therefore: se, pe, st, ca.

Towards the end of the 20th Century, the use of mutants of floral development in the model plant *Arabidopsis thaliana* provided a surprisingly simple answer to the question of the identity of a flower's organs (Bowman *et al.* 1991). Three mutations eliminate whole organs, without altering the verticil's organization, and the eliminated organs are replaced by others (we speak of **homeotic mutations**). The first mutant, which has lost function A, has neither sepals nor petals and the four verticils are (from the outside in) ca, st, st, ca. The second mutant, which has lost function B, has neither petals nor stamens and the four verticils are se, se, ca, ca. The third mutant which has lost function C, has neither stamens nor carpels and the four verticils are se, pe, pe, se.

A very simple model (known as the ABC model) is formally deduced from this: to produce sepals, function A is necessary; to produce petals, both functions A and B are necessary; to produce stamens, both functions B and C are necessary; to produce carpels, function C is necessary. Phenotypes also imply that functions A and C are mutually exclusive, one taking the place of the other when the latter is absent.

This model is robust, because the phenotypes that can be predicted by combining the mutations verify it: the double mutant lacking A and B has flowers composed only of carpels (ca, ca, ca, ca); the double mutant lacking B and C has flowers composed only of sepals (se, se, se, se); the double mutant lacking A and C, as well as the triple mutant lacking A, B and C, have no floral organs, which are replaced by leaves. As further research was carried out, new mutants revealed the existence of an E function, so that A + E = sepals, A + B + E = petals, B + C + E = stamens, and C + E = carpels.

Box 5.5. *Identity of floral parts*

But there are also many other cases where the number of genes identified by mutations is insufficient to account for the entire genetic determinant of the observed phenotypic traits. We're talking about a lack of **heritability**. There are many reasons for this. Rare alleles are not included in the analysis for statistical reasons (see also section 5.8). But there are also two more fundamental reasons. Some alleles have incomplete ***penetrance****, that is, not all carriers of this allele necessarily develop the associated phenotypic trait. In pure genetic lines, where all individuals have the same "genetic background", the phenomenon is often explained by threshold effects that apply to inherently fluctuating molecular mechanisms. In natural populations, there is also the possibility of unknown interactions within "heterogenous genetic backgrounds". These are actually unidentified cases of epistasis. Similarly, the ***expressivity**** of the alleles must be taken into account, that is, in a genetically homogeneous population, the intensity of a quantitative trait may vary between carriers of the same allele because of the intrinsic noise of the molecular mechanisms involved in its phenotypic expression. Again, the possibility of unknown genetic interactions may contribute to the appearance of the phenomenon in the case of natural populations. In any case, the search for missing genetic determinants for interesting phenotypic traits remains one of the priorities for the various applications of genomics.

5.7. Natural populations: pan-, core-genomes and SNP

While genes can be gained or lost without major consequences, different individuals of the same species may not have exactly the same set of genes in their genome. This is regularly observed for the different species in which a sufficient number of individuals have now been sequenced, including humans. We will therefore distinguish a ***core-genome****, made of the sum of all the genes common to all the individuals of the species, and a ***pan-genome****, made of the sum of all the genes present in the species, but not carried by all the individuals of this species (Vernikos *et al.* 2016). Each individual carries in its genome an intermediate number of genes between that of the core-genome and that of the pan-genome. It should be noted that this definition is operational. It depends on the number of individuals that have been sequenced. But trends are emerging that distinguish "open" and "closed" species.

In the *E. coli* bacterium, an "open" species where the phenomenon is particularly spectacular, each isolate has about 4,000 protein-coding genes,

but when comparing a few hundred isolates, less than a thousand of these genes form the core-genome, while the pan-genome exceeds 16,000 genes! Much of this phenomenon is due to the existence of variable chromosome regions that correspond to episomes, bacteriophages integrated into the genome (prophages) or other elements horizontally transmitted between individuals. On the contrary, other bacterial species known as "closed", such as *Bacillus anthracis*, show little difference between core- and pan-genomes as an increasing number of isolates is studied.

The difference between the core- and pan-genomes is not specific to bacteria. In the yeast *S. cerevisiae*, where more than a thousand natural isolates have recently been fully sequenced, the core genome has about 4,800 genes for a pan-genome of ~7,800 genes (Peter *et al.* 2018). Since each yeast genome carries ~5,800 genes, it has therefore lost or gained one or two genes on average compared to the other members of the species. Similar figures look valid for the human population, although it is unlikely that the pan-genome will reach very high values for the entire world population due to stratification of the population (the same genes are affected for different individuals). In cultivated plants such as rice, corn, soybeans, cabbage and tomatoes, for which sequences of various varieties are available, and provided that gene relocations caused by transposons are ignored, the proportions of the core-genome to the pan-genome are fairly constant, in the order of 80%, which is high but still leaves several thousand genes specific to some varieties and absent in others. These are often genes involved in adaptive traits to environmental conditions, whether physical (climate, photoperiod, etc.) or biological (diseases, parasites, etc.) and other traits selected by humans. Note that the core genomes should not be confused with the minimum genomes, which are being constructed for different organisms (see Chapter 9), as there remains a degree of internal redundancy in the core genomes.

Another important factor in population genetics, more studied so far, concerns the polymorphism of sequences between alleles of the core genome. By aligning two or more ***orthologous**** sequences (see Box 5.6), one can be observe single-point differences called a ***SNP**** (single nucleotide polymorphism), or nucleotide differences at a given site, and *indels* (Insertion-Deletion) defined by the presence/absence of one or a few nucleotides at a given site. Indels are less abundant than SNPs. Since these differences are due to mutations that have accumulated over generations, counting them gives us a measure of the evolutionary distances between the

alleles examined. From these measurements, phylogenetic trees can be reconstructed for the sequences concerned. By combining data on the entire core genome, we have an estimate of the average genetic distance between individuals. For example, each individual human genome differs from the arbitrary reference by 4 to 5 million sites, that is, by about 0.15% of the total genome sequence. Much higher values of sequence polymorphism (several %) can be found in many animal, plant or microbial populations. Individuals whose sequences differ by more than 10% are still able to form viable and often fertile hybrids. In *Arabidopsis thaliana*, on a sample of 20 lines with a determined geographical origin representative of the species, the number of SNPs relative to the reference sequence is between 453 and 790 thousand, and the number of indels about 10 times less. In other words, on average, one point mutation is found for every 180 base pairs between two lines, a figure comparable to the difference between the genomes of human and chimpanzee. The corn genome behaves somewhat like the human genome: the sequencing of about 30 lines representative of the crop types reveals a total of 57 million SNPs, when compared to the reference. But in this case, it is the variations in the position of genes in the genome and the cases of presence/absence that are most striking, to such an extent that between two lines, nearly half of their genome may not be strictly collinear. The genetic polymorphism in natural populations therefore depends on two independent parameters: the presence/absence of genes and the difference of sequences between the genes present. The importance of this polymorphism will be discussed in Chapter 7.

```
..HKVGPNLHGIFGRHSGQAEGYSYTDANIKK..
    *       *    **    *       *  **
..HKTGPNLHGLFGRKTGQAPGYSYTAANKNK..
```

Alignment of fragments of sequences of cytochrome c in yeast (top) and humans (bottom). The actual proteins are about four time longer than the segments illustrated. Amino acids are represented by the one-letter code (inset below). Here, the two fragments of sequences are identical for 73% of positions. They differ by 8 amino acids (stars on the gray background).

A : L-Alanine	**Q** : L-Glutamine	**L** : L-Leucine	**S** : L-Serine
R : L-Arginine	**E** : L- Glutamic acid	**K** : L-Lysine	**T** : L-Threonine
N : L-Asparagine	**G** : Glycine	**M** : L-Methionine	**W** : L-Tryptophane
D : L-Aspartic acid	**H** : L-Histidine	**F** : L-Phenylalanine	**Y** : L-Tyrosine
C : L-Cysteine	**I** : L-Isoleucine	**P** : L-Proline	**V** : L-Valine

```
CAT AAG GTT GGT CCA AAC TTG CAT GGT ATC TTT GGC AGA CAC TCT
CAC AAG ACT GGG CCA AAT CTC CAT GGT CTC TTT GGG CGG AAG ACA

GGT CAA GCT GAA GGG TAT TCG TAC ACA GAT GCC AAT ATC AAG AAA
GGT CAG GCC CCT GGA TAC TCT TAC ACA GCC GCC AAT AAG AAC AAA
```

Alignment of the fragments of the corresponding coding sequences (yeast at the top, humans at the bottom). The codons have been artificially separated by intervals to facilitate reading. The two sequences are identical at 68% of the nucleotide positions. However the differences affect 19 of the 30 codons for only 8 amino acids replaced.

Alignments between **orthologous sequences** are the basis for genome **annotation** and for studies of **molecular phylogenies** between genes or organisms. Alignments can be made between protein sequences (top) or nucleic acid sequences (bottom). They can concern two sequences (as shown here) or multiple sequences. These alignments involve specialized **algorithms** that calculate the maximum number of similarities based on parameters that can be modulated according to the data examined.

Alignments are all the more reliable when the similarities between sequences are high. Alignments between nucleic acid sequences quickly become difficult as the divergence between sequences increases, because there are only four possibilities at each position. Alignments between protein sequences remain more robust as the divergence between sequences increases, because there are 20 possibilities at each position. In practice, these alignments are therefore preferred to compare organisms that are evolutionarily distant, often leaving aside all genes that do not code for proteins.

Box 5.6. *Alignment of orthologous sequences*

5.8. Population genomics

A lot of the sequence differences observed between individuals of the same species are of no functional importance, they represent a neutral polymorphism, but some are responsible for phenotypic distinctions between individuals, whether they are normal or pathological features. Identifying them is one of the major challenges of modern genetics, because, in the absence of pure genetic lines and experimental crosses as can be done with laboratory models, this is a very difficult problem. Examination of family pedigrees (in humans) provides information on the mode of inheritance of the trait under consideration (dominance, recessivity, autosomal, sex-linked or uniparental), but remains far from the precision necessary to reach the level of DNA sequences. And of course, there are no pedigrees for other natural populations. It is sometimes possible to refine the study by taking into account a possible genetic linkage with loci already described along the chromosomes. We thus obtain the location of the genetic determinant of the trait of interest on a factorial map graduated in centimorgans which, depending on the case, may be more or less well anchored on DNA sequences, but whose resolution is still insufficient to reach the gene level. This defines a more or less extensive chromosomal region that may facilitate the cloning of a gene if it is sufficiently reduced in size. The search can also be continued through the so-called “candidate gene” strategy, that is seeking

to predict, on the basis of DNA sequence annotations, which of the genes in the identified region is most likely to be involved in the trait under study. In addition to the cumbersome implementation of this strategy, it has proved to be generally poorly effective.

The alternative strategy, most widely used so far because it is well adapted to the rapid advances of DNA sequencing, consists of seeking an association between a phenotypic trait and the sequence polymorphisms observed between individuals. This strategy, called a genome-wide association study or ***GWAS****, has made it possible in a few years to identify the genetic determinants of a large number of interesting traits, mainly those of multifactorial inheritance (Visscher *et al.* 2012). In humans, a significant number of loci involved in various genetic diseases (Crohn's disease, prostate cancer, breast cancer, type 2 diabetes, etc.) could thus be identified. A catalog of approximately 25,000 SNP-trait associations is available (MacArthur *et al*. 2017). The success of such a statistical strategy obviously depends on the size of the available sample, that is, the number of individuals whose genomes are sequenced, but it also depends on other factors that cannot be controlled, such as the frequency of alleles in the studied population, their degree of phenotypic importance and, finally, the stratification of populations (the nature of alleles and their frequencies are not the same in each subpopulations and in the overall population because of the kinship between individuals). In practice, while GWAS has been effective to identify frequent alleles with high phenotypic impact, it is much less effective for alleles that are rare (present in less than 5% of individuals) and have low penetrance or expressivity. In most cases, therefore, there remains a lack of heritability for the traits studied. Another limitation is the existence of structural mutations producing gene copy number variations that are not taken into account in the GWAS method.

5.9. The genetics of genomes

The strategy of crossing pure lineages, which has enabled the early development of genetics, takes on a new dimension thanks to genomics, because the genomes of the parents and their descendants can now be compared at the nucleotide level. It now becomes possible to re-examine, with maximum precision, the phenomena of mutations, recombinations or epigenetic regulations and to identify the genes involved in these mechanisms. In the yeast *S. cerevisiae,* for example, it has been possible to construct hybrids between strains, whose genomes differ by several tens of

thousands of nucleotides distributed along all the chromosomes, from which all events occurring in mitoses or meioses can be studied with a previously unknown precision by sequencing the genomes of the mitotic or meiotic progenies. It was thus possible to map events of loss of heterozygosity (loss of a chromosome segment of one parent replaced by a copy of the other) or genetic recombination (exchange between chromosomes of both parents) in normal or mutated contexts for the molecular actors involved in these exchanges. While this strategy essentially applies to important model organisms such as yeasts, fruit flies, *Caenorhabditis elegans,* or *Arabidopis thaliana,* not to mention certain bacteria, it can also be used in the parent-child trios in humans, as already mentioned.

When the parental strains are chosen to show distinct phenotypic traits, the genetic determinants of these traits can also be directly identified by the relationship between the phenotypes and the genome sequences of the descendants. In the (rare) cases of traits with monogenic genetic determinism, the locus involved can be identified very quickly where, previously, laborious crosses were required to obtain a less precise result. In the more general case of traits with complex genetic determinism, the genetics of genomes makes it possible, if a sufficient number of meioses are studied, to identify quantitative traits loci or ***QTL**** as loci involved in the genetic determinism of the trait studied. A QTL is identified as the locus where a given allele is statistically significantly associated with the presence of the phenotypic trait examined among the meiotic progeny of the cross. Each QTL can then be associated with one or more genes or genetic elements interpreted from the DNA sequence depending on the accuracy of its mapping. This strategy, widely used to dissect complex traits in model organisms such as mice or *A. thaliana,* is particularly useful in the case of animal or plant species of agricultural importance.

5.10. Important ideas to remember

– Genomes are large, generally too large for the number of genes they carry, because they contain additional genetic elements related to their evolutionary history.

– These include active **mobile elements** and their remnants, which play an essential role in the evolutionary dynamics of genomes and contribute to a significant part of repeated sequences.

– By identifying genes through the phenotypic consequences of their mutations, classical genetics has focused unknowingly on genes of major importance. **Genomics** shows us that in reality only a minority of existing genes fall into this category, the inactivation by mutation of the others having little or no effect.

– **Genetic polymorphism** in natural populations includes the differences in allele sequences between individuals, but also the presence or absence of certain genes in each individual genome.

– Some phenotypic traits are determined by a small number of genes (sometimes only one), others by the **interaction** of multiple genes of different relative importance whose nature and/or even presence of alleles vary within natural populations.

– By comparing the **sequences** of individual genomes within **cohorts** chosen for a given trait, we can identify the **genetic determinants** of that trait more or less exhaustively depending on the populations and cohort sizes. Alternatively, the existence of a **genetic linkage** with a previously known locus or the sequencing of a large number of offspring from experimental crosses allows the identification of a chromosomal region to be explored.

5.11. References

Bowman, J.L., Smyth, D.R., Meyerowitz, E.M. (1991). Genetic interactions among floral homeotic genes of Arabidopsis. *Development*, 112, 1–20.

Denoeud, F., Aury, J.-M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M., Morgante, M., Valle, G., Wincker, P., Scarpelli, C., Jaillon, O., Artiguenave, F. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biology*, 9(12), R175.

Lynch, M. (2007). *The Origins of Genome Architecture*. Sinauer Associates, Sunderland.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z.M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., Parkinson, H. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, 45, D896–D901.

Ohno, S. (1970). *Evolution by gene duplication*. Springer Verlag, Berlin.

Peter, J., De Chiara, M., Friedrich, A., Yue, J.X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., Liti, G., Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature*, 226, 339–344.

Smith, D.R., Keeling, P.J. (2015). Mitochondrial and plastid genome architecture: Reoccurring themes, but significant differences at the extremes. *PNAS*, 112, 10177–10184.

Vernikos, G., Medini, D., Riley, D.R., Tettelin, H. (2016). Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23, 148–154.

Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, 90, 7–24.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8, 973–982.

Winckler, H. (1920). *Vererbung und Ursache der Parthenogenese im Pflanzen und Tierrich*. Fischer, Iéna.

Zhang, Y., Cheng, T.C., Huang, G., Lu, Q., Surleac, M.D., Mandell, J.D., Pontarotti, P., Petrescu, A.J., Xu, A., Xiong, Y., Schatz, D.G. (2019). Transposon molecular domestication and the evolution of the RAG recombinase. *Nature*, 469, 79–102.

6

Natural Dynamics and Directed Modifications of Genomes

6.1. The dynamics of genomes

As we have already said about genes, and for the same reasons, genomes are responsible for both the intergenerational permanence of genetic information and its opposite, the formation of genetic diversity. Over successive generations, DNA sequences diverge and reorganize within genomes. These are two distinct processes in terms of the molecular mechanisms involved and the scale of changes produced (see Chapter 1, Box 1.3).

Sequence divergence between genomes results from the accumulation of point mutations during successive replications or from poorly repaired DNA alterations (see Chapter 2). In a given lineage of organisms, this accumulation occurs at a regular average rate, so that the comparison of sequences originating from a same common ancestor (orthologous sequences) makes it possible to calculate the age of their separation and thus to reconstitute phylogenies based on "molecular clocks".

The reorganization of sequences within genomes results from the accumulation of structural mutations acting on chromosome segments of variable sizes by moving them relative to each other (translocations, inversions) or by modifying their number per loss (deletions) or gain (duplications, ***amplifications****). This accumulation occurs at an irregular rate, variable between evolutionary lineages, with the result that the order of loci along the chromosomes will not be identical for the different

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

descendants of a same common ancestor. We're talking about loss of ***synteny****. This loss is irregular, so that some regions, which maintain an ancestral organization, enable studies of species evolution and facilitate physical genome mapping.

These phenomena had long been known by classical genetics. The novelty of genomics lies in the accuracy of their description and their precise quantification. In particular, it has become clear that structural changes are much more frequent than previously imagined and that they play a crucial role in the evolution of genomes through gene loss and gain. Structural mutations also generate new DNA sequences at junctions between rearranged segments. These new sequences may result from clean cuts in the pre-existing sequences, followed by their ligations. But, in general, they are less accurate because of the molecular mechanisms of re-joining DNA fragments. New junctions can occur between genes as well as within genes, or can involve transposable elements. The junctions within genes can lead to their inactivation or to the fusion of different genes. In the latter case, new products may be evolutionary innovations or, on the contrary, the cause of more or less severe anomalies. For example, there are cases of cancer due to accidental fusions between genes whose products are, on the one hand, kinases and, on the other hand, transcription regulators. Chapter 8 will present spectacular examples of structural rearrangements of chromosomes linked to cancers.

Beyond the sequence divergences and structural rearrangements mentioned above, genomes also evolve by polyploidization, that is, increasing the number of copies of homologous chromosomes in eukaryotic nuclei. This phenomenon is common in the plant world. Remember that the first mutants described by de Vries were nothing other than genome duplications. These duplications seem rarer in the animal world. However, in vertebrates, teleostean fishes have inherited a complete duplication of their genome compared to tetrapods (Jaillon *et al.* 2004). This ancestral event seems to have coincided in time with the phenomenal species diversification of these fishes. Similarly, a genome triplication coincides with the separation of monocotyledons from dicotyledons among the angiosperms (Jaillon *et al.* 2007). The evolutionary impact of these rare events can therefore be considerable.

One of the reasons for this importance is to be found in the massive loss of gene copies in the lineages which follow such events, as shown by

detailed studies carried out in yeasts or parameciums. Immediately after a whole genome duplication event, all genes are in double number of copies and the random losses of supernumerary copies should therefore have no or limited effect on the organism carrying them. However, such losses are likely to increase the genetic diversity between descendants of the same ancestor and accelerate speciation, since the remaining copies of the same gene may no longer be located on the same chromosome. In this case, chromosome reassortment during meiosis will result in the complete absence of the gene in some gametes and zygotes and thus make sexual reproduction between descendants of the same genome duplication event impossible (Scannell *et al.* 2006).

The origin of whole genome duplications remains imperfectly understood. In the plant world, the doubling (or more) of the number of chromosomes is found at the origin of several of our cultivated species (durum wheat, bread wheat, rapeseed, cotton, brown mustard, tobacco, etc.), which are in these cases interspecific hybrids that have become sexually fertile again. It is likely that they are plants coming from tetraploid embryos derived either from the fertilization between (frequent) diploid gametes derived from abnormal meiosis, or from the ***endoreplication**** of initially diploid embryos. The restoration of meiotic fertility by endoreplication has been demonstrated experimentally using artificial yeast hybrids (Wolfe 2015). There is a strong positive selection for whole genome duplication, prior to or after interspecific hybridization, a necessary condition for normal meiosis and therefore for the sexual fertility of these hybrids. The use of colchicine to double the number of chromosomes has been a common practice for plant variety breeders performing interspecies hybridizations for about 80 years. There are also many plant species that result from duplications of the same perfectly fertile genome, such as bananas (3 copies), groundnuts, potatoes, alfalfa (4), sweet potatoes (6) or sugar cane (more than 10). They are "autopolyploids". Autopolyploidy of plants often leads to a greater capacity for colonization under environmental changes. Examples can be seen after the retreat of glaciers in mountainous areas. It can also ensure their success in geographical expansion at different latitudes, as observed in the case of 180 wild potato species in South America depending on whether they have 2, 3, 4, 5, or 6 copies of the genome (Hijmans *et al.* 2007). The presence of several alleles per locus (genetic redundancy promoting heterozygosity, including in gametes) is probably responsible for these benefits. Initial hybrids or polyploids can also reproduce clonally if their meiotic fertility is reduced, which is consistent

with a large number of examples of genome duplications in plants, fungi and some protozoa compared to their greater rarity in metazoans with mandatory sexual reproduction.

6.2. Hereditary acquisitions

Another key lesson from genomics is the evolutionary importance of horizontal genetic transmissions between different organisms (Soucy *et al.* 2015). In contrast to the regularity of vertical hereditary transmission from parents to their descendants – clonal or sexual – on which genetics is based, the acquisition of alien DNA sequences in a lineage is always accidental in nature. However, its consequences can be considerable. Conceptually, this phenomenon is very important, because if it shows that acquisitions can accidentally become hereditary, this should not be confused with what has been called "the inheritance of acquired characters", which assumed a directed action of the environment on the genome, totally excluded by the facts. With genome sequencing, foreign sequences are found in almost all species studied, indicating that horizontal transfer of genetic material is a general phenomenon throughout the living world. It concerns both bacteria (where it is very common) and eukaryotes. It results from various mechanisms ranging from the direct transformation of cells by DNA to cellular endosymbioses, not to mention the role of viruses and mobile genetic elements.

6.2.1. *Transformation by DNA and horizontal gene transfer*

How is it that a DNA sequence from one organism can integrate into the genome of another without any sexual exchange? The answer to this question remains incomplete. Bacterial transformation, accidentally discovered as early as the 1920s (see Chapter 1), was studied in detail during the 1970s. Its mechanisms differ slightly according to the bacterial species. But in all cases, DNA fragments from the external environment enter the cell where they, or their copies, become inserted into the chromosome through the normal mechanisms of DNA replication, recombination and repair. The transformation of eukaryotic cells, discovered later, proceeds from the same molecular mechanisms, even if, in this case, the foreign DNA must also find its way to the cellular compartments that contain the genome: mainly the nucleus, but also mitochondria or chloroplasts in some cases. The nucleus

seems particularly suitable for incorporating any DNA present in the cell and inserting it into the chromosomes, particularly because the repair of double-stranded breaks in chromosomal DNA can be done by using foreign DNA fragments that happen to be located near the break. Thus, mitochondrial DNA fragments (called NUMT) or plastid DNA fragments (NUPT) of variable size are inserted into nuclear chromosomes, where they can exert a mutagenic action. Our own genome contains several hundred NUMTs. But genetic transfers between organisms do not necessarily involve DNA molecules. Exchanges of messenger RNA have been shown between host plants and their parasites (e.g. plant or fungus), which opens up another way of gene acquisition after reverse transcription of these RNAs.

The frequency of horizontal gene transfers is increased by the ecological proximity of organisms. This is particularly clear in cases of parasitism or symbiosis. For example, the sweet potato, *Ipomea batatas*, has acquired and expresses several genes of the bacteria responsible for crown gall, *Agrobacterium tumefaciens*. Similarly, in ruminants, rumen protists have genes from bacteria of the digestive tract in their genomes. The intracellular bacterium *Wolbachia pipientis*, found in 20% of insects and some nematodes, has transferred all or part of its genome to 70% of its hosts. Trypanosomes have about 50 bacterial genes. Functional genes of alpha- and beta-tubulin have been acquired by bacteria such as *Prosthecobacter*. The *URA1* gene, which allows *Saccharomyces* yeasts to live in anaerobiosis (and therefore to make wine!), has been acquired from a bacterium to replace the ancestral gene, *URA9,* which exists in other yeasts. There are many horizontal gene transfers from host plants to parasitic plants: for example, from sorghum to *Striga hermonthica* in Africa, from Brassicaceae to dodders, from *Tetrastigma sp* to its parasite of the genus *Rafflesia sp*. In the case of *Amborella trichopoda*, hundreds of genes from various dicotyledons and epiphytic mosses have been acquired, particularly mitochondrial sequences. Yeasts and other fungi have acquired nitrate assimilation genes, a specialty of the plant world. The examples could continue almost indefinitely. However, these are only the horizontal genetic exchanges that have persisted in the recipient genomes after successive generations. The others fade away at the speed of neutral mutations and it is all the more difficult to recognize them in the sequences as their traces are older.

6.2.2. *Primary endosymbioses of eukaryotes*

As already mentioned in Chapter 5, the existence of mitochondrial and plastid genomes confers a functional semi-autonomy to these organelles, with some functions depending on both these genomes and the nuclear genome. This genetic semi-autonomy has revived a hypothesis put forward at the beginning of the 20th Century, according to which these cytoplasmic organelles are the result of very distant ancestral endosymbioses. According to this hypothesis, a primitive cell would have absorbed another cell, which became its endosymbiont. Fragments of the latter's genome would have gradually been transferred to that of the host, while its remains would constitute the genomes of the current organelles (Gray 1993). Biochemical (nature of organelle membranes), cytological (number of membranes surrounding these organelles) and genetic (structure of organelle genes, and their transcription and translation mechanism) arguments support this hypothesis.

What would then be the ancestors of the chloroplasts and mitochondria that lived more than a billion years ago, before the formation of eukaryotic cells? It is obviously very difficult to answer this question precisely. On phylogenetic bases, it is currently believed that mitochondria would have derived from the phagocytosis of an α-proteobacterium (actually a related lineage of that time) by a primitive cell 1.5 billion years ago, thus giving rise to a eukaryotic cell capable of breathing molecular oxygen. The subsequent phagocytosis of a cyanobacterium (from that time) would then have given rise to cells equipped with chloroplasts and therefore capable of photosynthesis. In both cases, most of the endosymbiont genes would have been transferred to the host's nuclear chromosomes, explaining the current semi-autonomy of organelles (Zimorski *et al.* 2014). This assumption is obviously simplistic. Other endosymbioses have probably occurred, giving rise, for example, to red algae on the one hand and the Chromalveolata lineage that contains brown algae on the other. Horizontal gene acquisitions have also occurred in the genomes of organelles of certain lines. Similarly, secondary losses of organelles with their genomes have also occurred in other eukaryotic lines (rare for mitochondria, more frequent for chloroplasts). In some evolutionary lineages devoid of photosynthesis, for example *Plasmodium,* the agent of malaria (a member of the *Apicomplexes*), we find traces of an ancestral plastid DNA that is now strongly altered, which suggests that this group derives from ancestors which had been photosynthetic.

6.2.3. *Viruses and transposable elements*

Viruses are obviously the preferred vectors for horizontal gene transfer, particularly retroviruses whose cycle involves the integration of DNA complementary to their RNA into the host cell genome (see Chapter 2). In bacteria, gene transfer from one bacterium to another via a virus was observed as early as the early 1950s (Lederberg *et al.* 1951). The hereditary transmission of viral genomes inserted into that of their hosts has long been known for bacteria. By the 1960s, it was known that temperate bacteriophages inserted their genome into the genome of the bacteria they infected and were thus silently transmitted to subsequent bacterial generations, creating a phenomenon called lysogeny (the potential for lysis of a bacterial culture) that had played an important role in the emergence of molecular genetics. But do they contribute to the emergence of new host functions? A spectacular example of this is provided by the syncytins of the mammalian placenta, the organ that ensures nutrient and respiratory exchanges between mother and embryo during gestation[1].

It is now known that the genes for human[2] syncytins 1 and 2 come from a retrovirus (HERV-W) that infected a primate ancestor about 45 million years ago. In other mammalian lineages, these genes are also present, but even more surprisingly, they come from different retroviruses. Depending on the lineages, retroviral genes were therefore captured at variable dates during their evolution between about 70 and 25 million years ago. In mice, the gene comes from the IAPE-A virus and its inactivation makes gestation impossible and lethal. The acquisition of mammalian viviparity – which seems so natural to us – is therefore the result of a real "bombardment" of their ancestors' genomes by retroviruses during their evolution. This phenomenon continues before our eyes. In Australia,

1 These are eutherian mammals, more primitive mammals such as marsupials have a rudimentary placenta that only allows a gestation period of a few days.

2 Syncytin 1 has the property of fusing the membranes, causing the spontaneous fusion of adjacent cells into a syncytium (hence its name) bringing together the different cellular contents and, consequently, many nuclei. Strongly expressed in the placenta, it causes cytotrophoblasts to fuse into a syncytiotrophoblast. This is where nutrient exchanges and immune regulation between mother and fetus take place, as well as the synthesis of hormones that affect its development. Another essential role of this protein in pregnancy derives from its immunosuppressive properties: it prevents rejection reactions by the mother of an embryo that would be recognized as an exogenous antigen because of the paternal contribution to its genome.

the koala is submitted to an endogenization of the KoRV retrovirus, which reaches germ cells and is currently spreading, so that more than 100 copies of the viral genome may be found in the genomes of some animals[3]. It should be recalled here that, if acquisitions become hereditary, it is not an oriented evolution towards an end. The functions acquired by horizontal transfer may be advantageous (placenta), but their acquisition remains accidental (infection).

Another example of the accidental nature of acquired sequences is provided by the ***CRISPR**** bacterial system that has become well known in recent years because it forms the basis of genome manipulation tools (see section 6.3). In this case, bacteria use a special endonuclease, called *Cas*, to cut a small segment of DNA from the viruses that infect them and integrate it into a particular locus of their own genome, where it will be transcribed into small RNAs. Thus, bacteria transmit to their descendants the information of their accidental encounter with this virus. Using the small RNAs, the progeny bacteria will use the same endonuclease to recognize and destroy viruses from families already encountered by their ancestors, even though they have never encountered them themselves. They inherit an immunization. It should be noted that here, as in all cases where the acquisition becomes hereditary, the molecular mechanisms that make the phenomenon possible are pre-existing (viral DNA cleavage and integration), but that the nature of the acquisition depends on fortuitous circumstances (infection of a cell by a virus).

What role do the transposable elements play in these horizontal exchanges? Are they also "flying elements", as their phylogenetic distribution seems to indicate? The analysis of their sequences leads us to the conclusion that these elements propagate between living species by horizontal transfer, without a clear understanding of the mechanisms that allow them to do so. For example, the grapevine and the clementine tree share retrotransposons, which have 94% sequence identity, while

3 The integration of retroviral genes or whole retrovirus genomes into host genomes has other effects. It can cause gene inactivation, but also may have ***enhancer*** effects by integrating upstream of promoters. It is estimated that more than 10,000 genes in the human genome are thus subjected to greater expression. In addition, the dispersion of identical sequences in the genome is likely to create a coordinated expression of certain genes, which will then constitute functional networks.

orthologous genes between these two species have only 79% identity on average. The mobile elements are therefore closer to each other than the genomes carrying them. The role of parasitism in horizontal DNA transfers is certainly important. For example, it has recently been discovered that *Bracoviruses* that live in symbiosis with parasitic wasps of lepidopteran caterpillars are injected into the host body at the same time as wasp eggs (Drezen *et al.* 2017). Viral DNA is then integrated into the genome of the Lepidoptera along with wasp genes and can be transmitted to its germ line cells, and therefore its offspring, if the animal survives. This gene vector between Hymenoptera and Lepidoptera has been active for about 100 million years, and these exchanges play a role in the arms race between these species and their common pathogens, *Baculoviruses*.

Apart from parasitism, pathogens common to several species can probably play the same role, whether viruses, bacteria, or fungi. After accidental introduction of transposable elements into a new genome by any of the mechanisms already mentioned, they can multiply or be eliminated over generations, leaving only traces that are increasingly difficult to identify over time. At the rate of neutral yeast mutations, for example, 50% of the nucleotides in the genome are replaced at least once in just over one billion generations, which corresponds to only a few million years. Values of the same order of magnitude (half-life of 1.7 million years) have been calculated for transposons in rice, for example. The random acquisition of genes in genomes is therefore probably much more frequent than what can be detected by examining the sequences, the vast majority of these events having no consequences helping to preserve them during the evolution of the lineages.

6.3. Directed manipulations of genomes: principles and tools

We see that nature has a wide range of molecular tools and functional processes to modify genomes. But there is absolutely no indication that these changes have a purpose, or that they are driven by conditions outside the organism. Mutagenesis, which geneticists have traditionally used extensively, is a random process. Although the physical or chemical nature of the mutagen makes it possible to choose the type of mutation obtained, there is no way to target the genes that will be affected and spare others. The choice of interesting mutants is made *a posteriori*, not *a priori*. And when an interesting mutant is obtained, whether from microbes, experimental models

or plants and animals of agronomic interest, all secondary mutations induced in the genome by the mutagenesis must first be removed before clean mutant lineages of interest are obtained, which involves many successive crosses.

The situation is quite different if the mutagen is DNA itself. Of all mutagens, DNA is the most powerful, but also the only one that is truly specific. Once we learned how to transform cells with DNA (which nature also does spontaneously; see section 6.2), it became possible to use DNA to produce *a priori* selected mutations. This is called **directed mutagenesis**. Directed mutagenesis became possible from the 1980s onwards, when chemical DNA synthesis began to simplify and the first synthetic oligonucleotides could be produced at a reasonable cost (see Chapter 4). It introduced a major change in genetics, as the mutations produced were no longer random, but targeted and determined by the mutagen. At the same time, the development of specialized cloning vectors, such as those derived from the M13 bacteriophage, considerably facilitated the steps of directed mutagenesis and very many experiments were carried out. But while, the expected mutations in the genes studied were actually obtained, for the first time in history, they were still carried by the cloning vector, they were not in their original genome. The experimenter could then attempt to place the mutant genes obtained in the desired genomes, but this procedure remained difficult because, with few exceptions, such as yeast and some bacteria, transgenes were mainly randomly integrated into the genomes, with a very small proportion of transformants receiving the transgene at the targeted locus. The screening of transformants then became the tedious, but essential, step in any directed genetic construction, considerably limiting therapeutic or agronomic applications (see chapters 7 and 8). Advances in our understanding of the molecular mechanisms of homologous recombination facilitated the experiments, but did not really change the fundamentals on which they were based.

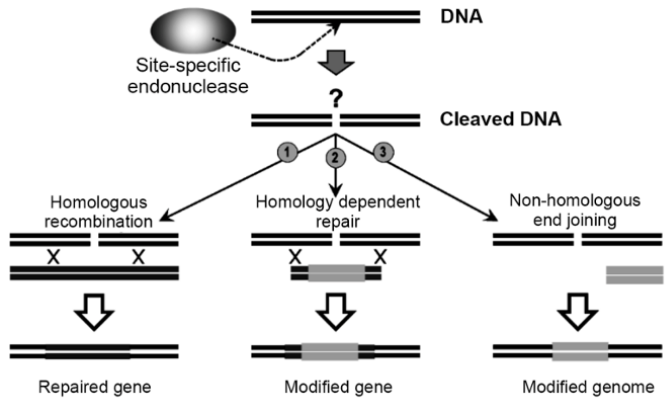
The situation was to change from the late 1980s onwards, with the discovery of new endonucleases whose recognition specificity of DNA sequences was such that they offered hope to target a single specific locus in whole eukaryotic genomes such as those of plants or mammals. The first endonucleases of this type were derived from mitochondrial introns of yeast (Dujon 2005). They received the name homing endonucleases, because they are responsible for a natural phenomenon of ***gene drive****, called intron

homing[4]. Homing endonucleases recognize long DNA sequences, typically 18 to 24 nucleotides, through specific interactions between amino acids and nucleotides, and therefore allow a single site to be cleaved in complex genomes such as the human genome, provided that their natural site is present or previously integrated. Many artificial homing endonucleases were synthesized during the 2000s, considerably broadening the range of available target sites.

Two other classes of artificial nucleases were successively constructed during the same period, by fusing the gene of a type II bacterial endonuclease with artificial proteins capable of recognizing long DNA sequences derived from bacterial transcription factors. The first were based on multiple "zinc finger" proteins, each "finger" being composed of a few amino acids capable of recognizing sequences of a few nucleotides (Carroll 2014). The second were based on proteins made of successions of very short amino acid patterns, each recognizing sequences of a few nucleotides. In both cases, the artificial assembly of amino acid patterns allowed the construction of new nucleases able to recognize, within certain limits, the new DNA sequences that we wanted to target. Regardless of the type of endonuclease used, the principles remained the same. Cleaved DNA was picked up by either of the molecular repair mechanisms available in the cell (see Box 6.1), resulting in different types of genetic modifications depending on the type of experiment performed.

A few years ago, a new class of specific endonuclease was discovered in bacteria whose DNA sequence recognition is guided not by amino acid interactions but by short single-stranded RNA molecules that bind to the endonuclease (Makarova *et al.* 2011). It is a natural system used by bacteria to recognize and destroy bacteriophages that their ancestors have already encountered (see section 6.2). This system, quickly modified for applications, gave rise to the CRISPR/Cas family of tools, which quickly overturned other genome editing techniques because of their ease of use. Their ease of use is due to the fact that the recognition of DNA sequences depends only on the small RNA that guides the Cas nuclease, and it is very easy to have such RNAs synthesized by a cell: you need only to transform the cell with a synthetic DNA of the appropriate sequence.

4 In crosses, the allele with the intron replaces all alleles devoid of introns, thus ensuring a very effective spread of these introns in populations.



A specific **endonuclease** (gray oval) recognizes a precise sequence of the genome (dotted arrow) and cleaves the DNA at this location by hydrolysis of the phosphodiester bonds between the nucleotides. The cleaved DNA will be repaired in one of three ways, depending on the other DNA molecules present in the cell:

– in the presence of a homologous DNA molecule (e.g. in the case of a diploid), the cut will be repaired using the homologous DNA as a matrix by a mechanism of homologous recombination. The result may be identical to the original;

– in the presence of a DNA molecule carrying homologous sequences on either side of a heterologous sequence, the cut will be repaired using the homologous sequences, resulting in the insertion of the heterologous DNA at the chosen site;

– in the absence of homologous sequence, the cut will be repaired by the non-homologous end-joining mechanism which, in the presence of heterologous DNA, may lead to its integration into the genome at the chosen site.

Currently, most genome engineering experiments use a bacterial endonuclease from the **Cas** family that is guided to its target sequence by a small **RNA**, ensuring the specificity by complementarity of its bases with genomic DNA. The system is called ***CRISPR-Cas****. Previously, artificial endonucleases had been developed by fusing bacterial restriction endonucleases with artificial proteins capable of recognizing DNA sequences. The latter, derived from *TALE* proteins (transcription activator-like effector) or "zinc finger" families, gave rise to the **Talen** and **Zing-finger** nucleases, respectively. Even before that, the natural endonucleases encoded by Group I introns (**homing endonucleases**) had been manipulated to modify the DNA sequences they are able to recognize.

Box 6.1. *Genome engineering*

The most commonly used system, CRISPR/Cas9, is derived from the bacteria *Streptococcus pyogenes*. Other similar systems have been found in many species of bacteria or archaea, and there is no doubt that the list is far from closed. Among these, the Cas enzymes of *Staphylococcus aureus*, *Streptococcus thermophiles* and *Neisseria meningitidis* have the advantage of a smaller size, which facilitates the cloning of the corresponding nucleic acid sequences in virus vectors. The variants of the enzyme Cpf1 found in *Francisella novicida*, *Acidaminococcus* species and *Lachnospiraceae bacterium* have the same advantage, in addition to a greater positioning versatility on the target DNA and a cleavage of the two DNA strands leaving single stranded ends that subsequently facilitate the insertion of donor DNA.

Genome editing offers great opportunities for investigation in all branches of biology and their applications, whether in terms of human health or the selection of microorganisms, animals or plants useful to humans. These are targeted DNA modifications, at the base pair level, that come at the right time to exploit our rapidly advancing knowledge of genes, of their sequences and functions. Whether made with one or the other of the endonuclease classes mentioned above, genome editing relies on the same rules that apply to the repair of cleaved DNA. If the DNA supplied to the cell carries at both ends homologous sequences to those flanking the cleavage site, this DNA will be integrated into the cleaved site by the cellular mechanisms of homologous recombination (HR) or homology dependent repair (HDR). Thus, a gene or gene fragment can be modified, repaired or inactivated as desired by playing with the sequence of the donor DNA that can be chemically synthesized. If no sequence homologous to those flanking the cleavage site is present in the cell, the cut will be repaired by the mechanism of non-homologous end joining (NHEJ) which eliminates or adds a few nucleotides at the junction. In the absence of any other cleaved DNA in the cell, the mechanism will therefore result in the inactivation of a gene (e.g. by causing a frameshift in a coding sequence) or another active genetic element (e.g. a promoter). When another cleaved DNA is present in the cell, it can be integrated into the chromosome at the cut site with modifications of some nucleotides at the junctions. A foreign DNA fragment can therefore be integrated into a genome at a chosen site. In practice, the effectiveness of the operation depends on the efficiency of cleavage of the nuclease used and its specificity in recognizing DNA sequences, two properties that are antagonistic to each other (Jiang and Doudna 2017). A highly active or highly concentrated nuclease will efficiently cut the target site, but may generate parasitic cuts on other sequences, more or less similar,

elsewhere in the genome. A less active nuclease or one produced in lower concentration will tend to be more specific. The difference in cleavage specificity also depends on the targeted sequence. The fewer similar sequences in the genome there are, the greater the cleavage specificity will be. Computer programs allow the selection of best targets from complete genome sequences.

6.4. Directed manipulations of genomes: applications

The applications of directed genome modifications are countless and recent developments of CRISPR-Cas systems have considerably accelerated them by facilitating the procedures and making it possible to target not only one site per genome, but several simultaneously. In the field of health, several human genetic diseases are currently being studied in animal models, particularly in mice. Genome editing in these animals has shown its effectiveness in correcting hereditary tyrosinemia (hepatocyte fumaryl-acetoacetate hydrolase gene), Duchenne muscular dystrophy (dystrophin gene) and hemophilia B (factor IX gene). Genome editing of stem cells is being studied, with the hope of therapeutic developments to repair affected tissues or organs. Experiments have already been carried out on human embryos that are not viable because they are triploid. Assays based on artificial bacteriophages carrying the sequences for guide RNA and Cas nucleases have been designed try to control bacterial infections by modifying their resistance to antibiotics. Similarly, one could probably think to act on the intestinal microbiota (see Chapter 8).

In the field of biotechnology, the CRISPR-Cas defense system has been introduced into lactic acid bacteria that did not have it, improving their resistance to bacteriophages, whose infections are problematic in the industry of milk-derived products. Similarly, the industrial production of some metabolites has been improved. For example, by simultaneously inactivating 5 genes, the production of mevalonate (precursor of the antimalarial drug, artemisinin) and, independently, farnesene (diesel substitute) was increased by a factor of 41 times.

In the field of animal breeding, genome editing has been used to increase muscle mass by inactivating the myostatin gene (cattle, goats, sheep, pigs), resistance to diseases (cattle, pigs), elimination of allergens from milk (beta-lactoglobulin gene in cattle and goats) or eggs (ovalbumin gene), elimination

of a prion[5] protein (goats). In plant breeding, dozens of examples include the inactivation of vacuolar invertase in potatoes (frying without acrylamide production), apple phenol oxidase (fruits that no longer turn black when cut), corn inositol-phospho-kinase (reduction of phytic acid in the grain that blocks phosphate assimilation), genes from the caffeine biosynthesis pathway in coffee (Robusta without caffeine), the three *Mlo* genes of common wheat (powdery mildew resistance), the transcription factor Os ERF922 (rice resistance to rice blast), the gene *CsLob1* (citrus resistance to bacterial canker), etc. International initiatives have been launched in developing countries, particularly in Africa, to create genomic resources and apply modern methods of genetic improvement, including genome editing, in about 100 economically and socio-culturally important[6] species. China, the United States and Russia want the forefront of this methodological revolution in selection. The application of these methods is hampered in some parts of the world, particularly in Europe, by restrictive regulations and hostile public opinions wanting to ignore the evolution of genetic knowledge.

Within current genome editing technologies one can also imagine spontaneously propagating the desired genetic modifications to entire natural populations through the reproductive modes of each species. This is now called **gene drive**. Again, gene drive is a phenomenon that exists in nature (see section 6.3), but has received little attention so far. Particular alleles are transmitted preferentially to the offspring to the detriment of the other alleles, ignoring Mendelian proportions. If crosses are random, their frequency increases in populations over successive generations and can quickly lead to their ***fixation**** in populations.

The CRISPR/Cas system offers the possibility of constructing artificial gene drive systems. For example, to eliminate a gene it is sufficient to introduce into it an artificial sequence producing the Cas9 protein and a guide RNA targeting the natural allele. When the two alleles are present in the same diploid genome, the produced Cas9 protein will cleave the normal allele which will be repaired using the allele that carries the artificial sequence, thus propagating the latter. This strategy is used, for example, to try to reduce populations of undesirable species, such as insect vectors of

5 Prion proteins are infectious agents responsible for alterations in the nervous system (Creutzfeldt-Jakob, scrapie, mad cow disease).

6 http://africanorphancrops.org.

diseases. By targeting a gene necessary for female fertility, their fertilization by males carrying the manipulated allele will sterilize populations. Similarly, populations of mosquitoes carrying infectious agents such as the malaria parasite or the "zika" virus can be replaced by populations unable to transmit these infectious agents due to the inactivation of the genes necessary for their presence in the insect. We can also imagine eradicating invasive species by spreading harmful genes or, on the contrary, promoting the propagation of endangered species. Many projects are being studied in laboratories, but it is still too early to implement them in nature, as many uncertainties persist, the greatest concern being to control the process in time and space so that it does not escape its purpose, for example by spreading to other species. Early applications could target invasive species, such as rodents, in islands isolated from the continents. The development of an international regulatory framework is necessary. Another problem is the time response of the targeted populations, as spontaneous mutations of resistance to gene drive are expected to appear and will be automatically selected.

Finally, the CRISPR/Cas system is not limited to cleaving the targeted DNA sequence. Many changes have already been made to the Cas protein. For example, by grafting on it a cytidine deaminase, a particular cytidine of the target gene sequence can be converted *in situ* to uridine, resulting in a mutation of the GC base pair to AT. The grafting of an adenine deaminase allows the reverse mutation to be carried out. GC to TA mutations can also be obtained by chemical mutagenesis with EMS, but in this case they occur randomly on the different GC base pairs throughout the genome, whereas with the modified CRISPR/Cas system, only one base pair is reached, the one chosen *a priori*. It can therefore be seen that, contrary to the ideas often conveyed in public opinion, the editing of genomes is much more reliable and less risky than the procedures used previously. Similarly, by grafting a methylase or demethylase onto the Cas protein, the methylation of DNA from a promoter region can be modified as desired to study epigenetic regulations or any other reason. The current abundance of research on these enzymes systems suggests that directed genome manipulation is only at the beginning of its potential.

6.5. Important ideas to remember

– Far from immutable and optimized structures, genomes are subject to changes due to spontaneous **structural mutations** or accidental encounters

with other organisms (hybridizations). These changes may or may not have immediate phenotypic consequences, but they are of great importance on the evolutionary scale.

– There are many and varied situations in which important genes for an organism are inherited from **accidental acquisitions** in the genomes of its ancestors through horizontal transfer mechanisms from alien sources.

– The gradual erasure of ancestral sequences at the rate of **neutral mutations** is a rapid phenomenon on the evolutionary scale.

– A wide range of natural molecular mechanisms are known to specifically modify genomes in a targeted manner. Some of them are used as **tools for genome editing**.

6.6. References

Carroll, D. (2014). Genome engineering with targetable nucleases. *Annual Review of Biochemistry*, 83, 409–439.

Drezen, J.M., Josse, T., Bézier, A., Gauthier, J., Huguet, E., Herniou, E.A. (2017). Impact of lateral transfers on the genomes of *Lepidoptera*. *Genes*, 8(11), E315.

Dujon, B. (2005). Homing endonucleases and the yeast mitochondrial ω locus – A historical perspective. In *Homing Endonucleases and Inteins*, Belfort, M. (ed.), 11–31. Springer Verlag, Berlin.

Gray, M.W. (1993). Origin and evolution of organelle genomes. *Current Opinion in Genetics & Development*, 3, 884–890.

Hijmans, R.J., Gavrilenko, T., Stephenson, S., Bamberg, J., Salas, A., Spooner, D.M. (2007). Geographical and environmental range expansion through polyploidy in wild potatoes (Solanum section Petota). *Global Ecology and Biogeography*, 16, 485–495.

Jaillon, O. *et al*. (2004). Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431, 946–957.

Jaillon, O. *et al*. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449, 463–467.

Jiang, F., Doudna, J.A. (2017). CRISPR-Cas9 structures and mechanisms. *Annual Review of Biophysics*, 46, 505–529.

Lederberg, J., Lederberg, E.M., Zinder, N.D., Lively, E.R. (1951). Recombination analysis of bacterial heredity. *Cold Spring Harbor Symposia on Quantitative Biology*, 16, 413–443.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J., Wolf, Y.I., Yakunin, A.F., Van der Oost, J., Koonin, E.V. (2011). Evolution and classification of the CRISPR-Cas systems. *Nature Reviews Microbiology*, 9, 467–477.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, 440, 341–345.

Soucy, S.M., Huang, J., Gogarten, J.P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16, 472–482.

Wolfe, K.H. (2015). Origin of the yeast whole-genome duplication. *PLOS Biology*, 13(8), e1002221.

Zimorski, V., Ku, C., Martin, W.F., Gould, S.B. (2014). Endosymbiotic theory for organelle origins. *Current Opinion in Microbiology*, 22, 38–48.

7

Of Genes and Humans

Since the birth of civilizations, humanity has increasingly made its mark on the biosphere, particularly through domestication and selection of needed species. Within the human species itself, some characteristics essential to its existence have been selected during ancient migrations and changes of lifestyles. These changes can be deduced from the study of current genomes, where they have left traces over time. But recently, thanks to the improvement of techniques to purify and sequence ancient DNA, present in bones or plant remains several tens of thousands of years old, it has become possible to go back "directly" in time. In the space of a few years, paleogenomics has revolutionized our understanding of the genealogical links between human populations and our knowledge of migrations from Africa to Europe, Asia, Oceania and the Americas for about 600,000 years. Paleogenomics in the same way regards the history of other animal or plant species, including extinct species.

In the genomes of our contemporaries we find the traces of the genetic selections that took place due to the lifestyles of their ancestors and in response to the adversity of climates and diseases. In this way, we can write the evolutionary history of mankind, in other words find its origins, trace its migrations, understand its adaptations to the various environments encountered, altitudes, climates, pathogens and types of food that have enabled it to survive. This recent history can also be tracked in wildlife species, such as forest trees, which were harvested but not cultivated. European oaks have been particularly studied in this perspective.

The Neolithic period is considered as the pivotal period that saw the multiplication of domesticated species for agriculture and livestock. Plant

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

domestication continued with a more or less conscious and directed selection under different soil and climate conditions, in different latitudes, where isolation, migration and exchanges played major roles. In this respect, the relatively recent discovery of the New World was an episode of great importance for the diversity of cultivated species.

Animal selection through the choice of breeding stock began very early. There was no mystery about how to conduct animal reproduction (of mammals), while for plants it was not until the Enlightenment that their reproduction began to be understood. It was finally during the 20th Century that the rational selection of plants and animals was developed, making a decisive contribution to the quantitative and qualitative progress of world agricultural production. Another revolution in the improvement and diversification of animal breeds and plant varieties is beginning with the study of genomes.

7.1. Ancient DNA and human history

A new phase in our knowledge of human history began a few years ago. The turning point was a first draft of the Neanderthal human genome (Green *et al.* 2010), obtained in 2010 from bones from the Vindija cave in Croatia, dated between 44,500 and 40,000 years ago, to which were added other samples from Spain (El Sidron, 49,000 years ago), Germany (Feldhofer, 42,000 years ago) and the Russian Caucasus (Mezmaiskaya, between 70,000 and 60,000 years ago). This was a truly amazing feat, because this DNA is highly fragmented into molecules of 200 base pairs on average, chemically modified by deamination of cytosines into uracils, and contaminated by the DNA of microorganisms, and possibly by that of the discoverers of these bones as well.

In 2014, the first "complete" version of the Neanderthal genome was published (Prüfer *et al.* 2014), taking advantage of a toe phalanx at least 50,000 years old, found in 2010 in the Denisova cave in Altai. Its DNA was that of a female individual whose high degree of consanguinity (for one-eighth of its genes, the two alleles were identical) indicates that her two parents could have been half-brother and half-sister, for example. On the same site, three teeth and a phalanx belonging to another individual provided a very different genome (more than 1% SNP divergence between the sequences) and was therefore attributed to "the Denisovan human", hitherto unknown, because no other skeletal element has yet been discovered that

would allow us to compare his/her morphology with those of Neanderthals or modern humans: another hominid existed, contemporary of Neanderthals but only known by its DNA (Reich *et al*. 2010; Meyer *et al*. 2012).

The genomes of these hominids were immediately analyzed in comparison with the genomes of today's humans, whose sequences can now be counted by the thousands. As early as 2010, the first surprise was to discover fragments of the Neanderthal genome in contemporary genomes, signatures of ancient inter-breeding. It was therefore no longer possible, as it had been done until then, to talk about two species. More interestingly, these fragments, which represent on average 2–3% of an individual's genome, are not always the same from one individual to another, so that we can estimate that 20% of the Neanderthal genome is preserved in the current human population. These Neanderthal traces which are probably the result of multiple inter-breeding, are more numerous in Central Asian populations than in European populations. The Neanderthal genome of the Altai also contains signatures of the modern human genome, reinforcing the idea of frequent inter-breeding. Similarly, the Oceanian populations (Melanesian, Papuan and Aborigines) have preserved signs of ancient genetic exchanges with the Denisova population. Some characteristics, such as the light color of skin or hair or elements of the immune system (Deschamps *et al*. 2016), may have benefited modern humans. Representing long term adaptations of the ancestral indigenous populations to Eurasian conditions, they were quickly transmitted to the newcomers through these genetic exchanges. Nevertheless, Neanderthal or Denisova alleles are much rarer in chromosomal regions containing genes strongly expressed in the testicles, or carried by the X chromosome. Low fertility of mixed-race males may be the cause of this phenomenon (Sankararaman *et al.* 2016). Current African populations do not show these signs of ancient genetic exchanges, consistent with the confinement of Neanderthals and Denisovans in Eurasia.

A scenario now emerges based on these data supported by numerous studies undertaken since then, in particular on genomes isolated from the skeletons of modern humans from all the continents that lived during the last 50,000 years (Nielsen *et al.* 2017). A first migration of populations from Africa occurred around 600,000 years ago, from which are derived Neanderthals and Denisovans, who occupied Eurasia around 300,000 years ago in the form of dispersed populations of small size. Between 190,000 and 120,000 years ago, new populations of modern morphological types left Africa, where they had differentiated between 350,000 and 260,000 years

ago in this continent. They dispersed and multiplied in Eurasia, where they mixed with their predecessors, Neanderthals and Denisovans. From generation to generation, this inter-breeding "diluted" the genomes of the latter into the genomes of modern humans whose numbers or fertility were higher, until the "pure" Neanderthal types disappeared. There is no need to imagine furious wars or competitions between these different human types: rather the opposite!

Between 55,000 and 47,500 years ago, the populations of modern humans who were in contact with Denisovans populated Oceania, those who were in contact with Neanderthals, populated Europe and Asia. From Asia, the populations of Pre-Columbian America migrated through the Behring Strait between 23,000 and 15,000 years ago. In the same period, western Europe experienced at least three waves of migration, a first one around 43,000 years ago, from populations who left Africa between 65,000 and 55,000 years ago and settled in the Middle East. These populations disappeared before the last glacial maximum and were replaced around 8,000 years ago by farmer-breeders from Anatolia, supplemented during the Bronze Age by nomads from the Pontic and Caucasian steppes.

It is likely that the growing number of studies on this subject will quickly detail these scenarios. The study of ancient human genomes has progressed rapidly in recent years. After studies on single genomes, genomic studies have developed on populations comprising hundreds of ancient individuals. This has provided a wealth of information on human evolutionary genomics, but this field is still in its infancy. What can we expect from the development of these studies on the ancient DNA of hominids? Will we see the discovery of new archaic lineages ignored until now by prospecting under-explored regions of the world, such as the contemporary African populations? Will we be able to disentangle the relative contributions of population isolation, migration and mixing in the biological and cultural variation of modern human groups? For many parts of the world, particularly in Asia and Africa, our understanding of the historical events that have led to the current distribution of genetic and cultural variations is still very fragmentary, especially since soil and climate conditions are often unfavorable to the conservation of ancient DNA. Continued efforts to sequence and analyze the genomes of modern and ancient humans, with a focus on under-sampled regions of the world, will help us to build a more complete picture of the events that have shaped the cultural and genetic variation of contemporary populations.

7.2. Traces of the past in today's human genome

The "International HapMap" project, using genotyping technologies, and more recently the "1,000 genomes" project, using next generation sequencing techniques (see Chapter 4, Box 4.3.), have made it possible to characterize the genetic variation of current human populations throughout the world. The "1,000 genomes" project, which involved 26 populations from the five inhabited continents, revealed more than 80 million genetic variations from the reference genome. In other words, a random individual shows on average between 4 and 5 million differences with the reference genome, mostly changes involving one nucleotide (SNP) or short DNA insertion-deletions and more than 2,000 larger structural changes, such as large deletions or inversions. What is also striking in the results of these studies is that, on average, each of us has lost the function of 250 to 300 known genes and that we carry 50 to 100 alleles involved in genetic diseases, generally in the heterozygous state. These studies also show how natural selection pressures, such as climate, latitude, altitude, diseases, or socio-cultural factors such as hygiene conditions or food security, have shaped the genetic variation of these populations (Quintana-Murci 2016). Indeed, to follow a "Darwinian" reasoning, adverse mutations in a certain environment will tend to disappear and conversely, favorable mutations in a particular context will be preferred and will spread among the population (Barreiro *et al.* 2008).

In order to detect these selection events in human populations, several tests are available to measure whether there is a deviation from the assumption of neutrality, that is, no positive or negative selection. This may be exemplified by the distribution of allele frequencies in a given population where an excess of rare alleles in the overall population is a sign of positive or negative selection. The genetic differentiation between populations will tend to increase if they are subject to different and geographically defined selection pressures. Other tests are based on comparing the frequency of a given allele, taking into account the length of the particular DNA sequence where it lies (haplotypic block). Indeed, due to meiotic recombinations, haplotypic blocks decrease in size from generation to generation. Older mutations are associated with short blocks and newer mutations with longer blocks. If the mutation is advantageous, it will therefore be positively selected, and a long haplotypic block will be unusually common in the population.

7.2.1. *Adaptations to the world's regions*

An example of genetic adaptation to changing environments is provided by the exposure of ancestral populations to colder climates and lower levels of solar radiation after their first migrations outside Africa. These changes in climatic conditions have led to a variation in the quantity, type and distribution of melanin, leading to different levels of skin pigmentation. Darker skins are seen in areas with high UV radiation, due to the protection they provide against sun damage (e.g. sunburn, melanomas, carcinomas). An even greater advantage comes from the fact that darker skins protect against UV-induced photolysis of folic acid. This metabolite is essential for the embryonic development of the neural tube and for spermatogenesis. Conversely, lighter skins can confer a real selective advantage in areas with low levels of UV radiation, with a higher level of vitamin D photosynthesis. A number of genes involved at different stages of skin pigmentation have been identified, from melanogenesis to the production and maintenance of melanosomes and the distribution between the production of eumelanin and pheomelanin. White skins in the Eurasian continent can result from different mutations of different genes (*ASIP*, *OCA2*, *SLC24A5*, *MATP*, *TYR*), while oriental skin colors depend on other mutations of other genes (*ADAM17*, *ATRN*).

Another striking example is the adaptation to life at high altitude of the Andeans on Altiplano and Tibetans on the Himalayan plateau, where partial oxygen pressure can decrease by 40%. They must avoid the physiological stress of hypoxia. The genetic basis for these adaptations to extreme environmental conditions has only recently begun to be deciphered. Tibetans on one side or Andeans on the other have selected different alleles of the *EGLN1* gene and different genes, *EPAS1* in the former and *PRKAAI* and *NOS2A* in the latter. This clearly shows that Tibetans and Andeans have followed different evolutionary trajectories to adapt to high altitude, illustrating the phenomenon of convergence.

7.2.2. *Adaptations to lifestyles*

Many examples of the genetic adaptation of human beings to diet have been described: milk consumption, starchy diets, perception of bitter taste. Adaptation to milk consumption, through the persistence of lactase in adulthood, is probably one of the best known examples of natural selection

in humans. A high concentration of lactase ensures that lactose is effectively digested during the first few weeks of life. The lactase gene is then gradually repressed, resulting in residual concentrations (5 to 10% of neonatal concentrations) of the enzyme in adults. However, some populations, particularly those that have traditionally bred livestock, maintain their ability to digest milk as adults. This trait of lactase persistence is common in Europe and in some populations in West Africa and the Middle East. It can reach frequencies of 90% in Northwest Europe. This type of mutation was not selected in other populations of herders consuming fermented milk (cheeses, yoghurts). The persistence allele 13910T is found in Europeans, and it is another mutation, 14010C, of the same gene and with the same effect that is detected in populations in East Africa. These mutations are located in the promoter part of the gene, and result in the increase of its expression. These selections have taken place over the past 7,000 years, with pastoralism, cattle or sheep breeding, and have not affected other populations, such as those in the Far East where, for example, 90% of the Chinese population is lactose intolerant. In total, these results show that the cultural characteristic of milk consumption has conferred a selective advantage in terms of human survival in different parts of the world. The most obvious selective advantage provided by the persistence of lactase is the ability of lactose-tolerant individuals to access a valuable food source in the event of food shortage, while lactose-intolerant individuals present diarrheal syndromes with dehydration.

7.2.3. *Adaptations to diseases*

Another important selective pressure that humans have faced over time is that imposed by pathogens and infectious diseases. Indeed, pathogens have been and are still the main causes of death in regions where antibiotics and vaccines are still lacking, thus exerting strong selective pressure on the human genome (Quintana-Murci and Barreiro 2010; Fumagalli *et al*. 2011). A balanced selection caused by pathogens has been clearly demonstrated for the Human Leukocyte Antigen (HLA) gene, whose very high diversity is strongly correlated with residence in areas where a large number of pathogenic species exist. The high frequency of some hemoglobinopathies has been correlated with greater resistance to *Plasmodium falciparum*, the agent of malaria. The HbS allele or "sickle cell trait" is an example of positive selection, as it increases resistance to life-threatening forms in heterozygotes. Another example is glucose-6-phosphate dehydrogenase

(G6PD) deficiency. Patients with this disease have abnormally low levels of G6PD, an enzyme that is particularly important for red blood cell metabolism. More than 100 alleles can lead to this deficiency, some of which are selected because they provide greater protection against *P. falciparum* or *Plasmodium vivax*. An extreme example is provided by the *DARC* gene, whose null allele leading to the absence of protein, prevents *P. vivax* from entering host cells. Positive selection for the null allele has been demonstrated in sub-Saharan Africa, and it is almost fixed in some Central African populations, while it is practically absent in populations from other parts of the world. *P. vivax* and other pathogens using the same mode of entry into host cells have been identified as very likely sources of selective pressure.

7.2.4. *Maladaptation following past selections*

It is increasingly clear that some diseases in modern societies, such as obesity, hypertension, inflammatory, autoimmune and allergic diseases, or even cancers, may simply be the result of past adaptation to other selective forces, while recent lifestyle changes have been too rapid for the spreading of the genetic modifications necessary to adapt to them. For example, some alleles were selected because they promoted the accumulation of large amounts of fat, significantly increasing the probability of survival in the event of malnutrition or famine. For example, a very rare allele of the *CREBRF* gene in the general human population is extremely common in Samoans and is believed to have contributed to the expansion of these populations in Polynesia facing long periods of crossing between the Pacific islands. Today, the transition to a sedentary lifestyle and an abundance of food have increased the risk of developing type II diabetes in individuals with these alleles. This is also the case for Native American populations that have recently and rapidly moved to a "Western lifestyle" in the United States. Similarly, a strong and exacerbated immune response, which may have been the best way to survive in pathogen-rich environments in the past, appears to have become a burden in modern societies, where it increases the risk of developing inflammatory and autoimmune diseases (Brinkworth and Barreiro 2014). This hypothesis, known as the "hygiene hypothesis" (Sironi and Clerici 2010), is supported by the observation that pro-inflammatory alleles are more frequent in populations of tropical origin than in those who have lived in temperate regions for a long time. For example, ethnicity was

found to be more important than environmental differences in the prevalence of asthma.

7.2.5. *Conclusion*

Population genetic studies have greatly improved our knowledge of how humans have genetically adapted to changes in environmental pressures and lifestyle changes over time. It is now possible to exploit all available genome data on different human populations and their association with several diseases or traits of interest. Multi-disciplinary efforts are also needed to improve our understanding of the evolutionary mechanisms that have led to current differences in the susceptibility, resistance or progression of diseases observed in humans. Together with epidemiological studies on populations with different lifestyles and completely different environments, they are essential to determining the contributions of genotypic, epigenetic and environmental variables to the current risk of many diseases, and to facilitate their diagnosis, prevention and treatment.

7.3. Traces of past climates in the trees of our forests

Species that are immobile and have a long reproductive cycle in natural ecosystems, such as primary forests or coral reefs, are the most vulnerable to climate change, especially if it occurs suddenly. Weather conditions have changed constantly in the past. It is estimated that 89 botanical genera of trees disappeared in Europe during the transition from the tertiary to the quaternary era 2.58 million years ago[1]. The examination over a long period of wide fluctuations over the past 320,000 years of Hungarian forest species makes it possible to distinguish between those that have disappeared locally (genera *Quercus*, *Fagus*, *Ulmus*, *Carpinus*, *Sequoia*, *Nyssa* and *Parrotia*) and those that have shown strong resilience to alternance of cold and hot periods (genera *Pinus*, *Picea*, *Abies*, *Betula*, *Zelkova*).

What consequences does the rapid climate change we are witnessing in the 21st Century have for forest species? Can past data, obtained by palaeo-botanical analysis, associated with the resulting genetic structure of current populations, allow us to predict them? The history of climate during the

1 It should be noted that the period from 66 million years ago to the present day is now grouped by geologists under the term Cenozoic.

quaternary period can serve as an example: at least 15 glacial and interglacial cycles have followed one another, with episodes of sudden changes in temperature or composition of the atmosphere (Petit *et al.* 2008). During this period, the species of the genus *Quercus* (oaks), major forest species in the northern hemisphere, developed mechanisms that reduced their risk of extinction. Other species have not been able to persist. This is the case, for example, of a spruce tree, *Picea cretchfieldi*, abundant in North America during the last glacial maximum and which disappeared during the rapid following deglaciation.

The first mechanism that allowed oaks to avoid extinction is the possibility, in the northern hemisphere, of migration of the species from north to south during the ice age and from south to north during the interglacial period (Kremer 2016). The speed of these migrations, which are estimated to be in the order of 500 to 1,000 m per year, most probably involved the intervention of human populations, for which acorns were part of their diet, because movement by rodents such as squirrels, who also participated in this dispersion, was far too limited. The oaks followed the humans! However, these migrations have also been promoted by interspecific hybridizations between *Q. petraea* (sessile oak) and *Q. robur* (pedunculate oak). Indeed, the latter is a pioneering species, better able to colonize unoccupied areas: the progression of *Q. petraea* followed by wind transport of its pollen and hybridization with *Q. robur*. More competitive for fertilization, this pollen will carry out a real substitution of the genome of the pedunculate oak by that of sessile oak following successive crosses. This mechanism is attested in particular by the presence in both species of the same plastid genome (transmitted by the female gamete) when they occupy the same forest.

The second mechanism is that of local adaptation, for example to altitude and therefore temperature gradients, which is the result of a genetic evolution of these populations through the production of new alleles by mutations and especially new combinations of alleles by crosses. The resulting genetic diversity can be measured in a variety of ways. For example, at a given locus, the allelic diversity can be measured by the probability that a diploid individual will carry two different alleles. This allelic diversity will be all the higher when the number of alleles in the population is large and their frequency distribution is unbiased. At the genome-wide level, the average of allelic diversity values for the different loci provides an estimate of the population's genetic diversity. Whole

genome sequencing offers another direct measurement by comparing the sequences (see Chapter 5, Box 5.6). The proportion of nucleotides which are different for DNA fragments of the same locus between randomly selected individuals in the population defines nucleotide diversity. For example, the nucleotide diversity is about one nucleotide in 100 in Drosophila and one in 1,000 in humans. It appears that oak and similar species, which one would assume less susceptible to variation and therefore adaptation than species with shorter generation cycles and larger populations, show high levels of nucleotide diversity. The sequencing of a diploid oak genome revealed a nucleotide diversity of one in 90 nucleotides in the coding regions on average and one in 40 in the intergenic regions.

Low genetic diversity increases population susceptibility to risks of diseases, parasites or severe climatic constraints. In addition, and this is a general law, the resulting consanguinity can lead to a decrease in the vigor of individuals and a decrease in their reproductive capacity. Many endangered populations or species have significantly lower levels of genetic diversity than species that are not endangered, further increasing the fragility of their situation. By opposition, the high nucleotide diversity of the oak tree suggests that it will be able to adapt within a few generations, particularly to climate change. However, there is no guarantee that this diversity will be sufficient, particularly with regard to pests or diseases emerging as a result of climate change.

The history of forest populations over the past millennia is enriched by modern methods of sequencing ancient DNA (Neale and Kremer 2011). To this respect, aquatic environments such as lakes and shallow seas, that have preserved the piles of Neolithic lacustrine dwellings and the oak tree trunks that populated the shores now swallowed up by the rising waters since the end of the last Ice Age, offer interesting material to study. Ligneous species represent the majority of the continental biomass and are the essential components of ecosystems rich in biodiversity. Humans will have to help them adapt to current climate change, which is much faster than glacial cycles and can lead to sudden breakages and extinctions.

7.4. The domestication of cultivated plants

At the origins of agriculture can be found the domestication of species, which means for them a certain isolation from the wild world, but also changes that result from multiple causes and lead to very different results.

Plant domestication is a consequence of the first cultivations that followed gathering. Barley and oats have been cultivated for a long time before their domestication. Indeed, for these domestication processes to take place human intention was required with a practice (to cultivate, to breed) and a project (higher yield, longer harvest, easier handling). It is more common to talk about domestic animals than domestic plants. And yet, our cultivated plants that directly or indirectly provide most of our food often differ radically from their wild ancestors. They are truly domesticated and not only tamed, and therefore they are genetically modified.

As we have seen previously, nature constantly offers very many mutations, for example the immediate germination capacity of a seed without dormancy. By opportunism, or more often as a consequence of agricultural practices, humans propagated such mutations. Cultivated wheat has undergone mutations that have transformed the spike: their grains remain attached at maturity (indehiscence), become larger with more starch reserves, lose their husks (naked grains), the spike becomes more "square" by increasing the fertility of the spikelets. The massive cultivation of durum wheat (*Triticum turgidum* subsp. *durum*), itself a hybrid between two species, has led to its spontaneous hybridization with a third wild species (*Aegilops tauschii*) which has resulted in the common wheat (*Triticum aestivum*). Unlike normal hybridization, it took the addition of their genomes (most likely through the fusion of diploid gametes produced by deregulated meioses) for the new species to reproduce. How many events of this type have shaped the cultivated species on which our civilizations ultimately depend over the centuries? By opposition, triticale, which derives from recent human hybridization between wheat and rye, did not appear spontaneously despite centuries of mixed cultivation (meslin) of the two species. To discover the traces of these processes, archaeology, ecology, genetics and molecular biology interact, knowing that while common points can be found in the final result, designated "domestication syndrome", the scenario varied from one species to another and several parallel scenarios could affect the same species.

7.4.1. *Characteristics of domestication*

The "domestication syndrome" is defined by all the morphological or physiological characteristics that differentiate the cultivated plant, whatever the variety, from its wild ancestors (Martinez-Ainsworth and Tenaillon

2016). This goes back to the traits selected by the first farmers, as opposed to the diversity created by their successors. These characteristics can be directly related to harvesting (grains that remain attached to the spike of cereals or pods that do not open spontaneously in legumes, larger fruits), consumption (reduction of chemical defenses such as bitterness of squash or almonds or physical defenses such as barbs or spikes) or cultivation (loss of seed dormancy, plants with limited height).

It was long thought that domestication traits conferred such an advantage on the first farmers that they would have adopted these new plants very quickly. This is not the case: on the contrary, archaeological data show that the process was very slow for the cereals of the Fertile Crescent, where wild and domesticated types coexist in the harvest samples dating from 7250 to 4500 BCE, the latter only gradually becoming the majority. A form of agriculture therefore preceded the domestication of these species, which must be seen as an adaptation to these practices. In other cases where the trait is clearly visible, such as the formation of the ear of corn or the absence of seeds in the fruit of banana that makes it edible, it is reasonable to assume that a conscious choice could have been made.

One ought to refer to Nikolai Vavilov's[2] concept of "center of diversity", according to which the domestication of a plant species took place in the region where the wild species from which it is derived was not only present, but had developed a maximum of genetic diversity. Today there are 11 regions of the world regarded as "centers" of origin of cultivated plants, for example, the Fertile Crescent of the Middle East, which has produced oats, wheat, barley, rye, lentils and chickpeas. The combination of genetic, archaeological, paleo-climatic and linguistic data is necessary to reconstruct domestication scenarios. The scenario can be simple, as in the case of corn domesticated in Mexico in a single region, from a teosinte population. In the case of millet in the Sahel or barley in the Middle East, the area of domestication is larger and the analysis of their genomes shows that several wild populations have contributed to each of them. The common bean has experienced independent domestications in Mexico on the one hand and in the Andes on the other. Asian rice (*Oryza sativa*) consists of three main subspecies: *japonica*, *indica* and *aus*. They are derived from the wild species *O. rufipogon* by a first domestication in China, between 24,000 and

2 Supporting Mendelism against Lysenko and Stalin, he was sentenced to death in 1941 and died in prison in 1943.

13,000 years ago, which led to the subspecies *japonica*. Significant flows of domestication genes between japonica and other wild populations of *O. rufipogon* or *O. nivara* (*proto-indica* and *proto-aus*) have led to the domestication of these two other subspecies, *indica* and *aus* (Choi *et al.* 2017).

7.4.2. *The mutations that enabled domestication*

Mendelian genetic analysis, which consists of crossing a domesticated form with a wild form and studying the progeny, enables a first approach: thus G. W. Beadle, by crossing a primitive maize with a teosinte, obtained one plant in *ca.* 500 identical to corn and one in *ca.* 500 identical to teosinte among the 50,000 plants of the hybrid's progeny made by self-fertilization (F2) under study. These are the expected proportions if both parents differ by four or five independent major genes, which is consistent with the influence of early human selection (Doebley 2004).

This quantitative approach, revealing but a limited number of differences, has encouraged molecular biologists to look for the major mutations that explain them. This is how, for example, lists of domestication traits have now been established for wheat, cabbage, beans, maize, barley, rice, soybeans, sorghum, tomatoes, etc. It is striking to note that in the majority of cases (5 times out of 6), these are mutations in genes involved in transcriptional regulation (transcription factors) of one or more developmental genes. This is the case with the *Q* gene of common wheat, which is responsible for the square section of the spike, the *BoCAL* gene, which transforms the inflorescence into the curd of cauliflower, the *PvTFL1y* gene, which confers a determinate growth and flowering to the common bean, the *tb1* gene (teosinte branched), which suppress the ramifications of the corn stalk, the *HvVRS1* gene, which constructs the 6-row type of barley, the *sh4-1*, *SHAT1-5*, *SbSH1* genes, ensuring the absence of grain shattering respectively in rice, soybean or sorghum, the gene *fw2.2*, which increases the number of locules in the fruit of the tomato, etc. Very often, mutations in these genes are caused by the insertion of transposable elements leading to a kind of reshuffling of gene networks. Far from leading to loss of function, these changes create new developmental functions (Shiu *et al.* 2005; Doebley 2006).

Plant populations involved in the domestication process have represented only a very small fraction of their species' populations. A significant loss of genetic diversity was therefore inevitable. Depending on whether or not crops can be enriched by crossing with the wild forms with which they coexist, this loss, measured by the variability of genome sequences, will be limited if the species is ***allogamous**** (around 20% in maize), or, on the contrary, massive if the species is ***autogamous**** (80% in durum wheat). These orders of magnitude highlight the importance of preserving wild genetic resources for the improvement of cultivated species, keeping in mind that spontaneous mutations that accumulate over time in cultivated populations restore some variability.

7.5. Selection of livestock

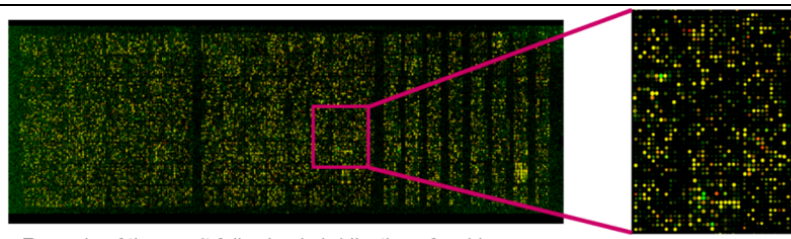
The choice of breeding animals of our domestic animal species has been made for millennia on visible traits: labor strength, physical appearance, coat color. With the advent of artificial insemination in the middle of the last century, the number of sires involved in the reproduction of the herd has become more limited. Therefore, the cattle selection rested essentially on them. But how, for example, do you choose the bull that will ensure the best milk production from its daughters? It was necessary to wait for the recording of the performances of its first daughters to estimate its genetic value and, from the results, to decide whether or not to use his semen for other inseminations. Thus, the improvement of dairy breeds was based for about 50 years on a vast system of performance recordings on all farms, and the best bulls were selected from the results of more than 100 females distributed in a large number of herds. Each bull was evaluated on about 40 traits, including milk production and composition, fertility, conformation, disease resistance, etc.

In 2001, in view of technical progress in genome sequencing, it became possible to take into account entire genomes for the selection of farm animals (Meuwissen *et al.* 2001). In 2007, after the first draft of the bovine genome sequence became available, a genotyping chip (see Box 7.1) was made available to the selectionists and the bulls undergoing traditional testing were immediately characterized. This chip made it possible to simultaneously characterize 54,000 loci regularly distributed over the genome and discriminative for a wide range of breeds. Their daughters'

performances were statistically correlated with the genome compositions revealed by the chip. Once these correlations were established in this first reference population, it became no longer necessary to continue the painful evaluation of the daughters of each bull, as bull selection could then be based solely on the value derived from their genotyping, performed using DNA chips (Boichard *et al.* 2016). It is sufficient to choose as breeder stock the bulls with the "best" genomic regions for the different selection criteria.

Genomic selection is changing selection practices (Boichard *et al.* 2012). It does not require physiological knowledge of a cause-and-effect relationship between a gene and a trait. It is a statistical method based on the correlation that can be established between the value of a trait in an individual or in its progeny and the presence, in the genome, of certain molecular markers. For each individual in the reference population, we therefore have a double entry table, where on the first dimension are the values obtained for each of the different characteristics subject to selection and on the second dimension is the presence or absence of the different molecular markers on the genome. Comparison of the different individuals in the reference population makes it possible to establish statistical correlations between characters and molecular markers. For each character's value, some markers will be positively correlated, others negatively correlated and a majority of others will be neutral. An equation predicting the reproductive value of each individual for a particular combination of traits can be developed based on its genotype, by integrating these data. By applying this equation, breeding candidates for livestock improvement will be chosen solely on the basis of their genotype.

It is also a recurring strategy: from generation to generation, new animals derived from the selection enrich the original reference population, thus completing genotypic and phenotypic databases. By increasing the number of populations, correlation values and molecular labelling become more accurate. In addition, functional genomics studies showing direct role of particular alleles in the value of certain traits are enriching the molecular labelling of genomes. All these improvements facilitate the establishment of prediction equations, using data from different breeds within which genomic selection can be generalized.



Example of the result following hybridization of a chip containing more than 20,000 DNA deposits. (Photo generously provided by Jean-Pierre Renou)

A genotyping chip consists of an **inert material**, such as glass, small in size on which single-stranded DNA fragments (synthetic **oligonucleotides**) are fixed in an orderly manner. These are either deposited in microdrops or synthesized *in situ* by a lithographic process.

The material is divided into a very large number of rows and columns forming hundreds of thousands of intersections at which there are a few hundred thousand DNA molecules of a particular sequence. Each sequence is chosen to correspond to a locus in the genome of the species studied where a **single nucleotide polymorphism** (SNP) has been identified in a reference population. Several SNPs can correspond to the sequence of the same gene. Two oligonucleotides are present for each SNP, one corresponding to the reference allele and the other to the variant allele. To ensure reproducibility of the results, these pairs are repeated several times in different regions of the chip (typically 5 times).

For a genome the size of that of a bovine, for example, one can identify by this method about a hundred thousand SNPs distributed over the entire genome in order to perform what is called **genotyping**. To do this, the animal's DNA (usually purified from blood) is fragmented, chemically labelled with a fluorophore, then denatured (separation of the two strands), and finally deposited on the chip under **conditions** where hybridization with the fixed oligonucleotides is capable of discriminating a perfect complementary sequence from a sequence carrying a single nucleotide difference. After rinsing the chip to remove all DNA fragments that have not found their exact complementary sequence among all fixed oligonucleotides, the automated reading of the fluorescence at each point on the chip will allow the animal's **genotype** to be deduced, that is the determination of each allele at each locus tested. By measuring the relative fluorescence intensities of each pair of oligonucleotides, we can further distinguish **homozygotes** from **heterozygotes**.

Box 7.1. *Genotyping chip. For a color version of this figure, see www.iste.co.uk/dujon/genetics.zip*

The consequences of genomic selection are numerous. This revolution in breeding practices has been implemented in different countries following the availability for dairy cows of the high-density genotyping chip (50,000 SNPs) since 2007. Following its rapid success, genomic selection has expanded to an increasing number of cattle breeds and other animal species where, similarly, selection takes place directly among exploited populations. For these animals, selection is fully integrated into the production process. For animal species such as fishes or for plants, selection precedes production.

In addition to the financial savings resulting from the disappearance of phenotypic tests on progenies, this strategy has made it possible to accelerate genetic gain at a lower cost, to extend selection to other more complex and difficult to select characteristics that can only be handled on the reference population and to rebalance the role of both sexes by also submitting females to selection.

At the same time, the number of genotyped cattle in France is increasing over time: there were 400,000 in databases in 2015, then 800,000 at the end of 2017. Therefore, the genetic gain per unit of time is raising, the increase in genotyped populations ensuring a better accuracy in the choice of genitors and the saving of female offspring testing reducing generation time. It is estimated that the genetic progress per unit of time is about twice as high as with the traditional method.

The genetic diversity of the whole livestock also benefits from genomic selection. At the same time, this selection method makes it possible to objectively and exhaustively characterize the biodiversity of populations, which provides tools to ensure its maintenance. The cost of selection is falling sharply, relying solely on genotyping, the cost per animal being increasingly low (around 30 euros at the end of 2016). As a result, this allows an increase in the number of reproductive animals and therefore reduces consanguinity. The selection criteria are multiplied and the balance between them is modified, with an emphasis on animal health, product quality, environmental impact and less on production potential. The multiplication of selection criteria results in the diversification of the genomic regions subject to selection. The increasing accuracy of correlations between genomic loci and selected traits combined with complete genome sequencing (Daetwyler *et al.* 2014) leads to the detection in the DNA sequence of novel causal variations, that may be rare in populations but

whose frequency can be increased by the selection. Genomic selection is also being implemented in France for "small" dairy breeds such as Tarentaise, Vosgienne and Simmental, which benefit from the possibilities of exploiting the data obtained from the majority breeds as well as from the practical experience acquired in its implementation.

7.6. Conclusion

Today's genomes are part of the history of their species, and the resulting diversity can be described in terms of variations in DNA sequences. This snapshot image of genetic variation allows the modeling of human genealogies for about 200,000 years, and, closer to us, their various migrations for about 20,000 years. All humans share the same ancestors beyond 200,000 years of age. European, Native American or Asian populations have passed through population bottlenecks, with effective genetic sizes[3] in the order of a thousand individuals between 20,000 and 15,000 years ago. The bottleneck effect on genetic variability was compensated, in part, by the rapid expansion of populations and the occupation of various continents where occasional inhospitable conditions led to the selection of favorable mutations in the subpopulations concerned. Populations in Africa where this effect was less pronounced are also those that currently show the greatest genetic diversity. A broader representation of the entire human species enriched by subsequent genome sequencing projects is needed to recognize more important functional variations, and in particular those related to health (Nature 2015).

A similar scenario applies to crop species. Domestication was carried out on small samples of plants, resulting in first cultivated populations of low genetic variability. Accompanying the human species in its migrations, the natural selection of spontaneous mutations and, later, the choices of agriculturists have produced genetic variability that is particularly palpable in the diversity of the fruits of certain species. The selection of crop varieties has intensified over the past 100 years, and some people may regard it as a genetic standardization. However, a global comparison of SNP polymorphisms between selected varieties and local wheat or maize populations, for example, shows a decrease of only a few percent. Indeed,

3 Size of a theoretical population that, under certain simplifying assumptions, would show a genetic polymorphism equivalent to that of the natural population.

artificial selection, by applying on numerous punctual targets, combines very large portions of the genomes of the original populations (Cavanagh *et al.* 2013). Yet, analysis of genome sequences reveals the loss of many genes during the domestication process, not just a loss of alleles. Consequently, a concerted global effort to protect, characterize and exchange genetic resources represented by domestic varieties and samples of wild relatives, as well as an intensification of breeding projects using modern genetic methods, will be needed to address environmental constraints and the expected doubling of food demand in the next 30 years (McCouch *et al.* 2013).

Livestock have also followed human migration. The current breeds originate from the selection implemented in the 18th Century, a period when they also began to be dispersed outside their region of origin. Under the same name, it is often a combination of several breeds. The idea of the ancient, sometimes wild origin of certain breeds is an urban legend, maintained by breeders: for example, the Salers cattle breed, with long horns, has nothing to do with the aurochs of the Lascaux rock paintings, because it is in fact related to the other traditional breeds of southern France, which derive from the first cattle herds introduced 10,000 years later!

Molecular studies indicate a relatively recent history and a partial genetic isolation of cattle breeds. Visible and marked differences between breeds are marginal compared to differences between animals of the same breed. Therefore, from a strictly genetic point of view, the disappearance of an animal breed cannot be considered as an irreversible loss of genetic resources for the species. If, for practical and economic reasons, animal breeds characterize production chains, they should not be treated as hermetic selection compartments, since this would lead to genetic impoverishment in the future (Felius *et al.* 2015).

7.7. Important ideas to remember

– The **sequencing** of the genomes of current populations makes it possible to reconstruct their ancestral history through the **genetic traces** found in the different individuals. This applies to the human species as well as to the various natural or domesticated species.

– The sequencing of the **ancient DNA** of populations that have now disappeared completes these reconstructions.

– Traces of recent phenomena (thousands of years) of **natural selection** are recognizable in the genomes of current human populations, the various alleles that result from them may now have beneficial or deleterious effects.

– The domestication of useful plants, animals or micro-organisms consists not only in the cultivation or breeding of wild varieties with interesting traits, but also in the genetic isolation of mutants better adapted to and becoming dependent on artificial ecological niches (cultivation or breeding). All domesticated varieties are therefore **genetically modified**.

– Domestication increases the variety of apparent traits (developmental mutations), while it tends to decrease the overall polymorphism of genomes relatively to natural populations (impoverishment of genetic diversity), hence the need to preserve natural **genetic resources**.

– By correlating genome polymorphism with phenotypic traits selected from reference cohorts, it has become possible to directly select breeding stocks solely on the basis of their genome (**genomic selection**), thereby significantly improving the selection of important farm animals such as cattle.

7.8. References

Barreiro, L.B., Laval, G., Quach, H., Patin, E., Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. *Nature Genetics*, 40(3), 340–345.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., Van Tassell, C.P., Van Raden, P.M., Viaud-Martinez, K.A., Wiggans, G.R. (2012). Design of a bovine low-density SNP array optimized for imputation. *PLOS One*, 7.

Boichard, D., Ducrocq, V., Croiseau, P., Fritz, S. (2016). Genomic selection in domestic animals: Principles, applications and perspectives. *Comptes Rendus Biologies*, 339(7/8), 274–277.

Brinkworth, J.F., Barreiro, L.B. (2014). The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. *Current Opinion in Immunology*, 31, 66–78.

Cavanagh, C.R. *et al.* (2013). Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *PNAS*, 110, 8057–8062.

Choi, J.Y., Platts, A.E., Fuller, D.Q., Hsing, Y.I., Wing, R.A., Purugganan, M.D. (2017). The rice paradox: Multiple origins but single domestication in Asian rice. *Molecular Biology and Evolution*, 34(4), 969–979.

Daetwyler, H.D. *et al*. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, 46, 858–867.

Deschamps, M., Laval, G., Fagny, M., Itan, Y., Abel, L., Casanova, J.L., Patin, E., Quintana-Murci, L. (2016). Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *American Journal of Human Genetics*, 98(1), 5–21.

Doebley, J. (2004). The genetics of maize evolution. *Annual Review of Genetics*, 38, 37–59.

Doebley, J. (2006). Unfallen grains: How ancient farmers turned weeds into crops. *Science*, 312, 1318–1319.

Felius, M., Theunissen, B., Lenstra, J.A. (2015). Conservation of cattle genetic resources: The role of breeds. *Journal of Agricultural Science*, 153, 152–162.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L., Nielsen, R. (2011). Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLOS Genetics*, 7(11).

Green, R.E. *et al*. (2010). A draft sequence of the Neandertal genome. *Science*, 328, 710–722.

Kremer, A. (2016). Microevolution of European temperate oaks in response to environmental changes. *Comptes Rendus Biologies*, 339(7/8), 263–267.

Martinez-Ainsworth, N.E., Tenaillon, M.I. (2016). Superheros and masterminds of plant domestication. *Comptes Rendus Biologies*, 339(7/8), 268–273.

McCouch, S. *et al*. (2013). Agriculture: Feeding the future. *Nature*, 499, 23–24.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157, 1819–1829.

Meyer, M. *et al*. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, 338, 222–226.

Nature (2015). The 1000 Genomes Project Consortium. *Nature*, 526, 68–74.

Neale, D.B., Kremer, A. (2011). Forest tree genomics: Growing resources and applications. *Nature Review Genetics*, 12, 111–122.

Nielsen, R., Akey, J.M., Jakobsson, M., Pritchard, J.K., Tishkoff, S., Willerslev, E. (2017). Tracing the peopling of the world through genomics. *Nature*, 541, 302–310.

Petit, R.J., Hu, F.S., Dick, C.W. (2008). Forests of the past: A window to future changes. *Science*, 320, 1450–1452.

Prüfer, K. *et al*. (2014). The complete genome sequence of a Neanderthal from the Altaï mountains. *Nature*, 505, 216–219.

Quintana-Murci, L. (2016). Genetic and epigenetic variation of human populations: An adaptive tale. *Comptes Rendus Biologies*, 339(7–8), 278–283.

Quintana-Murci, L., Barreiro, L.B. (2010). The role played by natural selection on Mendelian traits in humans. *Annals of the New York Academy of Sciences*, 1214, 1–17.

Reich, D. *et al*. (2010). Genetic history of an archaic hominin group from Denisova cave in Siberia. *Nature*, 468, 1053–1060.

Sankararaman, S., Mallick, S., Patterson, N., Reich, D. (2016). The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Current Biology*, 26, 1241–1247.

Shiu, S.H., Shih, M.C., Li, W.H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiology*, 139, 18–26.

Sironi, M., Clerici, M. (2010). The hygiene hypothesis: An evolutionary perspective. *Microbes and Infection*, 12(6), 421–427.

8

Genetics and Human Health

The link between genetics and human health is as old as genetics itself. As early as 1902, even before the term "gene" was introduced, Garrod understood that alkaptonuria, a metabolic disorder due to the toxic accumulation of homogentisic acid, was a recessively inherited condition. It is now known to be a mutation of a gene on chromosome 3 that directs the synthesis of homogentisate-1,2-dioxygenase, an enzyme of the phenylalanine and tyrosine degradation pathway. Similarly, it had been understood since ancient times that hemophilia was a hereditary problem, affecting mainly men, but transmitted by women (there are several forms of hemophilia, but the two most common ones are the deficiency of blood coagulation factors VIII or IX, both of which are products of genes carried by the X chromosome). However, compared to the major conceptual advances made possible by model systems, such as flies, fungi, bacteria or viruses, humans have only played a secondary role in the progress of genetics for a long time. The criminal wanderings of eugenics until the end of the Second World War are probably responsible in part for the late development of human genetics in the first half of the 20th Century.

It was in the late 1950s that **Victor McKusick** began to develop the first catalog of genetic diseases in humans, carefully listing the phenotypes that could be linked to chromosomal loci. We were very far from knowing the number of genes in the human genome and the importance of the interactions between genes. This catalog, called MIM (Mendelian Inheritance of Man), focused on monogenic traits. In its early days, it had very few genes. It was only in the 1980s, with the cloning and sequencing of

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

genes, that its content was to be expanded, so that when the human genome sequencing project came into discussion, there were about 2,000 loci more or less precisely linked to phenotypic traits, each attributed to a linkage group, but not yet positioned relatively to each other. With the considerable progress made by genomics, the catalog, renamed OMIM (Online MIM), now has about 16,000 described genes, of which about 3,900 are linked to some 6,000 phenotypic traits[1].

8.1. "Mendelian" and multifactorial diseases, a continuum of complexity

To what extent does knowledge of the genome permit us to anticipate diseases – that is, to predict a phenotype from a genotype? The answer cannot be simple, because, as we already glimpsed upon in the introduction, most phenotypic traits depend on the interaction of several genes, the entire list of which is generally not known. This is the missing heritability phenomenon already mentioned. Moreover, in humans, we are in the polymorphic context of a natural population and many alleles have variable penetrance depending on this genetic polymorphism.

The first case of genetic determinism allowing a pre-symptomatic diagnosis was Huntington's chorea, a fatal neurological degeneration due to an autosomal dominant mutation with very high penetrance, identified in 1993. There are other cases that allow a good prediction of the phenotype from the DNA sequence. Cystic fibrosis is caused by the alteration of a gene on chromosome 7 directing the synthesis of a transmembrane transporter of chloride ions. With about 2,000 alleles already described, each associated with a specific phenotype, we are able to interpret the gene sequence in terms of functional prediction, which gives, in this case, excellent meaning to prenatal genetic counseling. But in the majority of cases, functional prediction remains only statistical. For example, it is known that some mutated alleles of the *BRCA1* and *BRCA2* genes are responsible for a significant increase in the incidence of breast and ovarian cancers in women. But other genes are also involved in this phenomenon and our current knowledge is not sufficient to explain all the cases observed. In addition, these alleles have incomplete penetrance. Not all women who have them will necessarily develop these cancers. We are talking about genetic

1 https://www.omim.org.

predisposition. The precautionary advice is to have regular medical follow-ups.

8.2. Interpretation and use of DNA sequences

Interpreting DNA sequences is therefore a difficult problem that depends closely on the state of knowledge acquired. Currently, we are working with cohorts of a few thousand individual genomes. It is clear that the greater the number of individual genomes sequenced, the more precise the interpretations will become. In addition to studies devoted to some specific diseases, projects to sequence entire populations are planned. They should be very useful provided that a reliable system of computerized medical files exists to link genotypes to phenotypes, which is currently not the case in most countries, including France.

Faced with the presence of a new mutation in a sequence compared to all already known variations, the interpretation will be all the simpler as the consequence of this mutation is more drastic. A stop mutation in a coding sequence or a deletion leading to the absence of a protein will generally be easier to interpret than a missense mutation. On the opposite, effects of mutations in intronic or intergenic sequences remain the most difficult to predict. They are nevertheless the most numerous, since these regions occupy much more space in our genome than the coding sequences. As a result, many sequence data remain, for the time being, under-exploited. Similarly, sequence interpretation will be much easier for monogenic diseases than for multifactorial syndromes. Therefore, there remains much to be done in this field.

In the meantime, with the exception of the few alleles known to be directly responsible for well-defined syndromes or particular traits, one relies on an arbitrary classification of the observed sequence variants, ranging from “pathogenic” to “benign”, with the intermediates “probably pathogenic”, “probably benign” or, in most cases, VUS (variant of unknown significance). Pathogenic variants are defined by their association with a morbid trait in more than 90% of cases. In contrast, benign variants are those found associated with a morbid trait in less than 10% of cases. The interpretation of the sequences is not limited to the so-called “punctual” variants (affecting only one or very few nucleotides), the simplest to study. To be complete, it must also take into account the so-called “structural”

variants (deletions, duplications, inversions, translocations, loss of heterozygosity), which can encompass long sequence segments. The interpretation of the latter is often trickier. Some have direct effects, such as intellectual disability syndromes, autism, epilepsy, attention deficit, hyperactivity or schizophrenia, resulting from recurrent deletions or duplications at precise chromosomal loci (1q21.1, 16p11.2, 16p13.11, 15q13.3). But others have indirect effects. A deletion or a loss of heterozygosity may, for example, reveal recessive alleles. For example, the TAR syndrome (thrombocytopenia with absent radius) is associated with a 120 kb deletion of chromosome 1 which does not in itself lead to the syndrome, but can do it depending on which allele is present on the homologous chromosome.

With the multiplication of individual genomic data, some people (in fact many) are discovering themselves to be carriers of genetic deficiencies they did not expect (Chen *et al.* 2012). What should we do with these data? Of course, very few people are able to interpret their own sequence themselves, transferring important responsibilities to the providers of the genetic tests. The American College of Medical Genetics and Genomics (ACMG) has defined a number of genes called "actionable" for which it should be proposed to inform the patient if an abnormal allele is discovered, as these genes are associated with serious conditions such as a predisposition to type 2 diabetes or coronary risk. But what should we do with the person's descendants, if any, or with the other members of his/her family who are genetically related to him/her? Should they be informed if they are not themselves requesting genetic information, with the risk of revealing intimate health data of patients to third parties?

For all these reasons, and because it is still too new a science, the clinical use of DNA sequences remains limited, even though it is crucial information. It is conceivable that the situation will change rapidly and that genetic testing will become routine, even in the absence of specific symptoms. For example, knowing that there is an allele responsible for deep vein thrombosis (by alteration of factor V) in high frequency in the European population, systematic genetic testing would seem appropriate before the prescription of oral contraceptives that increase this risk. The same applies to the therapeutic aspect. Apart from oncology, where remarkable progress has been made using new molecules specific to certain types of causal mutations, the field of pharmacogenomics is currently underdeveloped. One

of the reasons is the need to develop molecules on a case-by-case basis, which makes their cost very high. For example, there is now a drug treatment for cystic fibrosis, but it only works on 5% of alleles. Similarly, there is an inhibitor of the product of *PCSK9* gene one allele of which is responsible for hypercholesterolemia, but its cost is high. As genetics makes it possible to discover the causes of diseases, we realize the extent of their diversity and the question arises of the desirable balance between the specificity of treatments and their possible spectrum of application.

8.3. Autism

The case of autism is of particular interest to human genetics because it is not a single neuropsychiatric disorder, but a set of disorders characterized, to varying degrees, by social communication problems, the presence of restricted interests and stereotypical and repetitive behaviors. We are talking about ASD (Autism Spectrum Disorder). ASDs show a continuum of phenotypic severity and a diversity of neural activity disorders.

After about 40 years of genetic research, most of the genes identified as responsible for ASD now converge towards a limited number of biological processes: chromatin remodeling, synthesis and degradation of proteins, and the synaptic function that ensures communication between different neurons or between neurons and muscle cells.

The first studies that identified the genetic components of ASDs examined how the risk of developing these disorders in an individual varies depending on whether or not similar cases exist in genetically related individuals. The results are clear. As expected for genetic components, this risk increases as the degree of kinship increases, from cousins, half-siblings, siblings or dizygotic twins, to monozygotic twins. For the latter, the risk is about 45%, giving us an idea of the overall penetrance of the alleles involved.

More recently, studies have focused on very large populations of several million people. The first studies showed chromosomal abnormalities and variations in gene copy numbers in 4 to 10% of ASD cases. Further studies confirmed the role of mutations in genes involved in the formation and function of synapses, such as *NLGN3*, *NLGN4*, *SHANK3*, etc. Finally, the complete sequencing of the genomes of thousands of families in which some

members have ASDs has revealed the existence of *de novo* mutations (not inherited from parents) preferentially involving genes involved in chromatin remodeling or genes deregulated in the fragile X syndrome (a rare genetic disease causing cognitive impairment).

To these genes, whose mutations can have a strong phenotypic effect, it is necessary to add the possible influence of a large number of genetic variations with low effects. Considering only one thousand common alleles (those whose presence in a population equals at least 5%) allows us to estimate that their joint supply from both parents accounts for 40% to 60% of the observed cases of ASD. We are therefore dealing with a complex genetic determinism. Cases of ASD can emerge from a single *de novo* mutation in a particular genetic background as well as from the random assembly of many parental alleles. This genetic architecture explains the variability of phenotypes: the same rare variant or *de novo* mutation may be harmful in one individual and without visible effect in another.

Much progress remains to be made in our understanding of the mechanisms involved in ASD. They will require quantitative measurements of gene activity, neural circuits and patient behavior. The continuum of severity of ASD obliges us to consider the general population in its genetic and cultural diversity as the normal control in these studies. This can only be achieved on a global scale with reliable data sharing given the required size of cohorts. Not forgetting that this research must be done with the best possible involvement of the patients and their families in order to improve, as much as possible, their quality of life and their integration into the activities of society, in the same way as exists for sensory or motor disabilities.

8.4. Gene therapy

The idea of repairing a genetic defect responsible for a disease began to emerge in the 1940s and 1960s as soon as it became clear that the genetic material consisted of DNA and the genetic code was deciphered. Yet, while we knew how to transform a few bacteria – hence the identification of the role of DNA (see Chapter 1) – the tools were far from being available for eukaryotic cells and, moreover, for targeted applications. The beginnings of recombinant DNA and gene cloning methods in the 1970s (see Chapter 4) rekindled this hope, but still did not provide controlled tools for the

transformation of eukaryotic cells. Genetic engineering therefore turned first to the production of therapeutic proteins rather than to gene therapy. Human insulin, for example, was produced by genetic engineering from 1978 onwards, followed by other proteins for therapeutic or vaccinal usage. Moreover, considering gene therapy requires the precise knowledge of the genetic cause of a disease at the molecular level, which was only exceptionally the case at the time (Friedmann and Roblin 1972).

To become a reality, gene therapy needed methods to deliver DNA into human cells capable of proliferating and to integrate it into their chromosomes. Yet, the transformation of eukaryotic cells (plants, yeasts, mammals, insects, etc.) only began to develop in the early 1980s and, except in the case of *Saccharomyces* yeast, the transforming DNA was randomly integrated into the chromosomes, rarely reaching the targeted gene. Transgenesis was achieved by adding the transgene (the chosen DNA) somewhere in the genome, with the risk of disrupting normal genes at or near the point of insertion into the chromosome. The idea quickly developed of using modified viruses as transgenesis vectors to facilitate the integration into chromosomes, but without solving the targeting problem. Moreover, some allowed the maintenance of the transgene in the cell nuclei as an episome, that is, without integration into a chromosome. Currently, gene therapy mainly uses vectors containing elements derived from lentivirus (LV) or adenovirus (AAV). Obviously, these are only fragments of viral genomes, unable to generate infection, because essential genes to the viral cycle are missing.

The first applications of gene therapy concerned hematopoietic stem cells, easily extracted from the organism, transformed by DNA *ex vivo* and reinjected into the patient where their proliferative capacity generates the different types of blood cells, some of which having a very long lifespan. For it is clear that gene therapy applies to diseases whose curation depends on long-lived cells in the body. Genetically correcting short-lived cells (a few days) would require the constant repetitions of long and costly protocols. Initially, about 20 patients with severe combined immunodeficiency X1 (SCID X1) were treated with gene therapy. The treatment proved effective, with the genetically corrected cells having an estimated lifespan of more than 50 years, most treated patients were able to return to a normal life during the last 20 years. However, unfortunately, six of them developed leukemia, because the vector used carried a retroviral enhancer that

accidentally activated oncogenes by integrating in their vicinity on the chromosomes. A new generation of vectors (called "auto-inactivated") was subsequently developed and about 100 patients with SCID X1 or deficient in adenosine deaminase (ADA) have now been successfully treated worldwide.

Other severe blood diseases, such as Fanconi anemia, beta-thalassemia and sickle cell anemia, are being treated with promising results. The treatment of liver cells with intravenously injected AAV8 vectors yields good results in hemophilia B (factor IX deficiency). A gene therapy strategy by sub-retinal inoculation of recombinant DNA is now available for the treatment of a mutation in the *RPE65* gene, responsible for a significant fraction of cases of Leber's congenital amaurosis, a severe dystrophy leading to blindness at an early age because the synthesis cycle of retinal pigments is altered. In this treatment, the recombinant DNA brings the normal allele of the *RPE65* gene to retinal cells, maintaining itself in their nuclei as an episome, and its expression restores the normal synthetic cycle of retinal pigments. Other therapeutic hopes concern myopathies and diseases of the nervous system. Gene therapy has therefore become a reality, but so far limited to a reduced number of situations and patients. Its extension should benefit from recent progress in gene editing (see Chapter 6) but also necessitates a reduction of costs, which are currently very high. In general, four types of gene therapy can be considered:

– adding a normal gene to compensate for a deficient (inactive) allele;

– inhibiting an expressed gene carrying a harmful mutation;

– repairing a gene carrying a harmful mutation;

– introducing a new function with therapeutic effect.

It is in this last category that remarkable recent advances have been made in cancer therapy by the genetic reprogramming of T lymphocytes. The original mechanisms ensuring the functions of these cells have been elucidated by numerous previous studies that have already made drug therapy possible. But recombinant DNA now makes it possible to modify these lymphocytes in a permanent way to make them act on targets of interest, such as the surface proteins of dividing cancer cells, and thus eliminate these cells. By genetically modifying these lymphocytes, they can be transformed into universal "drugs" no longer recognized by the recipient's immune system, such that the use of the patient's own

lymphocytes are no longer required. Leukemias have recently been successfully treated with this therapy, which also holds significant promises in the complex field of autoimmune diseases.

8.5. The multiple genetic causes of cancers

There is a wide diversity of cancers, depending on the primary tissues affected, the types of cancer cells and their ability to proliferate and spread throughout the body. While some cancers can be induced by infectious agents (hepatitis B virus for liver cancer, papillomavirus for cervical cancer, *Helicobacter pylori* or *Fusobacterium* for gastric or colorectal cancers), the majority results from genetic alterations in somatic cells. These mutations can be spontaneous or induced by various physical or chemical agents acting directly or indirectly on DNA. Their probability of occurrence over time and, to some extent, their molecular nature depends on each individual's genetic background. For example, the alteration of genes involved in DNA repair predisposes to the development of certain cancers early in life. But, more than the high-penetrance alleles of a few particular genes responsible for a significant proportion of family cases, it is the combination of low and medium-penetrance alleles of many genes that is the main cause of later onset cancers.

The discovery of genetic factors responsible for the development of cancers is more than four decades old. By transforming contact-inhibited *in vitro* cultures of murine or human cells with random fragments of cognate genomic DNA, foci of cell proliferation were identified that correspond to the presence of certain genes or sequences in the DNA fragments. The genes responsible for the malignant transformations were called **oncogenes** and several dozen of them were quickly identified in the human genomes or those of other vertebrates. These genes were often conserved among vertebrates, performing important functions in the regulation of cell cycles and divisions (e.g. genes producing kinases) or in normal developmental processes. Since their normal function was obviously not the production of tumors, these genes were renamed **proto-oncogenes** to indicate that it is by mutation that the normal proto-oncogene becomes an oncogene capable of triggering cancer. The discovery of proto-oncogenes was an important conceptual advance in our understanding of cancers, as it demonstrated that a specific and limited genetic change was capable of disrupting the entire complex machinery of cellular regulations (Logan and Cairns 1982). There were therefore critical points in this machinery that we could hopefully

identify. It was then discovered that, in addition to proto-oncogenes, genomes also contained tumor suppressor genes, named **onco-suppressors**, whose inactivation by mutation promoted the development of cancers. The most classic of these is the *TP53* gene, whose product controls the expression of many other genes through interaction with different signaling pathways. It has been known since the late 1980s that this gene is very frequently mutated in a wide variety of cancers.

There are two main types of mutations that convert proto-oncogenes to oncogenes: those that alter gene products, by modifying the amino acids of a protein or the sequences of non-coding RNAs, and those that modify gene expression without altering its products. In the latter case, it may be the overexpression of a gene normally expressed at a low level or the activation of a gene normally silent in normal differentiated cells. For example, the level of expression of the *IGF2* gene is 300 times higher than normal in colon cancer cells. Changes in gene expression can result from mutations in DNA sequences that control gene expression (promoter). But more often than not, it is the consequence of chromosomal rearrangements (translocations, inversions, deletions) that accidentally place a proto-oncogene near activation elements of other genes (we speak of diversion of activation sequences or enhancer highjacking) or fuse it to highly expressed genes. Note the similarity with the accidental activation of proto-oncogenes by insertion of viral genomes in chromosomes that was at the origin of their discovery, the viruses bringing strong promoters or activation elements in the vicinity of the proto-oncogene.

The idea that chromosomal rearrangements may be at the origin of cancers is an old one, since T. Boveri, one of the promoters of the chromosomal theory of heredity, had already contemplated it at the beginning of the 20th Century. Low-resolution cytological methods supported this hypothesis, since specific translocations were observed in the chromosomes of some cancer cells, as for example in the case of chronic myeloid leukemias (it is now known that these translocations lead to the formation of a *BCR-ABL* gene fusion at the junction point). But for a long time, the dominant idea remained that of the accumulation of point mutations in somatic clones, which led normal cells to a cancerous status through multiple precancerous states.

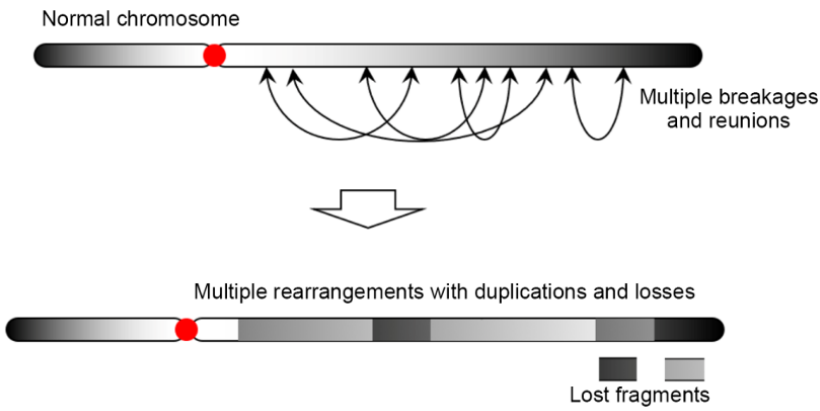
In this intellectual scheme, a distinction must be made between the main genes (drivers), whose mutations have a strong effect (such as *TP53*), and

the secondary genes (passengers), whose mutations are insufficient alone to trigger tumor transformation, but become so in the presence of pre-existing mutations in a strong gene or in other secondary genes. Mutations in the driver genes confer oncogenic properties to cells such as accelerated growth, invasive and metastatic capacity, stimulation of angiogenesis and escape to apoptotic mechanisms.

The advent of genomic methods for cancer characterization significantly clarified these concepts (Cowin *et al*. 2010). An international program (International Cancer Genome Consortium or ICGC) was set up in 2007 to study 50 types of cancers by sequencing 25,000 genomes. The first comparisons between genomes of cancerous and healthy cells from the same patient showed recurrent genetic alterations (probably highly causal) and a wide variety of occasional genetic alterations (possibly secondary). A catalog of mutations (COSMIC, for Catalogue Of Somatic Mutations In Cancer) was created in 2004 with the idea of listing all the mutations found in genomes of cancer cells in order to extract a list of genes (CGC, for Cancer Gene Census) whose mutations are primary causes of cancers. This list now includes 719 genes (Sondka *et al.* 2018).

However, genome sequences of cancer cells also confirmed that chromosomal rearrangements were not uncommon. Among them is a new and remarkable phenomenon, called "chromothripsis", discovered a few years ago in many cases of juvenile medulloblastoma or chronic lymphocytic leukemia (Korbel and Campbell 2013). Unlike conventional chromosomal rearrangements, chromothripsis corresponds to multiple rearrangements of fragments within a single chromosome arm. It only very rarely affects two chromosomes (see Box 8.1). The formation of these rearrangements involves several tens or hundreds of DNA breakage-repair events during which some fragments are lost and others duplicated. The absence of intermediate forms suggests a catastrophic phase during which all events occur simultaneously, probably during a single cell cycle. Similar structures have been found in other types of cancers. They are particularly common in bone cancer. The molecular mechanisms of chromothripsis remain to be elucidated, although experimental systems using *in vitro* cultivated cells have been developed. It should be noted that multiple chromosomal rearrangements of the chromothripsis type also exist in the plant world (Carbonell-Begerano *et al.* 2017). The fact that chromothripsis is only observed in patients who have inherited mutations in the *TP53* gene clearly shows an innate genetic susceptibility upon which somatic genetic

accidents are superimposed. According to ICGC data, a similar situation is found in more than a third of the cancers analyzed, which argues in favor of *a priori* genomic sequencing of every individual in order to offer appropriate preventive medical follow-ups to the people who have inherited mutated alleles of the main genes.



Multiple and **simultaneous** events of fragmentation and splicing of chromosome segments are observed in some cancers or in plant mutants. The molecular mechanisms responsible for these catastrophic events remain to be determined. Micronuclei, formed by abnormally distributed chromosomes in daughter cells after abnormal mitotic divisions, could be involved, since chromothripsis generally affects only one chromosome arm at a time, leaving intact the other chromosomes of the same cell.

Chromothripsis results in the formation of **mutated chromosomes** showing a very large number of entangled rearrangements with **loss** of some segments and **duplications** of others.

The diagram shows the general principle of chromothripsis. The centromere is schematized by the red circle joining the two chromosome arms (gray cylinders). Rearrangements are, in general, much more numerous than shown here.

Box 8.1. *Chromothripsis*

8.6. Microbiota

While some living organisms are able to live and proliferate in laboratories in the complete absence of other organisms (axenic

environments), they rarely do so in the natural environment. Animals, plants, algae or fungi are in contact with many microorganisms with which they have complex relationships, ranging from mandatory or optional symbiosis to negative interactions. Apart from major pathogens, the diversity of these microorganisms and their importance have remained poorly understood until recent advances in DNA sequencing (see Chapter 4) paved the way for direct metagenomic analysis of these microbial populations. These populations, now referred to as "microbiota", include bacteria, archaea, viruses and various unicellular eukaryotes.

In the human species, different microbiota live in contact with the skin, in the mouth, intestine, vagina, etc. With 10^{13} microbial cells in an adult – of the same order of magnitude as the total number of cells in the human body – the intestinal microbiota is the most abundant. It can reach two kilograms. Microbiota are formed from birth by contact with environmental microorganisms – including the mother's vaginal microbiota – and evolve with food diversification, hygiene, antibiotic treatments, etc. The genetic background of the host is also involved in the evolution of microbiota, although the precise mechanisms are not yet known.

The role of the intestinal microbiota in the normal functions and dysfunctions of the digestive system had been known for a long time, but its composition has remained largely unknown since most of the microorganisms it contains are not cultivable *in vitro*. Similarly, its role in the immune system was not anticipated. In addition, there was no clear explanation for the increasing incidence of chronic immunological diseases (type 1 diabetes, asthma, Crohn's disease, multiple sclerosis, etc.) in developed countries in recent decades. Changes in lifestyle (improved hygiene, especially perinatal hygiene), nutrition (presence of additives) and exposure to various environmental substances inexistent in the living world (xenobiotics) were discussed (Bach 2002). The hypothesis that these changes could act through their effects on the microbiota was put forward in 2006 (Blaser 2014). The first results of meta-genomic analyses of the intestinal microbiota supported this hypothesis, showing a loss of diversity in the microbial populations associated with certain symptoms.

We know that our intestinal microbiota contributes to our digestion by completing the assimilation of nutrients with a set of enzymes that we do not synthesize ourselves. These include enzymes that hydrolyze polysaccharides or ferment certain substrates, but also those that synthesize substances such

as vitamins K, B12 and B8, thereby enhancing their content in the diet. Animals kept in an axenic state in the laboratory have energy requirements 20–30% higher than control animals. We also know that our intestinal microbiota contributes to our immunity. Competition from commensal species reduces the presence of pathogenic species by producing bactericidal substances (bacteriocins), while our immune system regulates the composition of the microbiota and adapts to its variations. Axenic mice have an immature and incomplete immune system compared to mice raised normally.

With the new very high throughput sequencing methods, a large number of sequences are obtained from purified DNA from stool samples (for the intestinal microbiota) or from oral or vaginal samples for the corresponding microbiota. These sequences are compared with each other and with specialized databases using appropriate computer procedures. From these comparisons, we can deduce the presence in the DNA mixture of sequences of organisms already known or not and thus have a first qualitative description of the microbial flora. In addition, the relative abundance of the different organisms can be deduced from the number of times sequences belonging to their genome are found, relatively to their sizes. Observed differences are of several orders of magnitude between abundant and rare organisms. Finally, the stoichiometric presence of sequences that can be assembled provides a means of reconstituting whole genomes from previously unknown organisms. Comparing their sequences with the databases often suggest to which large evolutionary group they belong.

A first catalog of genes of the human intestinal microbiota was established in 2010 from 129 European adults (Li *et al.* 2014). It had more than three million genes belonging to more than a thousand distinct microbial species. Four years later, the catalog was extended to nearly 10 million genes by analyzing the microbiota of ten times as many individuals from three continents (Americas, Asia and Europe). Equivalent gene catalogs have been established for intestinal microbiota in pigs and mice. By taking into account the most abundant microbial species in each individual, it is possible to define **enterotypes**[2] and study their variation according to diet or pathological status (Lynch and Pedersen 2016). For example, patients with liver cirrhosis show a 25% decrease in intestinal

2 An international program defines standards for the analysis of human ***microbiomes**** to enable comparisons between samples from different sources.

microbial diversity, while about 30 species, rare in healthy individuals, may represent up to 40% of their total microbiota (Qin *et al.* 2014). These include bile-sensitive oral bacteria invading the intestine in liver cirrhosis (because bile production is reduced) and producing toxins such as those involved in hepatic encephalopathy, one of the complications of liver cirrhosis. Similarly, correlations have been established between enterotypes and metabolic disorders, leading to obesity without all causal relationships being clearly established (Le Chatelier *et al.* 2013). It should be stressed here that the microbiota analyzed so far mainly concerns the intestinal lumen and that many of the species interacting with the mucosa are ignored.

In the space of a few years, the composition of the intestinal microbiota has become a diagnostic tool. It has the great advantage over other methods of being totally non-invasive. Today, we can determine the stage of a liver cirrhosis without biopsy. Similarly, the observed correlations between the relative abundance of pro- or anti-inflammatory microbial species and the susceptibility to type 2 diabetes, liver and vascular complications or certain types of cancer are interesting to consider. Cases of acute inflammatory diseases such as Crohn's disease or hemorrhagic rectocolitis are correlated with an imbalance in the intestinal bacterial ecosystem (dysbiosis) with the predominance of certain bacterial families (*Enterobacteria*, *Fusobacteria*), and the rarefaction of others (*Clostridia*, *Faecalibacterium*). This suggests curative possibilities by controlling the microbiota of patients through the ingestion of probiotics (favorable microbial species), molecules that promote the development of beneficial microbial species, or even by the direct transplantation in intestines of healthy microbiota.

Finally, the presence of molecules of microbial origin in the intestine is not without influence on the whole body and particularly on the brain. In addition to the nerve signals sent to the brain by the intestine, the presence in the microbiota of neuroactive substances entering the bloodstream makes the intestine a **second brain**. The fact that antibiotic treatments can, against all logic, reduce autistic syndromes or, on the contrary, increase the severity of breast tumors in young women argues for active research in this area.

8.7. Important ideas to remember

– Identifying the **genetic determinants** of a given phenotypic trait in humans consists of searching for statistically significant correlations

between **variant alleles** (identified by genome sequence) and trait expression in well-chosen **cohorts**. The quality of the result depends on the number of genetic elements involved in the trait studied, the penetrance of the different alleles and the genetic polymorphism existing in the populations examined (nature and frequency of alleles).

– For many traits, the **lack of heritability**, combined with genetic polymorphism, leads to a statistical interpretation of the genotype established from the sequence of a genome.

– **Treatments now exist** for some genetic diseases by injecting recombinant DNA or genetically reprogrammed cells.

– Several hundred genes in the human genome are capable of inducing cancers when mutated or abnormally activated. Cancers are not transmissible to offspring, but genetic predisposition factors are.

– **Metagenomics** gives us complete and quantitative descriptions of the populations of microorganisms with which we live in harmony (or sometimes disharmony), including many species that had never been isolated before.

8.8. References

Bach, J.-F. (2002). The effect of infections on susceptibility to autoimmune and allergic diseases. *The New England Journal of Medicine*, 347, 911–920.

Blaser, M.J. (2014). The microbiome revolution. *Journal of Clinical Investigation*, 124, 4162–4165.

Carbonell-Begerano, P., Royo, C., Torres-Pérez, R., Grimplet, J., Fernandez, L., Franco-Zorrilla, J.M., Lijavetzky, D., Baroja, E., Martínez, J., García-Escudero, E., Ibáñez, J., Martínez-Zapater, J.M. (2017). Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiology*, 175, 786–801.

Chen, R. *et al*. (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 148, 1293–1307.

Cowin, P.A., Anglesio, M., Etemadmoghadam, D., Bowtell, D.D. (2010). Profiling the cancer genome. *Annual Review of Genomics and Human Genetics*, 11, 133–159.

Friedmann, T., Roblin, R. (1972). Gene therapy for human genetic disease. *Science*, 175, 949–955.

Korbel, J.O., Campbell, P.J. (2013). Criteria for inference of chromothripsis in cancer genomes. *Cell*, 152, 1226–1236.

Le Chatelier, E. *et al*. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500, 541–546.

Li, J. *et al*. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32, 834–841.

Logan, J., Cairns, J. (1982). The secrets of cancer. *Nature*, 300, 104–105.

Lynch, S.V., Pedersen, O. (2016). The human intestinal microbiome in health and disease. *The New England Journal of Medicine*, 375, 2369–2379.

Qin, N. *et al*. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513, 59–64.

Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I., Forbes, S.A. (2018). The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18, 696–705.

9

Now and Tomorrow

At this point, readers could be tempted to believe that they have already read everything about the complexity of the gene and its multiple facets in the living world and, as a result, that genetics can now turn resolutely towards its applications without risking new surprises. This reasoning is very naive. It forgets that, as modern genomics shows, only a small part of the true genetic diversity of the living world has been explored. Bacteria, opisthokontes (animals and fungi) and (plants and green algae) *viridiplantae* have largely dominated research and continue to do so, leaving aside almost all other branches of the tree of life (see Introduction, Box I.1). This reasoning also forgets that current organisms are the result of a contingent history, that of terrestrial evolution, and that, whatever their diversity, they will always remain but a tiny bit of the realm of possibilities, as illustrated in part by paleontology and now **xenobiology**. Finally, it forgets that, as Bateson pointed out as early as 1911, "genetics gives the human species a power that could never be predicted and which is extremely dangerous", and therefore confers on us a responsibility commensurate with this power, in which the evaluation of artificial risks only makes sense in comparison with those of natural phenomena.

9.1. A living world to be further explored

A first example of the surprises we encounter in exploring the diversity of the living world is given by the mitochondrial genomes of Diplonemids. Very few people, including biologists, have ever heard of these

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

monoflagellated[1] unicellular eukaryotes. Yet they are among the most abundant organisms on Earth, accounting alone for one seventh of marine eukaryotic micro-plankton (De Vargas *et al*. 2015). In these organisms, the mitochondrial genome does not consist of DNA molecules carrying identifiable genes as in ourselves and all animals, fungi or plants (see Chapter 3). And yet the products of such genes (enzymes of the respiratory chain and oxidative phosphorylation, RNA molecules) are well present in the mitochondria of Diplonemids. How are they synthesized without genes?

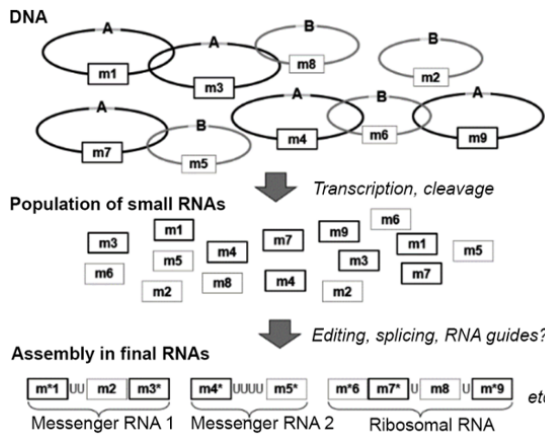
Recent work has revealed the existence of a very original mechanism (Marande and Burger 2007) that produces messenger RNA and functional ribosomal RNA molecules, similar to those of mitochondria of other living species, by the ordered assembly of "modules". One finds up to 11 modules at the origin of the same RNA molecule. These modules are themselves small RNAs formed by conventional sequence fragmentation and editing processes (see Chapter 2) from longer RNA molecules that are the transcripts of the small DNA circles present in the mitochondria of Diplonemids. These circles do not show any recognizable mitochondrial genes in their sequences. In *Diplonema papillatum*, there are about 100 copies per cell, divided into two categories (6 and 7 kb). Within each category, the sequences of the different molecules are identical to each other, with the exception of a short segment of a few hundred nucleotides. It is within these short segments that the genetic information is found which, after complex maturation steps of the RNA transcripts, allows the formation of modules whose final assembly produces the functional RNAs (see Box 9.1).

This system may seem unnecessarily complicated, totally non-economical and unreliable. How do we ensure that the small DNA circles are properly transmitted to the offspring? It is, however, an evolutionary success, judging from the abundance of these organisms. How is the correct assembling of RNA modules guided? There are never any aberrant assemblages[2]. And finally, why such a genetic system? We don't have the answer. Perhaps it is the result of a late evolutionary change that fragmented existing genes by the massive insertion of mobile elements without deleterious effect because of

1 Diplonemids belong to the large group of eukaryotes called Excavates, alongside Euglenas and Trypanosomes, which are better known to the public, because some are pathogens to humans and animals in tropical areas (Chagas disease, sleeping sickness, etc.).

2 The precise mechanisms of molecular guidance are not yet known in this system, but they are probably not different in nature from those of intron splicing or precise editing of RNA sequences.

the existence of the RNA maturation machinery? But perhaps it is a remnant image of what the first genes may have been before the invention of DNA made it possible to transmit them already assembled to the progeny? And if the first geneticists had studied Diplonemids before the other organisms, wouldn't the question be why other genes are not fragmented in the way we have just seen? This example is sufficient to illustrate how incomplete our knowledge is.



In **Diplonemids** – nicellular marine eukaryotes belonging to the large eukaryotic group of Excavata – **mitochondrial DNA** is replaced by about 100 small DNA circles belonging to two different families, A and B. Within each family, the sequences are identical except for a very short variable fragment called a **module** (m_i boxes). No mitochondrial gene is recognizable and yet their products are present. These DNAs are transcribed into RNA molecules which, like any RNAs, will undergo fragmentation and multiple modifications. RNA modifications include **cleavages** and sequence **editings** (addition of U, deamination of C or A, marked by stars). Once mature, these small RNAs are **assembled** in a precise order to give the messenger RNAs and ribosomal RNAs necessary for mitochondrial protein synthesis. It is not clear how the assembly order is determined, but there are no outliers. These could be chemical reactions similar to the trans-splicing of introns, or involving small **guide RNAs (anti-sense)** as in the case of some sequence editing reactions. The final RNAs produced are indistinguishable from what would have been produced by the transcription of classical genes.

Box 9.1. *Cryptic genes*

Still in the marine domain, spectacular advances in metagenomics in recent years have made it possible to explore the "global microbiome" of the oceans by sequencing the DNA contained in water samples systematically taken from different depths at different points around the world. During the "Tara Oceans" expedition, the samples were filtered by increasing size, in order to separate the fractions corresponding approximately to viruses, bacteria and archaea, unicellular eukaryotes and, finally, very small multicellular eukaryotes. The total DNA of each of these fractions, once purified, was then sequenced (Sunagawa *et al*. 2015). The first results revealed an impressive wealth of new sequences. More than 80% of the sequences had never been encountered before when compared to current databases, which yet contain several hundred thousand different genomes. This is another example illustrating the amount of what remains to be discovered about the living world. This knowledge deficit, illustrated by the metagenomics of the oceans, is particularly significant for the world of viruses, whose actual diversity is probably 100 times greater than what is exemplified by the viruses already described (Brum *et al*. 2015). Albeit in a more limited proportion, the same phenomenon exists for bacteria and archaea, in which twice as many clades as previously known could be deduced from ribosomal RNA sequences (Yilmaz *et al.* 2016). Finally, a catalog of 116 million eukaryotic genes was established by combining metagenomic results with RNA sequencing (Carradec *et al.* 2018). Half of these genes are entirely new. They do not correspond to any gene families already known.

Considering their abundance and mode of multiplication, it is clear that viruses play the crucial role in the microbial populations that make up plankton. They influence their composition, size and evolution with a considerable geochemical impact, in which almost everything remains to be discovered (Roux *et al*. 2016). It was among them that another surprise awaited us. The viruses on which molecular genetics has primarily relied have small genomes (a few tens of thousands of nucleotides, not to mention those whose genomes are RNA molecules). In this world, a virus like Epstein-Barr's (EBV, a member of the herpes virus family associated with various cancers) appears as a giant with a genome of 172 kb. However, it is dwarfed in comparison to the virus found in an amoeba in 2003 (La Scola *et al.* 2003). The latter, called *Mimivirus*, with a DNA of 1.2 Mb carrying a thousand genes capable of encoding proteins (i.e. similar in complexity to that of a bacterium) raised the question of the boundary between the cellular world and that of viruses. With a particle diameter of 0.7 µm, this virus is

retained by conventional filters with a pore size of 0.2–0.3 μm. Since the discovery of the tobacco mosaic virus by **Dimitri Ivanoski** in 1892, this limit has served as the operational basis for the definition of viruses.

With the successive discoveries of other viruses of this type, many of them in marine samples, it has become clear that these non-filterable so-called "giant" viruses (or giruses) are far from being a simple curiosity. They are probably ubiquitous and show a great diversity of shapes, genome sizes and gene contents. Today, there are several known families (*Mimiviridae*, *Pandoraviridae*, *Pithoviridae,* etc.), whose genetic organizations differ. They infect unicellular eukaryotes belonging to very diverse evolutionary lineages, which could suggest that they already existed before the separation of such lineages. In addition to *Mimiviridae*, whose virions are icosahedral as in some other virus families, there are amphora-shaped virions up to 1.2 μm or 1.5 μm long in *Pandoraviridae* and *Pithoviridae,* respectively, and spherical virions 0.6 μm in diameter in *Mollivirus*. *Pandoraviruses*, initially discovered in amoebae on the Chilean and Australian coasts (Philippe *et al.* 2013), have genomes that can reach 2.8 Mb of DNA and contain 2,500 genes capable of encoding proteins, 90% of which have no similarity with what was previously known. *Pithoviruses*, true "living fossils", have been discovered in Siberian permafrost. Although 30,000 years old, they are still able to infect and destroy their host amoebas.

Why are these elements new viruses and not degenerated intracellular parasites? This issue has been the subject of heated debates, but their viral nature is now well established. First, their genomes show no trace of genes for a protein synthesis machinery (ribosomal proteins, ribosomal RNAs). They are therefore totally dependent on their hosts for the synthesis of their own proteins, which is the main signature of viruses. Then, because their molecular phylogeny (based on the rare genes they share with the rest of the tree of life, such as those of DNA polymerases) place them back to the very origin of eukaryotes. And, at that distant time, the different virus families were already differentiated from each other if we judge by the very small number of genes they share (about 20). They look therefore like actual living fossils that may one day inform us about what has happened in the early stages of cellular life.

Returning to the present world, another phenomenon poorly known by the public deserves reflection: that of secondary endosymbioses in

eukaryotes. They are called secondary with respect to the hypothesis of so-called "primary" endosymbioses, which are supposed to be at the origin of mitochondria and chloroplasts (see Chapter 6). But while these hypothetical primary endosymbioses are based on indirect and difficult to prove arguments, the secondary endosymbioses are factual in nature. Two organisms best illustrate this phenomenon: *Guillardia theta*, a cryptomonad related to the Chromalveolata group, and *Bigellowiella natans*, a chlorarachniophyte belonging to the Rhizaria group. Both are unicellular eukaryotes (with two flagellas for the first and one for the second), which, contrary to what their phylogenetic positions predict, have functional chloroplasts with plastid genomes similar to those of red (for *G. theta*) or green (for *B. natans*) algae (Curtis *et al*. 2012). It was previously noted that these chloroplasts were surrounded by four membranes instead of the traditional double membrane and that these organisms each had, in addition to the nucleus, a nucleomorph containing three small chromosomes. The genomic analysis of these two organisms has now shown that these small chromosomes are the remains of the nuclei of a red (for *G. theta*) or a green (for *B. natans*) alga, while many of the nuclear genes of these algae (especially those important for chloroplast function) have now been transferred to the respective host nuclei.

From these observations, it is logical to infer that these organisms are the result of sufficiently recent endosymbioses to leave traces of the former nuclei of the phagocyted algae (now the nucleomorphs), while their functional chloroplasts were maintained, conferring a considerable advantage to the hosts, as they became autotrophic. Other similar cases exist, where the nucleomorph has disappeared, suggesting older events of the same type of endosymbiosis.

This phenomenon of secondary endosymbiosis seems much more general in the eukaryotic world than generally imagined. It is thought that all Chromalveolata, the large group of eukaryotes that includes brown algae, diatoms and oomycetes, but also ciliates (such as paramecium) and apicomplexes (such as *Plasmodium*, the malaria agent) which do not have functional chloroplasts, are the result of several successive endosymbioses of green and red algae (Dorrell and Smith 2011). From the non-photosynthetic ancestor of this group, a first endosymbiosis of green algae would then have been replaced by a second endosymbiosis of red algae, the first event leaving many green algae genes in the nucleus of Chromalveolates, the second leaving chloroplasts that have persisted in some lines (Stramenopiles) but

have been lost in others (Alveolates). Traces of a non-functional chloroplast (called "apicoplast") can be found in *Plasmodium*. There are also spectacular cases of functional chloroplast endosymbioses in the animal world, such as in the photosynthetic sea slugs[3], but, unlike previous cases, these are only temporary associations that must be renewed with each generation and not heritable phenomena leading to new evolutionary lineages. Nevertheless, these temporary associations are favored places for gene transfers that, if reaching the germ lines, lead to their heritability (see Chapter 6).

9.2. Genome synthesis

Chapter 4 mentioned the beginnings of recombinant DNA technologies and the subsequent acceleration of applications with, for example, the production of human growth hormone by bacteria (avoiding the problems of contaminated human blood) or synthetic vaccines (eliminating any risk of infection). But it was then a matter of cutting and pasting existing DNA molecules, not of chemically synthesizing new genes as we are able to do routinely today.

The chemical synthesis of DNA required a long period of research and optimization before it could be effective. The first synthetic gene was that of a yeast transfer RNA published in 1972 (Khorana *et al.* 1972). It was a DNA fragment of only 77 pairs of nucleotides, but it was a feat that had taken more than five years of work in H.G. Khorana's laboratory, one of the decipherers of the genetic code. Actually, it was not even an entirely chemical gene synthesis, as it used an enzyme (DNA ligase) to assemble the 15 chemically synthesized oligodeoxyribonucleotides, each consisting of only 5 to 20 nucleotides since the chemical synthesis of DNA remained so ineffective[4]. A few years later, an artificial gene was synthesized allowing *E.*

3 In this case, the chloroplasts of the algae on which slugs feed are maintained in a functional state in the animal's digestive cells, providing them with a light photosynthetic capacity throughout their lives, but not transmissible to their offsprings.

4 The amount of polymer produced decreases exponentially with size, considering the efficiency of coupling of each of its elements. For example, an efficiency of 80% (traditionally obtained with the phosphotriester chemical method), leaves only 10% of product of 10 nucleotides long, 1% of product of 20 nucleotides long, etc. This problem has long limited the size of the synthetic polydeoxyribonucleotides, and it was not until the arrival of a new chemistry with a better coupling efficiency (based on phosphoramidites) that longer molecules could be obtained. Today, polymers are routinely synthesized of a length from

coli to produce the human somatostatin (Itakura *et al.* 1977). It was still a DNA fragment of only about 50 pairs of nucleotides, obtained by *in vitro* assembling, with the help of a DNA ligase, eight oligodeoxyribonucleotides whose sequences were complementary to each other and partially overlapping.

As the chemical synthesis of DNA and its automation[5] progressed in the 1980s, synthetic oligonucleotides gradually entered the laboratories, but they remained rare and expensive and they required delicate purifications. Initially, they were used for directed mutagenesis, recombinant DNA constructs *in vitro* and, increasingly, as primers for the amplification of DNA fragments by polymerase chain reaction (PCR) *in vitro.* The latter technique has considerably accelerated the synthesis of recombinant genes. But it was an essentially enzymatic synthesis, made from fragments of natural DNA used as a matrix. However, by 1981, an artificial gene with more than 500 nucleotides, allowing a bacterium to produce the human interferon, had been fully, chemically synthesized (Edge *et al.* 1981). But it was not until the 1990s and the increase in DNA synthesis efficiency that the cost of short oligonucleotides (about 20 nucleotides) fell, and long oligonucleotides (about 100 nucleotides) appeared. The synthesis of DNA molecules of a length corresponding to that of average genes (kilobases) then became possible thanks to the combination of partial overlaps between long oligonucleotides and their elongation *in vitro*. But there were still two to three orders of magnitude missing to reach the size of bacterial genomes or eukaryotic chromosomes.

It is thanks to the yeast *S. cerevisiae* and its exceptional recombination efficiency between homologous DNA sequences that this step has been achieved over the past ten years. Yeast tolerates well the presence of exogenous DNA fragments the same length as its own chromosomes and, when several fragments share identical sequences at their ends, it joins them into a single molecule. Thus, a first synthetic DNA of 582,970 nucleotides corresponding to the genome of *Mycoplasma genitalium* could be constructed in 2008 (Gibson *et al.* 2008). By applying these principles of homologous recombination in yeast in successive cycles, larger and larger

several tens to a hundred deoxyribonucleotides, commonly called "synthetic oligonucleotides", which serve as "bricks" for the synthesis of whole genomes.

5 It was in the early 1980s that phosphoramidites replaced phosphotriesters and the first automatic oligonucleotide synthesizers appeared.

DNA molecules were obtained. In three successive cycles of yeast transformation, an artificial genome of *Mycoplasma mycoides* measuring 1,077,947 nucleotides was constructed from one thousand synthetic DNA cassettes of 1,080 nucleotides each, sharing two by two, 80 nucleotides of identical sequences at their ends (Gibson *et al.* 2010). Transplanted into the cytoplasm of *Mycoplasma capricolum*, this chemically synthesized genome gave birth to a bacterial clone with the phenotypic characteristics of *M. mycoides*. Several platforms for assembling DNA fragments in yeast (*genome foundries*) have recently been built around the world to manufacture DNA molecules of up to one megabase length on demand.

Why synthesize such DNA molecules? Freed from existing sequences, the infinity of possibilities is dizzying, and applications are only limited by our imagination. In the current state of our knowledge, it is possible to construct metabolic pathways that enable compounds of interest to be produced by organisms that do not have them. For example, yeasts have now been constructed that produce opiate substances for medical use by introducing in their genomes about 20 artificial genes copied from plants, mammals, bacteria or other yeasts and capable of producing the enzymes necessary for the synthesis of these complex molecules (Galanie *et al.* 2015). A particular genome can also be massively modified by introducing many deliberate changes (codon changes, easily actionable sequence insertions, etc.). For example, a viable *E. coli* genome of nearly 4 Mb has been constructed, which now uses only 57 codons to produce its proteins, leaving the saved codons potentially usable to incorporate new amino acids into proteins after construction of artificial transfer RNA molecules (Ostrov *et al.* 2016). With this strategy, one could obtain organisms capable of producing entirely new proteins that may have biochemical or biological activities still unknown today. Currently, several laboratories are investigating the extent to which a genome can be reduced while maintaining the organism's properties (elimination of repeated sequences, mobile elements, non-essential genes, etc.), in order to produce minimum-genome "chassis" strains, on which it would be easier to construct. For example, a viable genome of *Mycoplasma mycoides* was produced having only 531 kb of DNA (Hutchison III *et al.* 2016).

The most spectacular developments currently underway are those of the synthesis of eukaryotic genomes that could lead to the complete chemical synthesis of genomes of new animals or even humans. For the time being, it is still the yeast *S. cerevisiae* that has been chosen as proof of concept, given

the many experimental advantages it offers. The idea is to replace all its natural chromosomes by modified synthetic sequences in successive steps while continuing to produce a normal yeast. The work, which began a few years ago with the chemical synthesis of chromosome III, is now almost completed by an international consortium of laboratories (Pretorius and Boeke 2018). The artificial genome is organized differently from the natural genome. Genes for stable non-coding RNAs (ribosomal RNAs, transfer RNAs) were placed on an independent chromosome, mobile elements were ignored, UAG stop codons replaced by UAA codons and, most importantly, approximately every 3 kilobases, chromosomes were scattered with sites (called LoxP) recognizable by a specific recombinase (called Cre). All these modifications are tolerated by the yeast without any significant alteration in cell morphology, life cycle, chromosome conformation and stability or gene expression. LoxP sites enable the generation of an infinite variety of combinatorial rearrangements of the chromosomes (deletions, inversions, duplications, translocations) by simple transient expression of Cre recombinase in yeast. This procedure, called SCRaMbLE (for Synthetic Chromosome Rearrangement and Modification by LoxP-mediated Evolution) is now used to produce yeast strains with a wide phenotypic diversity, making it possible to accelerate fundamental research on gene functions, the optimization of artificial metabolic pathways or the exploitation of interspecific hybrids.

Producing living organisms capable of reproduction and whose genomes are entirely or partially artificial raises regulatory issues similar to those discussed for recombinant DNA in the 1970s (see Chapter 4). The laboratories of the international synthetic yeast consortium, for example, have committed themselves to a written charter based on the recommendations of the Asilomar Conference. With the acceleration and internationalization of such work, the problem of their legal framework will inevitably arise quickly.

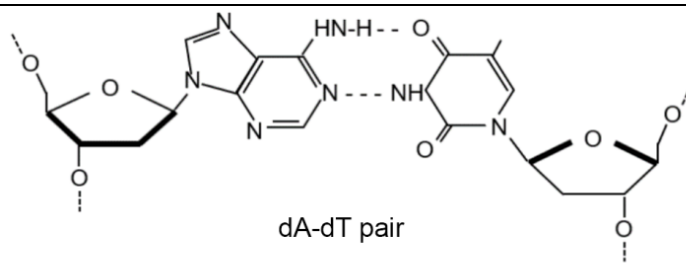
9.3. New lives

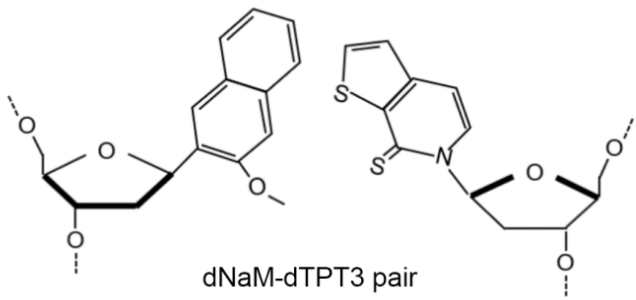
But why restrict ourselves to copying nature? Whether it is the editing of genomes or their chemical synthesis, the product is still DNA. But while the living world we know uses only two polymers – DNA and RNA – for the storage and processing of its genetic information, chemistry opens up a much wider field of possibilities.

Since **Steven Benner's** pioneering work in 1989, significant research activity has focused on modifications of the building elements of nucleic acids (nitrogenous bases, sugars and backbone structures), the details of which go beyond the needs of this book, but the results of which are giving a new dimension to genetics. We are now talking about xenobiology and xeno-nucleic acids or "XNA".

The chemistry of nucleic acids is very rich. Recently, for example, deoxyribonucleotidic chains have been synthesized with artificial bases unknown in natural DNA but which, by hydrogen bond pairings, allow the formation of double helixes as stable as natural DNA (Hoshika *et al.* 2018). By pairing small bases (pyrimidine radicals) or large bases (purine radicals) with each other, two double helixes are obtained, respectively narrower or wider than natural DNA, but which conserve the same principle of strand complementarity as the purine-pyrimidine pairing of the natural DNA. There was, therefore, no obvious chemical logic to the choice of DNA in evolution. But, of course, all enzymes in living cells that act on nucleic acids have evolved with DNA, so it is not surprising that XNAs are poor substrates for these enzymes. Rebuilding artificial biological systems with these analogs therefore requires to evolving these enzymes to modify their substrate specificity.

An example of such work is given by the DNA polymerase of *Thermotogus gorgonarius* (Pinheiro *et al.* 2012). After a few *in vitro* evolution cycles that selected several mutations, a mutant enzyme was obtained that was capable of using a DNA strand as template to synthesize a long enough XNA to carry a genetic information. This XNA was a nucleo-1,5-dianhydrohexitol acid. Similarly, a reverse transcriptase mutant was obtained able to use a strand of this XNA as template to synthesize a complementary DNA strand. Genetic information can therefore be conveyed through an XNA. Using the same principles, six different XNAs could be used as carriers of genetic information. In addition, they are able to evolve by mutation in *in vitro* selection systems. The two fundamental properties of life – heredity and evolution – are therefore not limited to DNA and RNA, but can be shared by different polymers, provided that they are able to convey information. This conclusion is not without significance for our speculations on the origin of terrestrial life or for the forms of life that may exist in the universe.





The artificial nucleotide pair dNAM-dTPT3, compared to the natural pair of dA-dT, shows comparable stereochemical volume that allows it to be incorporated into a genome by natural enzymes (Zhang and Romesberg 2018).

Many synthetic molecules have been produced which, mimicking the properties of nucleic acids, are capable of carrying **genetic information**. Some of these molecules can be incorporated into living cells leading to the formation of **semi-synthetic organisms**.

Shown are two artificial nucleotides, dNaM and dTPT3, able to pair with each other by **hydrophobic bonds**, which have been successfully incorporated into the *E. coli* genome. The DNA formed, which has six deoxyribonucleotides instead of four, can be transcribed into RNA molecules containing the NaM and TPT3 ribonucleotides in addition to the four natural ribonucleotides. Semi-synthetic organisms pave the way for the production of proteins containing unconventional amino acids that can expand the range of biochemical activities almost indefinitely.

The example shown leads to a **six-letter genetic code** instead of four, in which unconventional amino acids have already been assigned to the two artificial codons AXC and GXC (X = NaM) following the construction of **artificial transfer RNA** molecules capable of participating in protein synthesis.

Box 9.2. *New nucleic acid chemistry and genetic code expansion*

But beyond the *in vitro* evolution of molecules, it is the possibility of incorporating an XNA into the genetic material of living cells that represents the greatest potential of xenobiology. Through laboratory-directed evolution experiments, it was possible to evolve *E. coli* cultures to incorporate 5-chlorouracil in their DNA instead of thymine (5-methyl-uracil). 5-chlorouracil is normally toxic, but, when supplied in slowly increasing concentrations together with thymine to mutants unable to synthesize thymine, a gradual selection of mutants capable of incorporating 5-chlorouracil into their DNA in place of thymine is obtained (Marlière *et al.* 2011). Clones isolated after 6,000 generations carry a large number of mutations in their genome because, unlike thymine, 5-chlorouracil can also pair with guanine. While thymine now becomes toxic to them, they remain viable if one continues to supply them with 5-chloro-uracil.

9.4. Important ideas to remember

– Our current knowledge is based on a biased **sampling** of the living world and recent explorations of its actual diversity show us that we can probably expect many more surprises.

– The chemical synthesis of DNA produces **oligonucleotides** that can then be assembled *in vitro* into long DNA molecules or complete functional genes.

– Long, chemically-synthesized DNA molecules can be assembled into **synthetic chromosomes** using homologous recombination in yeast.

– Several fully **synthetic genomes** have already been constructed.

– **New synthetic molecules** are capable of replacing nucleic acids.

9.5. References

Brum, J.R. *et al*. (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science*, 348(6237), 1261498.

Carradec, Q. *et al*. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9, 973.

Curtis, B.A. *et al*. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, 492, 59–65.

De Vargas, C. *et al*. (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605.

Dorrell, R.G., Smith, A.G. (2011). Do red and green make brown? Perspectives on plastid acquisitions within chromalveolates. *Eukaryotic Cell*, 10, 856–868.

Edge, M.D., Green, A.R., Heathcliffe, G.R., Meacock, P.A., Schuch, W., Scanlon, D.B., Atkinson, T.C., Newton, C.R., Markham, A.F. (1981). Total synthesis of a human leukocyte interferon gene. *Nature*, 292, 756–762.

Galanie, S., Thodey, K., Trenchard, I.J., Filsinger Interrante, M., Smolke, C.D. (2015). Complete biosynthesis of opioids in yeast. *Science*, 349, 1095–1100.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison, C.A., Smith, H.O. (2008). Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science*, 319, 1215–1220.

Gibson, D.G. *et al*. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science*, 329, 52–56.

Hoshika, S., Singh, I., Switzer, C., Molt, R.W., Leal, N.A., Kim, M.-J., Kim, M.-S., Kim, H.-J., Georgiadis, M.M., Benner, S.A. (2018). "Skinny" and "fat" DNA: Two new double helices. *Journal of the American Chemical Society*, 140, 11655–11660.

Hutchison III, C.A. *et al*. (2016). Design and synthesis of a minimal bacterial genome. *Science*, 351, 6253.

Itakura, K., Hirose, T., Crea, R., Riggs, A.D., Heyneker, H.L., Bolivar, F., Boyer, H.W. (1977). Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science*, 198, 1056–1063.

Khorana, H.G., Agarwal, K.L., Büchi, H., Caruthers, M.H., Gupta, N.K., Kleppe, K., Kumar, A., Otsuka, E., RajBhandary, U.L., Van de Sande, J.H., Sgaramella, V., Terao, T., Weber, H., Yamada, T. (1972). Studies on polynucleotides. 103. Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast. *Journal of Molecular Biology*, 72, 209–217.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J.M., Raoult, D. (2003). A giant virus in amoebae. *Science*, 299, 2033.

Marande, W., Burger, G. (2007). Mitochondrial DNA as a genomic jigsaw puzzle. *Science*, 318, 415.

Marlière, P., Patrouix, J., Döring, V., Herdewijn, P., Tricot, S., Cruveiller, S., Bouzon, M., Mutzel, R. (2011). Chemical evolution of a bacterium's genome. *Angewandte Chemie International Edition*, 50(31), 7109–7114.

Ostrov, N. *et al*. (2016). Design, synthesis, and testing toward a 57-codon genome. *Science*, 353, 819–822.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., Abergel, C. (2013). Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, 341, 281–286.

Pinheiro, V.B., Taylor, A.I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J.C., Wengel, J., Peak-Chew, S.Y., McLaughlin, S.H., Herdewijn, P., Holliger, P. (2012). Synthetic genetic polymers capable of heredity and evolution. *Science*, 336, 341–344.

Pretorius, I.S., Boeke, J.D. (2018). Yeast 2.0 – connecting the dots in the construction of the world's first functional synthetic eukaryotic genome. *FEMS Yeast Research*, 18(4).

Roux, S. *et al*. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537, 689–693.

Sunagawa, S. *et al*. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science*, 348(6237), 1261359.

Yilmaz, P., Yarza, P., Rapp, J.Z., Glöckner, F.O. (2016). Expanding the world of marine bacterial and archael clades. *Frontiers in microbiology*, 6, 1524.

Zhang, Y., Romesberg, F.E. (2018). Semisynthetic Organisms with Expanded Genetic Codes. *Biochemistry*, 57, 2177–2178.

Zhang, Y., Ptacin, J.L., Fischer, E.C., Aerni, H.R., Caffaro, C.E., San Jose, K., Feldman, A.W., Turner, C.R., Romesberg, F.E. (2017). A semi-synthetic organism that stores and retrieves increased genetic information. *Nature*, 551, 644–647.

Conclusion

C.1. Risk and the perception of risk

Putting into action the possibilities now emerging through the current trajectories of genetics raises a legitimate apprehension. How do we predict the behavior of synthetic or semi-synthetic organisms when it is still so difficult to predict a phenotype from a genotype for the existing organisms? And how to anticipate the interactions of elements that did not yet exist with the terrestrial living world? Answering these questions implies first of all to clarify the boundary between the natural and the artificial, less clear than it seems. The same molecule, for example vitamin C, will be perceived differently by the non-specialist public depending on whether it is extracted from a fruit or chemically synthesized from glucose. Yet, it is the same ascorbic acid, $C_6H_8O_6$ or, more precisely, (2R)-2-[(1S)-1,2-dihydroxyethyl]-3,4-dihydroxy-2H-furan-5-one.

In this perception, the origin of the molecule prevails over its substance, as if there remained in it a trace of this origin, a kind of mysterious vital principle. This is obviously not the case. No trace of origin exists in any molecule, only its atomic structure defines it and contains in itself all its properties. But the illusion of a hierarchy of values between what is natural and what is artificial persists in people's minds for historical reasons. More than 250 years ago, **Carl Linnaeus** (Linneaus 1737) declared in his *Critica Botanica:*

> I distinguish the species of the Almighty Creator which are true, from the abnormal varieties of the Gardener: the former I

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

> reckon of the highest importance because of their Author, the latter I reject because of their authors [...].

He was the father of modern taxonomy who greatly helped to clarify biology, but his distinction does not survive subsequent genetic discoveries. The "abnormal varieties" of the gardeners result from the same natural mechanisms as his "true species".

The domestication of plants and animals offers us another dimension to reflect on. The perception of artifacts fades with time. Like the flowers in our gardens, the plant varieties and animal breeds used in our agronomy are not wild. Yet they are considered by the public as natural, as are agronomic practices such as artificial fertilization, grafts or hybrid production. The genetic changes involved, sometimes drastic, are assimilated to natural events because they are not deliberate (and often not even known). On the other hand, the same changes deliberately produced by rewriting or synthesizing genomes, or even by simply combining existing genes as in the case of drought-resistant maize lines obtained by CRISPR-Cas technology, will be considered artificial. Once again, the origin of the product takes precedence over its nature, in an irrational reasoning. If anything, the only difference between the natural and the artificial is the existence of a purpose. Now, if a purpose is present in the latter case, it is totally absent from the first.

The last illusion is that of the permanence of the natural, against which any artificial is perceived as a change. Yet, the greatest lesson of genetics is that permanent change is intrinsic to the living world. There is no real permanence. We are talking about *the* human genome or *the* oak genome when, in reality, there are as many versions of these genomes as there are people on the planet or trees in forests. Platonic uniqueness facilitates reasoning, it makes it possible to define universal references, but it cannot ignore the actual individual diversity which is such that even species have become practically impossible to define since we are able to decipher genomes. So what about varieties or breeds, especially when the diversity of populations is itself constantly changing? This does not mean that any artificial change is trivial, but it illustrates the context in which it must be considered in order to assess real risks. Inserting a new gene into a genome when, at the same time, nature inserts myriads of them is not in itself a problem. What needs to be examined is the exact nature of this gene. Once again, it is the product that is important, not its origin, although this idea contradicts the culture of patents that emphasizes processes.

If, as we currently imagine, the origin of life on Earth some 3.5 billion years ago was the capture of information by polymers capable of replication and therefore of evolving by mutation – we think of RNA – it is normal to find this essential property of information capture in the contemporary living world and to exploit it for the production of genetically modified organisms. But what is their relative importance when we know that one tenth of our own genome is made up of viral sequences accumulated during evolution from our distant ancestors, who had none or probably others in their genomes? Again, this does not mean that any directed modification of genomes is safe. One could imagine, for example, deliberate destructive acts through the construction of dangerous organisms (pathogens, toxin producers, etc.). This simply means that it is not the process that must be considered, but the manufactured product.

When it comes to synthetic genomes and, even more so, to new forms of life, the boundaries between artificial and natural become sharper as the differences increase. But as the limits become clearer, the risks decrease because, contrary to our intuition, in the living world danger increases with proximity. The AIDS epidemic is caused by our close relatedness with other primates, while we can eat oysters, full of viruses, without major risk. Allografts, such as blood transfusions, carry a higher risk of transmitting infectious agents than xenografts. On the other hand, a transgenic plant has no chance of altering the human genome, and a synthetic genome will have less and less chances to exchange its genes with the natural populations as its DNA is more deeply chemically modified.

C.2. Ethics and genetics

Science can always be exploited for political or ideological interests. Genetics has not escaped this.

From the very beginning, the eugenicist ideas initiated by F. Galton in 1883 tried to divert genetics from its purpose by applying to the social domain simplistic Darwinian principles of natural selection which, in reality, have no real genetic basis because they confuse genotype and phenotype. According to these principles, the social and medical advances brought by civilization, abolishing natural selection by protecting the survival of the weakest, would be responsible for the spread of defects in the human species such as weakness, physical malformation or deviant behaviors like

alcoholism. The society of the time, convinced of the inheritance of acquired characters under environmental influence during life, was permeable to these theories. The corollary application of the theory of eugenics in various countries such as the United States and England at the beginning of the 20th Century led to forced internments and sterilizations, where the arbitrariness and total lack of scientific basis are now obvious. Generally opposed to these practices on moral grounds, the main religions that already rejected natural selection and evolution also came to denounce genetics, even though it was separate from these abuses. Eugenics culminated in the horrific crimes of Nazism which, in addition to massive arbitrary sterilizations and executions, carried out shameful experiments on the illusory concept of races. Obviously, none of this had any scientific basis, but these dramatic events show how much vigilance is required in the face of the pseudo-scientific excesses from which the current world is unfortunately not exempt.

The notion of race, whatever the species, is an imaginary construct which, by highlighting particular characteristics, artificially brings together a certain number of individuals that other criteria would group otherwise. In the case of the human species, it is coupled with social prejudices devoid of any scientific basis. Human genomics shows that most of the genetic variation between individuals is distributed as gradients without precise limits between different populations. In addition, there is considerable overlap in the distribution of alleles between individuals belonging to different populations, making illusory any attempt of classification into subgroups. This reality is explained by human history since its inception, made up of migrations in diverse environments and continuous genetic exchanges between populations. These exchanges were superimposed on the multiple mutations accumulated over generations, as discussed in Chapter 7 (Hellenthal *et al.* 2014). While, for each individual, genomics now makes it possible to reconstitute the heritage of ancestral genomes, it primarily shows the multiplicity of these heritages and their heterogeneity within populations (Royal *et al.* 2010). Evoking genetics, as some are trying to do, to promote condemnable racist ideologies must therefore be denounced with the utmost assurance.

Of course, this does not exempt science from its own responsibilities. By clarifying our understanding of the living world and increasing the power of human intervention on it, advances in genetics require new reflection. The American National Academies of Sciences, Engineering and Medicine recently published a joint book proposing rules of science, ethics

and governance for new genome engineering technologies, particularly when applied to the human species (National Academies of Sciences, Engineering and Medicine 2017). While not binding, these common sense recommendations could provide the necessary reflections to establish informed regulations that different countries could adopt taking into account their own cultures. The fact that germline gene therapy has become technically possible with present genomic editing tools makes it necessary to rethink its easy prohibition from the time when it was simply impossible. The three Academies recommend that serious genetic diseases be eliminated with caution and under control in cases where no other medical solution exists (incurable diseases transmissible to offspring). On the other hand, they totally prohibit this practice for the so-called "comfort" genetic modifications (the choice of offspring). The same prohibition applies to the modification of somatic cells in the case of benign diseases (comfort treatments), but not for serious genetic diseases for which it is recommended to act according to existing local jurisdiction. Finally, concerned that these restrictions should not impede progress, the Academies advocate the authorization of genetic modifications carried out for research purposes, including those involving human embryos *in vitro*, within existing jurisdictions.

New genome modification techniques also make it possible to reprogram an epithelial cell *in vitro* so that it can later differentiate into various cell types (pluripotency). A recent example has been obtained by acting on only four genes encoding transcription factors (Takahashi and Yamanaka 2016). This opens up new therapeutic possibilities for the repair of accidentally or pathologically damaged tissues or simply aging tissues. Everyone could thus have their share of "spare parts". Will this lead to attempts to prolong life span or glorify intellectual capacities with the tools of molecular genetics, as some transhumanist currents suggest? Will these tools be used to genetically modify our human condition?

Another problem arises with the gene drive procedures because they are not limited to the genetically modified individuals themselves. By modifying the rules of hereditary transmission, gene drive procedures offer the ability to alter or even eliminate entire populations through their simple natural reproductive mechanisms (see Chapter 6). Logically, such applications would require the ability to assess in advance their consequences in terms of advantages and disadvantages for ecosystems, which is very difficult given our current state of knowledge. An alternative is to launch trials in confined

territories, such as oceanic islands, for terrestrial populations. Transparent public information is essential to build trust, while research continues in controlled laboratory environments. A responsible strategy would be to require, for each construction of a new gene drive tool, the construction of a reversible tool that could serve as a safety feature in the event of mishap in the field, as has already been modeled in the laboratory with microorganisms (DiCarlo *et al.* 2015).

Various countries are trying to adapt their recommendations to the accelerated advances in genetics and genomics that become applicable. These applications can bring about substantial improvements in the medical, agronomic or environmental fields, but they upset our understanding of the very nature of the human species and its place in the living world. Some are tempted to advocate research moratoria pending hypothetical progress, which is hard to imagine in the absence of research! The advances in genetics lie within the context of the remark attributed to **Albert Einstein**: "Life is like riding a bicycle. To keep your balance, you must keep moving." The ethical choice is not to decide whether genetics should continue or stop moving forward, but rather to decide which trajectory to follow.

C.3. References

DiCarlo, J.E., Chavez, A., Dietz, S.L., Esvelt, K.M., Church, G.M. (2015). Safeguarding CRISPR-Cas9 gene drives in yeast. *Nature Biotechnology*, 33, 1250–1255.

Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., Myers, S. (2014). A genetic atlas of human admixture history. *Science*, 343, 747–751.

Linneaus, C., *Critica Botanica*, (1737) Leyden. Translated by Arthur Hort, revised by M.L. Green (1938). Ray Society Series, Bernard Quaritch, Ltd. London.

National Academies of Sciences, Engineering and Medicine (2017). Human genome editing. Science, ethics and governance. Report, The National Academies Press, Washington D.C.

Royal, C.D., Novembre, J., Fullerton, S.M., Goldstein, D.B., Long, J.C., Bamshad, M.J., Clark, A.G. (2010). Inferring genetic ancestry: Opportunities, challenges and implications. *American Journal of Human Genetics*, 86, 661–673.

Takahashi, K., Yamanaka, S. (2016). A decade of transcription factor-mediated reprogramming to pluripotency. *Nature Reviews Molecular Cell Biology*, 17, 183–193.

Glossary

The terms which appear in the glossary are in ***bold italics*** with an asterisk* in the text when they first appear.

Activator (transcription): a molecule, generally a protein, whose action on DNA or chromatin activates the transcription of DNA from one or more promoters. Activators can act alone or within multi-molecular complexes. See *repressor*.

Addition (of nucleotide): insertion into a sequence of one or a few nucleotides during a mutation. See *indel*.

Allele: a form of a gene. In a given living species, a gene can take many forms, which differ from each other by the DNA sequence. Differences are at the root of the genetic polymorphism of the population. A locus on a chromosome can only carry one allele at a time. A diploid cell carries two alleles of each gene that may or may not be the same.

Allogamous: describes a species that reproduces preferentially or exclusively by cross-fertilization (between different individuals). See *autogamous*.

Allostery: the three-dimensional conformational change of macromolecules having several different stable states. This change is usually induced by the presence of effectors.

Amino acid: a simple organic molecule made of a central carbon atom (known as "alpha"), to which are chemically bonded a hydrogen, a

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

carboxylic acid (-COOH), an amine ($-NH_2$) and a variable radical according to the nature of the amino acid. There are many different amino acids in cellular metabolism, 22 of which are involved in the composition of proteins (see "The genetic code", section 2.4).

Amplification: multiplication of the copy number of a gene or chromosome segment by a factor higher than the duplication.

Anaphase: phase of mitosis where sister chromatids separate and migrate to the opposite poles of the cell. During the first division of meiosis, anaphase I separates homologous chromosomes and, during the second division, anaphase II separates sister chromatids.

Aneuploid: describes an organism, organ, tissue or cell whose chromosome number is not a multiple of the haploid number of the species.

Anisogamy: a form of fertilization in which the male and female gametes are of different size, physiology or morphology.

Anticodon: the three nucleotides of a transfer RNA molecule that pair with the corresponding codons of the messenger RNAs.

Apoptosis: the process by which cells trigger their death in response to a signal.

Archaea: formerly "archaebacteria", member of one of the two domains of the tree of life where cells have no nuclei. The other is formed by bacteria. The third domain, eukaryote, is made up of organisms whose cells have a nucleus.

Autogamous: describes a species that reproduces by fertilization between gametes from the same individual. See *allogamous*.

Auxotrophic: describes a living organism whose diet must consist of organic molecules that it is unable to synthesize. Opposite: prototrophic.

Bacteria: or eubacteria, member of one of the two domains of the tree of life where cells have no nuclei. The other is formed by the archaea. The third domain, eukaryote, is made up of organisms whose cells have a nucleus.

Bacteriophage: bacterial virus. The term has remained in use because of the history of discoveries, but also because of the very different molecular mechanisms that differentiate bacterial genomes from eukaryotic genomes.

Bivalent: a pair of homologous chromosomes matched during the prophase of the first division of meiosis.

Centromere: area of the chromosome that unites the two sister chromatids and connects them to the fibers of the spindle during mitosis and meiosis.

Chiasma: site of exchange between homologous chromatids during meiotic prophase.

Chloroplast: or, more generally, plastid. Cell organelle present in eukaryotes of the Archaeplastidae lineage (plants and algae) and, occasionally, other lineages. Where photosynthesis takes place.

Chromatid: each of the two copies resulting from the replication of a chromosome during the cell cycle and joined together at least at the level of the centromere before their separation.

Chromatin: chemical component of chromosomes. Chromatin is made up of proteins (called "histones") and DNA. In eukaryotic chromosomes, DNA is wrapped around a protein nucleus made up of four histones, forming a nucleosome. Nucleosomes are connected to each other by the DNA molecule in the manner of pearls on a necklace.

Chromosome: etymologically "colored body". The permanent unit element carrying all or part of the genetic material of a cell and transmitted independently of other chromosomes to daughter cells during cell division. Chromosomes are made of chromatin and each contains only one DNA molecule. Depending on the organism, chromosomes can be circular (bacteria, archaea, organelles of eukaryotic cells) or linear (nuclei of eukaryotic cells).

Cis: term used to designate two genetic elements carried by the same DNA molecule. Opposes *trans*. In practice, also refers to the interaction between two genetic elements carried by the same DNA molecule if this interaction involves the genes themselves and not their products.

Cistron: functional definition unit of the gene derived from the *cis-trans* test. A cistron corresponds to an end product of the gene. A single gene in the molecular sense of the term may have one or more end products and will therefore be composed of as many cistrons.

Clone: all cells or individuals derived from a single common ancestor by asexual multiplication exclusively. All the elements of a clone are genetically identical (with the exception of mutations that may have occurred during the growth of the clone).

CNV: copy number variant. Refers to variations in the number of copies of genes or chromosome segments observed by comparing genomes. These variations result from deletions, duplications or amplifications of DNA segments that can occur during cell divisions.

Codon: triplet of ribonucleotides having a meaning in the genetic code. Since there are four ribonucleotides, there are $4^3 = 64$ different codons.

Core-genome: all genes common to all individuals of the same species (or the sample studied).

CRISPR-Cas: a targeted genome modification technique, derived from bacteria, based on the guidance of an endodeoxyribonuclease to a DNA site by a short RNA.

Cross-over: phenomenon by which, during meiosis, homologous chromosomes exchange certain segments of their genetic material before separating.

Cytogenetics: study of chromosomes during mitosis or meiosis using cytological techniques.

Deletion: the loss of a continuous fragment of genetic material, from a single nucleotide to several genes. Deletion limits can split or merge genes.

Deoxyribose: five-carbon sugar derived from ribose by replacing the hydroxyl radical carried by carbon 2' with a hydrogen atom. In DNA, deoxyribose carries the nitrogenous bases on its 1' carbon.

Diplobiontic: describes a developmental cycle in which the diploid phase is more important than the haploid phase.

Diploid: describes a genome composed of two sets of homologous chromosomes and, by extension, to the organism that contains them. See *haploid*.

DNA: deoxyribonucleic acid. A macromolecule formed by two chains of four deoxyribonucleotides, A, C, G and T, whose sequences are complementary to each other. The two chains are maintained together by the hydrogen bonds established between the nitrogenous bases of the nucleotides, forming a double helix. DNA is the carrier of the genetic information of all cellular organisms and many viruses.

Dominance: property of a character that prevails over another character in a hybrid. By extension, property of an allele whose phenotype masks that of another allele at the same locus.

Duplication: additional copy of a gene or of all or part of the genome of the same cell. The limits of duplication can split or merge genes.

Dysgenesis (hybrid): malformation or death of the descendants of a cross between normal individuals following the activation of specific transposable elements during hybridization.

Editing or rewriting (of the genome): targeted modification of a genome using endodeoxyribonucleases which, fixed to a specific site of the DNA molecule defined by its sequence, make a cut of the two strands which will be repaired by homologous recombination or by non-homologous repair.

Editosome: an enzyme complex consisting of proteins and guide RNAs, which is responsible for the editing of RNAs.

Enantiomer: each of the two forms of the same asymmetric molecule which are images in a mirror of each other and therefore not superimposable (left- and right-hand model).

Endonuclease: an enzyme that cuts a nucleic acid into shorter fragments.

Endoreplication: process leading to a complete doubling of the chromosome number of a cell. Endoreplication is a rare event in which DNA replication is not followed by cell division. Multiple endoreplications can be observed, for example, in certain tissues of plants or insects.

Endosymbiosis: symbiotic association where one of the organisms, called "endosymbiont", is present inside the cells of its host.

Enhancer: a region of DNA that can bind proteins to stimulate gene transcription. The gene and the *enhancer* are not necessarily close to each other.

Enzyme: catalyst that accelerates or makes possible the biochemical reactions of the living cells. Most enzymes are proteins, but RNAs can also have catalytic actions. RNA-enzymes are called "ribozymes".

Epigenetics: the phenomenon of hereditary transmission of a phenotypic trait resulting from heritable changes in gene expression and not from changes in DNA sequence. This heritability results from the complexity of the molecular mechanisms of gene expression. It is only partial, epigenetic traits are reversible, but it can persist for large numbers of successive generations.

Episome: DNA molecule independent of chromosomes that replicates autonomously. Some episomes can also integrate into cellular chromosomes and be replicated in their continuity. On the contrary, chromosome fragments can be excised and persist in the cell as episomes. An episome may or may not be essential to cell life and, in the latter case, may be optional.

Epistasis: functional interaction between two or more different genes. For example, epistasis occurs when the allele of one locus masks or modifies the phenotypic effect of another allele at another locus.

Equational: describes the second division of meiosis where sister chromatids separate.

Eukaryote: member of the only domain of the tree of life where cells have a nucleus. The other two domains, archaea and bacteria, are made up of organisms whose cells are nucleus-free.

Exon: for expressed region. Segment of an RNA molecule that will be spliced to another exon during a splicing reaction. Each exon is therefore necessarily flanked by at least one intron (terminal exon), sometimes two introns (central exon). Exons may be non-coding or partially or totally

coding when contributing to a messenger RNA molecule. By extension, the corresponding segment of DNA in the genome (usually distinguished as coding, partially coding or non-coding exons). It is worth noting that, RNA splicing being often alternative, the same DNA segment can therefore be simultaneously intron and exon (coding or not), depending on how it is considered.

Exosome: protein complex capable of rapidly degrading the various RNA molecules present in cells that are not protected by specific signals (cap, poly-A tail, chemical modifications, three-dimensional structures, etc.) or with anomalies (incomplete transcription, incorrect splicing, etc.).

Expressivity: the phenotypic manifestation of a mutation or a genetic or chromosomal anomaly.

Fertilization: stage of sexual reproduction where the fusion of male and female gametes produces the cell called "zygote".

Fixation: occurs when, as a result of **genetic drift** or **selection**, a single allele becomes present at one locus in all members of a natural population.

Gamete: mature sex cell (normally haploid).

Gametophyte: haploid phase of the reproductive cycle of Archiplastidae (plants and algae).

Gene drive: an active process by which one allele eliminates another allele at the same locus when they meet in the same cell. In the case of sexual reproduction, this leads to ignoring the Mendelian proportions in the offspring, all or a majority of the descendants inheriting the allele of only one of the two parents. This process exists in nature in cases where special molecular machineries are active. It is now being artificially used to drive desired population change.

Genome: the complete hereditary material composed of nucleic acids (DNA or RNA) of a cellular organelle, a cell, an organism or a species. See *subgenome*.

Genotype: all or known part of the genetic information of an individual, a cell or a virus.

Genetic code: rules of equivalence linking a ribonucleotide sequence (RNA) to an amino acid sequence (polypeptide). The genetic code gives the meaning of each of the 64 possible triplets of ribonucleotides.

Genetic linkage: tendency of two alleles of two different genes to be transmitted together by a gamete. The genetic linkage generally corresponds to the fact that the two genes are carried by the same chromosome. It increases as the genes are closer each other.

Genome annotation: identifying and locating notable sequences of a genome and assigning them biological functions.

GWAS: for Genome Wide Association Study. Statistical analysis of the allelic distribution in a cohort of individuals with a particular phenotypic trait compared to the distribution of the same alleles in a control population. Applied to the entire genome, the method makes it possible to identify the alleles involved in the genetic determinism of a particular trait.

Haplobiontic: refers to a developmental cycle in which the haploid phase is more important than the diploid phase.

Haploid: refers to a genome composed of a single set of chromosomes and, by extension, the organism that contains them.

Haplotype: group of alleles at different loci on the same chromosome received together from the same parent.

Hemizygosity: a condition of a chromosome or part of it that, in a diploid cell, is found in a single copy by loss of its counterpart.

Heritability: the notion of heritability, which dates back to 1941, has acquired great importance with the recent development of genomics. It estimates the part of the variance of a phenotype attributable to genes within a population. By eliminating the proportion of phenotypic variance attributable to the environment (homogeneous conditions), heritability is the ratio of total genetic variance known to the observed phenotypic variance. The total genetic variance is the sum of the identified genetic variances.

Hermaphroditism: normal presence of both sexes in the same individual, animal or plant, which produces both categories of gametes, male and female.

Heterogametic: describes an organism whose sexual reproduction is based on morphologically distinct male and female gametes.

Heteroplasmic: describes a eukaryotic cell with a mixture of distinct mitochondrial genomes or distinct chloroplastic genomes.

Heterozygous: the presence of two distinct alleles at the same locus in a diploid genome.

Homogametic: describes an organism whose sexual reproduction is based on morphologically similar gametes.

Homoplasmic: describes a eukaryotic cell with only one type of mitochondrial or chloroplastic genomes.

Homozygous: the presence of two identical alleles at the same locus of a diploid genome. By extension, the presence of as many identical alleles at the same locus as the chromosomal number in a polyploid.

Indel: for insertion-deletion. All point mutations that result in the addition or loss of a single or very small number of nucleotides in DNA. Currently, indels are identified by comparing aligned DNA sequences, provided they are similar enough.

Intron: short for internal region. The internal segment of an RNA molecule that will be eliminated during splicing. Each intron is therefore flanked by two exons. After splicing, this RNA segment can be degraded or participate in various functions (formation of non-coding RNAs, translation into protein, etc.). By extension, the corresponding segment of DNA in the genome. However, RNA splicing is often alternative and an intron may contain genes or moving genetic elements. The same DNA segment can, therefore, give rise to an intron or an exon at the RNA level.

Inversion: reversal of a DNA or chromosome segment taking with it all the genetic elements it contains without changing their number. The limits of inversions can split or merge genes.

Isogamy: a form of fertilization in which male and female gametes are of identical size, physiology or morphology.

Karyotype: classification of chromosomes according to their size and position of centromere, from a microscopic image of the metaphase of cell division.

Locus: location on a chromosome.

Meiosis: succession of two modified and entangled mitotic divisions which, starting from a diploid cell, produce four haploid daughter cells. Meiosis is therefore reserved for eukaryotic cells. During meiosis, the set of chromosomes is divided by two, each daughter cell receiving only one of the two members of each pair of homologous chromosomes with or without exchanges between them (crossover).

Meristem: set of cells at the end of the stems and roots responsible for building plant organs.

Messenger RNA: RNA molecule of which part of the sequence is an open reading frame which, when associated with ribosomes, will be translated into protein according to the rules of the genetic code.

Metagenome: all the genomes present in a heterogeneous population studied.

Metaphase: phase of mitosis where duplicate chromosomes are aligned at the equator of the cell before the separation of chromatids. During meiosis, the bivalent chromosomes align in metaphase I and the sister chromatids in metaphase II.

Microbiome: complete set of genes present in a given microbiota.

Microbiota: all living species of microorganisms in a studied sample.

Mitochondria: cellular organelle present in all eukaryotic organism lineages. Seat of cellular respiration and oxidative phosphorylations, but also of many metabolic biosynthesis pathways.

Mitosis: eukaryotic cell division in which a mother cell produces two daughter cells that are genetically identical to itself. Mitosis occurs in both

haploid and diploid (or even polyploid) cells. Before each mitosis, the DNA is fully replicated, resulting in the formation of two sister chromatids for each chromosome which, after condensation, will be equally distributed between the two daughter cells. In the case of diploid cells, homologous chromosomes behave independently of each other, unlike what happens during meiosis.

Modification (of DNA): term used in connection with the phenomenon of DNA restriction (see *restriction*). It generally consists of methylation of the adenines or cytosines that are part of specific small sequences forming the sites of recognition and action of enzymes called methylases.

Monogenic: refers to a trait whose variation depends on only one gene. It is also sometimes referred to as "Mendelian" and is opposed to quantitative, which means polygenic.

Mutagen: refers to an agent that causes a mutation frequency higher than the spontaneous frequency. The different mutagens induce different spectra of different types of mutations.

Mutagenesis: production of mutants by the action of a mutagen. Conventional mutagens, whether chemical or physical, increase the mutation rate by acting randomly throughout the genome. Interesting mutations are sorted *a posteriori*. On the contrary, directed mutagenesis, carried out using synthetic oligonucleotides or site-specific nucleases (see *genome editing*), targets a specific site of the genome without increasing the random mutation rate.

Mutant: a cell or organism that has inherited one or more mutations. The term is also used as an adjective, for example, mutant allele or mutant phenotype.

Mutation: spontaneous or induced modification of genetic information. There are point mutations, affecting a single nucleotide or a small number of adjacent nucleotides, and structural mutations, affecting whole fragments of chromosomes. Point mutations include nucleotide substitutions, as well as additions or losses of a few nucleotides (indels). Structural mutations include deletions, duplications, inversions and translocations.

Non-coding RNA: An RNA molecule that is not translated into protein. This designation covers a disparate set of RNA molecules of very different functions (ribosomal RNA, transfer RNA, snRNA, interfering RNA, etc.).

Nucleic acid: polymer of nucleotides. There are two types of nucleic acids, ribonucleic acid (RNA) made of ribonucleotides, and deoxyribonucleic acid (DNA), made of deoxyribonucleotides.

Nucleus: organelle of eukaryotic cells containing chromosomes (and sometimes episomes) and where replication and transcription processes, but not protein synthesis, take place. The nucleus is separated from the cytoplasm by a nuclear membrane containing complex structures, called "nucleopores" through which a large number of macromolecules pass in both directions. This membrane generally disappears during cell divisions and re-forms immediately afterwards in the two daughter cells. The nucleus contains a special space, called "nucleolus", where ribosomal RNAs are concentrated during maturation. The nucleus is also the place where splicing of introns occurs.

Nucleomorph: remainder of the nucleus of a eukaryotic cell following secondary endosymbiosis in another eukaryotic cell. Nucleomorphs contain chromosome remains (usually small), but carrying active genes.

Nucleosome: 11 nm diameter discoidal particle formed by an octamer of histones (H2A, H2B, H3 and H4, each in duplicate) around which the DNA double helix is wrapped over two turns. The nucleosome is the basic unit of chromatin.

Nucleotide: an organic molecule made of a 5-carbon sugar with a nitrogenous base and which, when phosphorylated, constitutes a nucleic acid monomer. The sugar is a ribose in the ribonucleotides that form RNA and a 2' deoxyribose in the deoxyribonucleotides that form DNA. The nitrogen bases are purines (adenine and guanine) and pyrimidines (cytosine, uracil or thymine, thymine being a 5-methyl uracil). Free nucleotides carry one, two or three phosphates on the 5' carbon of the sugar. Only triphosphate nucleotides are used in nucleic acid synthesis.

Oogamy: a form of fertilization in which female gametes are large, immobile cells with reserves (called "ovules" in the animal world).

Open reading frame: ribonucleotide sequence segment located between two stop codons in the selected reading frame.

Operator: term of bacterial genetics. A DNA segment to which a regulatory protein binds. The binding of this protein to the operator can either stimulate (called an activator) or inhibit (called a repressor) the transcription of adjacent genes. The binding of the regulatory protein to the operator is modulated by an effector, an external signal, often a small molecule such as a sugar or amino acid.

Operon: term of bacterial genetics. A transcription unit consisting of a promoter, an operator and one or more structural genes. In the case where the unit contains several structural genes, the messenger RNA produced is polycistronic.

Organelle: element of the internal structure of eukaryotic cells. Nuclei, mitochondria, chloroplasts, nucleomorphs, peroxisomes, vacuoles are organelles. Peroxisomes and vacuoles do not contain DNA.

Ortholog: homologous genes or genetic elements derived from the same common ancestor and separated in different species. See *paralog*.

Pan-genome: all the genes present in the genomes of all the members of a species (or the sample studied).

Paralog: homologous genes or genetic elements derived from the same common ancestor after duplication, and maintained together in the same organism. See *ortholog.*

Penetrance: proportion of individuals expressing the phenotype determined by a given mutation or allele.

Phenotype: set of observable characteristics of an individual or a cell. The phenotype applies to different scales from the whole organism to the molecules that make it up.

Phylogeny: study of the kinships between organisms during evolution. Molecular phylogeny is based on comparisons of *orthologous* sequences of proteins or nucleic acids.

Phylum: (plural *phyla*) all past and present species sharing common characteristics considered representative of a common ancestry. Note that the word is also used in taxonomy to precisely designate the second hierarchical rank of the classification of living organisms.

Plasmid: a DNA molecule distinct from chromosomal DNA, capable of autonomous replication and not essential for cell survival. Term used mainly for bacteria, see *episome*.

Ploidy: number of chromosomal sets in a cell or organism. See *aneuploid*, *diploid*, *haploid* and *polyploid*.

Polygenic: refers to a phenotypic trait whose expression depends on several genes. See *monogenic*.

Polypeptide: chain of amino acids linked together by peptidic bonds (covalent bonds between carboxyl and amino radicals of two successive amino acids). Polypeptide synthesis occurs on ribosomes using the genetic information carried by messenger RNAs. See *genetic code* and *protein*.

Polyploid: refers to a nucleus, cell or organism with more than two sets of chromosomes. If they are homologous, we talk about autopolyploidy. If they are different, we talk about allopolyploidy.

Prokaryote: a generally unicellular microorganism whose cell is very small and lacks organelles (the only exception being thylakoids in cyanobacteria) and nuclei (as opposed to eukaryotes). Prokaryotes group bacteria and archaea.

Promoter: region of DNA where RNA transcription is initiated. Transcription can be monodirectional from the promoter (common in bacteria) or, on the contrary, bidirectional (common in eukaryotes). See *terminator*.

Prophase: first stage of eukaryotic cell divisions, mitosis or meiosis, where chromosomes condense. In meiosis, homologous chromosomes pair and exchange segments by recombination.

Protein: biological macromolecule formed from one or more identical or different polypeptide chains. See *polypeptide*.

Pseudogene: traces of a gene in a DNA sequence showing alterations in variable numbers. Pseudogenes have two possible origins: an accumulation of mutations that have inactivated or partially destroyed an old active gene, or an insertion into the genome of a DNA segment resulting from the reverse transcription of an RNA molecule.

QTL: acronym meaning Quantitative Trait Locus: a more or less large region of DNA (locus) whose genetic variation between individuals results in a measurable (quantitative) variation of a continuous phenotypic trait. Human height provides a good example of a measurable phenotypic trait with continuous variation. It is genetically determined by several hundred QTLs.

Reading frame: one of the three reading phases of the messenger RNA sequence used by ribosomes for protein synthesis.

Recessiveness: property of a parental trait that no longer appears in a hybrid and, by extension, of an allele whose phenotype can only be observed in the homozygous state.

Recombination (genetics): the process by which DNA segments from different molecules are assembled into a single new molecule, generating a new allelic combination. Recombination can occur between homologous sequences (case of meiotic recombination between chromosomes) or by joining DNA ends of different segments. Recombination requires DNA to be cut, by specialized proteins or accidentally.

Reductional: refers to the first division of meiosis where homologous chromosomes separate.

Replication: the process of duplication of a DNA molecule during which the two strands separate to form two double stranded DNA molecules identical in sequence. Replication involves complex molecular machinery in which an enzyme called DNA polymerase catalyzes the sequential polymerization of deoxyribonucleotides by following the complementarity of the bases with the DNA matrix strand. Each new DNA molecule is therefore made up of an old strand, inherited from the mother cell, and a newly synthesized strand.

Replicon: a unitary element of DNA replication. The replicon can be a plasmid, a whole chromosome (common in bacteria) or a chromosome segment (general case of eukaryotic nuclear chromosomes). A replicon contains at least one replication origin and, as the case may be, termination sites. Autonomous replicons may also contain genes necessary for their own replication.

Repressor: a molecule, usually a protein, whose action at the level of DNA or chromatin inhibits the transcription of DNA from one or more promoters. See *activator.*

Restriction (enzymatic): cleavage of a DNA molecule by the action of enzymes (endonucleases). Restriction involves cleaving the two DNA strands at or near the same site. It is, therefore, to be distinguished from the cleavage of only one of the two DNA strands, referred to as nicking. The term of DNA restriction is derived from bacteriophage host restriction, a phenomenon that led to the discovery of restriction endonucleases.

Restriction/modification site: short DNA sequence recognized by specific endonucleases or methylases.

Retrogene: a gene or gene fragment originating from the DNA copy of an RNA molecule.

Retrotransposon: a transposon whose propagation involves an RNA intermediate. Retrotransposons are DNA sequences transcribed into RNA molecules that are themselves retro-transcribed into DNA molecules, using reverse transcriptase translated from this RNA, and integrated into the genome using other enzymes translated by this RNA. Retrotransposons move by multiplying in the host genome. Inactive (mutated) traces remain visible in the genomes for some time and can possibly be reactivated by the arrival of new active retrotransposons of the same family. See *transposon*.

Retrovirus: RNA virus whose viral cycle includes a step of retro-transcription of viral RNA into DNA using reverse transcriptase and other enzymes encoded by this RNA. The DNA produced is made double-stranded and then integrated into the genome of the host cell.

Ribosome: very large macromolecular particle composed of RNA and proteins and involved in protein synthesis. Ribosomes are composed of two

independent subunits, one large and one small, which assemble in the presence of a messenger RNA to begin protein synthesis. Each subunit includes a large RNA molecule. The RNA of the large subunit is the enzyme that catalyzes the formation of the peptide bond.

Ribozyme: catalytic enzyme consisting of RNA. Some introns, the RNA of the large ribosome subunit, the RNA of the spliceosomes and others are ribozymes.

RNA: ribonucleic acid. A macromolecule formed by a chain of four ribonucleotides, A, C, G and U. The order of their linkage is the sequence that carries the genetic information. RNA is the carrier of genetic information in some viruses.

RNA editing: post-transcriptional modification of the sequence of an RNA molecule by specific addition or subtraction of nucleotides at specific sites or by chemical transformation from one nitrogenous base to another.

RNA interference: a post-transcriptional mechanism involving specialized small RNA molecules whose action on other RNA molecules interferes with their function or life span.

Sequence: order of succession of elementary components along a polymeric chain: amino acids in the case of proteins, ribonucleotides in the case of RNA and deoxyribonucleotides in the case of DNA. For informatic reasons, the biological sequences are represented by one-letter codes, each corresponding to an elementary component. Thus, DNA sequences are represented by successions of the letters A, C, G and T, RNA sequences by the letters A, C, G and U and protein sequences by the letters A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

Segregation (of chromosomes): ordered separation and migration to opposite poles of the mother cell of the sister chromatids resulting from DNA replication (mitosis) or of the pairs of homologous chromosomes (first division of meiosis).

Singleton: a gene or genetic element present in a single copy in a haploid genome and without a paralog.

SNP: acronym meaning Single Nucleotide Polymorphism: variation (polymorphism) of a *single* base pair (single nucleotide) at a specific locus between individuals of the same species.

Spliceosome: used to designate a macromolecular complex composed of RNA and proteins and involved in the splicing of introns called "spliceosomic". The spliceosome consists of five subunits each containing a small RNA (called U1, U2, U4, U5 and U6).

Splicing (of RNA): a process by which two non-contiguous RNA segments are spliced together to form a single polyribonucleotide chain that is chemically indistinguishable from a primary DNA transcript. The main splicing events concern the elimination of introns.

Sporophyte: diploid phase of the reproductive cycle of Archaeplastidae (plants and algae).

Subgenome: part of a hybrid genome corresponding to one of the parental lines. For example, bread wheat (hexaploid) contains three subgenomes referred to as A, B and D. Note that cytogeneticists tend to call subgenomes "genomes".

Substitution (of a nucleotide): replacement of one nucleotide by another during a mutation. It should be noted that for each nucleotide in a sequence, there are two possible types of substitutions. See *transition* and *transversion*.

Subtraction (of a nucleotide): elimination in a sequence of one or more nucleotides during a mutation (see *indel*).

Synteny: refers to the fact that genetic elements are carried by the same chromosome. In practice, the term also includes the order of these elements along the chromosome. Thus, the term "conservation of synteny" is used to describe chromosome segments of two distinct organisms carrying the same genetic elements in the same order.

Telomere: end of a linear chromosome. Telomeres are made up of particular DNA sequences that are essential for the maintenance of chromosomes, especially during cell divisions.

Terminator: region of a DNA sequence at which transcription into RNA stops. See *promoter*.

Tetrad: set of the four haploid cells resulting from meiosis.

Trait (or characteristic): one of the anatomical, physiological, molecular or behavioral aspects of a living organism that can be characterized.

Trans: term used to designate two genetic elements carried by two DNA molecules. Opposite of *cis*. In practice, trans also refers to the interaction between two genetic elements carried by the same DNA molecule if this interaction involves a gene product and not the gene itself.

Transcription: the process of synthesizing an RNA molecule from a DNA molecule. Transcription involves a complex molecular machinery in which an enzyme, called RNA polymerase, catalyzes the sequential polymerization of ribonucleotides using the complementarity of the bases with the DNA strand serving as matrix.

Transesterification: constant energy exchange between phosphodiester bonds of the same RNA molecule or, more rarely, of two RNA molecules. Transesterifications allow the splicing and editing of RNAs.

Transfer RNA: a small, non-coding RNA molecule (typically 60–80 nucleotides) carrying an anticodon, to which one of the proteinogenic amino acids chemically binds. This binding requires specific enzymes called "tRNA-synthetases".

Transformation (cellular): insertion of foreign DNA into a living cell leading to genetic modification of its progeny by addition or replacement of a gene or gene fragment.

Transgenesis: insertion of foreign DNA into the genome of a living cell leading to the genetic modification of its progeny by adding a gene or gene fragment at any chromosomal locus.

Transition: replacement of a DNA nucleotide by another with a base of the same chemical family, purine or pyrimidine.

Translation: process of synthesizing a polypeptide chain during which the genetic message provided by the nucleotide sequence of a messenger RNA molecule is interpreted by the genetic code in an amino acid sequence in a protein molecule.

Translocation: movement of a chromosome segment to another location of the same chromosome or to another chromosome. Translocations may be reciprocal or not.

Transposon: a DNA fragment that can move from one locus of the genome to another with or without duplicating itself. Autonomous transposons carry in themselves the genes necessary for their movement. Traces of mutated (inactive) transposons remain visible in the genomes for some time and can possibly be reactivated by the arrival of new active transposons of the same family. Note the distinction between Class I transposons, which propagate via RNA (see *retrotransposon*), and Class II transposons (or true transposons), which do not.

Transversion: replacement of a DNA nucleotide by another nucleotide carrying a base of the other chemical family, purine by pyrimidine or its opposite.

Trisomy: presence in a diploid organism of a chromosome in triplicate.

Vector: autonomous replicon capable of transporting a foreign DNA fragment into a host cell and replicating it as part of itself or integrating it into the host genome. There is a wide variety of vectors depending on the type of host cell targeted and the intended use. For practical reasons, these are now mostly artificial constructions.

Virus: particle endowed with genetic continuity, but incapable of activity without the help of a living cell. Alone, the particle is inert. Its genome, consisting of RNA or DNA, only functions after infection of a cell where it multiplies to form new infectious particles or integrates into the host genome. There is a very high diversity of viruses and the number of their particles is such that they probably represent, together, most of the total biological material on the surface of the planet.

Xenobiology: development of life forms different from terrestrial life in biochemical and informational terms.

Zygote: diploid cell resulting from the fusion of two gametes.

References

Académie des sciences (2016). *Les origines du vivant*. Gallimard, Paris.

Buican, D. (1984). *Histoire de la génétique et de l'évolutionnisme en France*. Presses universitaires de France, Paris.

Deutsch, J. (2012). *Le gène, un concept en évolution*. Le Seuil, Paris.

de Duve, C. (1990). *Construire une cellule*. De Boeck InterÉditions, Brussels.

Edelstein, S.J. (2002). *Des gènes aux génomes*. Odile Jacob, Paris.

Jacob, F. (1970). *La logique du vivant*. Gallimard, Paris.

Le Guyader, H. (2012). *Penser l'évolution*. Imprimerie nationale Éditions, Paris.

Morange, M. (1998). *La part des gènes*. Odile Jacob, Paris.

Mukherjee, S. (2016). *Il était une fois le gène*. Flammarion, Paris.

Pääbo, S. (2015). *Néandertal, à la recherche des génomes perdus*. Babel, Paris.

Rostand, J. (1971). *Les étangs à monstres*. Stock, Paris.

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

Index

Trajectories of Genetics, First Edition. Bernard Dujon and Georges Pelletier.

F, G

H

I

K, L

M

N

O

P

R

S

T

U, V, Y, Z

Other titles from

in

Biology

2019

BRAND Gérard
Discovering Odors

BUIS Roger
Biology and Mathematics: History and Challenges