UNIVERSITY OF AMAR TELIDJI - LAGHOUAT

FACULTY OF SCIENCES

****

DEPARTMENT OF AGRICULTURAL SCIENCES

Course notes

Intended for students of the
1st Year Academic Master's Degree :
Specialization : Plant Breeding

PLANT GENOMICS

Directed by :
Dr. Mériem MARFOUA

List of figures

FOREWORD

Plant Genomics is a semester-long module taught in the first semester of the Master's program in Plant Breeding (see appendix 1). It consists of lectures and practical work.

The module is four hours per week (1.5 hours of lectures and 2.5 hours of practical work).

The objectives of this module are to enable students to :

- ✓ Knowledge of the plant genome and its functioning.
- ✓ Relationship between plant genomics and species.

Plant genomics is a branch of botany that lies at the interface of biochemistry, genetics, and molecular biology.

Plant genomics essentially focuses on the study of an entire plant's genetic makeup, including its :

- **Structure :** How the genetic material (DNA) is organized within the plant cell.
- **Function :** How genes are expressed and how they influence the plant's traits and characteristics.
- **Variation :** How the plant genome differs between different species and even within a single species.

Applications of plant genomics :

- Understanding plant growth and development
- Improving crop yields and disease resistance
- Developing new and more nutritious food varieties
- Preserving endangered plant species

Further exploration :

- **Plant Genome Databases :** These databases store and provide access to the complete genetic information of various plant species. https://global-engage.com/
- **Plant Genomics Research Papers :** Scientific publications delve deeper into specific aspects of plant genomics research. You can find them through online academic databases.
- **Biotechnology Companies :** Many companies focus on utilizing plant genomics for agricultural advancements. Their websites might offer insights into practical applications.

INTRODUCTION

Summary

1.- What is Plant Genomics ?

Genomics is the comprehensive study of genomes, encompassing the entire set of **genes**, their **arrangement** on chromosomes, their **sequence**, **function**, and role. The genomes of living organisms are vast, ranging from hundreds of millions to billions of nucleotides (three billion for the human genome).

Plant genomics is a scientific discipline that studies the complete set of genes in a plant (its genome). It allows us to understand how these genes are organized, interact, and influence a plant's characteristics, from growth to disease resistance.

Why study plant genomes ?

- **Crop improvement:** Identify genes responsible for desirable traits (yield, disease resistance, drought tolerance) to develop new, higher-performing varieties.
- **Understanding evolution:** Study the relationships between plant species and trace their evolutionary history.
- **Conservation of biodiversity:** Identify endangered species and implement conservation strategies.
- **Development of biotechnology:** Utilize genomic knowledge to produce valuable molecules (medicines, biofuels).

2.- Key concepts in plant genomics

a- Genetics is the study of heredity. It examines the characteristics that are passed from one generation to the next.

b- A genome is the complete set of genetic material present in an organism. It's essentially the entire blueprint for an organism, containing all of its **genes**.

c- Genomics is the science that studies the entirety of genes. It focuses primarily on the physical mapping of genomes and gene structures, known as structural genomics, which was previously referred to as molecular genetics (locating genes, determining the number of exons and introns, identifying polymorphisms, etc.). Genomics also explores gene functions, known as functional genomics. One approach to studying these functions involves examining gene expression mechanisms, referred to as expression genomics.

d- A **gene** is a segment of DNA that codes for a specific protein. Proteins are the molecules that perform most of the work in cells and are responsible for the traits an organism exhibits.

e- DNA is the molecule that carries genetic information. It consists of four nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G).

f- High-throughput **genome sequencing** has led to a substantial advancement in structural genomics, while "**DNA microarray**" technology has propelled expression

genomics forward. Indeed, the automation of molecular biology techniques now enables the simultaneous study of nearly all transcripts (or messenger RNA) within a given tissue of an individual at a specific moment in time. This comprehensive set of transcripts is referred to as the **transcriptome**.

g- The suffix "**ome**" has been added to the object of study (transcripts) to signify that these studies are conducted on a **high-throughput scale**. Similarly, the development of other automated techniques enabling large-scale protein studies has led to **the study of the proteome (that is, the complete set of proteins)**.

Genomics generates a vast amount of data due to the automated and comprehensive nature of modern techniques. Alongside physiology, it enables an integrated view from genes to **phenotypes**, encompassing transcripts, proteins, **metabolites**, and more—thus referred to as **integrative biology**. This field cannot advance without the systematic storage and holistic analysis of data generated at different levels (transcripts, proteins, physiology). **These steps require the support of bioinformatics, another rapidly expanding science.**

3.- The origins of genomics

A brief look back highlights the speed at which our knowledge of genomes has developed, suggesting that future advancements may be equally remarkable.

- 1953: Discovery of DNA's double-helix structure, the carrier of heredity;

- 1973: Approximate date of the first genetic engineering experiments in laboratories;

- 1977: Development of DNA sequencing techniques;

- 1984: Sequencing of the first genome, the Epstein-Barr **virus**;

- 1990: Official launch of the international Human Genome Project;

- 1991: Creation of the first "**DNA microarray**" by the company AFFYMETRIX;

- 1996: Sequencing of the **yeast** genome;

- 1997: Sequencing of the **bacteria** *Escherichia coli* genome;

- 1998: Sequencing of the genome of the **nematode** *Caenorhabditis elegans*;

- 2000: Sequencing of the genomes of the **fruit fly**, mouse, and the plant *Arabidopsis*;

- 2001: First data on the complete sequencing of the human genome.

These advancements reveal some key insights: genetically, humans are 99.9% identical, but minor differences (0.1%) allow for **genetic fingerprinting** based on DNA. Indeed, two individuals differ by about 3 million nucleotides out of a total of 3 billion. These individual genetic profiles have numerous **applications**, such as paternity testing, forensic investigations, and **tracking** meat products in cattle.

Less than 5% of the genome corresponds to genes that can be expressed. Therefore, transcriptome studies focus on only a small portion of the genome, but certainly the most important part, as it is functional.

4.- Genomic concepts

Genomics relies on sequencing and analysis methods to study the entire genome of organisms.

It includes several subfields, such as **structural** genomics, **functional** genomics, and **comparative** genomics.

5.- Genomic tools

- **Sequencing:** Determination of the order of nitrogenous bases in DNA.
- **Assembly:** Reconstructing sequenced DNA fragments to obtain the complete genome.
- **Annotation:** Identification and characterization of genes within the genome.
- **Bioinformatics:** Use of computational tools to analyze genomic data.

6.- Goals of plant genomics

- Determine the genomic sequences of organisms.
- Define the location of all genes in a genome.
- Annotate all genes in a genome.
- Determine the function of every gene in a genome.
- Establish gene expression profiles of cells under different conditions.
- Identify all proteins that can be produced by a given genome.
- Compare genes and proteins between organisms to determine their evolutionary relationships.

7.- Applications of plant genomics

- **Molecular marking:** Identification of genetic markers linked to traits of interest.
- **Genetic engineering:** Modification of a plant's genome to give it new properties.
- **Omics:** Comprehensive study of different levels of gene expression (transcriptomics, proteomics, metabolomics).

8.- Challenges in plant genomics

- **Genome complexity:** Plant genomes are often larger and more complex than those of animals.
- **Gene-environment interaction:** Gene expression is influenced by environmental factors.
- **Ethics:** Applications of genomics raise ethical questions (GMOs, intellectual property).

In summary, plant genomics is a rapidly expanding field that offers numerous opportunities to enhance agricultural production, preserve biodiversity, and develop new technologies. By understanding the foundations of this discipline, you will be better equipped to address the current and future challenges associated with the study of living organisms.

STRUCTURAL GENOMICS

I

Contents

Structural genomics focuses on the three-dimensional organization of the genome, specifically how DNA is folded and packaged within the cell nucleus. This discipline also investigates the interactions between DNA, proteins, and other molecules, and how these interactions influence gene expression.

Structural genomics centers on studying the physical structure of genomes, including gene mapping, sequencing, and the identification of genetic variations. Understanding the genome is essential for plant improvement, as it allows for the identification of genes of interest (disease resistance, increased yield, tolerance to environmental stress, etc.) and facilitates marker-assisted selection (MAS) and genome editing approaches.

I.1.- Plant genome program: tools for plant genomics studies

The **plant genome program** is a large-scale international effort aimed at sequencing and analyzing the genomes of various plant species. This program has led to the development of numerous valuable tools for plant genomics research.

The plant genome program seeks to sequence, analyze, and understand the genomes of major plant species. These projects are driven by the growing needs of agriculture to improve production, resistance to abiotic (drought, salinity) and biotic (diseases, pests) stresses.

I.1.1- Definition and objectives

A genome program is a large-scale project aimed at sequencing the entire genetic material of an organism, whether it be a virus, bacterium, animal, or plant. In the context of plants, these programs have the following objectives:

- **Sequencing reference genomes:** Obtaining a complete and annotated sequence of the genome of a model species or an agronomically important species, serving as a reference for comparative studies.
- **Identifying genes and their functions:** Discovering new genes, predicting their functions by comparing them to known proteins, and identifying the regulatory regions of the genome.
- **Understanding genome evolution:** Studying the evolutionary relationships between species, identifying genes specific to a species or group of species, and understanding the mechanisms of genome duplication and rearrangement.
- **Improving crop plants:** Identifying genes associated with traits of agronomic interest (yield, quality, disease resistance) to develop new varieties through marker-assisted selection or genetic engineering.

- **Key tools :**

1. **Genome sequencing:** Next-generation sequencing (NGS) technologies such as Illumina, PacBio, and Oxford Nanopore.
2. **Physical and genetic mapping:** Locating genes on chromosomes.
3. **Bioinformatics analysis:** Tools and algorithms for assembling, annotating, and interpreting genomic sequences.

Objectives of genome programs

- Sequencing the genomes of model plants and major crop species (rice, maize, wheat).
- Identifying genes responsible for important agronomic traits.
- Providing genetic and bioinformatics resources to the scientific community.

The need for gene maps and the emergence of high-density molecular marker maps (such as microarray-based maps) have led to systematic genome sequencing efforts. This has provided a physical representation of chromosomes, showing the positions of markers and genes.

Genomics encompasses the analysis involved in locating, isolating, and sequencing genes, as well as studying their function. It's important to note that this approach is not limited to plants.

Some examples of model species

Arabidopsis thaliana has been chosen as a model species for sequencing programs due to its small genome size (100-130 megabases), rapid development cycle (2 months), and self-pollination. Many techniques, such as transgenesis, are easily implemented in this species. Additionally, its genome structure is conserved in related species, making discoveries directly applicable to crops like oilseed rape.

Rice is also a model plant for cultivated grasses, as it has the smallest genome among cultivated grasses. Moreover, it exhibits significant synteny with other species in this genus.

Maize is a more complex case due to its larger genome size. The estimated number of genes is about five times that of *Arabidopsis*.

The gene catalog

To gain access to the knowledge of model plant genomes, an exhaustive inventory of genes has begun using two approaches :

- The first approach is based on the inventory of expressed genes. The characterized sequences are called ESTs (*Expressed Sequence Tags*). However, this approach has its limitations. Indeed, a gene is not always active in a cell. In *Arabidopsis*, it is estimated that only 50% of genes could be identified by this method.

- The second approach focuses on direct genome sequencing. This work is longer but can provide the complete sequence of genes and their positions.

a.- Towards the construction of elite genotypes

Genomics should make it possible to increase knowledge in the plant field and thus to identify numerous plant genes. Thus, the applications are in four areas:

- Providing a large quantity of molecular markers linked to functions or phenotypes. This allows for marker-assisted selection for quality or agronomic value traits.

- Better control of gene regulation.

- Identifying a reservoir of candidate genes for the analysis of QTLs for major agronomic traits.

- Identifying favorable alleles.

To learn more

Genomics is the study of an organism's entire genome, that is, deciphering its DNA and understanding its function. Genetic material is generally organized into chromosomes, which are themselves organized in pairs in the case of diploid organisms: there are then two alleles, that is, two forms of each gene (one from each parent).

Compared to animal genomes, plant genomes are often larger but vary greatly in size: 130 Mb for *Arabidopsis*, 200 Mb for horse chestnut, 400 Mb for rice, 2.4 Gb for maize, 16 Gb for wheat...

Only a small portion of the genome codes for proteins. DNA is transcribed into mRNA, which is itself translated into proteins.

The majority of DNA has sequences whose role is not clearly understood, but which are the subject of intense research (The ENCODE Project Consortium, 2012).

Sequencing a genome and then determining its function makes it possible to know:

- The number of genes
- The organization of the genome
- The identification of regions that have been affected by genetic selection
- The evolution of the genome
- The role of genes
- The influence of external factors such as environmental factors on gene expression.

I.1.2.- Genome characterization using markers

Genetic markers provide information about an individual's genotype and are not modified by the environment. They can be used throughout an experiment and are observable at any stage of plant development and in any organ (the plant's genetic information is contained in its entirety in all cells). The main types of genetic markers used are:

- Biochemical markers (proteins, isozymes): Plant cell proteins can be easily extracted and analyzed. The most commonly used biochemical markers are isozymes. They correspond to different forms of the same enzyme and make it possible to determine the presence of the allele corresponding to each of these forms. In this sense, they are indicators of polymorphism between individuals for the coding sequences of the genome.

- Molecular markers are easily detectable DNA sequences. They are used as tags to study the polymorphism of an organism's DNA: nucleotide substitutions for a given locus or insertion-deletion of longer or shorter segments. They do not correspond to active genes but to sequences associated with active genes.

Applications: Thanks to genetic markers, it becomes possible to:

- Establish the genetic fingerprint of an individual, that is, to describe and define individuals and varieties for the purpose of registration, protection, and classification.

- Highlight and track genes involved in the expression of traits of agronomic or technological interest.

a.- Characteristics of a genetic marker

A good marker must be:

- **Neutral:** Its different alleles have no effect on the individual's phenotype.

- **Polymorphic:** Possessing numerous alleles to characterize different individuals.

- **Codominant:** The heterozygous individual can be distinguished because it simultaneously exhibits the characteristics of its homozygous parents, insensitive to the environment, non-epistatic, multiallelic.

I.1.3.- The genetic fingerprint of a plant

Molecular markers make it possible to establish the genetic fingerprint of a plant. DNA microarrays allow for the rapid visualization and measurement of thousands of expression differences at a single nucleotide level among numerous genes.

Applications: This identification can occur at different levels:

- **Variety identification**. These techniques make it possible to distinguish lines and recognize the parents of hybrids.
- **Variety purity control**. This involves detecting varietal impurities in seed lots, or testing the homogeneity of a population at any stage of a breeding program.
- **Variety protection**. This is an integral part of variety identification. It becomes possible to detect counterfeits or copies of genotypes. Thus, for issues related to plant breeders' rights, the concept of an essentially derived variety (EDV) has been introduced: it is a variety derived from an original variety by modification of a few small chromosomal regions.

I.1.4.- Complementary tools in genomics

Thanks to the miniaturization and automation of genomic tools, the cost and time required to characterize DNA fragments or genomes have been decreasing significantly year after year. However, to obtain biological knowledge from genomic data, it is necessary to perform complex computer analyses as well as numerous measurements and observations on plants. Specialized laboratories called platforms can handle these different aspects.

a.- Bioinformatics platforms

- They integrate data produced by sequencing instruments.
- They perform computer and statistical analysis of data, producing genomic maps or databases, and guide research towards the most interesting active or regulatory genes.

b.- Phenotyping platforms

Phenotyping is the observation and measurement of the phenotype, i.e., the trait of interest. To study the genes that control a trait, such as drought tolerance, it is necessary to measure this trait very precisely on numerous plants using a collection of mutants. By inactivating a gene and observing the physiological consequences, it is possible to correlate a gene with its function.

Phenotyping platforms allow these high-throughput measurements to be performed on hundreds of different plants, thanks to robots that measure, weigh, and photograph plants automatically and perform image analysis.

c.- Transcriptomics, proteomics, and metabolomics platforms

"Omics" technologies are high-throughput techniques that generate large amounts of data at multiple biological levels: from gene sequencing to protein expression and metabolic structures. This data can cover all the mechanisms involved in the variations that occur in cells and that influence the life of the organism.

They include transcriptomics (gene expression and regulation) (see chapter Molecular markers), proteomics (the study of proteins), metabolomics (the study of cellular metabolites produced during cellular reactions), and epigenetics (the study of regulations, sometimes heritable, of gene expression that are not accompanied by changes in the DNA sequence, via methylation or other chemical modifications, or non-coding RNAs).

Scientific studies have already been published on Arabidopsis, a model laboratory plant, but also on cereals, potatoes, and tomatoes. They can complement studies conducted to assess the substantial equivalence between two varieties.

I.1.5.- Molecular markers

a.- RFLP markers

Let's compare two individuals, A and B. Their DNA will be digested separately by given restriction enzymes, and then hybridized with a probe S.

Digestion with enzyme ♦ yields identical restriction fragments for both individuals. The two profiles revealed after electrophoresis do not allow them to be distinguished. With this enzyme ♦ - probe S pair, no polymorphism is revealed.

On the other hand, for enzyme ▼, individual B has a mutation at a restriction site, leading to the loss of this site. Thus, by digestion, individual A gives a smaller fragment than individual B.

We reveal the polymorphism between the two individuals: a fast fragment for A and a slower one for B. This enzyme ▼ - probe S pair reveals a polymorphism.

Creating markers

It is the enzyme/probe pair that constitutes the marker. The DNA fragment used as a probe reveals a polymorphic or monomorphic locus.

Restriction enzymes allow the cutting of genomic DNA and thus the visualization of the number of alleles detectable, in a population, at a given locus. In maize, 95% of loci are polymorphic, whereas in wheat, a self-pollinating plant, only 5 to 10% of loci are polymorphic.

Use of the technique

This molecular marker technique has been widely used, as it provides simple profiles that allow the characterization of a plant's genetic fingerprint or the construction of a genetic map. It is reliable, and the observed results can be repeated; however, it is gradually being abandoned in favor of techniques that are faster to implement and often based on PCR.

RFLP Technique: Restriction Fragment Length Polymorphism

Any modification of DNA sequences (mutation, addition, deletion) frequently reorganizes restriction sites. When restriction enzymes act, the size of the restriction fragments is then modified: a polymorphism is observed.

Steps of the technique

1. DNA extraction: The plant's DNA is extracted.

2. Restriction enzyme digestion: The DNA is subjected to digestion by one or more restriction enzymes. The size of the obtained fragments depends on the enzymes used.

3. Electrophoresis: The fragments are then separated according to their size by electrophoresis. When digesting genomic DNA, a smear is visualized on the gel, as there are a large number of fragments of different sizes that are impossible to separate.

4. Southern blotting: The DNA is transferred by capillary action onto a nylon membrane where it is denatured. This transfer technique allows the relative position of the DNA fragments to be preserved.

5. Hybridization: The membrane is placed in contact with a solution containing a probe labeled either by isotopes or chemically. This probe then hybridizes with the DNA fragment(s) with which it has complementary homology.

6. Detection: The position of the hybridization is revealed by placing the membrane in contact with a sensitive film, or by performing a colored enzymatic reaction (depending on the type of probe used).

Extraction, digestion, transfer, hybridization, and detection constitute the Southern blotting technique.

b.- Microsatellite markers

On the genome, there are sequences composed of repeated units of 1 to 4 nucleotides. These are microsatellites. The most common are (A)n'(TC)n' (TAT)n and (GATA)n', where the values of n can range from a few units to several dozen. These are referred to as tandem repeat sequences or SSRs.

The interest in these microsatellites lies in their polymorphism. This polymorphism is based on the variation in the number of repeat units that constitute the microsatellite.

Steps of the technique

PCR technique is used to reveal microsatellite polymorphism. A pair of primers specific to the right and left borders of a microsatellite is used to amplify the same microsatellite in different individuals. Indeed, each microsatellite is bordered by unique sequences that are specific to it.

The amplification fragments are then revealed by electrophoresis. An individual B, possessing more repeat units than A, has a distinct amplification product that migrates more slowly than A.

Creating markers

The pair of primers specific to the right and left borders of the microsatellite constitutes the marker.

Using the technique

It is necessary to know, synthesize, and test the primers bordering the microsatellite. This technique is simple to use because it relies simply on PCR. It allows for the development of numerous markers, notably for maize or rapeseed. However, it is not applicable to all species; the tomato, for example, does not possess polymorphism for microsatellites.

c. RAPD (Random Amplified Polymorphic DNA) markers

This technique involves performing PCR using a short primer of about ten nucleotides, with an arbitrary sequence. This primer will hybridize randomly within the genome. If two hybridization sites are close and on both DNA strands, amplification will occur, as is the case for individual A.

On the other hand, if these two sites are too far apart, amplification cannot occur, as is the case for individual B. Thus, an additional band is observed on the gel for A and the absence of the same band for B.

The revealed polymorphism is a polymorphism of primer hybridization sites. The primers therefore constitute the markers. For the entire genome, an average of ten fragments are amplified and then separated by electrophoresis.

Use of the technique

This method does not require digestion with a restriction enzyme, transfer to a membrane, or preparation of a radioactive probe. It is therefore rapid and requires little technical expertise. However, its reproducibility is difficult to achieve.

Indeed, the amplification obtained depends greatly on the PCR reaction conditions. Thus, the results are difficult to reproduce from one laboratory to another.

It is used for rapid preliminary analyses, such as the identification of a few markers in the vicinity of a gene of interest.

d. AFLP (Amplified Fragment Length Polymorphism) markers

This technique is based on the combined detection of restriction site polymorphism and polymorphism of hybridization of an arbitrary sequence primer.

Steps of the technique

Plant DNA is subjected to digestion by restriction enzymes. The sizes of the resulting fragments depend on the enzymes used.

Next, nucleotide adapters specific to the restriction enzymes used are added to the ends of the restriction fragments. These adapters have known sequences. The fragments are then amplified by PCR. An oligonucleotide complementary to the adapter sequence, extended by a few arbitrary nucleotides (1 to 3) called overhang bases, is used as a primer. Only fragments possessing bases complementary to these arbitrary bases are amplified. These are selective primers, reducing the number of amplified fragments to about a hundred: without these overhang sequences, thousands of fragments would be amplified. The bands are visualized by electrophoresis.

Creating markers

The combination of restriction enzyme and primer reveals the polymorphism between individuals. This combination constitutes the AFLP marker. The locus identified depends on the sequence of the restriction enzyme site and the arbitrary bases. There are numerous enzyme/primer combinations. Hundreds of restriction enzymes exist, but about ten are generally used.

Analyzing an AFLP profile

In the four lanes, the segregation for resistance to downy mildew in sunflower is shown. In the leftmost lane, it is the resistant parent (R), and in the rightmost lane, the susceptible parent (S). In the two central lanes are the F2 individuals: the second lane from the left corresponds to the mixed DNA of resistant F2 progeny and the third lane to the mixed DNA of susceptible F2 progeny. Among the hundred or so amplification fragments separated by electrophoresis, polymorphic loci can be visualized. It is particularly possible to identify bands present in the susceptible parent and absent in the resistant one, and vice versa. This distinction is also found in the progeny. These loci are therefore linked to the resistance gene by the absence or presence of a band. The other loci either do not reveal polymorphism or reveal polymorphic markers but do not allow the distinction between susceptible and resistant individuals, so they are not linked to resistance.

Use of the technique

It is used in particular for the selection of lines, and for the saturation of a region of the genome in the vicinity of a gene with a view to its cloning, that is to say, to position a large number of markers in this region.

Analyzing an AFLP profile

The four lanes highlight the segregation for sunflower downy mildew resistance. The leftmost lane represents the resistant parent (R), while the rightmost lane represents the susceptible parent (S).

The two central lanes show F2 individuals: the second lane from the left corresponds to a DNA pool of resistant F2 progeny, and the third lane corresponds to a DNA pool of susceptible F2 progeny.

Among the hundred or so amplification fragments separated by electrophoresis, polymorphic loci can be visualized. In particular, it is possible to identify bands present in the susceptible parent and absent in the resistant one, and vice versa.

This distinction is also found in the progeny. These loci are therefore linked to the resistance gene by the absence or presence of a band. The other loci either do not reveal polymorphism or reveal polymorphic markers but do not allow the distinction between susceptible and resistant individuals, so they are not linked to resistance.

Use of the technique

It is used in particular for the selection of lines, and for the saturation of a genomic region in the vicinity of a gene with a view to its cloning, that is to say to position a large number of markers in this region.

e.- DNA Microarrays

DNA microarrays are functional images of the genome. Using thousands of probes, it is possible to finely compare two plants. After laser excitation of the glass slides on which the plant samples have been deposited, the interpretation of the well colors allows the detection of genetic differences between two plants. RFLP markers also allow for the routine differentiation and characterization of varieties on a large number of agronomic traits.

Applications

DNA microarrays allow the study of the transcriptome, that is, the expression of the coding sequences of genes transcribed into messenger RNA (single strand). It is a functional image of the genome, since biochips or DNA microarrays will reveal the coding sequences of the genome that are actually expressed. The technique is based on the hybridization of two complementary nucleic acid fragments or their dissociation under the action of temperature and salt concentration of the medium.

Manufacturing of DNA Microarrays

A DNA microarray is a small, rigid support (glass or nylon) on which short DNA sequences, called probes, are fixed. These probes correspond to specific (unique) fragments of a single gene. Ideally, all the genes studied are deposited on the chip and each position is known. The probes are not labeled. The targets correspond to the messenger RNAs that one wishes to study. For this, after extraction of the messenger RNAs, researchers perform a reverse transcription to obtain complementary DNAs which are labeled with fluorophores (or radioactive markers).

Interaction between the probe and the target

This occurs over 12 hours at approximately 60°C through the interaction of the two complementary sequence strands, which corresponds to molecular hybridization. If the number of probes on the chip is much higher than the number of targets (coming to hybridize on the chip), it is possible to quantify the transcripts present, knowing that the detection signal is proportional to the target concentration.

If there is no signal, it means that molecular hybridization has not occurred.

Use of DNA microarrays

DNA microarrays allow the grouping of genes with the same expression profile under particular experimental conditions. They can, for example, be a powerful tool for the study of abiotic factors such as drought tolerance, which involve many metabolic pathways and therefore many genes, as several thousand, even several million, nucleotide substitutions can be studied simultaneously.

Different types of DNA microarrays

High-density filters (macroarrays) are nylon plates (12x8 cm) that allow the quantification of the presence of 2400 genes by radioactive labeling of a sample (total mRNA to be studied).

Glass slides (microarrays) allow the quantification of the presence of 10,000 genes by fluorescent labeling with two experimental conditions per slide (see infographic). Oligonucleotide microarrays (1.28cm x 1.28cm chip) contain 300,000 oligonucleotides (probes) per slide and allow the quantification of mRNA presence through fluorescent labeling of these targets from a single experimental condition.

I.2.- Physical structure of the nuclear genome in plants

I.2.1.- Definitions

The nuclear genome of plants is organized into several levels of organization, from the smallest to the largest scale: :

a. DNA

- DNA is the molecule that encodes genetic information.
- It consists of two long chains of nucleotides (nitrogenous bases) wound around each other to form a double helix.
- The sequence of nucleotides along the DNA determines the instructions for the construction and function of the organism.

b. Nucleosomes

- DNA is wrapped around proteins called histones to form nucleosomes.
- Nucleosomes are then packed into more compact structures called chromatin fibers.

c. Chromosomes

- Chromatin fibers are condensed into chromosomes.
- Chromosomes are structures visible in the nucleus of cells during cell division.
- The number of chromosomes varies among plant species.

d. Karyotype

- The karyotype is the complete set of chromosomes in a cell.
- The karyotype allows visualization of the structure and number of chromosomes in a plant species.

e. Heterochromatin and euchromatin

- Heterochromatin is a region of chromatin that is condensed and inactive.
- Euchromatin is a region of chromatin that is less condensed and active.
- Heterochromatin is generally located near centromeres and telomeres, while euchromatin is located in intergenic regions.

f. Transposable elements

- Transposable elements are DNA sequences that can move from one place to another in the genome.
- They can play an important role in genome evolution.

e. Non-coding DNA

- Non-coding DNA is DNA that does not code for proteins.
- It can play an important role in regulating gene expression.

In summary, the physical structure of the nuclear genome in plants is complex and dynamic. It is composed of several levels of organization that contribute to the regulation of gene expression and the evolution of the genome.

The **nuclear genome of plants** is contained within the cell nucleus and is characterized by great complexity. It consists of **linear chromosomes** that carry genetic

information in the form of DNA. Plant genomes exhibit a wide range of sizes, from a few hundred million base pairs (e.g., *Arabidopsis thaliana*) to several billion base pairs (e.g., wheat).

Genome elements

- **Coding genes:** DNA sequences that are transcribed into messenger RNA (mRNA) and then translated into proteins.
- **Non-coding regions:** Intergenic regions, introns, repetitive sequences, transposons.
- **Repetitive elements:** Approximately 50% of the plant genome consists of repetitive sequences, often associated with transposons, which contribute to plant evolution and diversity.

I.2.2.- Characteristics of plant genomes

- **Polyploidy:** Many plants exhibit polyploidy (multiplication of chromosome sets), which complicates the study of their genome.
- **Alternative splicing:** Genes can be spliced in different ways to produce multiple proteins from a single gene.

Numerous genomes, including that of humans, have now been sequenced, and many more are in the process of being sequenced. A large number of laboratories, both public and private, are participating (sometimes competitively) in this enormous effort that is revolutionizing both fundamental biology and biotechnology.

The purpose of this document is to provide some data on genome sequencing and its methods. For more information, see also the document on DNA sequencing..

I.2.3.- Sequenced genomes

A reminder about nucleotide sequences

Nucleotides, the elementary building blocks of DNA, can be of four different types in this molecule. They consist of a constant part (sugar-phosphate backbone) and a variable part, a base, from a chemical point of view. The four bases present in DNA are denoted A, T, G, and C (Adenine, Thymine, Guanine, and Cytosine).

The succession of bases along a strand of DNA is the sequence of that strand. We then speak of a nucleotide sequence (Figure below) **(Fig. 1)**.

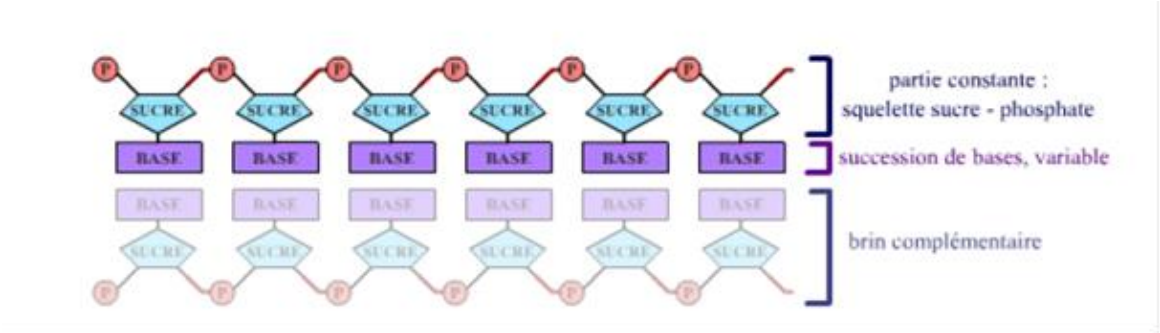


Figure 1 : Nucleotide sequence of a nucleic acid.

Therefore, the size of a sequence can be expressed in the number of bases - kilobases (kb) for thousands of bases, megabases (Mb) for millions of bases, gigabases (Gb) for billions of bases - and the size of a DNA molecule in nucleotides, or in base pairs, to remember that a DNA molecule is formed of two complementary antiparallel strands.

I.3.- Genome sequencing in plants

I.3.1.- What is genome sequencing?

Genome sequencing involves determining the nucleotide sequence of the DNA present in each cell of a given organism.

This determination is generally more difficult the larger the genome studied and the richer it is in repetitive sequences. Viruses, which have small genomes devoid of repetitive sequences (between 3,000 and 150,000 base pairs, often less than 10,000), were the first "organisms" to be sequenced, and still represent the majority of sequenced genomes today. The first bacterium was sequenced in 1995, and many other prokaryotes have since been fully sequenced. The size of their genome is on the order of a few million base pairs (megabases, Mb). The difficulty is quite different for eukaryotic organisms: the large size of their genome (2 to 3 billion base pairs for mammals, for example) requires prior mapping work and often a concerted effort from several sequencing centers. However, "small" eukaryotic genomes (such as that of the paramecium, which is "only" 100 Mb) can now be sequenced without prior mapping by a single large center.

While fully sequenced eukaryotes are still less numerous than prokaryotes and viruses, their number is constantly increasing. Some sequences are in a very fragmentary and incomplete state; others give a much more complete view of the genome. This reflects the sequencing effort undertaken as well as the strategy adopted by the sequencing authors, as we will see later. In March 2014, there were 12,919 completely sequenced cellular genomes (excluding viruses), 27,399 being sequenced, and 996 planned.

Regarding humans, the first sequence of the human genome, announced with great media fanfare at the end of 2000, was by no means a complete sequence. This complete sequence has been available since April 2003, with a few "holes" remaining.

The following list gives some examples of genomes sequenced among the first to be sequenced.

- Viruses: 3,778 viruses sequenced as of March 4, 2014, including HIV (Human Immunodeficiency Virus).

- Prokaryotes:

* Archaea: 319 fully sequenced genomes and 447 partially sequenced as of March 4, 2014.

* Bacteria: 12,286 fully sequenced genomes and 20,403 partially sequenced as of March 4, 2014. Examples: *Escherichia coli*, *Agobacterium tumefaciens*.

* *Haemophilus influenzae* Rd. (First cellular genome sequenced, in 1995)

- Eukaryotes: 314 fully sequenced genomes and 6,660 partially sequenced as of March 4, 2014. Examples (the first 5 listed are the first 5 published, with the reservation mentioned for humans):

* *Saccharomyces cerevisiae* (yeast, the first eukaryote to be sequenced in 1997, several strains sequenced today)

* *Caenorhabditis elegans* (nematode worm)

* *Drosophila melanogaster* (fruit fly)

* *Arabidopsis thaliana* (thale cress, a small plant in the cabbage family)

* *Homo sapiens* (us, the human species: several individuals sequenced including Watson)

* *Neurospora crassa* (ascomycete fungus)

* *Anopheles gambiae* (mosquito)

* *Takifugu rubripes* (fugu, pufferfish consumed in Japan)

* *Mus musculus* (mouse)

* *Plasmodium falciparum* (intracellular parasite responsible for malaria)

* *Oryza sativa* (rice: two subspecies sequenced; *japonica* and *indica*).

Many sequencing projects are currently underway or under consideration. Given the thousands of such projects, it would be too long to list them all here... For more information on this subject, the GOLD (Genome On Line Database) website lists all sequenced genomes, as well as ongoing projects.

Genome sequencing in plants is a rapidly developing field that aims to decipher the sequence of amino acids (nucleotides) that make up plant DNA. This complex and fascinating process allows us to better understand:

- Organization and structure of the plant genome
- Number and function of plant genes
- Evolution of the plant genome
- Relationship between genotype and phenotype

Genome sequencing methods

- **Sanger sequencing :** This historical method is based on dideoxy chain termination.
- **Next-generation sequencing (NGS) :** Newer and faster technologies, such as Illumina sequencing.

Applications of plant genome sequencing

- **Crop improvement:** Developing varieties that are more resistant to diseases, herbicides, and environmental stresses.
- **Development of new biotechnologies:** Production of biofuels, pharmaceuticals, and other plant-based products.
- **Understanding plant biodiversity:** Studying plant evolution and the relationships between species.
- **Preserving endangered plant species:** Identifying genes important for the conservation and restoration of species.

Genome sequencing involves determining the order of nucleotide bases (A, T, C, G) in an organism's DNA. Thanks to next-generation sequencing (NGS) technologies, plant genome sequencing has become faster and more affordable.

Main sequencing technologies

1. **Illumina:** Widely used for high-throughput sequencing. Provides short but highly accurate reads.
2. **PacBio:** Provides long reads, useful for assembling repetitive regions of the genome.
3. **Oxford Nanopore:** Portable technology that also enables long reads.

Applications in plant breeding

- **Sequencing model plants and crop varieties** to identify genes of agronomic interest.
- **Identification of mutations** and polymorphisms to develop molecular markers.
- **Genome mapping** and discovery of new genes involved in stress resistance.

I.3.2.- Genome sequencing: how does it work?

Since the late 1970s and the advent of molecular biology techniques, it has been possible to sequence a strand of DNA, meaning to read the order, or sequence, of the nucleotides that make up this molecule. In essence, this comes down to determining the succession of bases, the only variable part of nucleotides. However, current techniques can only read a maximum of a thousand bases at a time in each sequencing operation. Yet the "sequencable" part of the human genome comprises 2.9 billion base pairs (gigabases, Gb)! It is therefore impossible to read an entire genome in one go. Moreover, it is impossible to manipulate DNA molecules several tens or even hundreds of millions of bases long (the order of magnitude of those that make up human chromosomes).

The basic principle in any genome sequencing is to randomly fragment this genome - or large pieces of DNA derived from the genome - to obtain DNA fragments of a few thousand base pairs, which are easy to manipulate. The ends of a large number of these small fragments are then sequenced. The complete sequence of the genome - or the large piece of the genome - is then reconstructed from these individual sequences, or reads, based on the overlaps between the sequences (if the sequences overlap, it means that the DNA fragments from which they are derived have a portion of their length in common; since the breakage is random, the DNA molecules in the sample are not all broken in the same places) **(Fig. 2)**.

Figure 2 : General principle of genome sequencing.

However, this method presents certain challenges: first, to obtain sufficient overlapping sequences and to minimize sequencing errors, a certain level of redundancy must be achieved, that is to say, a quantity of random sequences representing several times the length of the sequence of interest must be produced. This leads to a very large number of sequences to be generated...

In many sequencing projects, the sequence of 10 times more DNA is thus determined than is contained in the genome studied: this is called a depth of 10X. In this case, each base of the target sequence has been read an average of 10 times, but some have been read more, others less, and still others not at all. Even at 10X, "holes" can therefore remain, leaving the final sequence slightly incomplete.

These parts of the target sequence that are not covered by the randomly performed reads constitute a second difficulty: due to these holes, the result of assembling the overlapping reads does not give a continuous sequence, but several blocks of continuous sequence, or "contigs", which can be difficult at first to orient and order relative to each other, and to assign to a location in the genome. Sequencing more improves the situation, but targeted work may be necessary to fill some gaps.

These difficulties explain why the first sequenced genomes were initially very small genomes: those of viruses. Technical advances (development of automatic sequencers, increased computer power, bioinformatics algorithms for sequence assembly...) have then made it possible to sequence larger and larger genomes: the first bacterial genome (*Haemophilus influenzae*) in 1995, then the first entire eukaryote (*Saccharomyces cerevisiae*) in 1996. The establishment of large sequencing centers, the influx of public or charitable funds and the reduction in sequencing costs, over the past decade, have made it possible to tackle the genomes of higher eukaryotes, including that of humans.

Schematically, two sequencing strategies are currently used:

1. The « *whole genome shotgun* » sequencing strategy ;

2. The « clone-by-clone strategy », or « hierarchical *shotgun* strategy », which involves the prior or concurrent construction of a physical map. This latter strategy was notably used by the international consortium in charge of sequencing the human genome.

In fact, more and more "mixed" strategies are being developed, such as that used by the public consortium responsible for sequencing the mouse genome. The seemingly clear opposition between these two strategies is becoming less and less sharp. For clarity, however, we will explain them separately **(Fig. 3)**.

Figure 3 : Construction of a genomic DNA library.

DNA microarrays and transcriptomics

The availability of complete, annotated genomes enables the creation of DNA microarrays, allowing for the study of the transcriptome beyond the genome itself. DNA microarrays are tools that enable the rapid measurement and visualization of gene expression differences on a whole-genome scale. This involves the direct and qualitative study of the transcriptome, which is the entire set of genetic material expressed in a given cell.

DNA microarrays (which have nothing to do with computers and electronic chips!) are slides coated with probes corresponding to the genes belonging to an organism's genome. Each probe, and therefore each gene, is placed at a specific and identified location on the plate. The microarray then allows for the comparison of gene expression between two cell strains (one strain serves as a control, the other corresponding to the study being conducted): to do this, the RNA from these cells is extracted, reverse-transcribed into DNA, and labeled with a fluorophore (green for one strain, red for the other). The whole is then incubated with the DNA microarray: the cDNA corresponding to the cell RNA hybridizes with the probes carried by the microarray. The microarray is then read, gene by gene, using a laser. We thus "read" three types of genes: (1) genes expressed more strongly in the first strain, whose probes have fixed a greater number of cDNA from the first strain, and which therefore fluoresce essentially in green; (2) genes expressed more strongly in the second strain, which appear rather in red; (3) genes expressed at comparable levels. **(Fig. 4)**.

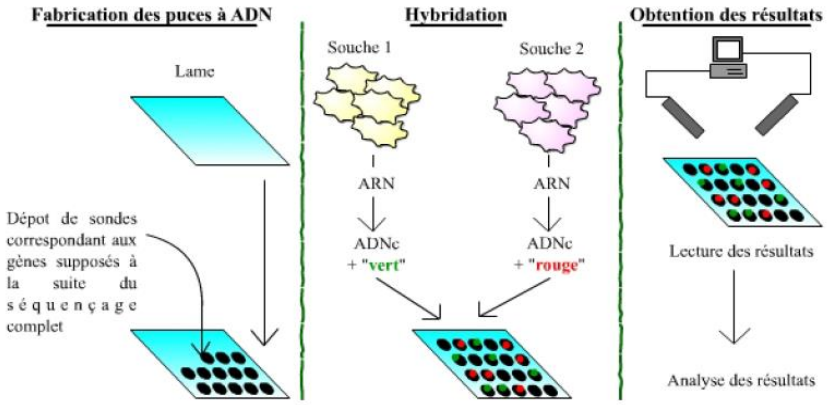


Figure 4 : DNA microarrays.

The three main steps of the process are schematized: microarray fabrication, hybridization, and data acquisition. Note that *this scheme is extremely simplified compared to reality*!

Numerous studies using DNA microarrays are already being conducted in yeast, whose genome has been known for several years. Such studies are expected to expand in humans, as well as for all model species whose genome is fully sequenced.

DNA sequencing, that is, the determination of the succession of nucleotides that compose it, is now a routine technique for biology laboratories. This technique uses the knowledge that has been acquired over the past thirty years about the mechanisms of DNA replication.

Sequencing techniques use specific enzymes: DNA polymerases. These enzymes are capable of synthesizing a complementary DNA strand from a template strand.

These reactions occur through the addition of deoxyribonucleotides (dNTP: deoxy Nucleoside Tri Phosphate). For sequencing, slightly different nucleotides are used: dideoxyribonucleotides (ddNTP). ddNTPs differ from dNTPs by the absence of a specific OH group **(Fig. 5)**.

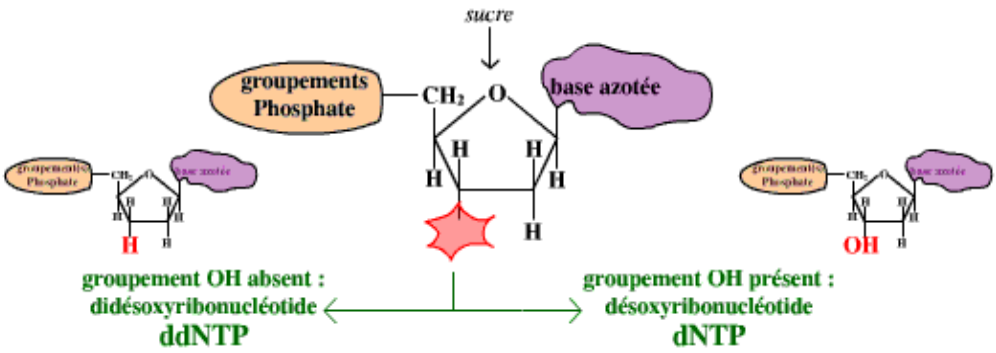


Figure 5 : Two types of triphosphate nucleotides.

Indeed, when a DNA polymerase incorporates a ddNTP instead of a dNTP, it can no longer add any more nucleotides to the chain: the synthesis of the DNA strand is thus halted.

Sequencing techniques rely on this principle. The process is as follows: a DNA polymerase synthesizes the complementary strand of the DNA to be sequenced. In the reaction mixture, there are a large number of dNTPs and a small proportion of a ddNTP (with Adenine, Guanine, Thymine, or Cytosine). At a completely random moment, a ddNTP will be added to the growing chain by the DNA polymerase. This synthesis will therefore stop at that point.

For example, if the reaction mixture contains a small proportion of dideoxyguanosine triphosphate (ddGTP), one will obtain, at the end of the reactions, a set of DNA strands of varying sizes, depending on where a ddGTP has been inserted and the elongation reaction has thus been stopped (which corresponds, due to base complementarity, to the presence of a Cytosine in the sequenced DNA strand). The same operation is repeated with a medium containing ddATP, a medium containing ddCTP, and a medium containing ddTTP **(Fig. 6)**.

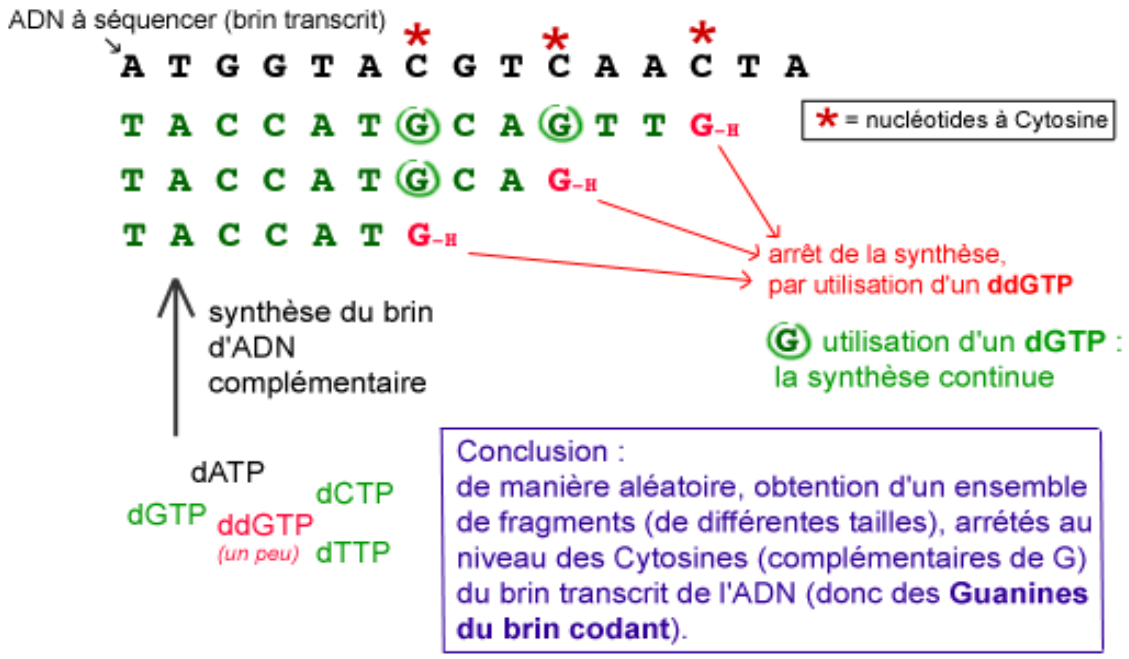


Figure 6 : Use of ddGTP in DNA sequencing.

Using a dideoxynucleotide (here, ddGTP) generates a set of DNA fragments of different sizes, corresponding to the positions of a specific nucleotide.

The sequence can then be "read" by separating these fragments on a gel according to their size. Each band on the gel corresponds to a specific fragment size (accurate to a single nucleotide). To read the four nucleotides of DNA, fragments from the four reaction mixtures (with ddATP, ddCTP, ddGTP, and ddTTP) are separated.

Some technical points...

A DNA polymerase is unable to initiate the synthesis of a DNA strand from scratch. It requires a short DNA fragment, called a primer **(Fig. 7)**. This primer is an oligonucleotide of 15 to 25 nucleotides, complementary to a known DNA sequence located just upstream of the DNA to be sequenced:

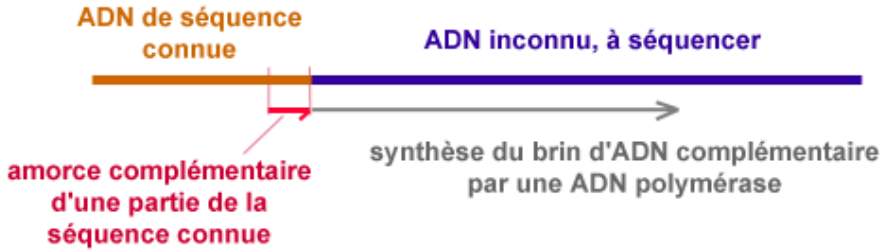


Figure 7 : The synthesis of a DNA strand requires a primer.

Sequencing reactions" (with ddNTPs) are quick (less than 15 minutes). The long part of the protocol is actually the reading of the results. The oldest method, still used in "manual" sequencing (see automation below), involves a 2 to 4-hour migration on an acrylamide gel. After that, one needs to be able to "see" the DNA fragments. To do this, some of the nucleotides used are labeled, either by adding a fluorescent molecule (to the primer used) or by using radioactively labeled nucleotides. The sequence is read either by observing the fluorescence or by exposing a photographic film to the gel (after development, dark bands appear where DNA was present: see the examples of results above, obtained using this method). The readable sequence is limited to approximately 200–300 nucleotides: beyond that, the bands become too crowded, and their order is no longer determinable...

I.3.3.- Sequencing automation

The vast majority of sequences generated and published today are produced using automated sequencers. These devices are capable of performing sequencing reactions and subsequently reading the results **(Fig. 8)**.

Figure 8 : DNA sequencing automation.

To achieve this, DNA fragments are labeled with fluorescent markers. Once the sequencing reaction is complete, the size of the resulting fragments is determined using chromatography. The sequencer detects the fluorescence emitted from the chromatography columns, thereby identifying the DNA fragments and their precise size. The most advanced systems can even read all four nucleotides from a single chromatography column.

The machine outputs the results as curves representing the detected fluorescence, which are then interpreted in terms of nucleotides **(Fig. 9)**. For example, here is a short excerpt from such a curve:

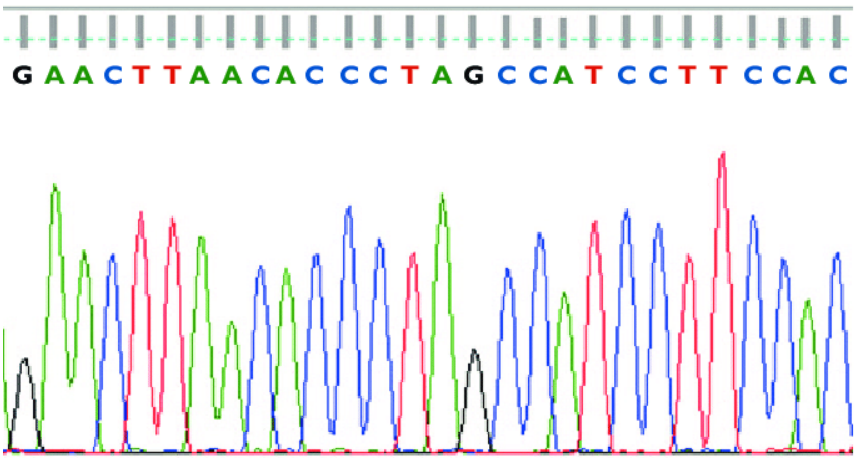


Figure 9: A sample result of automated sequencing.

The four curves represent the detected fluorescence of the obtained DNA fragments. Each peak corresponds to the detection of a specific nucleotide in the sequence; the interpretation is provided below the curves (blue: Adenine, green: Thymine, yellow: Guanine, red: Cytosine).

Automated sequencers offer numerous advantages: automation and the use of chromatography instead of electrophoresis result in significant time savings. The cost per sequence is much lower (once the initial investment in the machine has been recovered). Moreover, while it is difficult to accurately read more than 300 nucleotides using manual sequencing, sequencers can read several hundred nucleotides with very high quality, up to 1000 for the most performant devices! The only limitation to the use of automated sequencers is their high purchase cost, which in practice requires the establishment of shared sequencing services in research institutes.

I.3.4.- Visualizing a DNA fragment

a.- Restriction enzymes

Each restriction enzyme (discovered in bacteria) recognizes a specific DNA sequence, called a restriction site. It cuts both strands of the DNA molecule at this site.

The resulting DNA fragments are called restriction fragments. They are easier to use and identify than a whole DNA molecule. These enzymes are used like biological scissors to create recombinant DNA constructs and to analyze DNA, particularly its variability. Restriction enzymes are easy to use.

Restriction enzymes are the bacteria's defense mechanism against viruses. Each bacterium has its own restriction enzyme, for example, EcoRI for *Escherichia coli*.

b.- Electrophoresis

The action of a restriction enzyme on previously isolated genomic DNA generates restriction fragments of different sizes. These fragments can be separated according to their size on an electrophoresis gel. To do this, the DNA is loaded into wells located at one end of the gel. The gel is then subjected to an electric field. The DNA molecules migrate because they are negatively charged. As they pass through the pores of the gel, they separate according to their size. Thus, the larger molecules are retarded compared to the smaller ones.

DNA fragments can be visualized after electrophoresis using ethidium bromide, for example. This molecule intercalates between the bases of DNA and is fluorescent under ultraviolet light.

In the case of genomic DNA, a population of fragments is generated, whereas in the case of mitochondrial or chloroplast DNA, only a few fragments of a specific size are observed. After migration, the size of the fragments is estimated by comparing the migration distances with those of fragments of known sizes that serve as standards. By comparing the profiles obtained in each well, each corresponding to an individual, differences between individuals can be highlighted, i.e., the polymorphism of their DNA can be visualized, essentially after hybridization.

c.- Molecular hybridization

Molecular hybridization allows for the detection or search of a specific DNA fragment within a mixture of fragments. This method of study is based on two properties of the DNA molecule:

- DNA has the ability to reversibly transition from a double-stranded state to a single-stranded state, which is known as DNA denaturation or, conversely, renaturation. Under the action of temperature (95°C) and/or chemical agents, the bonds between the bases break, and the strands separate. Lowering the temperature causes the strands to renature, a process called hybridization.

- This pairing is not random ; it follows the law of complementarity between DNA bases. This is the second property of DNA molecules : only A-T and G-C pairs form.

The goal of hybridization is to reveal whether a DNA molecule contains a particular sequence or gene (this also applies to RNA). To do this, a probe is used, which is a DNA fragment corresponding to a known sequence.

The probe is generally chemically labeled (most often producing light). By using the phenomenon of renaturation and base pairing, it is possible to detect the DNA fragment containing the complementary sequence.

Description of the hybridization technique

The target DNA molecules to be characterized are first denatured into single strands and fixed onto a nylon hybridization membrane. The labeled single-stranded probe and the target molecule are brought into contact under conditions that allow renaturation. When the probe recognizes its homologue, hybridization occurs. A washing step removes the excess probe. Thus, only the (specific) DNA probe-DNA target hybrids remain labeled. If, as is most often the case, a chemical label producing photons is used, contact with a film will allow the detection of the hybridizations.

In the case of the Southern blot technique, the probe is a DNA fragment and the hybridization will be of the DNA/DNA type. To detect specific RNAs, the Northern blot technique is used for DNA/RNA hybridization.

Probes

Currently, mainly "cold" probes are produced, which use nucleotides labeled through complex chemical reactions that produce photons.

I.3.5.- Amplification of DNA fragments

PCR, Polymerase Chain Reaction, is a technique used to obtain large quantities of a specific DNA sequence from a DNA sample. This amplification is based on the replication of a double-stranded DNA template.

It consists of three phases: a DNA denaturation phase, a hybridization phase, and an elongation phase. The products of each synthesis step serve as templates for the following steps, thus allowing for an exponential amplification of the target sequence, which can then be easily visualized.

Applications

PCR allows for the multiplication of DNA in large quantities, thus opening the way for numerous applications:

- **Diagnosis :** For example, when genetically modifying plants, it is important to determine if the transferred DNA is integrated into the genetic makeup of the cell. PCR allows for a rapid diagnosis of the possible presence of the transferred DNA, on small amounts of DNA.

- **Marking :** PCR itself does not reveal polymorphism, but it leads to the creation of new types of markers that reveal a polymorphism of the amplified fragment. These are easy-to-use and often highly polymorphic markers.

Cloning and gene sequencing are also simplified by this technique.

a.- Denaturation

This is the separation of the two DNA strands at 95 degrees.

b.- Hybridation

The specific primers hybridize to the single-stranded DNA molecules at a given temperature that depends on the primer sequence. Primers are composed of short DNA sequences complementary to the sequence of the DNA to be amplified. It is always a pair of complementary primers flanking the DNA fragment to be amplified.

c.- Elongation

This is the synthesis of the complementary strand. A DNA polymerase enzyme, Taq polymerase (from the bacterium *Thermus aquaticus*), adds nucleotides present in the reaction mixture to the end of the primer, following the sequence of the DNA to be copied (template). A thermocycler, a type of water bath where temperature increases and decreases are programmed and very rapid, allows for a large number of cycles to be performed.

I.3.6.- How does Sanger sequencing work?

In the late 1970s, scientists developed the first two methods for DNA sequencing. Allan Maxam and Walter Gilbert developed chemical sequencing, also known as Maxam-Gilbert sequencing. This method uses chemicals to cleave DNA into smaller fragments in order to determine its sequence. While revolutionary at the time, this method is less efficient than those developed more recently.

In 1977, Frederick Sanger and his colleagues devised another DNA sequencing method. For approximately 40 years, it remained the most widely used method. Even though faster and less expensive methods have been developed since, Sanger sequencing is still extensively used.

Sanger sequencing is based on the process of DNA replication. Scientists create copies of DNA strands. Then, they note the nucleotides added. In this way, they can observe the sequence of nucleotides.

First, a DNA fragment is extracted from the sample. This fragment is then heated, causing the DNA to unwind. The two strands of the double helix separate into individual strands.

The next step is to lower the temperature and add a DNA primer. This is a short single-stranded DNA sequence. The DNA primer attaches to the strand of DNA being sequenced. It indicates the start of sequencing, much like a starting line.

Subsequently, the temperature is slightly increased. Then, free nucleotides and an enzyme called DNA polymerase are added. The free nucleotides contain one of the four following bases: cytosine, thymine, adenine, or guanine. Starting with the primer sequence, DNA polymerase builds a complementary (or reverse) DNA strand. It does this by adding one nucleotide at a time.

Four different sequencing reactions must occur, one for each of the four types of nucleotides. To obtain these reactions, chemically modified versions of each of the nucleotides must be added to the mixture. These versions indicate the termination of a chain. Each of these special nucleotides is labeled with a different dye. Thus, scientists can see them when they expose them to UV rays.

When DNA polymerase reaches a chain-terminating nucleotide, it stops the DNA sequence it was building. DNA polymerase adds the modified nucleotides randomly. Many DNA sequences of different lengths are therefore produced.

Next, the DNA segments undergo gel electrophoresis. This allows the separation of DNA fragments of different lengths. To do this, the DNA fragments must be added to a gel matrix, and then an electric current is passed through it. This causes the segments to align in the gel according to their size. Small fragments travel further than large ones. When the fragments have finished moving, the gel is examined using a radiography device or UV light. Thus, each segment becomes visible.

The gel can be "read" by looking at the dark bands in each column. There is one column for each type of nucleotide (G, C, A, T). By examining the sequence of the bands, the sequence of nucleotides can be determined **(Fig. 10)**.

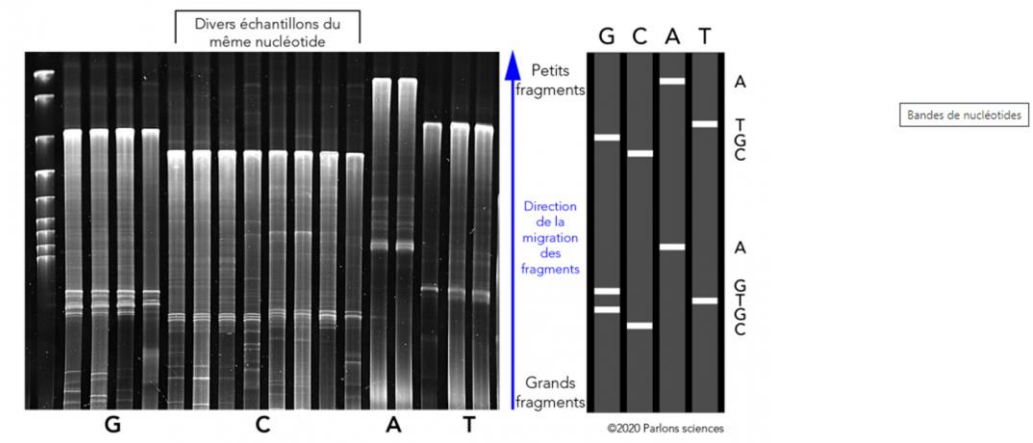


Figure 10: Left: autoradiogram showing the columns and bands for the four nucleotides. Right: the bands and how they can be used to determine the nucleotide order.

Scientists have modified the method developed by Sanger. Fluorochromes are now used. These are small chemical compounds that emit colored light. A different colored fluorochrome is added to each nucleotide. This allows sequencing to be performed in a single reaction mixture in a capillary tube.

To determine the sequence, the tube is illuminated with a laser. When light passes through each colored band, a detector records a peak on a graph. These peaks correspond to the nucleotide bases. With the use of fluorochromes, scientists can automate DNA sequencing. This speeds up the process **(Fig. 11)**.

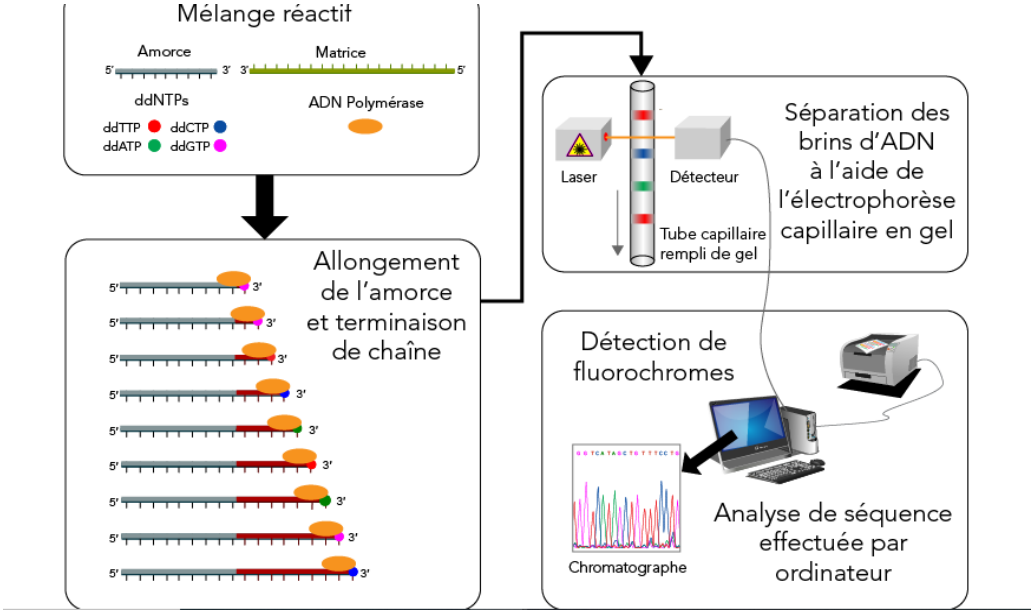


Figure 11: Capillary gel electrophoresis DNA sequencing.

Sanger sequencing has enabled scientists to sequence the DNA of numerous organisms, from bacteria to humans. Using fluorochromes and a computer, it is now possible to sequence a person's entire genome (3 billion base pairs) in just a few days. In the past, the same process took years!

I.4.- Genomics and bioinformatics

Genomics is the study of genomes, which are the complete set of DNA within an organism. **Bioinformatics** is the application of computer science to the analysis of biological data. These two disciplines are closely intertwined and mutually reinforcing.

Bioinformatics is essential to genomics for:

- **Assembling DNA sequences** obtained from high-throughput sequencing.
- **Annotating genomes,** that is, identifying genes and other functional elements.
- **Analyzing gene expression** and identifying gene regulatory networks.
- **Comparing genomes** with each other and identifying regions of conservation and divergence.
- **Developing new tools and methods** for genomic analysis.

Genomics provides bioinformatics with:

- **Large amounts of data** to analyze and interpret.
- **New challenges to address,** such as the analysis of non-coding data and the integration of data from different sources.
- **New applications** in fields such as medicine, agriculture, and the environment.

Concrete examples of applications of genomics and bioinformatics:

- **Human genome sequencing** has allowed the identification of genetic mutations responsible for many diseases.
- **Comparative genomics** has led to a better understanding of species evolution.
- **Bioinformatics applied to medicine** enables the development of new drugs and more accurate disease diagnosis.
- **Genomics applied to agriculture** allows the development of crops that are more resistant to diseases and environmental stresses.

Bioinformatics tools

The analysis of genomic data requires the use of powerful bioinformatics tools. Among the most commonly used are :

- **BLAST (Basic Local Alignment Search Tool):** Allows the comparison of a nucleotide or protein sequence to a database to identify similar sequences.
- **Geneious:** An integrated platform for assembling, annotating, and visualizing genomes.
- **BWA (Burrows-Wheeler Aligner):** Aligns short read sequences obtained by NGS to a reference genome.
- **SAMtools:** A set of tools for manipulating and analyzing SAM/BAM files containing sequence alignments**.**
- **GATK (Genome Analysis Toolkit):** A suite of tools for the analysis of genetic variation data.
- **Galaxy:** An open bioinformatics platform for analyzing genomic data.
- **Ensembl Plants:** A database for exploring plant genomes and their annotations.

Bioinformatics is essential for analyzing the vast amounts of data generated by genome sequencing. It employs software and algorithms to assemble sequences, predict genes, and compare genomes.

Main steps in bioinformatics analysis:

- **Genome assembly:** Reconstructing the genome from sequencing reads.
- **Gene annotation:** Identifying and predicting coding and non-coding genes.
- **Comparative analysis:** Comparing genomes from different species to study genetic similarities and differences.

Genomics and bioinformatics are two complementary disciplines that have a major impact on our understanding of life and the development of new technologies. Their collaboration is essential to meet the challenges of the 21st century.

I.5.- Databases

I.5.1.- Main databases

What are the main databases dedicated to plant genomes? (NCBI, Ensembl Plants)

A **database**, usually abbreviated as DB, is a structured and organized set of data that allows for the storage of large amounts of information in order to facilitate its use (adding, updating, searching for data).

A database is physically represented by a set of files located on a mass storage device (often a hard drive C).

Some databases can be accessed via networks, in which case they are called online databases.

It is important to know that there are genomic databases and protein databases **(Fig. 12)**. The three main nucleic acid databases are:

1. GenBank at NCBI (**N**ational **C**enter for **B**iotechnology **I**nformation): http://www.ncbi.nlm.nih.gov/ Created by IntelliGenetics in 1982. Up until October 2004, it contained 38 941 263 entries (or sequences per author).

2. EMBL at EMBO (**E**uropean **M**olecular **B**iology **O**rganization): http://www.ebi.ac.uk/embl/. The EMBL database contained 44 538 943 entries up until October 2004.

3. DDBJ: **D**NA **D**ata **B**ank of **J**apan: http://www.ddbj.nig.ac.jp/searches-e.html Created in 1986 and distributed by NIG (National Institute of Genetics, Japan). In October 2004, it contained 37 926 117 entries.

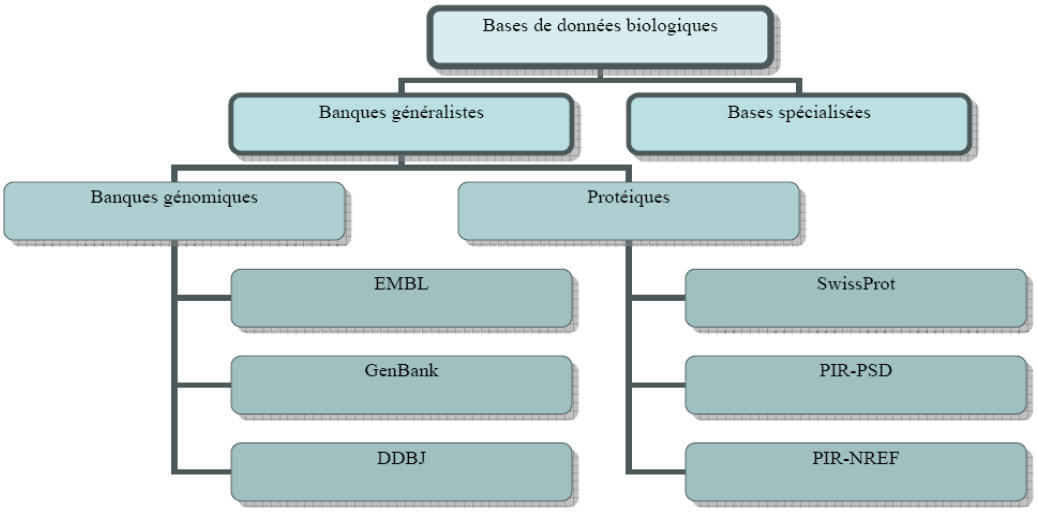
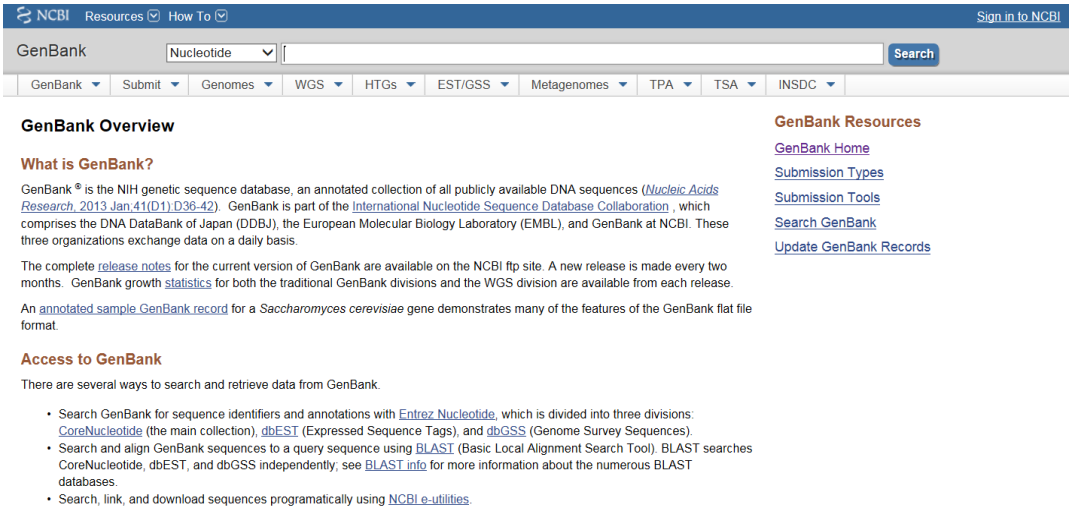


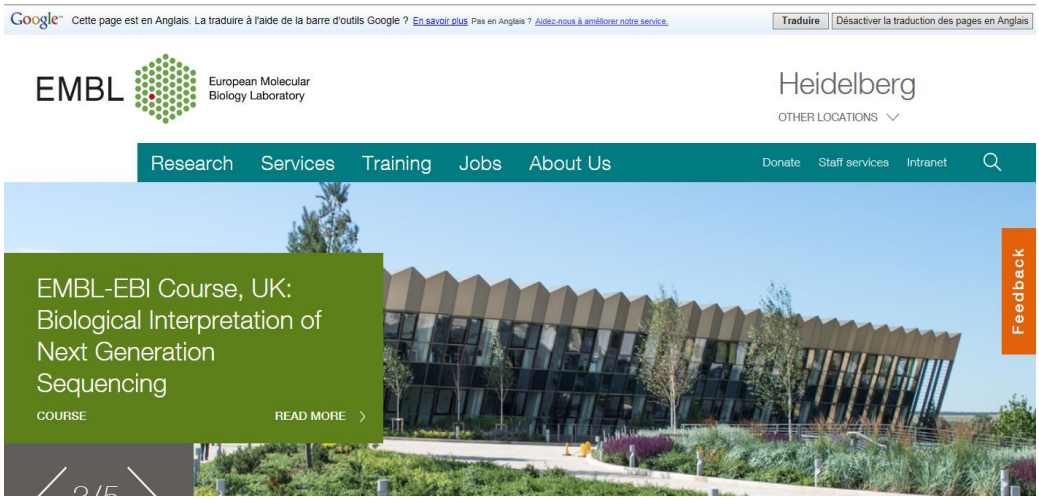
Figure 12 : A few examples of databases.

The interfaces of some nucleic acid databases

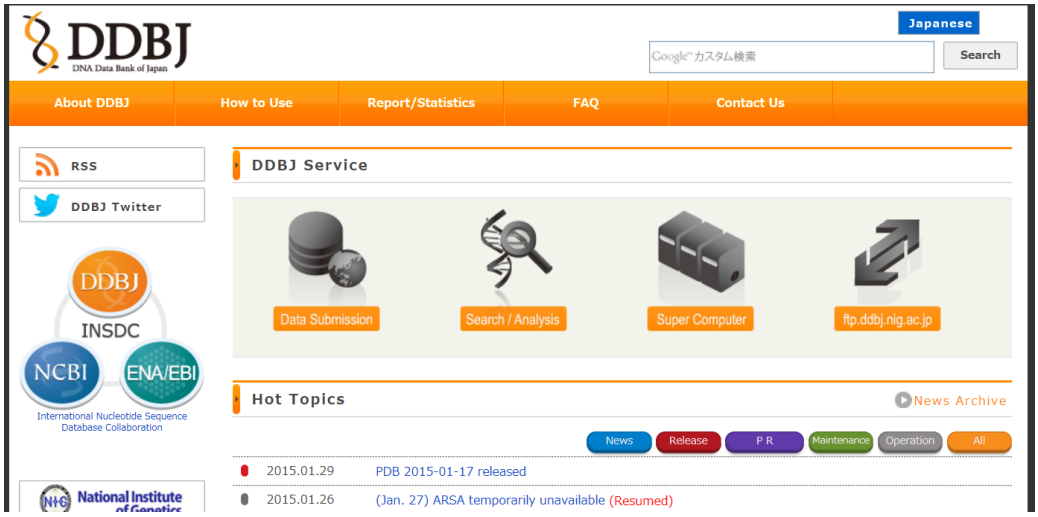
- **GenBank** : http://www.ncbi.nlm.nih.gov/genbank/



- **EMBL** : http://www.embl.de/



- **DDBJ :** http://www.ddbj.nig.ac.jp/



I.5.2.- Genomic data formats

What file formats are used to store genomic data? (FASTA, GFF, VCF)

Genomic data is stored in various file formats, each with its own specific characteristics and uses. Here's an explanation of the most common formats you mentioned:

- **FASTA**

Format: Text

Content: Nucleotide (DNA or RNA) or protein sequences.

Structure: Each sequence begins with an identifier (starting with ">" followed by a description). The nucleotides or amino acids are then listed on subsequent lines.

Use: Widely used to represent individual sequences, such as genes, entire genomes, or sequence fragments.

- **GFF (General Feature Format)**

Format: Tab-separated text

Content: Annotations on a genome, such as the position of genes, exons, introns, promoters, etc.

Structure: Each line represents a feature with specific fields for the identifier, sequence, feature type, start and end coordinates, etc.

Use: To store detailed information about the structure and organization of a genome.

- **VCF (Variant Call Format)**

Format: Tab-separated text

Content: Genetic variations compared to a reference sequence.

Structure: Each line represents a variation with fields for the chromosome, position, variation name, alternative alleles, quality, etc.

Use: To store the results of genomic variation analysis, such as those obtained from next-generation sequencing.

Other commonly used formats:

- **FASTQ:** Similar to FASTA, but also includes quality scores for each base, used for high-throughput sequencing data.
- **SAM/BAM:** For aligning read sequences to a reference genome.
- **BED:** A simpler format than GFF, often used to represent genomic regions.
- **GTF:** A variant of GFF, more specific to gene annotation.

Choosing the right format:

The choice of format depends on the specific application. For example:

- **FASTA** is ideal for storing raw sequences.
- **GFF/GTF** is used to store detailed annotations about genes.
- **VCF** is used to store information about genetic variations.

I.5.3.- Accessing genomic data

How can I access and retrieve available genomic data?

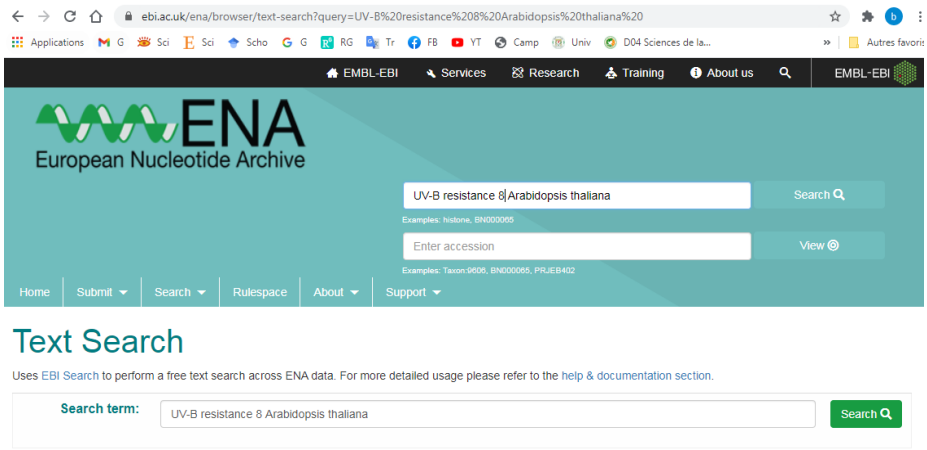
For example, let's do the following practical exercise: search the EMBL database to find the primary structure of the human alkaline phosphatase gene.

Steps:

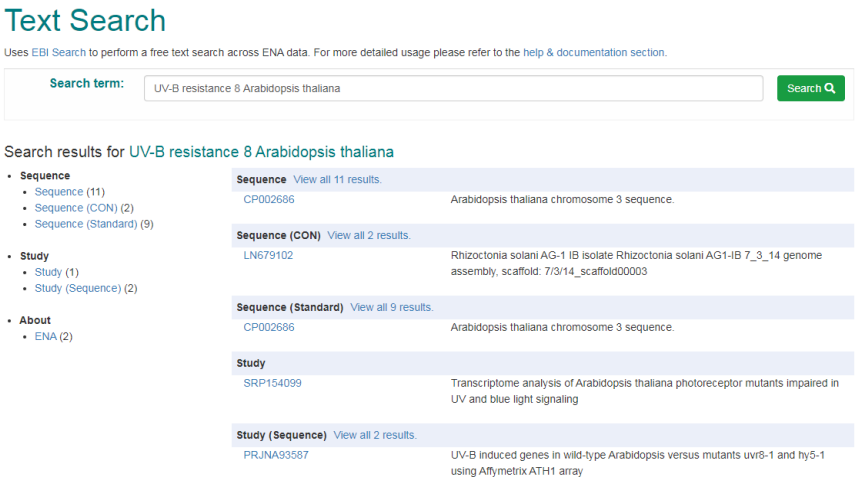
1. Go to your web browser and enter the following **EMBL** address : http://www.ebi.ac.uk/ena/
2. In the search bar, type "**ultraviolet-B receptor (UVR8)**" or "**UV-B resistance 8**" (always in English) and click the **Search** button.

Note: UVR8 has been identified as a crucial mediator of the *Arabidopsis thaliana* plant's response to UV-B. This plant exhibits hypersensitivity to UV-B.

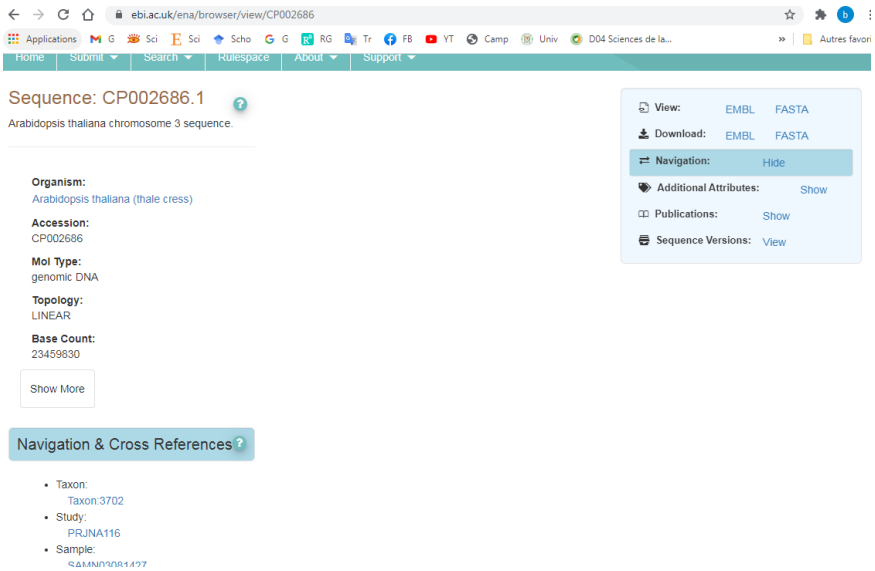
3. A new page will appear with suggested results:



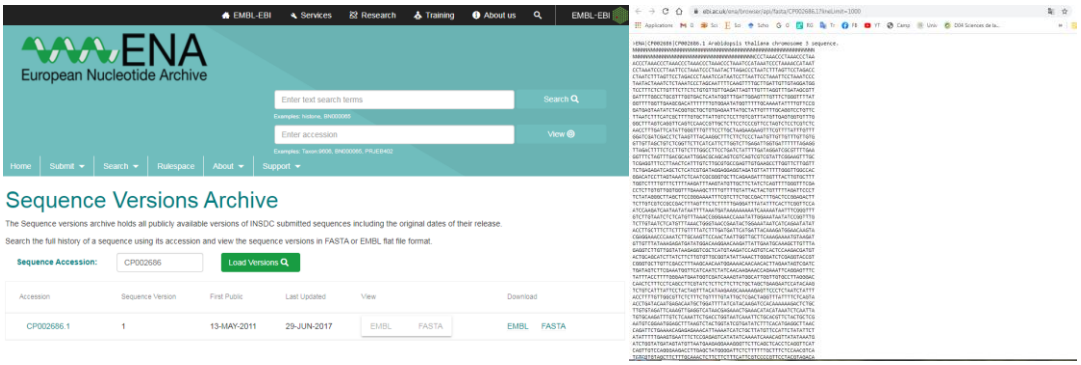
Within this interface, EMBL suggests scientific articles that were consulted to answer our query.



4. Click on one of these results. For example, click on the accession number "**CP002686**". This will take you to the sequence details page :



5. On this page, you can explore various options to study this sequence. For example, click on "**Sequence versions**" or "**view FASTA**" to see the sequence itself. This will take you to the following entry page :



Individual application 1:

Repeat the same process for the plant enzyme **Tyrosinase**. This enzyme is responsible for the browning of fruits and vegetables when exposed to air.

Genomic databases play a crucial role in storing, managing, and accessing plant genetic information. They provide researchers with access to genomic sequences, gene annotations, and polymorphisms.

Main databases for plant genomics:

- **GenBank**: A public database containing DNA sequences from thousands of organisms, including plants.
- **TAIR (The Arabidopsis Information Resource)**: A comprehensive resource dedicated to *Arabidopsis thaliana*.
- **Gramene**: A comparative database for cereals (rice, maize, wheat).

I.6.- Gene prediction

Gene prediction is the process of identifying DNA regions that encode proteins. It is a crucial step in genome annotation and understanding how organisms' function.

I.6.1.- What is genome annotation?

Initially, access to information contained within a genomic text was achieved through experimental genetic techniques that compared altered types (mutants) to the normal type (wild-type) based on observable characteristics (phenotypes). At the time, the observed phenotypes did not have a concrete physical reality. Today, we have direct access to this physical support, DNA (and thus an organism's genotype), where the sequence of a DNA region can be obtained even before characterizing the genes located within it.

Genome sequence annotation can be approached at different levels of analysis, with an initial, essential phase consisting of identifying the organism's genes, that is, locating their precise position on the genome sequence. In a second step, we then seek to assign one (or more) biological functions to each of these hypothetical genes. This second step is generally carried out by comparing the sequences of the hypothetical genes with the sequences of genes of known function.

Organisms whose genomes are now fully sequenced have revealed to us that nearly 40% of the genes identified during the first step have no assigned function, either because they do not resemble any known gene, or (for nearly half of them) because they resemble other genes but those genes themselves have unknown functions. With the sequencing of new genomes, the proportion of genes constituting 'conserved families' of unknown function tends to increase, suggesting that they fulfill important biological functions **(Fig. 13)**.

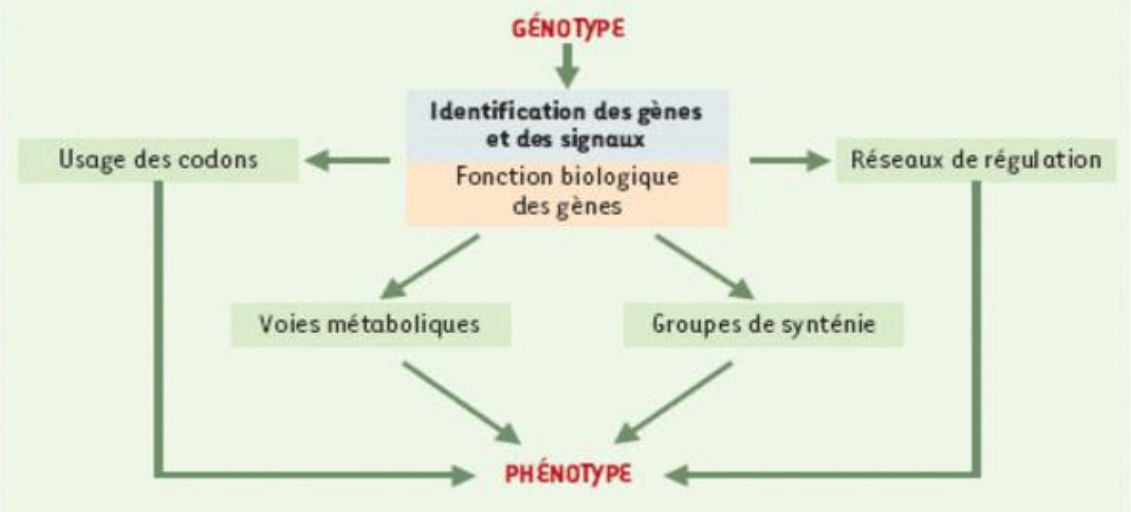


Figure 13: In silico genome annotation strategies.

Methods :

- **Similarity-based methods:** Compare the DNA sequence to known DNA sequences of genes.
- **Motif-based methods:** Search for DNA sequence motifs that are characteristic of genes.
- **Ab initio methods:** Predict genes using statistical models and information about gene structure.

Challenges :

- **Distinguishing genes from non-coding regions:** Non-coding DNA can resemble coding DNA.
- **Identifying exons and introns:** Genes are often composed of exons (coding) and introns (non-coding).
- **Predicting gene function:** The DNA sequence does not always provide a clear indication of a gene's function.

Applications :

- **Developing new drugs:** Identifying genes responsible for diseases and developing drugs to target them.
- **Improving crops:** Identifying genes that control important traits for agriculture, such as disease resistance and yield.
- **Understanding evolution:** Studying the evolution of genes and genomes.

Why annotate sequences?

Sequence annotation is a crucial step in genomics that assigns meaning to the raw data obtained through sequencing. In simpler terms, it's like translating a text written in an unknown language into one we understand.

- **Understanding gene function:** By identifying genes and their functions, we can better comprehend the biological mechanisms underlying life, diseases, and evolution.
- **Developing new drugs:** Knowledge of genes involved in diseases allows for the design of more targeted and effective treatments.
- **Improving crops:** By identifying genes responsible for important agronomic traits (disease resistance, yield, etc.), we can develop higher-performing plant varieties.
- **Studying evolution:** By comparing the genomes of different species, we can trace evolutionary history and understand the mechanisms of biodiversity.
- **Developing new technologies:** Sequence annotation is essential for many biotechnological applications, such as protein synthesis, creating genetically modified organisms, and bioinformatics.

What is annotated in a sequence?

- **Genes:** Identifying DNA regions that code for proteins.
- **Regulatory elements:** Identifying sequences that control gene expression.
- **Non-coding RNAs:** Identifying RNAs that do not code for proteins but have other important functions.
- **Epigenetic modification sites:** Identifying sites where DNA is chemically modified, which can affect gene expression.

How is sequence annotation performed?

Annotation is a complex process combining experimental methods and bio-informatics tools. It generally involves the following steps:

1. **Sequence assembly:** Short sequences obtained from sequencing are assembled to reconstruct the longer sequences of chromosomes.
2. **Gene prediction:** Computer algorithms are used to identify DNA regions that code for proteins.
3. **Functional annotation:** Predicted sequences are compared to databases of known proteins to assign a potential function to genes.
4. **Experimental validation:** Bioinformatics predictions are often confirmed by laboratory experiments.

Sequence annotation is a constantly evolving field of research, with new methods and tools being developed regularly to improve the accuracy and speed of this process.

I.7.- Protein sequence comparison

Protein sequence comparison is the process of analyzing the similarities and differences between the amino acid sequences of two or more proteins. It's an essential tool in biological and biochemical research.

Objectives :

- Determining Protein Homology: Identifying proteins that share a common ancestor.
- Studying Protein Evolution: Understanding how proteins have evolved over time.
- Predicting Protein Structure and Function: Using information from homologous proteins to predict the structure and function of new proteins.
- Identifying Protein-Protein Interactions: Comparing protein sequences to identify regions responsible for interactions between proteins.

Méthodes :

- **Determining protein homology:** Identifying proteins that share a common ancestor.
- **Studying protein evolution:** Understanding how proteins have evolved over time.
- **Predicting protein structure and function:** Using information from homologous proteins to predict the structure and function of new proteins.
- **Identifying protein-protein interactions**: Comparing protein sequences to identify regions responsible for interactions between proteins**.**

Applications :

- **Drug development:** Identifying drug targets and developing drugs that interact with proteins.
- **Crop improvement:** Identifying genes that control important traits for agriculture, such as disease resistance and yield.
- **Understanding diseases:** Identifying genetic mutations responsible for diseases.
- **Developing new technologies:** Developing biomaterials and biotechnologies based on proteins.

Tools :

- **BLAST :** https://blast.ncbi.nlm.nih.gov/Blast.cgi
- **ClustalW :** https://www.ebi.ac.uk/Tools/msa/clustalw2/
- **MUSCLE :** https://www.ebi.ac.uk/Tools/msa/muscle/

Conclusion:

Protein sequence comparison is a rapidly growing field with the potential to revolutionize our understanding of biology and medicine.

FUNCTIONAL GENOMICS

II

Contenent

Functional genomics is an exciting branch of genomics that focuses on understanding how genes function and influence organisms. Often referred to as "high-throughput genetics," it involves the simultaneous analysis of numerous genes. Below are the key points of functional genomics:

Objectives:

- **Determining Gene Functions**: Despite having sequenced genomes, the functions of many identified genes remain unknown. Functional genomics helps associate specific genes with biological roles, such as involvement in cellular processes or protein synthesis.
- **Analyzing Gene Expression**: Understanding when and where a gene is active in an organism is crucial. This includes studying expression variations across tissues, developmental stages, or environmental conditions.
- **Identifying Gene Interactions**: Genes do not function in isolation. Functional genomics explores interactions between genes to understand their collective influence on cellular and organismal functions.

Techniques used:

- **High-throughput Sequencing:** Enables the simultaneous sequencing of numerous genes.
- **DNA Microarrays:** Analyze the expression of thousands of genes at once.
- **RNA Sequencing (RNA-seq):** A newer, more precise technique for studying gene expression.
- **Mutation and Genetic Polymorphism Analysis:** Links genetic variations to specific phenotypes.

Applications:

- **Drug Development:** Identifying disease-related genes helps create targeted treatments.
- **Crop Improvement:** Functional genomics identifies genes responsible for desirable agricultural traits, such as disease resistance and yield.
- **Biotechnology and Bioproduction:** Enhances the production of valuable molecules (e.g., enzymes, biofuels).
- **Personalized Medicine:** Tailors medical treatments to a patient’s genetic profile.

Functional genomics is a rapidly evolving field of research with the potential to revolutionize our understanding of life and to develop new solutions for health, agriculture, and the environment.

Annotation efforts represent an initial step in identifying the functions of genes within the genome of a plant species. Additional tools are required for the precise determination of these genes' functions.

Annotation and gene function identification

a. Positional cloning

In species like *Arabidopsis thaliana*, which has been extensively studied, less than 50% of its genes have known functions. A universal approach for identifying the molecular basis of a hereditary trait is positional cloning. This method relies on extremely fine mapping of the target gene in relation to nearby molecular markers. These markers act as "mileposts" to precisely locate the gene within the DNA sequence. At URGV, this approach was employed to identify the *VAT* gene, which confers resistance to aphids in melon. A resistant genotype and a susceptible genotype were crossed, and segregation analysis of the resistance locus in the progeny allowed the construction of a DNA map near *VAT*. The *VAT* gene was identified as an NBS-LRR type, belonging to the large family of resistance genes against pathogens. This strategy is particularly well-suited for identifying genes responsible for agronomic traits in cultivated species whose genomes have not yet been sequenced. Once identified, these genes can be relatively easily combined within the same genotype to create elite varieties. For instance, these varieties might simultaneously carry genes for pathogen resistance and genes controlling fruit sugar content. This process is referred to as marker-assisted selection (MAS).

b. Gene tagging

Positional cloning is a labor-intensive method and poorly suited to high-throughput techniques. To analyze the function of a large number of genes, alternative tools are required. One such approach involves generating a collection of mutants through the random insertion of a known DNA sequence into the genome. This method, developed at INRA, takes advantage of high-throughput transformation techniques using *Agrobacterium tumefaciens*. This bacterium naturally inserts a fragment of its genome, called T-DNA, into plant cells. A collection of 50,000 *Arabidopsis thaliana* lines, each carrying a known T-DNA fragment in its genome, was generated. The collection was screened for mutants exhibiting alterations in various functions, such as reproductive systems, seed filling, pathogen resistance, and adaptation to water stress. When a mutant from the collection shows a defect in one of these functions, the target gene of the mutation can be easily identified because it is "tagged" by the inserted T-DNA.

c. Le TILLING

TILLING (Targeting Induced Local Lesions IN Genomes) is a reverse genetics approach (from gene sequence to function) that is particularly well-suited for studying gene function in cultivated plants. The principle is straightforward: mutations are induced in the genome of a homozygous plant through chemical mutagenesis of its seeds. In the second generation, the families resulting from mutagenesis are analyzed for mutations affecting the gene of interest. Instead of sequencing the gene in multiple plants from each family, which would be laborious and expensive, the DNA of each family is hybridized with an unmutated reference DNA. If a sequence difference exists between the test DNA and the reference DNA, the resulting hybrid DNA will exhibit a mismatch, which can be detected by a highly sensitive biochemical method. This technique facilitates the creation of a collection of mutations affecting a specific gene. The plants identified as mutants are subsequently studied for their phenotypes. The advantages of this technique are twofold: It works with non-GMO populations, eliminating the need for confined greenhouse facilities. The mutagenesis efficiency is independent of the genome size under study. For example, we have developed TILLING programs for *Pisum sativum* (pea), which is both resistant to transformation by *Agrobacterium* and possesses a very large genome.

d. Transcriptome analysis

A fourth approach to identify the function of a gene involves analyzing its expression characteristics. This is traditionally done using a "reporter gene" technique, which requires the production of transgenic plants. However, this method is considered "low-throughput." An alternative, known as transcriptome analysis, uses DNA microarrays. The 25,000 genes of *Arabidopsis* are individually deposited onto a glass slide (the DNA chip). The slides are then hybridized with preparations of DNA copies of the transcripts produced by different plant tissues. These DNA copies, labeled with a fluorochrome, hybridize with the genes on the slide. A reading device measures the intensity of hybridization signals for each gene, with the signal being stronger the more actively the gene is expressed. The analysis performed on various tissues and under different experimental conditions creates a "profiling" of the expression characteristics of each of the 25,000 genes in the Arabidopsis genome. These profiles are then compared and grouped into functional classes. Genes within these classes, whose functions are already known, allow for predictions about the functions of unknown genes with similar expression characteristics.

e. Convergent methods for identifying the possible role of genes with unknown functions

A fifth approach to analyzing the function of a gene involves studying the protein encoded by this gene. This protein can be localized, for example, using fluorescence techniques, in the cellular compartment where it performs its role. At the same time, it is possible to determine with which other proteins this protein interacts, using so-called "two-hybrid" methods, typically conducted in yeast.

The combination of all the data obtained helps to substantiate the hypothesis regarding the function of the gene under study. Mutation analysis, transcriptome analysis, identification of protein partners, and intracellular localization all contribute to describing the function of the gene. In our laboratory, this combined approach is used to analyze the PPR family, a multigene family nearly absent in the animal kingdom, whose role is to control the establishment of plastid and mitochondrial gene expression functions (such as splicing of transcripts, RNA sequence editing, and regulation of transcription of plastid or mitochondrial genes).

Conclusion

This brief overview of genomic approaches used to study plant genomes is incomplete. Every year, new tools for genome analysis are developed, adding to the existing methods (for example, tools for proteome analysis and next-generation sequencing techniques). This array of research, also supported by biodiversity analysis techniques of sequences, provides valuable insights into genome evolution and domestication processes. Genomics offers new tools to study plant biodiversity and to create organized and exploitable genetic resources, which is a crucial step in genetic improvement efforts. These tools profoundly renew the approach to the improvement of cultivated plants (such as introgression of traits from related species, marker-assisted selection, and the generation of new alleles). Plant genomics opens up a vast field of research and applications, with many emerging countries actively investing in it (Brazil, India, China, Mexico…). We hope that this will also be the case for a country that remains the third-largest global exporter in the seed industry.

II.1.- Plant transcriptome analysis

Structural genomics analyzes the structure of genes and other parts of the genome. Functional genomics focuses on the function of genes and other genomic elements. It includes the analysis of the transcriptome (messenger RNAs or mRNAs), also known as transcriptomics.

Plant transcriptome analysis focuses on the entire set of messenger RNAs (mRNAs) present in a plant at a given time. It is a powerful tool for decoding the complexity of gene expression and its influence on development, physiology, and response to environmental stimuli. This analysis provides crucial insights into how plants regulate gene expression in response to various environmental conditions or external stresses, helping to understand underlying biological processes such as development, stress adaptation, and resistance to pathogens.

II.1.1.- Objectives of transcriptome analysis

- **Identify expressed genes:** Determine which genes are active in a specific cell type, tissue, or organ at a given time.
- **Quantify gene expression:** Measure the amount of mRNA transcribed for each gene.

- **Understand gene expression regulation:** Identify the factors that control the activation and deactivation of genes.
- **Determine the role of genes in biological processes:** Link gene expression to specific phenotypes (observable characteristics) and biological functions.

Techniques used :

- **RNA sequencing (RNA-seq):** A cutting-edge technique that allows sequencing the entire set of mRNA present in a sample.
- **DNA microarrays:** A technology that enables the simultaneous measurement of the expression of thousands of genes.
- **Quantitative PCR (qPCR):** A technique that allows quantifying the expression of a specific gene.
- **Fluorescence in situ hybridization (FISH):** A technique used to visualize the localization of mRNA within cells.

II.1.2.- Applications of transcriptome analysis

- **Crop improvement**: Identifying genes that control traits of agronomic interest (disease resistance, yield, nutritional quality).
- **Development of new drugs**: Identifying therapeutic targets for plant-related diseases.
- **Understanding plant biology**: Discovering new mechanisms of gene expression regulation and biological processes.
- **Biotechnology and bioproduction**: Optimizing the production of valuable molecules in plants (enzymes, drugs, biofuels).

Transcriptome analysis has helped identify genes involved in plant resistance to water stress. This knowledge can be used to develop crops that are more drought-resistant, which is a major challenge in the context of climate change.

Transcriptome analysis is an expanding field of research that offers promising prospects for agriculture, medicine, and the understanding of the plant world.

II.2.- Mutants and reverse genetics

II.2.1.- Mutants

Mutants are organisms that exhibit a variation in their DNA sequence compared to the wild-type sequence. This variation can be natural or artificially induced by mutagenic agents.

A mutation is a permanent change in the DNA sequence of a gene. Mutations in the DNA sequence of a gene can alter the amino acid sequence of the protein encoded by the gene.

Nature of mutation :

Substitution mutations

Substitution mutations convert one type of base pair into another. Changes from G-C to A-T and A-T to G-C are called transition mutations (where a purine-pyrimidine base pair is replaced by another purine-pyrimidine pair). Changes from G-C to C-G, G-C to T-A, A-T to T-A, and A-T to C-G are called transversion mutations (where a purine-pyrimidine base pair is replaced by a pyrimidine-purine pair).

Although transitions are more frequent than transversions, both types of mutations can result from replication errors, chemical damage to DNA, and have been implicated as causative factors in hereditary genetic diseases and cancer. A single nucleotide change can modify a codon to encode a different amino acid, thereby altering the protein. Additionally, such changes can also create a stop codon, leading to premature truncation of the protein.

a. Point mutation

A point mutation is a simple change of a base in the sequence of a gene. It is equivalent to changing a letter in a sentence, like in this example where we change the "c" in "cat" to "h".

Original	The fat cat ate the wee rat
Point mutation	The fat **h**at ate the wee rat

b. Deletion

Mutations that result in the loss of DNA are called deletions. These can be small, such as the removal of a single "word," or larger, affecting a large number of genes on a chromosome. Deletions can also cause frameshift mutations. In this example, the deletion has removed the word "cat.".

Original	The fat cat ate the wee rat
Deletion mutation	The fat ate the wee rat

c. Insertion

Mutations that result in the addition of extra DNA are called insertions. They can also cause frameshift mutations and typically lead to a nonfunctional protein.

Original	The fat cat ate the wee rat
Insertion mutation	The fat **xlw** cat ate the wee rat

d. Frameshift mutation

In a frameshift mutation, one or more nucleotides are inserted or deleted, which is equivalent to adding or removing letters in a sentence. However, because our cells read DNA in "words" of three letters, adding or deleting a letter alters every subsequent word. This type of mutation can make the DNA meaningless and often leads to a truncated protein.

An example of a frameshift mutation using our example sentence is when the "t" from "cat" is deleted, but we keep the original spacing of the letters:

Original	The fat cat ate the wee rat
Mutation décalante	The fat c**aa tet hew eer at**

e. Inversion

In an inversion mutation, an entire section of DNA is reversed. A small inversion may involve only a few bases within a gene, while larger inversions can affect large regions of a chromosome containing multiple genes.

Original	The fat cat ate the wee rat
Inversion	The fat **tar eew eht eta tac**

II.2.2- Reverse genetics

Reverse genetics is an experimental approach that uses mutants to identify the function of genes. This approach involves modifying a specific gene in an organism and observing the effects of this modification on the organism's phenotype (observable characteristics).

Advantages of reverse genetics:

- **Direct approach:** It allows the study of the function of a specific gene in a direct and controlled manner.
- **High resolution power:** It enables the identification of precise functions of different domains of a gene.
- **Wide applicability:** It can be used to study genes in various organisms, including animals, plants, and microorganisms.

Challenges of reverse genetics:

- **Creation of mutants**: The process of creating mutants can be labor-intensive and time-consuming.
- **Interpretation of results**: The effects of a mutation can be complex and difficult to interpret.

- **Differences between mutants and wild-type organisms**: Mutants may exhibit characteristics that are not representative of wild-type organisms.

Examples of applications of reverse genetics:

- **Identification of genes responsible for genetic diseases**: Reverse genetics has helped identify genes responsible for numerous genetic diseases, such as cystic fibrosis and muscular dystrophy.
- **Understanding developmental mechanisms**: Reverse genetics has provided insights into the mechanisms of embryonic development and cellular differentiation.
- **Development of new drugs**: Understanding gene functions enables the development of new drugs to treat genetic diseases.

Conclusion:

Reverse genetics is a powerful tool in biological research. It has enabled significant advances in understanding how genes function and their role in diseases and development

II.3.- Proteomics and applications in plant biology

II.3.1.- Definition

Proteomics is the large-scale study of proteins, including their structure, function, localization, and interactions. It focuses on the entire set of proteins present in an organism, organ, cell, or subcellular compartment at a specific moment in time.

The goal of proteomics is to identify (and quantify) the entire set of proteins synthesized, known as the proteome, at a given time and under specific conditions within a tissue, cell, or subcellular compartment.

The proteome is highly complex for several reasons:

- Alternative splicing of primary transcripts
- Post-translational modifications and other protein maturations
- Developmental stage or phase of cellular activity
- Dynamics of protein synthesis rates and half-life times.

Proteomics provides answers that transcriptomics cannot offer:

- **Additional information on gene expression modalities** for organisms whose genome has not yet been sequenced or for which coding sequence prediction programs are less reliable. One example is aiding in the identification of exon boundaries, which in turn improves transcriptome annotation and genome annotation.

- **Quantitative estimation of synthesized protein concentrations** (using affinity markers containing an isotope for identification: ICAT).

- **Obtaining data on protein function and interactions between proteins or between proteins and other biological molecules** (using the yeast two-hybrid approach or the "tandem affinity purification by tag" - TAP/TAG approach).

The goals of these disciplines are therefore (non-exhaustive list) :

- Describe the organization of genes and locate regulatory motifs for gene expression (such as initiation or termination sites of transcription, etc.).

- Determine the overall structure of genes: identify open reading frames (ORFs), locate coding regions, locate splice sites at exon/intron boundaries, etc.

- Identify regions of genomes whose functions are still unknown and elucidate their roles, determine pseudogenes, transposable elements, etc.

- Study differences in gene product expression over time and across different tissue and cell types.

- Study the structure and function of proteins and RNAs encoded by the genes.

- Study methylation patterns.

- Study DNA-protein interactions (such as transcription factors, etc.).

- Integrate all this information into a larger system, that of all metabolic pathways (metabolome).

- Describe interactions among all these types of biological macromolecules (interactome).

- Obtain this data for as many organisms as possible ("gold: genomes Online database").

- Medical genomics.

- Metagenomics: study of the genome of an organism sampled directly from a complex environment (e.g., gut, ocean, soils, etc.), as opposed to a laboratory organism. The aim is to gather information about the impact of this environment. The prefix "meta" means "after, beyond, with, ...". ("metagenomics at EBI").

- Epigenetics and epigenomics: study of the influence of environment and individual history on changes in gene expression across generations. The prefix "epi" means "on, above, ...". **(Fig. 14)**.

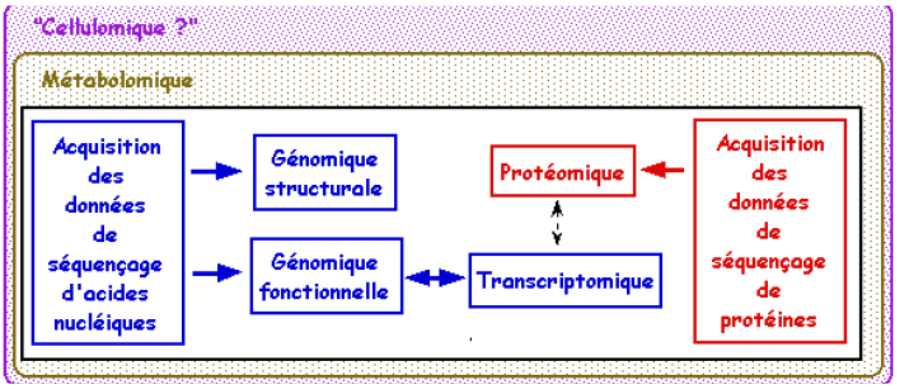


Figure 14 : The goals of proteomics and transcriptomics.

Proteomics techniques are numerous and aim, on one hand, to catalog the proteins within a specific cell type or cellular compartment, and on the other hand, to reveal interactions between proteins. The ultimate goal of these techniques is multifaceted, with the primary fundamental aim being to understand the functioning of complex molecular machines composed of multiple proteins. However, these techniques still face significant technical challenges, and in addition to issues specific to each method, they are not immune to broader generic problems, such as reproducibility and standardization. The large-scale practice of proteomics is still limited to a few platforms with expertise that is not yet widely available internationally. The information produced by these approaches is useful but primarily serves as a starting point for more targeted studies that should be conducted at the laboratory level, rather than on large platforms.

Many recent techniques and strategies for functional genome analysis have used yeast as a platform for validation due to the limited number of its genes and the availability of genetic tools (including mutant collections and gene collections cloned under different promoters in various vectors). Yeast has been instrumental in validating protein double-tagging technology for purifying native complexes (Tap-TAG), the first protein chips, ligand-binding assays, systematic tagging of proteins with artificial epitopes or GFP, and more. The aim is to quickly transfer these techniques to more complex systems, such as human cells. This is precisely where private companies have taken over, limiting the dissemination of knowledge in these fields.

Gene deletion mutant collections have been constructed in yeast and *Bacillus subtilis*, and these have proven to be valuable resources. In yeast, genetic interaction maps based on synthetic phenotypes are being developed. These maps, which may be general or more specific, involve very sophisticated experimental strategies, with expertise in genetics playing a critical role. However, such mutant collections are likely to remain limited to organisms where homologous recombination is effective for gene replacement (and where

the proportion of essential genes is low, which is probably the case for most organisms). For other organisms, such as *Arabidopsis thaliana* (and even yeast), gene inactivation by transposons or random mutagenesis by DNA insertion allows for the creation of nearly complete mutant collections. Depending on the difficulty of characterizing and maintaining these mutants, these collections may need to be centralized (*A. thaliana*) or can be non-centralized (yeast). For most organisms, functional gene inactivation via RNA interference (RNAi) seems to be the preferred method, as is the case for *Caenorhabditis elegans*, for example.

II.3.2.- Various approaches to proteomics

Proteomics is the global study of protein expression in a cell or organism. Complementary to genomics, which focuses on gene expression, the emergence of proteomics arises from the fact that proteins, the final products of genes, are more representative of a cell's function than the genes themselves.

Proteomics has been defined as the characterization of biological processes and the decoding of mechanisms controlling gene expression through the quantitative determination of gene expression at the protein level. This relatively satisfactory definition refers to a systematic study of proteins, based on their identification, quantification, characterization, and functional analysis.

To address this challenge, various separation and identification techniques have been developed, including:

- **Two-dimensional gel electrophoresis** developed by O'Farrell in 1975.
- **Mass spectrometry (MS)**.

II.3.3.- General overview (key reminders)

a- Protein structure

Each protein is made up of 20 "building blocks": amino acids (aa). A protein is a combination, in the form of a chain of varying length and orientation, of these 20 amino acids (ranging from 100 to 200 amino acids) organized in different structural levels **(Fig. 15)**.

- The sequence of amino acids in a protein constitutes its **primary structure**.
- Regions of the protein can adopt two specific shapes: helices or sheets, forming the **secondary structure**.
- The final three-dimensional structure adopted by the amino acid chain constitutes the **tertiary structure** of the protein.
- Multiple proteins can associate to form complex assemblies, which is referred to as the **quaternary structure**.

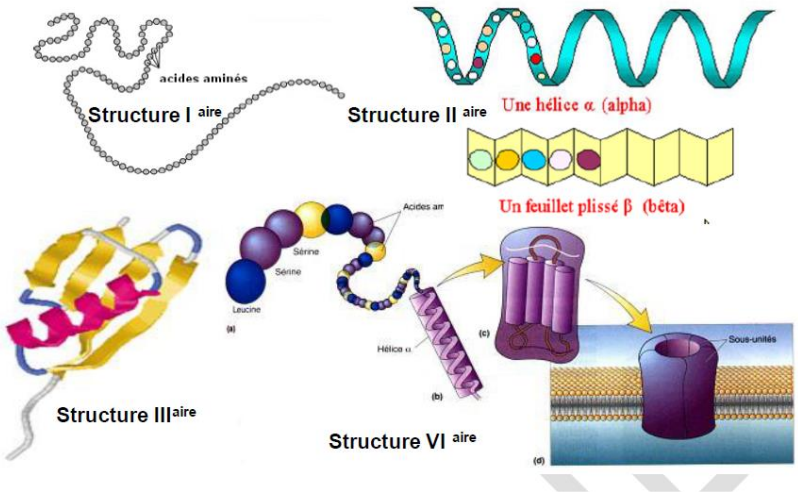


Figure 15 : Different levels of structural organization in proteins.

Primary Structure: Sequence of amino acids (aa) → Molecular weight and isoelectric point
Secondary Structure: Interactions between different chemical groups of the aa → Hydrophobicity
Tertiary Structure: Folding of the protein due to constraints imposed by the secondary structure → Hydrophobicity and apparent molecular weight
Quaternary Structure: Association of multiple protein subunits

b. Physicochemical properties of proteins

Due to their amino acid composition, proteins differ from one another in their molar mass and net charge at a given pH.

The net charge of a protein is characterized by its isoelectric point (pI).

Proteins exhibit:

- **pI values** ranging from as low as 2 (e.g., glucose-1-phospho-D-mannosylglycoprotein phosphodiesterase) to as high as 12 (e.g., cathepsin G).
- **Molar masses** ranging from 5 kDa to 590 kDa.

c- Role and location of proteins in the cell

- Structural proteins: Architectural elements within or outside the organism, such as actin, tubulin, and collagen.

- Enzymes: Catalysts of metabolic reactions, including enzymes involved in glycolysis and biosynthetic pathways of lipids, proteins, etc.

- Proteins involved in DNA and RNA synthesis and modification: Examples include DNA polymerase and RNA polymerase.

- Membrane channels and transporters: Regulate the exchange of substances between the cell and the extracellular environment, e.g., CFTR, ClC2, ENaC, Na/K ATPase.

- Receptors: Receive and process extracellular signals, such as EGFR (Epidermal Growth Factor Receptor) and TNFR (Tumor Necrosis Factor Receptor).

- Signaling proteins:

  - Protein-based hormones (e.g., insulin, growth hormone).

  - Intracellular signaling molecules involved in signal transduction.

  - Transcription factors that regulate gene expression.

- Defense proteins: Protect the organism against external threats, such as antibodies.

Applications in plant biology :

- **Understanding biological mechanisms:** Identify proteins involved in various biological processes such as photosynthesis, growth, stress response, etc.
- **Development of new technologies:** Develop new biomaterials, biosurfactants, and biocatalysts derived from plant proteins.
- **Crop improvement:** Identify proteins that control agronomically important traits (disease resistance, yield, nutritional quality).
- **Development of new drugs:** Identify therapeutic targets for diseases affecting plants.

Techniques used:

- **Gel electrophoresis:** A technique used to separate proteins based on their size and charge.
- **Mass spectrometry:** A technique used to identify proteins and determine their structure.
- **Chromatography:** A technique used to separate proteins based on their chemical properties.
- **Immunodetection:** A technique used to detect specific proteins using antibodies.

Concrete example:

Proteomic analysis has identified proteins involved in plants' resistance to water stress. This knowledge can be used to develop drought-resistant crops, which is a major challenge in the context of climate change.

II.4.- Metabolism and measurement of metabolic fluxes

Metabolism is the set of chemical reactions that occur within a living organism. It enables the organism to convert nutrients into energy and essential biomolecules required for growth and maintenance.

II.4.1.- Measurement of metabolic fluxes

Metabolic flux refers to the rate at which metabolic reactions occur. Its measurement allows for the quantification of the metabolic activity of an organism, organ, cell, or subcellular compartment.

Measurement techniques

- **Isotopic Methods:** Tracing radioactive or stable atoms to track the flow of metabolites in metabolic reactions.
- **NMR Spectroscopy:** Measuring the concentrations of specific metabolites.
- **Flow Cytometry:** Measuring metabolic fluxes at the cellular scale.
- **Mathematical Modeling:** Developing mathematical models to simulate metabolic fluxes.

Applications:

- **Understanding metabolic mechanisms:** Identifying the factors that control metabolic fluxes and the interactions between different metabolic pathways.
- **Development of new drugs:** Identifying therapeutic targets for metabolic diseases.
- **Crop improvement:** Identifying plant varieties with more efficient metabolic fluxes for better yield and nutritional quality.
- **Biotechnology and bioproduction:** Optimizing the production of valuable molecules by microorganisms.

Concrete example:

The measurement of metabolic flux has led to the identification of therapeutic targets for diabetes, a metabolic disease characterized by hyperglycemia.

Conclusion:

Metabolic flux measurement is a powerful tool in biological research. It enhances our understanding of metabolic mechanisms and opens up new applications in medicine, agriculture, and biotechnology.

II.5.- Genomics and analysis of tolerance to environmental stresses

Genomics is the study of the genome, which is the complete set of DNA of an organism. It allows for the identification and characterization of genes responsible for the variation in phenotypic traits, including tolerance to environmental stresses.

What are the prospects for the genetic improvement of drought-tolerant crops?

The prospects for the genetic improvement of drought-tolerant crops focus on several approaches aimed at enhancing the plants' ability to survive and grow under water stress conditions. Here are some key research and innovation directions in this field :

1. **Identification of drought tolerance genes** :
    - Advances in genomics allow researchers to identify genes involved in drought tolerance, such as those regulating stomatal closure, water management, and compatible solute accumulation.
    - The use of model plants like *Arabidopsis thaliana* or important crops like maize, wheat, and rice helps in understanding the genetic basis of drought tolerance.
2. **Genetic modification of water signaling pathways:**
    - Manipulating signaling pathways of hormones like abscisic acid (ABA) and stress response proteins, such as dehydration-responsive elements (DREB) or hyperosmotic stress proteins (HSP), can strengthen plant responses to water stress.
    - Controlling transpiration and stomatal opening through specific genes may help reduce water loss without affecting photosynthesis.
3. **Improvement of water use efficiency (WUE)** :
    - Selection and genomic editing strategies aim to improve water use efficiency, enabling plants to produce more biomass or yield with less water.
    - Research into enhancing root systems to access water, such as increasing root depth or density, can also contribute to better drought tolerance.
4. **Genetics of osmotic tolerance mechanisms and anti-stress proteins** :
    - Developing plants capable of maintaining high concentrations of compatible solutes (such as sorbitol or proline) and other protective molecules during dehydration could enhance their drought tolerance.

- Modifying the composition of the cell wall to make plants less susceptible to dehydration is also a promising avenue.

5. **Genome editing with CRISPR/Cas9** :

   - Genetic editing technologies like CRISPR/Cas9 offer immense potential to insert or modify specific genes responsible for drought tolerance. These technologies allow for targeting precise genes without affecting other important traits in the plant.

6. **Crossbreeding and traditional improvement:** :

   - The combination of genomics with traditional selection techniques, such as breeding resistant varieties with productive crops, remains a key approach to improving drought tolerance.

7. **Characterization and selection of genes associated with resistance to multiple stresses** :

   - Climate change leads to more varied stress conditions, such as drought combined with extreme temperatures. Genes that allow tolerance to multiple types of stress simultaneously are therefore of particular interest.

   - Identifying genes that control multiple tolerance mechanisms could lead to the creation of plants capable of withstanding more complex and prolonged stress conditions.

Conclusion :

Genetic improvement of crops for drought tolerance relies on the integration of genomic knowledge, advanced genome editing technologies, and traditional approaches. It offers promising prospects for developing crops that are more resilient to climate change and reducing dependence on irrigation, thus contributing to more sustainable and drought-resilient agriculture.

II.5.1.- Analysis of tolerance to environmental stresses

The complete continuity between genetics, ecology, and evolution is a recent development. Just a few years ago, it seemed unimaginable that the historical gap separating them could be bridged. The insights provided by genetic approaches play a central role not only in studies of evolution, with the genome itself being the obvious target, but also in most studies on biodiversity, the relationships of organisms to their environment, their responses and adaptations to stresses, the dynamics and stability of populations, parasitic or symbiotic interactions, and the organization of multispecific communities.

The goal of this analysis is to identify the genes and genetic mechanisms that enable organisms to survive and thrive under unfavorable environmental conditions.

Types of environmental stresses:

- **Abiotic stresses:** drought, salinity, extreme temperatures, UV light, etc.
- **Biotic stresses:** attacks by pathogens, herbivores, etc.

Genomic approaches:

- **Comparative genomics:** Comparing the genomes of species that differ in their tolerance to a given environmental stress.
- **Association genomics:** Identifying genetic variants associated with tolerance to an environmental stress in a population of organisms.
- **Transcriptomics:** Analyzing gene expression in organisms subjected to environmental stress.
- **Proteomics:** Analyzing proteins expressed in organisms subjected to environmental stress.

Applications:

- **Development of new crop varieties** resistant to environmental stresses.
- **Improving animal resistance** to diseases and parasites.
- **Developing new medicines** for human diseases linked to environmental stresses.
- **Understanding the mechanisms behind** organisms' adaptation to environmental changes.

Concrete example:

Genomic analysis has helped identify genes responsible for drought resistance in plants. These genes can be used to develop more drought-resistant crops, a major issue in the context of climate change.

Plants are a major component of our environment and play a significant role in oxygen production, carbon dioxide absorption, and the cycling of many mineral elements. Furthermore, in terms of health, many medications use active principles derived from plants (e.g., aspirin, antitumor agents such as taxol and vinblastine, morphine, etc.), and plant biodiversity has not been fully exploited by humans for nutritional purposes. Any advance in understanding the biosynthetic pathways of these products has, or will have, an impact on nutrition or the production of medicines, thus affecting health.

Practically, the development of plant genetics has fostered the growth of plant breeding and variety selection, which underpins all of our agriculture and part of plant biotechnology. Plant genetics and genomics are of great importance, both theoretically and practically. In terms of contributing to the acquisition of knowledge, plant genetics research initially helped establish the fundamental laws of genetics through Gregor Mendel's studies on peas. The work of Barbara McClintock on maize led to the discovery of transposable elements and the concept of genome fluidity. More recently, research into the development of transgenic plants revealed mechanisms for gene expression inactivation.

Pioneering German and English work on the floral development genetics of the snapdragon led to the identification of homeotic genes involved in floral development. These various studies have illustrated the peculiarities of developmental traits in higher plants. Generally speaking, plant genetics has allowed the analysis of developmental processes and signal transduction pathways of hormones or environmental cues over the past decade. These discoveries now guide strategies for plant improvement and input utilization.

Initially, plant genomics focused on species that had both "good genetics" and model value; this has now extended to major agronomic species, whose genomes are often much larger. Sequencing small genomes, like that of *Arabidopsis* (130 Mb) or rice (440 Mb), was realistic, but it was far less feasible to sequence that of maize (2,600 Mb) or wheat (16,000 Mb). For most agronomic species, the initial step involved creating saturated maps using RFLP markers and compiling gene catalogs from cDNA tags (ESTs).

II.5.2.- A sensitive topic: plant GMOs and plant biotechnology

Even though the creation of transgenic plants (GMOs) is just one of the many outcomes of plant genetics, both for research and practical applications, it is often equated with plant genetics. The negative perception of GMOs thus impacts the future development of plant genomics. Pioneering work in this field was primarily conducted in Europe during the 1980s, particularly under the leadership of J. Schell and M. Van Montagu in Ghent, and Jacques Tempé in Versailles. The cellular machinery of the phytopathogenic bacterium *Agrobacterium*, which allows it to insert a bacterial oncogene fragment into the genome of a plant cell, was used to develop effective methods for plant genetic transformation. This opened the door to numerous fundamental research projects, especially those that aim to establish the function of a gene by inactivating it or complementing a mutation. Notably, one major discovery was the mechanism of gene silencing (epigenetic gene inactivation).

However, these techniques also enabled the development of biotechnologies with agronomic benefits, such as herbicide resistance, the creation of new disease or insect resistances, modification of seed reserves, production of therapeutic molecules, and bioremediation. The advancement of genomics will also lead to the creation of more defined, more efficient, and more traceable GMOs. Despite this, these efforts have been,

and continue to be, significantly hindered in Europe by the moratorium on GMOs, while they are heavily supported in the United States. Today, 67.7 million hectares of GMOs, more than four times the arable land in France, are cultivated worldwide, with only a tiny fraction (less than one-thousandth) in Europe, without leading to the ecological catastrophes that were predicted.

Unfortunately, this has resulted in a disengagement from industries in this sector and a marked decline in the career interest of young people in plant research fields, as the number of applications for master's programs in plant molecular biology has decreased fivefold in five years. The solution to this issue will largely determine the competitiveness of our agribusiness industry and the future of this research sector, as well as our GDP and collective well-being.

MODEL PLANTS

III

Contenent

The very diversity of the plant kingdom reflects the great complexity of plant genomes: sometimes considerable size, partial or complete duplications, addition of genomes from related species, the result of natural evolution over millions of years, as well as various forms of selection (domestication) carried out by humans over the centuries. All of this necessitates the implementation of specific methods and tools.

Researchers have chosen to work on species known as "model species," with relatively small genomes, such as *Arabidopsis* (*Arabidopsis thaliana*) and rice. The integration of data obtained from these model species and their extrapolation to the study of agronomically important species, such as wheat, maize, or rapeseed, requires the use of high-throughput molecular biology techniques and the development of powerful bioinformatics tools. Large-scale exploration of populations of molecules, samples, or entire individuals thus opens entirely new perspectives.

Criteria for choosing model plants

Model plants are plant species used to study biological processes and develop plant technologies. They are chosen for a number of reasons, including :

- **Ease of growth and reproduction**: This allows researchers to rapidly generate large populations of plants for genetic studies.
- **Small genome**: This facilitates the sequencing and analysis of the plant’s genome.
- **Economic importance**: Model plants are important crops such as rice, maize, and soybean.
- **Fields of application**: Model plants are used to study a variety of biological processes, including photosynthesis, growth and development, stress response, and symbiosis.

Arabidopsis thaliana is one of the most widely used model plants. It is a small flowering plant native to Europe and Asia. Arabidopsis has a short life cycle, which facilitates growth and reproduction in the laboratory. Additionally, its genome has been fully sequenced and annotated.

Rice (*Oryza sativa*) is another important model plant. It is a grass cereal native to Asia. Rice is a major food crop for more than half of the world’s population. The rice genome has been sequenced and annotated, and a number of genetic tools are available to study this plant.

Maize (*Zea mays*) is another important model plant. It is a grass native to Central America. Maize is an important food and forage crop worldwide. The maize genome has been sequenced and annotated, and several genetic tools are available to study this plant.

Tomato (*Solanum lycopersicum*) is an important model plant. It is a flowering plant native to South America. Tomato is an important vegetable worldwide. The tomato genome has been sequenced and annotated, and a number of genetic tools are available to study this plant.

Model plants are a valuable tool for research in plant biology. They have allowed researchers to make significant progress in understanding biological processes and developing plant technologies.

III.1.- The model species *Arabidopsis thaliana*

Over the past 30 years, the plant *Arabidopsis thaliana* (commonly known as thale cress, a weed) has been adopted as a model organism by thousands of biologists **(Fig. 16)**. This scientific community has developed numerous tools, resources, and experimental methods that have significantly advanced plant research. The international scientific effort has culminated in sequencing the nuclear genome of *Arabidopsis*. Current research, including initiatives supported by GÉNOPLANTE, aims to characterize the function of the approximately 29,000 genes encoded by this genome. The unanimous acceptance of *Arabidopsis* as a model system can be attributed to several factors.

1/ **Small genome**: Its genome, consisting of 120 million base pairs, is one of the smallest among plants (for comparison, the wheat genome has about 16 billion base pairs).

2/ **Ease of cultivation**: It grows easily under laboratory conditions, completing its life cycle in just 8 weeks, from seed sowing to seed harvesting. Each plant produces thousands of seeds.

3/ **Natural diversity**: A wide variety of ecotypes exist, displaying diverse traits such as seed size and disease resistance, allowing researchers to utilize natural biodiversity for the characterization of agriculturally significant genes.

Description:

- A small flowering plant native to Europe and Asia.
- Short life cycle (approximately 6 weeks).
- Small genome (around 100 Mb).
- Widely used in plant biology research.

Advantages:

- **Ease of growth and reproduction**: Enables researchers to quickly generate large populations of plants for genetic studies.
- **Small genome**: Simplifies genome sequencing and analysis.

- **Large number of mutants**: Facilitates the study of gene functions.
- **Well-developed genetic tools**: Allows researchers to manipulate the plant's genome effectively.

Contributions to research:

- **Understanding fundamental biological processes**: Photosynthesis, growth and development, stress response, symbiosis.
- **Development of plant technologies**: Genetic engineering, biotechnology, agriculture.

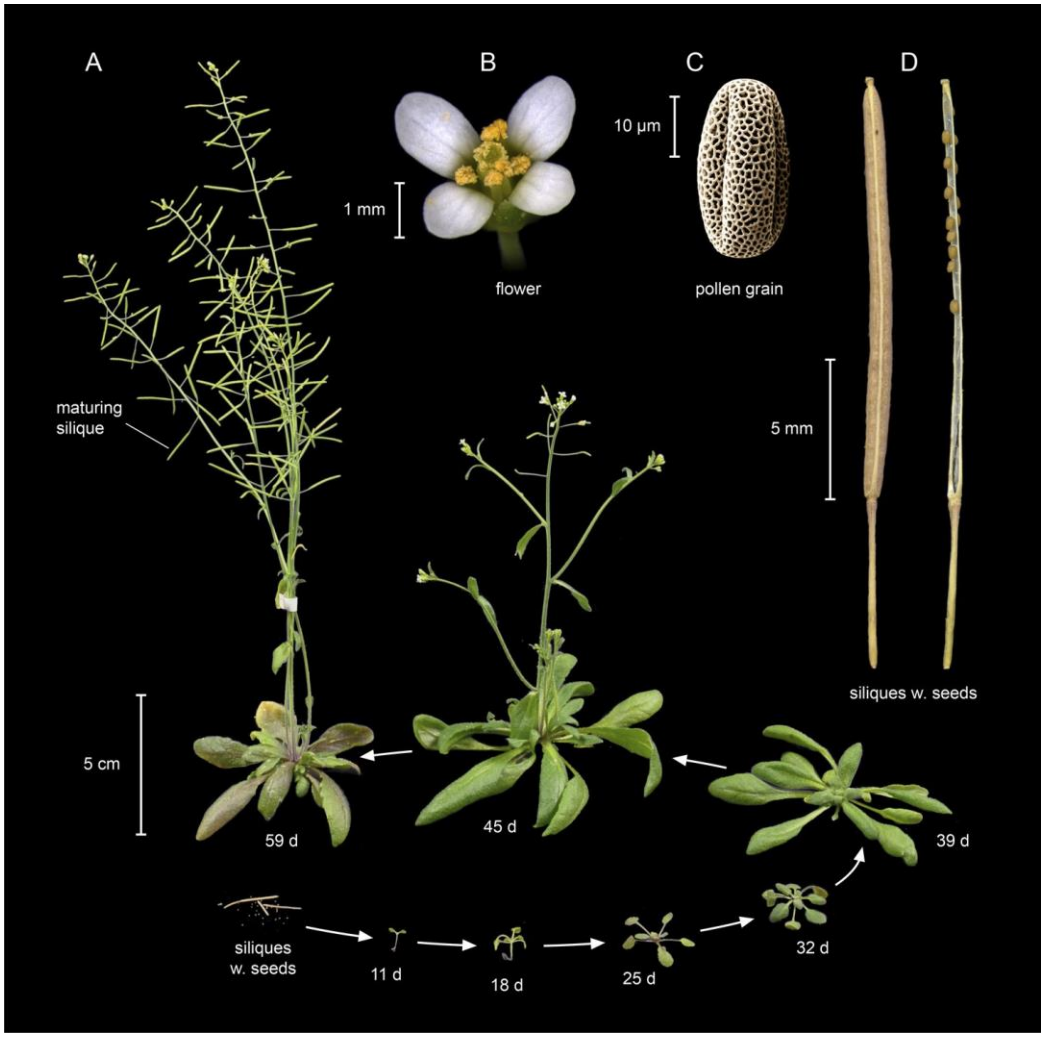


Figure 16 : Life cycle of *Arabidopsis thaliana*.

(A) *A. thaliana* of the accession Columbia (Col) at different stages of its life cycle, from seed (bottom left) to seedling (11 days), to vegetative growth (39 days), and to reproductive growth (45 days). Photographs of (B) a flower, (C) a pollen grain (scanning electron micrograph), and (D) mature siliques (seed pods; left: closed; right: open with a few remaining unshattered seeds) at higher magnification.

III.1.1.- Rapeseed (*Brassica napus*)

Rapeseed (*Brassica napus*, family Brassicaceae) is an annual plant cultivated for its seeds, which are rich in oil and used to produce meal suitable for livestock feed. Most varieties grown in Europe are winter varieties. After sowing in late August to early September, the plant reaches the rosette stage by early winter, a developmental phase during which it exhibits exceptional cold resistance. Moreover, low winter temperatures are necessary for the plant to flower when growth resumes in spring (vernalization). Fertilization, primarily self-pollinating, results in the formation of siliques containing the seeds.

The flowering period is extended (4 to 6 weeks), creating challenges related to the heterogeneity of harvested seed batches. Additionally, at maturity, the phenomenon of silique dehiscence (opening) can lead to seed loss and reduced yield.

Rapeseed is susceptible to several fungal diseases, including *Phoma*, *Sclerotinia*, and *Alternaria*. It is cultivated worldwide on nearly 25 million hectares, producing approximately 35 million tons (France: 1.3 million hectares; 4.4 million tons).

III.2.- Rice as a model plant for cereal genomics

Rice (family *Poaceae*) is a cereal crop grown in warm regions, with its grain forming the staple food for a large part of the world's population, particularly in Asia. Global production, approximately 600 million tons cultivated on nearly 150 million hectares, is dominated by Asia, which accounts for 90% of the world's output. Due to its major economic importance, the scientific community has focused on sequencing the nuclear genome of this species. Recently, the near-complete sequencing of its nuclear genome has been published by several research teams.

Moreover, rice is now considered a model plant because its gene repertoire and synteny (gene order) are largely conserved across all cereals. As a result, genomic research on rice directly benefits programs such as GÉNOPLANTE on maize and wheat.

Rice (*Oryza sativa*) is an essential model plant for cereal genomics for several reasons :

- **Economic importance:** Rice is the most widely consumed cereal in the world, feeding more than half of the global population.
- **Small genome:** The rice genome is relatively small compared to other cereals, making it easier to sequence and analyze.
- **Economic impact:** Rice is a crucial crop for many developing countries, and advances in rice research can significantly impact global food security.

- **Applications:** Rice is used to study a variety of biological processes, including photosynthesis, growth and development, stress response, and symbiosis.

Contributions to cereal genomics:

- **Genome sequencing**: The rice genome was one of the first cereal genomes to be sequenced, enabling the identification and characterization of many genes crucial for plant growth and development.

- **Development of genetic tools**: A variety of genetic tools have been developed for rice, allowing researchers to manipulate the plant's genome and study gene function.

- **Understanding biological processes**: Research on rice has significantly advanced the understanding of fundamental biological processes in cereals, such as photosynthesis, growth and development, stress responses, and symbiosis.

- **Crop improvement**: Insights gained from rice research have been applied to develop new rice varieties with enhanced resistance to diseases, pests, and environmental stresses.

Rice is a valuable model for cereal genomics research. Advances in rice research have improved the understanding of fundamental biological processes in cereals and led to the development of new rice varieties more resistant to diseases, insects, and environmental stresses.

Rice is an important model plant for cereal genomics. It has contributed to the understanding of fundamental biological processes in cereals and to the development of new rice varieties with enhanced resistance to diseases, pests, and environmental stresses. Research on rice will continue to play a crucial role in the fight against hunger and malnutrition worldwide

III.2.1.- Wheat

Two types of wheat (family *Poaceae*) are cultivated. Durum wheat (*Triticum durum*) is grown exclusively for its semolina (used in biscuits, cakes, couscous) and for the production of pasta. On the other hand, soft wheat (*Triticum aestivum*) is cultivated for both animal and human consumption and is the main cereal in temperate regions of the world.

With 220 million hectares, wheat significantly surpasses rice (150 million hectares) and maize (140 million hectares), with a total production of nearly 600 million tons. In France, the leading producer in Europe, wheat cultivation covers nearly 5 million hectares, with a production of around 40 million tons.

Wheat genomics aims to identify genes involved in the technological properties of flours, as well as to optimize yield under acceptable environmental conditions. Wheat is indeed susceptible to numerous fungal diseases. Therefore, it is important to identify resistant/tolerant varieties and introgress them into the best commercial varieties. Furthermore, the high yields currently achieved (100 q/ha) require substantial nitrogen fertilization, which can lead to groundwater pollution. It is essential, therefore, to analyze, at the genomic level, the necessity of these inputs in order to minimize their use.

III.2.2.- Maize

Originating from tropical regions, maize is an annual cereal introduced to Europe in the 16th century. Due to this origin, maize is a water-intensive plant (500 liters of water are required to produce 1 kg of seeds) and has specific temperature requirements (the germination threshold is around 6°C).

Maize can be cultivated for grain production (animal feed, starch production, semolina) or for forage (silage maize). This crop represents the largest global cereal production (600 million tons in 1998; 140 million hectares), with an average yield of around 45 quintals per hectare. The French production is approximately 15 million tons, accounting for about half of Europe's total maize production.

III.3.- *Medicago truncatula,* model plant for legumes and plant-microbe interactions

Description:

- Small flowering plant native to the Mediterranean region **(Fig. 17)**.
- Short life cycle (about 3 months)
- Small genome (around 450 Mbp)
- Widely used in legume research and plant-microbe interactions

Advantages:

- **Easy growth and reproduction**: Allows researchers to quickly generate large populations of plants for genetic studies.
- **Small genome**: Facilitates genome sequencing and analysis.
- **Nitrogen-fixing symbiotic system**: Enables the study of interactions between legumes and rhizobial bacteria.
- **Well-developed genetic tools**: Allow researchers to manipulate the plant's genome.

Contributions to research:

- **Understanding plant-microbe interactions:** *Rhizobia*, mycorrhizae, pathogens
- **Development of plant technologies:** Genetic engineering, biotechnology, agriculture
- **Crop improvement:** Development of legumes more resistant to diseases and environmental stresses.

Figure 27 : *Medicago truncatula* (barrel medic).

III.4.- Tomato as a model species for solanaceae and all fleshy fruits

Description:

- Herbaceous plant native to South America
- Widely cultivated for its fleshy fruit
- Short life cycle (about 4 months)
- Small genome (approximately 900 Mbp)
- Widely used in plant biology research

Advantages:

- **Ease of growth and reproduction**: Allows researchers to quickly generate large plant populations for genetic studies.

- **Small genome**: Facilitates sequencing and analysis of the plant's genome.
- **Fleshy fruit**: Enables the study of fruit development and maturation.
- **Well-developed genetic tools**: Allow researchers to manipulate the plant's genome

Contributions to research:

- **Understanding fundamental biological processes:** Photosynthesis, growth and development, stress response, symbiosis
- **Development of plant technologies:** Genetic engineering, biotechnology, agriculture
- **Crop improvement:** Development of tomatoes more resistant to diseases and environmental stresses

The tomato is a valuable model for plant biology research. It has contributed to the understanding of fundamental biological processes in plants and the development of new plant technologies. Research on tomatoes will continue to play an important role in improving tomato production and the sustainability of agriculture.

Additionally, the tomato is an important model for the study of fleshy fruits in general. The knowledge gained from tomato research has been applied to the study of fruit development and maturation in other fleshy fruits, such as apples, pears, and peaches.

In conclusion, the tomato is an important model species for plant biology research and the study of fleshy fruits. It has contributed to the understanding of fundamental biological processes in plants and the development of new plant technologies. Tomato research will continue to play a vital role in improving fruit production and agricultural sustainability.

In parallel with research on major cultivated species, the GÉNOPLANTE network has focused on several species of significant interest to consumers. These species were selected because they offer opportunities for generic research that is not feasible with the model species Arabidopsis and rice. Several projects were selected, focusing on vegetables, including leafy vegetables (e.g., celery) and fruit vegetables (e.g., tomato, pepper, melon).

The genetic improvement of these species aims to increase yields, improve disease resistance, enhance conservation ability, improve technological quality, and increase resistance to abiotic stresses. It combines classical genetic methods with genomics, thanks to the recent identification of genes of interest that the GÉNOPLANTE projects aim to clone.

GÉNOPLANTE has also decided to support an academic project aimed at establishing a consensus genetic map for grapevine. Grapevine is an important crop in France (producing 60 million hectoliters, representing a quarter of the global consumption). Protection against diseases and pests requires treatments ranging from 5 to 12 per year. Therefore, genomic projects aimed at characterizing disease resistance genes are essential.

Moreover, GÉNOPLANTE has supported a project on sorghum through CIRAD teams. Sorghum is a major cereal in Africa and India, and its cultivation is increasing in southern France as part of animal feed. It is also closely related to maize, self-pollinating, and has a genome four times smaller, making it a model for certain aspects of genetic analysis. This decision aligns with GÉNOPLANTE's charter, which aims to promote genomics research on species of major agronomic interest in developing countries.

BIBLIOGRAPHIC REFERENCES

- **Books**

[1] Morot-Gaudry Jean-François, Briat Jean-François, 2004 : La génomique en biologie végétale. Editions Quae, Paris. P.

[2] Douce R., 2000 : Le monde végétale. Du génome à la plante entière. Académie des sciences

[3] Elrod S., Stansfield W., 2003 : Génétique. Ed. : Dunod. Paris. 490P

[4] Gibson, G., 2003 : Précis de génomique. Ed. De Boeck University.

[5] Gontier, J. R., 2003 : La génétique en 1.001 QCM. Ed. Ellipses.

[6] Heberle E., 2001 : Génie génétique. Une histoire- un défi. Quae. 291P

[7] Henry, J. P., 2003 : Précis de génétique des populations : cours, exercices et problèmes résolus. Ed. : Dunod.

[8] Klug W., Cummings M. et Spencer C., 2006 : Génétique. Ed. : Pearson Education. France. 704P

[9] Lambert, G., 2003 : La légende des gènes : de l'origine de la génétique à la thérapie génique. Ed. Dunod.

[10] Maftah, A., 2003 : Génétique : DEUG SV, PCEM, Prépas, Pharmacie. Ed. Dunod.

[11] Ollivier, L., 2002 : Eléments de génétique quantitative. Ed. INRA.

[12] Schmid R.D., 2005 : Biotechnologie et de Génie génétique. Ed. : Flammation Médecine – Science. Paris. 335P

[13] Somouelian F., 2009 : Génétique moléculaire des plantes. Ed. Quae. 230p

[14] Winter P.C., Hickey G.I. et Fletcher H.L., 2000 : Génétique. Ed.: Port Royal Livres. Paris. 401P.

- **Web sites :**

https://www.gnis-pedagogie.org/sujet/sequencage-genome/

https://planet-vie.ens.fr/thematiques/manipulations-en-laboratoire/le-sequencage-d-un-adn

https://www.ncbi.nlm.nih.gov/

- **Articles**

[1] Albert, V. A., Barbazuk, W. B., Depamphilis, C. W., Der, J. P., Leebens-Mack, J., & Ma, H. (2013). The Amborella genome and the evolution of flowering plants. Science, 342(6165), 1241089. https://doi.org/10.1126/science.1241089

[2] Benfey, P. N., & Mitchell-Olds, T. (2008). Arabidopsis and rice as model systems for the genetics of agronomic traits. Nature Reviews Genetics, 9(7), 467-477. https://doi.org/10.1038/nrg2345

[3] Bolger, M. E., Scossa, F., Bolger, A. M., Lanz, C., Maumus, F., Tohge, T., & Usadel, B. (2014). The genome of the stress-tolerant wild tomato species Solanum pennellii. Nature Genetics, 46(9), 1034-1038. https://doi.org/10.1038/ng.3046

[4] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. Nucleic Acids Research, 40(Database issue), D1178-D1186. https://doi.org/10.1093/nar/gkr944

[5] Jansson, S. (2000). The light-harvesting chlorophyll a/b-binding proteins. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1461(1), 120-138. https://doi.org/10.1016/S0005-2728(00)00132-1

[6] Li, S., Lin, Y., Wang, P., Zhang, Z., Song, X., Gao, J., & Zhu, C. (2021). Transcriptomic, proteomic, and metabolomic analysis provide novel insights into cold-stress responses in maize seedlings. Frontiers in Plant Science, 12, 608245. https://doi.org/10.3389/fpls.2021.608245

[7] Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., & Katayose, Y. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485(7400), 635-641. https://doi.org/10.1038/nature11119

[8] Schatz, M. C., Witkowski, J. F., & McCombie, W. R. (2012). Current challenges in de novo plant genome sequencing and assembly. Genome Biology, 13(4), 243. https://doi.org/10.1186/gb-2012-13-4-243

[9] Schilling, R. K., Marschner, P., Shavrukov, Y., Berger, B., Tester, M., Roy, S. J., & Plett, D. C. (2014). Expression of the Arabidopsis vacuolar H+-pyrophosphatase gene AVP1 improves the shoot biomass and fruit yield of transgenic tomato under salinity stress. Functional Plant Biology, 41(4), 349-357. https://doi.org/10.1071/FP13242

[10] Tello-Ruiz, M. K., Stein, J., Wei, S., Preece, J., Olson, A., Naithani, S., & Ware, D. (2016). Gramene 2016: Comparative plant genomics and pathway resources. Nucleic Acids Research, 44(Database issue), D1133-D1140. https://doi.org/10.1093/nar/gkv1179

[11] Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 485(7400), 635-641. https://doi.org/10.1038/nature11119

[12] Udvardi, M., & Poole, P. S. (2013). Transport and metabolism in legume-rhizobia symbioses. Annual Review of Plant Biology, 64, 781-805. https://doi.org/10.1146/annurev-arplant-050312-120235

[13] Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nature Reviews Genetics, 10(10), 725-732. https://doi.org/10.1038/nrg2600

[14] Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. Trends in Plant Science, 10(12), 621-630. https://doi.org/10.1016/j.tplants.2005.10.004

[15] Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., & Paterson, A. H. (2011). The genome of the mesopolyploid crop species Brassica rapa. Nature Genetics, 43(10), 1035-1039. https://doi.org/10.1038/ng.919

[16] Weigel, D., & Mott, R. (2009). The 1001 genomes project for Arabidopsis thaliana. Genome Biology, 10(5), 107. https://doi.org/10.1186/gb-2009-10-5-107

[17] Yamaguchi-Shinozaki, K., & Shinozaki, K. (2006). Transcriptional regulatory networks in cellular responses and tolerance to dehydration and cold stresses. Annual Review of Plant Biology, 57, 781-803. https://doi.org/10.1146/annurev.arplant.57.032905.105444

[18] Yu, J., Hu, S., Wang, J., Wong, G. K. S., Li, S., Liu, B., ... & Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science, 296(5565), 79-92. https://doi.org/10.1126/science.1068037

GLOSSARY

Amino Acid: An amino acid is an organic molecule with a carbon skeleton and two functional groups: an amine (-NH2) and a carboxylic acid (-COOH). Amino acids are the basic structural units of proteins.

BAC: "Bacterial Artificial Chromosome" is a high-capacity vector (around 300 kilobases) with sequences that enable its replication and maintenance in *Escherichia coli*.

BLAST: The acronym for "Basic Local Alignment Search Tool," a heuristic method in bioinformatics used to identify similar regions between two or more nucleotide or amino acid sequences and align these homologous regions.

Codon: A codon is a triplet of nucleotides (A, C, U, or G) in messenger RNA.

Cosmid: A circular DNA molecule containing viral sequences (allowing it to be packaged into viral envelopes). It can clone inserts of up to 45 kilobases (45,000 base pairs).

DNA: Deoxyribonucleic acid (DNA) is a molecule found in all living cells. DNA is the carrier of heredity and genetic information, as it constitutes the genome of living organisms and is transmitted wholly or partially during reproduction. DNA determines protein synthesis.

Gene: A gene is a sequence of DNA that specifies the synthesis of a polypeptide chain or a functional RNA.

Genome: The genome is the entire genetic material of an individual or species encoded in its DNA (except for some viruses with RNA genomes).

Homolog: In evolutionary biology, homology refers to similarities between two traits (usually anatomical) in different species, inherited from a common ancestor. These traits are called homologous and can include anatomical or molecular features (e.g., homologous proteins).

Interactome: Refers to the entire set of interactions among proteins within a given cell.

Microsatellite: Short DNA sequences (2 to 6 base pairs) repeated multiple times. These sequences are found throughout the human genome and are highly polymorphic.

Nucleotide: Nucleotides are the building blocks of DNA (A, C, G, T) and RNA (A, C, G, U).

Ortholog: In evolutionary biology, two proteins are orthologous if they descend from a common ancestor following a speciation event.

Paralog: Two proteins are paralogous if they descend from a common ancestor after a duplication event.

Plasmid: A circular DNA molecule typically a few dozen kilobases in size. Found mostly in bacteria, plasmids are commonly used as vectors in gene cloning (e.g., pBR322, pBluescript). They can clone inserts up to 20–30 kilobases.

Polymorphic: A gene is polymorphic if it exists in multiple alleles. A gene or DNA fragment is said to be polymorphic if at least two distinct alleles (or forms) occur in a population with a frequency of at least 1%.

Protein: A protein is a macromolecule composed of one or more chains of amino acids linked by peptide bonds.

Proteome: The proteome represents the complete set of proteins produced by a genome.

Restriction Enzyme: An enzyme that digests DNA (endonuclease) at specific sequences, known as restriction sites. These sites can range from 2 to 20 base pairs. For example, the enzyme EcoRI recognizes and cleaves DNA at the sequence GAATTC.

RFLP: "Restriction Fragment Length Polymorphism" is a technique that involves digesting genomic DNA with restriction enzymes. Sequence differences among individuals result in different digestion patterns. After separation by electrophoresis, fragments of various sizes can be observed, creating markers across the genome.

Ribosome: A ribosome is a complex of proteins and ribosomal RNAs. Its function is to synthesize proteins by decoding the information in messenger RNA.

RNA: Ribonucleic acid (RNA) is a molecule similar to DNA, both structurally and functionally (materialization and processing of genetic information). There are various types of RNA, including messenger RNA (mRNA), which is transcribed from DNA as a copy. Its role is to transport genetic information from the nucleus to the cytoplasm, where it is translated into protein by ribosomes.

Transcriptome: The transcriptome is the complete set of transcripts (RNAs) present in a cell.

YAC: "Yeast Artificial Chromosome," an artificial chromosome created by combining centromeric and telomeric sequences from yeast chromosomes. These vectors can carry very large inserts (up to 1,000 kilobases). However, they are prone to rearrangements once integrated into yeast cells.

ANNEX

Annex 01: Teaching sheet for a course

Teaching unit (TU): Fundamental

Instructor(s) responsible for the TU:

Semester of the UE: Semester 1

Subject course: Plant genomics

Instructor(s) responsible for the course: Mériem MARFOUA

Weekly hourly volume: Lecture: 1h30 / Tutorial: 00 / Practical work: 1h30

Assessment methods:

1. **Duration of the end-of-semester exam:** Medium-length test (1h30) and Resit test (1h30).
2. **Continuous assessment:** (Specify the weighting of each element in the following table according to the subject composition):

Tutorials			**Practical work**			Presentation	
Attendance	Participation	Quiz	**Attendance**	**Participation**	**Reports**	Written	Oral
pts	pts	pts	**05 pts**	**05 pts**	**10 pts**	pts	pts

Course objectives (2 to 4 lines):

Gain knowledge of plant genomes and their functioning, and be able to make connections between plant genomics and cultivated species.

Course content (Chapters and paragraphs, maximum 10 lines):

Review of basic concepts

Chapter 1: Structural genomics

1. Plant genome program: tools for studying plant genomics.

2. Physical structure of the plant nuclear genome.

3. Plant genome sequencing.

4. Genomics and bioinformatics.

5. Biological databases and repositories.

6. Gene prediction.

7. Protein sequence comparison.

Chapter 2: Functional genomics

1. Analysis of the plant transcriptome.

2. Mutants and reverse genetics.

3. Proteomics and its applications to plant biology.

4. Metabolism and measurement of metabolic fluxes.

5. Genomics and analysis of environmental stress tolerance.

Chapter 3: Model plants

1. Model species *Arabidopsis thaliana*.

2. Rice as a model plant for cereal genomics.

3. *Medicago truncatula*, model plant for legumes and plant-microbe interactions.

4. Tomato as a model species for solanaceae and fleshy fruits.

Practical Work:

- TP 1: Extraction of genomic DNA from animal and plant cells.

- TP 2: Biological databases and repositories.

- TP 3: Genome sequence alignment.