الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -
aculté des Sciences de la Nature
t de la Vie et des Sciences de la Terre

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
- البويرة -
كلية علوم الطبيعة والحياة وعلوم الأرض

Democratic and Popular Algerian
Republic Ministry Of Higher Education
and Scientific Research
Akli Mohand Oulhadj University -
Bouira Faculty of natural and life
sciences and earth sciences

Educational handout

Biotechnological tools

Author: Dr. HADIDI Lila
Department: Biology
Faculty: Natural and Life
Sciences and Earth Sciences

Course intended for students of

Master I: Microbial biotechnology

Academic year: 2024/2025.

Preface

This handout is intended for students following the “Biotechnological Tools” course. It brings together the main theoretical and practical concepts covered during the program, as well as the methodologies and tools used in the field of modern biotechnologies. This document is intended to serve as both a teaching aid and a reference manual to support students in their learning.

This course, together with this handout, aims to provide students with a thorough understanding of the essential tools and techniques in this field, including genetic engineering, cell culture techniques, protein expression systems, and biological data analysis.

Each chapter of this document has been designed to address a specific aspect of biotechnological tools, in a progressive and detailed manner. In addition to theoretical notions, practical examples and case studies illustrate the concrete application of these technologies in the laboratory and in industry.

I recommend that students read this handout carefully and use it as a supplement to lectures and practical work sessions. The exercises (annexes) provided will help reinforce understanding of the concepts and effectively prepare for assessments.

I would like to thank my colleagues and students for their constructive feedback, which helped improve this document. I hope that this handout will be a valuable tool for students and that it will contribute to their success in studying biotechnology.

Happy reading and good luck to all!

Table of contents

Chapter III: Mapping and linkage

Chapter IV: Sequencing

Chapter V: Systematic expression and interaction studies

Chapter VI: Transcriptome, ChIp, ChIp on chip

Chapter VII: Proteomics

Chapter VIII: The double hybrid approach

Chapter IX: The TAP-tag technique

Chapter X: Systematic approaches to expression disruption

Chapter XI: RNA interference (RNAi) and systematic generation of mutants

Chapter XII: Expression of recombinant proteins

Chapter XIII: Methods and applications of protein engineering

Chapter XIV: Rational protein engineering

Chapter XV: Directed evolution of proteins and nucleic acids

List of abbreviations

2D-GE: Two-dimensional gel electrophoresis

bp: Base pairs

CAD: Computer-aided design

CAR: Chimeric antigen receptor

CBP: Calmodulin-binding peptide

ChIP: Chromatin immunoprecipitation

cM: Centimorgan

CNVs: Copy number variants

Co-IP: Co-immunoprecipitation

CRISPR: Clustered regularly interspaced short palindromic repeats

Cryo-EM: Cryo-electron microscopy

ddNTPs: Dideoxyribonucleotide triphosphates

DNA: Deoxyribonucleic acid

sRNA: Double-stranded RNA

E. coli: *Escherichia coli*

FISH: Fluorescence in situ hybridization

FRET: Fluorescence resonance energy transfer

GMOs: Genetically modified organisms

GWAS: Genome-wide association studies

HTS: High-throughput screening

iTRAQ: Isobaric tags for relative and absolute quantitation

LD: Linkage disequilibrium

LC: Liquid chromatography

LC-MS/MS: Liquid chromatography–tandem mass spectrometry

LFQ: Label-free quantification

LOD: Logarithm of the odds

MAS: Marker-assisted selection

mRNA: Messenger RNA

MS: Mass spectrometry

NGS: Next-generation sequencing

NMR: Nuclear magnetic resonance

PPIs: Protein–protein interactions

PTMs: Post-translational modifications

qPCR : Quantitative polymerase chain reaction

QTL : Quantitative trait loci

RNA : Ribonucleic acid

RNAi: RNA interference

RNA-seq: RNA sequencing

RIN: RNA integrity number

SMRT: Single molecule real-time

SNPs: Single nucleotide polymorphisms

SPR: Surface plasmon resonance

TAP: Tandem affinity purification

TFs: Transcription factors

TMT: Tandem mass tags

WES: Whole-exome sequencing

WGS: Whole-genome sequencing

Y2H: Yeast two-hybrid

List of figures

List of tables

Abstract

This handout provides an overview of modern techniques in molecular biology, genomics, and bioengineering. It highlights the tools used to analyze, manipulate, and understand genomes, proteins, and their molecular interactions. Emphasis is placed on approaches such as genomic library construction, high-throughput sequencing, gene mapping, gene expression analysis, proteomics, and gene expression disruption techniques such as RNA interference (RNAi).

The chapters progressively explore these topics, from the fundamentals of genetic engineering to advanced methods for studying molecular interactions (e.g., two-hybrid systems, TAP-tag). Protein engineering is also addressed, illustrating how rational or evolutionary approaches can be used to design molecules with enhanced or novel properties for biomedical, industrial, or research applications.

This document is intended for students with prior knowledge of molecular biology, biochemistry and genetics, and aims to link theoretical concepts with their practical applications in biomedical research, agriculture, and biotechnology.

Keywords: Functional genomics; high-throughput sequencing; molecular interactions; protein engineering.

The rise of molecular biology and genomics technologies in recent decades has paved the way for a deep understanding of the fundamental mechanisms that govern life at the molecular level. This course handout aims to provide a comprehensive overview of the main modern techniques and approaches used in molecular biology, genomics, and bioengineering, with an emphasis on methods for manipulating and analyzing genomes, proteins, and molecular interactions.

The development of tools for genome analysis, genomic library construction, gene mapping, and sequencing has accelerated the identification of genes and their function. These methods have been enhanced by systematic approaches to gene expression analysis and proteomics, which provide us with a global view of cellular activity at an unprecedented level. By integrating this information with expression disruption techniques such as RNA interference (RNAi) or systematic mutant generation, we can study gene and protein function in a targeted manner.

The following chapters detail these different approaches progressively. Each chapter explores an essential facet of current biomolecular research, whether it is the construction of genomic libraries (Chapter 2), high-throughput sequencing (Chapter 4), or protein engineering (Chapters 13 to 15). Advances in genetic mapping and linkage (Chapter 3) and the systematic study of molecular interactions via two-hybrid or Tap-tag (Chapters 8 and 9) are also discussed.

Protein engineering, discussed in the final chapters, provides a fascinating framework for designing molecules with improved or new properties, using rational or evolutionary approaches. These techniques create proteins with biomedical, industrial, or research applications, and illustrate how molecular biology and bioengineering influence many areas of biotechnology.

Understanding the concepts developed in this handout requires a good knowledge of basic molecular biology, as well as familiarity with the principles of genetics, biochemistry, and biotechnology.

This course strives to present theoretical concepts while emphasizing practical applications in biomedical research and biotechnology. Through these techniques, we seek to provide tools for fine-grained analysis of biological systems, with the ultimate goal of improving our understanding of biological processes and developing innovative solutions for human health, agriculture, and industry.

I. Genome analysis

Genome analysis is a crucial discipline in bioinformatics and molecular biology. It allows us to understand the structure, function and evolution of genetic sequences on a global scale. Today, with the rise of next-generation sequencing (NGS) technologies and the massive availability of genomic data, this analysis offers unknown eventuality to explore genetic diversity, identify mutations associated with diseases and understand evolutionary mechanisms.

I.1. Definitions and basic concepts

The DNA molecule is organized into units called genes which are successions of nucleotides determining the transmissible characteristics. On the functional level, a gene is a DNA sequence originating from the synthesis of a functional product which can be in the form of RNA or polypeptide. The genome is the entire genetic material of an organism (DNA or RNA in RNA viruses). It includes both coding (genes) and non-coding sequences and more precisely the DNA sequence corresponding to a haploid set of chromosomes. Each chromosome contains a DNA molecule and proteins associated (Figure I.1). A diploid human somatic cell therefore contains two genomes, a paternal and a maternal genome, whereas a sex cell (egg or sperm) contains only one. The term genome applies both to the DNA of the cell nucleus (nuclear genome) and to the DNA of the organelles: mitochondrial genome and chloroplast genome **[1].**

Genome analysis refers to the comprehensive examination of the complete DNA sequence of an organism’s genome. This analysis aims to identify, interpret, and compare genes, their functions, and variations across different organisms or within populations. Genome analysis plays a pivotal part in understanding natural processes, elaboration, and the inheritable base of conditions.

I.2. History

The sequencing of the complete human genome, completed in 2003, revolutionized our understanding of genomics **[2].**

1943-1953: DNA as a carrier of genetic information

1977: Modern DNA sequencing techniques sanger)

1981: Sequencing of the human Mitochondrial genome

1982: First sequence databases (GenBank, EMBL)

1990: Start of the human genome project mapping)

1995: First complete genome of a cellular organism (H. influenzae)

1999: Human chromosome 22

2001: First outline of the human genome

2003: Sequencing of the human genome completed

2007: First complete sequence of the genome of an individual

2021: Revolutionizing molecular biology by integrating omics layers at single-cell resolution and preserving spatial tissue architecture **[3,4].**

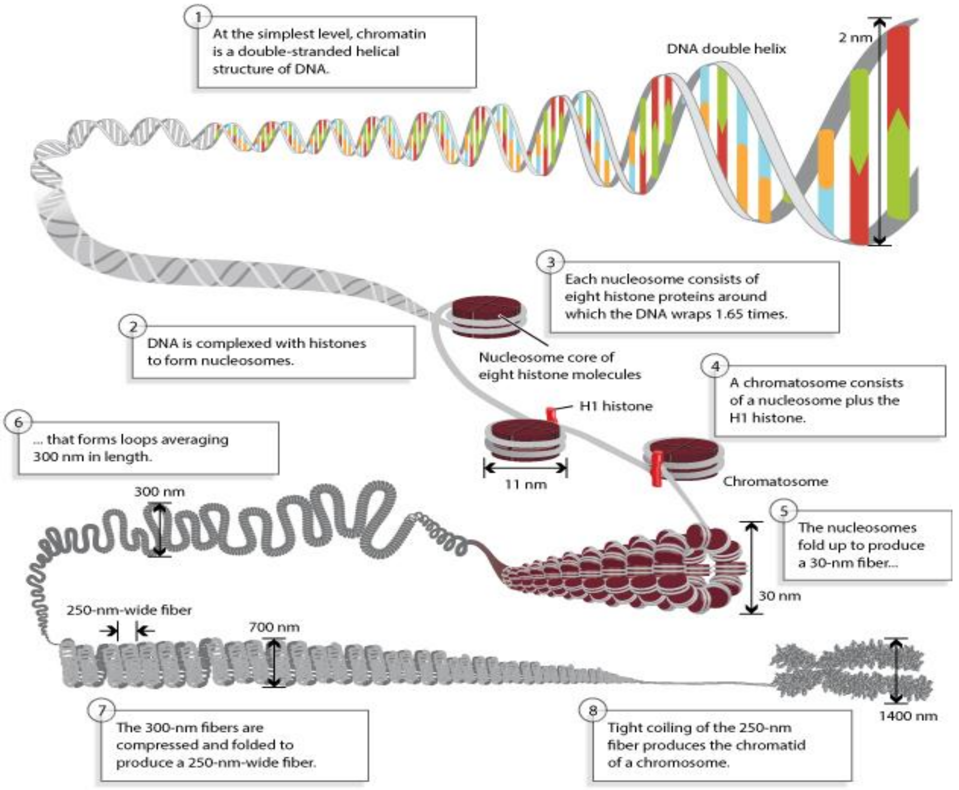

1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
2
DNA is complexed with histones to form nucleosomes.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
6
... that forms loops averaging 300 nm in length.
11 nm
Chromatosome
300 nm
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
1400 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.

Figure I.1: Diagram of the human genome (the double helix structure of DNA) [5].

I.3. Methods and techniques of genome analysis

I.3.1. Sequencing techniques

DNA sequencing constitutes a method whose aim is to determine the linear succession of the bases A, C, G and T taking part in the structure of DNA. Reading this sequence makes it possible to study the biological information contained therein.

There are two types of sequencing:

I.3.1.1. First-generation sequencing***:*** Sanger sequencing was the first widely used method for sequencing DNA, developed in the 1970s **[6].** It relies on the incorporation of dideoxynucleotide triphosphates (ddNTPs) that terminate the chain.

I.3.1.2. Next-generation sequencing (NGS): Next-generation sequencing (NGS) technologies, such as those from PacBio and Illumina, parallel sequencing, enable massive, thereby reducing the cost and time required to sequence a complete genome **[7]**. See further, the details of these two methods (sequencing chapter).

I.3.2. DNA assembly

I.3.2.1. De novo and reference-based assembly***:*** De novo assembly consists of reconstructing a genome without prior reference, using raw sequence fragments (reads). Reference-based assembly compares the reads to a previously sequenced genome **[6,8].**

With technologies still common in many laboratories, each sequencing only yields a read of a few thousand base pairs. Therefore, it is impossible to sequence DNA molecules as large as chromosomes in one go.

In order to reconstruct these immense sequences, a large number of sequencing operations (i.e. greater than the size of the chromosome) must be performed (figure I.2). Sequencing redundancy makes it possible to link the sequences to each other and to ensure the quality of the result of each reading. By connecting all of the contigs, we reconstruct sequences of several million to several tens of millions of nucleotides. These operations are carried out by bioinformatics programs.

I.3.2.2. Assembly tools

Different software like SPAdes **[9]** and Velvet **[10]** are commonly used to assemble genomes from Next-generation sequencing (NGS) data.

I.3.3. Genome annotation: Genome annotation consists of identifying functional elements, such as genes, exons, introns, and regulatory sequences. It allows predicting gene functions in a newly sequenced genome.

Annotation tools such as GenBank and RAST are used to annotate genomes. Theses software perform automatic annotation of sequences by comparing the data to existing databases **[11].**

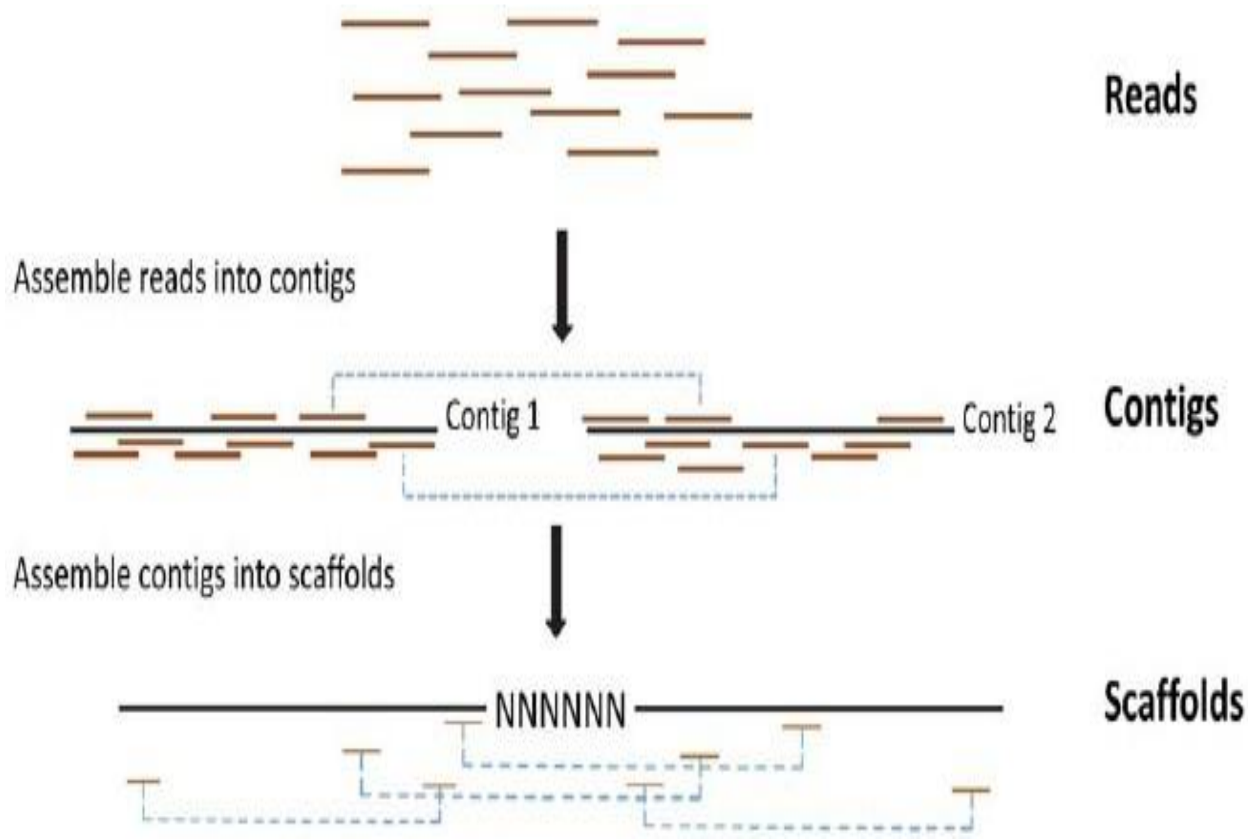


Figure I.2: Genome assembly *[12].*

I.4. Genomic variation analysis

I.4.1. Types of genomic variation

We can find several genomes' variations such as SNPs (Single Nucleotide Polymorphisms) and CNVs (Copy Number Variants) which are common examples of variations in genomes. These variations can affect phenotypic traits or be associated with diseases **[13].**

Genome-wide association studies (GWAS) allow us to study the relationship between specific genetic variations and phenotypic traits in human populations. **[14].**

I.5. Comparison of genomes

I.5.1. Comparative genomics

Comparative genomics is the field that focuses on comparing the genomes of different species in order to identify conserved and divergent genetic elements. By analyzing similarities and differences at the DNA level, this approach provides valuable insights into the evolutionary relationships between organisms and helps elucidate the functions of genes and regulatory elements **[15].**

Selection pressure (evolutionary theory) induces the conservation of sequences that play an important role in the functioning of the organism (genes, regulatory regions, etc.). This allows the detection of functional elements of a new genome compared to a genome that is already well known.

A basic approach in comparative genomics is reciprocal similarity analysis. For example, if gene A″ in genome 2 shows the highest similarity to gene A′ in genome 1, and vice versa, gene A′ in genome 1 is most similar to gene A″ in genome 2, we can conclude that gene A″ is likely the ortholog of gene A′. This reciprocal best-hit strategy is commonly used to predict gene function and homology.

Beyond simple gene-to-gene comparisons, more advanced techniques such as synteny analysis allow researchers to examine the conservation of gene order and chromosomal structure between species. These methods provide deeper insights into genomic rearrangements, duplications, and evolutionary divergence.

To perform such analyses, various bioinformatics tools are commonly used. For instance, BLAST (Basic Local Alignment Search Tool) enables the comparison of nucleotide or protein sequences to identify regions of similarity, while ClustalW allows for multiple sequence alignments, facilitating the detection of conserved motifs and phylogenetic relationships [**16**].

I.6.Applications of genome analysis

I.6.1. Personalized medicine

Genome analysis allows for the development of targeted therapies based on a patient's specific mutations. Advances in identifying disease-causing mutations have led to personalized treatments for diseases such as cancer [**17**].

I.6.2. Genomics in agriculture

Marker-assisted selection (MAS) uses genome analysis to identify genetic markers associated with traits of interest in plants and animals. This approach is used to improve yields and disease resistance **[18].**

II. Construction and validation of genomic databank

II. 1. Introduction to genomic banks

Construction and validation of genomic databank are crucial steps in genetic studies and genome analysis. These banks allow the storage and exploitation of DNA fragments representing all or part of an organism's genome. They are essential for applications such as sequencing, gene cloning or functional analysis.

II.1.1. Definition and importance of genomic banks

A genomic library is a collection of DNA fragments representing all or part of the genome of interest of an organism. It is used to store and manipulate genomic sequences in vectors. The importance of genomic banks lies in their ability to allow the study of genes, their function, and their regulation on a large scale. They are also essential for the mapping and sequencing of new genomes.

II.1.2. Types of genomic banks

There are two types of DNA banks:

II.1.2.1. Genomic DNA banks**:** Collection of clones representing the entire genome of an organism of interest obtained by partial digestion, using one or more restriction enzymes, of genomic DNA (genomic DNA fragments inserted into vectors: plasmids, BACs, YACs).

II.1.2.2. Complementary DNA or cDNA banks**:** Collection of clones representing all the mRNAs present at a given time in a given tissue or organ, allowing the study of expressed genes [**19**].

II.2. Construction of genomic libraries

*II.2.1. Isolation (extraction) of genomic DNA**:*** The first step in constructing a genomic library is to isolate (extract) genomic DNA from the organism of interest (plant, animal, bacteria, etc.). This includes cell lysis, deproteinization, and DNA precipitation [**20**]. DNA is typically isolated from cells or tissues and purified to ensure its quality and integrity (Figure II.1).

II.2.2. DNA fragmentation (digestion by restriction enzymes): Genomic DNA is too large to be manipulated directly, so it is fragmented into smaller pieces. This fragmentation can be done randomly using restriction enzymes, which recognize specific DNA sequences and cut them at specific sites, or by physical methods such as sonication in order to have fragments that can be cloned into vectors **[21]**.

II.2.3. Cloning of fragments: The DNA fragments obtained are inserted into cloning vectors, usually plasmids or bacteriophages, which will allow their amplification and storage. The vectors with the DNA fragments are then introduced into host cells, usually bacteria (such as *E. coli*), where each cell will contain a single DNA fragment [**22**].

II.2.4. Host cell transformation: After ligation of the DNA fragments into the vectors, the host cells (often E. coli) are transformed by electroporation or heat shock, allowing the incorporation of the vectors containing the genomic DNA (Figure II.1)**.**

II.2.5. Colony plating and selection: Transformed cells are plated on selective media containing antibiotics, allowing the selection of clones carrying vectors containing DNA fragments. Each bacterial colony contains a specific copy of a DNA fragment of the genome. Individual clones are then isolated to create a genomic library [**23**].

*II.2.6. Formation of the genomic library***:** Once the clones are selected, they form a genomic library or a DNA library. This library contains a series of clones representing random fragments of the organism's entire genome. This library can then be used to search for specific genes, study regions of the genome or perform sequencing (Figure II.1)**.**

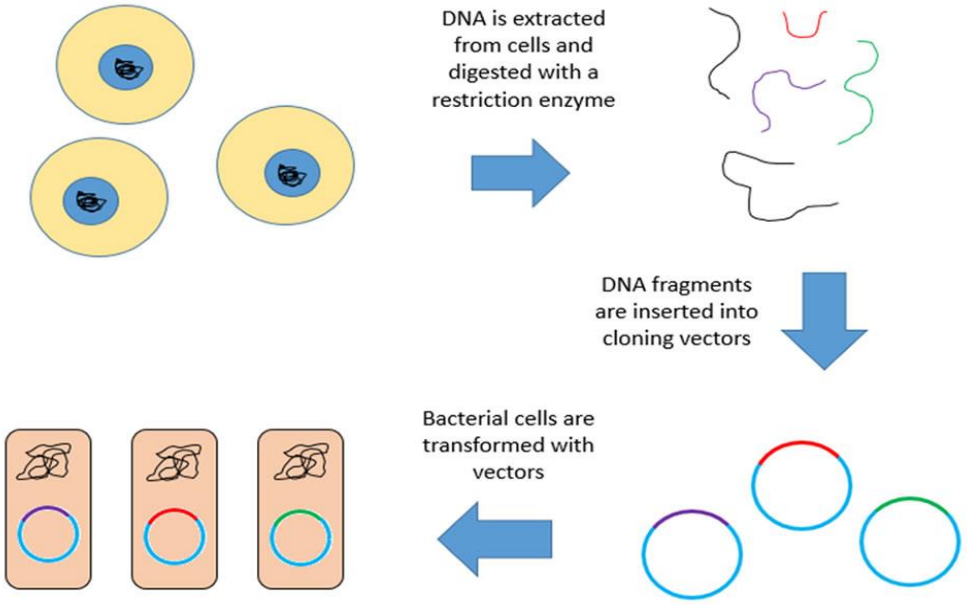

DNA is extracted from cells and digested with a restriction enzyme
DNA fragments are inserted into cloning vectors
Bacterial cells are transformed with vectors

Figure II.1. Steps of genomic library construction. *Adapted from Cepham Life Sciences [24].*

II.3. Validation et analysis of genomic banks

After construction, it’s critical to validate the library to ensure it is representative, accurate, and free from errors. Several steps are typically involved in the validation process:

II.3.1. Quality control of input material: Before constructing the library, the quality of the input DNA or RNA is assessed using techniques such as gel electrophoresis, nanodrop, or bioanalyzer to check for integrity, purity, and concentration.

- For RNA samples, RNA integrity number (RIN) scores are used to ensure the RNA is not degraded.

II.3.2. Size distribution: Fragmentation should yield DNA of the desired size range. This is validated by running a subset of the library on a gel or capillary electrophoresis to check the distribution of fragment sizes. For sequencing libraries, a typical range might be 200–500 bp.

II.3.3. Insert size validation: For cloning-based libraries, sequencing or PCR can be used to confirm that the inserted DNA fragments are of the correct size and match the expected regions of the genome or transcriptome **[25-28].**

II.3.4. Library complexity and coverage: For genomic libraries, coverage refers to how much of the genome is represented in the library. This can be validated by sequencing a small portion of the library to ensure that most of the genome is covered and that no regions are missing or underrepresented.

Library complexity, refers to the number of unique DNA fragments in the library. A high-complexity library should have minimal redundancy, meaning that each clone or sequence represents a different part of the genome.

II.3.5. Quantification: It is essential to quantify the final library, typically done using qPCR (quantitative PCR) or fluorometric methods (e.g., Qubit). This ensures that there is enough DNA to proceed with downstream processes like sequencing or cloning.

II.3.6. Adapter ligation efficiency: For libraries constructed for sequencing, the ligation of adaptors to DNA fragments is validated. Poor ligation can result in low sequencing yields. Adapter ligation can be checked using qPCR or specific sequencing primers.

II.3.7. Functional validation: For expression libraries (e.g., cDNA libraries), the functionality of the library can be tested by expressing the proteins encoded by the cDNA and validating that the expression patterns match expected biological functions.

For certain types of libraries (e.g., CRISPR libraries), functional screens can be performed to ensure the desired constructs are working properly in the context of the host system **[25-28].** Applications of genomic banks

*II.3.8. Genetic mapping and gene cloning***:** Genomic banks are crucial for genetic mapping, particularly for locating and cloning genes of interest. This has allowed, for example, the cloning of genes responsible for human genetic diseases [**29**].

II.3.9. Whole genome sequencing: Genomic banks also allow the sequencing of entire genomes, particularly before the advent of high-throughput sequencing technologies. BAC banks have been used in the sequencing of the human genome **[2].**

II.3.10. Applications in biotechnology and agronomy: Genomic libraries are used in biotechnology for the identification of genes of industrial interest, as well as for the improvement of plants and animals in agricultural genomics [**18**].

III. Mapping and linkage

III.1. Introduction to mapping and linkage

Mapping and linkage are key concepts in genetics that refer to the process of locating genes or genetic markers on chromosomes and understanding their inheritance patterns. These techniques are crucial for identifying the genetic basis of traits and diseases.

Here's an overview of both:

*III.1.1. Genetic mapping***:** in genetics, mapping refers to the process of determining the position of genes, markers, or sequences on a chromosome. The goal is to create a genetic or physical "map" that shows the relative positions of genes or markers. There are different types of genetic mapping, and they vary based on what is being mapped and how the information is obtained.

III.1.1.1. Types of mapping

A- Genetic mapping: Based on the recombination frequency between genes or markers during meiosis. The further apart two genes are on a chromosome, the more likely a recombination event will occur between them. - Genetic maps are measured in centimorgans (cM), where 1 cM corresponds to a 1% chance of recombination between two loci.

B- **Physical mapping**: Refers to the actual physical distance between genes or markers on a chromosome, typically measured in base pairs (bp). Physical maps provide more detailed and precise information compared to genetic maps. - Techniques such as restriction mapping, fluorescence in situ hybridization (FISH), and sequence-based approaches are used for physical mapping.

C- **Comparative mapping**: Involves comparing the genetic maps of different species to understand the conservation of gene order (synteny) and evolutionary relationships. This is useful for transferring knowledge about genes from model organisms to other species.

III.1.1.2. Applications of mapping

- **Gene discovery**: Helps identify the location of genes responsible for diseases or traits by associating markers with phenotypes.

- **Genome assembly**: Assists in constructing the overall sequence of an organism's genome, linking fragments of DNA to their correct locations.

-**Marker-assisted selection**: In agriculture, genetic maps are used to select desirable traits, like disease resistance or yield, by tracking associated markers.

Linkage refers to the tendency of genes or markers that are close to each other on a chromosome to be inherited together during meiosis. The closer two genes are, the less likely they are to be separated by recombination, meaning they are linked.

III.1.2. Key concepts in linkage

- ***Linkage group***: A group of genes or markers on the same chromosome that tend to be inherited together. The entire chromosome can be considered a linkage group, but closely linked genes show strong inheritance patterns.

- ***Recombination***: During meiosis, homologous chromosomes exchange segments in a process called crossing-over. The frequency of recombination between two loci gives an estimate of their genetic distance.

If two genes are located far apart on the chromosome, recombination is more likely to occur between them, meaning they will be inherited independently. If they are close together, they are more likely to be inherited as a unit.

- ***Linkage disequilibrium (LD)***: Describes the non-random association of alleles at different loci. In other words, certain combinations of alleles or genetic variants are inherited together more frequently than expected by chance because of their close proximity.

III.1.3. ***Linkage mapping (Linkage analysis):*** it uses the concept of recombination to estimate the distance between genes or markers. By analyzing how traits are inherited within families or controlled crosses, linkage mapping identifies the relative positions of loci (Figure III.1).

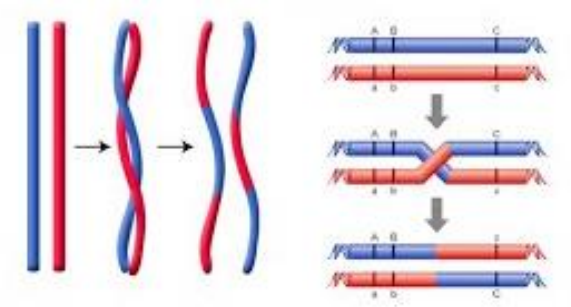


Figure III.1: **Linkage Mapping** *[30].*

III.1.3.1. Steps in linkage mapping

✓ ***Selection of a population***: Family pedigrees, controlled crosses in plants or animals, or natural populations are studied to observe inheritance patterns.

✓ **Genotyping**: Genetic markers, such as SNPs (Single Nucleotide Polymorphisms) or microsatellites, are used to genotype the individuals. The genotypes at these markers are analyzed to detect patterns of co-inheritance with traits of interest.

✓ ***Recombination frequency***: The recombination frequency between markers and traits is calculated. If two markers or a marker and a trait have a low recombination frequency, they are likely to be close to each other on the chromosome.

✓ **LOD scores (Logarithm of the odds (LOD))** score is a statistical measure used to determine whether two loci are linked. A LOD score of 3 or higher is typically considered evidence of linkage, meaning there is a 1000:1 chance that the loci are linked rather than assorting independently.

III.1.3.2. Applications of linkage mapping

- **Identifying disease genes**: Linkage mapping has been critical in finding genes responsible for inherited diseases, such as cystic fibrosis and Huntington's disease.

-**Quantitative trait loci (QTL) mapping**: Identifies regions of the genome associated with complex traits, like height, yield in crops, or susceptibility to diseases. QTL mapping is widely used in plant and animal breeding.

Marker-assisted selection: Like genetic mapping, linkage mapping can help breeders select for beneficial traits by tracking markers linked to those traits.

III.1.4. Differences between mapping and linkage

Mapping is a broader term referring to the general process of determining the positions of genes or markers, either genetically or physically. It encompasses different approaches like genetic mapping, physical mapping, and comparative mapping. Linkage specifically refers to the inheritance pattern of genes or markers that are physically close on the chromosome and tend to be transmitted together. Linkage analysis is a specific type of genetic mapping based on recombination frequencies.

- ✓ **How locations of various genetic markers were determined in chromosomes?**

The recombination frequencies between marker pairs are estimated from suitable mapping populations and are converted to map or genetic distances. Based on the genetic distance, the markers are grouped into linkage groups, and their order in the linkage group is depicted as the linkage map. The formula for estimating recombinant frequency is given hereunder:

$$\text{Recombination frequency} = \frac{\text{Number of recombinant prognies}}{\text{Total number of proginies}} \dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

The map units are centimorgan (cM). One cM is the distance over which, on average, one crossover occurs per meiosis.

- ✓ *Can we use recombinant frequency directly as genetic distance?*

Usage of recombinant frequency directly as the genetic distance will not be possible as in the case of double or more crossovers and interferences. For example, in the case of double crossover parental types will be more than recombinant types that will result in less recombinant frequency. In this case, if we use recombinant frequency directly as the genetic distance will result in false estimation. To overcome this error, mapping functions were used to convert the recombinant frequency to genetic distance.

IV. Sequencing

IV.1. General information on sequencing

Since the description of the structure of DNA in 1955 to the present day, biology has experienced a series of remarkable technological advances, of which sequencing is one of the key events.

Sequencing is a fundamental technique in molecular biology used to determine the exact order of nucleotides (A, T, C, G) in a DNA or RNA molecule. It provides detailed genetic information, which is essential for understanding the structure, function, and evolution of organisms. Sequencing plays a critical role in many areas, such as genomics, medicine, biotechnology, and agriculture **[31].**

IV.2. Types of sequencing

IV.2.1. Sanger sequencing (Chain termination method)

Developed by Frederick Sanger in 1977, this method involves using dideoxynucleotides triphosphates **(ddNTPs)**, which terminate DNA chain elongation when incorporated. It produces DNA fragments of varying lengths.

- Steps*:* The different steps of this technique are as follows:

1. DNA replication is initiated using normal nucleotides (dNTPs) and a small amount of ddNTPs labeled with a marker.

2. The ddNTPs randomly incorporate into the growing DNA chain, causing it to stop elongating.

3. These fragments are separated by electrophoresis and the sequence is read by detecting the labeled ddNTPs (Figure IV.1).

- Applications: While largely replaced for high-throughput sequencing, Sanger sequencing is still used for specific applications like validating small DNA regions or sequencing single genes **[31]**.

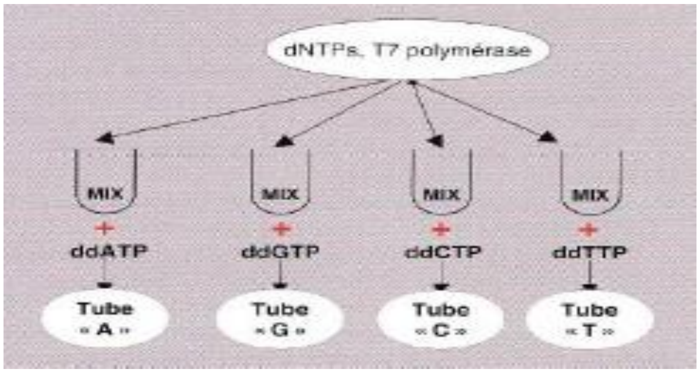


Figure IV.1. Sanger sequencing steps *[32].*

IV.2. 2. Next-generation sequencing (NGS)

Next-generation sequencing NGS is a group of high-throughput sequencing methods that can sequence millions to billions of DNA fragments simultaneously. It's faster, cheaper, and scalable compared to traditional methods. we find there:

IV.2.2.1. Sequencing by synthesis (Illumina): Bridge PCR is a method used in high-throughput sequencing platforms (notably Illumina/Solexa) to amplify DNA fragments *in situ* on a slide. Millions of DNA fragments are immobilized on a solid surface and amplified through bridge amplification cycles. During sequencing, fluorescently labeled nucleotides are incorporated one by one, and a high-resolution camera detects the emitted fluorescence at each step, enabling base-by-base sequence determination (Figure IV.2).

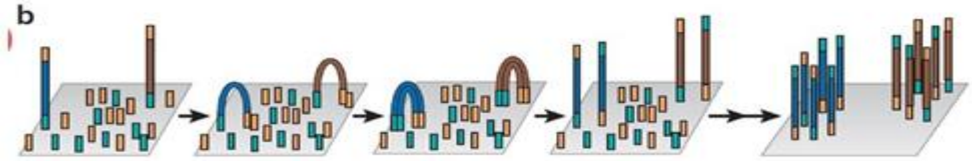
- **Steps illustrated in the figure**

1- **Adapter Attachment:** The DNA fragments (shown in orange and blue) are flanked at their ends by specific adapters. These adapters bind to primers attached to the surface of a slide (shown in light gray), forming a solid attachment base.

2- **Hybridization and Extension (First Loop):** A DNA fragment hybridizes to a complementary primer attached to the surface. DNA polymerase synthesizes a complementary strand from this primer.

3- **Bridge Formation:** The newly synthesized strand remains attached to the surface at one end. It bends (like a bridge, hence the name) to hybridize to a second attached primer.

4- **Cyclic Amplification:** This process is repeated in several cycles. At each cycle, the fragments duplicate exponentially, forming clusters of cloned DNA (colonies visible in orange and blue in the last image).



- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by illumina/Solexa.

Figure IV.2: Diagram illustrating the principle of bridge PCR used in Illumina sequencing technologies *[33].*
DNA strands are amplified directly on a slide through fixed primers, forming clusters. Adapted from Illumina, Inc. (n.d.)**.**

- **Applications**: Widely used for whole-genome sequencing (WGS), whole-exome sequencing (WES), RNA sequencing (RNA-seq), and microbial metagenomics.

IV.2.2.2. Ion torrent sequencing: Measures changes in pH when nucleotides are incorporated into the DNA strand. No fluorescent markers are used; instead, each addition of a nucleotide releases an ion, changing the pH, which is measured to determine the sequence (Figure IV.3).

- **Applications**: Suitable for smaller sequencing projects like bacterial genomes or targeted sequencing.

IV.2.2.3. Pyrosequencing: Based on detecting the release of pyrophosphate (PPi) during nucleotide incorporation. This triggers a series of reactions that generate light, which is detected to infer the sequence (figure IV.4).

- **Applications**: Used for short-read sequencing applications requiring high accuracy.

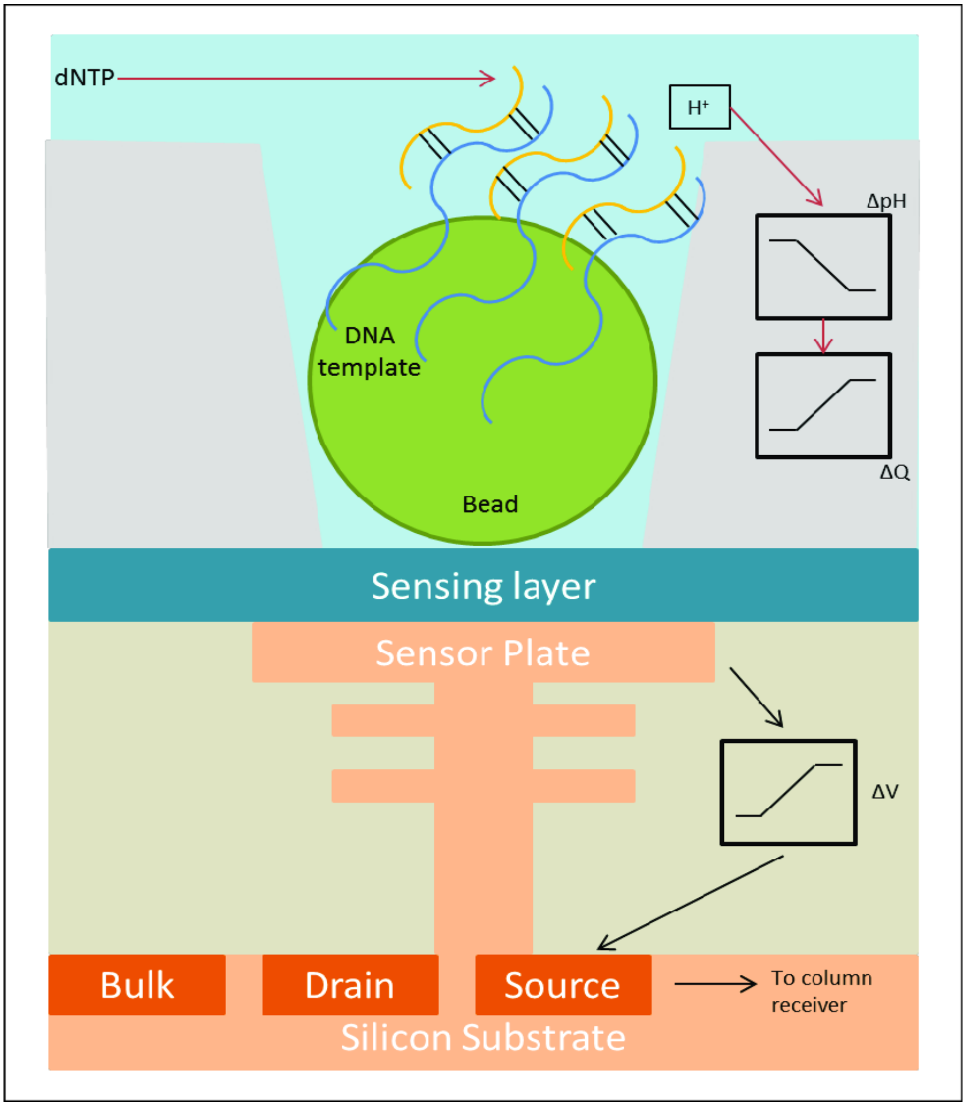


Figure IV.3: Schematic representation of a single well of an Ion Torrent sequencing chip *[34]*. The well harbors Ion Spheres particles containing DNA template. When a nucleotide incorporates, a proton (H+) is released and the pH of the well changes (ΔpH). A sensing layer detects this change of charges (ΔQ) and translates the chemical signal into a digital signal (ΔV).

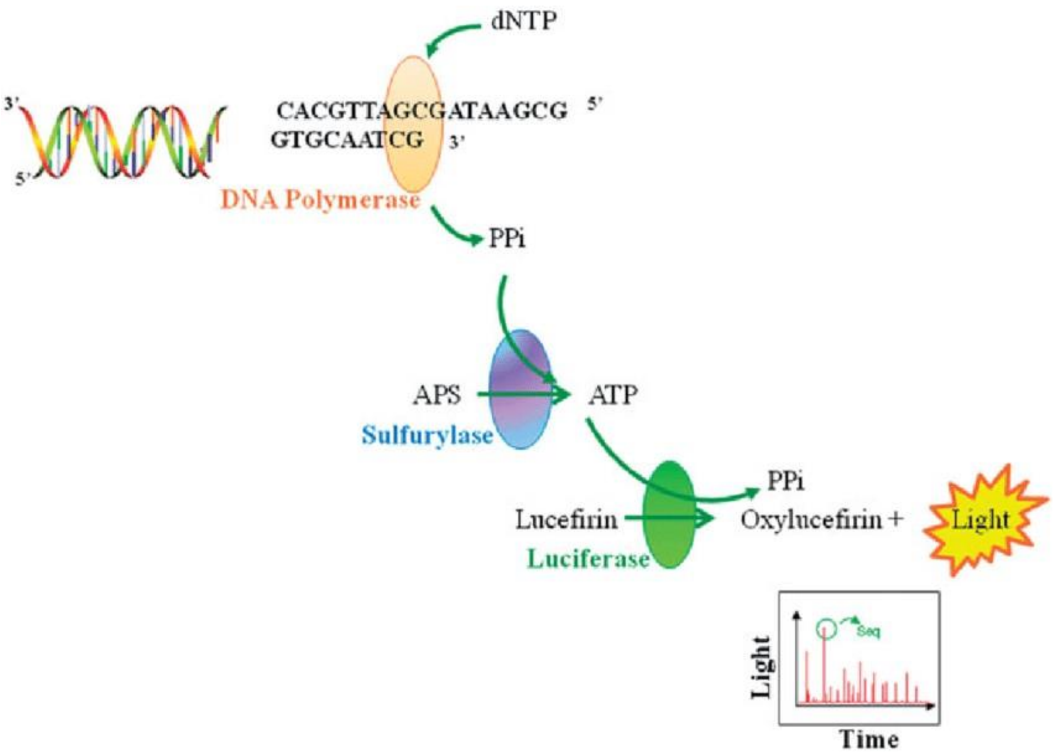


Figure IV.4: Pyrosequencing principle *[35].*

IV.2. 3. Third-generation sequencing

These technologies can sequence long DNA or RNA molecules in real-time, which is particularly useful for studying complex genomic regions.

IV.2.3. 1.. Nanopore sequencing (oxford nanopore): A single DNA molecule is passed through a nanopore embedded in a membrane. As nucleotides pass through the pore, they cause specific disruptions in the electric current, which is read to determine the sequence.

- **Applications**: Suitable for long-read sequencing, detecting epigenetic modifications, and sequencing samples in remote locations (portable, fast).

IV.2. 3. 2. PacBio (Single Molecule Real-Time, SMRT) sequencing: This technique sequences a single DNA molecule in real-time. As nucleotides are incorporated into the growing DNA strand, fluorescent signals specific to each base are recorded.

- **Applications**: Ideal for long-read sequencing, resolving highly repetitive regions of the genome, and detecting structural variations.

IV.3. Applications of sequencing

IV.3.1. Genomics (Whole-Genome Sequencing (WGS)): Determines the complete DNA sequence of an organism. It is used for studying genetic variation, discovering new genes, and understanding evolutionary relationships.

IV.3.2. Personalized medicine: Sequencing is used clinically to identify mutations responsible for genetic diseases, predict an individual's response to drugs (pharmacogenomics), and assess the risk of developing certain conditions.

IV.3.3. Transcriptomics (RNA-seq): Sequencing RNA reveals which genes are being expressed in a cell or tissue at a given time, providing insights into gene regulation, development, and disease mechanisms.

IV.3.4. Microbiome and metagenomics: Sequencing is used to analyze the genetic material of microbial communities from various environments (soil, water, human body), providing insights into biodiversity and ecological function.

IV.3. 5. Evolution and population genetics: Sequencing helps trace evolutionary relationships between species, understand genetic diversity within populations, and identify natural selection pressures.

IV.3.6. Exome sequencing: Focuses on sequencing the exome, which includes the coding regions of the genome (the exons). Exome sequencing is often used to identify mutations responsible for genetic disorders

IV. 4. Innovations and challenges in sequencing

IV. 4. 1. Cost reduction: The cost of sequencing has dropped dramatically, making it accessible for large-scale projects such as sequencing the human genome. For example, the cost of sequencing a human genome has fallen from about $3 billion in 2003 to under $1,000 today.

IV. 4. 2. Data analysis: The major challenge in modern sequencing is not generating the data, but analyzing and interpreting the massive amounts of information generated. Bioinformatics tools are essential for managing, visualizing, and understanding sequencing data.

IV. 4. 3. Structural variation detection: Newer sequencing technologies are better equipped to detect large-scale genetic variations (such as deletions, duplications, inversions), which are important for understanding diseases like cancer.

V. Systematic expression and interaction studies

Systematic expression and interaction studies are comprehensive approaches used to investigate gene expression patterns and protein interactions within an organism. These studies provide a global view of cellular processes, helping to decipher how genes are regulated and how proteins interact to perform biological functions. They are critical for understanding molecular networks, disease mechanisms, and the development of new therapeutic strategies.

V.1.Systematic expression studies (transcriptomics)

Systematic studies of gene expression, commonly known as transcriptomics, involve analyzing all RNA transcripts present in a cell or tissue at a given time. These studies are crucial for understanding which genes are active or repressed under various conditions, such as different environmental stimuli, developmental stages, or disease states.

V.1.1. Main analysis techniques used

V.1.1.1. RNA-seq (RNA sequencing): RNA-seq uses high-throughput sequencing to analyze the entire transcriptome, providing a quantitative view of gene expression. It allows the identification of differentially expressed genes, transcript variants, and post-transcriptional modifications.

- **Applications**: Used to study gene expression in various conditions (e.g., stress, drug treatment), compare normal and diseased tissues (e.g., cancer vs. healthy cells), and analyze gene expression during development or immune responses.

V.1.1. 2. Microarrays: Microarrays use DNA probes on a chip to hybridize with RNA transcripts. The intensity of the fluorescence signal corresponds to the level of gene expression.

- **Applications**: Although less common today due to the rise of RNA-seq, microarrays remain a cost- effective option for large-scale expression studies.

V.1.1.3. qPCR (Quantitative Polymerase Chain Reaction**)**: qPCR provides precise quantification of specific RNA molecules. It is often used to validate findings from RNA-seq or microarray studies.

- Applications: Used for measuring the expression of key genes, such as biomarkers in cancer or genes involved in specific cellular responses.

V.1.2. Benefits of systematic expression studies

- **Global view of gene activity**: These methods provide a comprehensive snapshot of which genes are turned on or off in different biological contexts.

- **Comparative studies**: They enable comparisons of gene expression between healthy and diseased states, aiding in the identification of potential therapeutic targets.

V.2. Systematic interaction studies (interactomics)

Interaction studies, also known as "interactomics", focus on analyzing the protein-protein interactions (PPIs) and other molecular interactions (e.g., RNA-protein, protein-DNA) that occur in cells. These interactions are key to understanding how cellular processes are coordinated, as proteins often work together in complexes or signaling pathways.

V.2.1. Main analysis techniques used

V.2.1.1. Yeast Two-Hybrid (Y2H): This technique detects interactions between two proteins in yeast cells. If two proteins interact, they reconstitute a functional transcription factor that activates a reporter gene, which is then detected.

-**Applications**: Used to map protein interaction networks on a large scale, often applied in model organisms to understand cellular pathways.

VCo-immunoprecipitation (Co-IP): Co-IP uses antibodies to capture a target protein and any interacting partners. The proteins are then analyzed by methods like mass spectrometry or Western blot.

- **Applications**: Commonly used to confirm interactions between proteins or to study specific protein complexes in different cellular conditions.

V.2.1.2. Mass spectrometry-based proteomics: Mass spectrometry allows the identification and quantification of proteins and their interactions. It is used to detect and characterize large-scale protein complexes or post-translational modifications.

-**Applications**: Widely used for systematic mapping of protein interaction networks in cells and for studying how signaling pathways change in response to different stimuli.

V.2.1.3. RNA-protein interaction (RIP, CLIP, PAR-CLIP): These methods identify interactions between RNA molecules and proteins using immunoprecipitation followed by RNA sequencing.

- **Applications**: Used to map interactions between RNA-binding proteins and their target RNAs, which are critical for gene regulation, splicing, and other post-transcriptional processes.

V.2.2. Benefits of systematic interaction studies

- **Mapping biological networks**: These methods allow researchers to map complex networks of interactions within cells, helping to understand how proteins and other molecules cooperate to regulate cellular processes.

- **Target discovery**: Identifying key interaction hubs or nodes in this network can reveal potential therapeutic targets, especially in diseases where protein interactions are altered.

V.3. Combined approaches (expression and interaction studies)

V.3.1. Gene expression and interaction networks: By integrating gene expression and protein interaction data, researchers can build models of how changes in gene expression affect protein-protein interactions and biological pathways. For example, transcriptomic data may reveal which genes are overexpressed in a cancer cell, while interactomics studies can show how these genes influence signaling networks and cellular responses.

V.3. 2. Transcriptional regulation networks: Transcription factors (TFs) are proteins that regulate gene expression by binding to specific DNA sequences. Combining interactomic studies with transcriptomic data can provide insights into how transcription factors and their co-regulators control gene expression. These approaches are essential for understanding developmental processes, immune responses, and disease progression.

V.4. High-throughput screening (HTS)

HTS involves the rapid testing of thousands of samples for activity against a target. In the context of protein interactions, it can identify genes or proteins that regulate specific pathways.

CRISPR and RNAi Screening techniques are used [**36**]:

- **CRISPR Screening**: Using CRISPR-Cas9 for genome-wide screens;
- **RNAi Screening**: Using RNA interference to knock down gene expression in a systematic way.

V.5. Quantitative methods for analyzing interactions

Quantitative approaches are essential for understanding the strength and dynamics of interactions. **Surface Plasmon Resonance (SPR):** A technique used to measure the binding affinity between proteins. **Fluorescence Resonance Energy Transfer (FRET):** Used to study protein interactions in live cells [**37**].

V.6. Applications of systematic expression and interaction studies

V.6.1. Cancer research: Identifying oncogenes and tumor suppressors through differential expression studies, while interaction studies can reveal how these genes alter signaling pathways or protein complexes involved in cell growth and survival.

V.6.2. Drug development: Mapping interaction networks helps identify drug targets by pinpointing critical proteins in disease pathways. Expression studies allow researchers to track how these proteins change in response to treatment, aiding in the design of more effective therapies.

V.6.3. Understanding genetic diseases: Inherited disorders often arise from mutations that affect both gene expression and protein interactions. Systematic studies can elucidate how genetic mutations disrupt normal cellular functions, providing insights into potential therapeutic approaches.

*V.6. 4. Functional genomics***:** These studies are used to assign functions to newly discovered genes or proteins by analyzing their expression patterns and interaction partners. This is especially important in model organisms like yeast, mice, and fruit flies.

VI. Transcriptome, ChIp, ChIp on chip

In this chapter, we will discuss the notions of transcriptome, ChIP (Chromatin Immunoprecipitation), and ChIP-on-chip, focusing on their definitions, methodologies, applications, and relevance in genomics and molecular biology.

VI.1. Transcriptome definition

The transcriptome refers to the complete set of RNA molecules, including messenger RNA (mRNA) and non-coding RNA, that are transcribed from the DNA of an organism or specific tissue at a given time. The study of the transcriptome helps researchers understand gene expression and how genes are regulated under various biological conditions.

VI.1.1. Analysis techniques

VI.1.1. 1.Microarray: utilizes hybridization to measure gene expression levels. It consists of probes corresponding to known genes or genomic sequences and provides relative quantification of gene expression.

VI.1.1. 2. RNA-seq: A high-throughput sequencing technique that offers a comprehensive view of the transcriptome, including the identification of novel transcripts, alternative splicing events, and the quantification of transcript levels with higher sensitivity and specificity compared to microarrays **[38].**

-Applications: Transcriptome analysis helps in identifying gene expression changes during development, differentiation, and disease. For example, studying cancer transcriptomes has revealed novel biomarkers and therapeutic targets **[39].**

VI.2. Chromatin immunoprecipitation (ChIP) definition

ChIP is a powerful technique for mapping protein-DNA interactions *in vivo*, allowing researchers to identify specific DNA regions bound by proteins, such as transcription factors or histone modifications [**40**].

VI.2. 1. *Chromatin immunoprecipitation (ChIP) steps*

a- **Crosslinking**: Formaldehyde is typically used to crosslink proteins to DNA, preserving the protein-DNA interaction.

b- **Chromatin shearing**: This step involves breaking the chromatin into smaller fragments, usually by sonication or enzymatic digestion.

c- **Immunoprecipitation**: An antibody specific to the protein of interest is used to precipitate the protein-DNA complex.

d- **Purification**: The crosslinks are reversed, and the DNA is purified for downstream analysis using quantitative PCR (qPCR), microarrays, or sequencing (ChIP-Seq) [**41**].

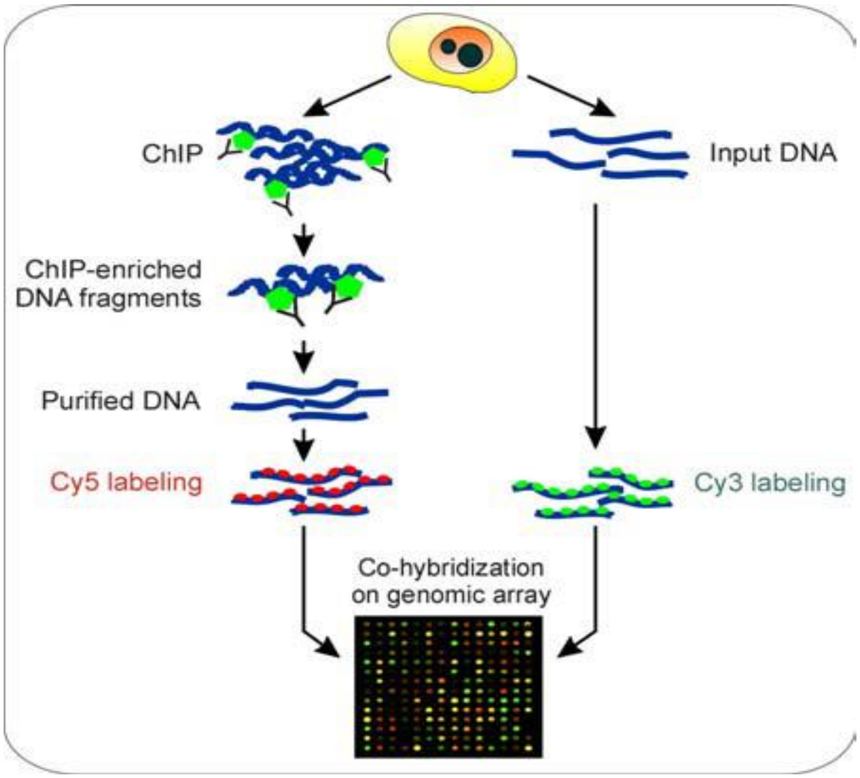
-**Applications**: ChIP is widely used to map transcription factor binding sites and histone modifications, aiding in the understanding of gene regulatory mechanisms. For example, ChIP has been essential in studying epigenetic changes in various diseases [**42**].

*VI.3. **ChIP-on-Chip (ChIP-Chip) definition***

ChIP-on-chip is a genome-wide approach combining ChIP with DNA microarrays to identify protein-DNA interactions across the entire genome [**43**].

VI.3.1. ChIP-on-Chip process:

Following the standard ChIP protocol, the enriched DNA is labeled and hybridized onto a microarray containing probes corresponding to specific genomic regions. The microarray data provides a readout of the regions where the protein binds to the DNA (Figure VI.1).


ChIP
Input DNA
ChIP-enriched DNA fragments
Purified DNA
Cy5 labeling
Cy3 labeling
Co-hybridization on genomic array

Figures VI.1: Chip-on-Chip technique *[44].*

The ChIP-chip approach. A protein of interest (green pentagon) is selectively immunoprecipitated by ChIP. The ChIP-enriched DNA is amplified by PCR and fluorescently labeled with, e.g., Cy5. An aliquot of purified input DNA is labeled with another fluorophore, e.g., Cy3. The two samples are mixed and hybridized onto a microarray containing genomic probes covering the whole genome or to a high-resolution tiling array covering a region of interest. Binding of the precipitated protein to a target site is inferred when intensity of the ChIP DNA (red Cy5 labeling) significantly exceeds that of the input DNA (green Cy3 labeling) on the array.

VI.3.2. Technique's advantages

ChIP-on-chip allows for the study of protein-DNA interactions across the entire genome, offering an advantage over traditional ChIP when only small portions of the genome are targeted. However, it has largely been replaced by ChIP-Seq due to the latter's higher resolution and sensitivity.

-**Applications**: ChIP-on-chip has been used to map global transcription factor binding sites and histone modifications, particularly in model organisms like yeast and Drosophila [**45**].

Innovative work illustrating the evolution of ChIP-on-chip based methods towards more advanced technologies such as ChIP DIP, which allows the simultaneous analysis of hundreds of protein-DNA interactions at the whole genome scale **[46].**

Despite the rise of ChIP-Seq, ChIP-on-chip is still useful in certain contexts where microarrays may be more accessible.

VII. Proteomics

Proteomics is considered as the large-scale study of proteins, the key functional molecules in biological systems. In opposition to the genome which remains identical and constant in every cell of an organism, the proteome is changing (dynamic) depending on conditions, developmental phases or cell types.

The objectives of proteomics can be described as:

- **Protein identification**: In a biological sample, determination of the identity and composition of proteins is available.
- **Quantification of proteins**: The relative abundance of proteins in different samples or conditions is measured.
- **Study of post-translational modifications (PTMs)**: Characterization and identification of different modifications such as glycosylation, phosphorylation, that occur after protein synthesis.
- **Analysis of protein-protein interactions**: This analysis allows us to understand the interaction that exists between proteins and other biomolecules, which is essential for cellular signaling and function.

VII. 1. Definition and scope of proteomics

Proteomics refers to the comprehensive study of the proteome, which is the complete set of proteins expressed by a genome, cell, tissue, or organism at a given time. This field encompasses the study of protein expression, modification, localization, interactions, and functions [**47**].

Proteins are responsible for most cellular processes, and understanding the proteome is critical for elucidating mechanisms in health and disease. Due to its dynamic nature, the proteome reflects the functional state of the cell better than the genome or transcriptome [**48**].

VII. 2. Proteomic techniques

Proteomics involves a variety of technologies for protein identification, quantification, and functional analysis.

VII.2.1. Protein identification

a- **Mass spectrometry (MS):** This is the cornerstone of proteomics, allowing for the identification and quantification of proteins in complex mixtures. MS-based proteomics involves protein digestion into

peptides, ionization, and analysis based on mass-to-charge ratio. Key advancements like tandem MS (MS/MS) and high-resolution MS have revolutionized proteomics [**49**].

b- **Two-Dimensional Gel Electrophoresis (2D-GE):** Proteins are separated by isoelectric point in one dimension and by molecular weight in the second. Proteins are then visualized and identified using techniques like MS. Although powerful, 2D-GE has been largely replaced by more sensitive and high-throughput methods [**50**].

VII.2.2. Protein quantification

a- **Label-Free Quantification (LFQ):** Quantification of protein abundance based on MS signal intensity or spectral counting without the use of labeling.

b- **Isobaric Tag for Relative and Absolute Quantitation (iTRAQ) and Tandem Mass Tags (TMT):** Chemical labeling techniques that allow simultaneous quantification of proteins in multiple samples by comparing MS signal intensities **[51].**

c- **Post-Translational Modifications (PTMs):** PTMs like phosphorylation, ubiquitination, and glycosylation are essential for regulating protein function. Specialized proteomic methods, such as phosphoproteomics, aim to map and quantify these modifications in cells, which is crucial for understanding cellular signaling pathways **[52].**

VII. 3. Functional and interaction proteomics

VII. 3.1. Protein-Protein Interactions (PPIs)**:** Understanding PPIs is fundamental to elucidating protein function in cellular processes. Techniques like Yeast Two-Hybrid (Y2H), Affinity Purification-Mass Spectrometry (AP-MS), and Proximity Ligation Assays (PLA) allow for the identification and mapping of protein networks [**53**].

VII. 3.2. Structural proteomics: Structural proteomics aims to determine the three-dimensional structures of proteins and protein complexes. Techniques such as X-ray crystallography, Nuclear Magnetic Resonance

(NMR), and Cryo-Electron Microscopy (Cryo-EM) are used to analyze protein structure at high resolution [**54**].

VII. 4. Applications of proteomics

a- **Disease biomarker discovery**: Proteomics has become a key tool in identifying disease-specific biomarkers, particularly in cancer, cardiovascular diseases, and neurodegenerative disorders. By comparing protein profiles of diseased and healthy tissues, researchers can identify potential diagnostic

and therapeutic targets [**55**].

b- **Personalized medicine**: The identification of protein signatures associated with specific diseases can lead to more precise diagnostics and tailored therapies based on an individual's proteome. This is especially promising in oncology, where proteomics can reveal the heterogeneity of tumors [**56**].

c- **Drug development**: Proteomic techniques can help identify drug targets, understand drug mechanisms of action, and study drug resistance mechanisms. The integration of proteomics into the drug development pipeline is enhancing the discovery of new therapeutics [**57**].

VII. 5. Challenges in proteomics

a- **Proteome complexity**: The dynamic range of protein concentrations, varying from highly abundant proteins to low-abundance signaling molecules, poses a significant challenge. Proteins also undergo extensive modifications and exist in diverse isoforms.

b- **Data integration**: With the vast amount of data generated from proteomics experiments, computational tools and bioinformatics are essential for data analysis and interpretation. Proteomic data must often be integrated with genomic, transcriptomic, and metabolomic data to gain a comprehensive understanding of biological systems [**58**].

VIII. The double hybrid approach

VIII .1. Generalities

In proteomics, the double hybrid approach generally refers to a strategy that combines multiple techniques to analyze protein structures, functions, and interactions. While the term "double hybrid" is more commonly associated with quantum chemistry, in the context of proteomics, a similar strategy may involve combining mass spectrometry (MS) with other high-throughput techniques such as yeast two-hybrid (Y2H) systems or chromatography-based methods to gain more comprehensive insights into protein-protein interactions, post-translational modifications, and protein quantification.

VIII .2. Double hybrid approach in proteomics

VIII.2.1. Mass Spectrometry (MS) and Chromatography

One of the most powerful techniques in proteomics is the combination of liquid chromatography (LC) with tandem mass spectrometry (LC-MS/MS). This approach separates complex protein mixtures using chromatography before identifying and quantifying proteins via mass spectrometry. The double hybrid aspect here refers to the synergy between separation techniques (like LC) and analytical methods (like MS).

VIII.2.2. Mass spectrometry and yeast two-hybrid (Y2H) systems

The Yeast Two-Hybrid (Y2H) system is a molecular biology technique used to discover protein-protein interactions by expressing two proteins in yeast cells to see if they interact. When combined with mass spectrometry for protein identification and interaction mapping, this hybrid technique provides a robust method for mapping out complex interaction networks (Figure VIII.1).

VIII.2.3. Quantitative proteomics

Stable isotope labeling by amino acids in cell culture (SILAC) and isobaric tags for relative and absolute quantitation (iTRAQ) are examples of techniques that, when used alongside mass spectrometry, form a "double hybrid" approach. SILAC and iTRAQ introduce isotopic labels into proteins that can be tracked during MS analysis, allowing precise quantification of proteins in different biological conditions or states.

VIII .3. Applications in proteomics

The double hybrid approach in proteomics enables comprehensive analysis of biological systems, such as:

a- **Protein identification and quantification**: Through MS-based proteomics combined with chromatographic separation techniques.

b- **Protein-protein interaction mapping:** By integrating Y2H assays and MS to identify interactions in cellular pathways.

c- Post-translational modifications: Using MS in combination with affinity purification to study modifications like phosphorylation, ubiquitination, etc.

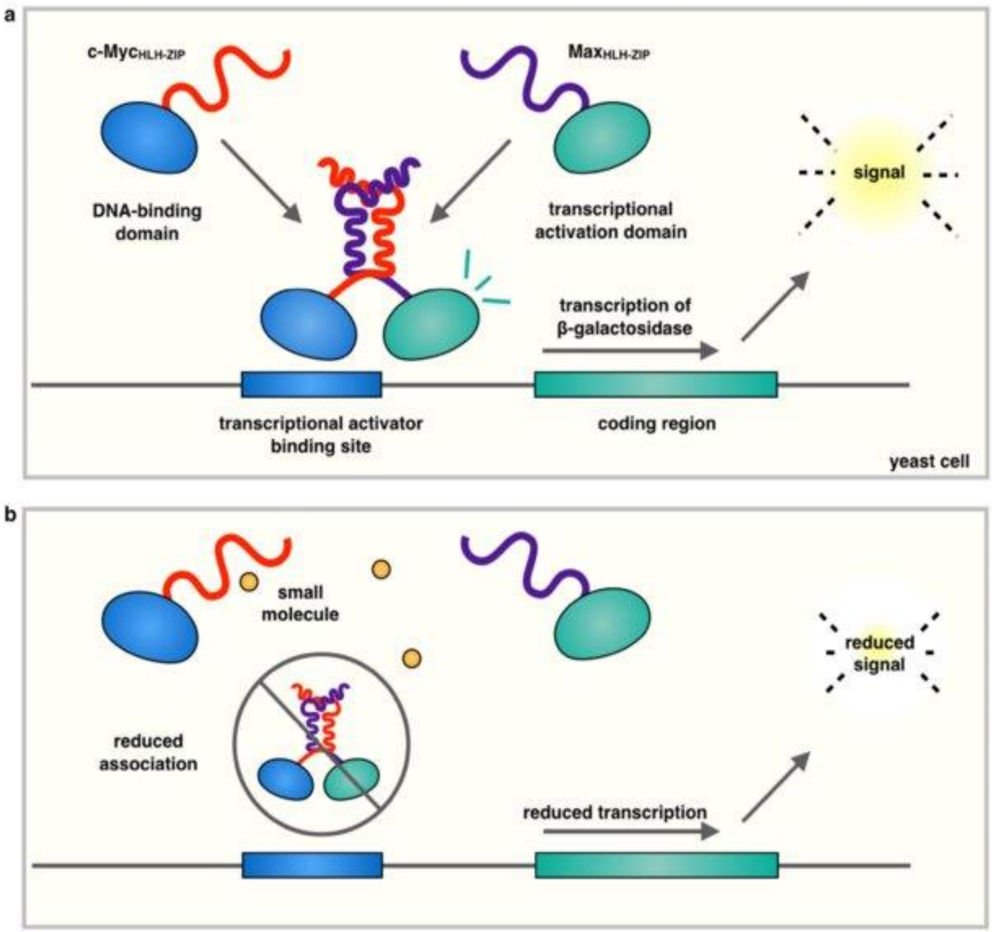


Figure VIII.1: Yeast Two-Hybrid (Y2H) system principle *[59]*.

In the type of yeast two-hybrid system used to identify inhibitors of c-Myc/Max dimerization, recombinant genes encoding the HLHZip domain of c-Myc fused to the DNA-binding domain and HLHZip domain of Max fused to the transcriptional activation domain are introduced into a yeast cell (a). Upon c-Myc/Max association, the transcriptional activation domain induces expression of b-galactosidase in a quantitative manner (b)

VIII .4. Example workflows

a- LC-MS/MS proteomics: Proteins are first digested into peptides. These peptides are then separated via liquid chromatography and analyzed using tandem mass spectrometry, allowing for both identification and quantification.

b- Y2H + MS proteomics: The Y2H system is used to identify potential protein-protein interactions. These interactions are then validated and characterized using mass spectrometry.

Aebersold & Mann [**60**], pioneers in proteomics, highlight the integration of MS with other technologies as crucial for understanding protein functions. Whereas, Gavin et al. [**53**], discuss the use of hybrid approaches like Y2H combined with MS for large-scale protein interaction studies.

By applying these double hybrid techniques, proteomics research advances the understanding of protein networks, signaling pathways, and the proteome's dynamic behavior in health and disease.

IX. The TAP-tag technique

IX.1. Generalities

The Tandem Affinity Purification (TAP)-tag technique is a widely used method in proteomics for isolating and identifying protein complexes under near-physiological conditions. It was first introduced as a way to purify native protein complexes with high specificity and minimal contamination, allowing researchers to study protein-protein interactions and complex compositions in a relatively unbiased manner.

IX 2. TAP-tag technique overview

The TAP-tag technique involves genetically fusing a tandem affinity purification tag to the protein of interest. This tag consists of two affinity tags separated by a protease cleavage site. These tags facilitate a two-step purification process, which helps in minimizing contaminants and isolating proteins and their interaction partners with high purity.

IX 3. Components of the TAP-tag

This technique is compounded of:

1. Protein A (for IgG binding).
2. Calmodulin-Binding Peptide (CBP) (for binding to calmodulin in the presence of calcium).
3. TEV protease cleavage site (to release the tagged protein between the two purification steps).

IX 4. TAP-tag purification process

a. **First affinity purification (IgG):** The tagged protein, along with its interacting partners, is expressed in cells. The protein complex is then extracted and passed through a column containing IgG. The Protein A tag binds to IgG, and after washing away non-specific proteins, the complex is eluted using TEV protease, which cleaves the tag at the TEV site (Figure IX.1).

b. **Second affinity purification (calmodulin):** The eluate is subjected to a second purification step involving calmodulin beads in the presence of calcium. The Calmodulin-Binding Peptide (CBP) binds to the calmodulin beads, and after additional washing to remove contaminants, the protein complex is eluted using a calcium chelator (such as EDTA), releasing the purified complex (Figure IX.1).

This sequential purification ensures that only specific interactors of the tagged protein are isolated, reducing background noise and improving the reliability of downstream analysis, such as mass spectrometry (MS).

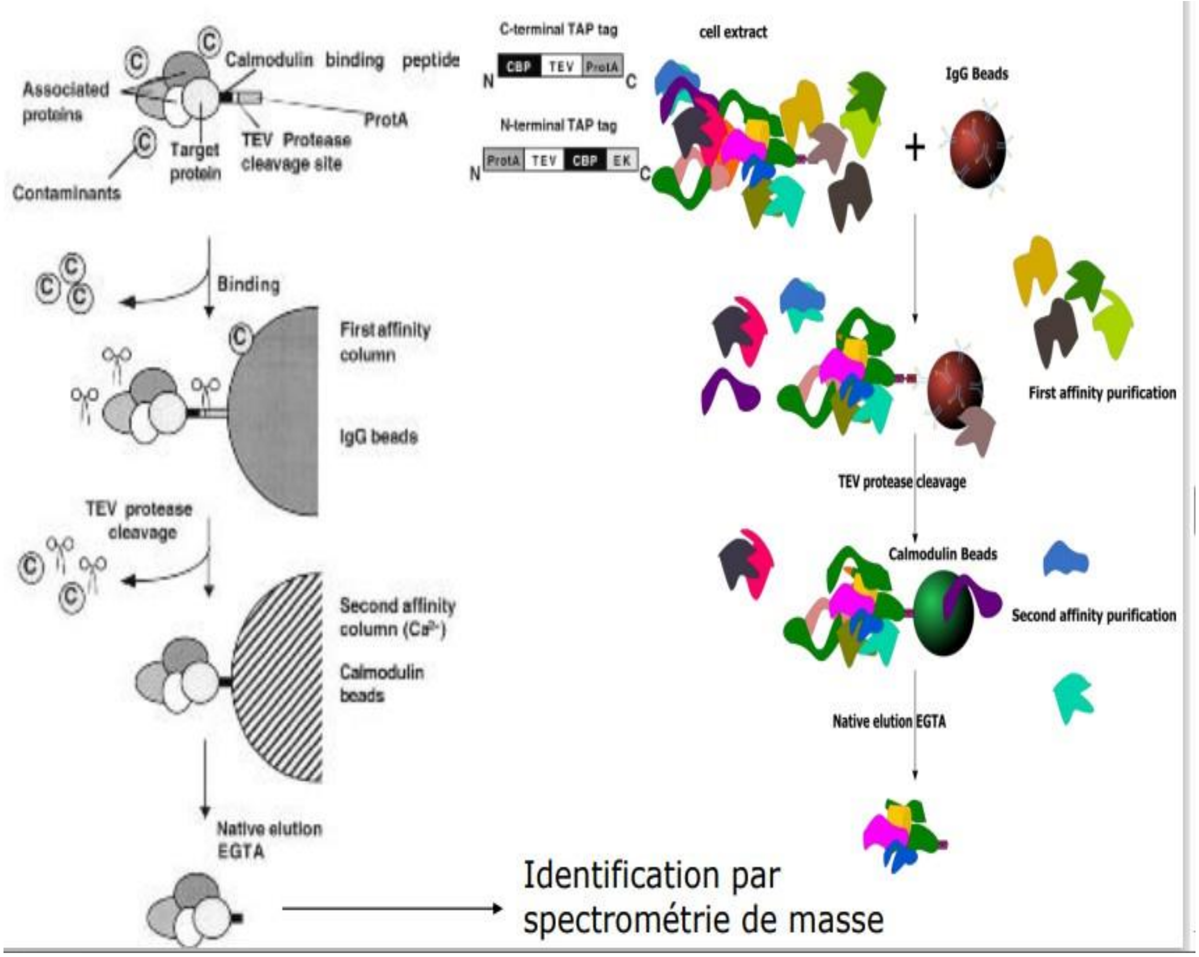


Figure IX.1: TAP-tag process *[61].*

IX 5. Applications of TAP-tag in proteomics

a-Identification of protein complexes: The TAP-tag method allows for the isolation of multi-protein complexes in their native state, which can then be identified and characterized by mass spectrometry (MS).

B-protein-protein interaction networks: TAP-tag is often used in large-scale studies to map protein- protein interaction networks within cells, revealing how proteins collaborate to perform biological functions.

C-Functional proteomics: This method helps in understanding how protein complexes change in response to various conditions or treatments, providing insight into cellular signaling pathways and other dynamic processes.

IX 6. Advantages of TAP-tag

a- High purity: The two-step purification process reduces non-specific protein binding and ensures high-purity isolation of protein complexes.

b- Minimal perturbation: Since the TAP tag is relatively small and purification is done under mild conditions, native protein interactions are preserved.

c- Versatility: The technique can be applied to a wide range of organisms, from yeast to mammalian cells.

Rigaut et al. [**61**], originally developed the TAP-tagging method for studying protein complexes in yeast, demonstrating its effectiveness in isolating native complexes. Otherwise, Gavin et al. [**62**], used TAP-tag to map protein interaction networks in yeast, revealing insights into the modularity and organization of the yeast proteome.

- Example of TAP-tag application: In yeast, researchers used TAP-tagging to identify over 500 protein complexes and more than 1,700 proteins, illustrating the technique’s power to unravel complex cellular interaction networks.

X. Systematic approaches to expression disruption

X.1. Systematic approaches to expression disruption definition

Systematic approaches to expression disruption are techniques employed in molecular biology to study gene function by altering their expression. These methods allow the analysis of the consequences of activating or repressing specific genes on biological networks, metabolic pathways, and cellular behaviors.

X.2. Main approaches to expression disruption

X.2.1. RNA interference (RNAi)

RNAi is a cellular mechanism that uses small double-stranded RNAs such as small interfering (siRNAs), to specifically degrade target mRNAs, thereby inhibiting gene expression. The process (Figure X.1) involves the cutting of double-stranded RNA into siRNAs by the enzyme Dicer, and then loading these siRNAs into the RISC complex, which recognizes and cleaves the target mRNA [**63**].

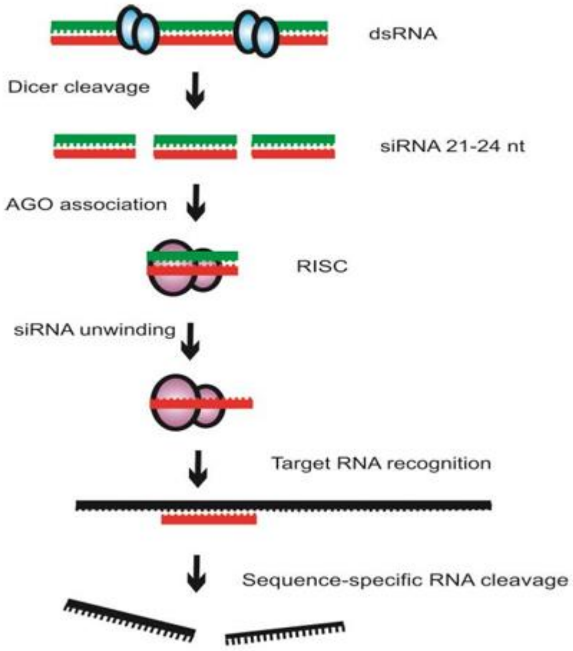


Figure X.1: Mechanism of RNA interference (RNAi) *[63]*.

RNAi is initiated by the enzyme Dicer that cleaves Figure 2. Mechanism of RNA interference (RNAi). RNAi is initiated by the enzyme Dicer that cleaves double-stranded RNA (dsRNA) into short fragments of approximately 21- to 24-nucleotide double-stranded RNA (dsRNA) into short fragments of approximately 21- to 24-nucleotide short interfering RNA (siRNA). The siRNA is unwound into single-stranded RNA and the sense RNA (green) is further cleaved and degraded by the enzyme Argonaute (AGO). The antisense RNA (red) is recruited into the RNA-induced silencing complex (RISC) that binds to the target sense RNA through the specificity of the complementary antisense RNA.

Applications: Used to silence specific genes in loss-of-function studies in order to observe phenotypic effects, support the development or validation of therapeutic targets, and serve as tools in agricultural biotechnology.

X.2.2. CRISPR/Cas9

CRISPR/Cas9 is a genome editing technology that uses a single guide RNA (sgRNA) to target a specific DNA sequence and induce a double-strand break via the Cas9 nuclease (Figure X.2). The break is repaired either by non-homologous end joining (NHEJ), which is often error-prone and leads to mutations, or by homology-directed repair (HDR), which enables precise gene modification or insertion [**64**].

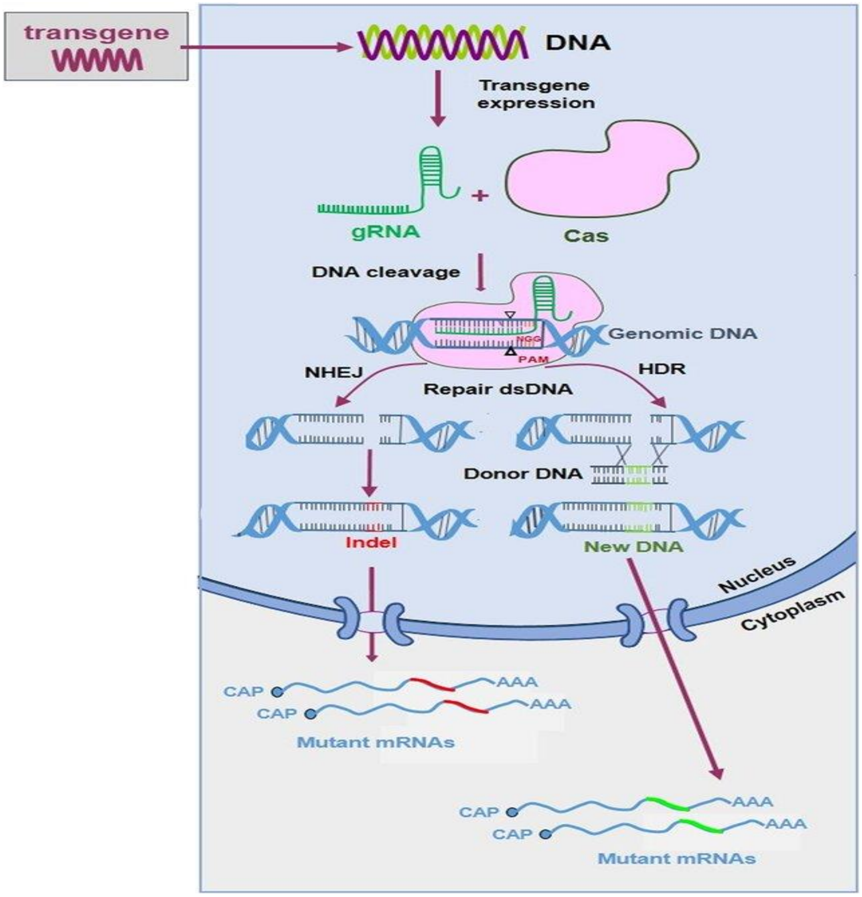


Figure X.2: Schematic of the CRISPR-Cas9-mediated genome editing process *[64].* CRISPR-Cas9 requires expression of a gRNA (green line) and the Cas9 endonuclease (pink shape). The gRNA instructs Cas9 for cleavage of a complementary DNA target with an adjacent PAM sequence. Cas generates a dsDNA break that is repaired by the NHEJ or HDR pathways.

- **Applications**: Used for loss-of function studies or to introduce specific modifications in genes of interest.

X.2.3. Overexpression systems

These systems involve inserting genes of interest into a strong expression vector (viral or plasmid) that are then introduced into cells, leading to increased expression of the gene (overproduction of the target protein).

- **Applications**: Allows researchers to study the effects of gene activation on cellular pathways or specific phenotypes.

X.2.4. Inducible systems

In these systems, we use controllable systems that allow genes to be turned on or off in response to external signals, such as drugs (e.g., tetracycline-inducible systems).

- **Applications**: Useful for studying gene function in a time-dependent or dose-dependent manner, providing deeper insights into their roles. These systems support temporal analysis, "all-or-none" gene activation models, and fine-tuning of expression in both research and therapeutic contexts.

X.2.5. Protein inhibitors

The system used small chemical molecules to block the activity of specific proteins, which in turn influences the expression or function of other related genes. The figure below (Figure X.3) illustrates the main modes of action of small protein inhibitor molecules, notably as modulators of intracellular signals [**65**].

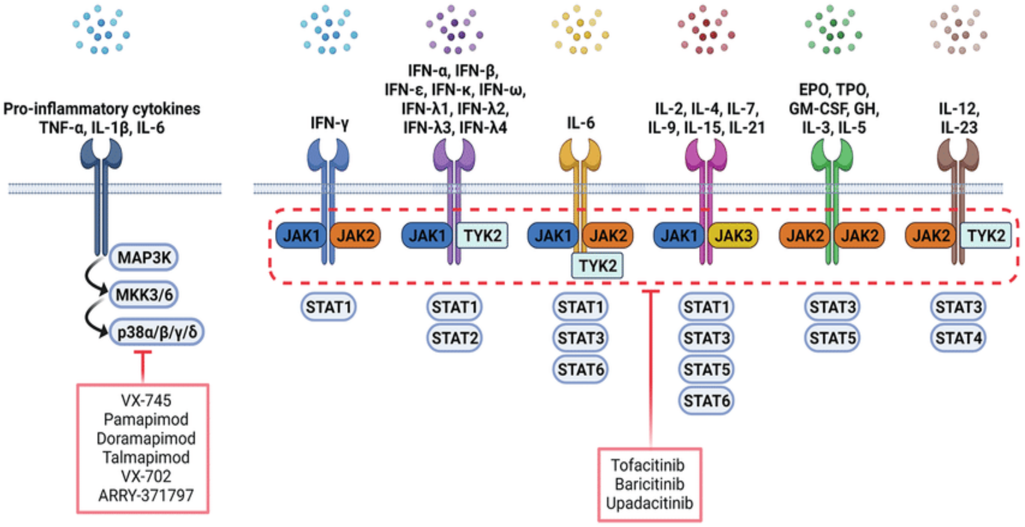


Figure X.3: Overview of small molecule inhibitors targeting the p38 MAPK pathway and JAK/STAT pathway in RA treatment *[65]*.

Each cytokine receptor recruits and activates a specific combination in MAPK and JAK/STAT cascades. Tofacitinib is a pan-JAK inhibitor, selective for JAK3 and JAK1 with minor activity for JAK2 and TYK2. Baricitinib is selective for JAK1 and JAK2 and less selective for JAK3 and TYK2.

- **Applications**: Allows for the study of signaling pathways and protein-protein interactions.

X.3. Complementary methods

X.3.1. Transcriptomic analysis

Transcriptomic analysis measures mRNA levels across the whole transcriptome, especially after gene expression has been altered (disrupted). Using RNA sequencing (RNA-seq), it helps identify which genes are up- or down-regulated, offering valuable insight into how gene expression changes under various conditions (e.g. following RNAi or CRISPR-based disruption) [**66**].

Applications: Used to determine the global effects of disrupting a gene on the cell's transcriptome.

X.3.2. Proteomic analysis

Proteomic analysis allows evaluates protein levels and post-translational modifications following disruption, providing a comprehensive view of changes at the protein level. This approach helps reveal the functional consequences of gene expression disruption.

X.3.3. Phenotypic studies

Phenotypic studies involve observing the phenotypic effects (the physical or developmental changes) resulting from gene expression disruption, enabling the study of specific traits.

In a study conducted on the entomopathogenic fungus *Beauveria bassiana*, the deletion of Bbsmr1 using CRISPR-Cas9 resulted in mutants overproducing the red pigment oosporein, displaying accelerated colony growth and enhanced virulence, clearly demonstrating the phenotypic consequences of gene disruption [**67**].

- **Applications**: Useful in disease research and developmental biology.

Systematic approaches to expression disruption are crucial for understanding the role of genes in cellular biology and physiological processes. By combining different analytical methods, researchers can elucidate complex interaction networks and underlying mechanisms of biological phenomena. These techniques pave the way for advancements in biomedical research and personalized medicine.

XI. RNA interference (RNAi) and systematic generation of mutants

XI.1. Generalities

RNA interference (RNAi) and systematic generation of mutants is a powerful molecular biology technique used to silence specific genes, providing insights into gene function and regulation. The systematic generation of mutants using RNAi allows to create a library of loss-of-function phenotypes, facilitating comprehensive studies of gene roles in various biological processes.

XI.2. RNA interference (RNAi) definition

RNAi is a biological process in which small RNA molecules inhibit gene expression or translation, effectively silencing specific genes. This process can be harnessed as a powerful tool for functional genomics [**63**].

XI.3. Mechanism of RNAi

RNAi reacts by adopting a mechanism of action (Figure X.1) as follows [**63**]:

a-Dicer enzyme: The RNAi process begins with the cleavage of long double-stranded RNA (dsRNA) into small interfering RNAs (siRNAs) by the enzyme Dicer.

b-siRNA incorporation: The siRNAs are then incorporated into a multi-protein complex known as the RNA-induced silencing complex (RISC).

c-Target mRNA degradation: Within the RISC, the siRNA guides the complex to complementary mRNA sequences, leading to the degradation of the target mRNA and preventing its translation into protein.

XI.4. Systematic generation of mutants using RNAi

The systematic generation of mutants through RNAi involves creating libraries of siRNAs targeting a wide array of genes within an organism, allowing for high-throughput screening of gene function [**68**].

XI.5. Steps in the systematic generation of mutants

The different steps in the systematic generation of mutants are summarized as follows:

a- **Library construction**: A library of siRNAs is designed to target a specific set of genes, often using bioinformatics tools to predict optimal siRNA sequences. Then, the siRNAs can be synthesized chemically or cloned into expression vectors for transfection into target cells.

b- **Transfection**: The siRNA library is introduced into the target cells (e.g., cultured cells, model organisms) via methods such as lipid-mediated transfection, electroporation, or viral delivery systems.

c- **Screening for phenotypes**: Following transfection, cells are cultured, and the effects of gene silencing are observed. Researchers assess the resulting phenotypes to identify genes whose disruption leads to specific outcomes, such as changes in growth, differentiation, or response to stress.

d-**Validation of hits**: Candidate genes identified through screening are further validated by reintroducing siRNAs targeting those genes or using alternative methods like CRISPR/Cas9 for confirmation of the phenotype.

e-Functional characterization: Once validated, the roles of the disrupted genes can be characterized through additional experiments, including transcriptomic and proteomic analyses to understand downstream effects [**68**].

XI.6. Applications of RNAi in systematic mutant generation

a-Gene function discovery: RNAi enables the identification of gene functions in various biological contexts, from basic cellular processes to complex developmental pathways.

b-Disease research: Systematic RNAi libraries can be used to identify potential therapeutic targets by elucidating the roles of genes in disease models, such as cancer or neurodegenerative diseases.

c-Synthetic biology: RNAi can be employed to create custom genetic circuits or regulatory systems in synthetic biology applications [**69,70**].

XI.7. Advantages and disadvantages

X1.7.1. Advantages

-High-throughput capability: RNAi allows for the simultaneous targeting of multiple genes, facilitating large-scale functional genomic studies.

-**Reversible effects**: The effects of RNAi are temporary, allowing researchers to study gene functions without permanent changes to the genome.

- **Cost-effectiveness**: Creating siRNA libraries is generally less expensive compared to generating stable knockout mutants.

XI.7.2. Disadvantages

-Off-target effects: siRNAs may inadvertently silence non-target genes, leading to ambiguous results.

-Incomplete knockdown: RNAi may not completely abolish gene expression, which can complicate the interpretation of phenotypic effects.

-Cell type-specific responses: The efficiency and effectiveness of RNAi can vary significantly between different cell types and organisms.

XII. Expression of recombinant proteins

XII.1. Expression of recombinant proteins generalities

Recombinant protein expression is a critical technique in molecular biology and biotechnology that involves producing proteins by inserting a gene encoding that protein into a host organism. This method is widely used for various applications, including research, drug development, and vaccine production. Here's an overview of the steps, expression systems, and applications associated with recombinant protein expression.

XII.2. Recombinant protein concept

A recombinant protein is a protein produced from genetically engineered cells that contain a cloned gene inserted into their DNA. This gene is usually derived from a different species and expressed in a host to generate the protein.

XII.3. Steps in recombinant protein expression

The different steps in recombinant protein (Figure XII.1) are summarized as follows:

a- **Cloning the gene of interest**: The gene encoding the target protein is isolated and cloned into an appropriate expression vector. This vector contains essential elements for expression, such as a promoter, a selection marker, and termination sequences.

b- Transformation of the host: The recombinant vector is introduced into a host organism, such as bacteria (e.g., *E. coli*), yeast, mammalian cells, or insect cells, through methods like transformation, transfection, or viral infection.

c- Cell culture: Transformed cells are cultured under controlled conditions, allowing for the growth and proliferation of cells expressing the recombinant protein.

d- **Induction of expression:** In some systems, protein expression is induced by adding specific inducers (e.g., IPTG for *E. coli* systems).

e- **Extraction and purification**: Once expressed, the recombinant protein is extracted from the cells and purified using techniques such as affinity chromatography, precipitation, or electrophoresis [**71**].

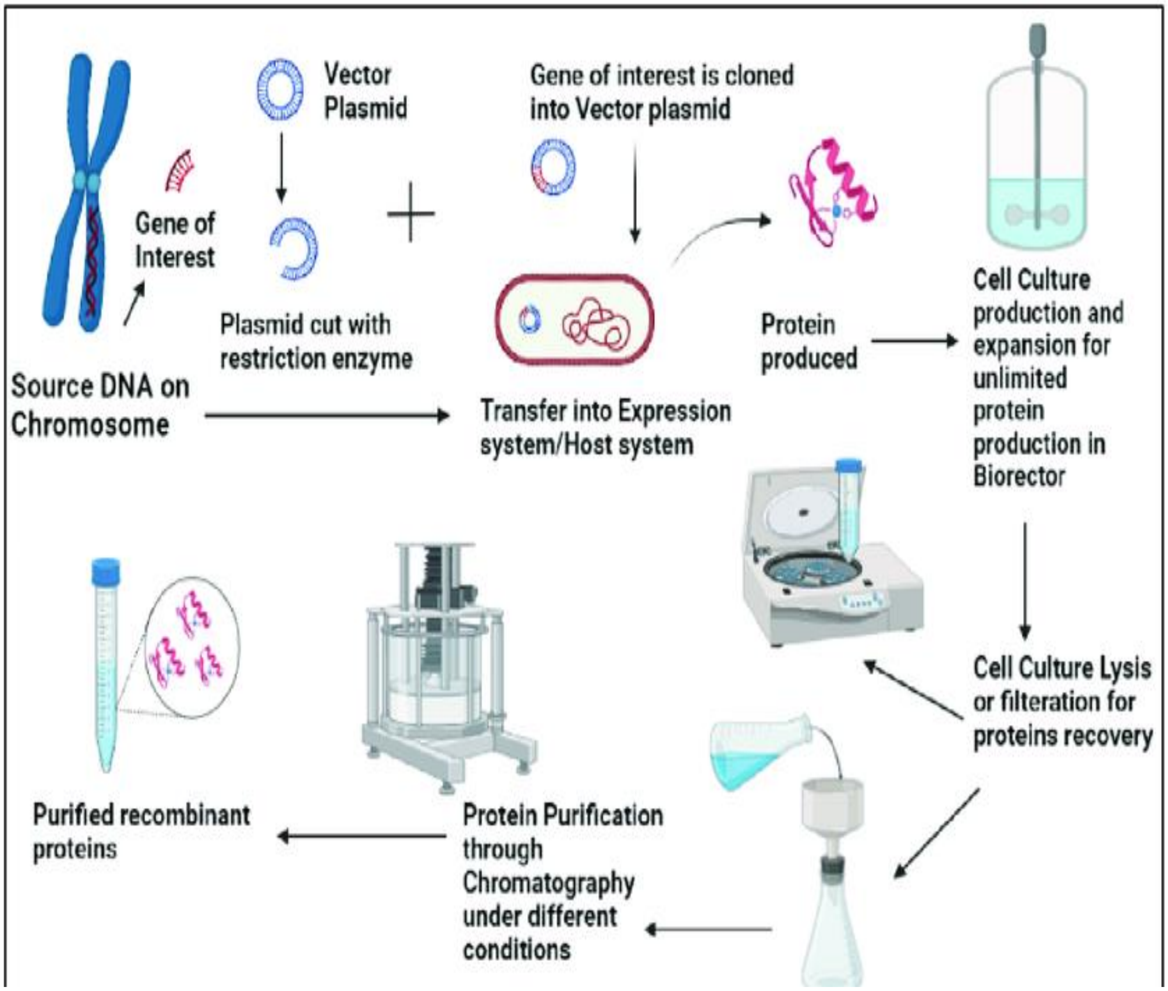

Vector Plasmid
Gene of interest is cloned into Vector plasmid
Gene of Interest
Plasmid cut with restriction enzyme
Source DNA on Chromosome
Transfer into Expression system/Host system
Protein produced
Cell Culture production and expansion for unlimited protein production in Biorector
Cell Culture Lysis or filteration for proteins recovery
Purified recombinant proteins
Protein Purification through Chromatography under different conditions

Figure XII.1: General methodology for recombinant protein production in expression system and afterward purification *[71].*

XII.4. Expression systems of recombinant protein

Different expression systems can be utilized to produce recombinant proteins [**72**], each with its own advantages and disadvantages (Table XII.1).

Table XII.I: Different expression systems of recombinant proteins *[72]*.

Expression systems	**Advantages**	**Disadvantages**
Bacteria (E. coli)	Rapid growth, low cost, and high yield.	Lack of post-translational modifications (such as glycosylation) that some proteins require for functionality.
Yeast (Saccharomyces cerevisiae)	Some post-translational modification capabilities, rapid growth.	Limited ability to process complex proteins.
Mammalian cells	Capable of performing complex post- translational modifications, producing functional proteins.	Higher cost and longer culture times compared to prokaryotic systems.
Insect cells (Baculovirus system)	Ability to express complex and glycosylated proteins.	More complex and costly processes than prokaryotic systems.

XII.5. Applications of recombinant proteins

-**Fundamental research**: Study of protein functions and biomolecular interactions.

-**Drug production**: Production of therapeutic proteins, such as monoclonal antibodies, hormones (like insulin), and growth factors.

-**Vaccines**: Development of subunit vaccines based on recombinant proteins for infectious diseases.

-**Industrial enzymes**: Use in agriculture, food biotechnology, and bioremediation processes.

XII.6. Advantages and disadvantages

XII.6.1. Advantages

-**High throughput**: Capable of producing large amounts of protein quickly.

-**Flexibility**: Ability to produce various proteins from different species.

-**Temporary effects**: The expression of recombinant proteins can be controlled and is reversible, allowing researchers to study gene functions without permanent changes to the genome.

XII.6.2. Disadvantages

- **Off-target effects**: Recombinant proteins may interact with unintended targets, complicating results.

- **Incomplete folding**: Proteins may misfold in bacterial systems, affecting their functionality.

- **Variability by cell type**: Expression efficiency can vary significantly between different cell types and organisms.

XIII. Methods and applications of protein engineering

XIII.1. Generalities

Protein engineering is a multidisciplinary field that involves designing, modifying, and optimizing proteins to enhance their functions or introduce new properties. This discipline leverages techniques from molecular biology, biochemistry, and structural biology to achieve specific goals.

XIII.2. Methods of protein engineering

The main protein engineering strategies [**73**]: directed evolution, rational design, and semi-rational or computational design are illustrated as a flowchart in the figure XIII.1, as follows:

XIII.2.1. Directed evolution

A technique that mimics natural selection to evolve proteins or enzymes with desirable traits. Libraries of protein variants are created through mutagenesis, and those with improved functions are selected through screening.

- **Application**: Commonly used to enhance enzyme activity, stability, or specificity for industrial applications.

XIII.2.2. Site-directed mutagenesis

This method allows specific amino acid substitutions in a protein sequence through techniques like PCR (Polymerase Chain Reaction) or the use of synthetic oligonucleotides.

- **Application**: Used to study the effects of specific mutations on protein function or to create proteins with altered properties.

X.III.2.3. Fusion protein technology

Involves the fusion of two or more protein domains or genes to create a chimeric protein. This can enhance stability, activity, or provide new functionalities.

- **Application**: Used in the development of novel enzymes or therapeutic proteins that combine multiple activities.

XIII.2.4. Structure-based design

Utilizes knowledge of a protein's three-dimensional structure to design modifications that improve its function or stability. Techniques include X-ray crystallography and molecular modeling.

-**Application**: Optimizing drug-binding sites or engineering proteins with improved catalytic efficiency.

XIII.2.5. Protein domain shuffling

Involves rearranging and recombining different protein domains to create new proteins with desired functionalities.

- **Application**: Useful for generating proteins with novel activities or improved properties by combining beneficial features from different proteins.

XIII.2.6. High-throughput screening

Utilizes automated techniques to rapidly assess large libraries of protein variants for specific properties or activities.

- **Application:** Efficiently identifies promising candidates for industrial enzymes, therapeutic proteins, or antibodies.

XIII.2.7. CRISPR-Cas9 and genome editing

Employs the CRISPR-Cas9 system for targeted modification of genes in organisms, allowing precise editing of protein-coding regions [**73**].

- **Application**: Used to create model organisms with specific traits or to engineer cell lines for the production of modified proteins.

XIII.3. Applications of protein engineering

XIII.3.1. Biotechnology and industry

-**Industrial enzymes**: Engineering enzymes for applications in food production, biofuels, and waste management, often focusing on enhancing stability and efficiency under industrial conditions.

- **Bioremediation**: Development of proteins that can degrade environmental pollutants, facilitating cleanup efforts [**73**].

XIII.3.2. Medicine and pharmaceuticals

- **Therapeutic proteins**: Production of engineered proteins, such as monoclonal antibodies and hormones (e.g., insulin), for treatment of various diseases.

- **Vaccine development**: Use of recombinant proteins to create subunit vaccines against infectious diseases [**74**].

XIII.3.3. Research and development

- **Studying protein functions**: Engineering proteins to investigate specific interactions and functions within biological systems, contributing to our understanding of cellular mechanisms.

- **Synthetic biology**: Designing proteins for constructing biological circuits and systems that can perform specific tasks, such as biosensing or metabolic engineering.

XIII.3.4. Immunotherapy

-**Targeted cancer therapy**: Engineering antibodies or other proteins to specifically target cancer cells, enhancing the efficacy of cancer treatments.

- **Chimeric antigen receptor (CAR) T-cell therapy**: Creating engineered T-cells that express CARs to recognize and attack cancer cells.

XIII.3.5. Agriculture

- **Genetically modified organisms (GMOs)**: Engineering proteins in plants to confer resistance to pests or diseases, improve nutritional content, or enhance yield.

Biopesticides: Development of proteins that can act as environmentally friendly alternatives to chemical pesticides [**73**].

XIII.3.6. Diagnostics

- **Biosensors**: Engineering proteins to serve as biosensors for detecting specific biomolecules, pathogens, or environmental toxins in medical and environmental applications.

- **Diagnostic tools**: Creating proteins for use in assays to detect diseases, such as using engineered antibodies for specific antigen detection [**74**].

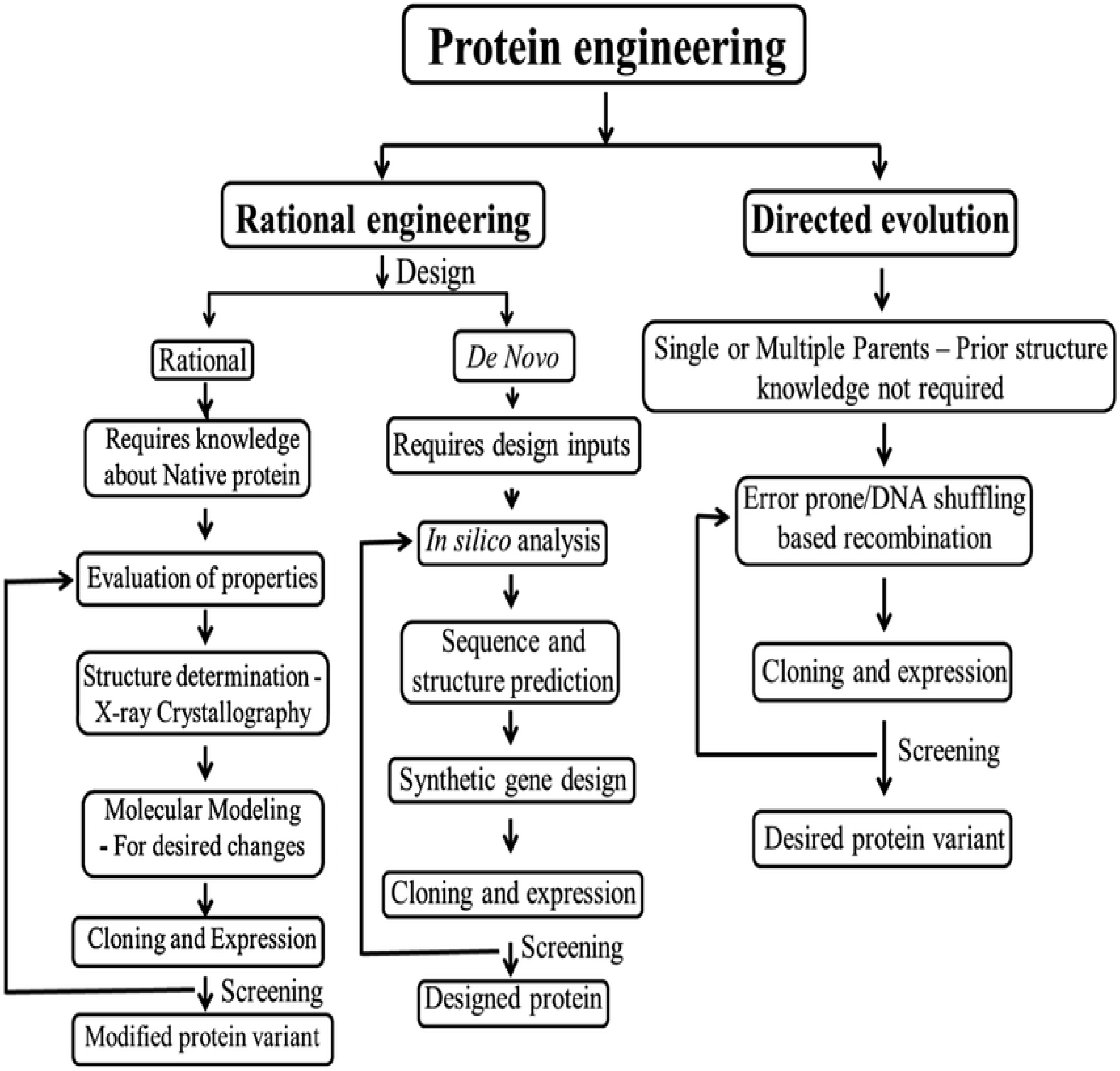


Figure XIII.1: Structural outline of protein engineering strategies *[75].*

Chapter XIV. Rational protein engineering

XIV.1. Rational protein engineering generalities

Rational protein engineering is a systematic approach used in molecular biology and biotechnology to design and optimize proteins or biomolecules based on scientific principles and experimental data. Unlike random approaches or directed evolution, rational engineering relies on structural models, biological mechanisms, and sequence data to guide modifications to proteins [**73**].

XIV.2. Concepts of rational protein engineering

The three main concepts of rational protein engineering are as follows:

a-**Structural modeling*:*** Analyzing the three-dimensional structure of proteins using techniques such as X-ray crystallography, NMR (nuclear magnetic resonance), and cryo-electron microscopy helps to understand the relationship between protein structure and function.

Structural models aid in identifying active sites, binding sites, and regions amenable to modification.

b-**Bioinformatics*:*** The use of software and algorithms to analyze genetic and proteomic data enables the prediction of the effects of mutations on protein structure and function.

Phylogenetic tools and sequence alignment help identify conserved motifs and variations.

c-**Molecule design:** Computer-aided design (CAD) approaches allow the simulation of sequence or structural modifications, facilitating the design of new proteins with desired properties.

Protein design can also involve optimizing physicochemical properties to enhance solubility, stability, and activity [**73**].

XIV.3. Methods of rational protein engineering

a-**Site-directed mutagenesis:** Targeted modification of amino acids to study their impact on protein function, typically performed by PCR or gene synthesis.

b-**Mutation analysis**: Systematic evaluation of the effects of mutations on enzymatic activity, stability, and protein-protein interactions [**76**].

c- **Condition optimization***:* Use of statistical methods to determine the best experimental conditions for protein expression and purification [**73**].

d-**Enzyme design:** Creation of enzymes with enhanced catalytic properties by modifying amino acid residues in the active site [**76**].

XIV.4. Applications of rational protein engineering

a- ***Medicine***

-Gene therapy: Designing therapeutic proteins or growth factors to treat genetic diseases or cancers.

- Drug development: Creating new drug targets by modifying proteins to increase their specificity and efficacy [**70**].

b-***Biotechnology***

-Enzyme production: Engineering enzymes for industrial applications, such as biomass degradation or biofuel production.

- Diagnostic enzymes: Development of sensitive enzymes for medical diagnostic tests [**73**].

e- ***Vaccines*****:** Using rational engineering to design recombinant proteins as vaccine candidates against specific pathogens.

c- *Agriculture*

- Biopesticides: Developing proteins with insecticidal properties to reduce the use of chemical pesticides.

-Crop improvement: Engineering proteins to enhance plant resistance to diseases or environmental stresses.

d-Nanotechnology: Using proteins to create nanostructures or drug delivery systems, leveraging their specificity and ability to interact with other biomolecules [**73**].

XV. Directed evolution of proteins and nucleic acids

XV.1. *Generalities*

Directed evolution is a powerful method used in molecular biology to generate variants of proteins or nucleic acids that possess improved properties or new functions. This approach simulates the natural selection process by creating genetic diversity within a population of biomolecules and selecting those that exhibit desired characteristics. Here's a summary of the concepts, methods, and applications of directed evolution [**77**].

XV.2. Concepts of directed evolution

XV.2.1. Natural selection

The fundamental principle of directed evolution is based on natural selection, where the best-adapted variants survive and proliferate. In the laboratory context, this means identifying and isolating variants of biomolecules that function better or exhibit new activities.

XV.2.2. Genetic diversity

Creating a broad genetic diversity allows for the exploration of a vast functional space. This can be achieved through random or targeted mutagenesis, resulting in modifications to the amino acid sequences of proteins or the sequences of DNA [**77**].

XV.2.3. Selection and screening

The generated variants undergo a selection process where those that display the desired properties are isolated and enriched within the population [**77**].

XV.3. Methods of directed evolution

Directed evolution relies on key methodologies such as random mutagenesis, mutant library construction, high-throughput screening, selection systems, and phage display. Together, these approaches make it possible to efficiently generate and identify biomolecule variants with enhanced or new functionalities [**77**].

XV.3.1. Random mutagenesis

Introducing random mutations into the gene of interest using methods such as PCR (polymerase chain reaction), recombination, or chemical mutagens. This method generates a variety of variants that can be tested for their functions (Figure XV.1).

XV.3.2. Mutant libraries

Creation of libraries of mutants (Figure XV.1), where each variant is a modified version of the original protein or nucleic acid. These libraries can contain millions of variants, increasing the chances of finding candidates with the desired characteristics [**77**].

XV.3.3. High-throughput screening

Utilizing high-throughput screening methods (Figure XV.1) to rapidly evaluate thousands or millions of variants for specific properties. This can include enzymatic assays, affinity tests, or fluorescence assays [**77**].

XV.3.4. Selection systems

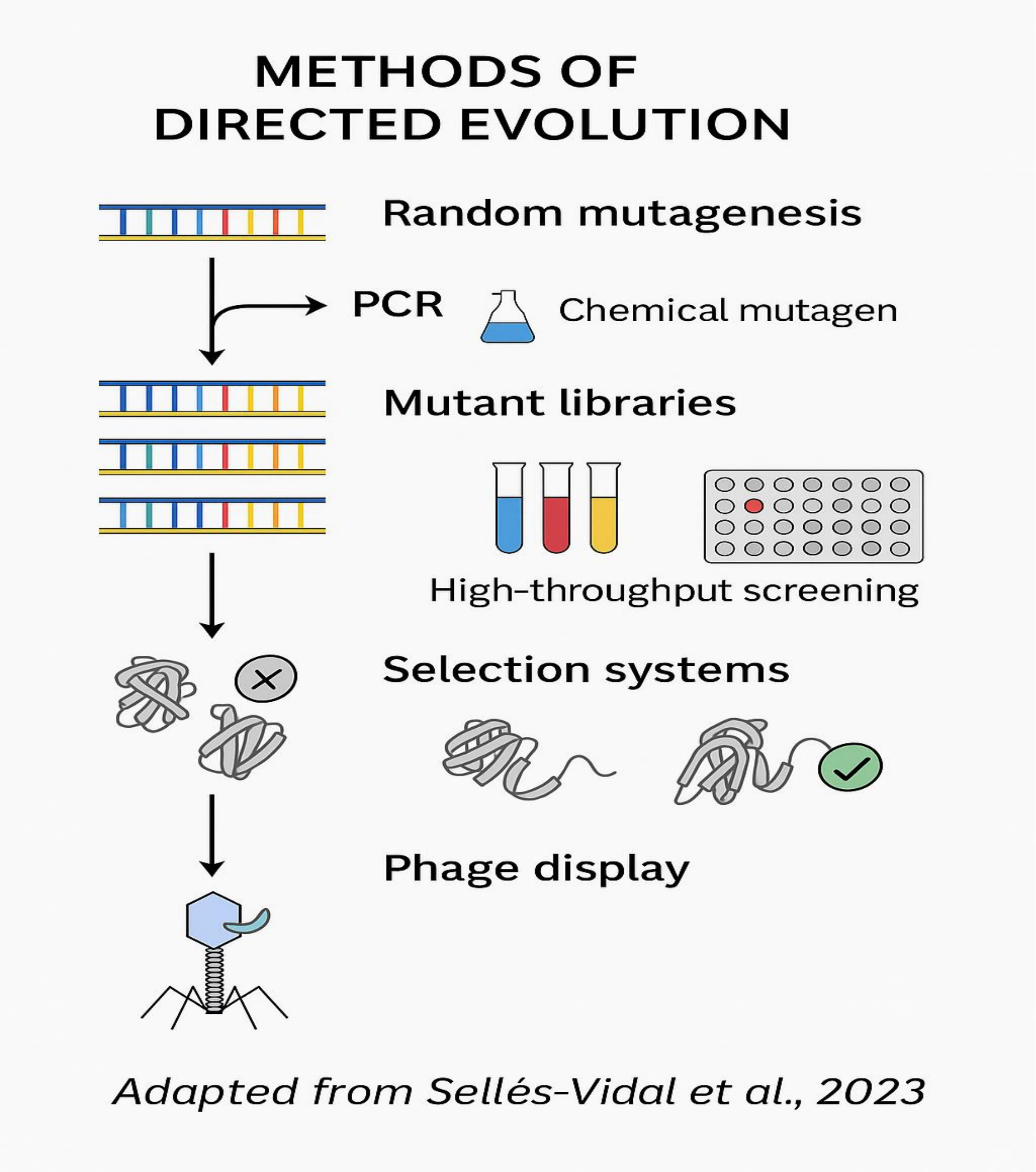
Implementing selection systems (Figure XV.1) based on specific criteria, such as affinity for a ligand, enzymatic activity, or the ability to bind to a target. These systems allow for the efficient selection of the most promising variants.

XV.3. 5.Phage display

A technique in which phages (Figure XV.1) are used to display peptides or proteins. Phages displaying strong interactions with a target of interest can be isolated and amplified [**77**].

XV.4. Applications of directed evolution

A summary table (Table xv.1) of the main applications of directed evolution is presented below.



Adapted from Sellés-Vidal et al., 2023

Figure XV.1: Overview of Key Methods Used in Directed Evolution for Protein Engineering *[77].*

Tableau XV.1: Applications of Directed Evolution.

Field	Specific Application	
Biotechnology	- Industrial enzymes (enhanced stability, activity, thermal resistance)	**[78].**
	- Green biocatalysis	**[79].**
Medicine	- Monoclonal antibody optimization	**[80].**
	- Therapeutic proteins	**[81].**
	- Gene therapy	**[80].**
Fundamental Research	- Protein function studies	**[82].**
	- Gene network analysis	**[83].**
Agriculture	- Transgenic plants with increased resistance	**[84].**
	- Improved nutrient uptake	**[85].**
Nanotechnology	- Protein design for drug targeting	**[86].**
	- Assembly of protein-based nanostructures	**[87].**

This course handout explored the many facets of genomics and molecular biology, covering methods and technologies that have transformed our understanding of biological systems. From genome analysis to protein engineering, each chapter provided theoretical and practical tools for studying and manipulating genes, proteins, and their interactions in various biological contexts.

Genome mapping, construction of genomic libraries, and high-throughput sequencing are the first key steps in identifying genes and functional elements. These foundations are essential for understanding the structure and organization of genomes, which pave the way for discoveries about the role of genes in health and disease. Linkage studies and genetic mapping have led to major advances in identifying genes responsible for complex traits, whether human diseases or agricultural traits.

Then, the integration of gene expression data via global transcriptome studies, proteomics, and molecular interactions (double hybrid, Tap-tag) offers a dynamic vision of cellular activity. These approaches not only allow us to identify genes and proteins in action, but also to better understand how they interact in complex functional networks. Highlighting protein-protein interaction networks or transcriptional regulation is crucial for deciphering biological processes and abnormalities associated with diseases.

Expression disruption techniques such as RNA interference (RNAi) and systematic mutant generation have proven important for exploring gene functions. These tools provide direct approaches to disrupt gene expression and identify specific contributions of genes to biological processes.

Finally, the last sections focused on protein engineering, with methods such as rational engineering and directed evolution. These techniques allow proteins to be manipulated to create variants with improved or new properties, with applications ranging from basic research to biotechnology and medicine. Protein engineering is a rapidly expanding field, with major implications for the development of new therapies, industrial enzymes, and biomedical innovations.

In short, this handout has provided an opportunity to address the most advanced approaches in the study and manipulation of genomes and proteins. These tools now allow not only to understand biological systems better systems, but also to develop innovative solutions in the fields of health, agriculture, and biotechnology. The future of molecular biology and genomics lies in the integration of these technologies on a large scale, paving the way for discoveries that will transform our lives in the years to come.

This conclusion highlights the coherence between the different themes covered in the course, while emphasizing their importance for progress in molecular biology and their practical applications.

I would like to express my heartfelt gratitude to everyone who contributed to the completion of this course material.

First and foremost, I am deeply grateful to Professor OUCHEMOUKH Salim, Dr. BOURNINE Lamine and Dr. YALAOUI-GUELLAL Drifa for their guidance, insights, and encouragement throughout this journey. Their expertise and passion for the subject matter have been a true source of inspiration.

I would also like to thank my colleagues and fellow students, whose valuable feedback and discussions have greatly enriched the content of this document. The collaborative spirit and shared enthusiasm have made this experience more rewarding.

Special thanks to my family and friends for their unwavering support and patience during the many hours spent working on this project.

Lastly, I acknowledge the support of Akli Mohand Oulhadj - Bouira University

, whose resources and environment have been instrumental in making this work possible.

Thank you all !

[1] ROMANAS, BONNEFONT, J.P, CAVAZZANA, M, CAVAZZANA-CALVO, M, MALAN, V, JAÏS, J.P. Méthodes d'étude et d'analyse du génome (Cours + QCM). United Kingdom: Elsevier Health Sciences France.2012.

[2] VENTERJ. C., et al. "The sequence of the human genome." Science, 291(5507). 2001, 1304-1351.

[3] MARXV. (2021). Method of the Year: Spatially resolved transcriptomics. Nature Methods, 18, 9–14. https://doi.org/10.1038/s41592-020-01033-y

[4] HAOY, HAO, S, ANDERSEN-NISSEN, E, MAUCK, W. M., Zheng, S., Butler, A., ... & Satija, R. (2021). Integrated analysis of multimodal single-cell data. Cell, 184(13), 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

[5] ANNUNZIATOA. (2008) DNA Packaging: Nucleosomes and Chromatin. Nature Education 1(1):26

[6] SANGERF, NICKLEN, S, COULSON, A. R. "DNA sequencing with chain-terminating inhibitors." Proceedings of the National Academy of Sciences, 74(12). 1977, 5463-5467.

[7] MARDISE. R. "Next-generation DNA sequencing methods." Annual Review of Genomics and Human Genetics, 9.2008, 387-402.

[8] SCHATZM. C., LANGMEAD, B., & SALZBERG, S. L. "Cloud computing and the DNA data race." Nature Biotechnology, 28(7). 2010, 691-693.

[9] BANKEVICHA, ET A.. "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of Computational Biology, 19(5). 2012, 455-477.

[10] ZERBINOD. R., & BIRNEY, E. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Research, 18(5), 2008, 821-829.

[11] TATUSOVAT, ET A.. "RefSeq microbial genomes database: new representation and annotation strategy." Nucleic Acids Research, 41(D1). 2013, D595-D603.

[12] JOHNSONM.T.J, CARPENTER, E.J, TIAN, Z, BRUSKIEWICH, R, BURRIS, J.N, ET A.. "Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes." PLoS ONE, 7(11): e50226, 2012.

[13] THE 1.P.C.. (2015). "A global reference for human genetic variation." Nature, 526(7571), 68-74.

[14] VISSCHERP. M., et al. "10 years of GWAS discovery: biology, function, and translation." The American Journal of Human Genetics, 101(1). 2017, 5-22.

[15] HARDISONR. C. "Comparative genomics." PLoS Biology, 1(2), e58, 2003.

[16] ALTSCHULS. F., et al. (1990). "Basic local alignment search tool." Journal of Molecular Biology, 215(3), 403-410.

[17] COLLINSF. S., & VARMUS, H. "A new initiative on precision medicine." The New England Journal of Medicine, 372(9). 2015, 793-795.

[18] VARSHNEYR. K., et al. "Genomics-assisted breeding for crop improvement." Trends in Plant Science, 19(7). 2014, 370-380.

[19] SAMBROOKJ, RUSSELL, D. W. "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory Press. 2001.

[20] GREENM. R., & SAMBROOK, J. "Molecular Cloning: A Laboratory Manual." Cold Spring Harbor Laboratory Press. 2012.

[21] ROBERTSR. J. "How restriction enzymes became the workhorses of molecular biology." Proceedings of the National Academy of Sciences, 102(17). 2005, 5905-5908.

[22] KIMU. J., et al. "Construction and characterization of a human bacterial artificial chromosome library." Genomics, 34(2). 1996, 213-218.

[23] DOWERW. J., MILLER, J. F., & RAGSDALE, C. W. "High efficiency transformation of E. coli by high voltage electroporation." Nucleic Acids Research, 16(13). 1986, 127-6145.

[24] CEPHAMLIFE S.. Genomic libraries: construction and applications [Internet]. Rockville (MD): Cepham Biosciences; 2022 [cité le 22 juill. 2025]. Disponible sur:

[25] SAIKIR. K., et al. "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase." Science, 239(4839). 1988, 487-491.

[26] SOUTHERNE. M. "Detection of specific sequences among DNA fragments separated by gel electrophoresis." Journal of Molecular Biology, 98(3). 1975, 503-517.

[27] METZKERM. L. "Sequencing technologies - the next generation." Nature Reviews Genetics, 11(1). 2010, 31-46.

[28] ALTSCHULS. F., et al. "Basic local alignment search tool." Journal of Molecular Biology, 215(3). 1990, 403-410.

[29] COLLINSF. S., et al. "A DNA polymorphism discovery resource for research on human genetic variation." Genome Research, 13(4). 2003, 1011-1020.

[30] NEXTGENERATION S.-.S.E.. YouTube. [cité 2025 juill. 24]. Disponible sur :

[31] GAUDRIAULTS, VINCENT, R. Génomique. Editions D Boek Université, Bruxelles, Belgique.2009.

[32]

[33] NEXTGENERATION S.(.. SlidePlayer. [cité 2025 juill. 24]. Disponible sur :

[34] JUAREZP. (2017). Regulatory mechanisms of mexEF-oprN efflux operon in Pseudomonas aeruginosa: From mutations in clinical isolates to its induction as response to electrophilic stress (Thèse de doctorat). ResearchGate. ...

[35] SCHEMATIC R.O.T.. ResearchGate. [cité 2025 juill. 24]. Disponible sur : https://www.researchgate.net/figure/Schematic-representation-of-Pyrosequencing-Technique-During-the-incorporation-of_fig1_266264269

[36] SHALEMO, SANJANA, N.E, ZHANG, F. High-throughput functional genomics using CRISPR–Cas9. Nature Reviews Genetics. 2015.

[37] JARES-ERIJMANE.A, JOVIN, T.M. FRET imaging. Nature Biotechnology. 2003.

[38] WANGZ, GERSTEIN, M, SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1). 2009, 57-63.

[39] WEINSTEINJ. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R., OZENBERGER, B. A., ELLROTT, K., ... & al. The Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. Nature Genetics, 45(10), 1113-1120.

[40] ORLANDOV. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends in Biochemical Sciences, 25(3). 2000, 99-104.

[41] FUREYT. S. ChIP–seq and beyond: New and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics, 13(12). 2012, 840-852.

[42] GIFFORDC. A., & MEISSNER, A. Epigenetic regulation in pluripotent stem cells: A gateway to cell fate. Nature Reviews Genetics, 14(7). 2013, 431-444.

[43] RENB, ROBERT, F, WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ... & YOUNG, R. A. Genome-wide location and function of DNA binding proteins. Science, 290(5500). 2000, 2306-2309.

[44] COLLASP, DAHL, J. A. (2008). Chip-based methods for transcription factor mapping. Methods, 44(1), 3–9.

[45] [ANONYMOUS 5.

[46] PÉREZ-PINERAP, KOCAK, D. D., Vockley, C. M., Adler, A. F., Kabadi, A. M., Polstein, L. R., ... & Gersbach, C. A. (2013). RNA-guided gene activation by CRISPR–Cas9-based transcription factors. Nature Methods, 10(10), 973–976.

[47] DOMINGUEZA. A., Lim, W. A., & Qi, L. S. (2016). Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. Nature Reviews Molecular Cell Biology, 17(1), 5–15.

[48] NISHIMASUH, ET A.. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell, 156(5), 935–949.

[49] JINEKM, ET A.. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science, 337(6096), 816–821.

[50] CONGL, ET A.. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science, 339(6121), 819–823.

[51] MALIP, ET A.. (2013). RNA-guided human genome engineering via Cas9. Science, 339(6121), 823–826.

[52] DOUDNAJ. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR–Cas9. Science, 346(6213), 1258096.

[53] BARRANGOUR, DOUDNA, J. A. (2016). Applications of CRISPR technologies in research and beyond. Nature Biotechnology, 34(9), 933–941.

[54] GILBERTL. A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell, 154(2), 442–451.

[55] QIL. S., et al. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell, 152(5), 1173–1183.

[56] ESVELTK. M., et al. (2013). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. Nature Methods, 10(11), 1116–1121.

[57] SHALEMO, ET A.. (2014). Genome-scale CRISPR–Cas9 knockout screening in human cells. Science, 343(6166), 84–87.

[58] WANGT, ET A.. (2014). Genetic screens in human cells using the CRISPR–Cas9 system. Science, 343(6166), 80–84.

[59] HELLERR. C., & Marians, K. J. (2006). Replisome assembly and the direct restart of stalled replication forks. Nature Reviews Molecular Cell Biology, 7(12), 932–943.

[60] MG. Rosenfeld, C. J. Lunyak, and J. F. Glass, “Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response,” Genes & Development, vol. 20, no. 11, pp. 1405–1428, 2006.
https://doi.org/10.1101/gad.1424806

[61] RA. Young, “Control of the embryonic stem cell state,” Cell, vol. 144, no. 6, pp. 940–954, 2011. https://doi.org/10.1016/j.cell.2011.01.032

[62] KD. Makova and R. C. Hardison, “The effects of chromatin organization on variation in mutation rates in the genome,” Nature Reviews Genetics, vol. 16, no. 4, pp. 213–223, 2015. https://doi.org/10.1038/nrg3890

[63] KOPPÉEL, MA, L, SPANJAARD, B, ET A.. (2021). Simultaneous epitope and transcriptome measurement in single cells. Nature Methods, 18(3), 282–289.

[64] HENERAP, ET A.. (2021). Single-cell RNA-seq in cancer: Advances and clinical implications. Critical Reviews in Oncology/Hematology, 157, 103170.

[65] KIMD, ET A.. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915.

[66] TZEC-JIMÉNEZL, ET A.. (2022). Advances in transcriptomics for understanding plant stress responses. Frontiers in Plant Science, 13, 876543.

[67] MASCARNHASA. P., et al. (2020). Proteomic approaches in plant stress signaling. Proteomics, 20(15), e1900332.

[68] MOHRS, ET A.. (2013). Thermal proteome profiling: Unbiased assessment of protein–drug interactions. Nature Protocols, 8(4), 749–762.

[69] GAOY, ET A.. (2022). Application of proteomics in food safety and quality. Trends in Food Science & Technology, 126, 41–54.

[70] EVERST. M. J., et al. (2019). Quantitative proteomics reveals molecular networks in Alzheimer’s disease. Molecular Systems Biology, 15(6), e8831.

[71] SAHREENS, ET A.. (2021). Antioxidant and anti-inflammatory effects of Zizyphus jujuba extracts. Biomedicine & Pharmacotherapy, 137, 111351.

[72] RENAMS, ET A.. (2020). Nutraceutical and pharmacological importance of jujube. Journal of Functional Foods, 65, 103729.

[73] CHENA. F. A., et al. (2018). Development of plant-based dairy alternatives: A review. Advances in Food and Nutrition Research, 85, 47–96.

[74] EBRAHIMM. T., et al. (2021). Production of plant-based milk and dairy analogs. Journal of Food Processing and Preservation, 45(10), e15720.

[75] BRINDHAP, ET A.. (2022). Directed evolution: Protein engineering for improved biocatalysts. Evolutionary Bioinformatics, 18, 1–10.

[76] SONGC, ET A.. (2021). Machine learning-assisted directed evolution. Trends in Biotechnology, 39(12), 1262–1273.

[77] SELLEEE, ET A.. (2021). Applications of directed evolution in synthetic biology. ACS Synthetic Biology, 10(5), 1051–1062.

[78] ZEYNERF, ET A.. (2022). Protein engineering through rational design and directed evolution. BioTechniques, 73(4), 156–170.

[79] BORUSHM, ET A.. (2021). Enzyme engineering for industrial biocatalysis. Biotechnology Advances, 49, 107752.

[80] CHENJ. J. M. S., et al. (2022). Improving protein thermostability by directed evolution. Journal of Molecular Modeling, 28(3), 85.

[81] WONGT. S., et al. (2006). Combining directed evolution and rational design to improve enzyme properties. Biotechnology Advances, 24(3), 243–248.

[82] ARNOLDF. H. (2018). Directed evolution: Bringing new chemistry to life. Angewandte Chemie International Edition, 57(16), 4143–4148.

[83] CURRANK. A., et al. (2015). Enabling high-throughput screening of engineered microbial strains. Biotechnology Journal, 10(10), 1644–1652.

[84] MENZELLAH. G. (2011). Comparison of two codon optimization strategies to enhance recombinant protein production in Escherichia coli. Microbial Cell Factories, 10(1), 15.

[85] CHENY, ET A.. (2021). Machine learning-guided protein engineering. Nature Communications, 12, 4335.

[86] LIH, ET A.. (2022). Applications of deep learning in protein engineering. Bioinformatics, 38(6), 1455–1463.

[87] BRINDHAP, ET A.. (2019). Directed evolution strategies and applications in industrial biotechnology. RSC Advances, 9, 10496–10512.

Exercise 1:

Given the following bacterial DNA sequence:

5'- ATTTACGGGCCTTAATGGCATAACCGCCTAATGGTTAACCGCTAGCGCG - 3'

Q1- Give the sequence of the corresponding double-stranded DNA?

Q2- Under what condition would this double-stranded DNA be transcribed in vivo?

Q3- Give the sequence of the possible transcript?

Exercise 2:

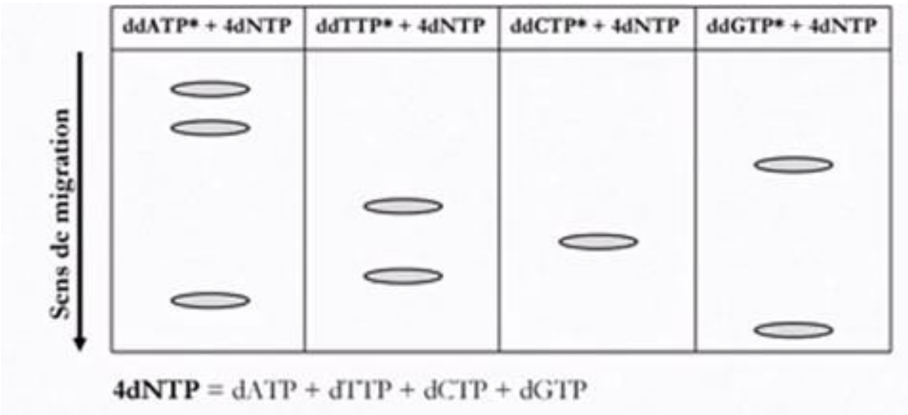
Nucleotide sequencing is a technique based on the following principle:

a single-stranded DNA fragment is incubated in the presence of DNA polymerase with a primer and a mixture of nucleotides in the presence of the radioactive dideoxy form of the same nucleotide and the three other non-radioactive nucleotides.

4 separate incubations are carried out (one nucleotide is marked each time):

a. ddATP* + (dATP, dTTP, dCTP, dGTP)
b. ddTTP* + (dATP, dTTP, dCTP, dGTP)
c. ddCTP* + (dATP, dTTP, dCTP, dGTP)
d. ddGTP* + (dATP, dTTP, dCTP, dGTP)

The mixtures from each incubation are separated by polyacrylamide gel electrophoresis.



1. Deduce the sequences of the complementary strand from this electrophoretic profile, then deduce the strand studied?
2. Specify the role of the dideoxy nucleotide form?

Exercise 3:

The EcoRI recognition site is described as follows G/AATTC:

-What is the meaning of the symbols: [/] and *? You have the following enzymes

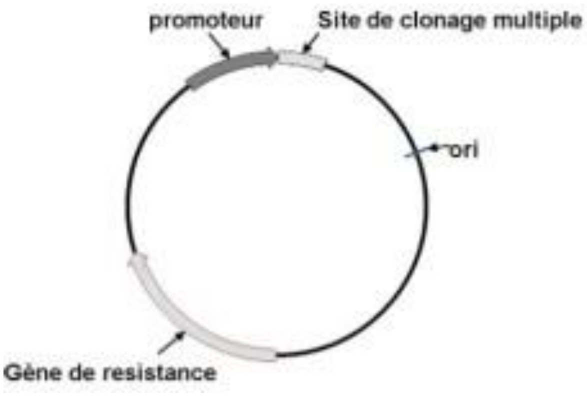
***Kpn*I** : 5'GGTAC/C3'   ***Acc65*I :** 5'G/GTACC3'

Bam*HI** (G / GATCC)Mbo*I** (/GATC).

-What is the relationship between the couple: [*Kpn*I, *Acc65*I] and the couple [*Bam*HI, *Mbo*I]?

Exercise 4:

- Define precisely the role of the four parts of the vector indicated by the arrows in the figure below?

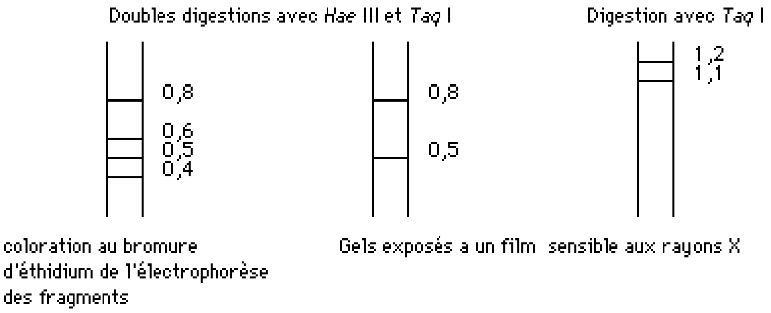


Exercise 5: Restriction maps of a cDNA and the corresponding genomic clone.

Genomic clones and cDNA of a phosphatase enzyme were isolated. From the following results, the structural features of the gene and its transcript can be determined.

cDNA map:

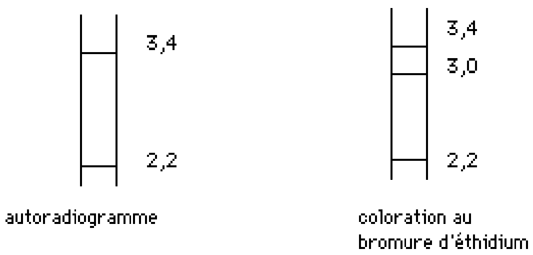
The cDNA fragment was removed from the plasmid and its ends were labeled with 32P. It was then digested with restriction enzymes. The analyses gave the following results:


Doubles digestions avec *Hae* III et *Taq* I
Digestion avec *Taq* I
0,8
0,6
0,5
0,4
0,8
0,5
1,2
1,1
coloration au bromure d'éthidium de l'électrophorèse des fragments
Gels exposés a un film sensible aux rayons X

- Determine the cDNA map?
- Determine the genomic DNA map?

The genomic DNA fragment was extracted from a lambda phage clone by Eco RI digestion, and its ends were labeled with 32P. It was then digested with restriction enzymes. The analyses gave the following results:

Digestion *Bam* HI complète et électrophorèse



b. Draw the genomic map of the fragment, positioning the restriction sites.

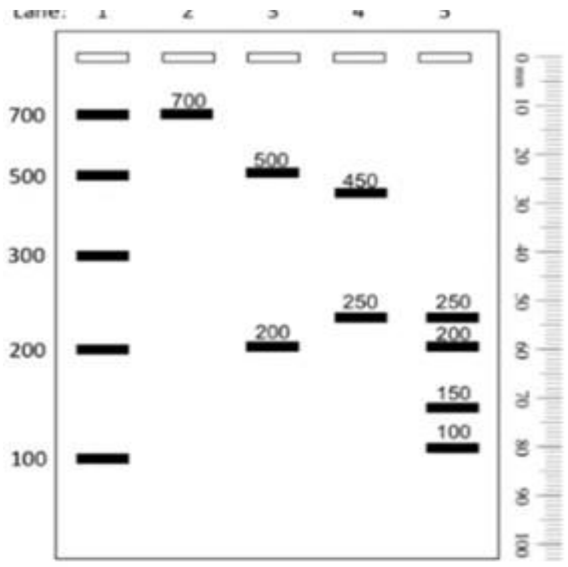
c. How much of the gene does the 3.0 kb genomic fragment represent?

d. A labeled cDNA probe hybridizes to the 3.4 and 2.2 kb genomic fragments. The 1.2 kb Taq I fragment hybridizes to the 3.4 kb genomic fragment.

- If the phosphatase gene is present in single copy, to which genomic fragment(s) will the 1.1 kb Taq I fragment hybridize?

Exercise 6:

1- The sizes of the fragments obtained after digestion of a circular plasmid with cut sites by the EcoRV and BamHI enzymes are in the gel below:



Numéro de puits	Echantillon
1	standards
2	indigéré
3	Digéré par EcoRV
4	Digéré par BamHI
5	EcoRV & Bam HI

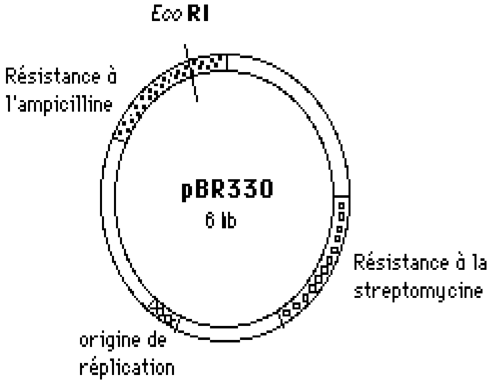
Les tailles des fragments sont indiquées sur chaque bande.

-Reconstruct the restriction maps of this plasmid by positioning the sites of each enzyme: Eco RV and Bam HI, respectively from the restriction products? Then the total map?

Exercise 7:

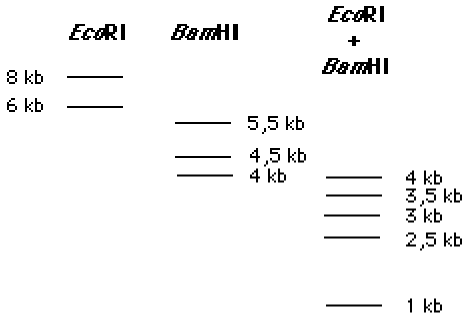
We want to study the mouse gene homologous to the beef M gene that has been cloned in another laboratory. To do this, we try to clone at the EcoRI site of the plasmid vector pBR330 (see diagram)

an EcoRI-EcoRI fragment of mouse genomic DNA carrying the gene homologous to the beef M gene.



a- Propose a cloning protocol and indicate how you select recombinant clones.

b- One of the recombinant plasmids containing the M gene (called pBM1) is digested with the restriction enzymes Bam HI and Eco RI. After migration and separation of the DNA fragments on agarose gel and then staining with ethidium bromide, the following restriction profiles are obtained:



-Give the restriction map of the recombinant plasmid pBM1?

c- A transfer on nitrocellulose membrane is prepared from the piece of the gel from question b corresponding to the Eco RI-Bam HI double digestion.

The membrane is hybridized with a probe consisting of the plasmid pBR330 labeled with 32P. Among the 5 fragments generated during the Eco RI-Bam HI double digestion, to which fragment(s) will the probe hybridize?

Exercise 8: Analyzing gel results post Tap-tag purification

We have to analyze the results of protein purification by Tap-tag technique using SDS-PAGE.

1. You have performed a Tap-tag purification and obtained an SDS-PAGE gel with several bands in the purified sample.

2. Compare the results of the **pre-purification** sample and the **post-purification** sample.

3. Your goal is to determine if your protein of interest has been successfully isolated and assess the purity.

Questions:

- Which band corresponds to your protein of interest?

- Was the purification successful? How can you tell?

- Are there any non-specific bands present? If so, what purification steps can you adjust to improve purity?

Exercise 9: Understanding the mechanism of RNAi

Read through the following statements about RNAi and determine if they are **True** or **False.** Provide an explanation for each statement.

1. RNAi involves the degradation of messenger RNA (mRNA) to prevent translation.

2. The RNAi pathway is activated by double-stranded RNA (dsRNA) molecules, which are recognized by Dicer enzymes.

3. The primary result of RNAi is an increase in the expression of the target gene.

4. Small interfering RNAs (siRNAs) are incorporated into the RNA-induced silencing complex (RISC), where they guide the complex to the target mRNA.

5. RNAi can be used to silence both coding and non-coding genes.

Exercise 10: Interpreting sequencing data

You have obtained the following chromatogram (electropherogram) from a Sanger sequencing reaction. The peaks correspond to the following bases:

• **A**: 0.5 **T**: 0.3 • **G:** 0.2 1. **C**: 0.1

1. How would you interpret this chromatogram?

2. What could cause poor peak resolution in a chromatogram?