Dr.Umayal Ramanathan College for Women
Accredited with B+ Grade by NAAC
Affiliated to Alagappa University
(Run by Dr.Alagappa Chettiar Educational Trust)
Karaikudi- 630003
Department of Microbiology and Clinical Lab Technology

7BMC6C3

RECOMBINANT DNA TECHNOLOGY

Dr.B.Eswara Priya
Asso.Prof & Head
Department of Microbiology and CLT
Dr.Umayal Ramanathan College for Women
Karaikudi- 630003

III YEAR – VI SEMESTER

COURSE CODE: 7BMC6C3

CORE COURSE - XIV – RECOMBINANT DNA TECHNOLOGY

Unit I

History of rDNA Technology - Enzymes in rDNA Technology – Ribonuclease-H (RNase-H), Klenow enzymes or klenow Fragment, SI Nuclease, Taq DNA Polymearse, Restriction Endonucleases, Terminal Nucleotidyl Transferase, Alkaline Phosphatase, Polynucleotide Kinase, DNA ligases and Methyl transferase. Coupling Tools – Linkers and Adaptors. Construction and of Applications rDNA.

Unit II

Gene cloning: Strategies in gene cloning. Plasmids – Introduction and classification. Gene cloning vectors: pBR322, pUC, ColE1 plasmid. Cosmids and phagemid as vectors. Shuttle vectors, Expression vectors.

Unit III

Gene transfer techniques: Microinjection, Electroporation, Microprojectile, Shot Gun method, Ultrasonication and Liposome fusion. Selection of recombinant Bacteria - Direct selection, Antibiotic resistance and lacZ complementation (Blue-white selection).

Unit III

Construction of genomic and cDNA libraries. Site directed mutagenesis, Chromosome jumping. Safety regulations in rDNA techniques.

.

Unit V

Genetically Engineered Microorganisms(GEMOs). Production of Healthcare products from GEMOs-Insulin, Human growth hormone, Interferons, Blood products and Vaccines. ``

Books for Reference:

1. Principles of Gene Manipulation and Genomics - Primrose, S.B. and Twyman,R.M. 2006. 7th Edition. Blackwell Publishing Company.
2. Recombinant DNA Second Edition - James D. Watson, Micheal Gilman, Mark Zoller, 2001. W.H. Freeman and Company, New York.
3. Biotechnology, Satyanarayana. U, (2008), Books and Allied (p) Ltd.
4. A Text Book of Biotechnology. R.C. Dubey. S.Chand& Co Ltd, New Delhi.
5. Genomes 3 by T.A.Brown, Third Edition (Garland Science Publishing), 2007.

♣♣♣♣♣♣♣♣♣♣

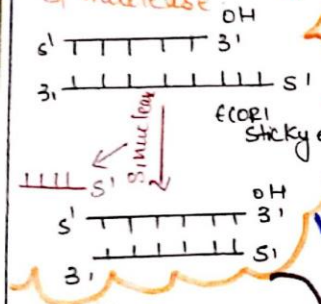
Unit I

Unit-1.

Enzyme Used in rDNA Technology

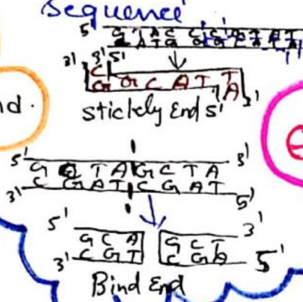
S1 Nuclease

The single standard extension present on the hairpin loop It is cleave by S1 nuclease



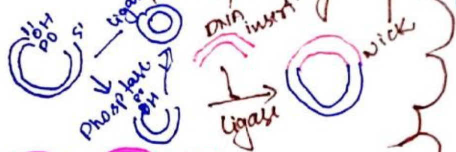
Restriction endonuclease

* Restriction endonuclease are Special class of endonucleases cleaves DNA molecules only at specefic nucleotide sequence



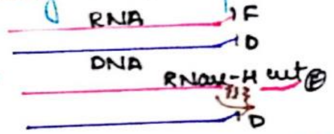
Alkaline Phosphatase

Alkalene Phosphate represent a group of enzyme that remove phosphate group.



Ribonuclease-H (RNase-H)

* An enzyme that cleave the RNA primer from okazaki fragments during DNA replication



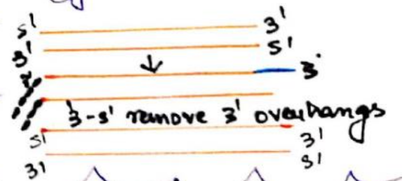
Polynucleotide Kinase

Polynucleotide Kinase add the Phosphate group at 5' end

Klenow enzyme (ou)

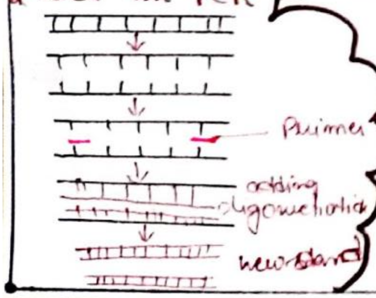
Klenow fragment

* A breakdown product of DNA Polymerase I.
* The Klenow fragment that act as klenow enzyme.



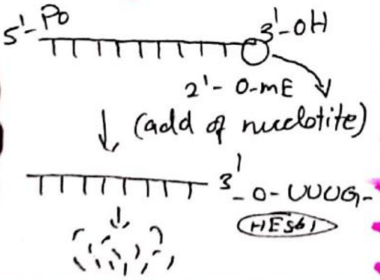
Tag Polymerase Enzyme

* It is extracted from thermus aquaticus.
* It add oligonucleotides.
* Used in PCR



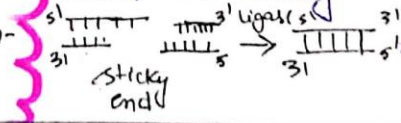
Terminal nucleotidyl Transferase

* It adds the nucleotide to the terminal end of the strand



DNA ligase

Is a specefic type of enzyme. It covalently joins the Phosphate backbone of DNA with blunt or sticky.



Methyl transferase

Methyl group is transferred s-Adenosyl methionine methylation process.

I-History of rDNA Technology

Introduction:

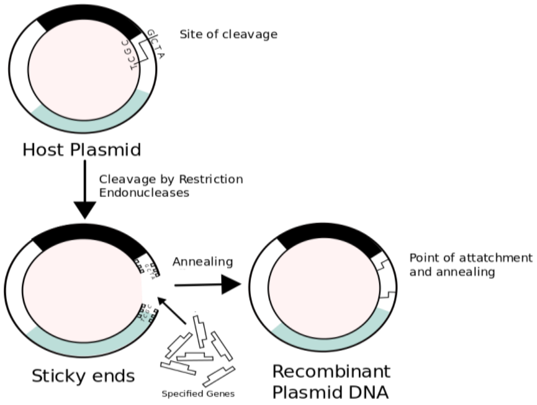
Recombinant DNA (rDNA) molecules are DNA molecules formed by laboratory methods of genetic recombination (such as molecular cloning) to bring together genetic material from multiple sources, creating sequences that would not otherwise be found in the genome.

Recombinant DNA is the general name for a piece of DNA that has been created by combining at least two strands. Recombinant DNA is possible because DNA molecules from all organisms share the same chemical structure, and differ only in the nucleotide sequence within that identical overall structure. Recombinant DNA molecules are sometimes called chimeric DNA, because they can be made of material from two different species, like the mythical chimera. R-DNA technology uses palindromic sequences and leads to the production of sticky and blunt ends.

The DNA sequences used in the construction of recombinant DNA molecules can originate from any species. For example, plant DNA may be joined to bacterial DNA, or human DNA may be joined with fungal DNA. In addition, DNA sequences that do not occur anywhere in nature may be created by the chemical synthesis of DNA, and incorporated into recombinant molecules. Using recombinant DNA technology and synthetic DNA, literally any DNA sequence may be created and introduced into any of a very wide range of living organisms.

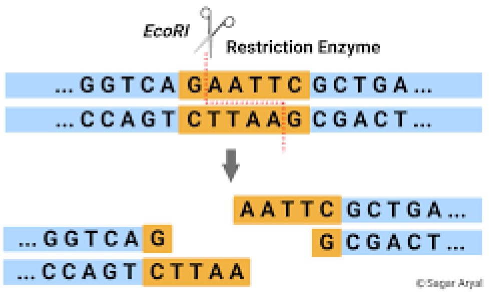
Proteins that can result from the expression of recombinant DNA within living cells are termed *recombinant proteins*. When recombinant DNA encoding a protein is introduced into a host organism, the recombinant protein is not necessarily produced.[1] Expression of foreign proteins requires the use of specialized expression vectors and often necessitates significant restructuring by foreign coding sequences.[2]

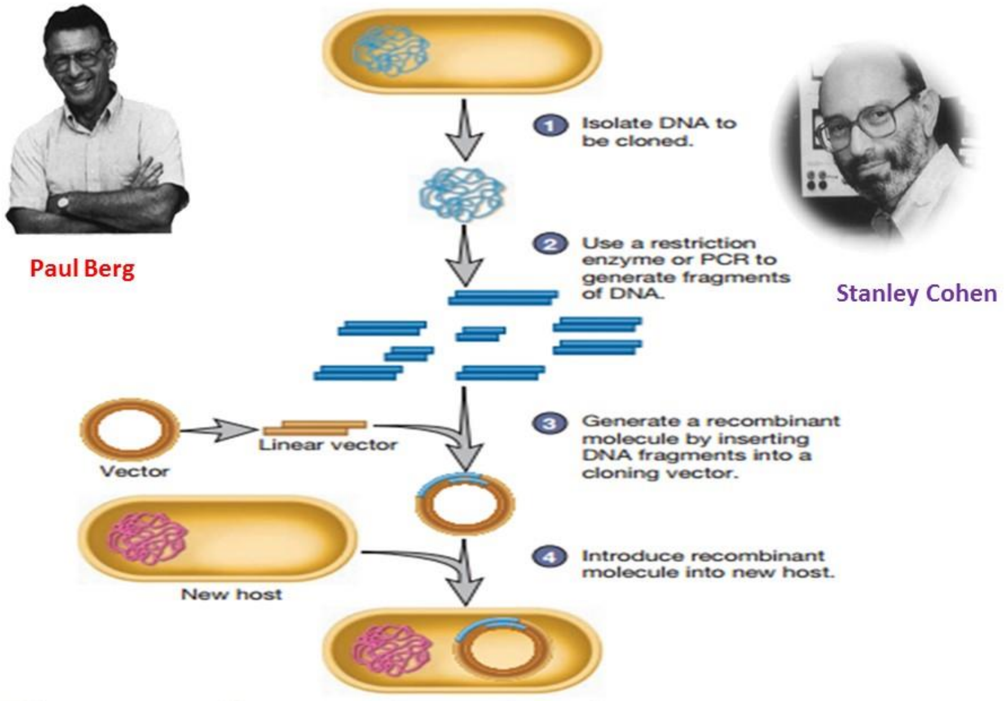
Recombinant DNA differs from genetic recombination in that the former results from artificial methods in the test tube, while the latter is a normal biological process that results in the remixing of existing DNA sequences in essentially all organisms.



History of Recombinant DNA Technology

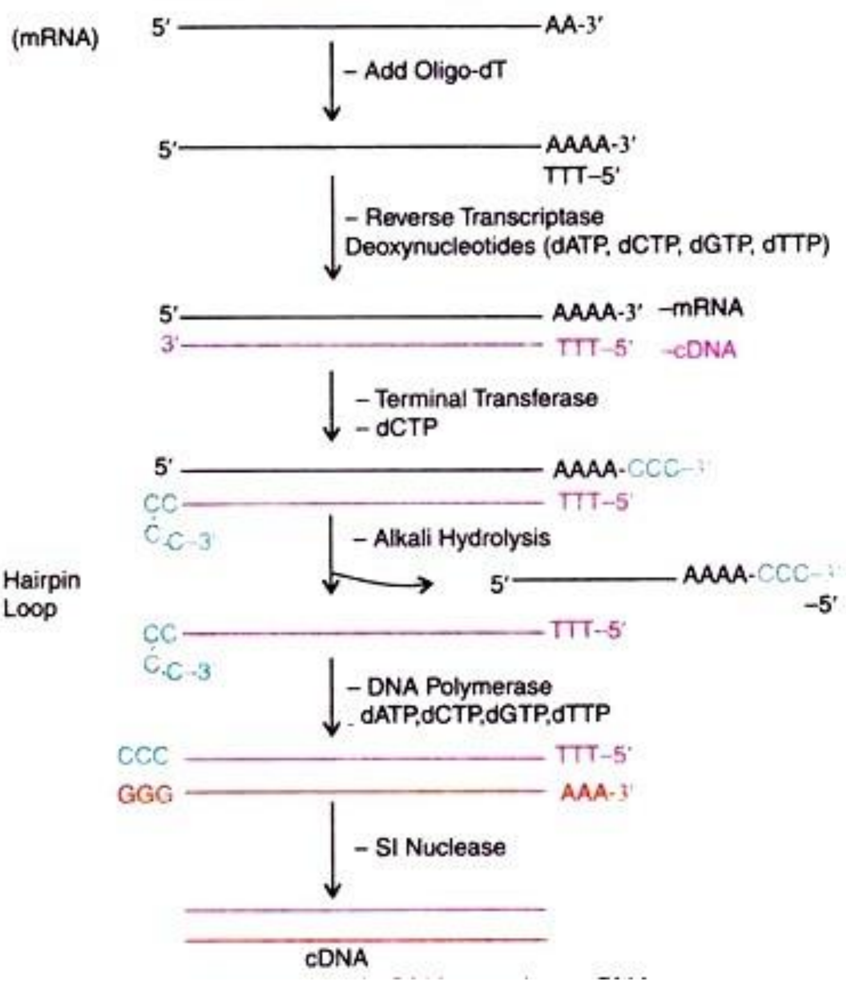
- Over the past decades, the development of new and powerful techniques for studying and manipulating DNA has revolutionized genetics. These techniques have allowed biologists to intervene directly in the genetic fate of organisms for the first time. Hence, this is called as the gene technology or recombinant DNA technology.
- Recombinant DNA technology refers to the joining together of DNA molecules from two different species that are inserted into a host organism to produce new genetic combinations that are of value to science, medicine, agriculture, and industry.
- Recombinant DNA technology was invented largely through the work of American biochemists Stanley N. Cohen, Herbert W. Boyer, and Paul Berg.
- The first break through of rDNA technology occurred with the discovery of restriction endonucleases (restriction enzyme) during the late 1960s by Werner, Arber and Hamilton Smith. The restriction enzymes were discovered in microorganisms. These enzymes protect the host cell from the bacteriophage.
- In 1969, Herbert Boyer isolated restriction enzyme EcoRI from E. coli that cleaves the DNA between G and A in the base sequence GAATTC as below:




Paul Berg
Stanley Cohen
1 Isolate DNA to be cloned.
2 Use a restriction enzyme or PCR to generate fragments of DNA.
Vector
Linear vector
3 Generate a recombinant molecule by inserting DNA fragments into a cloning vector.
New host
4 Introduce recombinant molecule into new host.

Steps in Gene Cloning

- In 1970 Howard Temin and Davin Baltimore independently discovered the enzyme reverse transcriptase from retroviruses. Later on this enzyme was used to construct a DNA called complementary DNA (cDNA) from any mRNA.



Synthesis of cDNA by using mRNA

- In, 1972 David Jackson, Robert Symons and Paul Berg successfully generated rDNA molecules. They allowed the stickly ends of complementary DNA by using an enzyme DNA ligase.

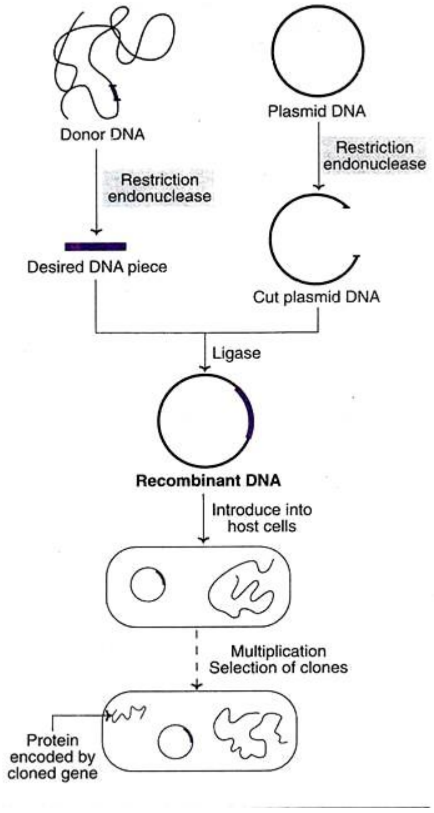
- In 1973 for the first time S.Cohen and H. Boyer developed a recombinant plasmid (p^{SC101}) which after using as vector replicated well within a bacterial host.

- In, 1975, Edwin M.Southern developed a method for detection of specific DNA fragments for isolation of a gene from complex mixture of DNA. This method is known as the Southern blotting technique.

- 1976 – First prenatal diagnosis by using gene specific probe.

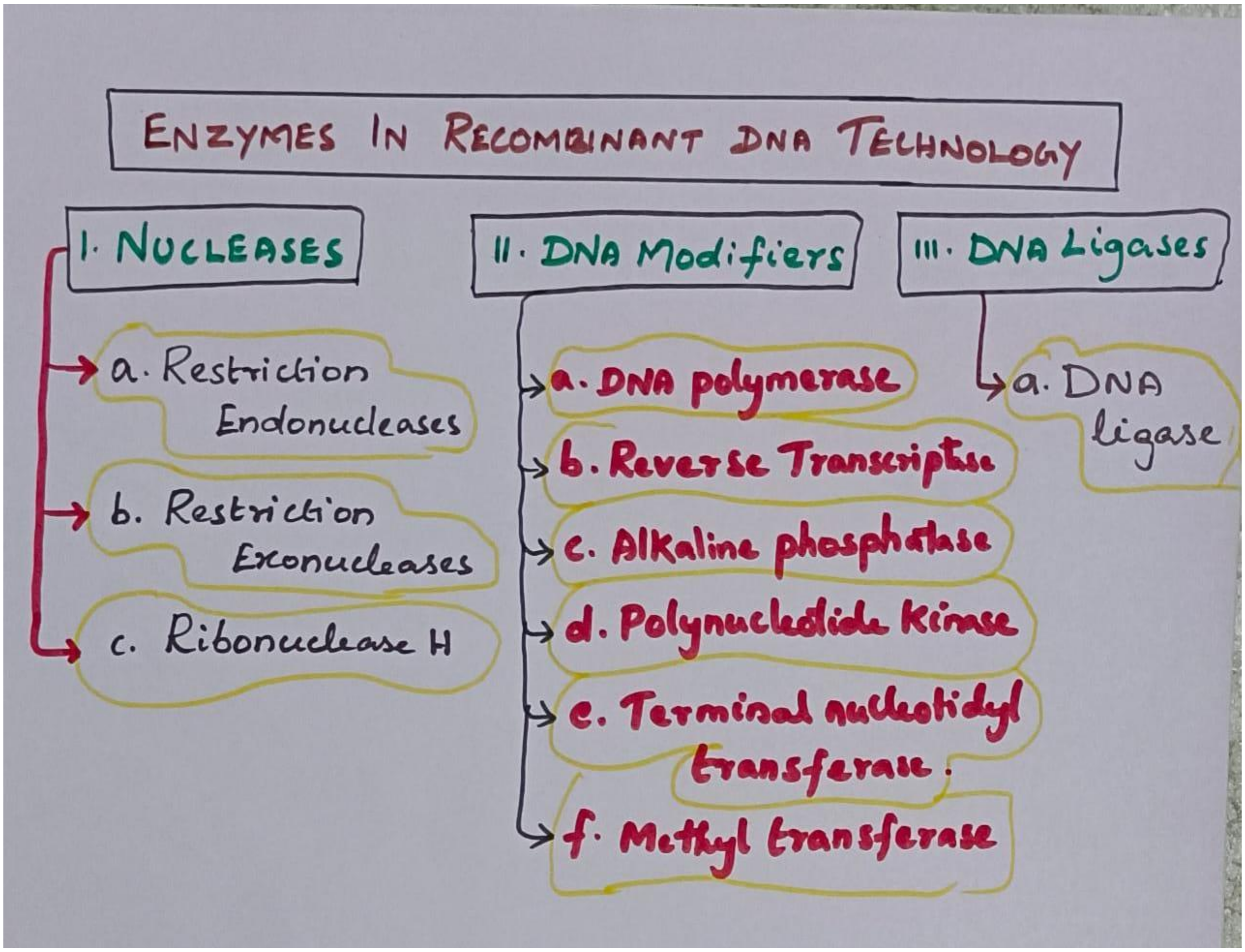
- 1977 – Methods for rapid DNA sequencing, discovery of split genes and somatostanin by rDNA.

- 1979 – Insulin synthesized by using rDNA; first human viral antigen.
- 1981 – Foot and mouth disease viral antigen cloned.
- 1982 – Commercial production of coli of genetically engineered human insulin, Isolation, cloning and characterization of human cancer gene.
- 1983 – Engineered Ti-plasmid used to transform plants.
- 1985 – Insertion of cloned gene from Salmonella into tobacco plant to make resistant to herbicide glyphosphate; Development of PCR technique.
- 1986 – Development of gene gun.
- 1989 – First field test of genetically engineered virus (baculovirus) that kills cabbage looper caterpillars.
- 1990 – Production of first transformed com.
- 1991 – Production of first transgenic pigs and goats, manufacture of human haemoglobin, first test of gene therapy on human cancer patients.
- 1994 – The Flavr Savr tomato introduced; the first genetically engineered whole food approved for sale. Fully human monoclonal antibodies produced in genetically engineered mice.
- 1997 – World’s first mammalian clone (Dolly) developed from a non-reproductive cell of an adult animal through cloning by nuclear transplantation.
- 2003- Human genome sequenced



II. Enzymes in rDNA Technology / Molecular Tools of Genetic Engineering

An engineer is a person who designs, constructs (e.g. bridges, canals, and railways) and manipulates according to a set plan. The term genetic engineer may be appropriate for an individual who is involved in genetic manipulations. The genetic engineer's toolkit or molecular tools namely the enzymes most commonly used in recombinant DNA experiments are briefly described.



Discovery of Enzymes:

Genetic engineering was born because scientists learned to manipulate DNA. This skill was derived mainly from the field of nucleic acid enzymology. Prior to 1970, there was simply no technique available for cutting a duplex (double-stranded) DNA molecule into distinct fragments. Discovery of DNA metabolising enzymes granted scientists to propose and initiate genetic engineering.

All sort of DNA research developed from the ability, to cut DNA molecules at defined sequences. In other words, it was based upon the discovery of type II restriction endonuclease enzymes.

The isolation of the first restriction endonuclease enzymes, such as Hind II and Hind III, was a result of an interesting discovery by Hamilton O. Smith and his coworkers (1971) that Haemophilus influenza extracts contained activities that cut large DNA molecules into defined fragments.

*H. influenza*e is a non-motile, gram-negative, facultative, anaerobic, pathogenic, rod shaped bacterium which is associated with human respiratory infections, conjunctivitis and meningitis.

Smith’s discovery gave rise to recombinant DNA research. In addition, an entire industry developed with the main purpose of discovery, characterization, purification and marketing of over 100 different site-specific restriction enzymes.

The bringing together of DNA fragments to form covalently linked chimeric molecules is the basis of recombinant DNA research. This step is essential in genetic engineering. This is attained by ligation which is catalysed by DNA ligase, an enzyme which was discovered much prior to that of restriction enzymes.

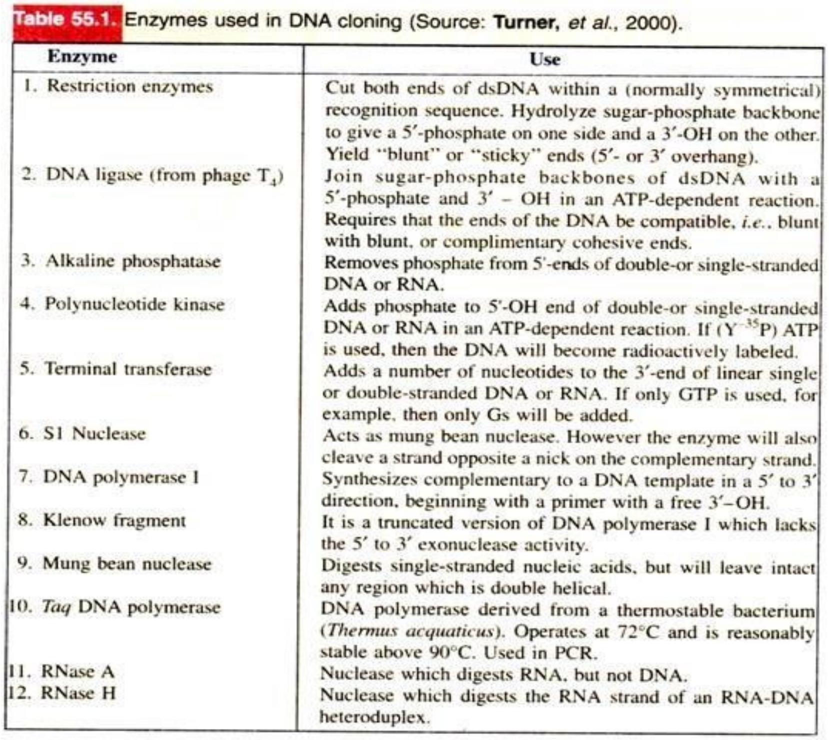
Before 1970 the existing central dogma in molecular biology was that genetic information transfer occurred from DNA to RNA, and then to protein. The proof that RNA-to-DNA information transfer did occur is based on the discovery and characterization of reverse transcriptase enzyme by Temin and Baltimore (1970).

Reverse transcriptase enzyme allows scientists to generate DNA copies (cDNA) of mRNA subsequent to cloning. The generation of cDNA, containing direct protein coding information is the normal step in cloning of eukaryotic genes.

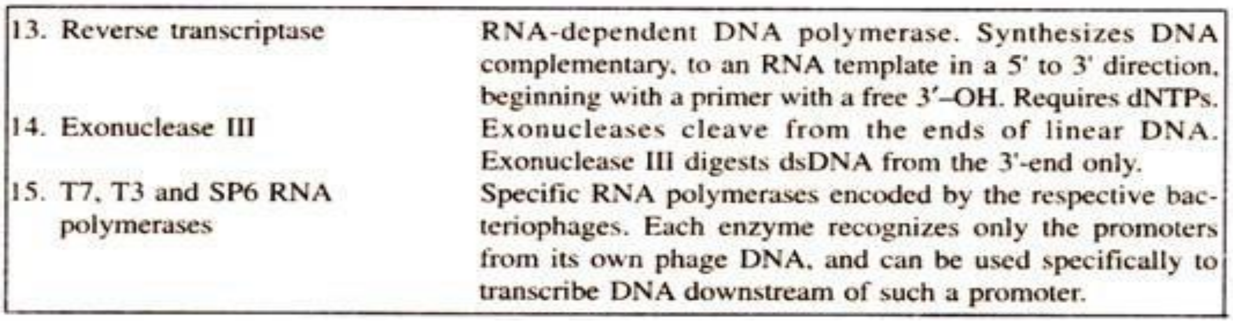
In fact the most revolutionary and most simplistic molecular biological technical development, is PCR (i.e., polymerase chain reaction). PGR is a direct application of DNA polymerase to permit the test tube binomial amplification of specific DNA sequences.

The discovery of type II restriction enzymes demonstrated the enormous power and utility of site specific DNA cleavage reagents. This article deals with the restriction enzymes and other useful enzymes which are commonly used in genetic engineering (Table 55.1).

Table 55.1. Enzymes used in DNA cloning (Source: **Turner,** *et al.*, 2000).



Enzyme	Use
1. Restriction enzymes	Cut both ends of dsDNA within a (normally symmetrical) recognition sequence. Hydrolyze sugar-phosphate backbone to give a 5′-phosphate on one side and a 3′-OH on the other. Yield “blunt” or “sticky” ends (5′- or 3′ overhang).
2. DNA ligase (from phage T_4)	Join sugar-phosphate backbones of dsDNA with a 5′-phosphate and 3′ – OH in an ATP-dependent reaction. Requires that the ends of the DNA be compatible, *i.e.*, blunt with blunt, or complimentary cohesive ends.
3. Alkaline phosphatase	Removes phosphate from 5′-ends of double-or single-stranded DNA or RNA.
4. Polynucleotide kinase	Adds phosphate to 5′-OH end of double-or single-stranded DNA or RNA in an ATP-dependent reaction. If ($Y^{-35}P$) ATP is used, then the DNA will become radioactively labeled.
5. Terminal transferase	Adds a number of nucleotides to the 3′-end of linear single or double-stranded DNA or RNA. If only GTP is used, for example, then only Gs will be added.
6. S1 Nuclease	Acts as mung bean nuclease. However the enzyme will also cleave a strand opposite a nick on the complementary strand.
7. DNA polymerase I	Synthesizes complementary to a DNA template in a 5′ to 3′ direction, beginning with a primer with a free 3′–OH.
8. Klenow fragment	It is a truncated version of DNA polymerase I which lacks the 5′ to 3′ exonuclease activity.
9. Mung bean nuclease	Digests single-stranded nucleic acids, but will leave intact any region which is double helical.
10. *Taq* DNA polymerase	DNA polymerase derived from a thermostable bacterium (*Thermus acquaticus*). Operates at 72°C and is reasonably stable above 90°C. Used in PCR.
11. RNase A	Nuclease which digests RNA, but not DNA.
12. RNase H	Nuclease which digests the RNA strand of an RNA-DNA heteroduplex.



13. Reverse transcriptase	RNA-dependent DNA polymerase. Synthesizes DNA complementary, to an RNA template in a 5' to 3' direction, beginning with a primer with a free 3′–OH. Requires dNTPs.
14. Exonuclease III	Exonucleases cleave from the ends of linear DNA. Exonuclease III digests dsDNA from the 3'-end only.
15. T7, T3 and SP6 RNA polymerases	Specific RNA polymerases encoded by the respective bacteriophages. Each enzyme recognizes only the promoters from its own phage DNA, and can be used specifically to transcribe DNA downstream of such a promoter.

a. **Ribonuclease:**

Generally RNase A and RNase T1 enzymes are used in genetic engineering techniques. Both enzymes cleave the phosphodiester bond between adjacent ribonucleotides. RNase A cleaves next to uracil (U) and cytosine (C) in such a way that phosphate remains with these pyrimidines. The nucleotide present on the other side of phosphate is dephosphorylated. RNase A enzyme is isolated from the bovine pancreas.

RNase T1 cleaves specifically next to guanine. The phosphate group at the 3′ end of the nucleotide remains with the cut end. This enzyme is isolated from Aspergillus oryzae.

Ribonuclease H (RNase H):

The enzyme RNase H is an endoribonuclease that degrades the RNA portion of the RNA- DNA hybrids. RNase H enzyme cuts the RNA into short fragments.

Applications of RNase H:

1. RNase H is the key enzyme in the cDNA cloning technique. In this case, it is used to remove the mRNA from the RNA-DNA hybrid.

2. RNase H enzyme is used to detect the presence of RNA-DNA hybrid.

3. RNase H enzyme is used to remove poly (A) tails on mRNA.

b. **DNA Polymerase I: Holoenzyme:**

This enzyme has two-fold activities:

5′ → 3’exonuclease activity and DNA synthesis (acts as 5’ -3’ polymerase). Such a bifunctional activity enables the DNA polymerase I enzyme to use nicks or gaps in double- a stranded DNA as a starting point of DNA synthesis.

The 5′-exonuclease activity degrades that DNA strand which is complementary to the template strand and thus forming nick. DNA synthesis begins at 3′-end of nick and produces a new strand of DNA complementary to the template.

The new result is the movement of nick along the template strand (nick translation) until all the DNA complementary to the template strand (starting from the site of the origin of nick to the 5′-end of the template strand) is replaced.

Uses of DNA Polymerase I:

1. DNA polymerase I enzyme is used with radioactive or biotinylated nucleotides to prepare labelled DNA of high specific activity.

2. DNA polymerase I enzyme can also catalyse de novo DNA synthesis.

3. The enzyme DNA polymerase I has 3’→5′ proof-reading exonuclease activities on a single polynucleotide chain.

b.1- DNA Polymerase I: Klenow Fragment:

Treatment of DNA polymerase I holoenzyme of E. coli with protease enzyme results in the production of two protein fragments. The larger fragment is called Klenow fragment and it does not show 5′-exonuclease activity (i.e., the 5′- exonuclease activity is exhibited by the intact enzyme) This Klenow fragment is used to synthesize DNA when there is no need of removing the DNA strand which is complementary to the template strand.

Uses of Klenow Fragment:

Klenow fragments are used in the following ways:

1. In DNA sequencing by dideoxy method.

2. For the production of second strand of cDNA.

3. Radiolabelling by filling in 5′-single stranded extension on double-stranded DNA.

4. Mutagenesis of DNA with synthetic oligonucleotides.

5. In labelling the DNA by random primer method.

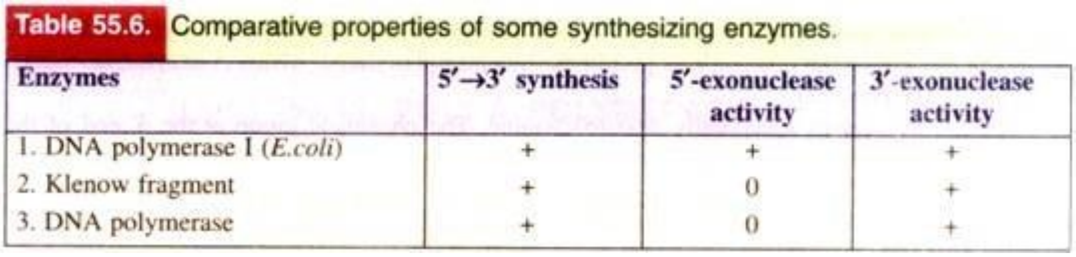


Table 55.6. Comparative properties of some synthesizing enzymes.

Enzymes	5′→3′ synthesis	5′-exonuclease activity	3′-exonuclease activity
1. DNA polymerase I (*E.coli*)	+	+	+
2. Klenow fragment	+	0	+
3. DNA polymerase	+	0	+

+ stands for presence of particular activity by the enzyme.

0 presents absence of such activity.

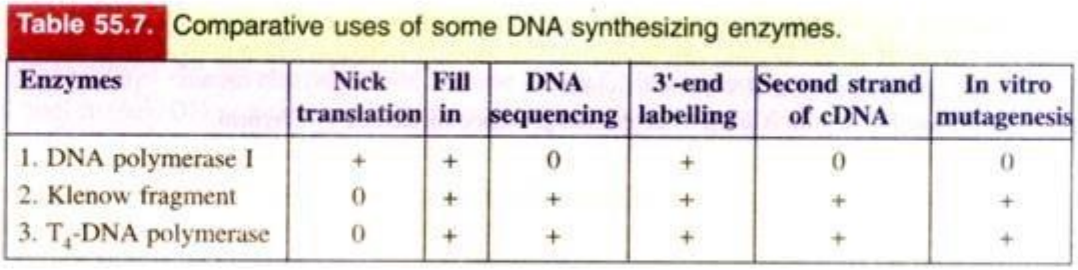


Table 55.7. Comparative uses of some DNA synthesizing enzymes.

Enzymes	Nick translation	Fill in	DNA sequencing	3'-end labelling	Second strand of cDNA	In vitro mutagenesis
1. DNA polymerase I	+	+	0	+	0	0
2. Klenow fragment	0	+	+	+	+	+
3. T_4-DNA polymerase	0	+	+	+	+	+

c. S1 Nuclease Enzyme:

The S1 nuclease enzyme is single- strand specific endonuclease which cleaves DNA to release 5′-mono and 5′-oligonucleotides. Normally, double- stranded DNA, double- stranded RNA and DNA-RNA hybrids are resistant to action of S1 nuclease enzyme.

However, very large amounts of S1 nuclease enzyme can completely hydrolyze double- stranded nucleic acids. The enzyme hydrolyzes single stranded regions in duplex DNA such as loops and gaps.

S1 nuclease enzyme can also cleave single stranded areas of super helical DNA at torsional stress points where DNA may be unpaired or weakly hydrogen bonded. Once the super-helical DNA is nicked, S1 nuclease enzyme can cleave the second strand near the nick to generate linear DNA.

S1 nuclease enzyme is a monomeric protein with 3800 dalton molecular weight. It requires Zn^{2+} for its activity and is relatively stable against denaturing reagents such as urea, SDS and formamide. The optimum pH requirement lies between 4 to 4.5.

Uses of S1 Nuclease Enzyme:

1. S1 nuclease enzyme is used to analyse DNA-RNA hybrid structures to map transcripts.

2. It can be used to remove singles stranded tails from DNA fragments to produce blunt ends.

3. Hair pin loop structures formed during synthesis of double-stranded cDNA is digested by this enzyme.

4. S1 nuclease enzyme is also used for DNA mapping called SI nuclease mapping Turner.

d. Taq DNA Polymerase Enzyme:

The enzyme Taq DNA polymerase is isolated from the thermophilic bacterium Thermus aquaticus. Taq enzyme has the highest DNA polymerase activity at a pH of 9 and temperature around 75°C.

Activity of Taq DNA polymerase is resistant to incubation at high as 95°C. Taq enzyme consists of a single polypeptide chain with a molecular weight of 95000. It lacks 5′ to 3′ and 3′ to 5′ exonuclease activity.

The highly thermostable Taq DNA polymerase from Thermus aquaticus is ideal for both manual and automated DNA sequencing because it is fast, highly progressive, has little or no 3’ – exonuclease activity and is active over a broad range of temperature.

Application of Taq Polymerase:

1. Taq enzyme is used in DNA sequencing studies.

2. Taq enzyme is used in ‘Polymerase chain reaction’ or PCR as it can withstand high temperatures.

e. **Restriction Endonuclease Enzyme:**

Restriction endonucleases (RE) are special class of endonucleases which cleave DNA molecules only at specific nucleotide sequences, called restriction sites. These specific sequences are of four to six nucleotides.

A tetranucleotide sequence will occur more frequently in a given molecule than hexanucleotide, therefore, more fragments will be produced by an enzyme which recognizes tetranucleotide sequence. At these restriction sites, restriction endonucleases cut the DNA by cleaving two phosphodiester bonds one within each strand of the double stranded DNA.

The term 'restriction endonuclease' was coined by Lederberg and Meselson (1964) to describe the nuclease enzymes that destroy ('restrict') any foreign DNA entering the host cell. However, the first restriction endonuclease enzyme to be isolated and studied was E. coli K12 by Meselson and Yuan (1968). Now these enzymes have been classified into three different types viz., Type I, Type II and Type III.

The discovery of these enzymes led to Nobel Prize for W. Arber, H. Smith and D. Nathans in 1978. In gene manipulation technology, restriction endonuclease enzymes are popularly called molecular knives, molecular scissors or molecular scalpels.

Restriction Enzymes:

Steward Linn and Werner Arber (1963) isolated two enzymes which restricted the growth of bacteriophage in bacterium E. coli. One of these enzymes added methyl groups to DNA and second one cut DNA. The second enzyme was named as **"restriction endonuclease."**

H.O. Smith, K.W. Wilcox and T.J. Kelley (1968) isolated restriction endonuclease whose working depended on a particular nucleotide sequence. They isolated this enzyme from bacteria Haemophilus influenzae and called is as Hind II. It was observed that Hind II always cut DNA molecules at specific place by identifying a particular sequence of six base pairs.

Restriction enzymes belong to a larger class of enzymes called nucleases.

They are of two types:

(i) Exonucleases:

They remove nucleotides from the ends of DNA.

(ii) Endonucleases:

They make cuts at specific positions within DNA.

Thus, a restriction enzyme (or restriction endonuclease) recognizes a specific base pair sequence in DNA called a restriction site and cleaves the DNA (hydrolyzes the phosphodiester back bones) within the sequence. Restriction enzymes are widely found in prokaryotes and provide protection to host cell by destroying foreign DNA that makes entry into it.

Here they act as a part of defence mechanism called Restriction Modification System.

It bears two components:

(a) First component is a restriction enzyme that selectively identifies a specific DNA sequence and degrades any DNA bearing that sequence.

(b) In second component is a modification enzyme. It adds a methyl group to one or two bases within the sequence identified by restriction enzyme. If a base in DNA is modified due to addition of methyl group, restriction enzyme cannot identify and cleave that DNA. By this method bacteria are able to protect their chromosomal DNA from cleavage by restriction enzymes.

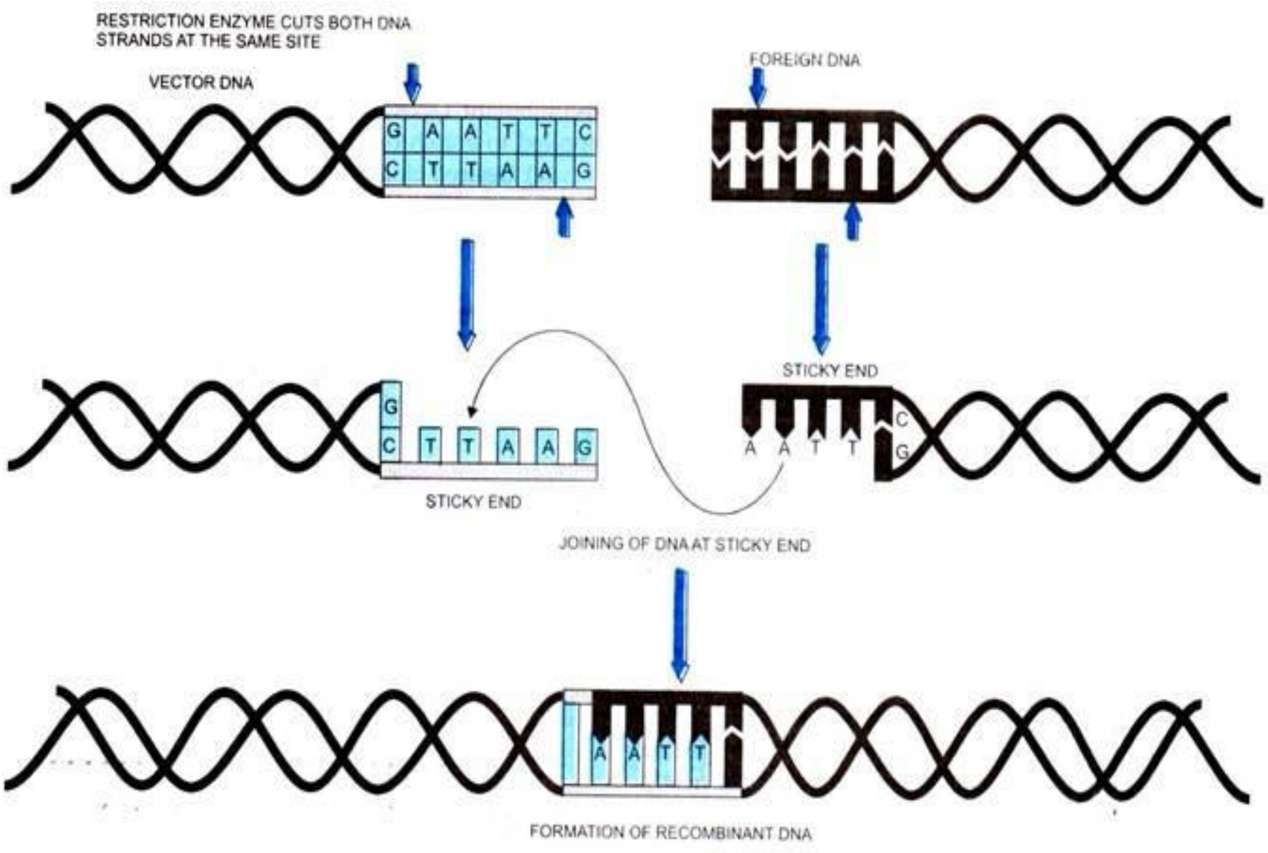

RESTRICTION ENZYME CUTS BOTH DNA STRANDS AT THE SAME SITE
VECTOR DNA
G A A T T C
C T T A A G
FOREIGN DNA
STICKY END
G
C T T A A G
STICKY END
A A T T
C
G
JOINING OF DNA AT STICKY END
A A T T
FORMATION OF RECOMBINANT DNA

Fig. 11.2. Action of restriction enzyme : *EcoR1*-cuts the DNA between bases G and A only when the sequence GAATTC is present in the DNA.

Thus, bacteria bear sets of restriction endonucleases and corresponding methylases.

Endonucleases are enzymes that produce internal cuts called cleavage in DNA molecules. Endonucleases cleave DNA molecules at random sites. A class of endonucleases cleaves DNA only within or near those sites with specific base sequences called restriction endonucleases. Sites recognised by them are called recognition sites or recognition sequences. These sites differ for different restriction enzymes.

Restriction endonucleases serves as the tools for cutting DNA molecules at predetermined sites, which is the basic requirement for gene cloning or recombinant DNA technology.

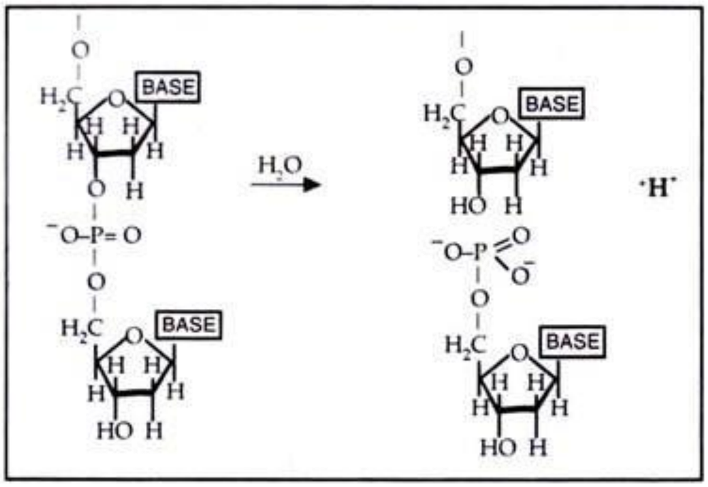
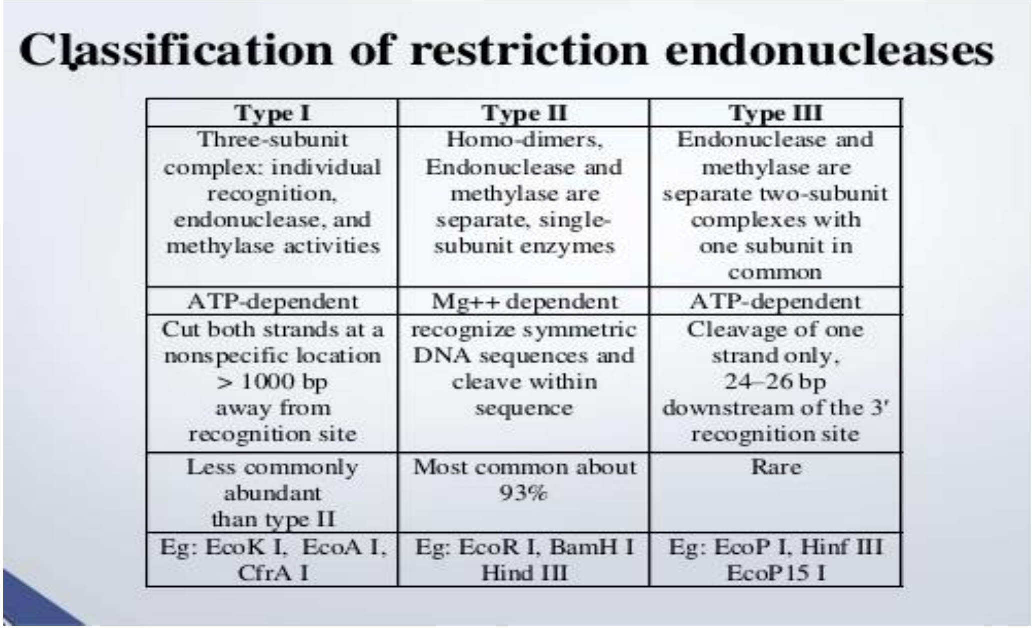


Fig. 11.3. Hydrolysis of 3′-O-P bond in a nucleotide by a restriction endonuclease.

Types of Restriction Endonucleases:

Three main types of restriction endonucleases i.e., Type I, Type II and Type III are known with slightly different mode of action. Type II restriction enzymes are used in rDNA technology because they can be used in vitro to identify and cleave within specific DNA sequences usually having 4-8 nucleotides.

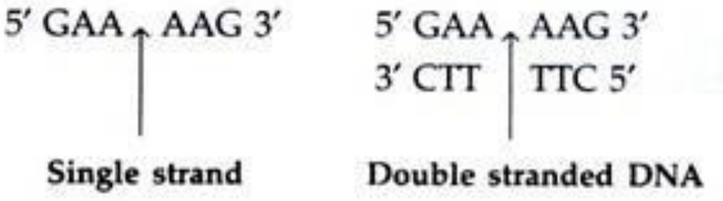
More than 350 different type II endonucleases with 100 different recognition sequences are known. They need Mg^{2+} ions for cleavage. The first type II enzyme isolated was Hind II in 1970.



Classification of restriction endonucleases

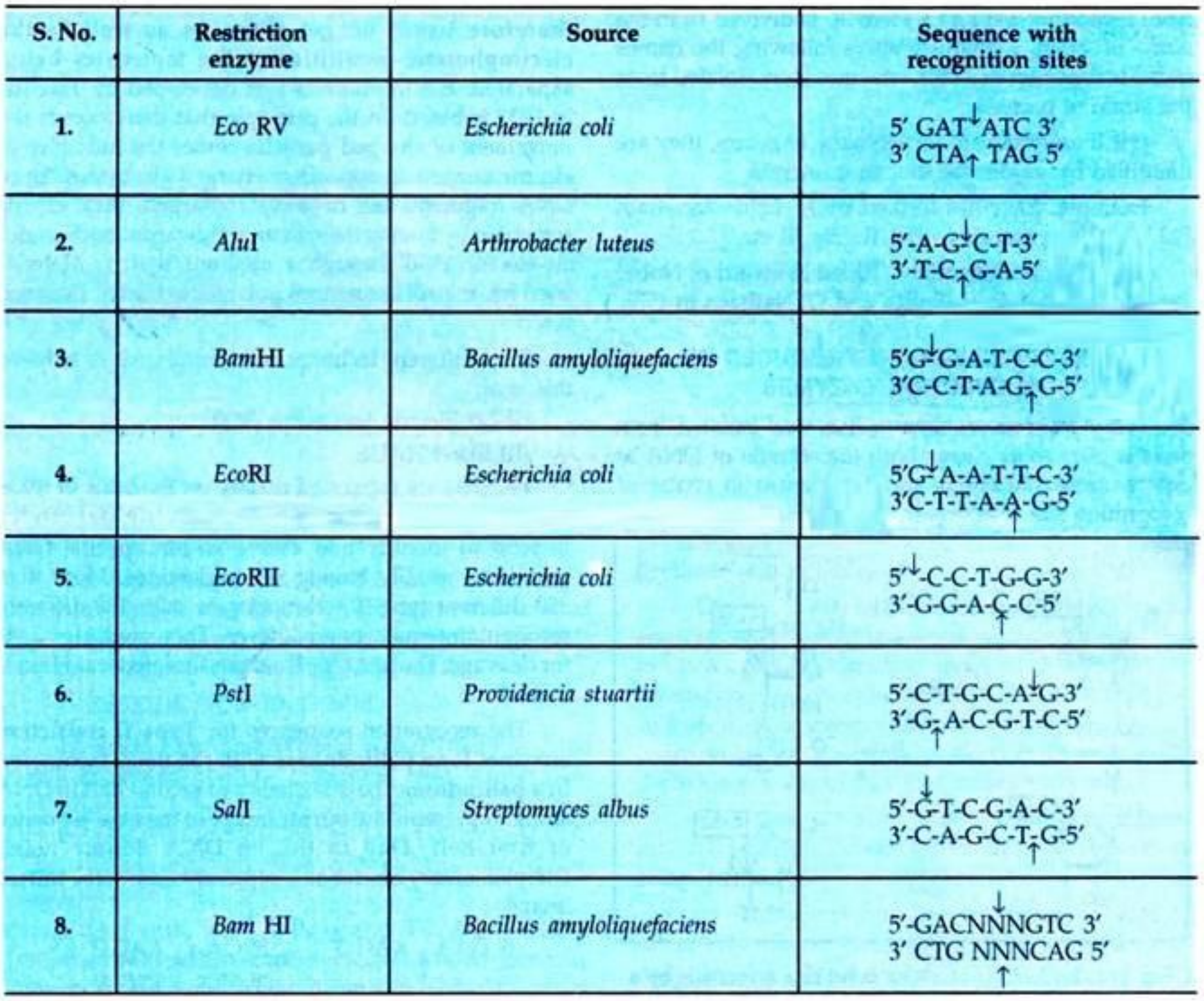
Type I	Type II	Type III
Three-subunit complex: individual recognition, endonuclease, and methylase activities	Homo-dimers, Endonuclease and methylase are separate, single-subunit enzymes	Endonuclease and methylase are separate two-subunit complexes with one subunit in common
ATP-dependent	Mg++ dependent	ATP-dependent
Cut both strands at a nonspecific location > 1000 bp away from recognition site	recognize symmetric DNA sequences and cleave within sequence	Cleavage of one strand only, 24–26 bp downstream of the 3′ recognition site
Less commonly abundant than type II	Most common about 93%	Rare
Eg: EcoK I, EcoA I, CfrA I	Eg: EcoR I, BamH I Hind III	Eg: EcoP I, Hinf III EcoP15 I

The recognition sequences for Type II restriction enzymes form pallindromes with rotational symmetry. In a pallindrome, base sequence of second half in DNA strand represents the mirror image of the base sequence of first half. Due to this in DNA double helix, complementary strand also represents the same mirror image.



Pallindromes are groups of letters that form the same words when read both forward and backward e.g., 'MALAYALAM'. As against a pallindrome when same word is read in both the directions, pallindrome in DNA is a sequence of base pairs that reads same on the two strands when orientation of reading is kept the same.

Table 11.1. Some examples of type II restriction enzyme along with sources and recognition sites on sequences.



S. No.	Restriction enzyme	Source	Sequence with recognition sites
1.	*Eco* RV	*Escherichia coli*	5′ GAT↓ATC 3′ 3′ CTA↑TAG 5′
2.	*Alu*I	*Arthrobacter luteus*	5′-A-G↓C-T-3′ 3′-T-C↑G-A-5′
3.	*Bam*HI	*Bacillus amyloliquefaciens*	5′G↓G-A-T-C-C-3′ 3′C-C-T-A-G↑G-5′
4.	*Eco*RI	*Escherichia coli*	5′G↓A-A-T-T-C-3′ 3′C-T-T-A-A↑G-5′
5.	*Eco*RII	*Escherichia coli*	5′↓-C-C-T-G-G-3′ 3′-G-G-A-C↑-C-5′
6.	*Pst*I	*Providencia stuartii*	5′-C-T-G-C-A↓G-3′ 3′-G↑A-C-G-T-C-5′
7.	*Sal*I	*Streptomyces albus*	5′-G↓T-C-G-A-C-3′ 3′-C-A-G-C-T↑G-5′
8.	*Bam* HI	*Bacillus amyloliquefaciens*	5′-GACNN↓NGTC 3′ 3′ CTG NN↑NCAG 5′

Nomenclature:

Nomenclature of restriction enzymes is usually done by following technique:

(i) The first letter of the genus is taken in which said enzyme was discovered. This letter is written in capital.

(ii) Then, first two letters of species of that organism are written.

(iii) All the above three letters should be written in italics.

Examples:

Eco from Escherichia coli, Hin from Haemophilus inflenzae and Hpa from Haemophilus parainfluenzae.

(iv) This followed by strain or type identification e.g., Eco K.

(v) When the enzyme is encoded by plasmid, the name of plasmid is written e.g. Eco RI i.e., Eco RI comes from Escherichia coli RY13. Here 'R' is derived from the name of strain. Roman numbers following the names indicate the order in which enzymes were isolated from the strain of bacteria.

(vi) If an organism forms many enzymes, they are identified by sequential Roman numerals.

Example:

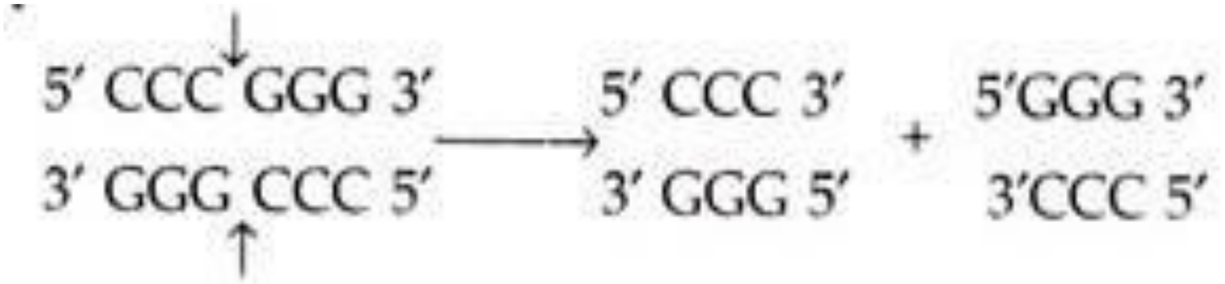
Enzymes formed by H. influenzae strain RD have been named as Hin II, Hin III etc.

Discovery of Enzyme Eco RI led to award of Nobel Prizes to W. Arber, H. Smith and D. Nathans in 1978.

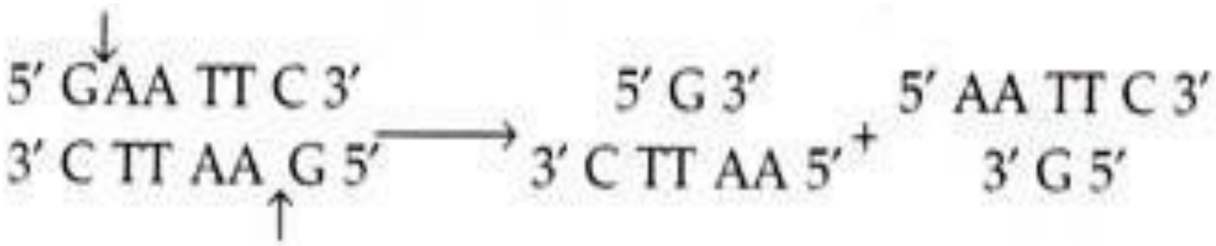
Types of Cleavage Produced By Restriction Enzymes:

Many restriction enzymes like Smal isolated from Serratia marcescens cleave both the strands of DNA at exactly same nucleotide position almost in centre of recognition site resulting in blunt or flush end.

Smal recognizes the 6 nucleotide palindromic sequence and cleave at both the ends.



Still some other restriction enzymes cleave the recognition sequence asymmetrically. Thus due to cleavage, they produce short, single stranded hanging structures. Such ends are called sticky or cohesive ends because base pairing between them can stick the DNA molecule again. 6 nucleotides palindromic nucleotide sequences recognised by Eco RI cleave both strands at different points.



f. Terminal Deoxynucleotidyl Transferase Enzyme:

The enzyme deoxynucleotidyl transferase catalyses the repetitive addition of monodeoxynucleotide units from a deoxynucleoside triphosphate to the terminal 3′-hydroxyl group of a DNA molecule. This enzyme has a molecular weight of 32000 and consists of two subunits each with a molecular weight of 26500 and 8000. This enzyme is isolated from calf thymus.

Uses of Terminal Transferase Enzyme:

1. The enzyme terminal transferase is used to add homopolymer tails of DNA fragmeirts. Using a technique called homopolymer tailing, sticky ends can be built up on blunt-ended DNA molecules.

For examples, one preparation of DNA could be treated with the enzyme terminal transferase in the presence of dATP, resulting in the addition of a poly (dA) chain to each DNA strand. There is another preparation of DNA which provides 3 tails of poly (T) using same enzyme with TTP.

When both types of DNA preparations DNA fragments with poly A tails and DNA fragments with poly T tails, are mixed, there takes place base pairing between complementary sticky ends, which could then be ligated. One advantage of this method is that ligation does not take place between fragments from the same DNA preparation.

2. Terminal transferase enzyme is used for 3′-end labelling of DNA fragments

3. Terminal transferase enzyme is also used for the addition of single nucleotides to the 3- end of DNA for in vitro mutagenesis. Purine and pyrimidine polymerization rates, by using terminal transferase enzyme, depend on the addition of Mg^{2+}, Mn^{2+} or Co^{2+} in the reaction mixture.

g. Alkaline Phosphatase Enzyme:

The enzyme alkaline phosphatase (AP) catalyses the removal of the 5′-terminal phosphate residues from nucleic acids (RNA, DNA and ribo- and deoxyribonucleotide triphosphates). This enzyme is isolated from bacteria (BAP) or calf intestine (CAP).

This enzyme is a dimeric glycoprotein with a molecular weight 14,000. It is made up of two identical or similar subunits each with a molecular weight of 6900. It is a zinc-containing enzyme with four atoms of Zn^{2+} per molecule.

Uses of Alkaline Phosphatase Enzyme:

1. Linearized cloning vectors can be prevented from recircularizing by dephosphorylation with alkaline phosphatase enzyme.

2. The free 5′-OH can be phosphorylated with polynucleotide kinase and Υ^{-32P} ATP to produce 32P end labelled nucleic acid.

3. AP enzyme is used for mapping and DNA fingerprinting studies.

h. Phosphonucleotide/ Polynucleotide Kinase Enzyme:

The enzyme phosphonucleotide kinase catalyses the transfer of the terminal phosphate group of ATP to the 5′-hydroxylated terminal of DNA or RNA. This enzyme is frequently used to end-label the nucleic acids with ^{32}P (i.e., it adds the phosphate back to 5′-termini of DNA).

This can be accomplished by any method among following:

1. Forward reaction:

Transfer of labelled ϒ-phosphate form (ϒ$^{-32}$P)-ATP to the free 5′-hydroxyl group of substrate-

$$5' - \text{OH} - \text{DNA} + [^{-32}\text{P}]\ \text{ATP}5'^{32} \rightarrow \text{PO} - \text{DNA} + \text{ADP}$$

Substrate lacking a free 5′-hydroxyl requires prior dephosphorylation by alkaline phosphatase.

2. Exchange reaction:

In the initial step, the terminal 5′-phosphate is transferred from substrate to ADP present in the reaction mixture. Then, the labelled ϒ-phosphate from [ϒ-^{32}P]-ATP is transferred to free hydroxyl group of substrate.

$$5' - \text{PO} - \text{DNA} + \text{ADP} \rightarrow 5'\text{HO} - \text{DNA} + \text{ATP}$$

$$5' - \text{HO} - \text{DNA} + [\Upsilon^{-32}\ \text{P}] - \text{ATP} \rightarrow 5'^{32}\ \text{PO} - \text{DNA} + \text{ATP}$$

Uses of Polynucleotide Kinase Enzyme:

The enzyme polynucleotide kinase is used to label 5′-termini of DNA and RNA with [ϒ-^{32}P]-ATP by phosphorylation of 5′-hydroxyl groups or by the exchange reaction. This 5′-terminal labelling is used in mapping of restriction sites, DNA or RNA fingerprinting, hybridization studies and sequence analysis of DNA.

i. Ligase Enzymes:

While exact cutting of DNA molecule is very useful for DNA cloning, its full potential is only exhibited when the fragments produced are joined together to give a new structure, known as recombinant DNA. This joining or ligation is achieved by the use of a DNA ligase enzyme.

The most common ligase enzyme is isolated from the bacterial virus (i.e., T4 bacteriophage). Thus, DNA ligases form a group of enzymes which mediate annealing, sealing or joining of DNA fragments. Primarily, ligase enzymes are involved in the repair of DNA molecule where sealing or union of DNA fragments takes place.

DNA ligases also play active part in processes such as DNA replication and recombination. These enzymes are widely used in genetic engineering for the production of hybrid DNA. Since ligase enzymes join DNA fragments or seal the nicks in the chain, they are called molecular structures.

Activity of DNA Ligase Enzymes:

(i) Ligation of DNA molecules with sticky or cohesive ends:

If two different DNA preparations are treated with the same restriction enzyme to give fragments with sticky ends, these ends will be identical in both preparations. Thus, when the two sets of DNA fragments are mixed, base pairing between sticky ends will result in the coming together of fragments which were derived from different molecules.

Also there will be pairing of fragments derived from the same molecule. Such pairing are temporary, owing to the weakness of hydrogen bonding between the few bases in the sticky ends.The pairing can be stabilised by the use of DNA ligase, which forms a covalent bond between the 5′-phosphoryl at the end of one strand and 3′-hydroxyl of the adjacent strand. Polynucleotide ligase enzyme of T4 bacteriophage catalyzes the end to end joining of DNA duplexes at the base paired end. This reaction could occur intermolecularly or intramolecularly. Researches confirm that intermolecular mode of reaction is correct.

The ligation reaction is driven by ATP and is carried out at 4°C to lower the kinetic energy of molecules. This reduces separation of paired sticky ends and are later stabilised by ligation. However, long reaction time is required to compensate for the low activity of DNA ligase enzyme in the cold. The enzyme concentration is kept high and polyethylene glycol is added to reaction mixture for Stimulation.

Since ligation reconstructs the site of cleavage, recombinant DNA molecules produced by ligation of sticky ends can be cleaved again at the joints, using the same restriction endonuclease enzyme that was used to generate the fragment initially. As a result, a fragment can be inserted into a vector DNA, and recovered again after cloning of the recombinant molecules.

(ii) Ligation of DNA molecules with blunt ends:

Fragments of blunt-ended DNA can be ligated, but since there is no base-pairing to hold fragments together temporarily, concentrations of DNA and ligase enzyme must be high. However, blunt-end ligation is a useful way of joining together DNA fragments which have not been produced by the same restriction enzyme, and which therefore have mismatched sticky ends. These ends are removed prior to ligation, using the enzyme S_1 nuclease, which digests single-stranded DNA.

In case of ligation of blunt-ends, a restriction site will not be regenerated and this may prevent recovery of a fragment after cloning. For this reason, short DNA duplexes, called linkers, are frequently used for joining DNA.

Linkers are short, double-stranded oligonucleotides, with blunt ends, containing at least one restriction site (i.e., Eco RI palindrome) within their sequence. These linkers can be joined to one preparation of DNA by blunt- ended ligation and then sticky ends can be created by cleavage of the linkers with a suitable restriction enzyme.

The linker is chosen so that the sticky end it produces is identical to that on the other DNA preparation. Consequently, the two can then be joined by ligation of their sticky ends. Some very versatile linkers are available which contain restriction sites for several different enzymes within a sequence of only eight to ten nucleotides. Blunt- ended molecules can also be ligated after building sticky ends.

Thus, with this method it is now possible to insert a foreign DNA segment at a particular site in the linker region of the vector and then retrieve this foreign DNA segment whenever necessary.

Sources of DNA Ligases:

DNA ligases are isolated from E. coli and T_4 bacteriophage. The ligase enzyme isolated from E. coli is a polypeptide chain with a molecular weight of 75 kDa. It requires NAD^+ as cofactor. The ligase obtained from T_4 bacteriophage is 68 kDa. It requires ATP as a cofactor and a source of energy.

For the routine laboratory requirement, T_4 DNA ligase is obtained from an induced lysogen of lambda T_4 lig phage. This enzyme has the capacity to ligate a variety of cohesive and blunt-ended DNA fragments. The enzyme concentration is kept higher and a fusogen called polyethylene glycol, is added to reaction mixture for stimulation.

Application of DNA Ligase Enzymes:

DNA ligase enzymes play an important role in genetic engineering. In the absence of DNA ligase enzymes, recombinant DNA technology cannot be successful.

The important functions of DNA ligase enzymes are as follows:

1. Genetic engineering experiments involve joining of DNA fragments to produce recombinant DNA molecules. Ligase enzymes are used in the joining process.

2. Ligase enzymes help in ligation of vector and inserting recombinant DNA.

3. They help in ligation of linkers or adapter molecules at the blunt ends of DNA fragments.

4. They help in sealing nicks in double- stranded DNA.

5. Ligase enzymes requires 3′ OH and 5′ P0^4 group for ligation.

This requirement can be advantageous as:
(i) Self ligation of DNA can be prevented by dephosphorylation (using alkaline phosphatase) of intended donor fragments prior to ligation.

(ii) Dephosphorylation of vector DNA will prevent recircularization of the vector in cloning procedure.

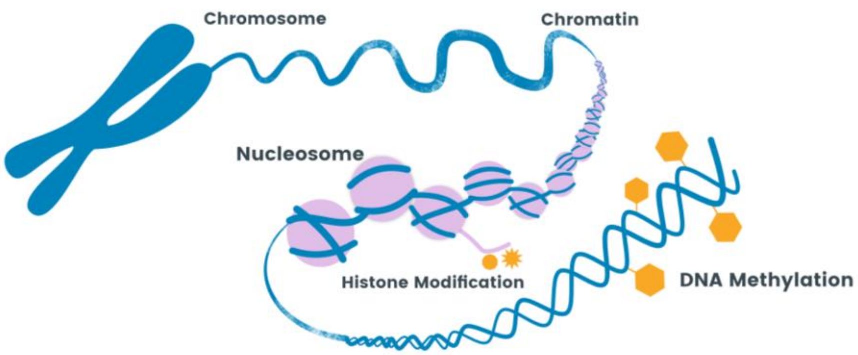
j. Methyl Transferase

- Methyltransferases are a class of enzymes that catalyze the transfer of a methyl group from the methyl donor S-adenosyl-l-methionine (SAM) to their substrates.

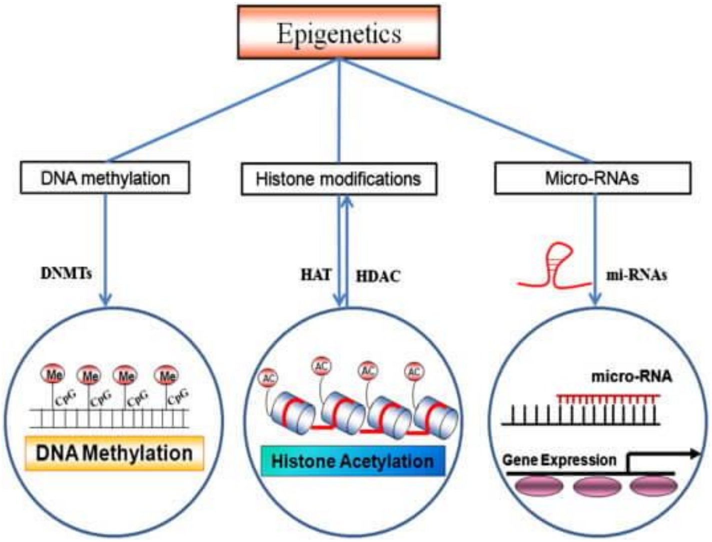
- The methyltransferases are an eclectic mix of enzymes of which the majority, over 95%, uses *S*-adenosyl-l-methionine (Ado-Met) as the methyl donor.

- DNA methyltransferase (cytosine-5′-methyltransferase, EC 2.1.1.37) plays an important role in controlling the profile of gene expression in mammalian cells.

- DNA methylation plays key roles in gene expression and regulation. **Regulation of DNA methylation by methyltransferases**



- DNA methylation involves the addition of a methyl group to the 5-carbon of the cytosine ring, which results in 5-methylcytosine or 5-mC. These methyl groups inhibit transcription by occupying the DNA's major groove. 5-mC constitutes about 1.5% of genomic DNA.



- DNA methylation is catalyzed by DNA methyltransferases (DNMTs) and is controlled at many different points in cellular processes. Three types of DNMTs, namely, DNMT1, DNMT3a, and DNMT3b are needed to establish and maintain DNA methylation patterns.

Significance of DNA methylation

- The role of DNA methylation in gene expression varies across different kingdoms of organisms. 5'—C—phosphate—G—3′ (CpG) methylation is distributed fairly globally in mammals, whereas among invertebrates, the methylation pattern is generally "mosaic," with heavily methylated DNA regions being interspersed with regions that are not methylated.

- The significance of 5-mC as a key epigenetic modification in gene expression is widely recognized. For instance, a decrease in global DNA methylation or DNA hypomethylation is likely to be the result of methyl deficiency caused by various environmental factors and it has been suggested as a molecular marker in many biological processes, including cancer.

III. Coupling Tools – Linkers and Adaptors

- DNA ligation is the process of joining two DNA molecules together, forming phosphodiester bonds. The enzyme called DNA ligase catalyzes this reaction. It is one of the critical steps in modern molecular biological fields such as recombinant DNA technology and DNA cloning.

- The ligation efficiency depends on the ends of DNA molecules to be ligated. There are two types of DNA ends as sticky ends and blunt ends. Ligation efficiency is high with sticky ends than with blunt ends. If the target DNA molecules have blunt ends, molecules called adaptors or linkers will be useful.

- Adaptors and linkers are chemically synthesized oligonucleotide molecules that help in DNA ligation. They have internal restriction sites as well. Adaptor has one sticky end and one blunt end, while linker has two blunt ends.

a. Linker

- Linker is a chemically synthesized oligonucleotide sequence that is double-stranded. Linker has two blunt ends. Linker is used to ligate DNA molecules that have blunt ends to vectors. It contains one or more internal restriction sites. These restriction sites work as recognition sites for restriction enzymes.

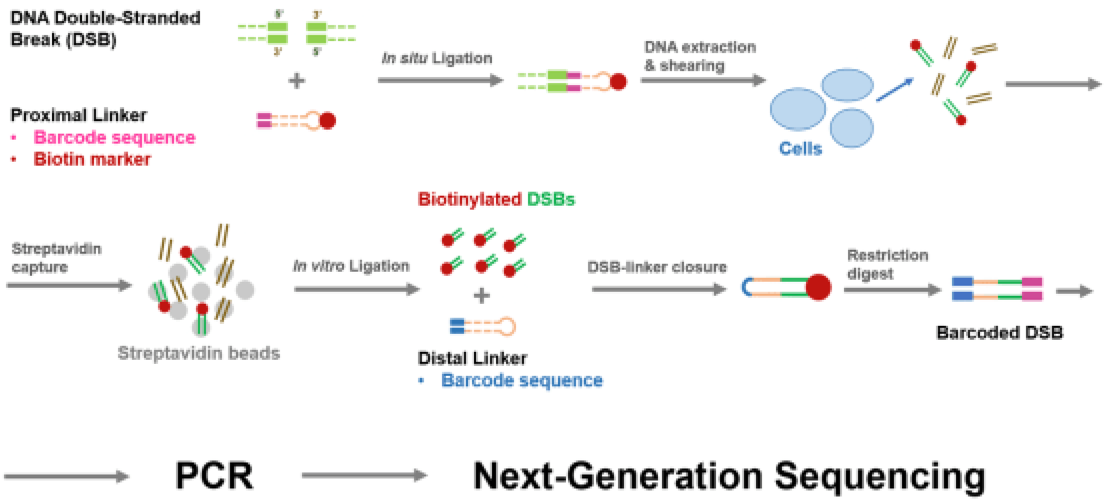


Figure 01: Linker

- After ligation, DNA is restricted again with restriction enzymes to produce cohesive ends. EcoRI-linkers and sal-I linkers are commonly used linkers.

b. Adaptor

- An Adaptor is a double-stranded oligonucleotide sequence used to link two DNA molecules together. It is a short sequence with one blunt end and one sticky or cohesive end. Therefore, it consists of a single-stranded tail at one end, which enhances the efficiency of DNA ligation.

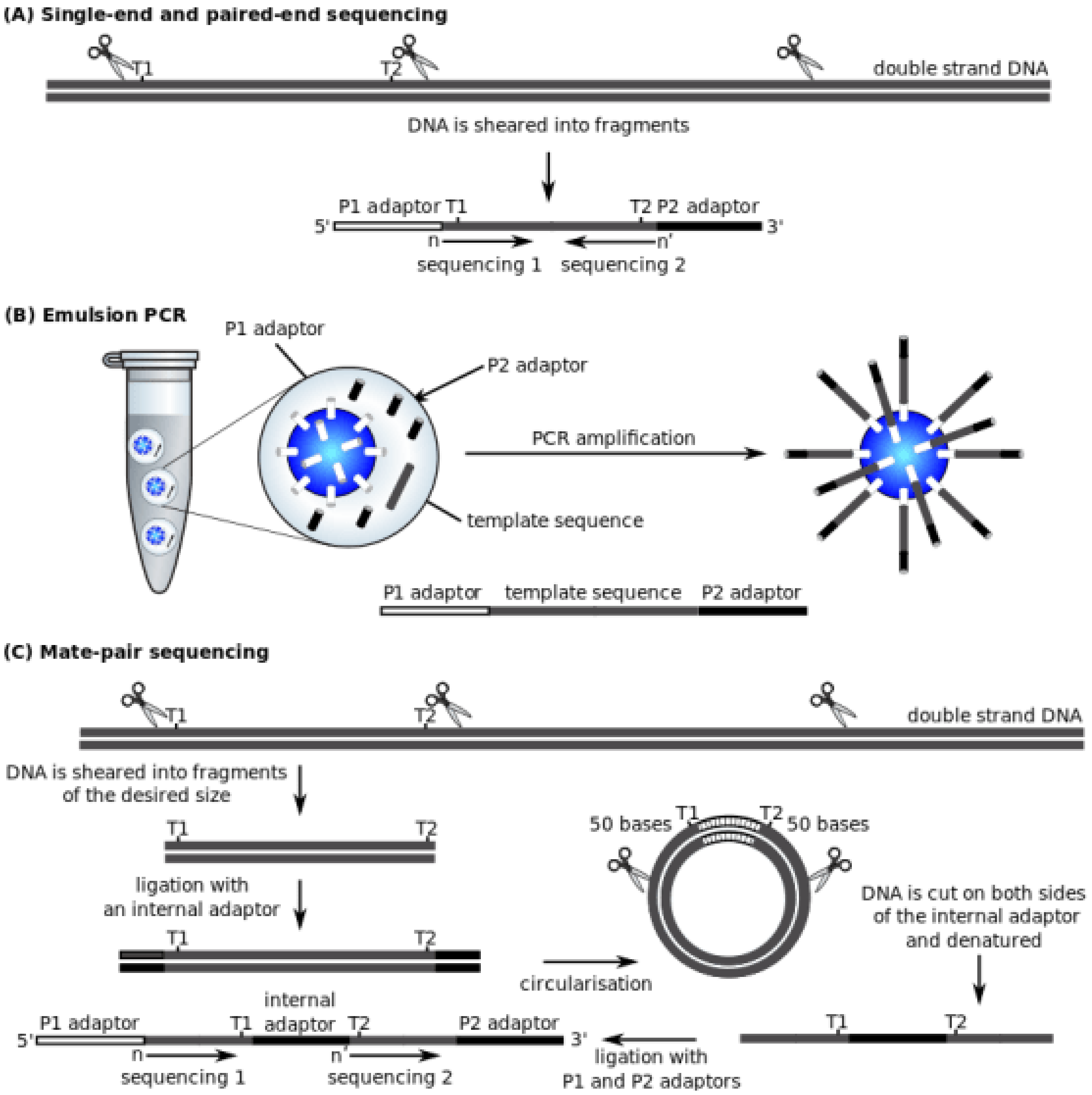


Figure 02: DNA Ligation by an Adaptor

- Moreover, the adaptor has internal restriction sites. Therefore, after ligation, DNA can be restricted with appropriate restriction enzymes in order to create a new protruding terminus. One disadvantage of adaptors is that two adaptors can form dimmers by base pairing with themselves. This can be avoided by treating them with the enzyme called alkaline phosphatase.

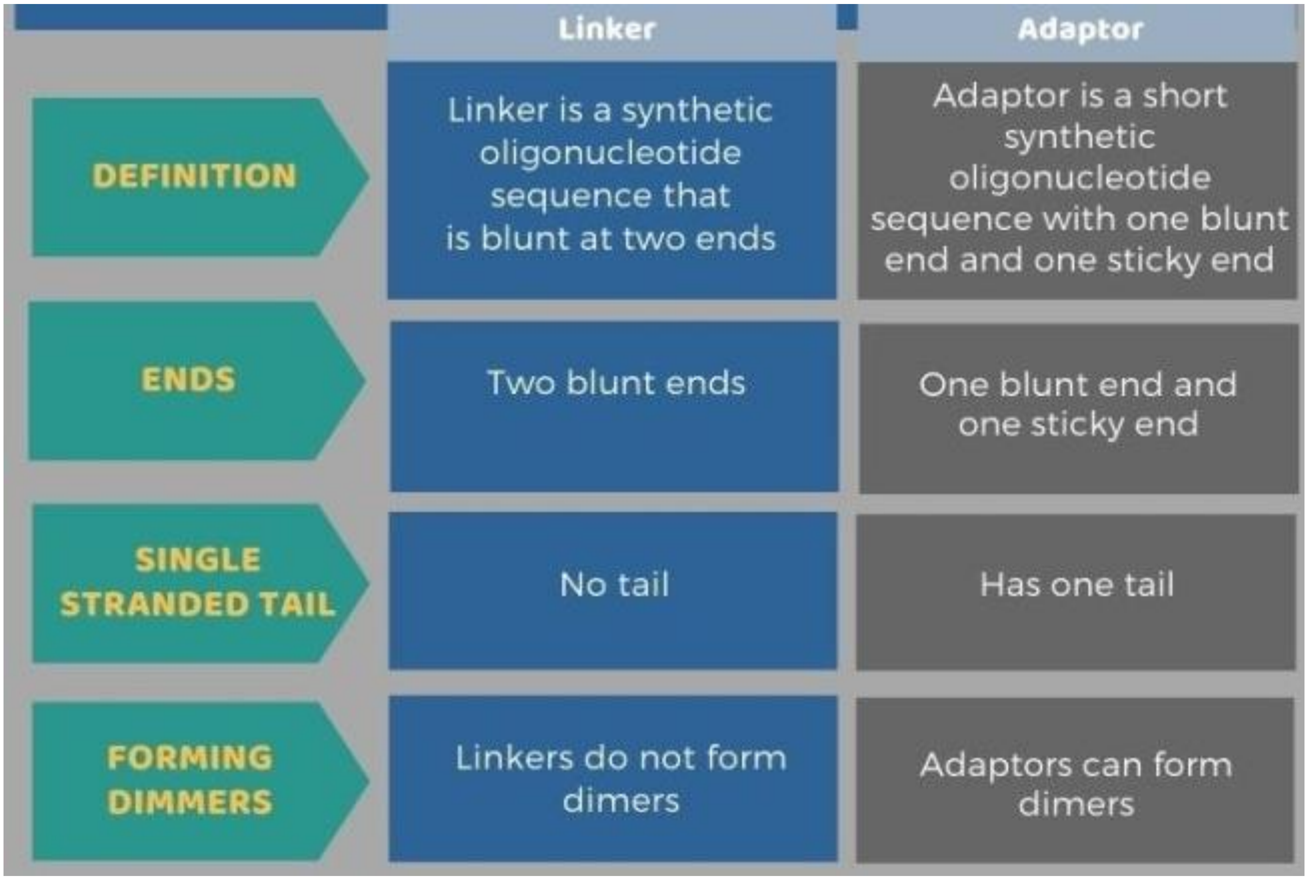
Similarities between Linker and Adaptor

- Both linker and adaptor are double-stranded short oligonucleotide sequences.
- They carry internal restriction sites.
- Moreover, they are chemically synthesized DNA molecules and are synthetic molecules.
- They can link two DNA molecules together.
- After ligation of linkers and adaptors, the DNA is again restricted with restriction enzymes in order to produce sticky ends.

Difference between Linker and Adaptor

A linker is a chemically synthesized short oligonucleotide duplex with two blunt ends. An adaptor is a chemically synthesized short oligonucleotide duplex with one sticky end and one blunt end. Thus, this is the key difference between linker and adaptor. Moreover, adaptors can form dimers, while linkers do not form dimers. So, this is another significant difference between linker and adaptor.

Below is a summary of the differences between linker and adaptor in tabular form.



	Linker	Adaptor
DEFINITION	Linker is a synthetic oligonucleotide sequence that is blunt at two ends	Adaptor is a short synthetic oligonucleotide sequence with one blunt end and one sticky end
ENDS	Two blunt ends	One blunt end and one sticky end
SINGLE STRANDED TAIL	No tail	Has one tail
FORMING DIMMERS	Linkers do not form dimers	Adaptors can form dimers

Summary – Linker vs Adaptor

- Linker and adaptor are two types of chemically synthesized oligonucleotides that are useful in ligating blunt-end DNA. Linker has two blunt ends, while adaptor has one blunt end and one cohesive end. So, this is the key difference between linker and adaptor. They are double-stranded molecules that have internal restriction sites. They are widely used in recombinant DNA technology and DNA cloning.

IV- Construction of Recombinant DNA Technology

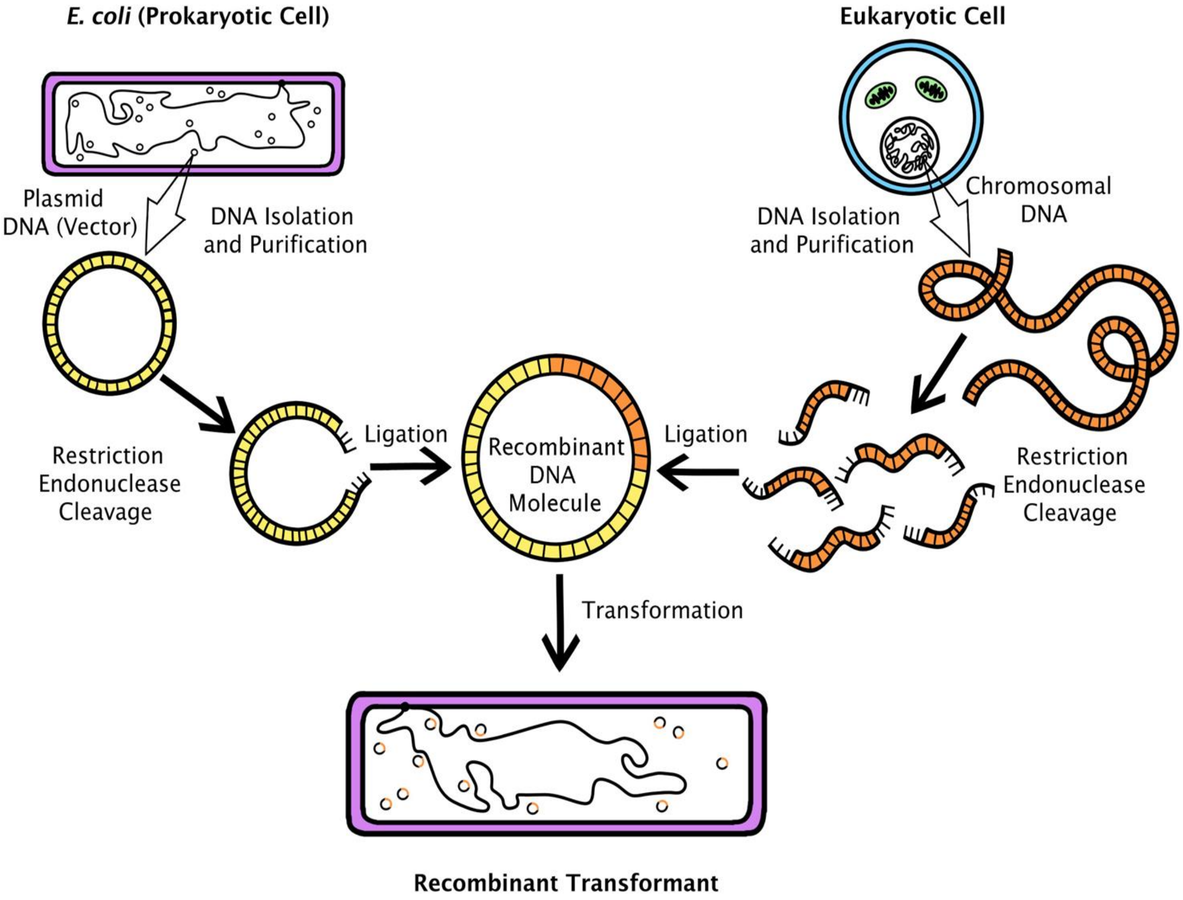
An outline of recombinant DNA technology:

- Recombinant **DNA** technology refers to the joining together of DNA molecules from two different species that are inserted into a host organism to produce new genetic combinations that are of value to science, medicine, agriculture, and industry.

- Recombinant DNA (rDNA), on the other hand is the general name for a piece of DNA that has been created by the combination of at least two strands.

- They are DNA molecules formed by laboratory methods of genetic recombination (such as molecular cloning) to bring together genetic material from multiple sources, creating sequences that would not otherwise be found in the genome.

- Recombinant DNA in a living organism was first achieved in 1973 by Herbert Boyer, of the University of California at San Francisco, and Stanley Cohen, at Stanford University, who used *E. coli* restriction enzymes to insert foreign DNA into plasmids.


E. coli (Prokaryotic Cell)
Eukaryotic Cell
Plasmid DNA (Vector)
DNA Isolation and Purification
DNA Isolation and Purification
Chromosomal DNA
Restriction Endonuclease Cleavage
Ligation
Recombinant DNA Molecule
Ligation
Restriction Endonuclease Cleavage
Transformation
Recombinant Transformant

The basic principle of recombinant DNA technology

Steps of Genetic Recombination Technology

There are many diverse and complex techniques involved in gene manipulation. However, the basic principles of recombinant DNA technology are reasonably simple, and broadly involve the following stages.

1. Generation of DNA fragments and selection of the desired piece of DNA (e.g. a human gene).

2. Insertion of the selected DNA into a cloning vector (e.g. a plasmid) to create a recombinant DNA or chimeric DNA (Chimera is a monster in Greek mythology that has a lion's head, a goat's body and a serpent's tail. This may be comparable to Narasimha in Indian mythology).

3. Introduction of the recombinant vectors into host cells (e.g. bacteria).

4. Multiplication and selection of clones containing the recombinant molecules.

5. Expression of the gene to produce the desired product.

1. **Isolation of Genetic Material**

- The first step in rDNA technology is to isolate the desired DNA in its pure form i.e. free from other macromolecules.
- Since DNA exists within the cell membrane along with other macromolecules such as RNA, polysaccharides, proteins, and lipids, it must be separated and purified which involves enzymes such as lysozymes, cellulase, chitinase, ribonuclease, proteases etc.
- Other macromolecules are removable with other enzymes or treatments. Ultimately, the addition of ethanol causes the DNA to precipitate out as fine threads. This is then spooled out to give purified DNA.

2. **Restriction Enzyme Digestion**

- Restriction enzymes act as molecular scissors that cut DNA at specific locations. These reactions are called 'restriction enzyme digestions'.
- They involve the incubation of the purified DNA with the selected restriction enzyme, at conditions optimal for that specific enzyme.
- The technique 'Agarose Gel Electrophoresis' reveals the progress of the restriction enzyme digestion.
- This technique involves running out the DNA on an agarose gel. On the application of current, the negatively charged DNA travels to the positive electrode and is separated out based on size. This allows separating and cutting out the digested DNA fragments.
- The vector DNA is also processed using the same procedure.

3. **Amplification Using PCR**

- Polymerase Chain Reaction or PCR is a method of making multiple copies of a DNA sequence using the enzyme – DNA polymerase in vitro.
- It helps to amplify a single copy or a few copies of DNA into thousands to millions of copies.

PCR reactions are run on ‘thermal cyclers’ using the following components:

1. Template – DNA to be amplified
2. Primers – small, chemically synthesized oligonucleotides that are complementary to a region of the DNA.
3. Enzyme – DNA polymerase
4. Nucleotides – needed to extend the primers by the enzyme.

- The cut fragments of DNA can be amplified using PCR and then ligated with the cut vector.

4. **Ligation of DNA Molecules**

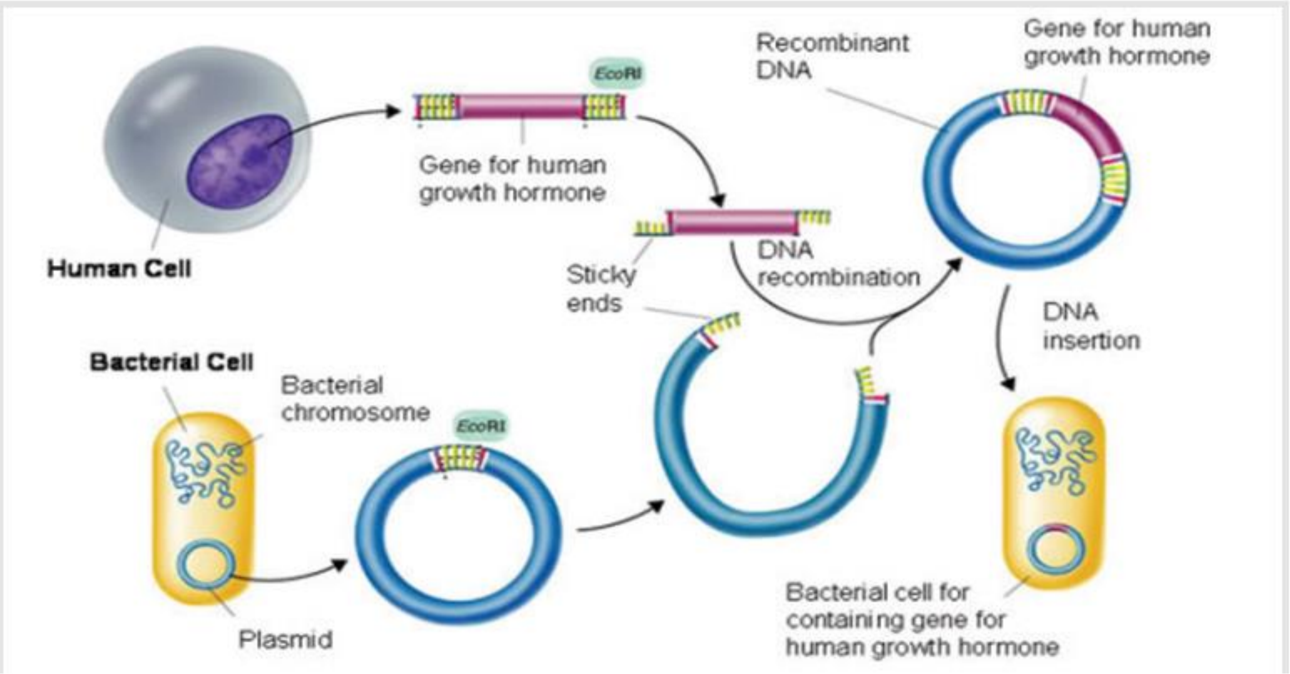
- The purified DNA and the vector of interest are cut with the same restriction enzyme.
- This gives us the cut fragment of DNA and the cut vector, that is now open.
- The process of joining these two pieces together using the enzyme ‘DNA ligase’ is ‘ligation’.
- The resulting DNA molecule is a hybrid of two DNA molecules – the interest molecule and the vector. In the terminology of genetics this intermixing of different DNA strands is called recombination.
- Hence, this new hybrid DNA molecule is also called a recombinant DNA molecule and the technology is referred to as the **recombinant DNA technology**.

5. **Insertion of Recombinant DNA Into Host**

- In this step, the recombinant DNA is introduced into a recipient host cell mostly, a bacterial cell. This process is ‘Transformation’.
- Bacterial cells do not accept foreign DNA easily. Therefore, they are treated to make them ‘competent’ to accept new DNA. The processes used may be thermal shock, Ca^{++} ion treatment, electroporation etc.

6. **Isolation of Recombinant Cells**

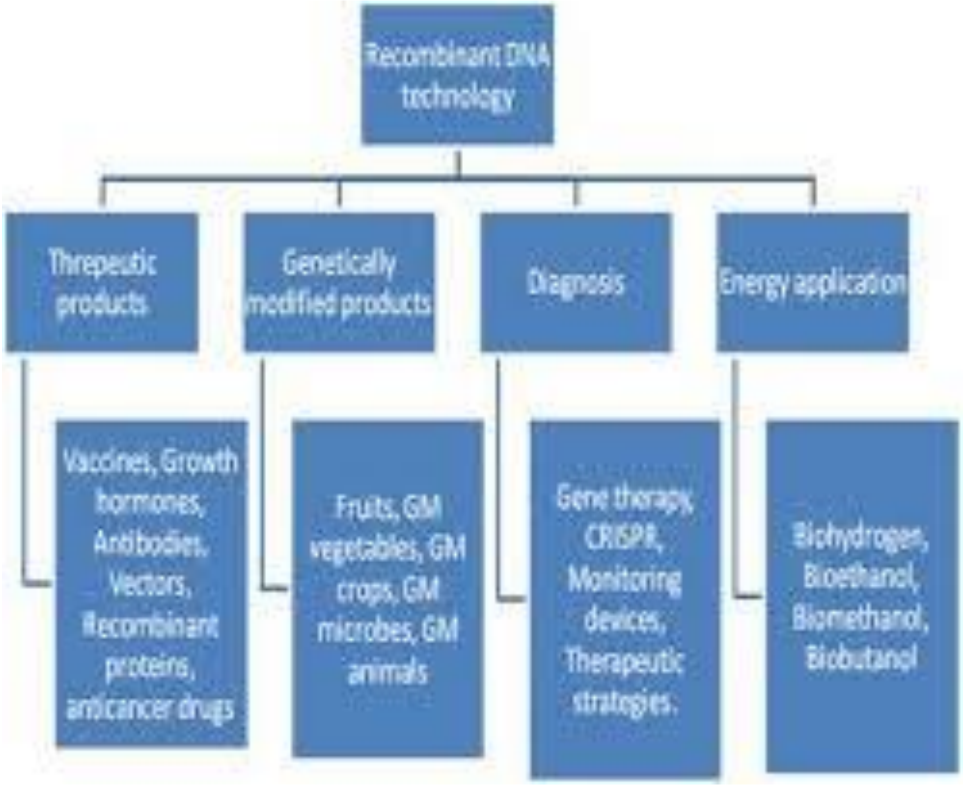
- The transformation process generates a mixed population of transformed and non-transformed host cells.
- The selection process involves filtering the transformed host cells only.
- For isolation of recombinant cell from non-recombinant cell, marker gene of plasmid vector is employed.
- For examples, PBR322 plasmid vector contains different marker gene (Ampicillin resistant gene and Tetracycline resistant gene. When pst1 RE is used it knock out Ampicillin resistant gene from the plasmid, so that the recombinant cell become sensitive to Ampicillin.


Human Cell
Gene for human growth hormone
EcoRI
Recombinant DNA
Gene for human growth hormone
Sticky ends
DNA recombination
DNA insertion
Bacterial Cell
Bacterial chromosome
EcoRI
Plasmid
Bacterial cell for containing gene for human growth hormone

Application of Recombinant DNA technology

- Recombinant DNA is widely used in biotechnology, medicine and research.
- The most common application of recombinant DNA is in basic research, in which the technology is important to most current work in the biological and biomedical sciences.
- Recombinant DNA is used to identify, map and sequence genes, and to determine their function.
- Recombinant proteins are widely used as reagents in laboratory experiments and to generate antibody probes for examining protein synthesis within cells and organisms.
- Many additional practical applications of recombinant DNA are found in industry, food production, human and veterinary medicine, agriculture, and bioengineering.

1. DNA technology is also used to detect the presence of HIV in a person.
2. Application of recombinant DNA technology in Agriculture – For example, manufacture of Bt-Cotton to protect the plant against ball worms.
3. Application of medicines – Insulin production by DNA recombinant technology is a classic example.
4. Gene Therapy – It is used as an attempt to correct the gene defects which give rise to heredity diseases.
5. Clinical diagnosis – ELISA is an example where the application of recombinant DNA is possible.


Recombinant DNA technology
Threpeutic products
Genetically modified products
Diagnosis
Energy application
Vaccines, Growth hormones, Antibodies, Vectors, Recombinant proteins, anticancer drugs
Fruits, GM vegetables, GM crops, GM microbes, GM animals
Gene therapy, CRISPR, Monitoring devices, Therapeutic strategies.
Biohydrogen, Bioethanol, Biomethanol, Biobutanol

Limitations of Recombinant DNA technology

- Destruction of native species in the environment the genetically modified species are introduced in.
- Resilient plants can theoretically give rise to resilient weeds which can be difficult to control.
- Cross contamination and migration of proprietary DNA between organisms.
- Recombinant organisms contaminating the natural environment.
- The recombinant organisms are population of clones, vulnerable in exact same ways. A single disease or pest can wipe out the entire population quickly.
- Creation of superbug is hypothesized.
- Ethical concern about humans trying to play God and mess with the nature's way of selection. It is exaggerated by the fear of unknown of what all can be created using the technology and how is it going to impact the civilization.
- Such a system might lead to people having their genetic information stolen and used without permission.
- Many people worry about the safety of modifying food and medicines using recombinant DNA technology.

****

References:

1. Lodish H, Berk A, Zipursky SL et al. (2000) Section 7.2, Constructing DNA Libraries with λ Phage and Other Cloning Vectors. *Molecular Cell Biology*. 4th edition. New York: W. H. Freeman.

2. Brown TA (2002) Chapter 6, Sequencing Genomes. *Genomes*. 2nd edition. Oxford: Wiley-Liss.

3. Lander ES, Linton LM, Birren B et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

4. Olson MV (2001) The maps. Clone by clone by clone. *Nature* 409(6822):816–818.

5. Venter JC, Adams MD, Myers EW et al. (2001) The sequence of the human genome. *Science* 291(5507):1304–1351.

6. https://nptel.ac.in/courses/102103017/pdf/lecture%2035.pdf

7. https://www.nature.com/subjects/genetic-vectors

8. https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/vector-molecular-biology

9. https://www.britannica.com/science/recombinant-DNA-technology#ref964476

10. https://en.wikipedia.org/wiki/Vector_(molecular_biology)

11. https://rajusbiology.com/cloning-vector-definition-features-and-types/

12. https://microbiologynotes.org/vector-properties-types-and-characteristics/

Assessment:

Brief the following:

1. Role of ribonuclease H.

2. Types of Restriction Enzymes.

3. Types of DNA ligase.

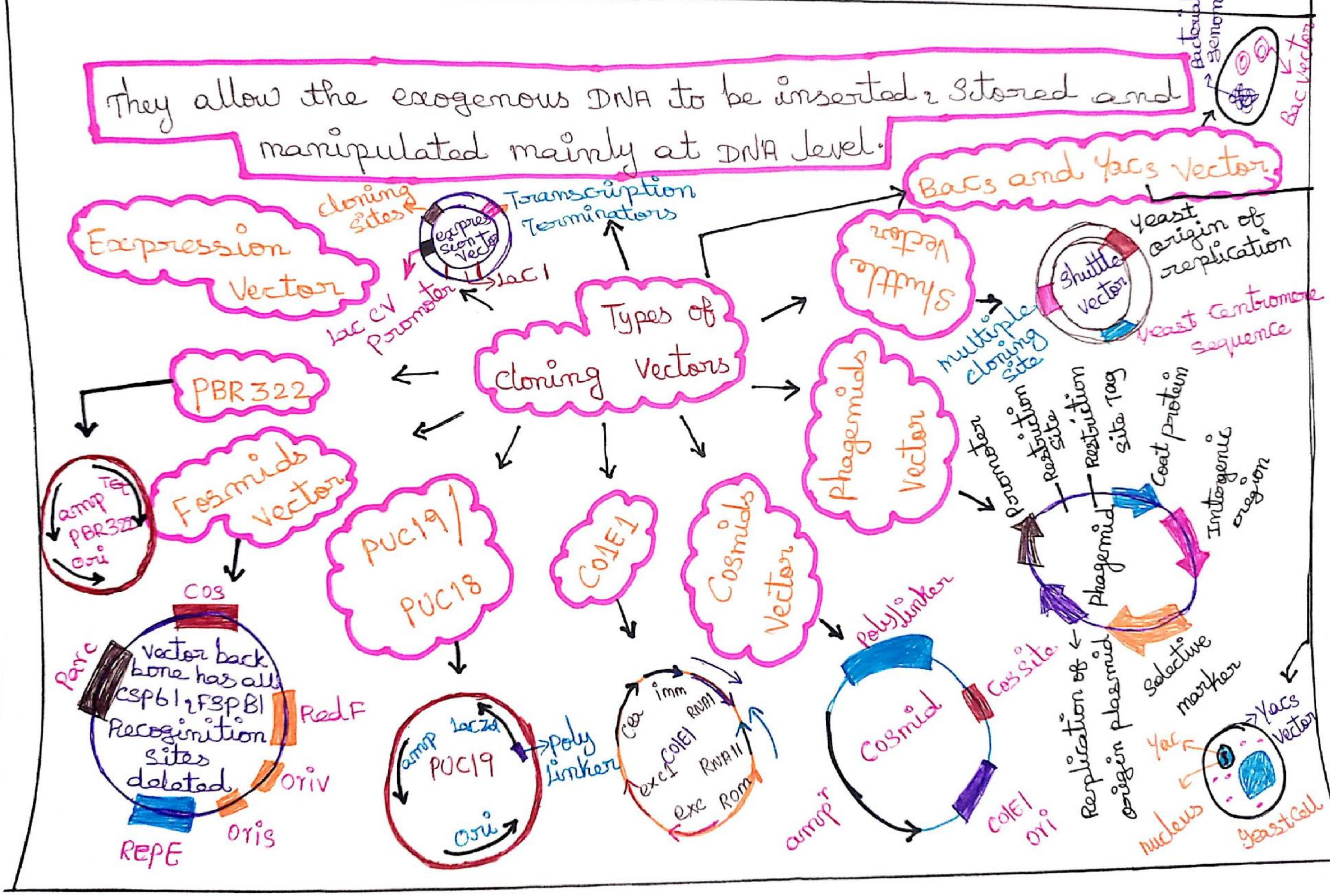
Detail the following:

4. Application of rDNA.

5. History of rDNA

Unit II

Unit-2



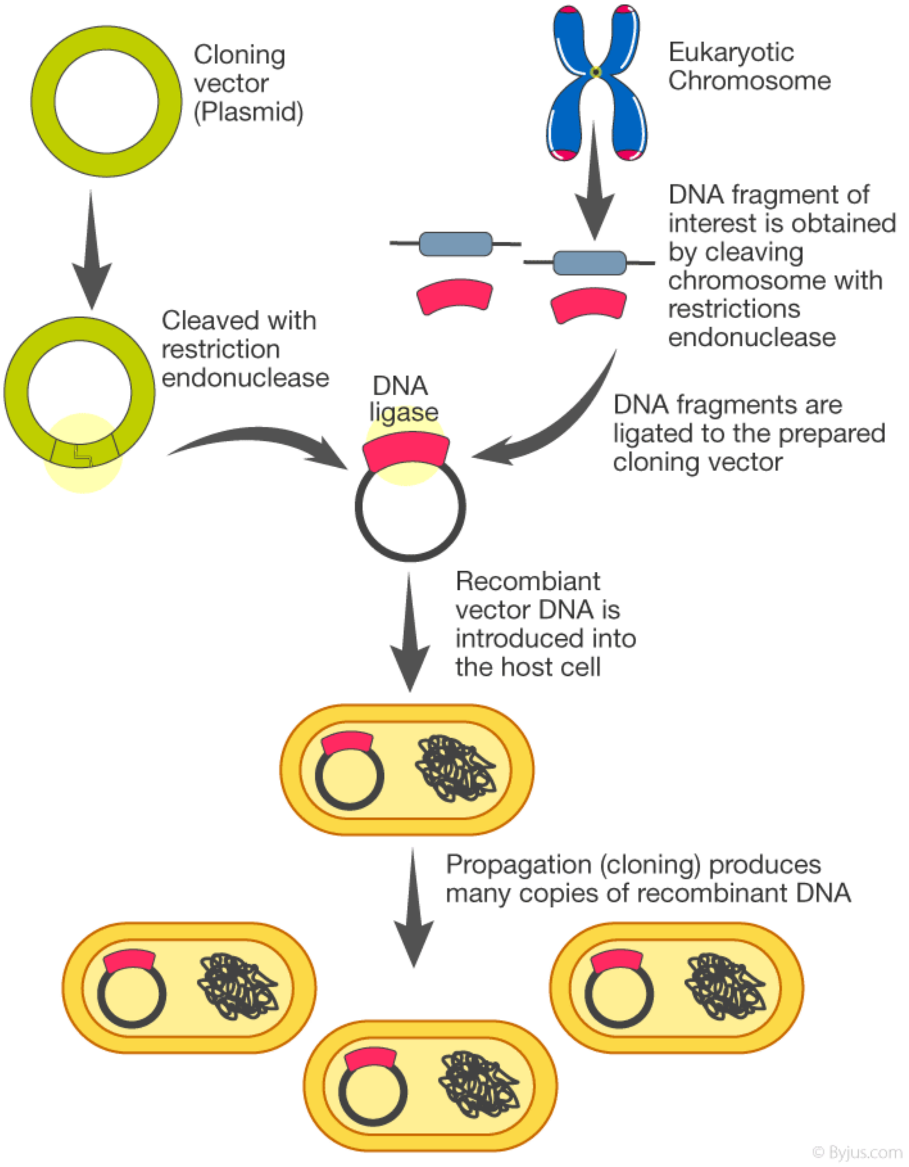
They allow the exogenous DNA to be inserted, Stored and manipulated mainly at DNA level.

Gene cloning: Strategies in gene cloning:

- "Gene cloning is a molecular biology technique which is used for the creation of exact copies or clones of a particular gene or DNA."

- Gene cloning is the process of making multiple copies of a particular segment of DNA. During this technique, the selected DNA fragment is inserted into a plasmid (the circular piece of DNA) using enzymes. Restriction enzymes and DNA ligase are used in the process.

- The restriction enzymes are used to cut the DNA fragments at specific sequences and DNA ligase enzymes are used to join the nicks. The recombinant DNA thus produced is introduced into bacteria. These bacteria reproduce and produce an exact copy of the plasmid. These copies are known as clones.


Cloning vector (Plasmid)
Eukaryotic Chromosome
DNA fragment of interest is obtained by cleaving chromosome with restrictions endonuclease
Cleaved with restriction endonuclease
DNA ligase
DNA fragments are ligated to the prepared cloning vector
Recombiant vector DNA is introduced into the host cell
Propagation (cloning) produces many copies of recombinant DNA
© Byjus.com

Gene Cloning

DNA Cloning Steps

DNA Cloning takes place in the following steps:
Cutting and Pasting DNA
Two types of enzymes are used in this method:

- Restriction enzymes
- DNA ligase

The restriction enzymes cut the DNA at specific target sequences. The target gene is inserted into the cut site and is ligated by DNA ligase. This is known as a recombinant plasmid.

Bacterial Transformation and Selection

The recombinant plasmid is introduced into bacteria such as E.coli. The bacteria are subjected to very high temperatures which compel them to take up the DNA. This process is known as transformation. The plasmid contains an antibiotic resistance gene which helps them to survive in the presence of antibiotics. The plasmid containing bacteria are selected on a nutrient containing antibiotics. The transformed bacteria survive, while the ones without a plasmid die.

Protein Production

The plasmid containing the bacteria are cultured and the bacteria are provided with a chemical signal that helps them to target the protein. After protein production, the bacteria are split open to release it. The protein is purified and the target protein is isolated from other contents of the cell.

Importance of DNA Cloning

The DNA molecules produced through the cloning techniques are used for many purposes which include:

1. DNA cloning can be used to make proteins such as insulin with biomedical techniques.
2. It is used to develop recombinant versions of the non-functional gene to understand the functioning of the normal gene. This is applied in gene therapies also.
3. It helps to analyse the effect of mutation on a particular gene.

Cloning applications and methodologies

Cloning methods rely on molecular biological processes that occur in nature. The techniques are continually being refined and simplified; therefore, many strategies nowadays permit cloning of sequences of interest from their sources more efficiently. These cloning strategies include:

- PCR cloning strategies
- Subcloning basics
- Library construction essentials
- Shotgun cloning and sequencing method

PCR cloning strategies

- PCR cloning is a method in which double-stranded DNA fragments amplified by PCR are ligated directly into a vector. PCR cloning offers some advantages over traditional cloning which relies on digesting double-stranded DNA inserts with restriction enzymes to create compatible ends, purifying and isolating sufficient amounts, and ligating into a similarly treated vector of choice (see insert preparation).

- With PCR amplification, this cloning technique requires much less starting template materials which include cDNA, genomic DNA, or another insert-carrying plasmid (see subcloning basics). Furthermore, PCR cloning provides a simpler workflow by circumventing the requirement of suitably-located restriction sites and their compatibility between the vector and insert. Nevertheless, there are a number of considerations related to: PCR primers and amplification conditions, the cloning method of choice and the cloning vectors used, and, finally, confirmation of successful cloning and transformation.

- With respect to PCR amplification of a sequence of interest, primers must be designed and PCR conditions (components and cycling) optimized for efficient and specific amplification of the template. Primer design tools are available to bioinformatically evaluate and select suitable target-specific primer sequences for amplification. Ligation requires that either the insert or vector has 5′-phosphorylated termini; therefore, if the cloning vector lacks 5′-phosphorylated ends, 5′-phosphate groups must be added to the PCR primers during synthesis or by T4 polynucleotide kinase for successful ligation. For PCR optimization, reaction component concentrations, annealing temperatures, and template amounts are of importance.

- TA cloning and blunt-end cloning represent two of the simplest PCR cloning methods. Their choice depends upon the nature of the vector and the type of PCR enzymes used in cloning. TA cloning employs a thermostable Taq DNA polymerase capable of amplifying short DNA sequences. This enzyme lacks 3′→ 5′ proofreading activity and features a terminal transferase activity that adds an extra deoxyadenine at the 3′ end of the amplicons (3′ dA). The resulting PCR products with 3′ dA overhangs are readily cloned into a linearized TA cloning vector containing complementary 3′ deoxythymine (3′ dT) overhangs (Figure 1). While relatively straightforward, the limitations of this method include the length of insert (up to 5 kb), the inability to clone inserts directionally, and the high error rate associated with Taq DNA polymerase.

- Blunt-end cloning involves the ligation of an insert into a linearized vector where both DNA fragments lack overhangs. Blunt-end inserts can be produced using high-fidelity DNA polymerases with 3′→5′ exonuclease or proofreading activity. Their proofreading activity improves the sequence accuracy of the amplified products; however, limitations include lower ligation efficiencies when inserting into blunt-end cloning vectors and the inability to clone directionally. Ligation efficiency can be improved by incubating the amplicons with a Taq DNA polymerase and dATP in a procedure called “3′ dA tailing” (incubate 20–30 minutes at 72°C), then purifying the 3′ dA-tailed products (Figure 1).

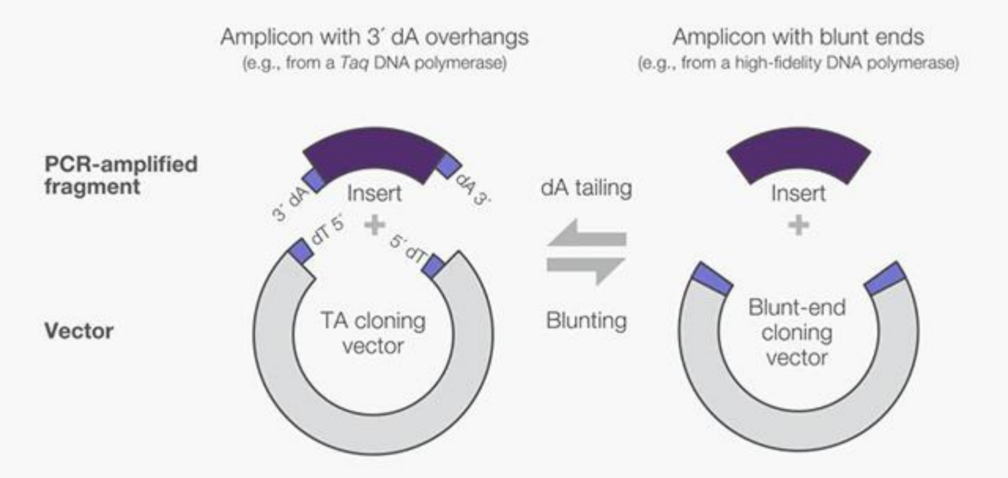


Figure 1. Common PCR cloning strategies

- To further simplify and streamline the cloning workflow, specialized vectors have been developed to place an insert into vector, for example, without using a ligase. One such class of vectors includes the Invitrogen TOPO cloning vectors which contain covalently linked DNA topoisomerase I that functions as both a restriction enzyme and a ligase (learn more about TOPO cloning technology). Compared to conventional PCR cloning vectors, these vectors result in shorter ligation reaction times (e.g., 5 minutes) and greater cloning efficiencies (e.g., >95% positive clones) and with a much simpler protocol. Furthermore, directional cloning of the PCR products can be achieved with a specially designed TOPO vector using a specific primer design.

- Regardless of the cloning method choice, cloning efficiencies are significantly improved by purification of PCR amplicons prior to the ligation reaction. PCR clean-up helps remove salts, nucleotides, nonspecific amplicons, and primer-dimers. After ligation and transformation into the appropriate competent cells, the resulting colonies need to be screened carefully for the correct insert, as well as its proper frame and orientation for subsequent studies to analyze gene fusions and/or protein expression.

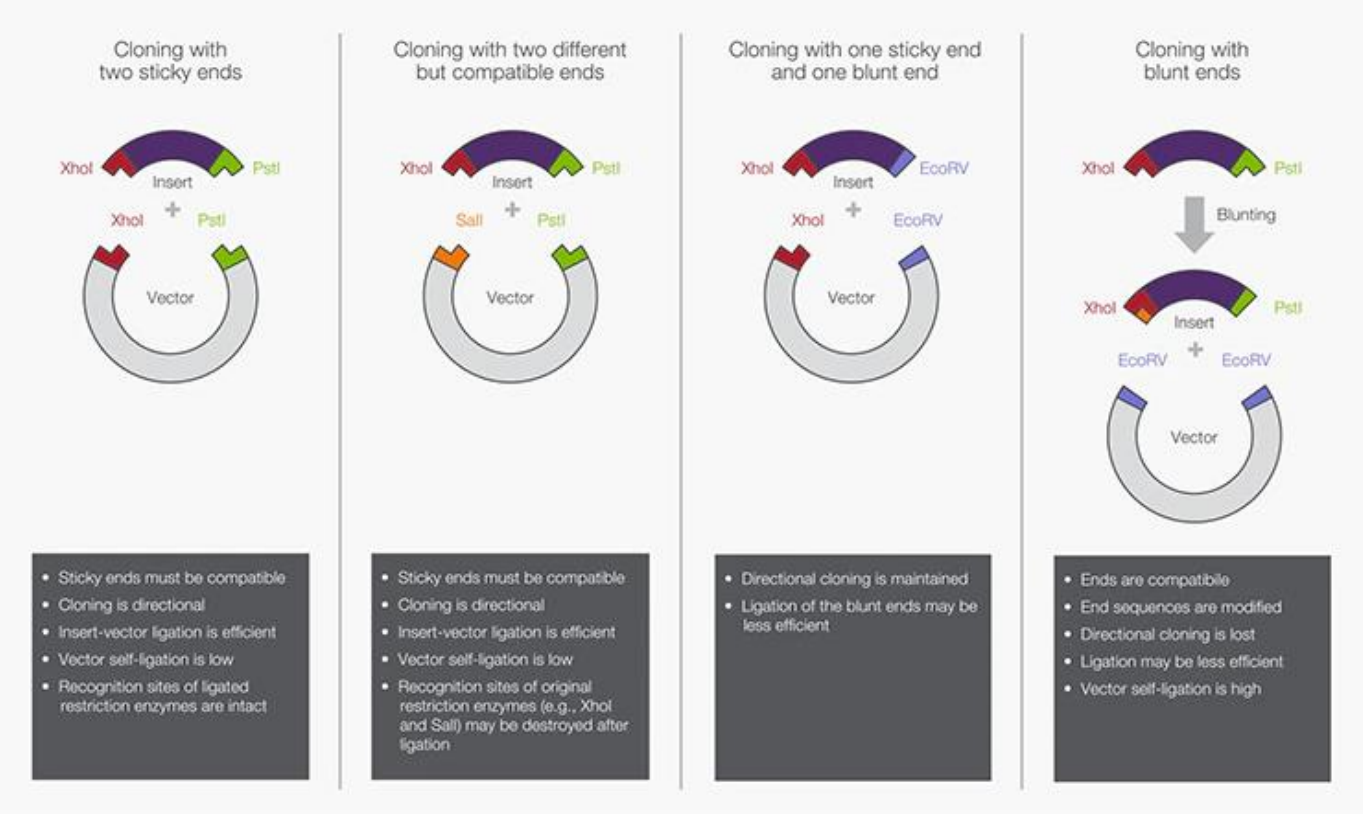
- To further simplify and streamline the cloning workflow, specialized vectors have been developed to place an insert into vector, for example, without using a ligase. One such class of vectors includes the Invitrogen TOPO cloning vectors which contain covalently linked DNA topoisomerase I that functions as both a restriction enzyme and a ligase (learn more about TOPO cloning technology). Compared to conventional PCR cloning vectors, these vectors result in shorter ligation reaction times (e.g., 5 minutes) and greater cloning efficiencies (e.g., >95% positive clones) and with a much simpler protocol. Furthermore, directional cloning of the PCR products can be achieved with a specially designed TOPO vector using a specific primer design.

- Regardless of the cloning method choice, cloning efficiencies are significantly improved by purification of PCR amplicons prior to the ligation reaction. PCR clean-up helps remove salts, nucleotides, nonspecific amplicons, and primer-dimers. After ligation and transformation into the appropriate competent cells, the resulting colonies need to be screened carefully for the correct insert, as well as its proper frame and orientation for subsequent studies to analyze gene fusions and/or protein expression.

Sub-cloning basics

- Subcloning refers to moving one fragment of a plasmid into another plasmid that can serve as a vector. There are a variety of reasons why it is necessary to transfer the fragment of interest into a different vector backbone. For instance, the new vector may possess a specific marker for antibiotic selection or fluorescent expression. Subcloning may also be performed to move a cloned fragment to an expression vector of a more suitable host for the study (e.g., bacteria, mammals, insects, plants, etc.); to place the gene of interest under a different expression promoter (e.g., a constitutive to inducible promoter); or to tag or fuse the experimental gene with another protein or a marker. Whatever the goal of the experiment may be, the two most common approaches to subcloning rely on restriction digestion and/or PCR cloning.

- Subcloning by restriction digestion is the more traditional of the two methods. In this workflow, fragments from the vector and the insert are double-digested with two restriction enzymes that generate sticky or cohesive ends (Figure 2). Since the vector and the insert can ligate in only one orientation (i.e., directional) and the digested ends of the vector are incompatible for self-ligation, this is arguably the preferred and most common method among other possible restriction enzyme options (see insert preparation for some options). For subcloning in protein expression or gene regulation studies, the selected restriction enzyme(s) should allow in-frame cloning of the fragment of interest with close proximity to the start codon as appropriate.



- Sticky ends must be compatible
- Cloning is directional
- Insert-vector ligation is efficient
- Vector self-ligation is low
- Recognition sites of ligated restriction enzymes are intact

- Sticky ends must be compatible
- Cloning is directional
- Insert-vector ligation is efficient
- Vector self-ligation is low
- Recognition sites of original restriction enzymes (e.g., XhoI and SalI) may be destroyed after ligation

- Directional cloning is maintained
- Ligation of the blunt ends may be less efficient

- Ends are compatibile
- End sequences are modified
- Directional cloning is lost
- Ligation may be less efficient
- Vector self-ligation is high

Figure 2. Subcloning restriction digest strategies.

- A second popular approach uses PCR to amplify the region of interest from the plasmid. The resulting PCR product is then cloned into the desired vector. TA cloning or blunt-end cloning methods can be used as described in the PCR cloning section, but neither approach maintains directionality of the insert. To achieve directional cloning, restriction sites that are present in the destination vector for subcloning can be incorporated into PCR amplicons by using PCR primers designed with the restriction sites in the 5′ end of the PCR primers. Following the PCR reaction, PCR products are restriction digested, purified, and subcloned into the restriction sites of the vector.

- There are a few considerations when designing the PCR primers with restriction enzymes sites. It is imperative that the introduced restrictions sites are unique and not present within the sequence of the fragment to be subcloned. The restriction sites should also be carefully designed to allow in-frame expression of the subcloned DNA. The cleavage efficiency of most restriction enzymes is greatly reduced when their recognition sites are close to termini of linear DNAs. To ensure proper digestion of the PCR fragments, a sequence with an extra 4–8 nucleotides (sometimes called “leader” or “spacer” sequence) is recommended at the 5′ end of the restriction sites on the primers **(Figure 3)**. Although there is no consensus on the optimal spacer sequence, a general recommendation is to avoid sequences that may result in primer-dimers or secondary structure formation (e.g., palindromes and inverted repeats). Furthermore, the primer recognition sequence design should be longer than those of the restriction site and the spacer combined to ensure specificity and proper binding to the target. When calculating the T_m of the primers, only sequences that are perfect matches to the template should be included. Finally, purification of the primers may be necessary to ensure full-length DNA oligonucletoides when using long primer sequences.

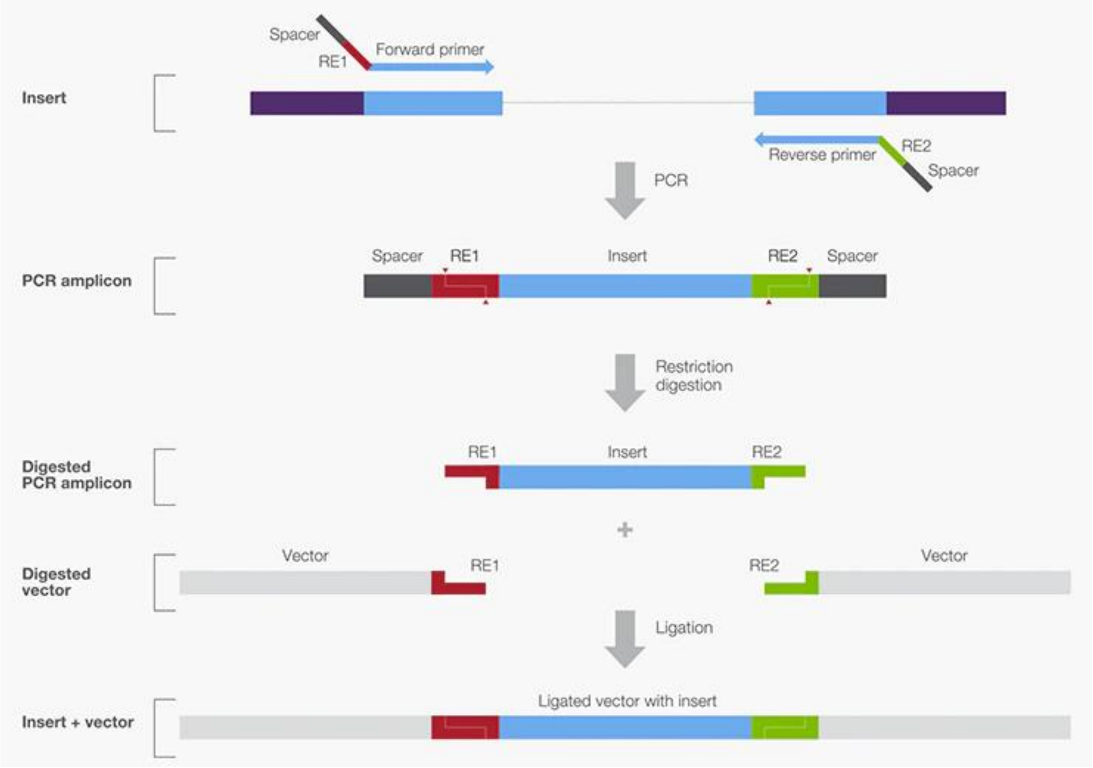


Figure 3. Schematic workflow of PCR subcloning in combination with restriction digestion (RE = restriction enzyme site).

- Other subcloning strategies have been devised to take advantage of special vectors that do not require the use of restriction enzymes or a ligase. One such example is Invitrogen Gateway cloning, which exploits unique recombination activities of the family of Invitrogen Clonase enzymes **(Figure 4)**. This method involves use of specially designed Gateway-specific plasmids and Gateway-compatible insert ends (*att* sites) for recombination.

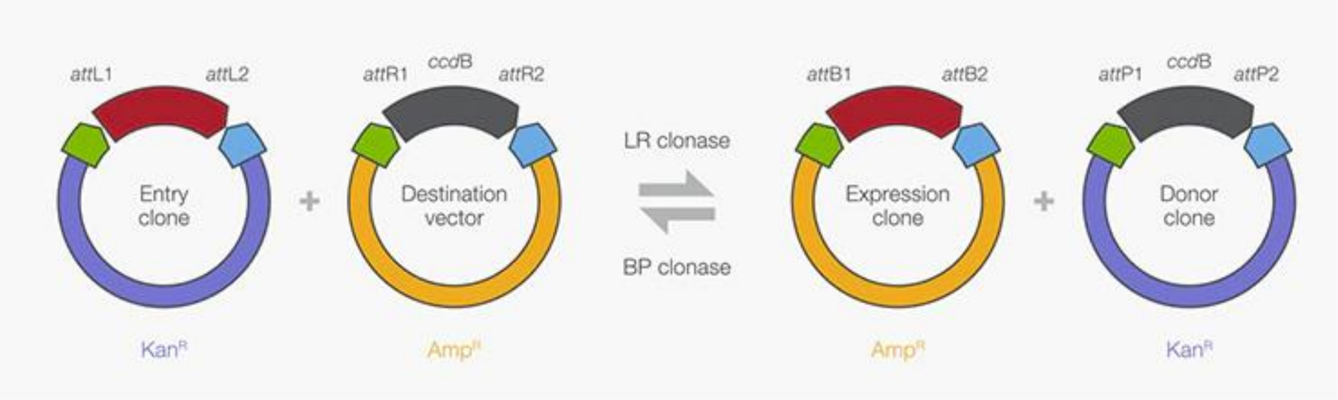


Figure 4. Gateway cloning strategies. ccdB is a toxic gene used in bacterial cell selection.

Library construction essentials

- In molecular cloning, DNA library construction refers to the creation of clones that carry DNA fragments representing the complete genomic DNA (gDNA) of a species, or the complementary DNA (cDNA) of RNA transcripts representing the expressed genome. By constructing DNA libraries, thousands of genetic fragments can be conveniently archived and expanded for downstream applications, such as genotyping and phenotypic screening. gDNA libraries serve as helpful tools to study the genetic composition of different species or gene mutations that occur in diseases such as cancer. cDNA libraries, on the other hand, are useful for expression analyses of genes and transcript variants based on the cell type and tissue origins (spatial), as well as time points (temporal).

- The construction of gDNA and cDNA libraries shares many similarities but also some important differences. Both strategies include nucleic acid purification, sample preparation (e.g., restriction digestion), vector cloning, vector introduction into a suitable host (e.g., transformation or transduction), and clonal selection. As the starting materials are different between the gDNA library and the cDNA library, their purification and preparation employ different approaches; however, once the gDNA or cDNA fragments are cloned into the desired vector, the same workflow may be followed.

- For genomic library preparation, gDNA is purified from the organism, tissues, or cells of interest. Extracted gDNA is then digested, isolated, and ligated into the vector of interest with compatible ends. Partial digestion of the genome is often carried out with a restriction enzyme with prevalent cutting sites to allow sequence overlaps between fragments for mapping of the cloned inserts **(Figure 5)**.

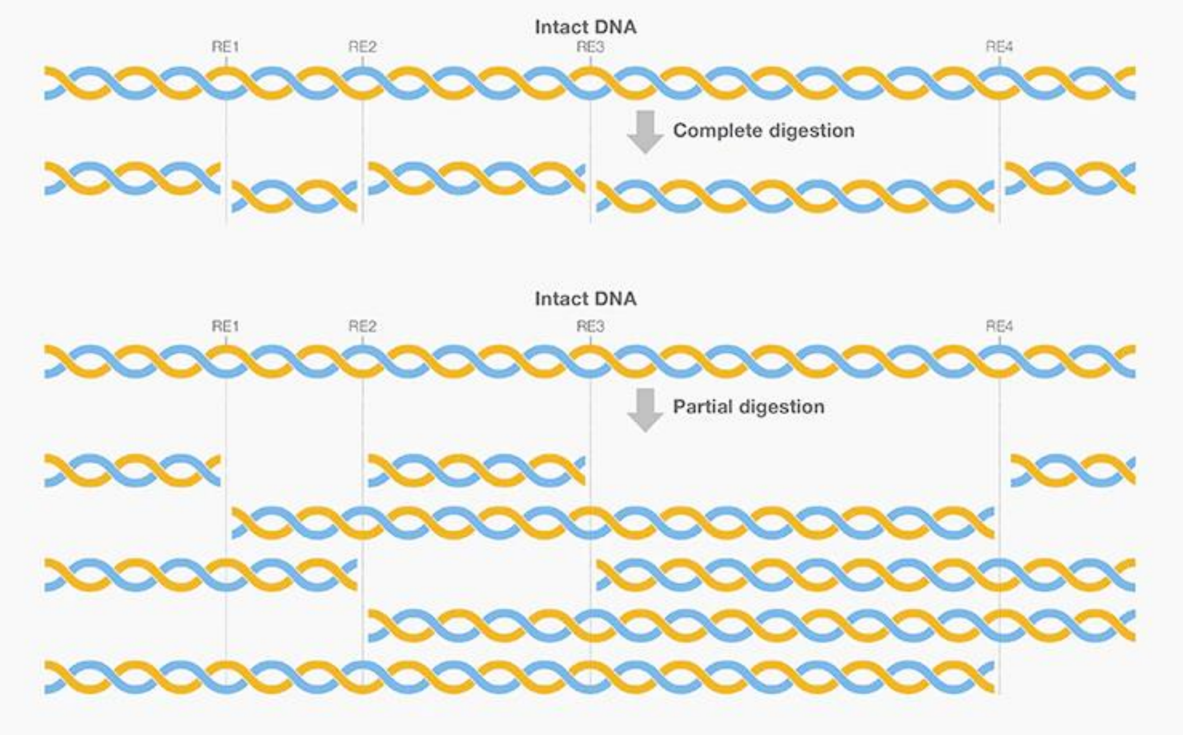


Figure 5. Schematic diagram of complete vs. partial digestion of a fragment by a restriction enzyme with four cutting sites.

Partial digestion results in overlapping sequences among fragments for mapping. (Only some possible partially-digested fragments are shown here for simplicity.)

- Vector selection for gDNA libraries is an important consideration because the gene fragments used in the library constructions are often large (e.g., >20 kb). The choice of cloning vector, in turn, determines the method to deliver insert-carrying vectors into the host **(Table 1)**.

Table 1. Common vector types, cloned fragment lengths, and vector delivery methods in library construction.

Vector type	**Cloned DNA (kb)**	**Vector delivery method**
Plasmid	20	Transformation
λ phage	25	Transduction
Cosmid	45	Transduction
P1 phage	100	Transduction
BAC (bacterial artificial chromosome)	300	Electroporation
YAC (yeast artificial chromosome)	1,000	Transformation (yeast)

- Ligation products or recombinant DNA can be introduced directly into bacterial cells via transformation or packaged into bacteriophage for infection or “transduction” of the host cells **(Figure 6)**. The transformed or transduced cells are intended for subsequent archiving, expansion, and sequencing in downstream experiments. Whole-genome sequences of many organisms, including the first whole human genome sequence, were determined using this basic strategy in early 2000.

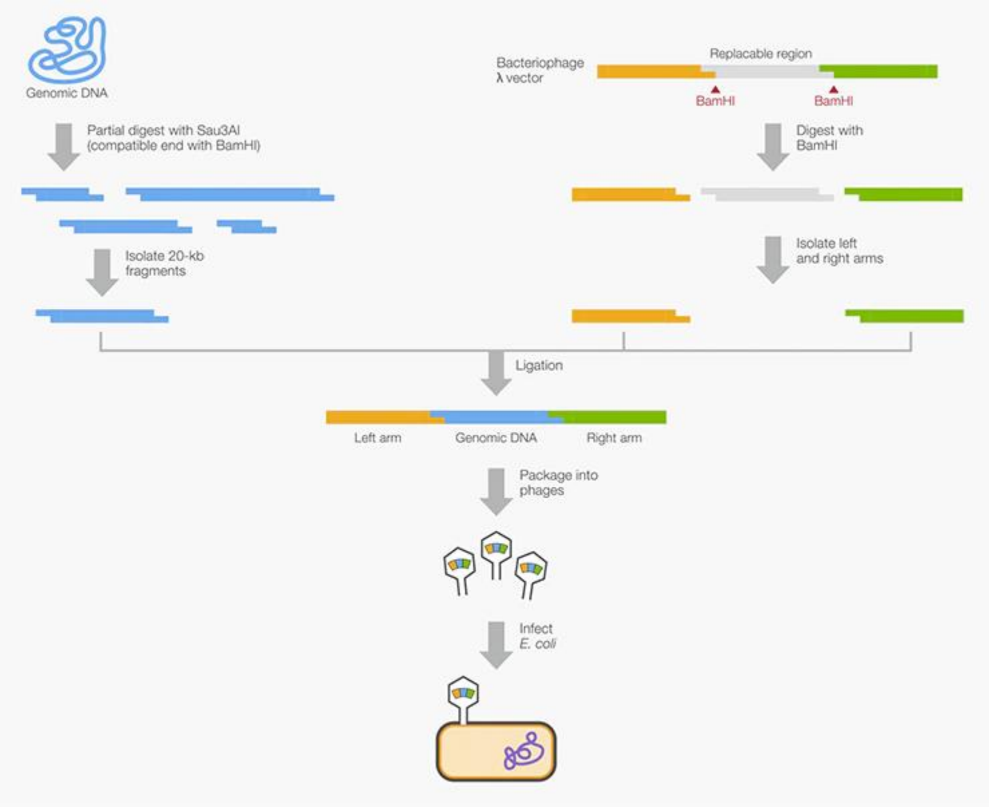


Figure 6. Schematic workflow of genomic library preparation using a λ phage vector. A genomic DNA sample is partially digested with Sau3AI, after which ~20-kb fragments (ideal size for viral packaging) are isolated for ligation with the viral gene fragments. The left and right arms of the λ vector comprise essential components for viral growth in the bacterial cells.

- For cDNA library preparation, total RNA is extracted from a biological source (e.g., cells, tissue, etc.), after which mRNA is reverse transcribed into complementary DNA (cDNA). This process is known as first-strand cDNA synthesis. The second strand is then synthesized to obtain the double-stranded cDNAs. The resulting double-stranded fragments may be ligated directly into a blunt-end cloning vector (random cloning), or “tagged” at the ends with restriction sites for directional cloning **(Figure 7)**.

- cDNA libraries that provide good, faithful representation of the expressed genome depend on several factors including the quality and integrity of the source mRNA population. For the reverse transcription steps, it is also crucial that the reverse transcriptase is capable of synthesizing cDNA from a mixed and complex population, including long RNA templates and rare RNA transcripts, for adequate coverage within the libraries (see reverse transcriptase choices).

- Using the basic strategy outlined in Figure 7, many cDNA library preparations were used to construct comprehensive collections including the Mammalian Gene Collection (MGC), the largest NIH-sponsored public collection of cDNA clone libraries of mammalian species including human, mouse, and rat.

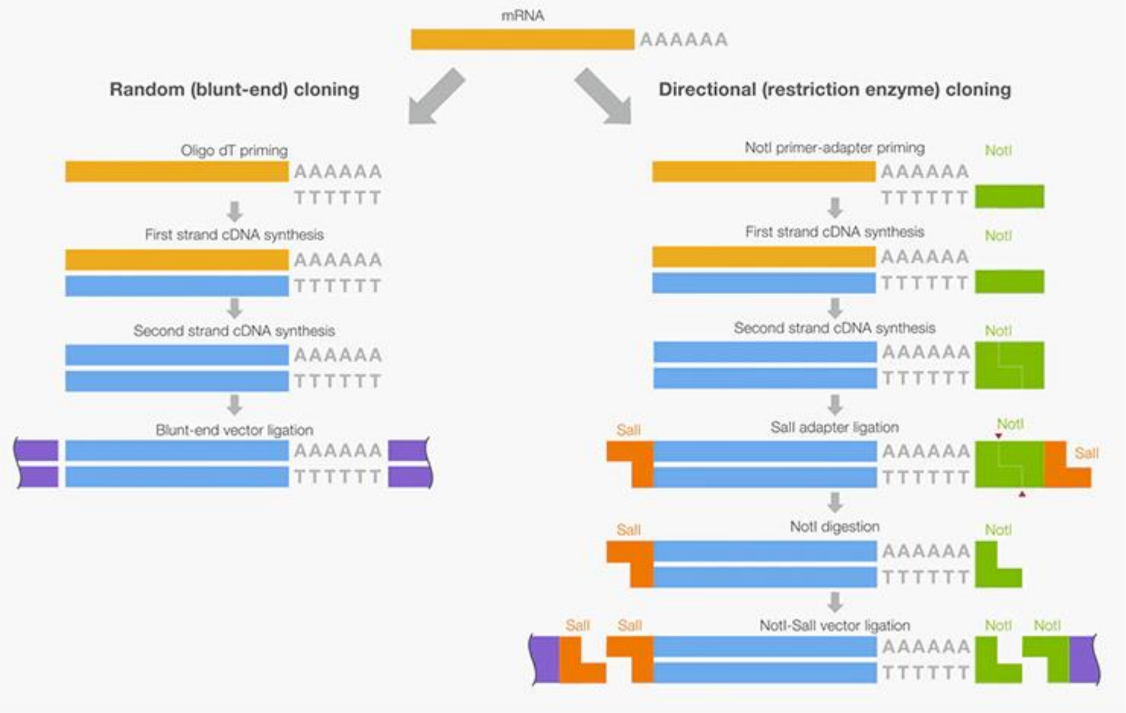

mRNA
AAAAAA
Random (blunt-end) cloning
Directional (restriction enzyme) cloning
Oligo dT priming
AAAAAA
TTTTTT
First strand cDNA synthesis
Second strand cDNA synthesis
Blunt-end vector ligation
NotI primer-adapter priming
NotI
SalI adapter ligation
SalI
NotI digestion
NotI-SalI vector ligation

Figure 7. cDNA cloning strategies using mRNA with a poly-A tail. In random (non-directional) cloning, double-stranded cDNA are ligated directly to a blunt-end cloning vector. In directional cloning, adapters with rare restriction sites (e.g., NotI and SalI) are ligated to the double-stranded cDNA ends to clone into a vector with compatible ends.

Shotgun cloning and sequencing method

- Following library construction, one of the goals is to characterize the clones by sequencing the inserts. Insert sizes represented within these libraries can often range from 25 kb to 300 kb, depending on the type of vectors and the genome size of the organism of interest.

- For Sanger sequencing, once the most widespread method for DNA sequencing, the upper limit of a sequencing reaction with good-quality reads is generally less than 1 kb.

- To overcome this dilemma, researchers can turn to shotgun cloning and sequencing. In this approach, the large cloned inserts are further fragmented by physical or enzymatic means and subcloned into another vector; the smaller cloned fragments are then sequenced. These sequences are reassembled thereafter based on sequence overlaps (termed contiguous or “contigs”) using bioinformatics programs to ultimately obtain the original long sequence (**Figure 8**).

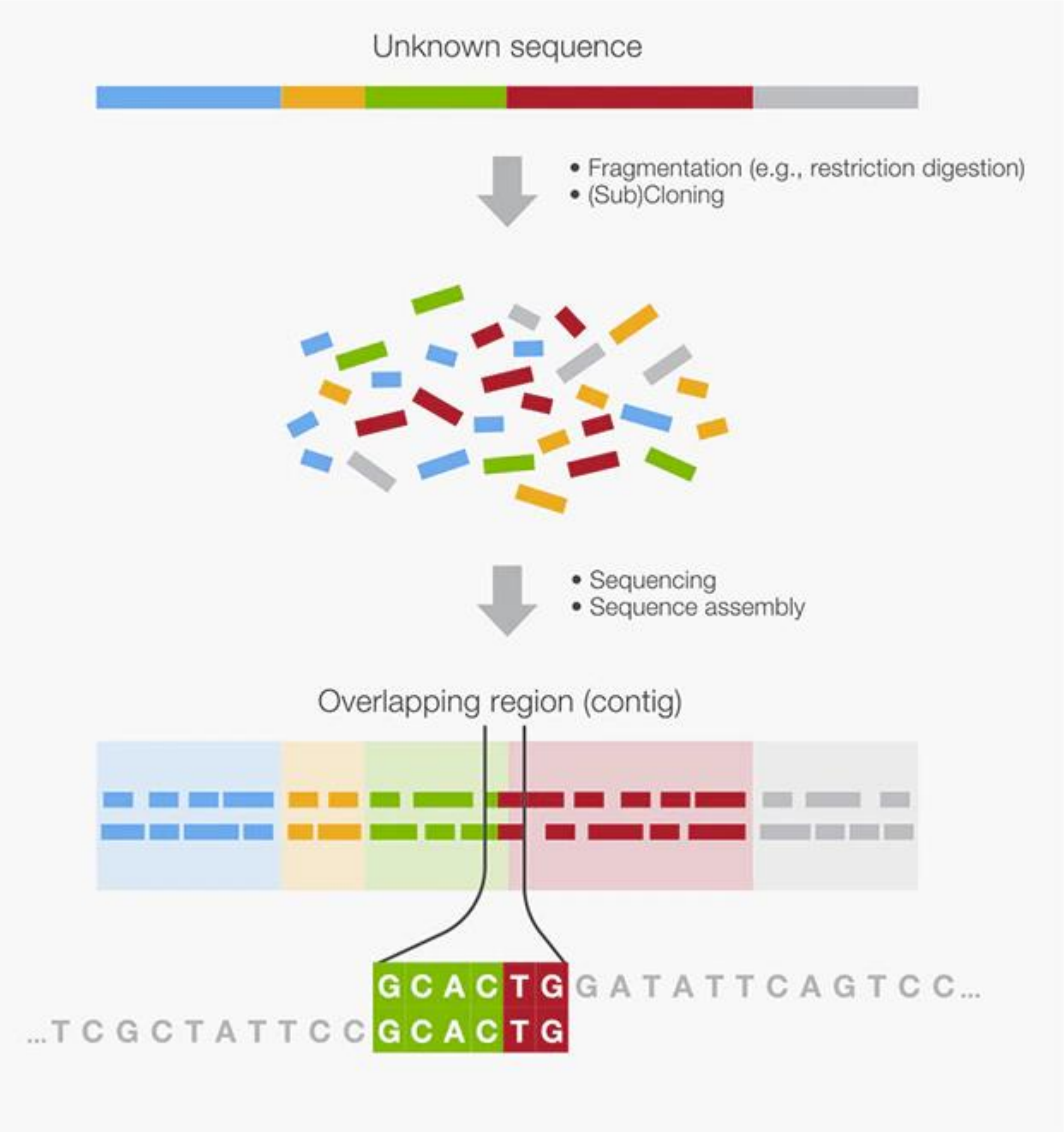


Figure 8. Schematic workflow of shotgun cloning.

- Shotgun sequencing is instrumental in whole-genome sequencing of many organisms, ranging from viruses and bacteria to human. The method can be used to sequence the genome *de novo*, as well as improve quality of already-sequenced genome by verifying reads and filling in gaps.

- During the first sequencing of the human genome, the publicly funded Human Genome Project employed shotgun sequencing of large gene fragments that had been cloned into a bacterial artificial chromosome or BAC vector. The genomic positions of the cloned fragments had been defined prior to shotgun cloning, making their shotgun sequence assembly easier. Hence, this method is known as **hierarchical shotgun sequencing** (**Figure 9A**). It is also called **clone-by-clone sequencing** due to the use of BAC clones as a source.

- Concurrent with the Human Genome Project, another privately funded whole genome sequencing project led by Craig Venter used shotgun sequencing strategies directly on the human genome DNA (instead of cloned fragments that had already been mapped). This process is known as the **whole-genome shotgun approach** (**Figure 9B**).

- In theory, shotgun sequencing requires no prior information about the genome or genetic maps, and would save time and resources. Nevertheless, it is helpful to have reference genetic maps during sequence assembly because a large amount of computational power is required in the whole-genome shotgun approach, especially for organisms with sizable genomes. Genetic mapping or fingerprinting is routinely carried out using restriction enzymes [4], as in the methods of RFLP and AFLP.

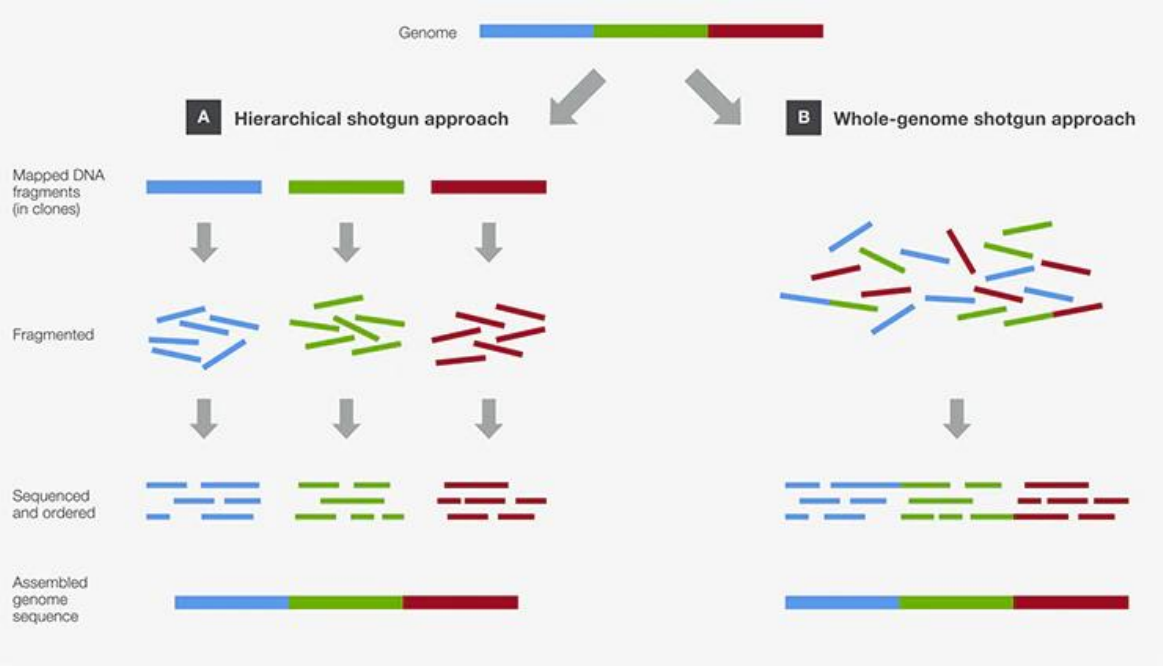
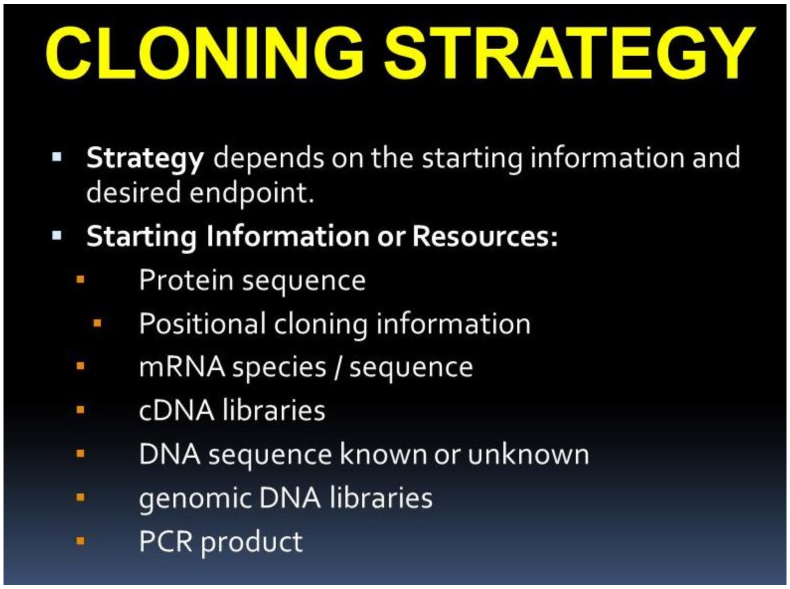

Genome
A
Hierarchical shotgun approach
B
Whole-genome shotgun approach
Mapped DNA fragments (in clones)
Fragmented
Sequenced and ordered
Assembled genome sequence

Figure 9. Schematic workflow of two shotgun sequencing approaches used in whole human genome sequencing.



CLONING STRATEGY

- **Strategy** depends on the starting information and desired endpoint.
- **Starting Information or Resources:**
  - Protein sequence
  - Positional cloning information
  - mRNA species / sequence
  - cDNA libraries
  - DNA sequence known or unknown
  - genomic DNA libraries
  - PCR product

Plasmid

- Plasmids are the extrachromosomal genetic elements found in bacteria.
- They are circular pieces of DNA that are extra genes.
- About 1-20 copies of plasmids are present in one bacterial cell.
- Episomes are the type of plasmid that can be inserted into the bacterial chromosome and can replicate with it.
- For normal life and functioning, a plasmid is not required in the bacteria. But their presence confers new properties in the bacteria. **Example:** Drug resistance, toxigenicity

Properties/Characteristics of bacterial plasmids:

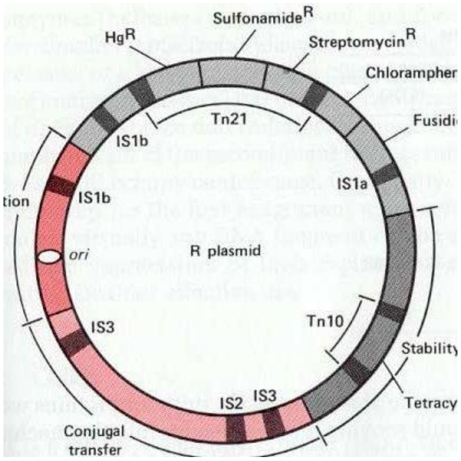
1. **Physical properties:**
    - Plasmid is a double-stranded circular and supercoiled DNA.
    - Within a cell, it can exist autonomously. It can replicate independently of the bacterial chromosome.
    - It has a molecular weight of 10^6-10^8 which may encode from 40-50 genes.
    - It has about 1-3% of the weight of the bacterial chromosome consisting of 1500-400,000 base pairs.
    - Plasmid as large as 2 million base pairs can occur in some bacteria.
2. **Replication:**
    - It contains genes for self-replication.
3. **Curing:**
    - It can be lost spontaneously or by curing agents.
4. **Incompatibility:**
    - In the same cell, two members of the same group cannot co-exist.
5. **Transferability:**
    - Some plasmids are self-transferable.
6. **Recombinations:**
    - Episome can integrate with host chromosome.
7. **Mobilisation:**
    - By the process of integration, the self-transferable plasmid can mobilize the chromosomal gene or other plasmids.

Types of plasmid:

- Based on their function, plasmids are of five types:
- Resistance (R) plasmid
- Fertility (F) plasmid
- Bacteriocinogen or Col plasmid
- Degradative plasmid
- Virulence plasmid

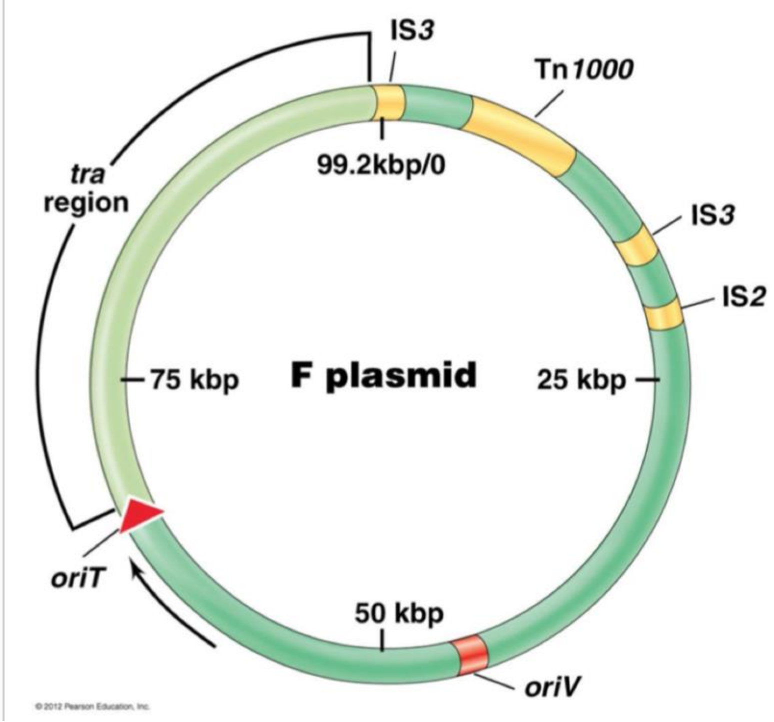
1. R-plasmid (R-factor):

- They are circular with double-stranded plasmid.
- R factor occurs in two sizes:
    - large plasmids (mol. wt. 60 million)
    - small plasmids (mol. Wt. 10 million)
- Large plasmids are conjugative ‘R’ factors. To code for the conjugation process, it contains extra DNA.
- Small plasmids contain only the ‘r’ genes. They are not conjugative.
- **It consists of two components.**
    - Resistance transfer factor (RTF): carries the genes that govern the process of intercellular transfer.
    - Resistant determinant (R-determinant): carries resistant genes for each of the several drugs.
- The drug resistance is not transferrable in the case when RTF dissociates from the R-determinant.
- For the spread of the multiple drug resistance in the bacteria, R factor plays a vital role.
- Antibiotics can be destroyed and the membrane transport system can be modified.
- R-factor may carry the resistance genes either one, two, or more than these.
- They may also carry the gene resistance for the metal ions.
- They also carry resistance to certain bacteriophages by coding for the enzymes.



2. F-plasmids:

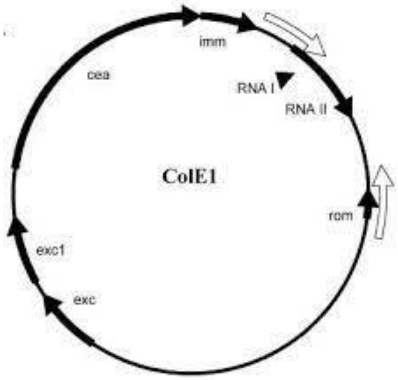
- It is a transfer factor or F-factor.
- It contains genetic information, which controls the mating process of the bacteria during the conjugation.
- It contains the basic genetic information necessary for:
    - Extra-chromosomal existence
    - Self-transfer
    - Synthesis of sex-pilus.
- F-plasmid carries some fourteen genes which include the structural gene for the pilin.
- Pilin is the pilus protein that functions in sex pilus formation.
- Strains of bacteria having the F plasmid are called F$_+$ and function as donors.
- Strains of bacteria lacking the F plasmid are called F- and function as recipients.
- It is also called the conjugative plasmid.
- The conjugative function is determined by the cluster of at least 25 transfer (tra) genes.
- These genes determine:
    - Expression of pili
    - Synthesis and transfer of DNA during mating
    - Interference with the ability of F$_+$ bacteria to serve as recipients.



F- Plasmid of E.coli

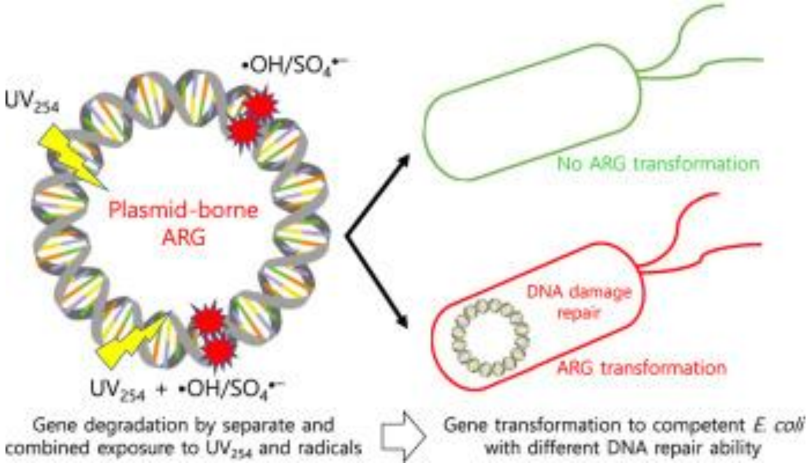
3. Bacteriocinogen or Col plasmid:

- Coliforms produce extracellular colicins.
- In the several species of coliform, the colicinogenic (col) factors are present.
- These bacterial factors are the lethal toxins for the closely related species or even for the different strains of the same species.
- Some bacterial substances are produced not only by the coliforms but also by the other bacteria.
- This group of substances is called bacteriocins.
- Colicins are produced by *coli*
- Pyocin are produced by *Pseudomonas aeruginosa.*
- Marscesins are produced by *Serratia marcescens.*
- Diphthericin is produced by *Corynebacterium diphtheria.*
- Bacteriocin produced by the different bacterial strains helps in the interspecies typing of organisms.



4. Degradative plasmids:

- From the dead plants and animals, degradative plasmid helps in the degradation and digestion of the dead organic matter.
- It is then used in the biosynthesis process.
- It will make energy and will recycle further.

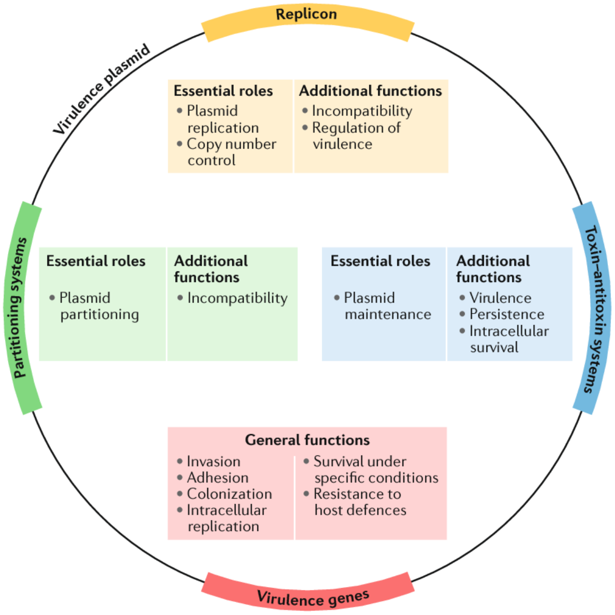

UV$_{254}$
•OH/SO$_4^{•-}$
Plasmid-borne ARG
UV$_{254}$ + •OH/SO$_4^{•-}$
Gene degradation by separate and combined exposure to UV$_{254}$ and radicals
No ARG transformation
DNA damage repair
ARG transformation
Gene transformation to competent *E. coli* with different DNA repair ability

5. Virulence plasmids:

- With the help of this plasmid, bacteria will be transformed into a pathogen.
- It carries the genes which are responsible for causing disease.

6. Cryptic plasmids:

- Cryptic Plasmids do not have any apparent effect on the phenotype of the cell harboring them. They just code for enzymes required for their replication and maintenance in the host cell.

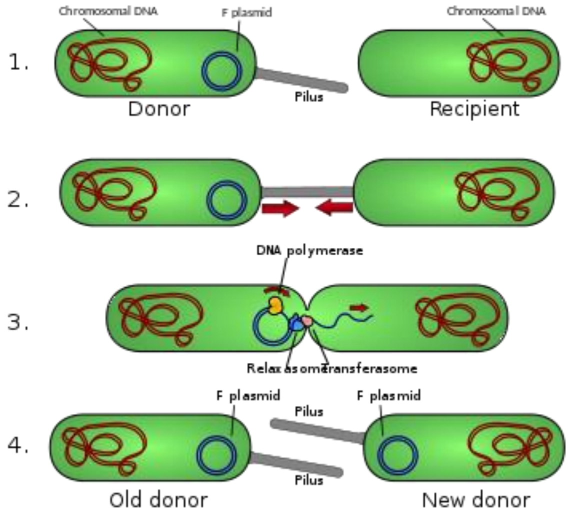


Based on the role in conjugation, plasmids are of two types:

- Conjugative plasmid
- Non-conjugative plasmid

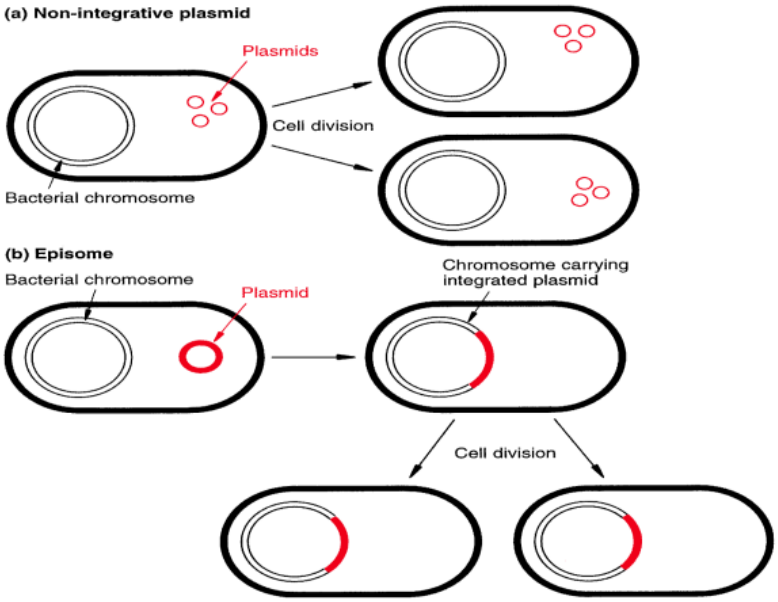
i. Conjugative plasmids:

- These large plasmids (F plasmids) carry genes that are responsible for transferring themselves to other cells.
- It includes the genes that direct the synthesis of sex pilli.



ii. Non-conjugative plasmids:

- These plasmids are present in Gram-positive bacteria, especially in the Gram-positive cocci.
- It is also present in the Gram-negative organism. Example: *Haemophilus influenza, Neisseria gonorrhoeae.*
- They are usually small, 1-10 dal.
- In each bacterium, multiple copies (more than 30) may be present.
- When the same bacterium carries both the conjugative and non-conjugative plasmids, they can be mobilized for transfer to another cell.
- When the conjugation is established then the donor can transfer non-conjugative plasmids.



Replication strategies- a. Plasmid b. Episome

Functions/applications of plasmid:

- The main function of the plasmid is the spread of antibiotic-resistant genes. These resistant genes are carried within the plasmid and are transferred from one cell to another.
- Plasmid is used in recombinant DNA technology.
- To deliver the desired drug into the body, a plasmid is used.
- For the insertion of the human insulin on the body
- Insertion of human growth hormone in mammalian cells of animals.
- Plasmids are used in Gene Therapy:
  - For the insertion of the therapeutic genes in the human body. It helps to fight against diseases.
  - Easy manipulation and can be replicated in bacterial cells easily.
  - Targeting the defected cells easily and triggering the therapeutic genes in them.
- Plasmids carry the genes involved in metabolic activities. They aid in the digestion of pollutants from the environment.
- Plasmids can produce antibacterial proteins.
- Plasmid can carry genes that increase the pathogenicity of the bacteria.
- When the nutrients are scarce, the plasmid can help bacteria by:

***** Fix the nitrogen   ***** Degrade organic compounds

Sizes of representative plasmids.

	SIZE		
PLASMID	**NUCLEOTIDE LENGTH (kb)**	**MOLECULAR MASS (MDa)**	**ORGANISM**
pUC8	2.1	1.8	*E. Coli*
ColE1	6.4	4.2	*E. Coli*
RP4	54	36	*Pseudomonas* and others
F	95	63	*E. Coli*
TOL	117	78	*Pseudomonas putida*
pTiAch5	213	142	*Agrobacterium tumefaciens*

Natural Plasmids .

Plasmid	Size (kb)	Origin	Host range	Antibiotic resistance	Additional marker genes showing insertional inactivation
RSF1010	8.6	*E.coli* (strain K-12)	Broad host range	Streptomycin and sulfonamides	None
ColE1	6.6	*E.coli*	Narrow host range	None	Immunity to colicin E1
R100	94.2	*E.coli*	*E.coli* K-12, *Shigella flexneri* 2b	Streptomycin, chloramphenycol, tetracycline	Mercuric (ion) reductase, putative ethidium bromide (EtBr) resistant protein.

Characteristics of ideal plasmid vectors

1. **Size:** plasmid must be small in size. The small is helpful for easy uptake of cDNA by host cells and for the isolation of plasmid without damage. **Ideal vector should be less than or equal to 10kb.** The small size is essential for easy introduction in cell by transformation, transduction and electroporation.
2. **Copy number:** the plasmid must be present in **multiple copies**.
3. **Genetic markers:** plasmid must have **one or few genetic markers**. These markers help us for the selection of organism that has recombinant DNA
4. **Origin of replication:** the plasmid must have **its own orogin of replication and regulatory genes for the self-replication.**
5. **Unique restriction sites:** the plasmid must have unique restriction sites common restriction enzymes in use.
6. **Multiple cloning sites:** This property permits the insertion of gene of interest and plasmid re-circularization.
7. **Insertional inactivation:** the plasmid must have unique sites for restriction enzymes in marker genes. This will help us for the selection of recombination by insertional inactivation method.
8. **Pathogenicity:** the plasmid should not have any pathogenic property.
9. **Should not be transferred by conjugation:** This property of vector molecule prevents recombinant DNA to escape to natural population of bacteria.
10. **Selectable make gene:** Vector molecules should have some detectable traits. These traits enable the transformed cells to be identified among the non-transformed ones. eg. antibiotic resistance gene.

Host Range of plasmid

- The host range of a plasmid means the types of bacteria in which the plasmid can replicate.
- It is usually determined by the *ori* region from where the replication starts.

Plasmid having the narrow host range includes:

- ColE1 plasmid type; Example: pBR322, pET, and pUC.
- Replication of these plasmids occurs only in *coli.*
- It may occur in *Salmonella* and *Klebsiella* also which are closely related bacteria.

Plasmids having the broad host range include:

- RK2
- RSF1010 plasmids
- RC plasmids; Example: pBBR1MCS .
- Plasmids with the *ori* region of RK2 can replicate in most types of Gram-negative proteobacteria.
- Plasmids with the RSF1010-derived plasmids can replicate in Gram-positive bacteria too. Example: *Firmicutes*.
- Replication of the same plasmid can occur even in the distantly related bacteria.
- Broad-host-range plasmids do not depend on the host cell because they encode their proteins. These proteins are essential for the initiation of replication.
- Broad host-range plasmids should be for gene expression in many types of bacteria.

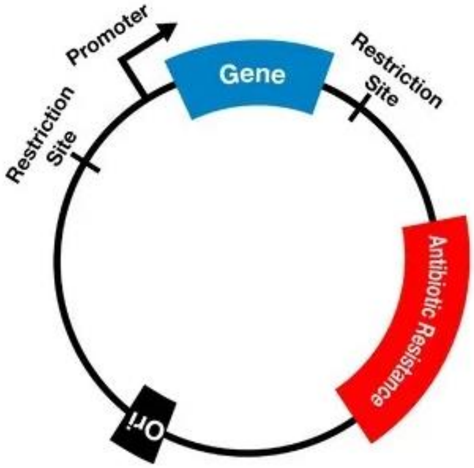
Determining the Host Range

- It is sometimes difficult to ensure the particular plasmid will replicate on the other host or not. So, The actual host ranges of most plasmids are unknown.
- Initially, plasmids need to be introduced to the other bacteria.
- So, for this process, a system has been developed which is known as transformation.
- By this method, the plasmid can be introduced into the bacteria to see if it could replicate or not.
- But it has limitations as it can’t be applied to all types of bacteria.
- Similarly, to introduce DNA into cells, electroporation can be used.
- Plasmids can be introduced into other types of bacteria by the conjugation process.
- It is also found that the expression of the gene present in one plasmid does not function well or doesn’t get expressed in the other bacterium.
- Sometimes the selected gene can be introduced into the different bacteria.
- A bacterium might possess resistance to any antibiotic due to the presence of a particular gene. Such resistance property can be transferred to other bacteria too when they will uptake those resistant genes.

- For example, the kanamycin resistance gene, which is first found in the Tn*5* It can be expressed in most Gram-negative bacteria. Then, it will make resistant to kanamycin antibiotic.
- By this property, a marker gene can be cloned in the plasmid. i.e making numerous copies.
- A transposon carrying a selectable marker into the plasmid can also be introduced by this method.
- Care must also be taken to ensure that the plasmid has not recombined into the host chromosome.
- Determining the host range of a plasmid is laborious too. Many barriers hinder the transfer of plasmid into the host. The same method can't be approached to all sorts of plasmids and bacteria.

Some important components of plasmids are

- **Origin of replication (Ori)**: A DNA sequence that allows bacteria to make more copies of the plasmid as they grow and divide.
- **Antibiotic resistance gene**: A specific gene that allows bacteria with the plasmid to grow in the presence of an antibiotic specific to the gene.
- **Gene**: A DNA sequence encoding a particular protein that a researcher has inserted into the plasmid to study.
- **Promoter**: A DNA sequence that allows the cell to produce the protein encoded by the gene.
- **Restriction sites**: DNA sequences that allow a researcher to cut and paste components of plasmids together



Types & Functions of Plasmids

As a single plasmid may carry many different genes, the classification of a plasmid in a single phenotypic category is difficult. Some of the notable types of plasmids and their functions are:

- Conjugative plasmids: Examples include F plasmid present in *E.coli*, conjugative P plasmid of *Vibrio cholerae*.

- Resistance plasmids (R plasmids): R plasmids confer resistance to antibiotics and various inhibitors of growth.
    - It carries a variety of antibiotic resistance genes which encode proteins that either inactivate the antibiotic or affect its uptake into the cell. Example: Plasmid R100 carries resistance genes for sulfonamides, streptomycin, fusidic acid, chloramphenicol, and tetracycline.
    - R100 also carries several genes that confer resistance to mercury.
    - Resistant strains can transfer resistance to sensitive strains via cell-to-cell contact. R100 can transfer itself between enteric bacteria of the genera *Escherichia, Klebsiella, Proteus, Salmonella,* and *Shigella* but does not transfer to the non-enteric bacterium *Pseudomonas.*
- Plasmids that code for virulence characteristics & toxins
    - Adherence/Colonization: Some of the plasmids code for the proteins that increase the ability of the organism to attach and colonize in specific sites within the host e.g. colonization factor antigen (CFA) of *E.coli.*
    - Toxin production: Toxin production in various pathogenic bacteria is found to be linked with the presence of plasmids. For example:
        - Hemolysin (lyse RBCs) and enterotoxin (induces extensive secretion of salt and water in the bowel) property of Enteropathogenic *Escherichia coli* (EPEC) are governed by plasmids.
        - Production of coagulase, hemolysin, fibrinolysin, and enterotoxin property of *S.aurues* is linked to the presence of the plasmid.
- Bacteriocins: Many bacteria produce peptides that inhibit or kill closely related species or even different strains of the same species. The gene responsible for this peptide and or its post-translational modification is coded in plasmids. For e.g colicin of *E.coli* is coded by Col plasmids.

Although plasmids carry useful genes such as genes that confer “antibiotic resistance”, as mentioned above, they do not carry genes that are essential to the host under all conditions.

The presence of plasmids in a cell can also have other biological significance such as:

- Nodulation and symbiotic nitrogen fixation: *Rhizobium*
- Transfer genetic information for a biochemical pathway for the degradation of organic compounds such as octane, camphor, naphthalene, salicylate etc: *Pseudomonas*.
- Pigment production: *Erwinia*, *Staphylococcus*
- Lactose, sucrose, urea utilization, nitrogen fixation: Enteric bacteria

Plasmids can be constructed artificially (artificial plasmids are called vectors) and are used to introduce foreign DNA into another cell of interest. Plasmids play crucial roles in genetic engineering, molecular cloning and various areas of Biotechnology.

Plasmid Vector

Plasmids and bacteriophages are frequently used as cloning vectors in DNA recombinant technology.

- The ease with which plasmids can be modified and replicated makes it a great tool in genetic engineering and biotechnology
- For genetic engineering purposes, plasmids are artificially prepared in the lab
- The lab-grown plasmids, which are used as a vector contain an origin of replication, cloning site and selection marker

Vector Element	Description
Origin of Replication (ORI)	DNA sequence where initiation of replication starts
Selectable Marker	For selecting bacteria containing desired plasmid, e.g. **antibiotic resistance genes** and other specific genes
Multiple Cloning Sites (MCS)	Recognition sites to insert foreign DNA fragment by using restriction enzymes, a few or single recognition site is preferred to avoid getting several fragments
Promoter Region	Promotes transcription of the target gene to get the desired protein
Primer Binding site	The sequence of DNA used as a start point for PCR amplification and sequence verification

Gene Cloning Vector

- A vector is a DNA molecule that is used to carry a foreign DNA into the host cell. It has the ability to self replicate and integrates into the host cell. These vectors have helped in analysing the molecular structure of DNA.
- Vectors can be a plasmid from the bacterium, a cell from the higher organism or DNA from a virus. The target DNA is inserted into the specific sites of the vector and ligated by DNA ligase. The vector is then transformed into the host cell for replication. ***"A cloning vector is a small piece of DNA into which a foreign DNA can be inserted for cloning purposes."***

Features of Cloning Vectors

The cloning vectors possess the following features:

1. A cloning vector should possess an origin of replication so that it can self-replicate inside the host cell.
2. It should have a restriction site for the insertion of the target DNA.
3. It should have a selectable marker with an antibiotic resistance gene that facilitates screening of the recombinant organism.
4. It should be small in size so that it can easily integrate into the host cell.
5. It should be capable of inserting a large segment of DNA.
6. It should possess multiple cloning sites.
7. It should be capable of working under the prokaryotic and eukaryotic systems.

Ideal Characteristics of Cloning Vectors

1. Origin of Replication (ori)

- A specific set/ sequence of nucleotides where replication initiates.
- For autonomous replication inside the host cell.
- Foreign DNA attached to ori also begins to replicate.

2. Cloning Site

- Point of entry or analysis for genetic engineering.
- Vector DNA at this site is digested and foreign DNA is inserted with the aid of restriction enzymes.
- Recent works have discovered plasmids with multiple cloning sites (MCS) which harbour up to 20 restriction sites.

3. Selectable Marker

- Gene that confers resistance to particular antibiotics or selective agent which, under normal conditions, is fatal for the host organism.
- Confers the host cell the property to survive and propagate in culture medium containing the particular antibiotics.

4. Marker or Reporter Gene

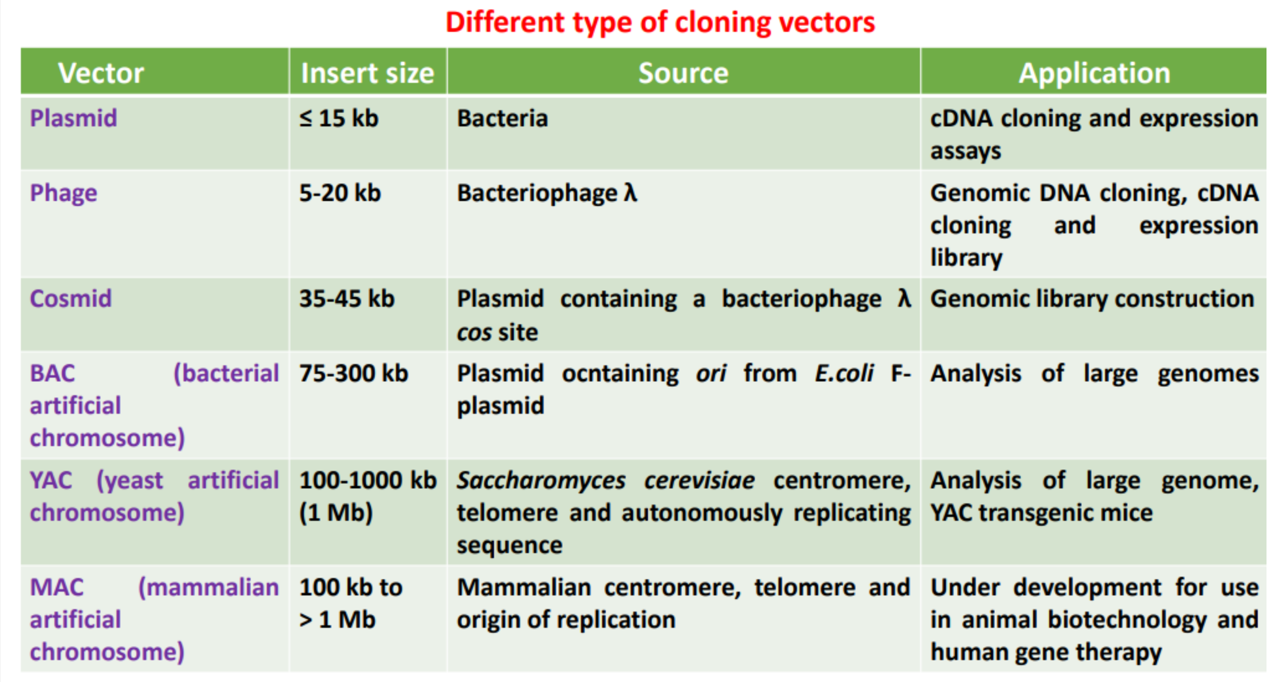
- Permits the screening of successful clones or recombinant cells.
- Utilized extensively in blue-white selection.

5. Inability to Transfer via Conjugation

- Vectors must not enable recombinant DNA to escape to the natural population of bacterial cells.

Types of vector

- **Vectors are of two types:**
  - a) **Cloning vector**
  - b) **Expression vector**



Different type of cloning vectors

Vector	Insert size	Source	Application
Plasmid	≤ 15 kb	Bacteria	cDNA cloning and expression assays
Phage	5-20 kb	Bacteriophage λ	Genomic DNA cloning, cDNA cloning and expression library
Cosmid	35-45 kb	Plasmid containing a bacteriophage λ *cos* site	Genomic library construction
BAC (bacterial artificial chromosome)	75-300 kb	Plasmid ocntaining *ori* from *E.coli* F-plasmid	Analysis of large genomes
YAC (yeast artificial chromosome)	100-1000 kb (1 Mb)	*Saccharomyces cerevisiae* centromere, telomere and autonomously replicating sequence	Analysis of large genome, YAC transgenic mice
MAC (mammalian artificial chromosome)	100 kb to > 1 Mb	Mammalian centromere, telomere and origin of replication	Under development for use in animal biotechnology and human gene therapy

Cloning vectors

- Cloning vectors are small piece of DNA which have the ability and used to introduce foreign gene of interest into the host cell.
- They can be stably maintained insides the host cell.
- Cloning vector are generally used to obtain multiple copies of desired foreign gene.
- Example- Plasmid, Cosmid and Phages, BACs, YACs.
- These type of vectors generally contains selectable marker, origin of replication and a restriction site.

Expression vector

- Expression vector is a type of vector which not only introduces a gene of interest into the host cell but also aids in the analysis of the foreign gene via relevant protein product expression.
- It is type of vector which is used to obtain or analyses the gene product, which may be RNA or protein of the inserted desired gene.
- Example- Only plasmid vector.
- Expression vector contains enhancer, promoter region, start/stop codon, transcription initiation, selectable marker, ori sites, and restriction site.

Shuttle vector

- Shuttle vectors are created to replicates in cell of different type of species.
- They contain two origin of replication, in which one is particular for each host species, also those genes required for their replication and not provided by the host cell.
- These types of vectors are developed by recombinant techniques.

Types of Cloning Vectors

There are the following different types of cloning vectors:

Plasmids

- These were the first vectors used in gene cloning.
- They are naturally occurring and autonomously replicating extra-chromosomal double-stranded circular DNA molecules. However, not all plasmids are circular in origin.
- These are found in bacteria, eukaryotes and archaea. These are natural, extrachromosomal, self-replicating DNA molecules.
- The size of plasmids ranges from 1.0 kb to 250 kb.
- DNA insert of up to 10 kb can be cloned in the plasmids.
- The plasmids have high copy number which is useful for production of greater yield of recombinant plasmid for subsequent experiments and also possess antibiotic-resistant genes.
- The low copy number plasmids are exploited under certain conditions like the cloned gene produces the protein which is toxic to the cells.

- Plasmids only encode those proteins which are essential for their own replication. These protein-encoding genes are located near the ori.

Examples: pBR322, pUC18, F plasmid, Col plasmid.

Based on the origin or source of plasmids, they have been divided into two major classes: such as natural and artificial.

i) Natural plasmids: They occur naturally in prokaryotes or eukaryotes. Example: ColE1.

ii) Artificial plasmids: They are constructed in-vitro by re-combining selected segments of two or more other plasmids (natural or artificial). Example: pBR322.

Some of the phenotypes which the naturally occurring plasmids confer on their host cells:

- Antibiotic resistance
- Degradation of aromatic compounds
- Sugar fermentation
- Heavy metal resistance
- Induction of plant tumors
- Antibiotic production
- Hemolysis production
- Enterotoxin production
- Bacteriocin production
- Hydrogen sulphide production

Advantages of using Plasmids as vectors:

- Easy to manipulate and isolate because of small size.
- More stable because of circular configuration.
- Replicate independent of the host.
- High copy number.
- Detection easy because of antibiotic-resistant genes.

Disadvantages of using Plasmids as vectors:

- Large fragments cannot be cloned.
- Size range is only 0 to 10kb.
- Standard methods of transformation are inefficient.

Bacteriophage

- These are more efficient than plasmids for cloning large DNA inserts.
- Bacteriophages or phages are viruses which infect bacterial cells.
- The most common bacteriophages utilized in gene cloning are Phage λ and M13 Phage.
- A maximum of 53 kb DNA can be packaged into the phage.
- If the vector DNA is too small, it cannot be packaged properly into the phage.

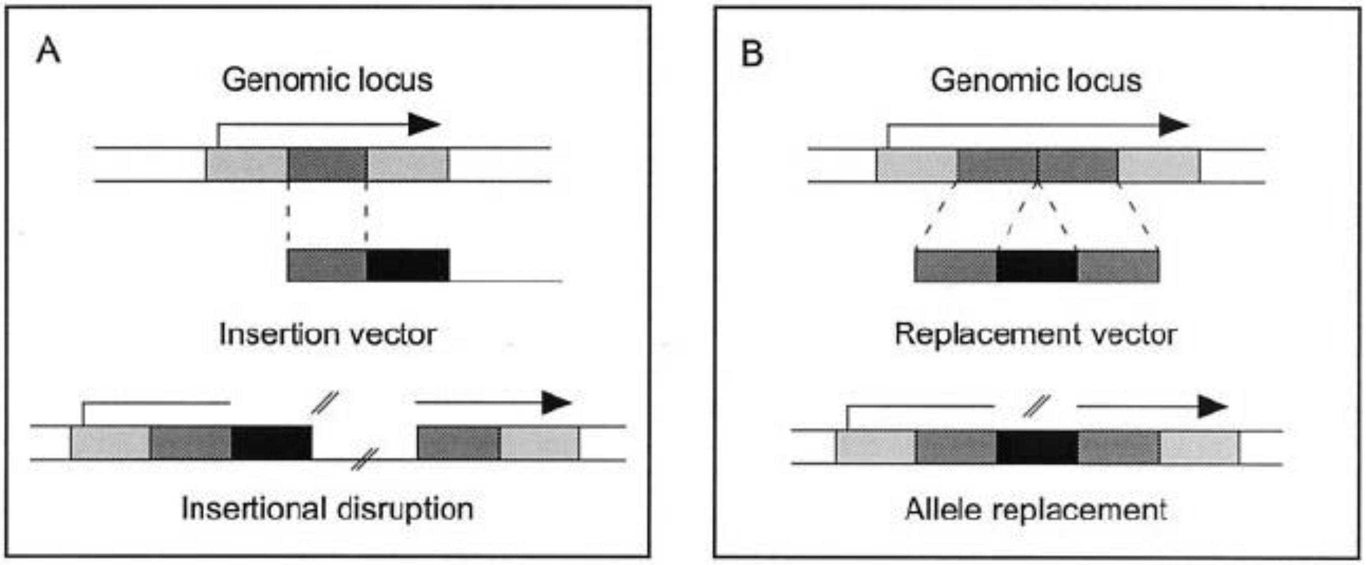
Examples: Phage Lambda, M13 Phage, etc.

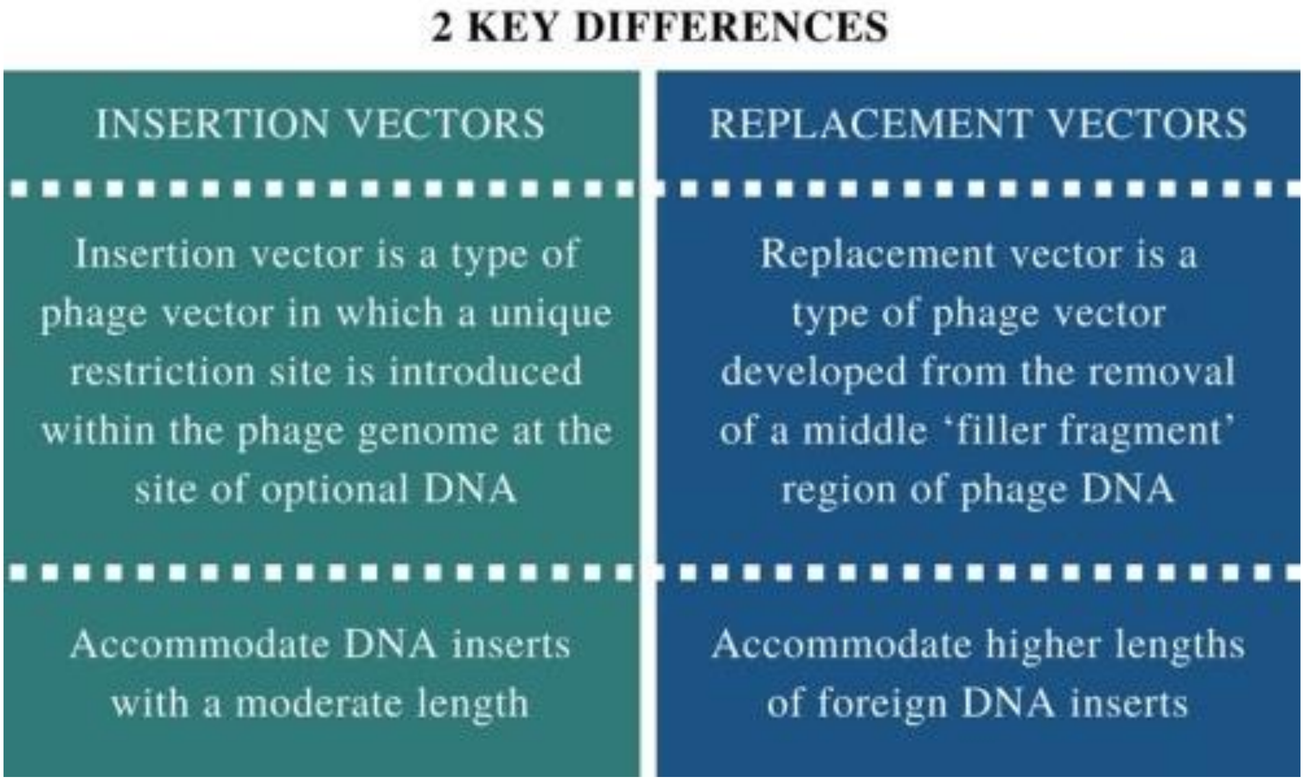
Types of Phage Vectors

- There are 2 types of phage vectors:
- Insertion vectors – these contain a particular cleavage site where the foreign DNA of up to 5-11 kb can be inserted.
- Replacement vectors – the cleavage sites flank a region which contains genes not necessarily important for the host, and these genes can be deleted and replaced by the DNA insert.

Advantages of using Phage Vectors

- They are way more efficient than plasmids for cloning large inserts.
- Screening of phage plaques is much easier than identification of recombinant bacterial colonies.





2 KEY DIFFERENCES

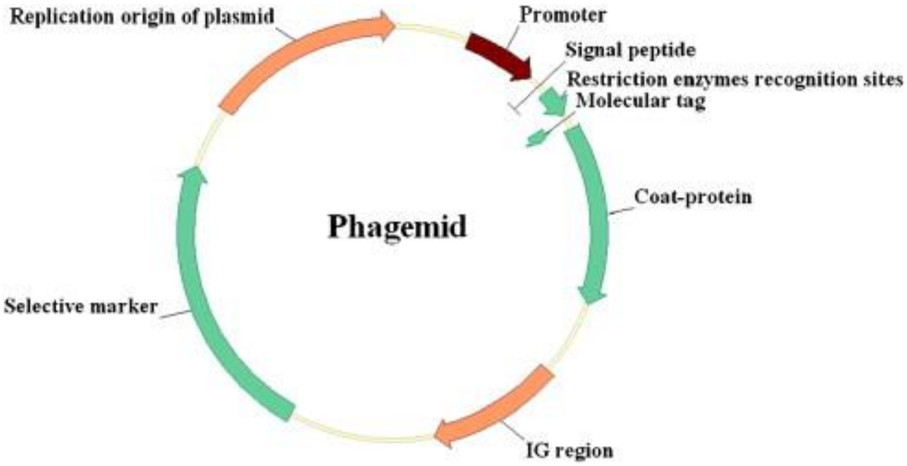
INSERTION VECTORS	REPLACEMENT VECTORS
Insertion vector is a type of phage vector in which a unique restriction site is introduced within the phage genome at the site of optional DNA	Replacement vector is a type of phage vector developed from the removal of a middle 'filler fragment' region of phage DNA
Accommodate DNA inserts with a moderate length	Accommodate higher lengths of foreign DNA inserts

Phagemids or Phasmid

- They are prepared artificially.
- Phasmid contains the F1 origin of replication from F1 phage.
- They are generally used as a cloning vector in combination with M13 phage.
- It replicates as a plasmid and gets packaged in the form of single-stranded DNA in viral particles.

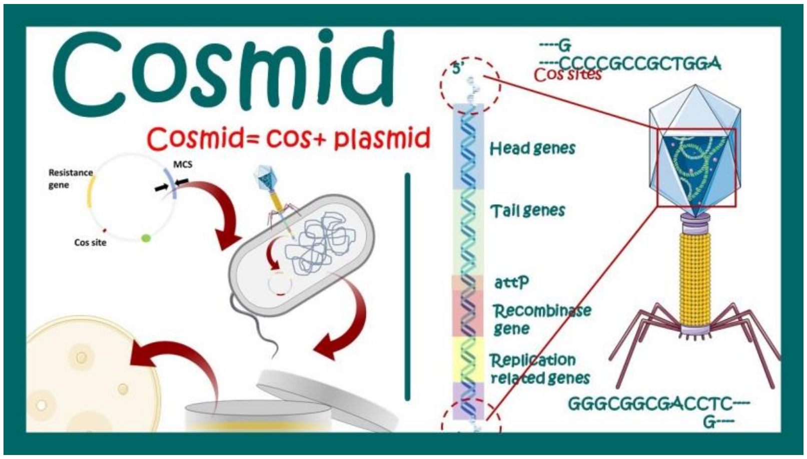
Advantages of using Phagemids:

- They contain multiple cloning sites.
- An inducible lac gene promoter is present.
- Blue-white colony selection is observed for identification.



Cosmids

- Cosmids are plasmids.
- They are capable of incorporating the bacteriophage λ DNA segment. This DNA segment contains cohesive terminal sites (cos sites).
- Cos sites are necessary for efficient packaging of DNA into λ phage particles.
- Large DNA fragments of size varying from 25 to 45 kb can be cloned.
- They are also packaged into λ This permits the foreign DNA fragment or genes to be introduced into the host organism by the mechanism of transduction.



Advantages of using cosmids as vectors:

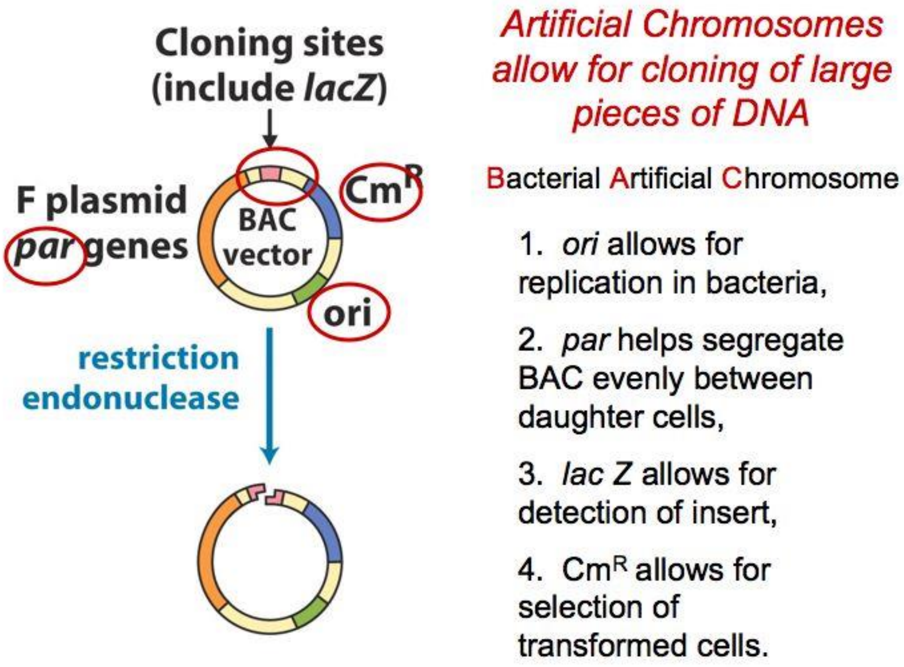
- They have high transformation efficiency and are capable of producing a large number of clones from a small quantity of DNA.
- Also, they can carry up to 45 kb of insert compared to 25 kb carried by plasmids and λ.

Disadvantages of using cosmids as vectors:

- Cosmids cannot accept more than 50 kb of the insert.

Bacterial Artificial Chromosomes

- Bacterial artificial chromosomes are similar to E. coli plasmid vectors.
- They contain ori and genes which encode ori binding proteins. These proteins are critical for BAC replication.
- It is derived from naturally occurring F' plasmid.
- They can accommodate large DNA sequences without any risk The DNA insert size varies between 150 to 350 kb.
- These are used to study genetic disorders.



Artificial Chromosomes allow for cloning of large pieces of DNA

Bacterial Artificial Chromosome

1. *ori* allows for replication in bacteria,
2. *par* helps segregate BAC evenly between daughter cells,
3. *lac Z* allows for detection of insert,
4. Cm^R allows for selection of transformed cells.

Advantages of BACs:

- They are capable of accommodating large sequences without any risk of rearrangement.
- BACs are frequently used for studies of genetic or infectious disorders.
- High yield of DNA clones is obtained.

Disadvantages of BACs:

- They are present in low copy number.
- The eukaryotic DNA inserts with repetitive sequences are structurally unstable in BACs often resulting in deletion or rearrangement.

Yeast Artificial Chromosomes (YACs)

- A large DNA insert of up to 200 kb can be cloned.
- They are used for cloning inside eukaryotic cells. These act as eukaryotic chromosomes inside the host eukaryotic cell.
- It possesses the yeast telomere at each end.
- A yeast centromere sequence (CEN) is present which allows proper segregation during meiosis.
- The ori is bacterial in origin.

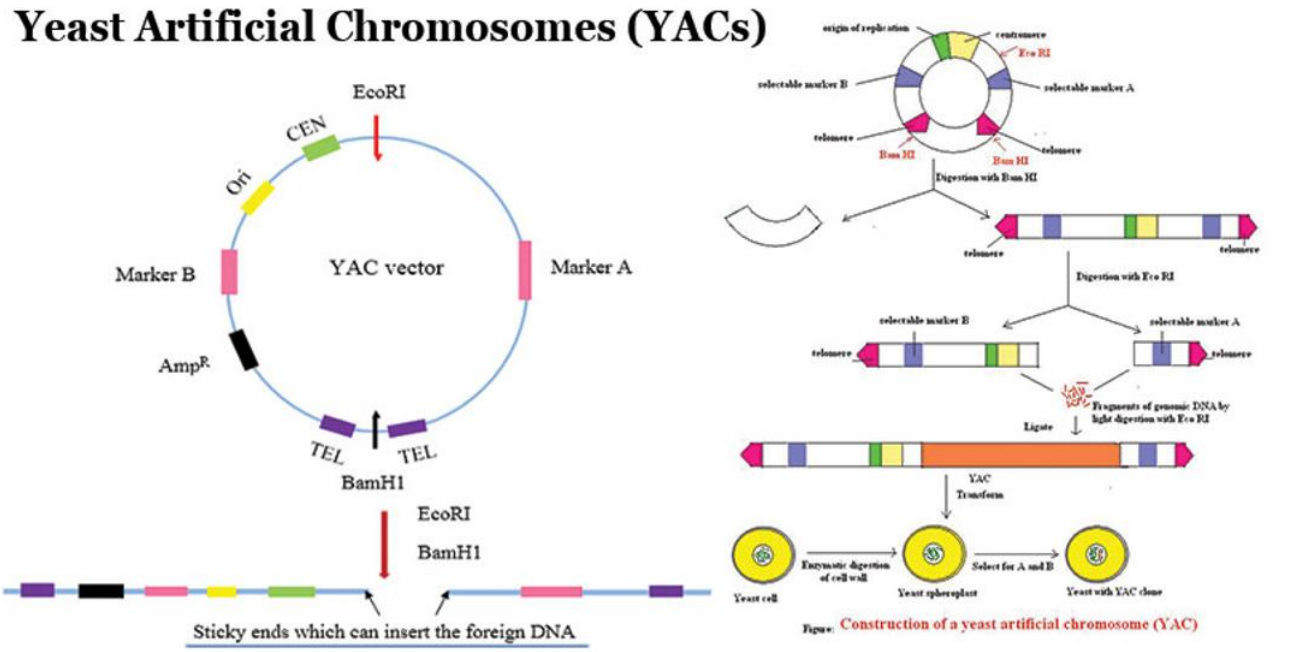


Figure: Construction of a yeast artificial chromosome (YAC)

- Both yeast and bacterial cells can be used as hosts.

Advantages of using YACs:

- A large amount of DNA can be cloned.
- Physical maps of large genomes like the human genome can be constructed.

Disadvantages of using YACs:

- Overall transformation efficiency is low.
- The yield of cloned DNA is also low.

Human Artificial Chromosome (HACs)

- Human artificial chromosomes are artificially synthesized.
- They are utilized for gene transfer or gene delivery into human cells.
- It can carry large amounts of DNA inserts.
- They are used extensively in expression studies and determining the function of the human chromosomes.

Advantages of using HACs:

- No upper limit on DNA that can be cloned.

- it avoids the possibility of insertional mutagenesis.
- Human Artificial Chromosome

Retroviral Vectors

- Retroviruses are the virus with RNA as the genetic material.
- Retroviral vectors are used for introduction of novel or manipulated genes into the animal or human cells.
- The viral RNA is converted into DNA with the help of reverse transcriptase and henceforth, efficiently integrated into the host cell.
- Any gene of interest can be introduced into the retroviral genome. This gene of interest can then integrate into host cell chromosome and reside there.

Advantages of using retroviral vectors:

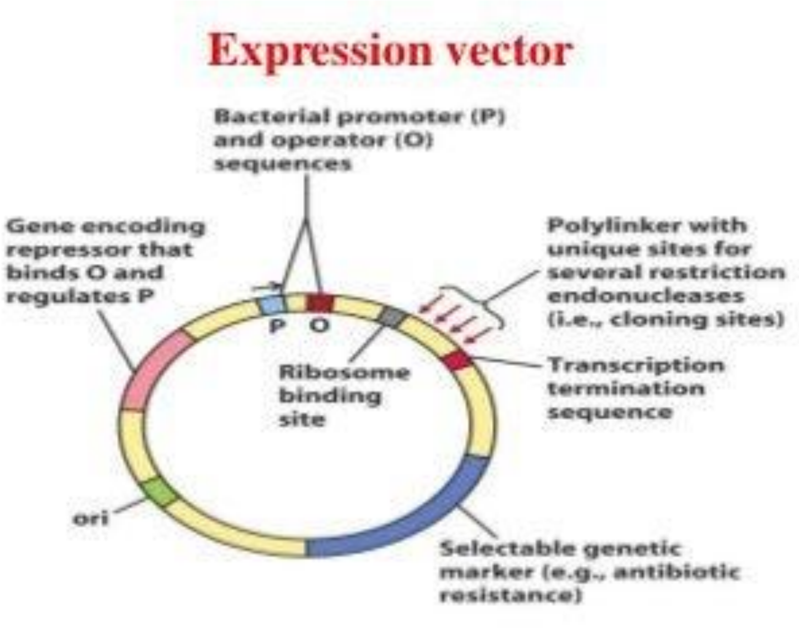
- They are widely used as a tool to study and analyze oncogenes and other human genes.

Other Types of Vectors

- All vectors may be used for cloning and are therefore cloning vectors, but there are also vectors designed especially for cloning, while others may be designed specifically for other purposes, such as transcription and protein expression.

Expression Vectors

- Vectors designed specifically for the expression of the transgene in the target cell are called expression vectors, and generally have a promoter sequence that drives the expression of the transgene. Expression vectors produce proteins through the transcription of the vector's insert followed by a translation of the mRNA produced.


Expression vector
Bacterial promoter (P) and operator (O) sequences
Gene encoding repressor that binds O and regulates P
Polylinker with unique sites for several restriction endonucleases (i.e., cloning sites)
P
O
Ribosome binding site
Transcription termination sequence
ori
Selectable genetic marker (e.g., antibiotic resistance)

Transcription Vectors

- Simpler vectors called transcription vectors are only capable of being transcribed but not translated: they can be replicated in a target cell but not expressed, unlike expression vectors. Transcription vectors are used to amplify their insert.

Uses of Vectors

- Vectors have been developed and adapted for a wide range of uses. Two primary uses are:
  (1) to isolate, identify, and archive fragments of a larger genome
  (2) to selectively express proteins encoded by specific genes.
- Vectors were the first DNA tools used in genetic engineering, and continue to be cornerstones of the technology.

Conclusion

Cloning vectors are utilized to insert foreign DNA into another cell and create multiple copies of the same. The foreign DNA is duplicated and expressed utilizing the host cell machinery. It amplifies one copy of DNA into multiple copies.

<u>Assessment:</u>

Brief the following:

1. pUC vectors.
2. pBR322 vector.
3. ColE1 Vector.

Detail the following:

4. Gene cloning strategies.
5. Gene cloning vectors.

UNIT III

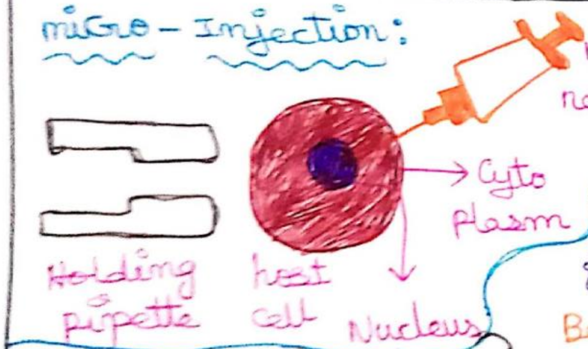
Unit - 3

Gene transfer Technique

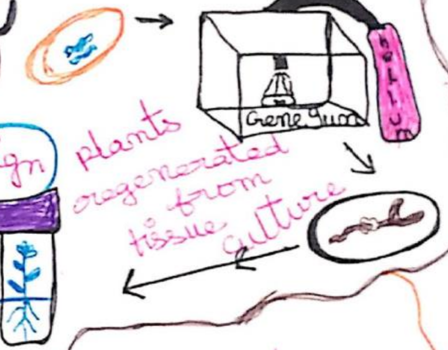
- natural gene Transfer
- Artificial gene Transfer
  - physical method → electroporation, Biolistic Transformat Gene Gun, protoplast fashion, micro Injection.
  - chemical method → Liposome mediated Gene transfer, Calcium phosphate mediated gene transfer, DEAE - Dextral mediated gene transfer, polyethylene glycol mediated gene transfer.

- Bacterial Transformation.
- Conjugation.
- phase transduction.
- Retroviral transduction.
- Transposition by transposons.
- A. Tumefaciens mediated.

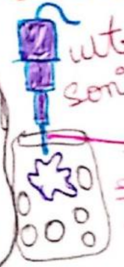
micro - Injection:



Gene gun:



ultrasonication :-

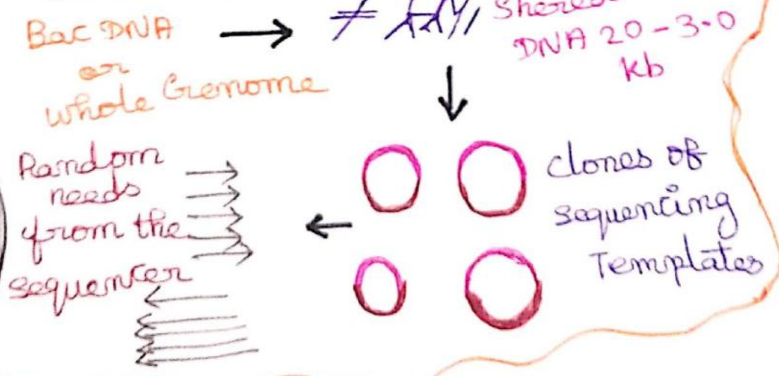


→ electric probe supplie sound energy which gives rise to unstable growth.

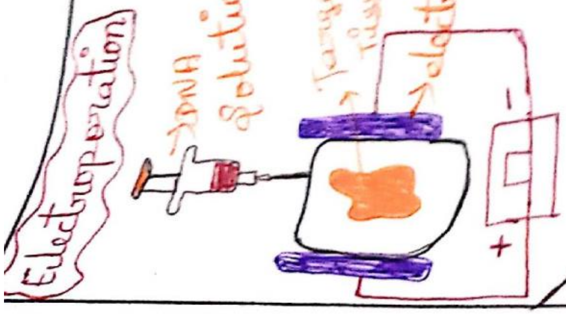
oil droplets

ultrasound causing cultivation.

shot gun :-


Bac DNA or whole Genome
Shered DNA 20 - 3.0 kb
clones of sequencing Templates
Random needs from the sequencer

Electroporation



GENE TRANSFER TECHNIQUES

GENOME

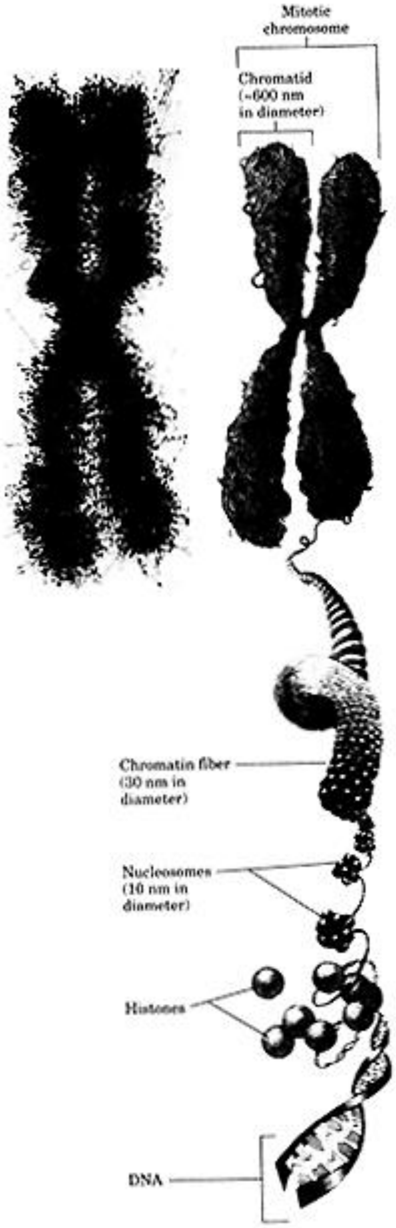
A genome is the complete set of genetic information of an organism. It contains all the instructions for creating and maintaining life. Every living organism consists of a genome.

A human genome consists of nuclear and mitochondrial DNA. On the contrary, the genome of a virus comprises RNA as the genetic material.

Our genome contains around 20,000 genes. They make up 1-5% of our genome. The DNA between the genes is involved in gene regulation.

Genome Organization

- The human haploid genome consists of about 3 x 10^9 base pairs of DNA.
- Genomic DNA exists as single linear pieces of DNA that are associated with a protein called a nucleoprotein complex.
- The DNA-protein complex is the basis for the formation of chromosomes; virtually all of the genomic DNA is distributed among the 23 chromosomes that reside in the cellular nucleus.
- A very small fraction of the genome is also found in a 16,000 base pair circular piece of DNA that is found in the mitochondria.
- The double helical DNA of the chromatin is replicated with the chromatin fiber condensing into discrete bodies, the chromosomes, each consisting of two identical chromatids.
- The two sister chromatids separate, one moving to each pole of the cell, where they become part of the newly formed nucleus of each daughter cell. The cells that make up most of the body of a multicellular organism, the somatic cells, have two copies of each chromosome and are said to be diploid (2n). Egg and sperm for example, produced by meiosis and having only one copy of each chromosome, are haptoid (n).
- The DNA of chromatin and chromosomes is bound tightly to a family of positively charged proteins, the histones, which associate strongly with the many negatively charged phosphate groups in DNA. The histones and DNA associate in complexes called nucleosomes in which the DNA strand winds around a core of histone molecules.


Mitotic chromosome
Chromatid (~600 nm in diameter)
Chromatin fiber (30 nm in diameter)
Nucleosomes (10 nm in diameter)
Histones
DNA

Functional Elements and Distribution of DNA within the Genome

- The major function of genomic DNA is to carry and store genetic information that is expressed as RNA and then as functional proteins.

- For gene expression to correctly occur there must be regulatory elements present on the genome and the genome must be faithfully replicated and segregated between daughter cells.

DNA Elements Required for Replication and Segregation of the Genome

- Based on studies with unicellular eukaryotes (yeast) at least three types of DNA elements are required for replication and stable inheritance of chromosomes: autonomously replicating sequences (ARS), centromeres and telomeres.

- Autonomously Replicating Sequences (ARS) are the sites at which DNA replication is initiated on the chromosomes.

- Centromeres are DNA sequences that are required for segregation of replicated chromosomes to daughter cells.

- Telomeres - Telomerase recognizes the tips of chromosomes also know as telomeres. The DNA sequences of telomeres have been determined in several organisms and consist of numerous repeats of a 6 to 8 base long sequence, [TTGGGG]n.

- Yeast Artificial Chromosomes or YAC's can be constructed by combining large segments of human DNA (50,000 base pairs or longer) with a selectable marker and the three essential elements described above. These artificial chromosomes can then be propagated and amplified in yeast cells. This technology is being used in the sequencing of the human genome.

Unique Sequences

- Greater than 50% of the eukaryotic genome consists of DNA that is unique in sequence and the human genome encodes for about 100,000 proteins.

- The average coding portions of a gene (the exons) consist of about 2,000 base pairs of DNA that is unique in sequence.

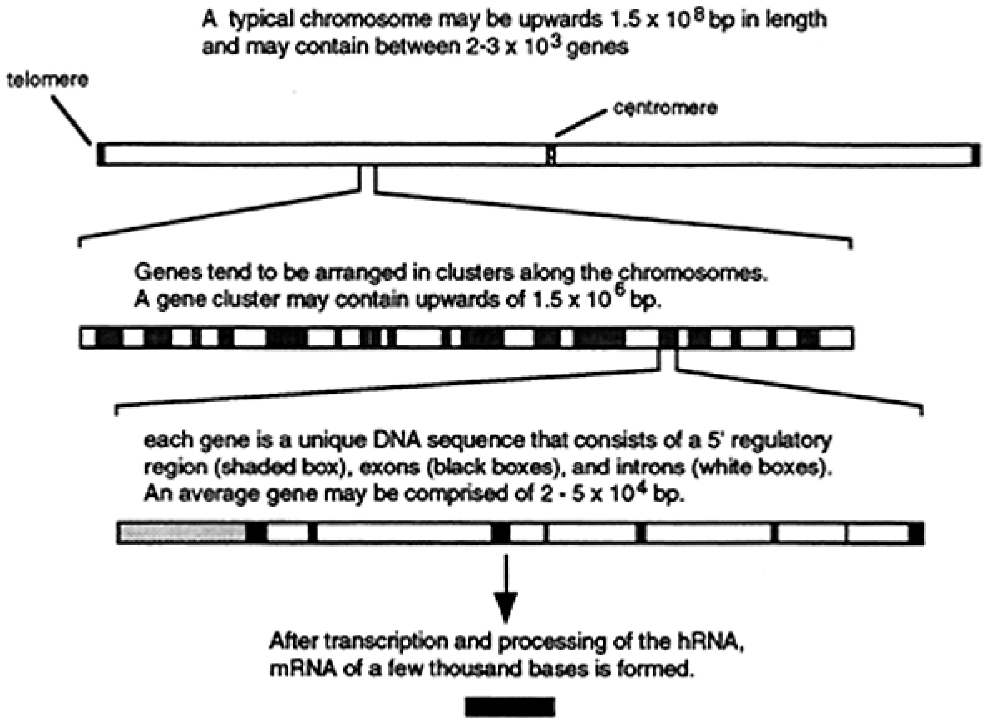
- This number represents less than 7% of the total DNA comprising the human genome and less than 14% of that DNA is unique.

- Most of the coding sequences are interrupted by from 1 to 50 noncoding sequences or introns.

- The total length of the introns that interrupt a gene generally far exceeds the total length of the exons.

- Since sequences that regulate gene expression also account for some of the unique sequences the actual amount of DNA coding for functional gene products is probably less than 3% of the total genomic DNA.

- The spatial distribution of genes, exons, introns and regulatory sequences along each chromosome is shown below.


A typical chromosome may be upwards 1.5 x 10^8 bp in length and may contain between 2-3 x 10^3 genes
telomere
centromere
Genes tend to be arranged in clusters along the chromosomes. A gene cluster may contain upwards of 1.5 x 10^6 bp.
each gene is a unique DNA sequence that consists of a 5' regulatory region (shaded box), exons (black boxes), and introns (white boxes). An average gene may be comprised of 2 - 5 x 10^4 bp.
After transcription and processing of the hRNA, mRNA of a few thousand bases is formed.

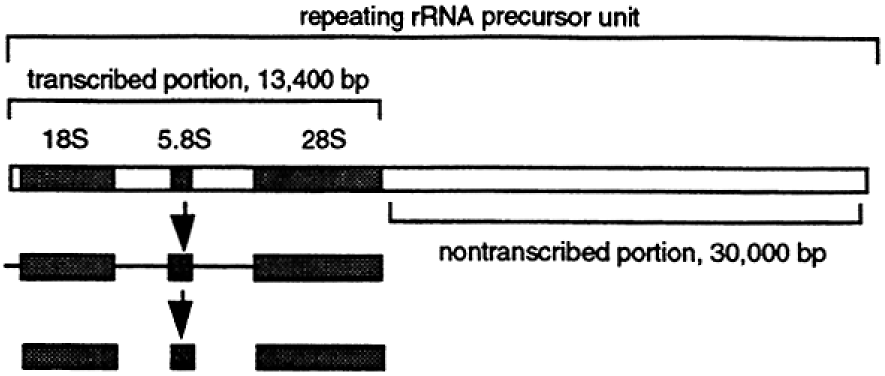
Repetitive Sequences

- There are multiple classes of repetitive DNA, two of these classes include: highly repetitive and moderately repetitive DNA.

- The function of repetitive DNA is not really known but approximately 30% of the human genome consists of repetitive DNA.

Highly Repetitive DNA consists of several different sets of short repeated polynucleotides, generally the repeats range from 5 to 500 base pairs in length and exist in tandem arrays. Highly repetitive DNA comprises about 10-15% of the total genomic DNA, is present in over a million copies and is transcriptionally inactive. Some of the highly repetitive DNA is clustered in structural regions of chromosomes particularly in the cetromeric and telomeric regions.

Moderately Repetitive DNA contains a large variety of repeated sequences ranging from a few hundred to tens of thousands of base pairs with different characteristics. Moderately repetitive DNA can be clustered at specific chromosomal locations or distributed throughout the genome. One type of moderately repetitive human DNA sequence is the rRNA precursor gene. Each rRNA precursor gene is contained in a DNA segment of about 43,000 base pairs. The actual transcript is 13,400 bases which is processed into the mature 28S, 18S and 5.8S rRNA's (see "RNA Synthesis and Processing" lecture). This means that at least 30,000 base pairs are not transcribed and apparently serve as spacer DNA. About 280 copies of the rRNA precursor gene are distributed in clusters on five chromosomes and account for about 0.4% of the genomic DNA.


repeating rRNA precursor unit
transcribed portion, 13,400 bp
18S
5.8S
28S
nontranscribed portion, 30,000 bp

Most types of moderately repetitive DNA are short about 300 base pairs in length, are interspersed with unique sequences, are often transcribed but do not code for gene product.

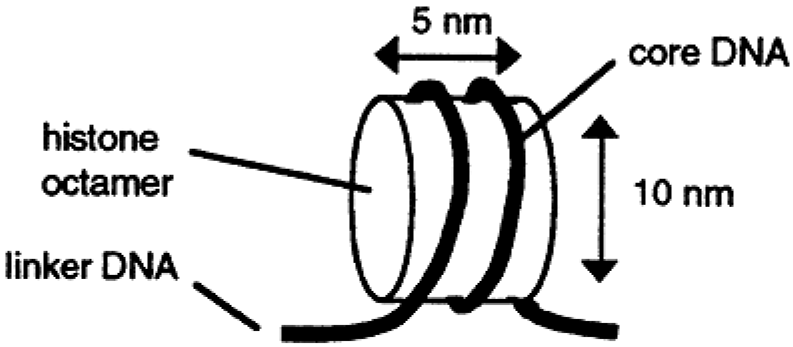
Chromosomal Structure

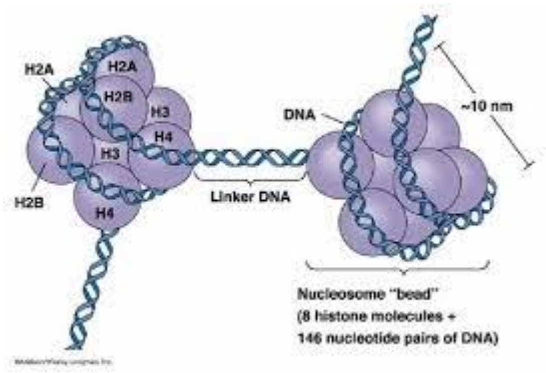
A typical human cellular nucleus is between 5 and 10 mM in diameter and the diploid human genome is over 2 meters long! Obviously to make the DNA fit into the nucleus it must be compacted, think of it as trying to put a piece of thread 6 miles long into a ping-pong ball. Fully compacted DNA can not be transcribed so consequently the cell must be able to selectively expose ARS elements so that replication can be initiated at the correct time in the cell cycle. In order to accomplish all of these tasks, compaction, transcription, replication the DNA is associated with a special set of structural proteins that form a nucleo- or DNA-protein complex called chromatin.

Composition and Structure of Chromatin

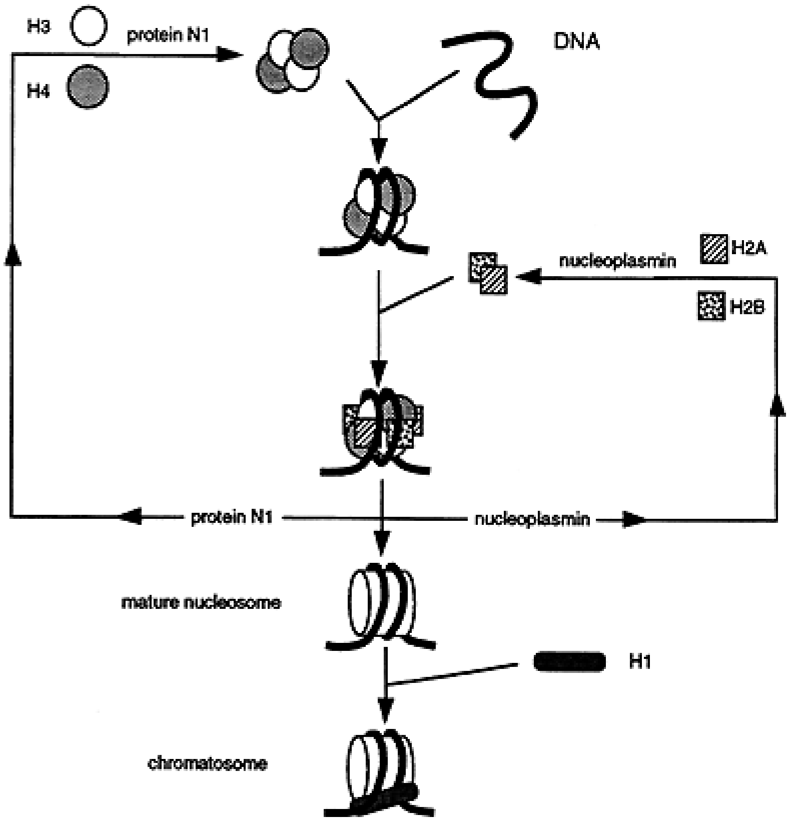
Chromatin contains two classes of protein: histones and nonhistone proteins. The overall purpose of histones is to condense the DNA though many nonhistone proteins are involved with transcription, DNA replication and maintenance of chromatin structure.

Histones are the most abundant proteins found in chromatin. There are five major types: H1, H2A, H2B, H3 and H4. The histones are small basic proteins composed mostly of Lys and Arg. The positive charge (basicity) of the histones allows the negatively charged DNA to "wrap" around it forming a nucleosome.





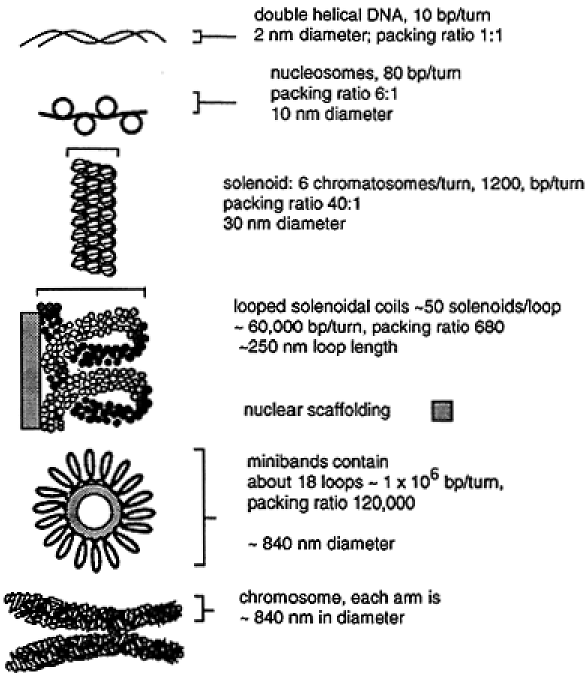
Chromatin consists of a linear chain of nucleosomes each linked to its neighbor by a segment of DNA that is between 20 and 100 base pairs in length. Nucleosomes that are bound to H1 are called chromatosomes. The assembly of nucelosomes is believed to require the participation of the nonhistone proteins, N1 and nucleoplasmin.



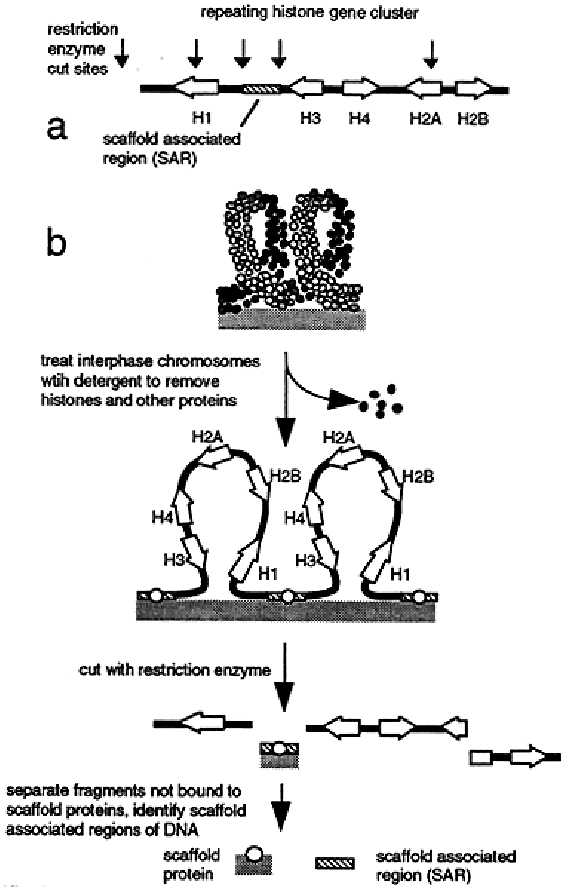
Nucleosome Assembly

The assembly of the nucleosome requires the nonhistone proteins N1, binds to a tetramer of H3 and H4, and nucleoplasmin which binds to dimers of H2A and H2B. The resulting H32H42 tetramer and H2AH2B dimers associate with the DNA while N1 and nucleoplasmin are released and recycled. H1 then adds to the structures forming a chromatosome.

Chromatin can be further compacted into higher order structures including a solenoidal coil with about six chromatosomes per turn and the resulting DNA fibril. The fibril forms loops anchored to a nonhistone protein scaffold, the looped structures forming the interphase chromosomes. During mitosis the looped structure further condense by coiling upon themselves to form minibands. Each miniband is comprised of about 18 loops, each loop containing over a million base pairs. The DNA in these minibands has been compacted by about 10,000 fold! The minibands are arranged along a central axis and form the arms of the mitotic chromosome.



Treatment of mitotic chromosomes with dextran sulfate followed by special detergents strips off the histones and most other proteins. Additional treatment with restriction enzymes cuts most of the DNA which can then be separated from the scaffold. When the scaffolding is then analyzed short segments of DNA are found attached to the scaffolding between genes, not within regions of transcribed DNA. These sequences are called scaffold associated regions or SAR's.

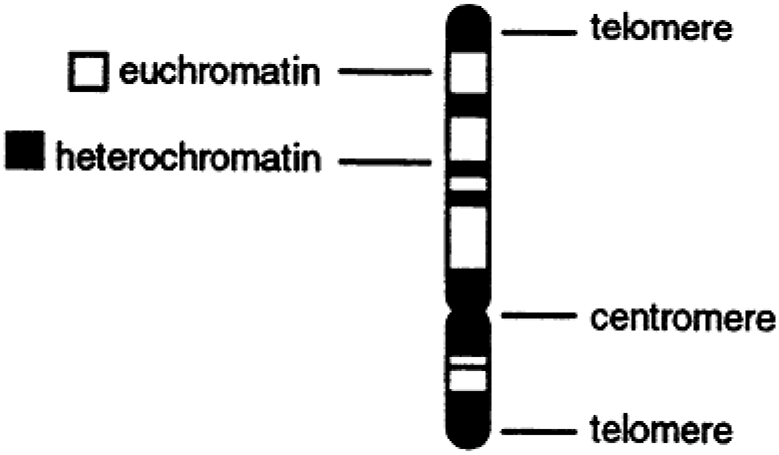


repeating histone gene cluster
restriction enzyme cut sites
a
H1
H3
H4
H2A
H2B
scaffold associated region (SAR)
b
treat interphase chromosomes wtih detergent to remove histones and other proteins
H2A
H2B
H4
H3
H1
H2A
H2B
H4
H3
H1
cut with restriction enzyme
separate fragments not bound to scaffold proteins, identify scaffold associated regions of DNA
scaffold protein
scaffold associated region (SAR)

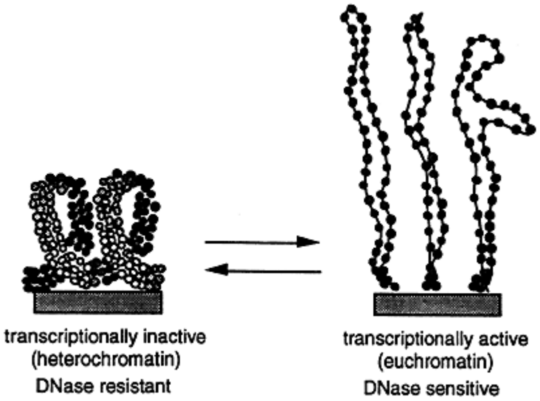
The major scaffold protein is topoisomerase II which regulates the extent of supercoiling in the DNA. Supercoiling, seen in circular DNA (mitochondrial DNA) and nucleosomes (DNA wrapped around something else), results when double stranded DNA twists upon itself. Topoisomerase II maintains the level of supercoiled DNA at a constant value because supercoiling can affect the efficiency of transcription, DNA replication and the integrity of chromatin.

Chromatin Dynamics

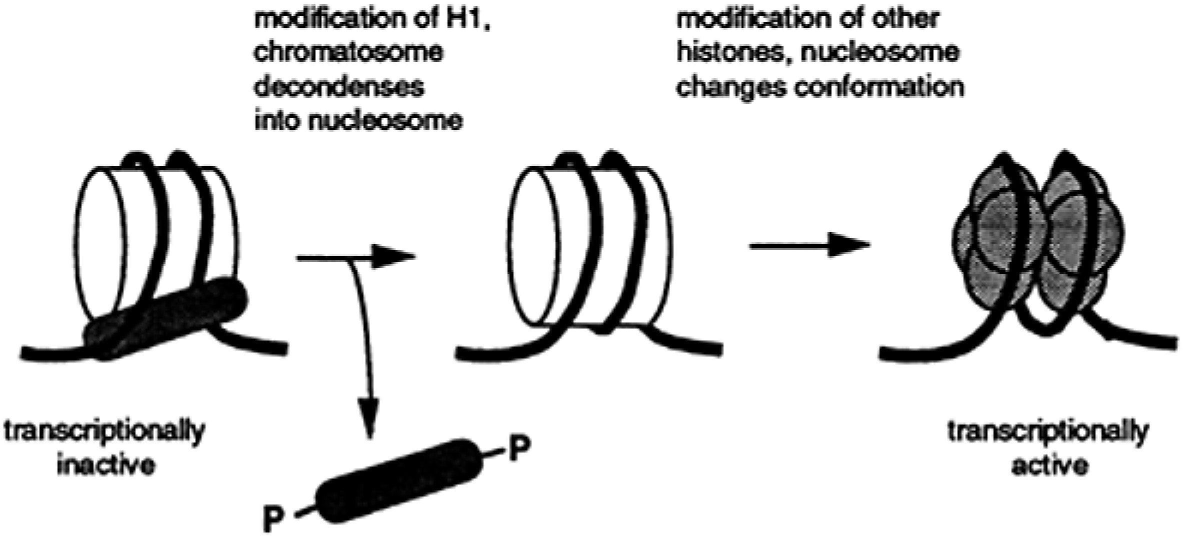
The higher order structure of chromatin varies and is determined by factors such as tissue type, sex and the developmental state of the cell. If chromosomes are stained with a dye and then analyzed microscopically numerous dark bands are seen. The dark bands correspond to the highly condensed and transcriptionally inactive heterochromatin. Heterochromatin is generally found at or near the centromere and telomeres and consists of highly repetitive DNA. The lighter bands are the less condensed, transcriptionally active euchromatin.

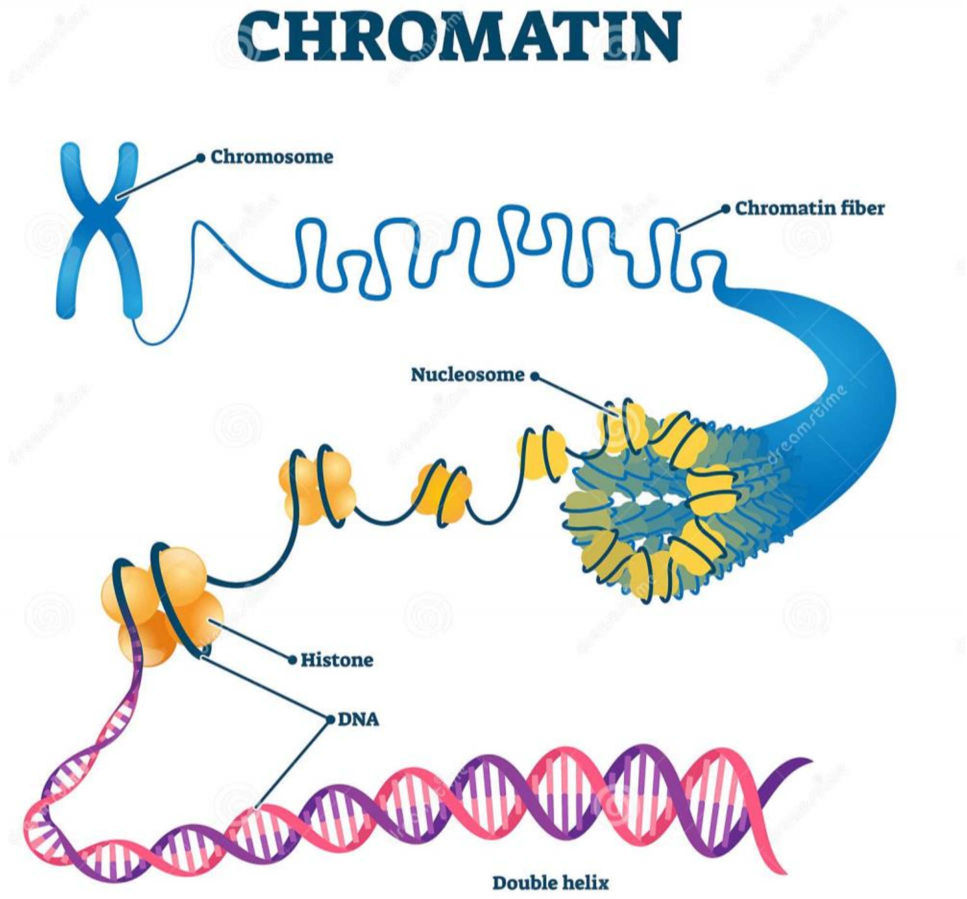


In order for DNA replication to occur the chromatin must be dynamically restructured or "decondensed" allowing the replication "machinery" to gain access to the DNA. Transcriptionally active genes are sensitive to digestion by DNase while inactive genes are insensitive to digestion. This suggests that the chromatin has "decondensed" during transcription which also allows access to the DNase.



Numerous subtypes of histones have been identified. Analysis of these histones indicates that histones are subject to chemical modification via: acylation, phosphorylation, ADP-ribosylation and ubiquination. Some of these modified histones appear to be associated with actively transcribing genes suggesting that the modifications may affect the structure of the nucleosome making the DNA more accessible to the enzymes required for regulating and carrying out transcription, replication and repair.


modification of H1, chromatosome decondenses into nucleosome
modification of other histones, nucleosome changes conformation
transcriptionally inactive
P
P
transcriptionally active



GENOME SEQUENCING METHODS

What is a Genome?

A genome is the complete set of genetic information of an organism. It contains all the instructions for creating and maintaining life. Every living organism consists of a genome.

A human genome consists of nuclear and mitochondrial DNA. On the contrary, the genome of a virus comprises RNA as the genetic material.

Our genome contains around 20,000 genes. They make up 1-5% of our genome. The DNA between the genes is involved in gene regulation.

Genome Sequencing

Genome is a unique sequence of DNA. It is sequenced by certain machines to identify the cause of a particular disease. Some diseases are caused by very little variation in the DNA. Sequencing the genome can help us identify which DNA changes are causing the problem.

The genome of the tumour cells is altered when compared to normal cells. By comparing the genome of the normal and cancer cells we can get clues about ways to treat cancer.

The sequencing of a human genome takes about a day. However, its analysis takes a longer time.

DNA Sequencing

“DNA sequencing is a process of determining or identifying the order of nucleotides present in a DNA sequence.”

Nitrogenous bases, sugar and phosphate are three ingredients of the DNA in which Adenine, Thymine, Cytosine and Guanine are bases. A functional piece of DNA is known as a gene that encodes proteins.

For understanding the structure and function of a gene, It is very important to study its nucleotide sequence. DNA sequencing serves this purpose. In a sequential manner, the long chain of DNA is read by the sequencer machine.

History of DNA sequencing:

The story of DNA begins when *Watson* and *Crick* discovered the structure of DNA in the year 1953. In 1964, *Richard Holley* who performed the sequencing of the tRNA was the first attempt to sequence the nucleic acid.

Using the technique of Holley and *Walter Fieser*, they sequenced the genome of bacteriophage MS2 (RNA sequencing). The sequenced molecules were RNA; Yet DNA sequencing was not performed.

In the year 1977, *Fredrick Sanger* postulated the first method for sequencing the DNA, named a **chain termination method**.

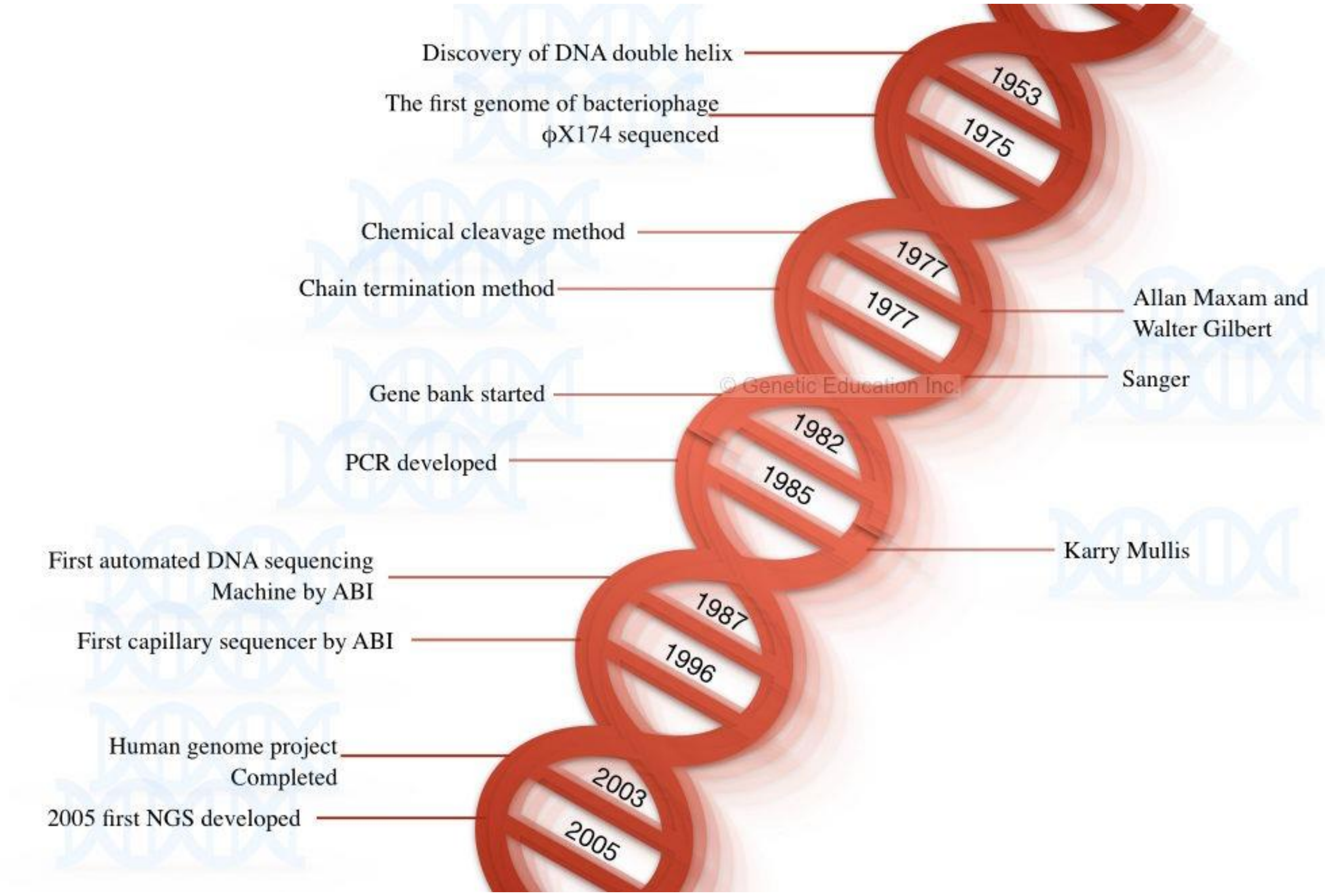
In the same year, the chemical method of DNA sequencing was explained by *Allan Maxam* and *Walter Gilbert*. The genome of bacteriophage X174 was sequenced in the same year using the chemical degradation method.

Because of the lack of automation, Both the methods (chemical degradation and chain termination) were tedious and time-consuming. The first semi-automated DNA method was developed by *Lorey* and *Smith* in the year 1986.

Later on, in the year 1987, the Applied Biosystem had developed a fully automated machine-controlled DNA sequencing method. After the development of fully automated machines, the era of the 2000s become a golden period for sequencing platforms.

Furthermore, in 1996, Applied Biosystem developed another innovative sequencing platform known as capillary DNA sequencing. After that, the human genome project was completed by using the combination of these methods in the year 2003.

A fast, accurate, reliable, and highly efficient next-generation sequencing platform was postulated in the year 2005 by Solexa/Illumina. Some of the milestone into the DNA sequencing is shown in the figure below,


Discovery of DNA double helix
1953
The first genome of bacteriophage ϕX174 sequenced
1975
Chemical cleavage method
1977
Chain termination method
1977
Allan Maxam and Walter Gilbert
Sanger
© Genetic Education Inc.
Gene bank started
1982
PCR developed
1985
Karry Mullis
First automated DNA sequencing Machine by ABI
1987
First capillary sequencer by ABI
1996
Human genome project Completed
2003
2005 first NGS developed
2005

What are the steps in DNA sequencing?

The steps mentioned below are the generalized representation of DNA sequencing; it may vary from platform to platform.

- Sample preparation (DNA extraction)
- PCR amplification of target sequence
- Amplicons purification
- Sequencing pre-prep
- DNA Sequencing
- Data analysis

Different methods of DNA sequencing:

Various methods of DNA sequencing are explained here.

- Maxam and Gilbert method
- Chain termination method
- semiautomated method
- automated method
- Pyrosequencing
- The whole-genome shotgun sequencing method
- Clone by the clone sequencing method
- Next-generation sequencing method

Two main methods are widely known to be used to sequence DNA:

1. **The Chemical Method** (also called the Maxam–Gilbert method after its inventors).
2. **The Chain Termination Method** (also known as the Sanger dideoxy method after its inventor).

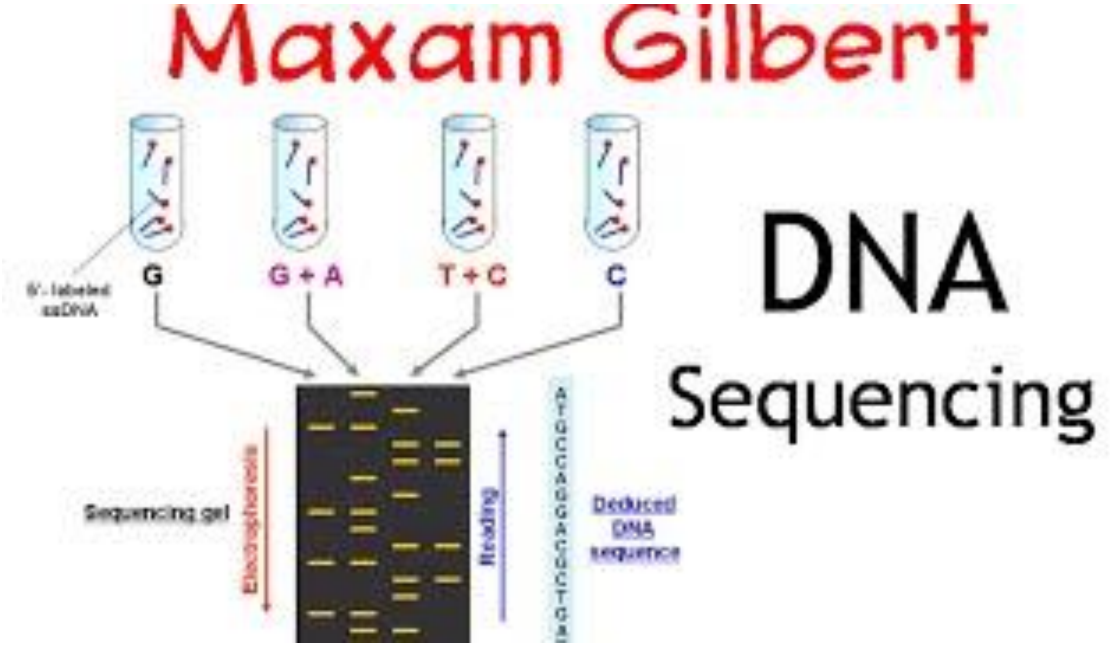
Chemical Cleavage Method (Maxam–Gilbert Method)

- In 1976-1977, Allan Maxam and Walter Gilbert developed a DNA sequencing method based on chemical modification of DNA and subsequent cleavage at specific bases.
- The method requires radioactive labeling at one end and purification of the DNA fragment to be sequenced.
- Obviously, DNA extraction is the very first step. After that, the DNA is denatured using the heat denaturation method and single-stranded DNA is generated.
- The phosphate (5’ P) end of the DNA is removed and labeled by the radiolabeled P32. The enzyme named phosphatase removes the phosphate from the DNA and simultaneously, the kinase adds the 32P to the 5’ end of it.
- 4 different chemicals are used to cleave DNA at four different positions; hydrazine and hydrazine NaCl are selectively attack pyrimidine nucleotides while dimethyl sulfate and piperidine attack purine nucleotides.
    - Hydrazine: T + C
    - Hydrazine NaCl: C
    - Dimethyl sulfate: A + G
    - Piperidine: G

- An equal volume of 4 different ssDNA samples is taken into 4 different tubes each containing 4 different chemicals. The samples are incubated for some time and electrophoresed in polyacrylamide gel electrophoresis. The results of the chemicals cleavage of four different tubes are shown in the figure below.
- Autoradiography is used to visualize the separation of DNA fragments. Due to the radiolabelled 32P end of the DNA, the DNA bands are visualized through autoradiography.

Key Features

- Base-specific cleavage of DNA by certain chemicals
- Four different chemicals, one for each base
- A set of DNA fragments of different sizes
- DNA fragments contain up to 500 nucleotides



Sanger Sequencing Steps & Method

What Is Sanger Sequencing?

Sanger sequencing, also known as the “chain termination method”, is a method for determining the nucleotide sequence of DNA. The method was developed by two time Nobel Laureate Frederick Sanger and his colleagues in 1977, hence the name the Sanger Sequence.

How Does Sanger Sequencing Work?

Sanger sequencing can be performed manually or, more commonly, in an automated fashion via sequencing machine (**Figure 1**). Each method follows three basic steps, as described below.

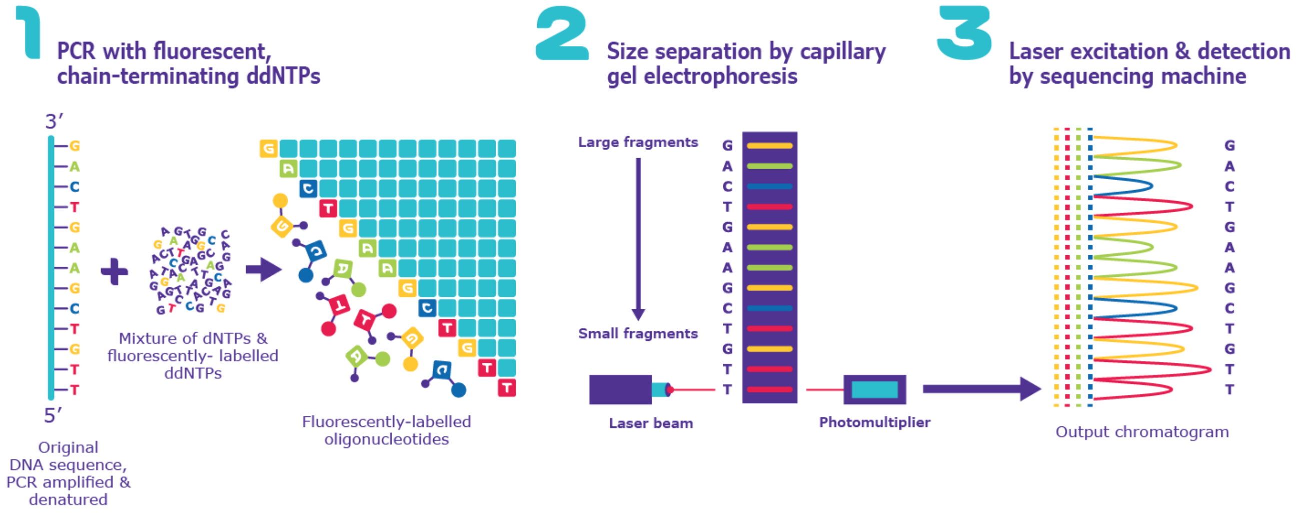

1
PCR with fluorescent, chain-terminating ddNTPs
3'
G A C T G A A G C T G T T
5'
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
2
Size separation by capillary gel electrophoresis
Large fragments
Small fragments
G A C T G A A G C T G T T
Laser beam
Photomultiplier
3
Laser excitation & detection by sequencing machine
G A C T G A A G C T G T T
Output chromatogram

Figure 1. Three Basic Steps of Automated Sanger Sequencing.

SANGER SEQUENCING STEPS

There are three main steps to Sanger sequencing.

1. DNA SEQUENCE FOR CHAIN TERMINATION PCR

The DNA sequence of interest is used as a template for a special type of PCR called chain-termination PCR. Chain-termination PCR works just like standard PCR, but with one major difference: the addition of modified nucleotides (dNTPs) called dideoxyribonucleotides (ddNTPs). In the extension step of standard PCR, DNA polymerase adds dNTPs to a growing DNA strand by catalyzing the formation of a phosphodiester bond between the free 3'-OH group of the last nucleotide and the 5'-phosphate of the next (Figure 2).

In chain-termination PCR, the user mixes a low ratio of chain-terminating ddNTPs in with the normal dNTPs in the PCR reaction. ddNTPs lack the 3'-OH group required for phosphodiester bond formation; therefore, when DNA polymerase incorporates a ddNTP at random, extension ceases. The result of chain-termination PCR is millions to billions of oligonucleotide copies of the DNA sequence of interest, terminated at a random lengths (n) by 5'-ddNTPs.

In **manual** Sanger sequencing, four PCR reactions are set up, each with only a single type of ddNTP (ddATP, ddTTP, ddGTP, and ddCTP) mixed in.

In **automated** Sanger sequencing, all ddNTPs are mixed in a single reaction, and each of the four dNTPs has a unique fluorescent label.

2. SIZE SEPARATION BY GEL ELECTROPHORESIS

In the second step, the chain-terminated oligonucleotides are separated by size via gel electrophoresis. In gel electrophoresis, DNA samples are loaded into one end of a gel matrix, and an electric current is applied; DNA is negatively charged, so the oligonucleotides will be pulled toward the positive electrode on the opposite side of the gel. Because all DNA fragments have the same charge per unit of mass, the speed at which the oligonucleotides move will be determined only by size. The smaller a fragment is, the less friction it will experience as it moves through the gel, and the faster it will move. In result, the oligonucleotides will be arranged from smallest to largest, reading the gel from bottom to top.

In **manual** Sanger sequencing, the oligonucleotides from each of the four PCR reactions are run in four separate lanes of a gel. This allows the user to know which oligonucleotides correspond to each ddNTP.

In **automated** Sanger sequencing, all oligonucleotides are run in a single capillary gel electrophoresis within the sequencing machine.

3. GEL ANALYSIS & DETERMINATION OF DNA SEQUENCE

The last step simply involves reading the gel to determine the sequence of the input DNA. Because DNA polymerase only synthesizes DNA in the 5’ to 3’ direction starting at a provided primer, each terminal ddNTP will correspond to a specific nucleotide in the original sequence (e.g., the shortest fragment must terminate at the first nucleotide from the 5’ end, the second-shortest fragment must terminate at the second nucleotide from the 5’ end, etc.) Therefore, by reading the gel bands from smallest to largest, we can determine the 5’ to 3’ sequence of the original DNA strand.

In **manual** Sanger sequencing, the user reads all four lanes of the gel at once, moving bottom to top, using the lane to determine the identity of the terminal ddNTP for each band. For example, if the bottom band is found in the column corresponding to ddGTP, then the smallest PCR fragment terminates with ddGTP, and the first nucleotide from the 5’ end of the original sequence has a guanine (G) base.

In **automated** Sanger sequencing, a computer reads each band of the capillary gel, in order, using fluorescence to call the identity of each terminal ddNTP. In short, a laser excites the fluorescent tags in each band, and a computer detects the resulting light emitted. Because each of the four ddNTPs is tagged with a different fluorescent label, the light emitted can be directly tied to the identity of the terminal ddNTP. The output is called a chromatogram, which shows the fluorescent peak of each nucleotide along the length of the template DNA.

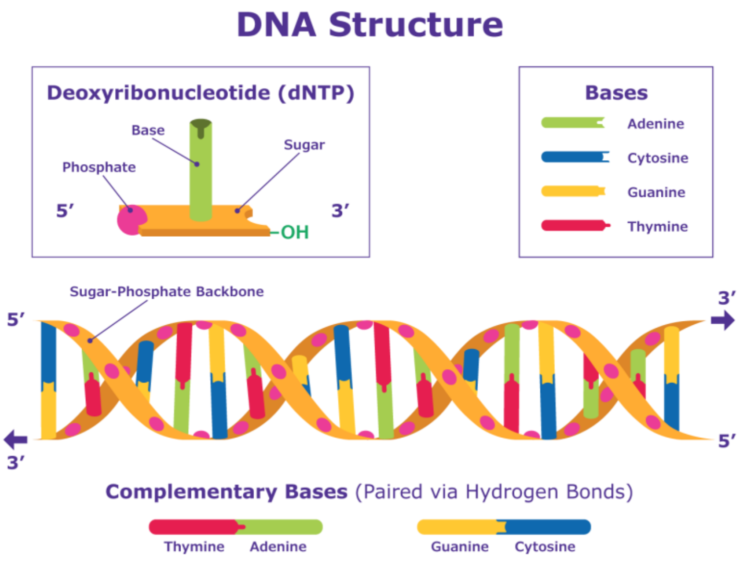


Figure 2.DNA Structure Schematic. DNA is a molecule composed of two strands that coil around each other to form a double helix. Each strand is made up of a string of molecules called deoxyribonucleotides (dNTPs).

Each dNTP contains a phosphate group, a sugar group, and one of four nitrogenous bases [adenine (A),thymine (T), guanine (G), or cytosine (C)]. The dNTPs are strung together in a linear fashion by phosphodiester covalent bonds between the sugar of one dNTP and the phosphate group of the next; this repeated sugar-phosphate pattern makes up the sugar-phosphate backbone.

The nitrogenous bases of the two separate strands are bound together by hydrogen bonds between complementary bases to form the double-stranded DNA helix.

How to Read Sanger Sequencing Results

Reading the Sanger sequencing results properly will depend on which of the two complementary DNA strands is of interest and what primer is available. If the two strands of DNA are A and B and strand A is of interest, but the primer is better for strand B, the output fragments will be identical to strand A. On the other hand, if strand A is of interest and the primer is better for strand A, then the output will be identical to strand B. Accordingly, the output must be converted back to strand A.

So, if the sequence of interest reads “TACG” and the primer is best for that strand, the output will be “ATGC” and, therefore, must be converted back to “TACG”. However, if the primer is better for the complementary strand (“ATGC”), then the output will be “TACG”, which is the correct sequence.

In short, before starting, you need to know what you’re targeting and how you’re going to get there! So keeping this in mind, here is an example of the former example (TACG -> ATGC -> TACG). If the dideoxynucleotides labels are T = yellow, A = pink, C = dark blue, and G = light blue, you will end up with the short sequences primer-A, primer-AT, primer-ATG, and primer-ATGC. Once the fragments have been separated by electrophoresis, the laser will read the fragments in order of length (pink, yellow, light blue, and dark blue) and produce a chromatogram. The computer will convert the letters, so the final sequence is the correct TACG.

SHOTGUN GENOME SEQUENCING METHODS

Shotgun sequencing involves randomly breaking up DNA sequences into lots of small pieces and then reassembling the sequence by looking for regions of overlap.

- Large, mammalian genomes are particularly difficult to clone sequence and assemble because of their size and structural complexity. As a result clone-by-clone sequencing, although reliable and methodical, takes a very long time. With the emergence of cheaper sequencing and more sophisticated computer programs, researchers have therefore relied on whole genome shotgun sequencing to tackle larger, more complex genomes.

- Shotgun sequencing was originally used by Fred Sanger and his colleagues to sequence small genomes such as those of viruses? and bacteria?.

- Whole genome shotgun sequencing bypasses the time-consuming mapping and cloning steps that make clone-by-clone sequencing so slow.

- In whole genome shotgun sequencing the entire genome is broken up into small fragments of DNA? for sequencing.

- These fragments are often of varying sizes, ranging from 2-20 kilobases? (2,000-20,000 base pairs?) to 200-300 kilobases (200,000-300,000 base pairs).

- These fragments are sequenced to determine the order of the DNA bases?, A, C, G and T.

- The sequenced fragments are then assembled together by computer programs that find where fragments overlap.

- You can imagine shotgun sequencing as being a bit like shredding multiple copies of a book (which in this case is a genome), mixing up all the fragments and then reassembling the original text (genome) by finding fragments with text that overlap and piecing the book back together again.

- This method of genome sequencing was used by Craig Venter, founder of the private company Celera Genomics, to sequence the human genome. Venter wanted to sequence the human genome faster than the publicly funded effort and felt this was the best way. To assemble the sequence Venter used the clone-by-clone publically available data from the Human Genome Project.

- Now, as technologies are improving, whole genome shotgun sequencing is being used to improve the accuracy of existing genome sequences, such as the reference human genome.

- It is used to remove errors, fill in gaps or correct parts of the sequence that were originally assembled incorrectly when clone-by-clone sequencing was used.

- As a consequence the reference human genome is constantly being improved to ensure that the genome sequence is of the highest possible standard.

What are the advantages of shotgun sequencing?

- By removing the mapping stages, whole genome shotgun sequencing is a much faster process than clone-by-clone sequencing.

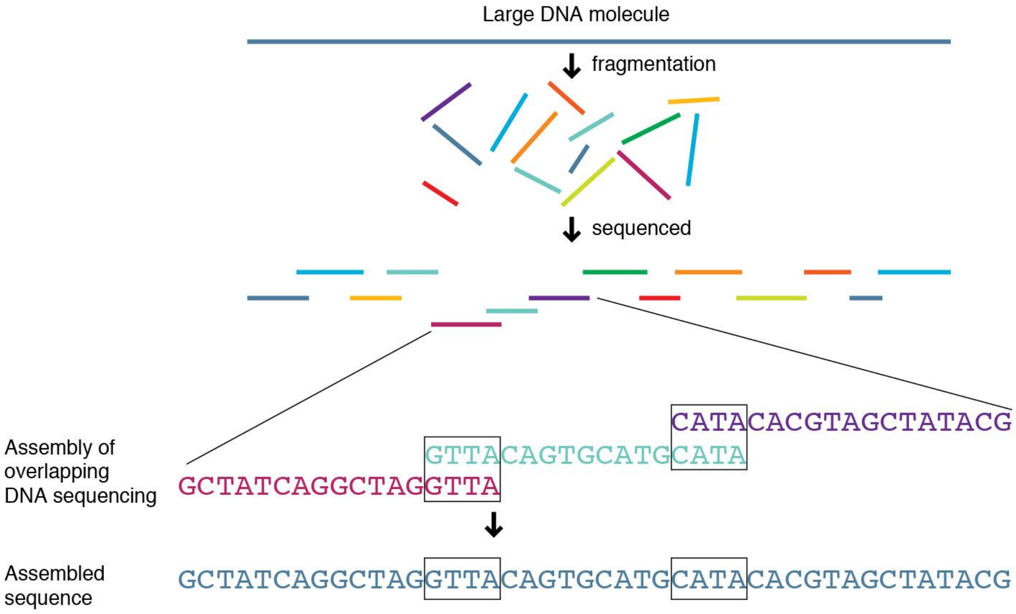
- Whole genome shotgun sequencing uses a fraction of the DNA that clone-by-clone sequencing needs.

- Whole genome shotgun sequencing is particularly efficient if there is an existing reference sequence. It is much easier to assemble the genome sequence by aligning it to an existing reference genome?.

- Shotgun sequencing is much faster and less expensive than methods requiring a genetic map.

What are the disadvantages of shotgun sequencing?

- Vast amounts of computing power and sophisticated software are required to assemble shotgun sequences together. To sequence the genome from a mammal (billions of bases long), you need about 60 million individual DNA sequence reads.
- Errors in assembly are more likely to be made because a genetic map is not used. However these errors are generally easier to resolve than in other methods and minimised if a reference genome can be used.
- Whole genome shotgun sequencing can only really be carried out if a reference genome is already available; otherwise assembly is very difficult without an existing genome to match it to.
- Whole genome shotgun sequencing can also lead to errors which need to be resolved by other, more labour-intensive types of sequencing, such as clone-by-clone sequencing.
- Repetitive genomes and sequences can be more difficult to assemble.



NEXT GENERATION SEQUENCING TECHNOLOGIES

- Next Generation Sequencing (NGS) is a powerful platform that has enabled the sequencing of thousands to millions of DNA molecules simultaneously.
- Next-generation sequencing (NGS), also known as high-throughput sequencing, is the catch-all term used to describe a number of different modern sequencing technologies.
- The high demand for low-cost sequencing has driven the development of high-throughput sequencing which produce thousands or millions of sequences at once.
- They are intended to lower the cost of DNA sequencing beyond what is possible with standard dye-terminator methods.

- Thus, these recent technologies allow us to sequence DNA and RNA much more quickly and cheaply than the previously used Sanger sequencing, and as such have revolutionized the study of genomics and molecular biology.

Classified to different generations, NGS has led to overcome the limitations of conventional DNA sequencing methods and has found usage in a wide range of molecular biology applications.

The generations it is classified into include:

First Generation

- Sanger Sequencing

Second Generation Sequencing

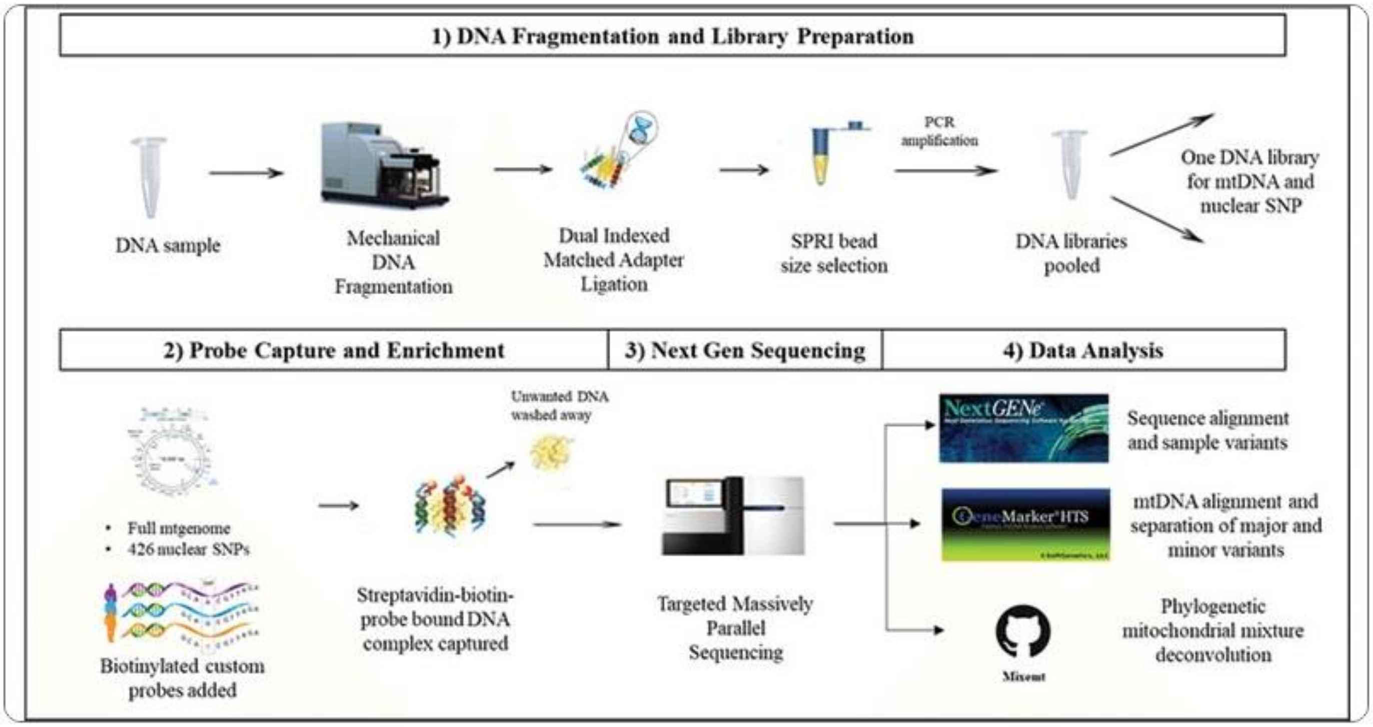
- Pyrosequencing
- Sequencing by Reversible Terminator Chemistry
- Sequencing by Ligation

Third Generation Sequencing

- Single Molecule Fluorescent Sequencing
- Single Molecule Real Time Sequencing
- Semiconductor Sequencing
- Nanopore Sequencing

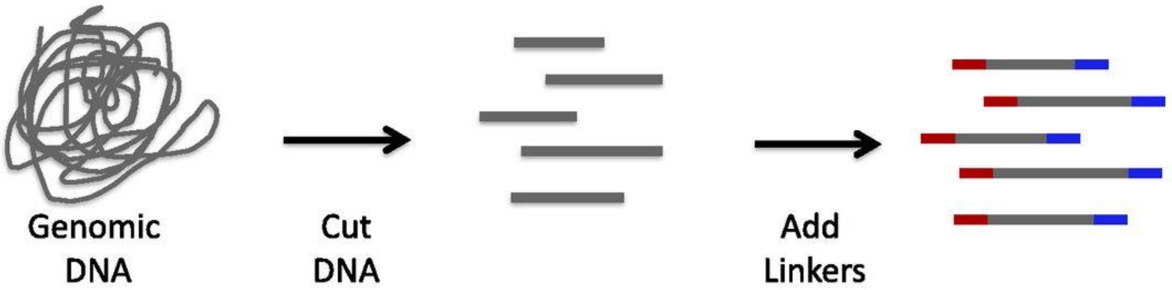
Fourth Generation Sequencing

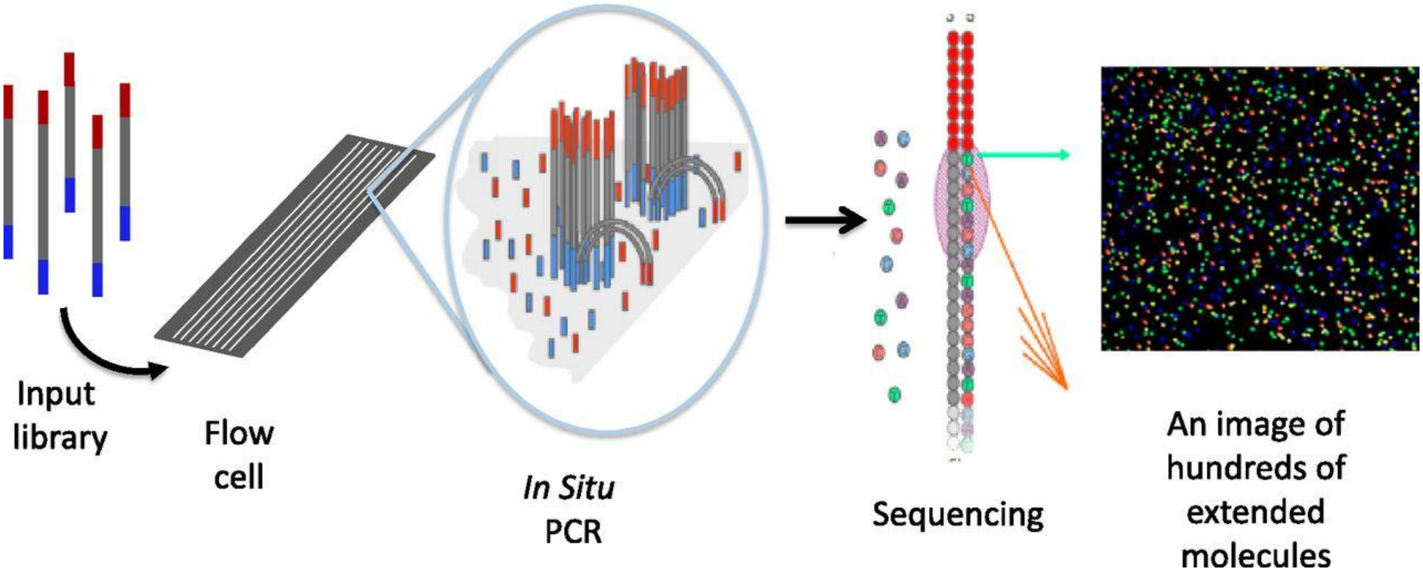
Aims conducting genomic analysis directly in the cell.


1) DNA Fragmentation and Library Preparation
DNA sample
Mechanical DNA Fragmentation
Dual Indexed Matched Adapter Ligation
SPRI bead size selection
PCR amplification
DNA libraries pooled
One DNA library for mtDNA and nuclear SNP
2) Probe Capture and Enrichment
3) Next Gen Sequencing
4) Data Analysis
Full mtgenome
426 nuclear SNPs
Biotinylated custom probes added
Unwanted DNA washed away
Streptavidin-biotin-probe bound DNA complex captured
Targeted Massively Parallel Sequencing
NextGENe
Sequence alignment and sample variants
GeneMarker HTS
mtDNA alignment and separation of major and minor variants
Mixemt
Phylogenetic mitochondrial mixture deconvolution

Important Next Generation Sequencing Techniques

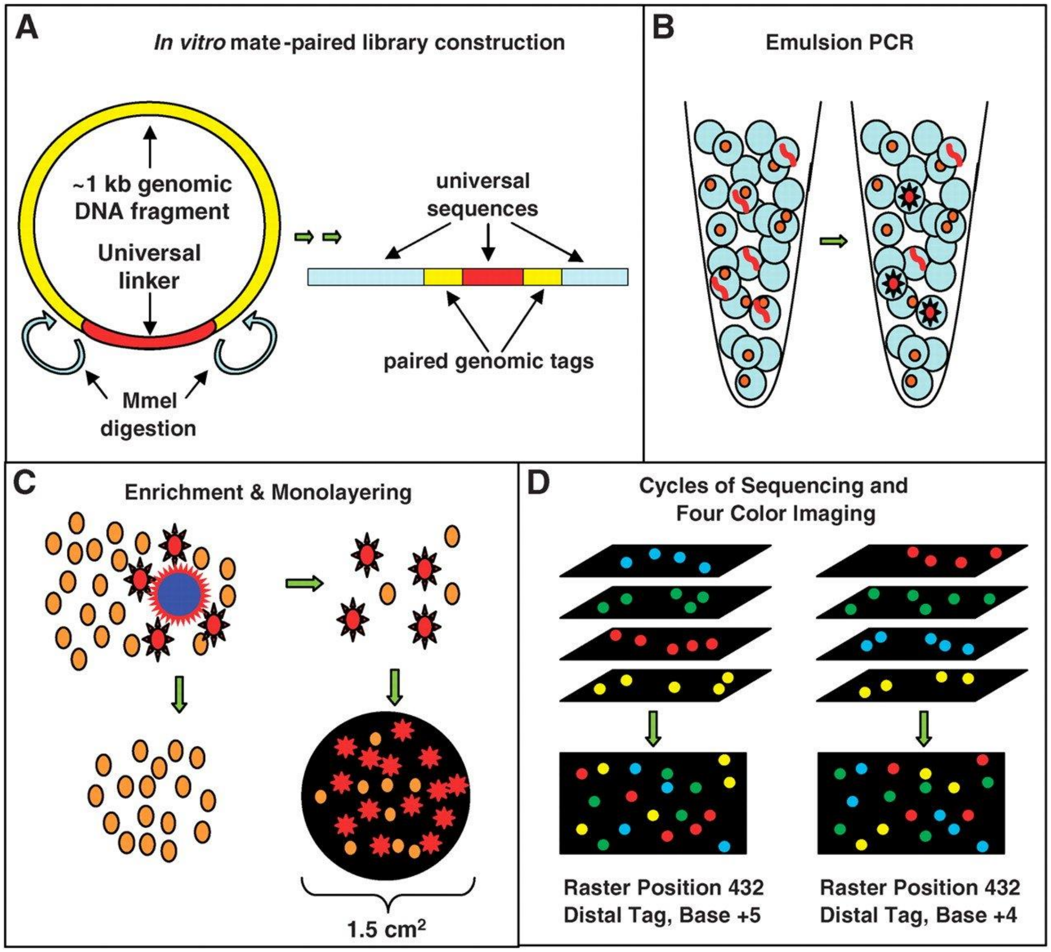
Lynx therapeutics' massively parallel signature sequencing (MPSS)





- It is considered as the first of the “next-generation” sequencing technologies.
- MPSS was developed in the 1990s at Lynx Therapeutics, a company founded in 1992 by Sydney Brenner and Sam Eletr.
- MPSS is an ultra high throughput sequencing technology.
- When applied to expression profile, it reveals almost every transcript in the sample and provide its accurate expression level.
- MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides; this method made it susceptible to sequence-specific bias or loss of specific sequences.
- However, the essential properties of the MPSS output were typical of later “next-gen” data types, including hundreds of thousands of short DNA sequences.
- In the case of MPSS, these were typically used for sequencing cDNA for measurements of gene expression levels.

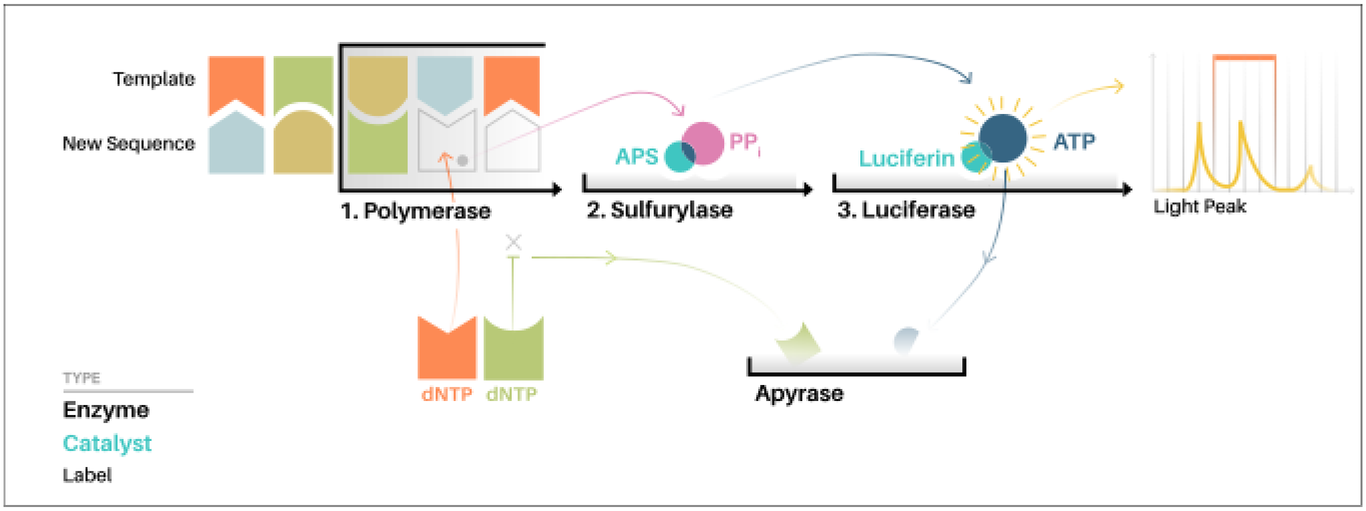
Polony sequencing


A
In vitro mate-paired library construction
~1 kb genomic DNA fragment
Universal linker
MmeI digestion
universal sequences
paired genomic tags
B
Emulsion PCR
C
Enrichment & Monolayering
1.5 cm^2
D
Cycles of Sequencing and Four Color Imaging
Raster Position 432 Distal Tag, Base +5
Raster Position 432 Distal Tag, Base +4

- It is an inexpensive but highly accurate multiplex sequencing technique that can be used to read millions of immobilized DNA sequences in parallel.
- This technique was first developed by Dr. George Church in Harvard Medical college.

- It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence an E. coli genome at an accuracy of > 99.9999% and a cost approximately 1/10 that of Sanger sequencing.

Pyro sequencing



- A parallelized version of pyrosequencing was developed by 454 Life Sciences, which has since been acquired by Roche Diagnostics.

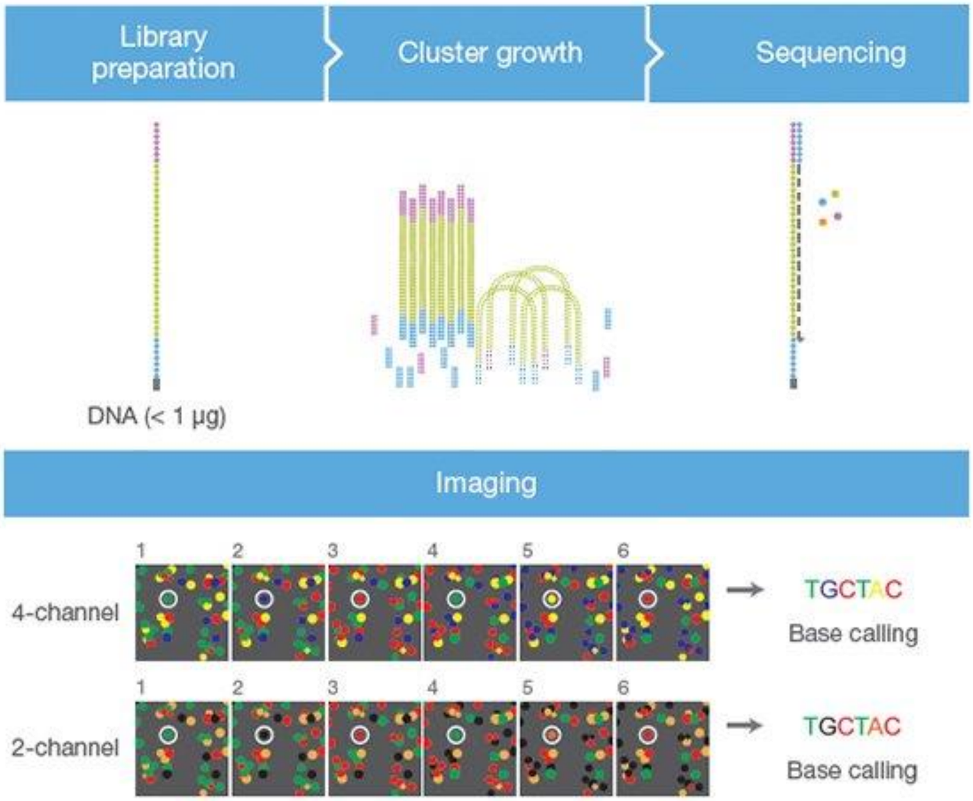
- The method amplifies DNA inside water droplets in an oil solution (emulsion PCR), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony.

- The sequencing machine contains many picolitre-volume wells each containing a single bead and sequencing enzymes.

- Pyrosequencing uses luciferase to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs.

- This technology provides intermediate read length and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other.

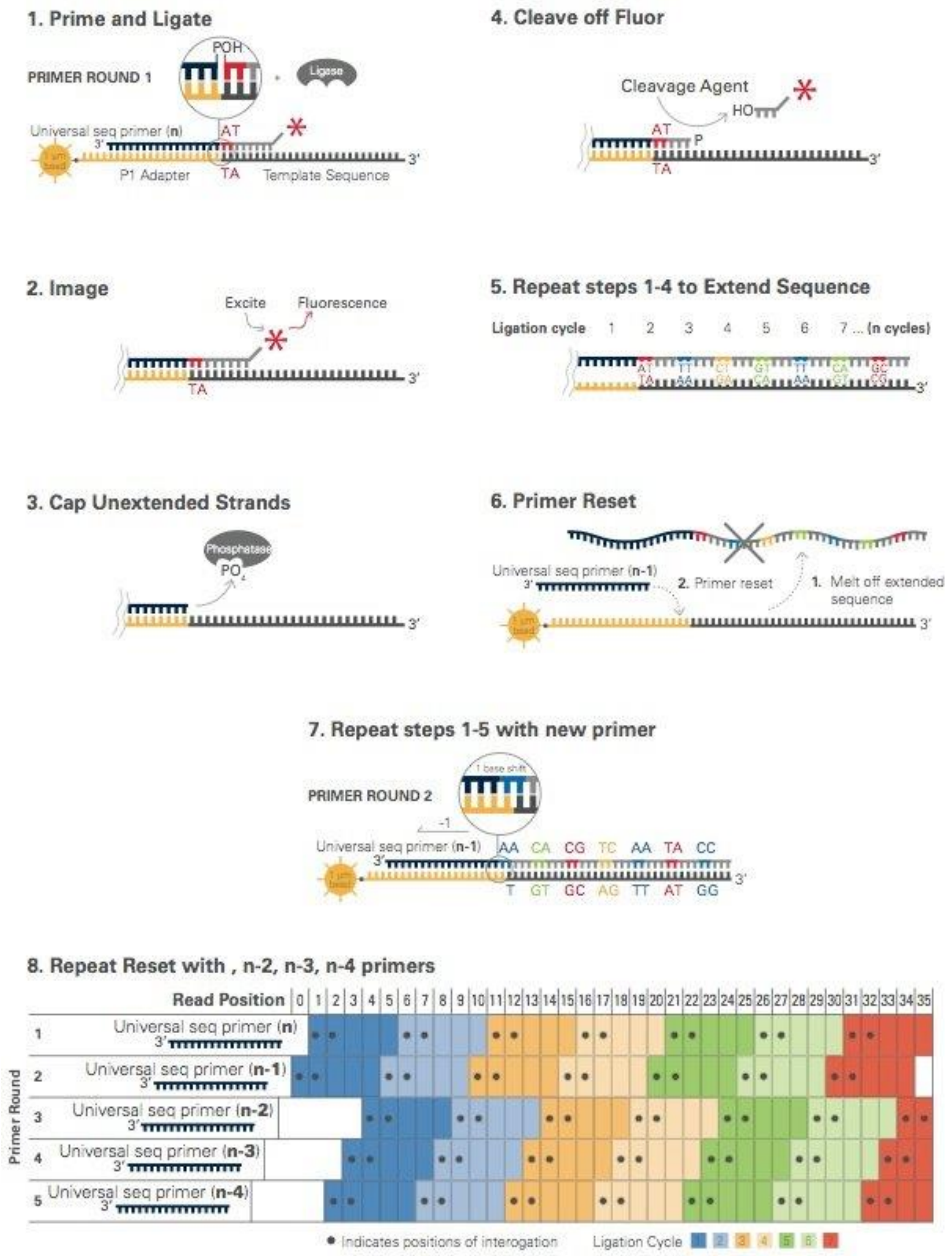
Illumina (Solexa) sequencing



- Solexa developed a sequencing technology based on dye terminators.
- In this method, DNA molecule are first attached to primers on a slide and amplified. This is known as bridge amplification.
- Unlike pyrosequencing, the DNA can only be extended one nucleotide at a time.
- A camera takes images of the fluorescently labeled nucleotides, then the dye along with the terminal 3′ blocker is chemically removed from the DNA, allowing the next cycle to commence.

SOLiD sequencing

- The technology for sequencing used in ABISolid sequencing is oligonucleotide ligation and detection.
- In this, a pool of all possible oligonucleotides of fixed length are labelled according to the sequenced position.
- This sequencing results to the sequences of quantities and lengths comparable to illumine sequencing.


1. Prime and Ligate
PRIMER ROUND 1
POH
Ligase
Universal seq primer (n)
3'
AT
TA
1 µm bead
P1 Adapter
Template Sequence
3'
4. Cleave off Fluor
Cleavage Agent
HO
P
AT
TA
3'
2. Image
Excite
Fluorescence
TA
3'
5. Repeat steps 1-4 to Extend Sequence
Ligation cycle 1 2 3 4 5 6 7 ... (n cycles)
3'
3. Cap Unextended Strands
Phosphatase
PO4
3'
6. Primer Reset
Universal seq primer (n-1)
3'
2. Primer reset
1. Melt off extended sequence
3'
7. Repeat steps 1-5 with new primer
PRIMER ROUND 2
1 base shift
-1
Universal seq primer (n-1)
AA CA CG TC AA TA CC
T GT GC AG TT AT GG
3'
8. Repeat Reset with , n-2, n-3, n-4 primers
Read Position 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n)
2 Universal seq primer (n-1)
3 Universal seq primer (n-2)
4 Universal seq primer (n-3)
5 Universal seq primer (n-4)
● Indicates positions of interogation
Ligation Cycle 1 2 3 4 5 6 7

DNA nanoball sequencing

- It is high throughput sequencing technology that is used to determine the entire genomic sequence of an organism.
- The method uses rolling circle replication to amplify fragments of genomic DNA molecules.
- This DNA sequencing allows large number of DNA nanoballs to be sequenced per run and at low reagent cost compared to other next generation sequencing platforms.
- However, only short sequences of DNA are determined from each DNA nanoball which makes mapping the short reads to a reference genome difficult.
- This technology has been used for multiple genome sequencing projects and is scheduled to be used for more.

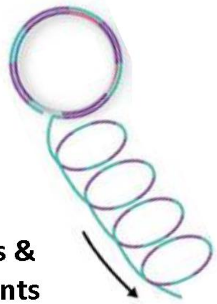
DNA Nanoball Sequencing

Workflow:

1. Isolate DNA

2. Fragment DNA (400-500 bp)

3. Attach adapters & circularize fragments



An iteration of the sequencing-by-ligation next-generation approach, developed by **Complete Genomics** as a human genome sequencing service.

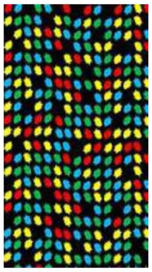
Pros

- *Highly accurate*
- *Low cost*- can generate ~45-87 fold coverage at a consumables cost of $4400/genome
- Nanoballs loaded in an organized array- *high # of reads per flow cell*

Cons

- Chemistry is complex and proprietary
- *Short reads* (35 base paired end)- complex data analysis, challenges with highly repetitive DNA
- Not optimized for a wide range of organisms

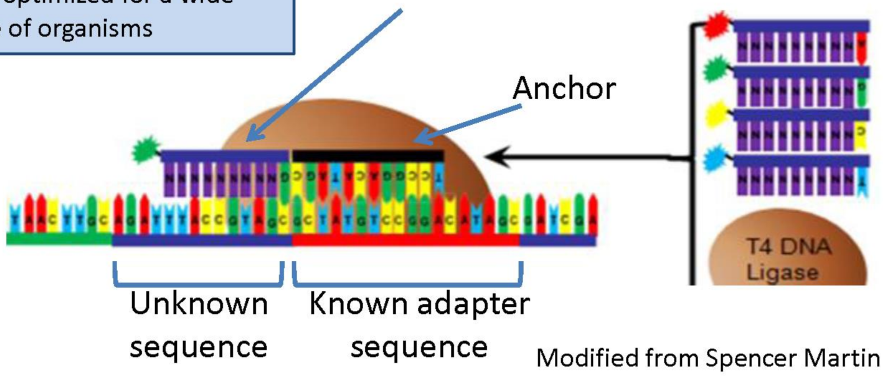
4. *Rolling circle replication*
- amplifies coils of ssDNA to form a chain of copies of the fragment
- the chain is compacted into a *DNA nanoball*- folds on itself due to hybridizing palindromic sequences in adapters



5. *Adsorption onto a silicon flow cell*- a highly ordered microarray

6. *Sequencing using cPAL technology*
-Combinatorial probe-anchor unchained ligation
- Flourescent detection of each hybridization & ligation reaction

Probe for position 1- Subsequent probes will interrogate other positions adjacent to the adapter

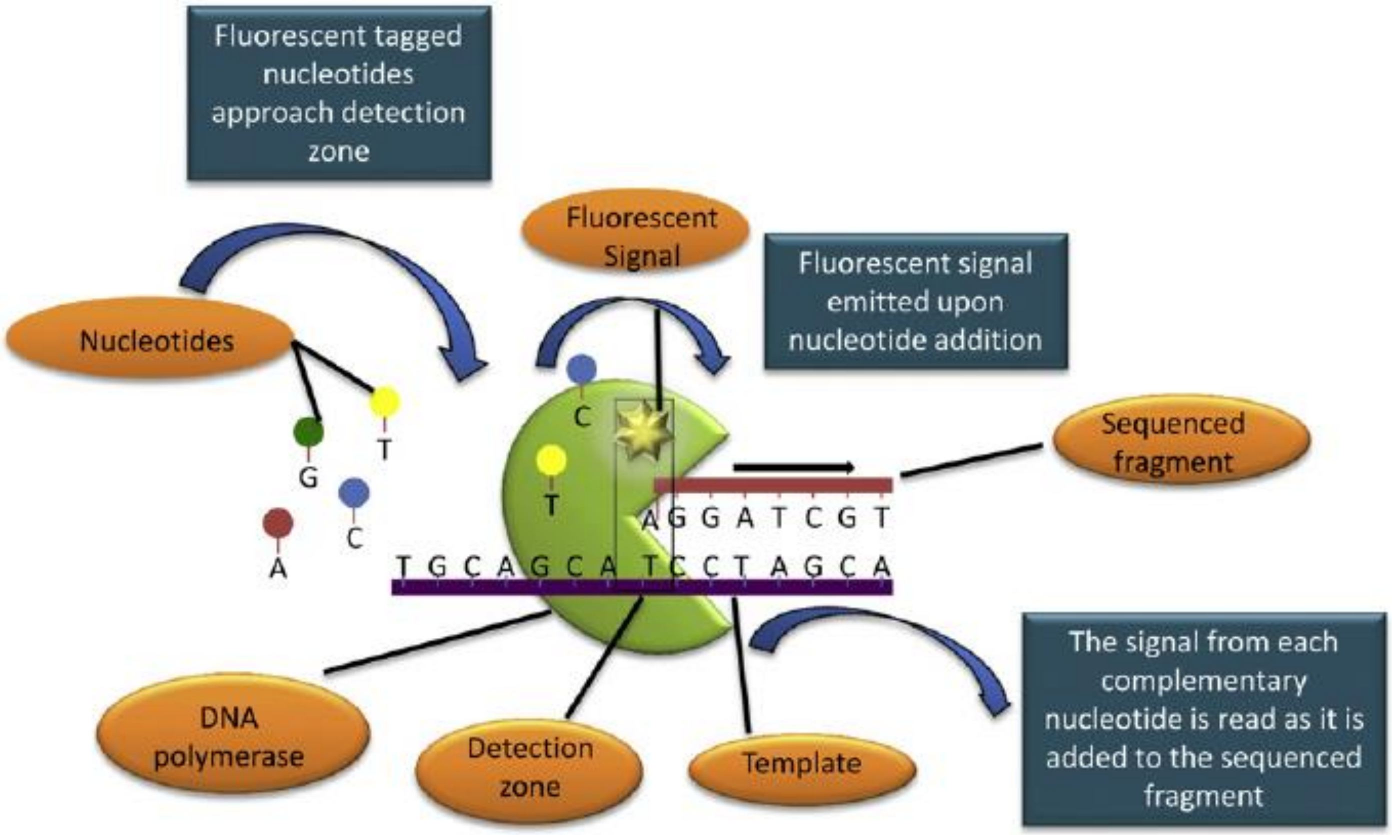


Modified from Spencer Martin

Helioscope single molecule sequencing

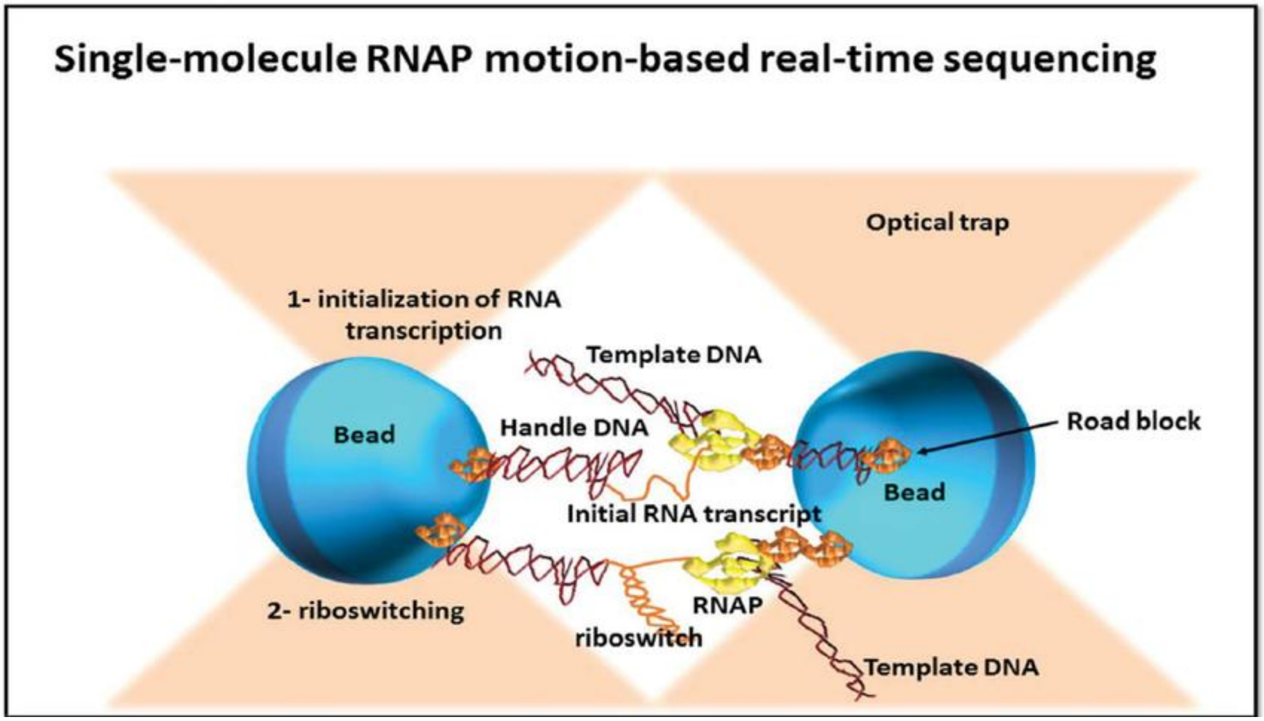
- Helioscope sequencing uses DNA fragments with added polyA tail adapters, which are attached to the flow cell surface.
- The next steps involve extension-based sequencing with cyclic washes of the flow cell with fluorescently labeled nucleotides.
- The reads are performed by the Helioscope sequencer.
- The reads are short, up to 55 bases per run, but recent improvement of the methodology allows more accurate reads of homopolymers and RNA sequencing.

Single molecule SMRT sequencing


Fluorescent tagged nucleotides approach detection zone
Fluorescent Signal
Fluorescent signal emitted upon nucleotide addition
Nucleotides
G
T
C
A
C
T
Sequenced fragment
A G G A T C G T
T G C A G C A T C C T A G C A
DNA polymerase
Detection zone
Template
The signal from each complementary nucleotide is read as it is added to the sequenced fragment

- SMRT sequencing is based on the sequencing by synthesis approach.
- The DNA is synthesisd in so called zero-mode wave-guides (ZMWs) – small well-like containers with the capturing tools located at the bottom of the well.
- The sequencing is performed with use of unmodified polymerase and fluorescently labelled nucleotides flowing freely in the solution.
- The wells are constructed in a way that only the fluorescence occurring by the bottom of the well is detected.
- The fluorescent label is detached from the nucleotide at its incorporation into the DNA strand, leaving an unmodified DNA strand.
- The SMTR technology allows detection of nucleotide modifications. This happens through the observation of polymerase kinetics.
- This approach allows reads of 1000 nucleotides.

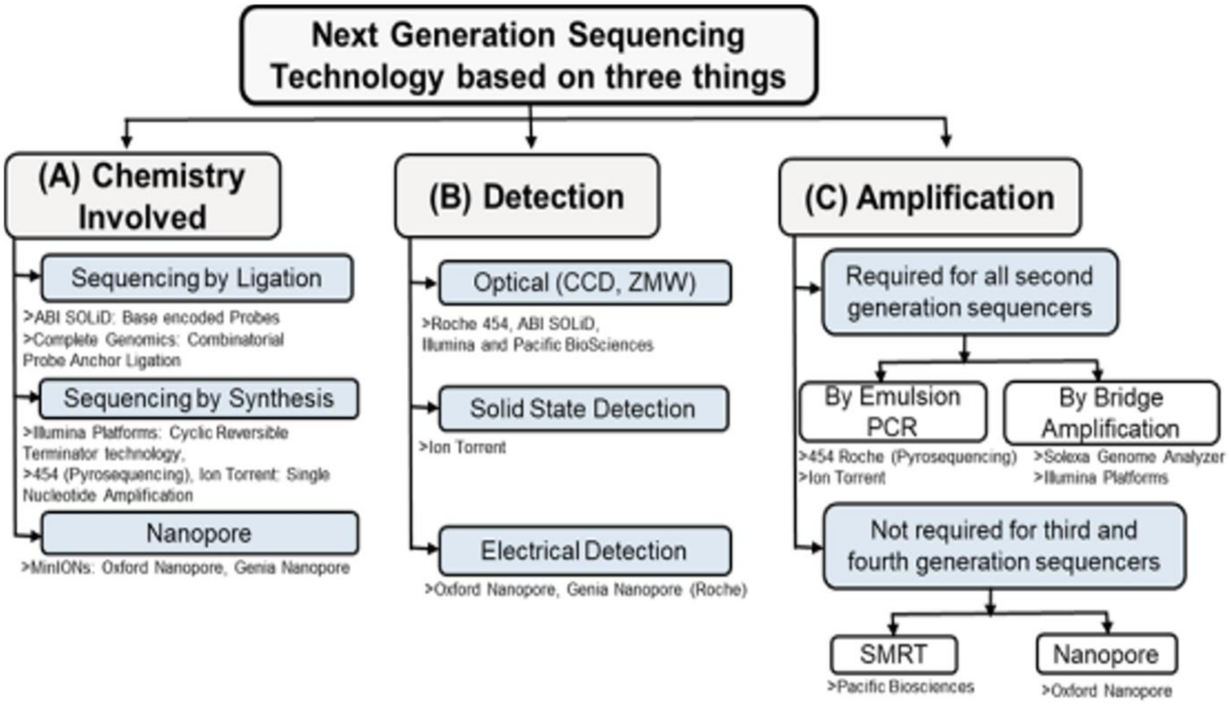
Single molecule real time (RNAP) sequencing



- This method is based on RNA polymerase (RNAP), which is attached to a polystyrene bead, with distal end of sequenced DNA is attached to another bead, with both beads being placed in optical traps.

- RNAP motion during transcription brings the beads in closer and their relative distance changes, which can then be recorded at a single nucleotide resolution.

- The sequence is deduced based on the four readouts with lowered concentrations of each of the four nucleotide types.


Next Generation Sequencing Technology based on three things
(A) Chemistry Involved
Sequencing by Ligation
>ABI SOLiD: Base encoded Probes
>Complete Genomics: Combinatorial Probe Anchor Ligation
Sequencing by Synthesis
>Illumina Platforms: Cyclic Reversible Terminator technology,
>454 (Pyrosequencing), Ion Torrent: Single Nucleotide Amplification
Nanopore
>MinIONs: Oxford Nanopore, Genia Nanopore
(B) Detection
Optical (CCD, ZMW)
>Roche 454, ABI SOLiD, Illumina and Pacific BioSciences
Solid State Detection
>Ion Torrent
Electrical Detection
>Oxford Nanopore, Genia Nanopore (Roche)
(C) Amplification
Required for all second generation sequencers
By Emulsion PCR
>454 Roche (Pyrosequencing)
>Ion Torrent
By Bridge Amplification
>Solexa Genome Analyzer
>Illumina Platforms
Not required for third and fourth generation sequencers
SMRT
>Pacific Biosciences
Nanopore
>Oxford Nanopore

GENETIC MAPS AND PHYSICAL MAPS

Genome Mapping

- Genome mapping is a process of identifying relative locations of genes, mutations or traits on a chromosome.
- It involves assigning/locating of a specific gene to particular region of a chromosome and determining the location of and relative distances between genes on the chromosome.
- **Linkage maps** show the arrangement of genes and genetic markers along the chromosomes as calculated by the frequency with which they are inherited together.
- **Physical maps** represent chromosomes and provide physical distances between chromosomal landmarks ideally measured in nucleotide bases.

Key Points

- Physical maps provide specified detail about the number of bases and physical distance that exists between genetic markers.
- Cytogenetic mapping is a method used to construct physical maps that uses stained sections of chromosomes to approximate the distance between genetic markers.
- Radiation hybrid mapping is a method used to construct physical maps that uses radiation or x-rays to break DNA into fragments to determine the distance between genetic markers and their order on the chromosome.
- Sequence mapping is a method used to construct physical maps that uses already-known locations of genetic markers to determine distances in number of base pairs.

Key Terms

- **cytogenetic**: of or pertaining to the origin and development of cells
- **physical map**: a map showing how much DNA separates two genes and is measured in base pairs
- **expressed sequence tag**: a short sub-sequence of a cDNA sequence that may be used to identify gene transcripts

Physical Maps

A physical map provides detail of the actual physical distance between genetic markers, as well as the number of nucleotides. There are three methods used to create a physical map: cytogenetic mapping, radiation hybrid mapping, and sequence mapping. Cytogenetic mapping uses information obtained by microscopic analysis of stained sections of the chromosome. It is possible to determine the approximate distance between genetic markers using cytogenetic mapping, but not the exact distance (number of base pairs). Radiation hybrid mapping uses radiation, such as x-rays, to break the DNA into fragments. The amount of radiation can be adjusted to create smaller or larger fragments. This technique overcomes the limitation of genetic mapping and is not affected by increased or decreased recombination frequency. Sequence mapping resulted from DNA sequencing technology that allowed for the creation of detailed physical maps with distances measured in terms of the number of base pairs. The creation of genomic libraries and complementary DNA (cDNA) libraries (collections of cloned sequences or all DNA from a genome) has sped up the process of physical mapping. A genetic site used to generate a physical map with sequencing technology (a sequence-tagged site, or STS) is a unique sequence in the genome with a known exact chromosomal location. An expressed sequence tag (EST) and a single sequence length polymorphism (SSLP) are common STSs. An EST is a short STS that is identified with cDNA libraries, while SSLPs are obtained from known genetic markers and provide a link between genetic maps and physical maps.

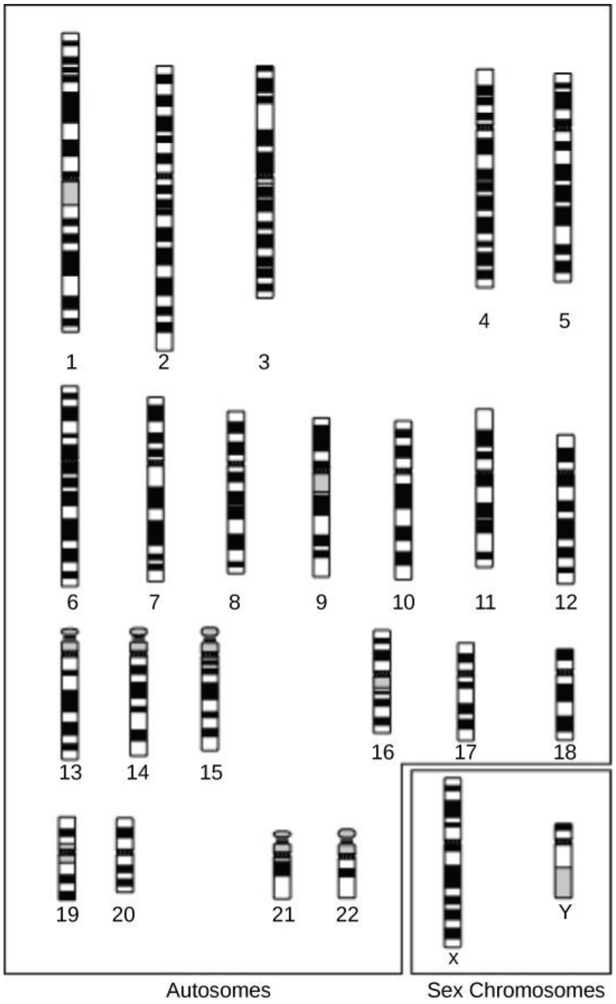
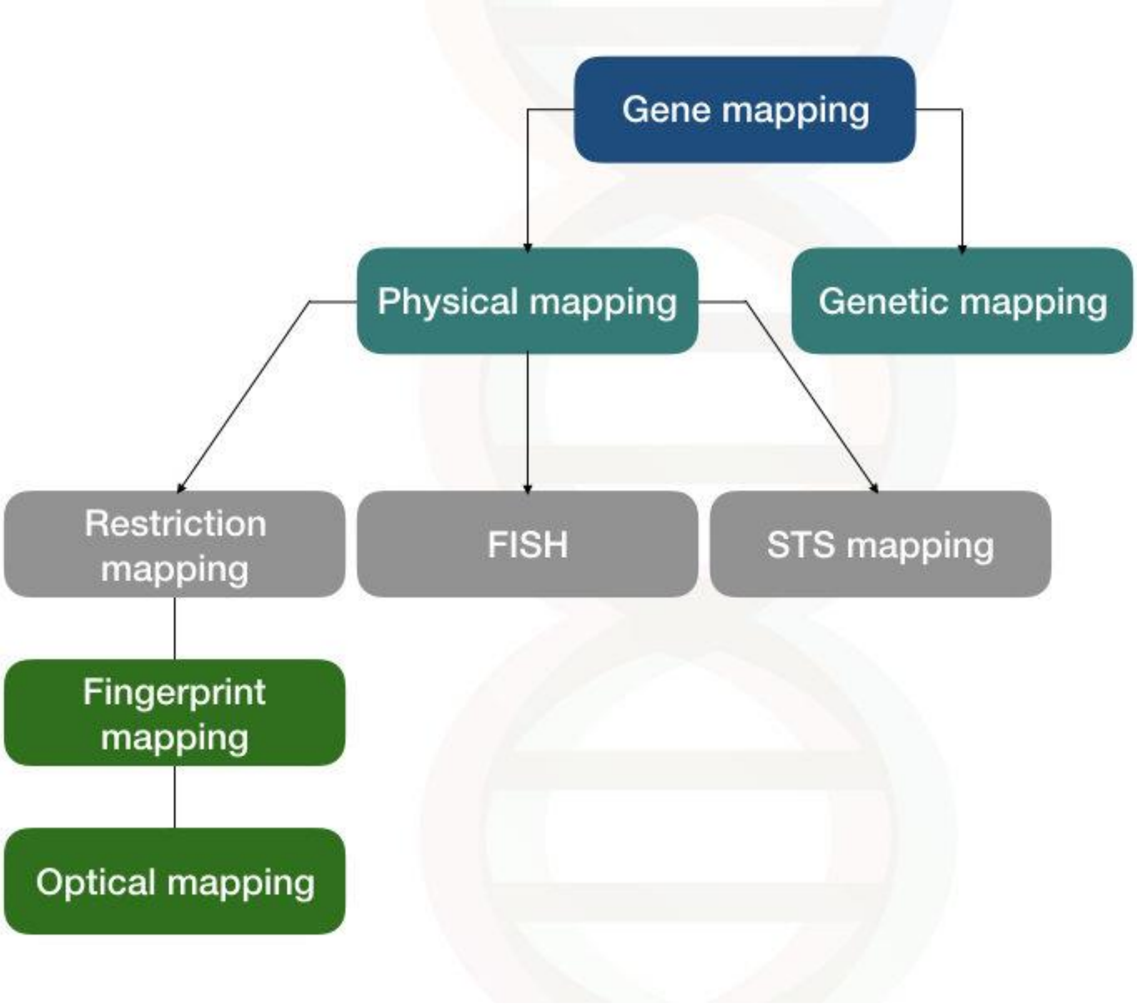


Figure 17.2B.117.2B.1: **Cytogenetic Map**: A cytogenetic map shows the appearance of a chromosome after it is stained and examined under a microscope.

Integration of Genetic and Physical Maps

Genetic maps provide the outline and physical maps provide the details. It is easy to understand why both types of genome mapping techniques are important to show the big picture. Information obtained from each technique is used in combination to study the genome. Genomic mapping is being used with different model research organisms. Genome mapping is an-ongoing process; as better techniques are developed, more advances are expected. Genome mapping is similar to completing a complicated puzzle using every piece of available data. Mapping information generated in laboratories worldwide is entered into central databases, such as GenBank at the National Center for Biotechnology Information (NCBI). Efforts are being made to make the information more easily accessible to researchers and the general public. Just as we use global positioning systems instead of paper maps to navigate through roadways, NCBI has created a genome viewer tool to simplify the data-mining process.



STS CONTENT BASED MAPPING

- Sequence-Tagged Site (STS) is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped.

- The STS concept was introduced by Olson et al (1989). In assessing the likely impact of the Polymerase Chain Reaction (PCR) on human genome research, they recognized that single-copy DNA sequences of known map location could serve as markers for genetic and physical mapping of genes along the chromosome.

- The advantage of STSs over other mapping landmarks is that the means of testing for the presence of a particular STS can be completely described as information in a database: anyone who wishes to make copies of the marker would simply look up the STS in the database, synthesize the specified primers, and run the PCR under specified conditions to amplify the STS from genomic DNA.

- STS-based PCR produces a simple and reproducible pattern on agarose or polyacrylamide gel. In most cases STS markers are co-dominant, i.e., allow heterorozygotes to be distinguished from the two homozygotes.

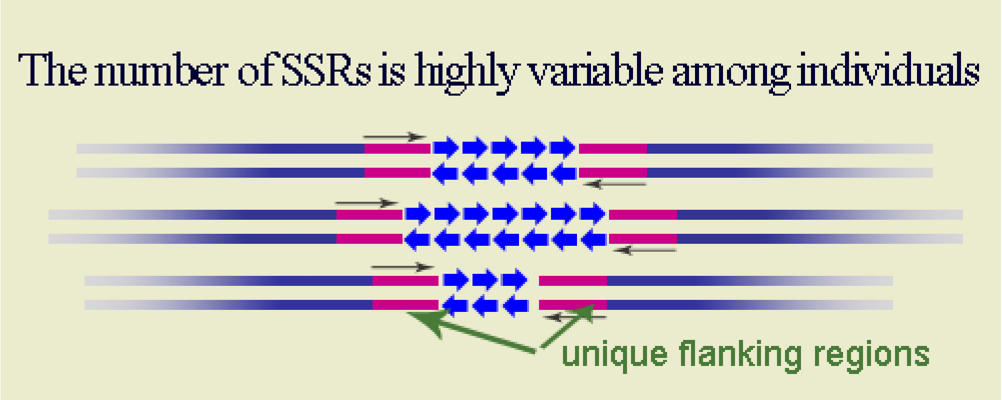
- The DNA sequence of an STS may contain repetitive elements, sequences that appear elsewhere in the genome, but as long as the sequences at both ends of the site are unique and conserved, researches can uniquely identify this portion of genome using tools usually present in any laboratory.

- Thus, in broad sense, STS include such markers as microsatellites (SSRs, STMS or SSRPs), SCARs, CAPs, and ISSRs.

Microsatellites

- Polymorphic loci present in nuclear DNA and organellar DNA that consist of repeating units of 1-10 base pairs, most typically, 2-3 bp in length, also called Simple Sequence Repeats (SSR), Sequence-Tagged Microsatellite Sites (STMS) or Simple Sequence Repeats Polymorphisms (SSRP).

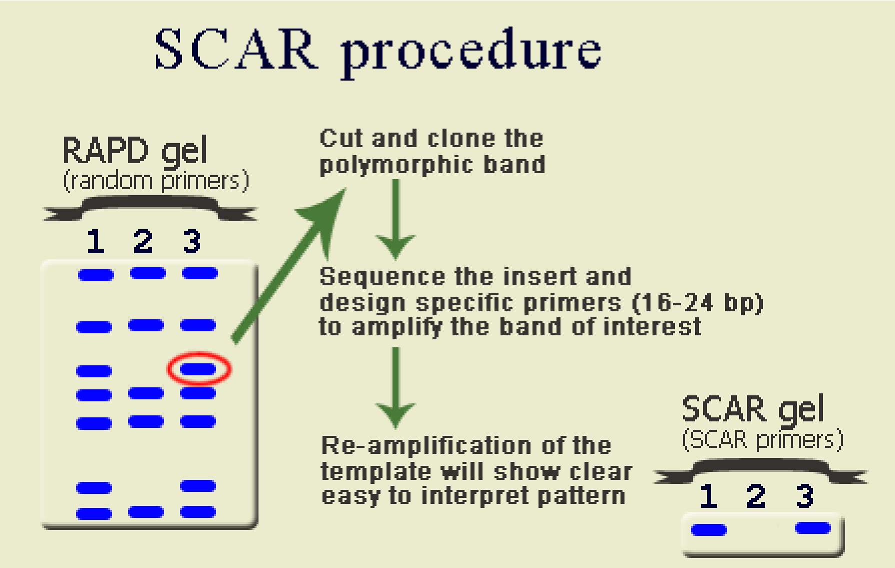
- SSRs are highly variable and evenly distributed throughout the genome. This type of repeated DNA is common in eukaryotes. These polymorphisms are identified by constructing PCR primers for the DNA flanking the microsatellite region. The flanking regions tend to be conserved within the species, although sometimes they may also be conserved in higher taxonomic levels.



Sequence Characterized Amplified Region (SCAR)

- DNA fragments amplified by the Polymerase Chain Reaction (PCR) using specific 15-30 bp primers, designed from nucleotide sequences established in cloned RAPD (Random Amplified Polymorphic DNA) fragments linked to a trait of interest.

- By using longer PCR primers, SCARs do not face the problem of low reproducibility generally encountered with RAPDs. Obtaining a co-dominant marker may be an additional advantage of converting RAPDs into SCARs.


SCAR procedure
RAPD gel
(random primers)
1 2 3
Cut and clone the polymorphic band
Sequence the insert and design specific primers (16-24 bp) to amplify the band of interest
Re-amplification of the template will show clear easy to interpret pattern
SCAR gel
(SCAR primers)
1 2 3

Cleaved Amplified Polymorphic Sequences (CAPS)

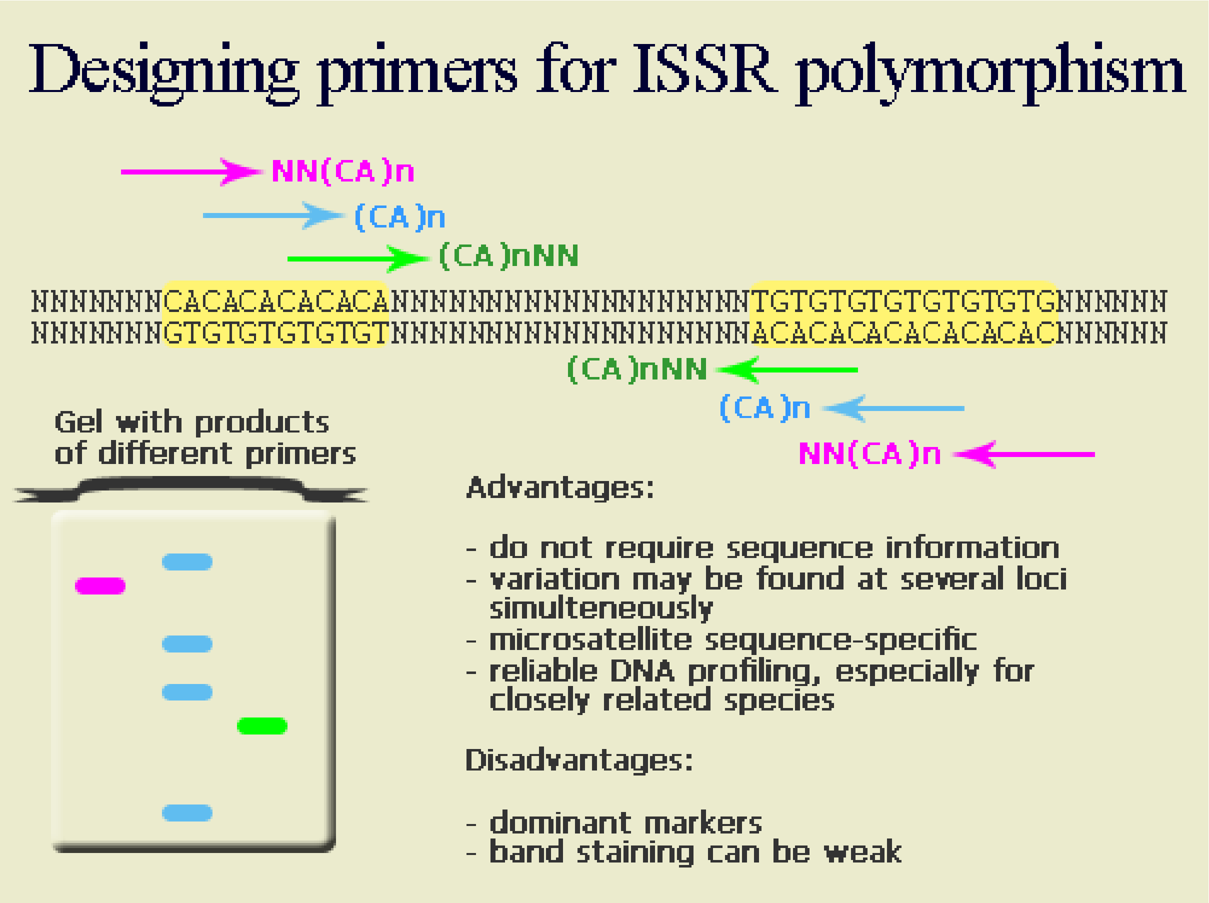
- STS polymorphisms that can be detected by differences in restriction fragment lengths caused by SNPs or INDELs that create or abolish restriction endonuclease recognition sites in PCR amplicons produced by locus-specific oligonucleotide primers.

- In other words this technique aims to convert and amplified band that does not show variation by length of PCR product into a polymorphic one. More about CAPS in Overview of CAPS technology.

Inter-simple Sequence Repeats (ISSRs)

- STS polymorphisms that are found between microsatellite repeats. Primers can be designed based on a microsatellite repeats exclusively, in which case this technique will target multiple loci due to known abundance of repeat sequences in the genome.

- Alternatively, primers can be extended outside or inside the ISSR in which case a unique region most likely will be amplified.



Designing primers for ISSR polymorphism

NN(CA)n
(CA)n
(CA)nNN
NNNNNNNCACACACACACANNNNNNNNNNNNNNNNNNNTGTGTGTGTGTGTGTGNNNNNN
NNNNNNNGTGTGTGTGTGTNNNNNNNNNNNNNNNNNNNACACACACACACACACNNNNNN
(CA)nNN
(CA)n
NN(CA)n
Gel with products
of different primers

Advantages:

- do not require sequence information
- variation may be found at several loci simulteneously
- microsatellite sequence-specific
- reliable DNA profiling, especially for closely related species

Disadvantages:

- dominant markers
- band staining can be weak

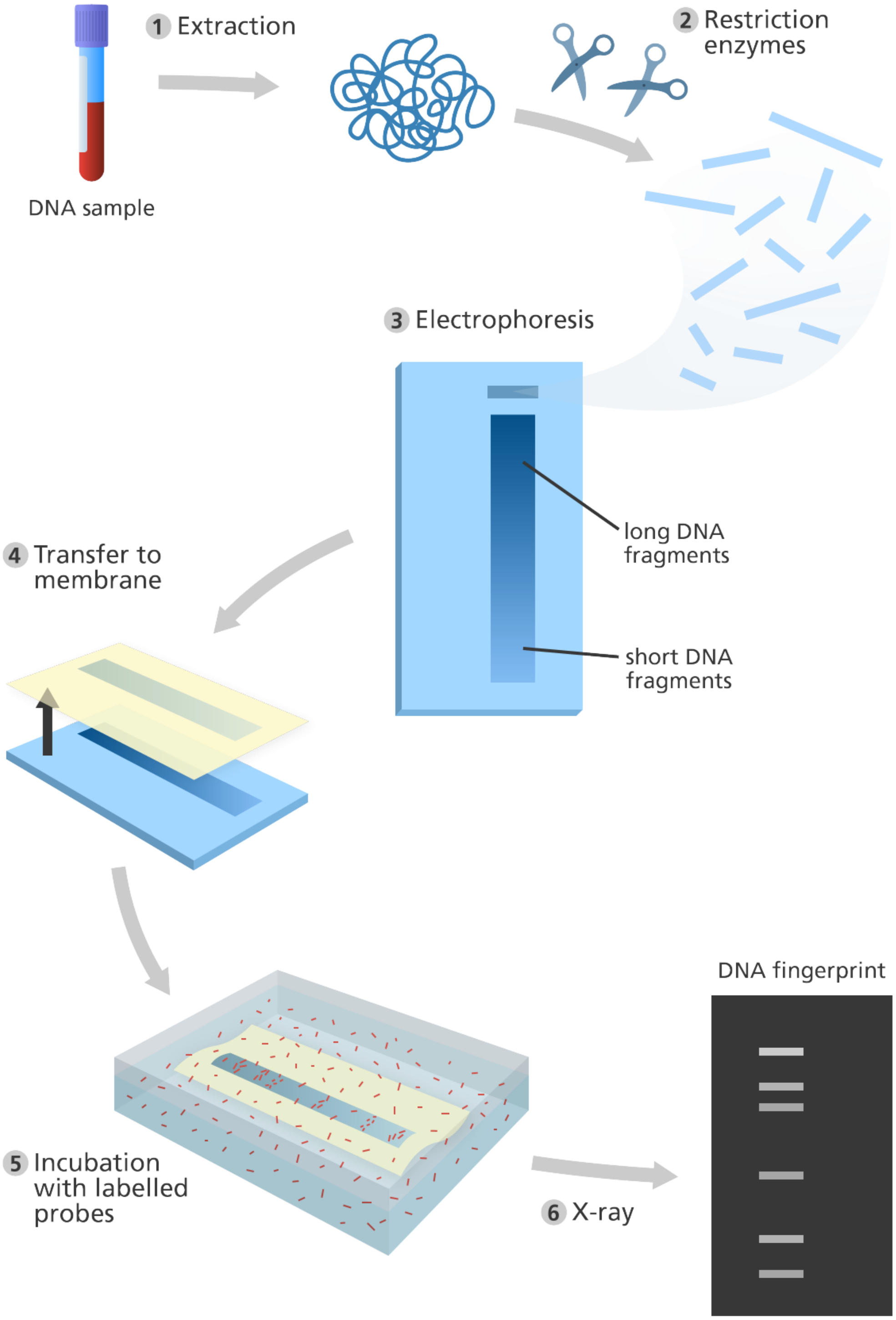
RESTRICTION ENZYME FINGER PRINTING

Restriction endonuclease fingerprinting (REF) is a modification of single-strand confirmation polymorphism (SSCP) that was developed to detect the presence of essentially all mutations in a 1-kb segment.

DNA Finger Printing

- DNA fingerprinting or DNA profiling is a process used to determine the nucleotide sequence at a certain part of the DNA that is unique in all human beings.
- The process of DNA fingerprinting was invented by Sir Alec Jeffrey at the University of Leicester in 1985.

Principle of DNA Fingerprinting



- The DNA of every human being on the planet is 99.9% same. However, about 0.1% or 3 x 10^6base pairs (out of 3 x 10^9 bp) of DNA is unique in every individual.

- Human genome possesses numerous small non-coding but inheritable sequences of bases which are repeated many times. They do not code for proteins but make-up 95% of our genetic DNA and therefore called the ―junk DNA.

- They can be separated as satellite from the bulk DNA during density gradient centrifugation and hence called satellite DNA.

- In satellite DNA, repetition of bases is in tandem. Depending upon length, base composition and numbers of tandemly repetitive units, satellite DNAs have subcategories like microsatellites and mini-satellites.

- Satellite DNAs show polymorphism. The term polymorphism is used when a variant at a locus is present with a frequency of more than 0.01 population.

- Variations occur due to mutations. These mutations in the non-coding sequences have piled up with time and form the basis of DNA polymorphism (variation at genetic level arises due to mutations).

- The junk DNA regions are thus made-up of length polymorphisms, which show variations in the physical length of the DNA molecule.

- At specific loci on the chromosome the number of tandem repeats varies between individuals. There will be a certain number of repeats for any specific loci on the chromosome.

- Depending on the size of the repeat, the repeat regions are classified into two groups. **Short tandem repeats (STRs)** contain 2-5 base pair repeats and **variable number of tandem repeats (VNTRs)** have repeats of 9-80 base pairs.

- Since a child receive 50% of the DNA from its father and the other 50% from his mother, so the number VNTRs at a particular area of the DNA of the child will be different may be due to insertion, deletion or mutation in the base pairs.

- As a result, every individual has a distinct composition of VNTRs and this is the main principle of DNA fingerprinting.

- As single change in nucleotide may make a few more cleavage site of a given nucleotide or might abolish some existing cleavage site.

- Thus, if DNA of any individual is digested with a restriction enzyme, fragments pattern (sizes) will be produced and will be different in cleavage site position. This is the basics of DNA fingerprinting.

Methods of DNA Fingerprinting

- Restriction fragment length polymorphism (RFLP) and polymerase chain reaction (PCR) amplification of short tandem repeats (STRs) are two main DNA tests widely used for DNA fingerprinting.

A. Restriction fragment length polymorphism (RFLP)

- The first step in this process is to isolate the DNA from the sample material to be tested. The sample size for RFLP test must be large enough to get the proper result.

- Once the required size of the sample is available, the DNA is isolated from the sample and is subjected to restriction digestion using restriction enzymes.

- The digested DNA sample is then separated by agarose gel electrophoresis, in which the DNA is separated based on the size.

- The next step is transfer of separated DNA from gel slab onto the nitrocellulose membrane to hybridize with a labeled probe that is specific for one VNTR region (radio activity labeled complimentary sequence for VNTR region nucleotide sequence).

- This technique of transferring and hybridizing DNA onto nitrocellulose membrane is known as southern blotting, a most widely used DNA detection technique by molecular biologists.

- After the hybridization with the radioactive probes, the X- ray film is developed form the southern blotting and only the areas where the radioactive probe binds will show up on the film.

- Now these bands when compared with the other known samples, will give the final result of the DNA fingerprinting.

Advantages

- The RFLP is considered to be more accurate than the PCR, mainly because the size of the sample used more, use of a fresh DNA sample, and no amplification contamination.

Limitation

- The RFLP, however, require longer time period to complete the analysis and is costly.

B. Polymerase Chain Reaction (PCR) amplification of short tandem repeats (STRs)

1. Thousands of copies of a particular variable region are amplified by PCR which forms the basis of this detection.

2. STR with a known repeat sequence is amplified and separated using gel-electrophoresis.

    - The distance migrated by the STR is examined.

3. For the amplification of STRs using PCR, a short synthetic DNA, called primers are specially designed to attach to a highly conserved common nonvariable region of DNA that flanks the variable region of the DNA.

4. By comparing the STR sequence size amplified by PCR with the other known samples, will give the final result of the DNA fingerprinting.

Advantages

- Small amount of specimen is sufficient for the test.

- Takes a shorter time to complete.

- Less costly.

Limitation

- Less accurate than RFLP.
- Possibility of amplification contamination.

RADIATION HYBRIDIZATION MAPPING

Radiation hybrid mapping is a genetic technique that was originally developed for constructing long-range maps of mammalian chromosomes. It is based on a statistical method to determine not only the distances between deoxyribonucleic acid (DNA) markers but also their order on the chromosomes. DNA markers are short, repetitive DNA sequences, most often located in noncoding regions of the genome , that have proven extremely valuable for localizing human disease genes in the genome.

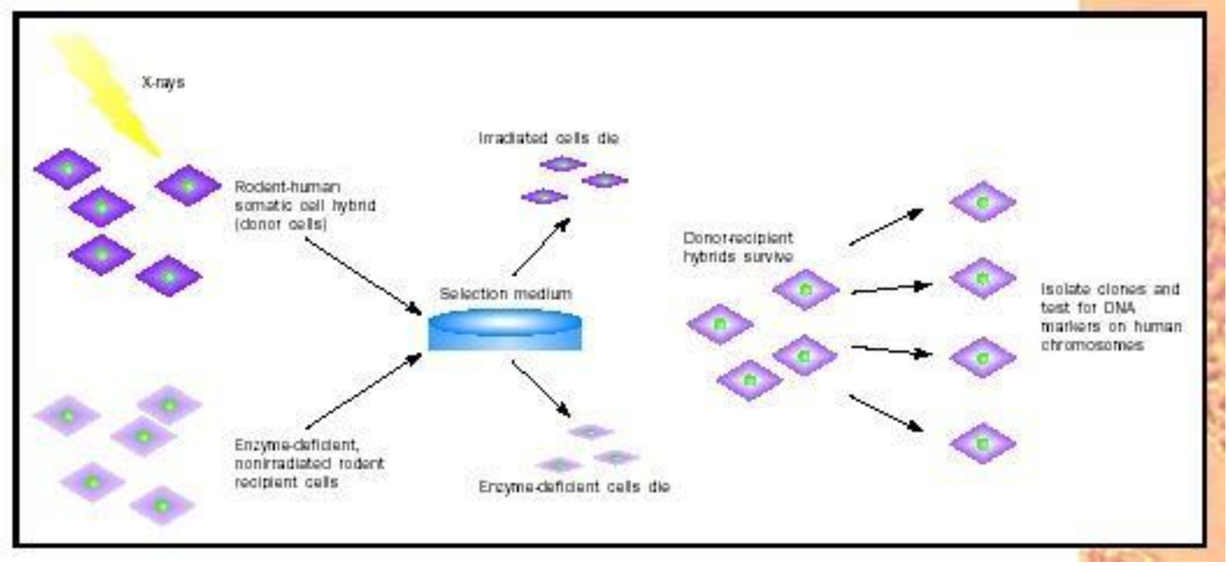
Theory and Application

In radiation hybrid mapping, human chromosomes are separated from one another and broken into several fragments using high doses of X rays. Similar to the underlying principle of mapping genes by linkage analysis based on recombination events, the farther apart two DNA markers are on a chromosome, the more likely a given dose of X rays will break the chromosome between them and thus place the two markers on two different chromosomal fragments. The order of markers on a chromosome can be determined by estimating the frequency of breakage that, in turn, depends on the distance between the markers. This technique has been used to construct whole-genome radiation hybrid maps.

Technique

A rodent-human somatic cell hybrid ("artificial" cells with both rodent and human genetic material), which contains a single copy of the human chromosome of interest, is

X-irradiated.


X-rays
Rodent-human somatic cell hybrid (donor cells)
Irradiated cells die
Selection medium
Donor-recipient hybrids survive
Isolate clones and test for DNA markers on human chromosomes
Enzyme-deficient, nonirradiated rodent recipient cells
Enzyme-deficient cells die

Radiation hybrid mapping process.

- This breaks the chromosome into several pieces, which are subsequently integrated into the rodent chromosomes. In addition, the dosage of radiation is sufficient to kill the somatic cell hybrid or donor cells, which are then rescued by fusing them with nonirradiated rodent recipient cells. The latter, however, lack an important enzyme and are also killed when grown in a specific medium. Therefore, the only cells that can survive the procedure are donor-recipient hybrids that have acquired a rodent gene for the essential enzyme from the irradiated rodent-human cell line (see Figure above).

- From these donor-recipient hybrids, clones can be isolated and tested for the presence or absence of DNA markers on the human chromosome of interest, and the frequencies with which markers were retained in each clone can be calculated. This process is complicated by the fact that hybrids may contain more than one DNA fragment.

- For example, two markers retained in one hybrid may result from retention of the two markers on separate fragments or from no break between the markers. However, the frequency of breakage, theta, can be estimated using statistical methods, and a lod score (logarithm of the likelihood ratio for linkage) can be calculated to identify significantly linked marker pairs.

OPTICAL MAPPING

- Optical mapping combined with a few other sequencing techniques such as PacBio, Oxford Nanopore, Nabsys, and next-generation sequencing (NGS) techniques provides a solution to de novo assembly of reference-quality whole genome sequencing.

- Optical genome mapping is a single molecule method to find the ordered length of sections of DNA between sequence specific labels.

- Fluorescence microscopy is used to visualize the entirety of single DNA strands using backbone stains.

- Various labeling methods, traditionally restriction enzymes, are used to find a small portion of the sequence of the DNA strand and cut it at each labeling point.

- The information acquired during optical mapping can be used to aid in sequencing of DNA, identifying pathogens, testing for diseases, and forensics.

- Optical mapping has the potential to supplant electrophoresis as the dominant technique in DNA profiling due to the additional information gained in addition to the benefits of using reduced sample sizes.

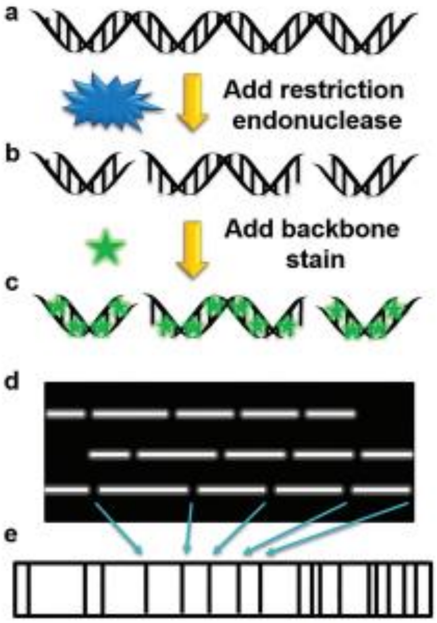


Fig. 1 Scheme of a traditional optical mapping protocol. (a) DNA before modification. (b) DNA with gaps after being cut by the restriction endonucleases. (c) Cut DNA with a backbone stain added to visualize the fragmented DNA strands. (d) Scheme of a fluorescence image of the digested DNA molecules. (e) A barcode generated from aligning fragments of cut DNA where each line represents a gap and the box represents the whole genome

ORF FINDING AND FUNCTIONAL ANNOTATION

DNA (Deoxyribonucleic acid) is the genetic material that contains all the genetic information in a living organisms. The information is stored as genetic codes using adenine (A), guanine (G), cytosine(C) and thymine (T). During the transcription process, DNA is transcribed to mRNA. Each of these base pairs will bond with a sugar and phosphate molecule to form a nucleotide. Three nucleotides that codes for a particular amino acid during translation is called as a codon. The region of a nucleotide that starts from an initiation codon and ends with a stop codon is called an Open Reading Frame(ORF). Proteins are formed from ORF. By analyzing the ORF we can predict the possible amino acids that might be produced during translation. The ORF finder is a program available at NCBI website. It identifies all ORF or possible protein coding region from six different reading frames.

DNA (Deoxyribonucleic acid) is the genetic material that contains the genetic information for development and helps in maintaining all the functions in living organisms. The information is stored as genetic codes using four different bases. They are adenine (A), guanine (G), cytosine(C) and thymine (T). In two strands of DNA, adenine always pair with thymine and guanine pair with cytosine. Each of these base pairs will bond with a sugar and phosphate molecule to form a nucleotide. The base pairing of DNA will result in a ladder shape structure of these strands which is called a double helix. RNA is differs from DNA only in 1 base pair i.e. in RNA it is uracil (U) instead of thymine(T). mRNA (messenger RNA) is a type of RNA which is formed from DNA transcription. During the transcription process, DNA is transcribed to mRNA in the nucleus and moves to the cytoplasm through the nuclear pores. This mRNA is translated to protein in the cytoplasm with the help of ribosomes. In mRNA, 3 nucleotides are considered at a time since a set of 3 nucleaotides (refered to as codon) codes for an amino acid. The region of a nucleotide that starts from an initiation codon and ends with a stop codon is called an Open Reading Frame(ORF).

An initiation codon is the triplet codon that codes for the first amino acid in the translation process. The translation process will start only with the initiation codon, ATG which codes for the amino acid methionine. The translation process stops when it comes across a stop codon.

There are three stop codons: TAA ("ochre"), TAG ("amber") and TGA ("opal" or "umber"). Any of these codons can stop the translation. Genetic codon can form 64 triplets(4^3) from the 4 nucleotides that codes for amino acids. Protein is formed from the ORF.

How to find ORF

1. Consider a hypothetical sequence:

CGCTACGTCTTACGCTGGAGCTCTCATGGATCGGTTCGGTAGGGCTCGATCACATCG
CTAGCCAT

2. Divide the sequence into 6 different reading frames(+1, +2, +3, -1, -2 and -3). The first reading frame is obtained by considering the sequence in words of 3.

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT

The second reading frame is formed after leaving the first nucleotide and then grouping the sequence into words of 3 nucleotides

FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

The third reading frame is formed after leaving the first 2 nucleotides and then grouping the sequence into words of 3 nucleotides

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT

The other 3 reading frames can be found only after finding the reverse complement.

Complement:
GCGATGCAGAATGCGACCTCGAGAGTACCTAGCCAAGCCATCCCGAGCTAGTGTAG
CGATCGGTA

Reverse complement:
ATGGCTAGCGATGTGATCGAGCCCTACCGAACCGATCCATGAGAGCTCCAGCGTAA
GACGTAGCG

Now same process as that of +1, +2 and +3 strands is repeated for -1, -2 and -3 strands with reverse complement sequence

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG

FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G

FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

3. Now mark the start codon and stop codons in the reading frames

FRAME +1: CGC TAC GTC TTA CGC TGG AGC TCT CAT GGA TCG GTT CGG TAG GGC TCG ATC ACA TCG CTA GCC AT

FRAME +2: C GCT ACG TCT TAC GCT GGA GCT CTC ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG CCA T

FRAME +3: CG CTA CGT CTT ACG CTG GAG CTC TCA TGG ATC GGT TCG GTA GGG CTC GAT CAC ATC GCT AGC CAT

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA GAG CTC CAG CGT AAG ACG TAG CG

FRAME -2: A TGG CTA GCG ATG TGA TCG AGC CCT ACC GAA CCG ATC CAT GAG AGC TCC AGC GTA AGA CGT AGC G

FRAME -3: AT GGC TAG CGA TGT GAT CGA GCC CTA CCG AAC CGA TCC ATG AGA GCT CCA GCG TAA GAC GTA GCG

4. Identify the open reading frame (ORF) - sequence stretch begining with a start codon and ending in a stop codon.

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA

FRAME -3: ATG AGA GCT CCA GCG TAA

5. Based on the amino acid table the peptide sequence is found

FRAME +2: ATG GAT CGG TTC GGT AGG GCT CGA TCA CAT CGC TAG

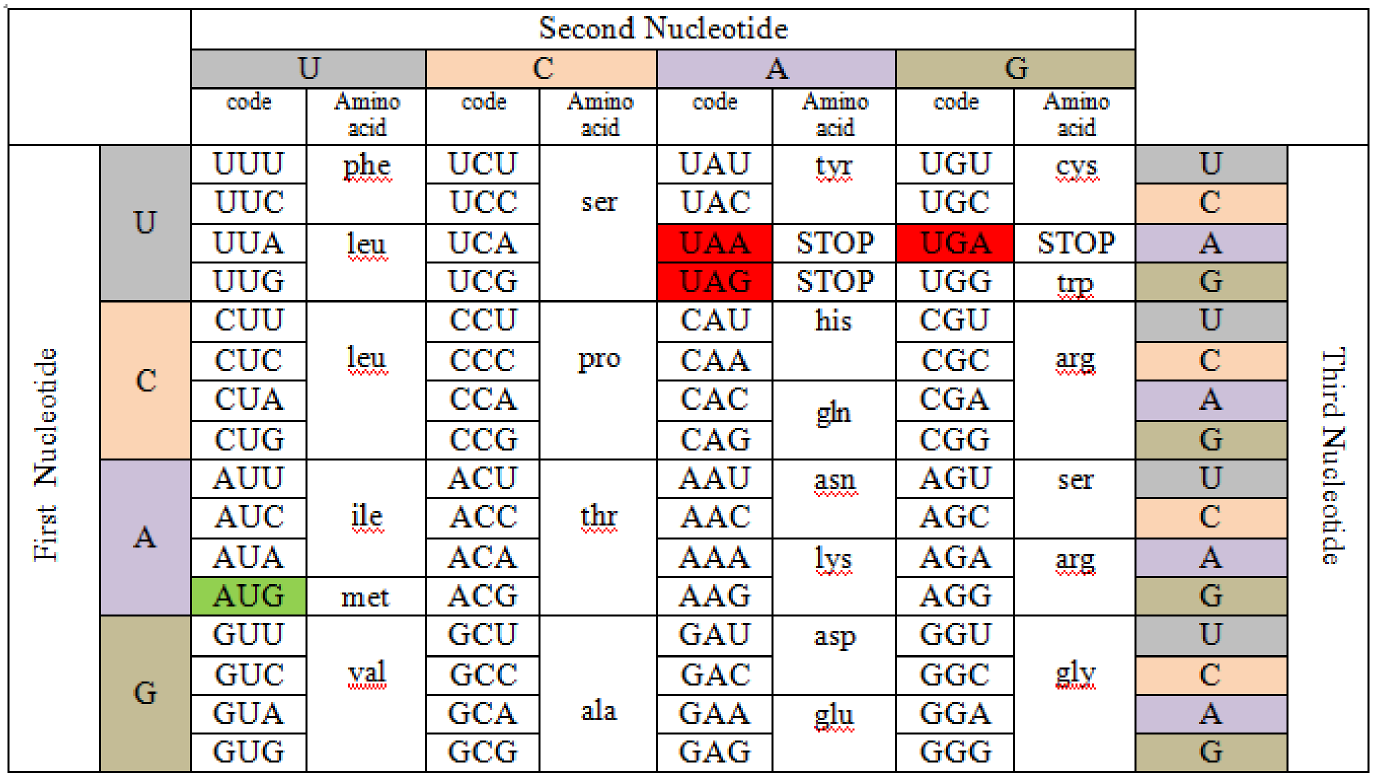
met asp arg phe gly arg ala arg ser his arg stop

FRAME -1: ATG GCT AGC GAT GTG ATC GAG CCC TAC CGA ACC GAT CCA TGA
met ala ser asp val ile glu pro tyr arg thr asp pro stop

FRAME -3: ATG AGA GCT CCA GCG TAA

met arg ala pro ala stop

By analyzing the ORF we can predict the possible amino acids that are producing during the translation process. The prediction of the correct ORF from a newly sequenced gene is an important step. Finding ORF helps to design the primers which are required for experiments like PCR, sequencing etc.



		Second Nucleotide									
		U		C		A		G			
		code	Amino acid	code	Amino acid	code	Amino acid	code	Amino acid		
First Nucleotide	U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U	Third Nucleotide
		UUC		UCC		UAC		UGC		C	
		UUA	leu	UCA		UAA	STOP	UGA	STOP	A	
		UUG		UCG		UAG	STOP	UGG	trp	G	
	C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U	
		CUC		CCC		CAA		CGC		C	
		CUA		CCA		CAC	gln	CGA		A	
		CUG		CCG		CAG		CGG		G	
	A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U	
		AUC		ACC		AAC		AGC		C	
		AUA		ACA		AAA	lys	AGA	arg	A	
		AUG	met	ACG		AAG		AGG		G	
	G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U	
		GUC		GCC		GAC		GGC		C	
		GUA		GCA		GAA	glu	GGA		A	
		GUG		GCG		GAG		GGG		G	

ORF Finder:

- The ORF finder is a program available at NCBI website. It identifies the all open reading frames or the possible protein coding region in sequence. It shows 6 horizontal bars corresponding to one of the possible reading frame.

- In each direction of the DNA there would be 3 possible reading frames. So total 6 possible reading frame (6 horizontal bars) would be there for every DNA sequence. The 6 possible reading frames are +1, +2, +3 and -1, -2 and -3 in the reverse strand.

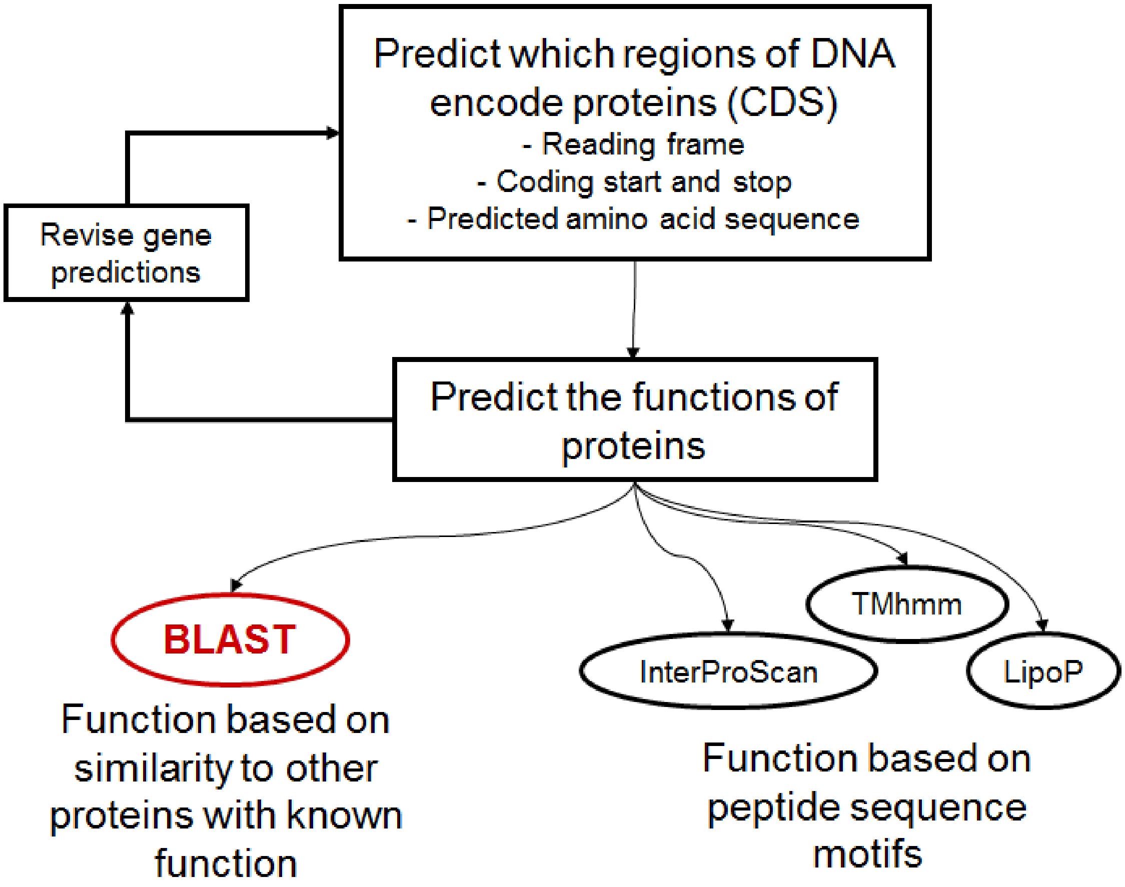
- The resultant amino acids can be saved and search against various protein databases using blast for finding similar sequences or amino acids. The result displays the possible protein sequence and the length of the open reading frame etc.

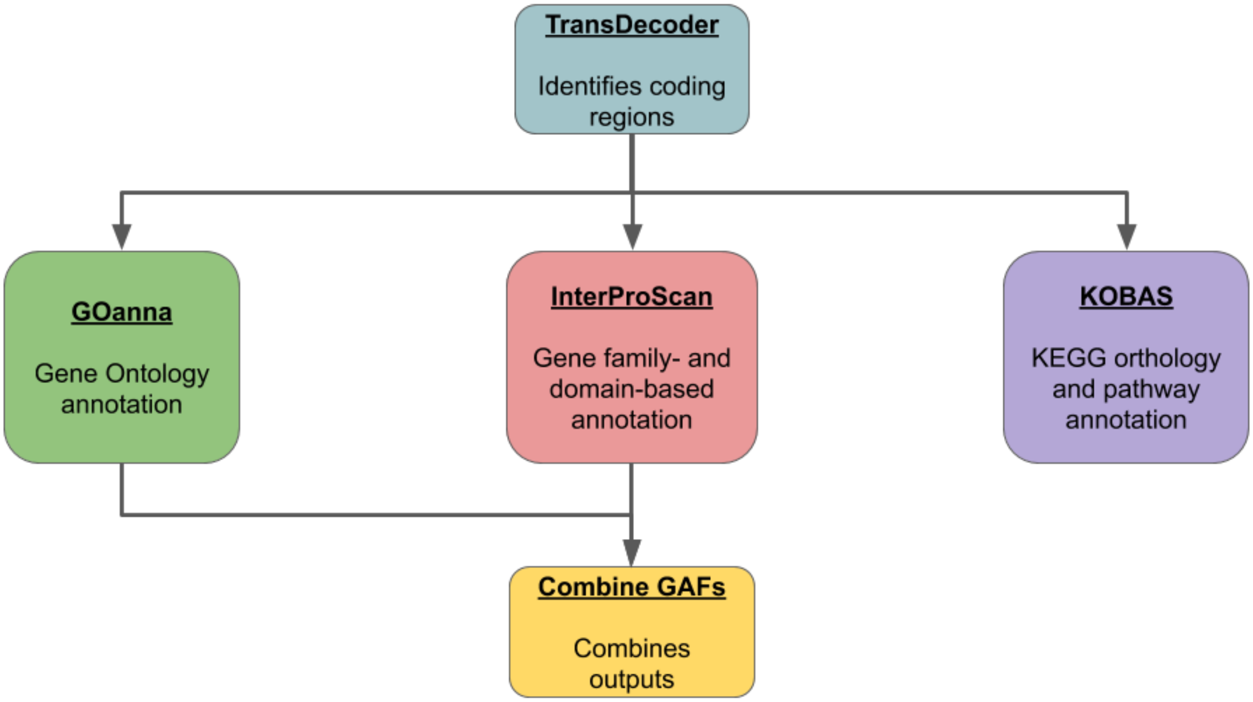
Genome Annotation

- Before the assembled sequence is deposited into a database, it has to be analyzed for useful biological features. The genome annotation process provides comments for the features.

- This involves two steps: gene prediction and functional assignment which both may be accomplished by bioinformatics tools.

FUNCTIONAL ANNOTATION


Predict which regions of DNA encode proteins (CDS)
- Reading frame
- Coding start and stop
- Predicted amino acid sequence
Revise gene predictions
Predict the functions of proteins
BLAST
Function based on similarity to other proteins with known function
TMhmm
InterProScan
LipoP
Function based on peptide sequence motifs


TransDecoder
Identifies coding regions
GOanna
Gene Ontology annotation
InterProScan
Gene family- and domain-based annotation
KOBAS
KEGG orthology and pathway annotation
Combine GAFs
Combines outputs

This functional annotation workflow employs three annotation tools:

1. **GOanna:** It performs a BLAST search and transfers gene ontology (GO) annotations from BLAST matches to the query gene products.
2. **InterProScan:** InterPro is a database which integrates together predictive information about proteins' function from a number of partner resources, giving an overview of the families that a protein belongs to and the domains and sites it contains. InterProScan can also provide GO and pathway annotations.
3. **KOBAS:** It uses BLAST to annotate the input with KEGG Orthology terms and KEGG pathways

- Genome annotation is the process of attaching biological information to sequences. It consists of two main steps: identifying elements on the genome, a process called gene prediction, and attaching biological information to these elements.
- Automatic annotation tools try to perform all of this by computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline (process).
- The basic level of annotation is using BLAST for finding similarities, and then annotating genomes based on that. However, nowadays more and more additional information is added to the annotation platform.
- The additional information allows manual annotators to deconvolute discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases rely on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline.
- Structural annotation consists of the identification of genomic elements: ORFs and their localization, gene structure, coding regions, and the location of regulatory motifs. Functional annotation consists of attaching biological information to genomic elements: biochemical function, biological function, involved regulation and interactions, and expression.
- These steps may involve both biological experiments and in silico analysis. Proteogenomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.
- A variety of software tools have been developed to permit scientists to view and share genome annotations. Genome annotation is the next major challenge for the Human Genome Project, now that the genome sequences of human and several model organisms are largely complete. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together"

Assessment:

Brief the following:

1. Micro injection.
2. Blue white selection.

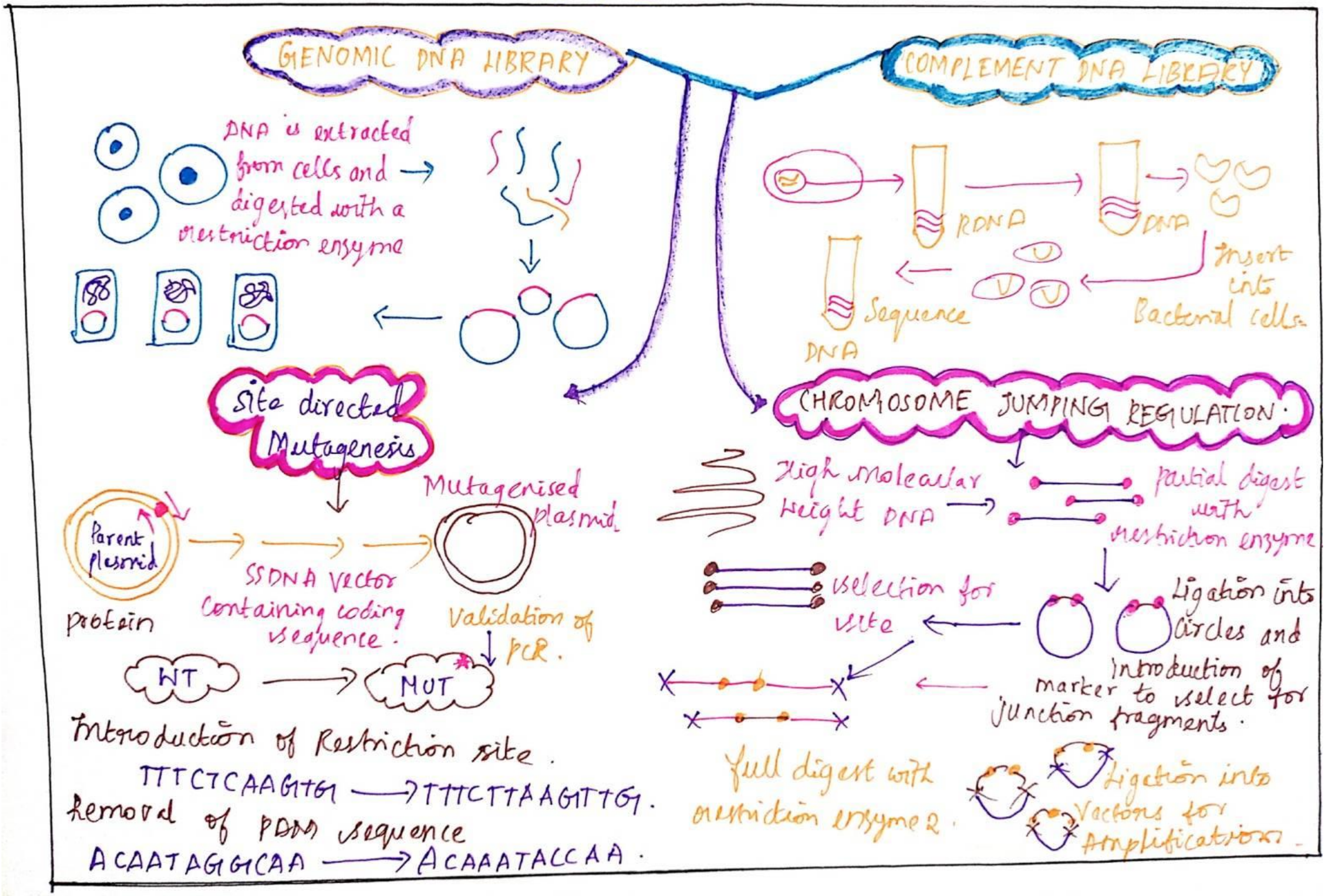
3. Micro projector.

Detail the following:

4. Selection of recombinant bacteria techniques.
5. Ultrasonication method.

UNIT IV

UNIT-IV.



GENOMIC DNA LIBRARY

DNA is extracted from cells and digested with a restriction ensyme

COMPLEMENT DNA LIBRARY

RDNA

DNA

Insert into Bacterial cells

Sequence DNA

Site directed Mutagenesis

Parent plasmid

protein

SSDNA Vector Containing coding sequence.

Mutagenised plasmid

Validation of PCR.

WT

MUT

Introduction of Restriction site.

TTTCTCAAGTGT → TTTCTTAAGTTGT.

Removal of PDNA sequence

ACAATAGGCAA → ACAAATACCAA.

CHROMOSOME JUMPING REGULATION.

High Molecular Height DNA

partial digest with restriction enzyme

Ligation into Circles and Introduction of marker to select for Junction fragments.

selection for site

full digest with restriction ensyme 2.

Ligation into Vectors for Amplifications.

DNA LIBRARIES

CONSTRUCTION OF GENOMIC LIBRARIES

1. Meaning of Genomic Libraries:

Genomic libraries are libraries of genomic DNA sequences. These can be produced using DNA from any organism.

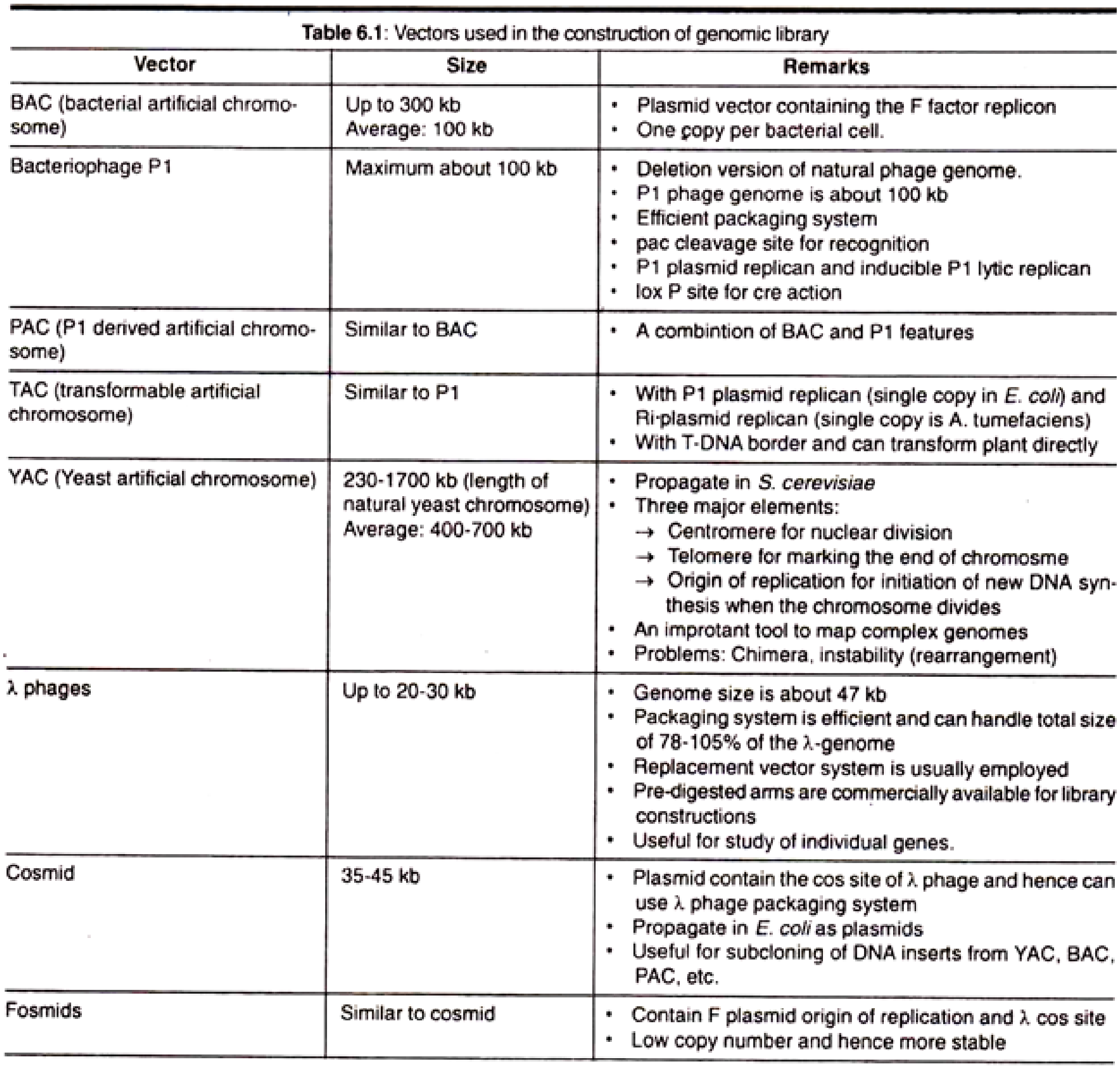


Table 6.1: Vectors used in the construction of genomic library

Vector	Size	Remarks
BAC (bacterial artificial chromosome)	Up to 300 kb Average: 100 kb	• Plasmid vector containing the F factor replicon • One copy per bacterial cell.
Bacteriophage P1	Maximum about 100 kb	• Deletion version of natural phage genome. • P1 phage genome is about 100 kb • Efficient packaging system • pac cleavage site for recognition • P1 plasmid replican and inducible P1 lytic replican • lox P site for cre action
PAC (P1 derived artificial chromosome)	Similar to BAC	• A combintion of BAC and P1 features
TAC (transformable artificial chromosome)	Similar to P1	• With P1 plasmid replican (single copy in *E. coli*) and Ri·plasmid replican (single copy is A. tumefaciens) • With T-DNA border and can transform plant directly
YAC (Yeast artificial chromosome)	230-1700 kb (length of natural yeast chromosome) Average: 400-700 kb	• Propagate in *S. cerevisiae* • Three major elements: → Centromere for nuclear division → Telomere for marking the end of chromosme → Origin of replication for initiation of new DNA synthesis when the chromosome divides • An improtant tool to map complex genomes • Problems: Chimera, instability (rearrangement)
λ phages	Up to 20-30 kb	• Genome size is about 47 kb • Packaging system is efficient and can handle total size of 78-105% of the λ-genome • Replacement vector system is usually employed • Pre-digested arms are commercially available for library constructions • Useful for study of individual genes.
Cosmid	35-45 kb	• Plasmid contain the cos site of λ phage and hence can use λ phage packaging system • Propagate in *E. coli* as plasmids • Useful for subcloning of DNA inserts from YAC, BAC, PAC, etc.
Fosmids	Similar to cosmid	• Contain F plasmid origin of replication and λ cos site • Low copy number and hence more stable

2. Principle of Genomic Libraries:

A genomic library contains all the sequences present in the genome of an organism (apart from any sequences, such as telomeres that cannot be readily cloned). It is a collection of cloned, restriction-enzyme-digested DNA fragments containing at least one copy of every DNA sequence in a genome. The entire genome of an organism is represented as a set of DNA fragments inserted into a vector molecule.

3. Vectors used for the Construction of Genomic Library:

The choice of vectors for the construction of genomic library depends upon three parameters:

1. The size of the DNA insert that these vectors can accommodate.

2. The size of the library that is necessary to obtain a reasonably complete representation of the entire genome.

3. The total size of the genome of the target organism.

In the case of organism with small genomic sizes, such as E. coli, a genomic library could be constructed by using a plasmid vector. In this case only 5000 clones (of average DNA insert size 5kb) would give a greater than 99% chance of cloning the entire genome (4.6 x10^6 bp).

Most libraries from organisms with larger genomes are constructed using lambda phage, BAC or YAC vectors. These accept DNA inserts of approximately 23,45,350 and 1000kb respectively. Due to this, fewer recombinants are needed for complete genome coverage in comparison to the use of plasmids.

4. Size of Genomic Library:

It is possible to calculate the number (N) of recombinants (plaques or colonies) that must be in a genomic library to give a particular probability of obtaining a given sequence.

The formula is:

N = In (1 – P)/ln (1 – f),

where ‘P’ is the desired probability and ‘f is the fraction of the genome in one insert. For example, for a probability of 0.99 with insert sizes of 20kb this values for the E. coli (4.6 x 10^6 bp) and human (3 x 10^9 bp) genomes are:

$N_{g\ coli}$ = In (1 – 0.99) / In [1 – (2 x 10^4/4.6 x 10^6)] = 1.1 x 10^3

N_{human} = In (1 – 0.99)/ In [1 – (2 x 10^4/3 x 10^9)] = 6.9 x 10^5

These values explain why it is possible to make good genomic libraries from prokaryotes in plasmids where the insert size is 5-10 kb, as only a few thousand recombinants will be needed.

5. Types of Genomic Libraries:

Depending on the source of DNA used forced construction of genomic library it is of following two types:

(a) Nuclear Genomic Library:

This is genomic library which includes the total DNA content of the nucleus. While making such a library we specifically extract the nuclear DNA and use it for the making of the library.

(b) Organelle Genomic Library:

In this case we exclude the nuclear DNA and targets the total DNA of either mitochondria, chloroplast or both.

6. Procedure in the Construction of Genomic Library:

1. Preparing DNA:

The key to generating a high-quality library usually lies in the preparation of the insert DNA. The first step is the isolation of genomic DNA. The procedures vary widely according to the organism under study. Care should be taken to avoid physical damage to the DNA.

If the intention is to prepare a nuclear genomic library, then the DNA in the nucleus is isolated, ignoring whatever DNA is present in the mitochondria or chloroplasts. If the aim is to make an organelle genomic library, then it would be wise to purify the organelles away from the nuclei first and then prepare DNA from them.

2. Fragmentation of DNA:

The DNA is then fragmented to a suitable size for ligation into the vector. This could be done by complete digestion with a restriction endonuclease. But this has a demerit. Digestion by the use of restriction endonuclease produces DNA fragments which are not intact.

To solve this problem we use partial digestion with a frequently cutting enzyme (such as Sau3A, with a four-base-pair recognition site) to generate a random collection of fragments with a suitable size distribution.

Once prepared, the fragments that will form the inserts are often treated with phosphate, to remove terminal phosphate groups. This ensures that separate rate pieces of insert DNA cannot be ligated together before they are ligated into the vector. Ligation of separate fragments is undesirable, as it would generate clones containing non-contiguous DNA, and we would have no way of knowing where the joints lay.

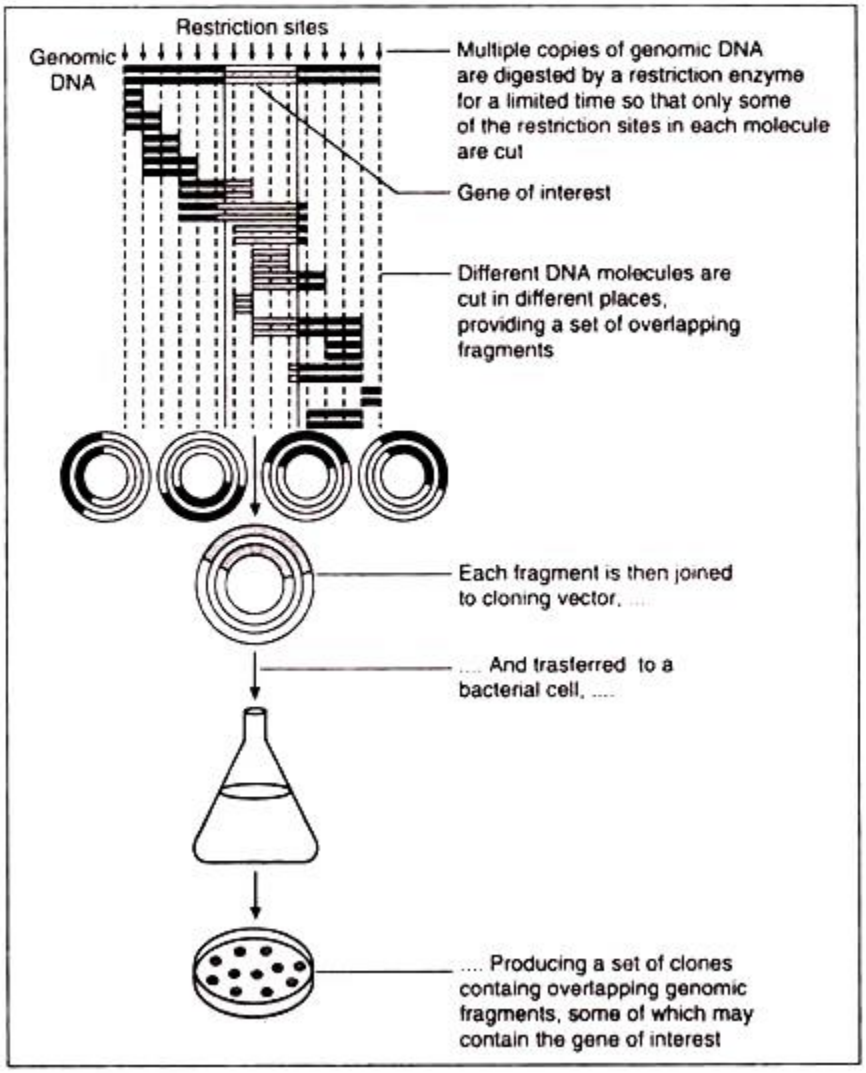

Restriction sites
Genomic DNA
Multiple copies of genomic DNA are digested by a restriction enzyme for a limited time so that only some of the restriction sites in each molecule are cut
Gene of interest
Different DNA molecules are cut in different places, providing a set of overlapping fragments
Each fragment is then joined to cloning vector, ...
.... And trasferred to a bacterial cell, ...
.... Producing a set of clones containg overlapping genomic fragments, some of which may contain the gene of interest

Fig. 6.1: Steps involved in the construction of genomic library

3. Vector Preparation:

This will depend on the kind of vector used. The vector needs to be digested with an enzyme appropriate to the insert material we are trying to clone.

4. Ligation and Introduction into the Host:

Vector and insert are mixed, ligated, packaged and introduced into the host by transformation, infection or' some other technique.

5. Amplification:

This is not always required. Libraries using phage cloning vectors are often kept as a stock of packaged phage. Samples of this can then be plated out on an appropriate host when needed. Libraries constructed in plasmid vectors are kept as collections of plasmid-containing cells, or as naked DNA that can be transformed into host cells when needed.

With storage, naked DNA may be degraded. Larger molecules are more likely to be degraded than smaller ones, so larger recombinants will be selectively lost, and the average insert size will fall.

7. Creation of a Genomic Library using the Phage-λ Vector EMBL3A:

High-molecular-weight genomic DNA is partially digested with Sau3Al. The fragments are treated with phosphatase to remove their 52 phosphate groups. The vector is digested with Bam/HI and EcoRI, which cut within the poly-linker sites.

The tiny BamHI/EcoRl poly-linker fragments are discarded in the iso-propanol precipitation, or alternatively the vector arms may be purified by preparative agarose gel electrophoresis. The vector arms are then ligated with the partially digested genomic DNA.

The phosphatase treatment prevents the genomic DNA fragments from ligating together. Non-recombinant vector cannot reform because the small poly-linker fragments have been discarded. The only package able molecules are recombinant phages. These are obtained as plaques on a P2 lysogen of sup+ E. coli. The Spi” selection ensures recovery of recombinant phage plaques.

8. Problems Associated with the Construction of Genomic Library:

In the making of a genomic library we digest the total genomic DNA with a restriction endonuclease, such as EcoRl, insert the fragments into a suitable phage X vector, and then attempt to isolate the desired clone. How many recombinants would we have to screen in order to isolate the right one?

Let us assume that EcoRI gives an average of about 4kb of DNA fragment, and given that the size of the human haploid genome is 2.8 x 106kb, it is clear that over 7 x 10^5 independent recombinants must be prepared and screened in order to obtain a desired sequence. In other words, we have to obtain a very large number of recombinants, which is a very labour intensive procedure.

There are three problems associated with the above approach:

1. The gene may be cut internally one or more times by Eco RI so that it is not obtained as a single fragment. This is likely if the gene is large.

2. Many times while making a library we want to obtain extensive regions flanking the gene or whole gene clusters. Fragments averaging about 4 kb are likely to be inconveniently short.

3. The obtained gene fragment may be larger than the size which the vector can accept. In this case the appropriate gene would not be cloned at all.

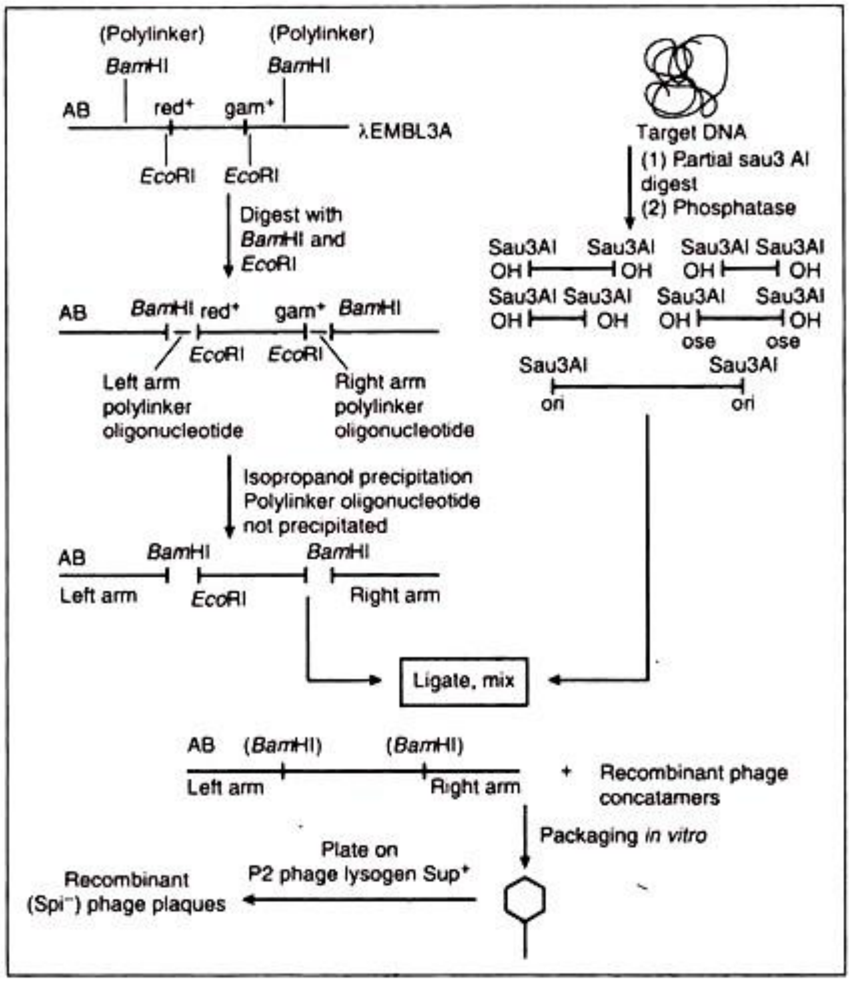


Fig. 6.2: Creation of a genomic DNA library using the phage-λ vector EMBL3A

These problems can be overcome by cloning random DNA fragments of a large size. Since the DNA is randomly fragmented, there will be no exclusion of any DNA sequence. Also in this case the clones will overlap one another allowing the sequence of very large genes to be assembled. Because of the larger size of each cloned DNA fragment fewer clones are required for a complete or nearly complete library.

Now again we have a problem. How can appropriately sized random fragments be produced? Various methods are available of which random breakage by mechanical shearing is the most appropriate one. This is because the average fragment size can be controlled. Along with this the insertion of the resulting fragments into vectors requires additional modification steps.

To achieve this the strategy devised by Maniatis et al. (1978) is the most followed one. In this method the target DNA is digested with a mixture of two restriction enzymes. These enzymes have tetra-nucleotide recognition sites which occur frequently in the target DNA. The restriction digestion by using these enzymes produces fragments having an average size of less than 1 kb.

However, only a partial restriction digest is carried out, and therefore the majority of the fragments are large (in the range 10-30 kb). Given that the chances of cutting at each of the available restriction sites are more or less equivalent, such a reaction effectively produces a random set of overlapping fragments.

These can be separated from each other on the basis of their size (size fractionation), e.g., by gel electrophoresis. This results in the generation of a random population of fragments of about 20kb which are suitable for insertion into a e replacement vector.

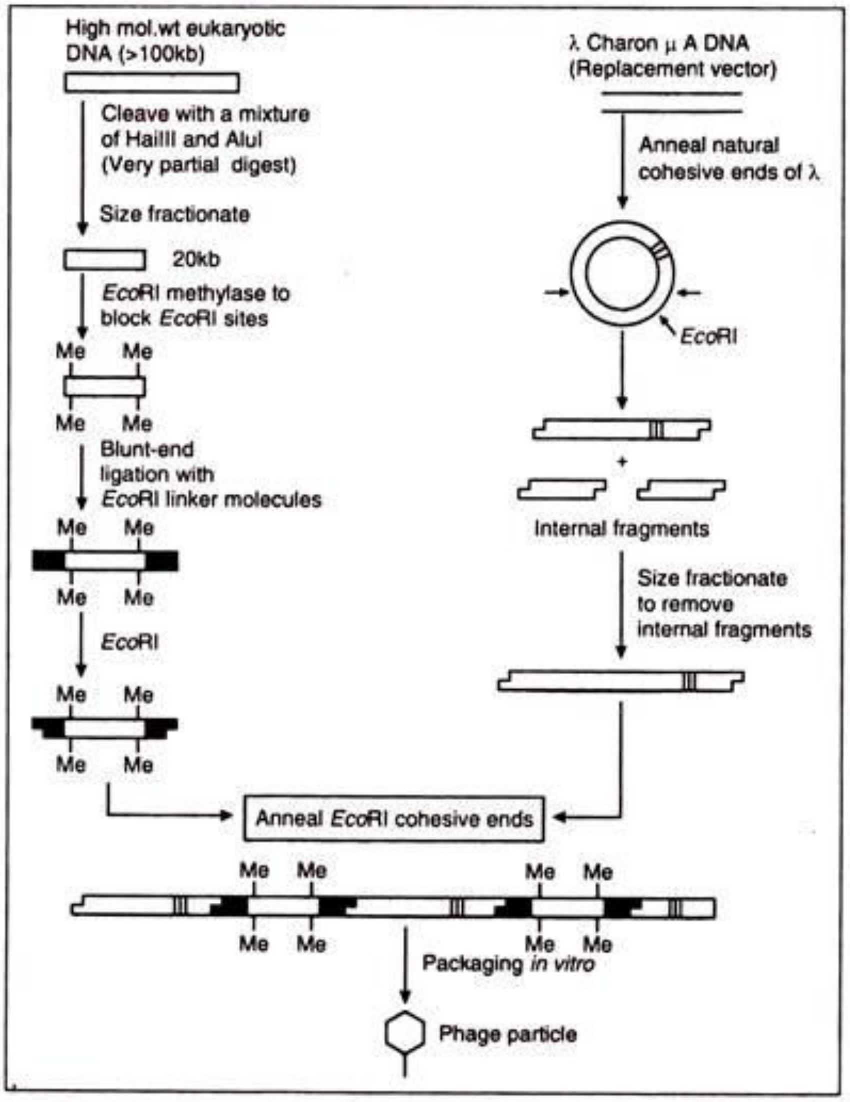

High mol.wt eukaryotic DNA (>100kb)
Cleave with a mixture of HaiIII and AluI (Very partial digest)
Size fractionate
20kb
EcoRI methylase to block EcoRI sites
Me
Blunt-end ligation with EcoRI linker molecules
EcoRI
λ Charon μ A DNA (Replacement vector)
Anneal natural cohesive ends of λ
EcoRI
+
Internal fragments
Size fractionate to remove internal fragments
Anneal EcoRI cohesive ends
Packaging in vitro
Phage particle

Fig. 6.3: Maniatis' strategy for the construction of genomic library

9. Storage of Genomic Library:

Once a genomic library has been made it forms a useful resource for subsequent experiments as well as for the initial purpose for which it was produced. Therefore, it is necessary to store it safely for future use. A random library will consist of a test tube containing a suspension of bacteriophage particle (for a phage vector).

The libraries are stored at – 80°C. Bacterial cells in a plasmid library are protected from the adverse effects of freezing by glycerol, while phage libraries are cryoprotected by dimethyl sulfoxide (DMSO).

10. Disadvantages of Genomic Library:

The main reason behind making a genomic library is to identify a clone from the library which encodes a particular gene or genes of interest. Genomic libraries are particularly useful when you are working with prokaryotic organisms, which have relatively small genomes.

On the face of it, genome libraries might be expected to be less practical when you are working with eukaryotes, which have very large genomes containing a lot of DNA which does not code for proteins.

A library representation of a eukaryotic organism would contain a very large number of clones, many of which would contain non-coding DNA such as repetitive DNA and regulatory regions. Also, eukaryotic genes often contain introns, which are un-translated regions interrupting the coding sequence.

These regions are normally copied into mRNA in the nucleus but spliced out before the mature mRNA is exported to the cytoplasm for translation into protein. Prokaryotic organisms are unable to do this processing so the mature mRNA cannot be made in E. coli and the protein will not be expressed.

If your screening method requires that the gene be expressed it will not work with a genomic library from a eukaryotic organism.

12. Applications of Genomic Library:

Genomic library has following applications:

1. It helps in the determination of the complete genome sequence of a given organism.

2. It serves as a source of genomic sequence for generation of transgenic animals through genetic engineering.

3. It helps in the study of the function of regulatory sequences in vitro.

4. It helps in the study of genetic mutations in cancer tissues.

5. Genomic library helps in identification of the novel pharmaceutical important genes.

6. It helps us in understanding the complexity of genomes.

CONSTRUCTION OF cDNA LIBRARIES

cDNA Library:

A cDNA library is defined as a collection of cDNA fragments, each of which has been cloned into a separate vector molecule.

Principle of cDNA Library:

In the case of cDNA libraries we produce DNA copies of the RNA sequences (usually the mRNA) of an organism and clone them. It is called a cDNA library because all the DNA in this library is complementary to mRNA and are produced by the reverse transcription of the latter.

Much of eukaryotic DNA consists of repetitive sequences that are not transcribed into mRNA and the sequences are not represented in a cDNA library. It must be noted that prokaryotes and lower eukaryotes do not contain introns, and preparation of cDNA is generally unnecessary for these organisms. Hence, cDNA libraries are produced only from higher eukaryotes.

Vectors used in the Construction of cDNA Library:

Both the bacterial and bacteriophage DNA are used as vectors in the construction of cDNA library.

The following table gives detailed information:

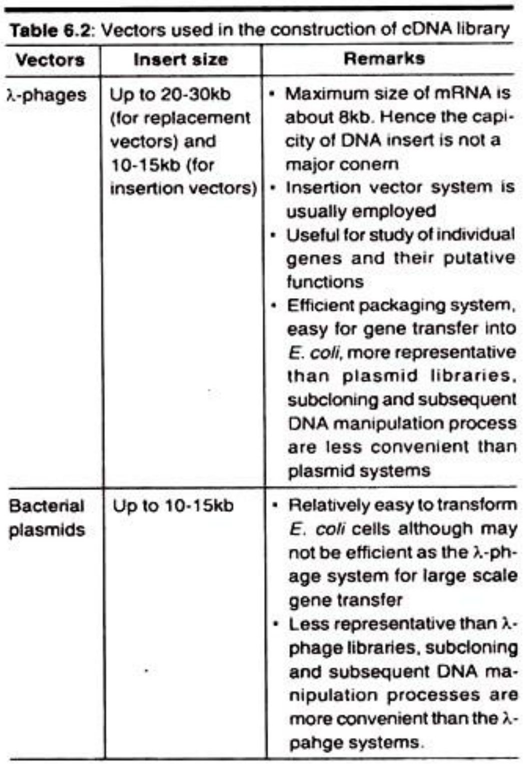


Table 6.2: Vectors used in the construction of cDNA library

Vectors	Insert size	Remarks
λ-phages	Up to 20-30kb (for replacement vectors) and 10-15kb (for insertion vectors)	• Maximum size of mRNA is about 8kb. Hence the capicity of DNA insert is not a major conern • Insertion vector system is usually employed • Useful for study of individual genes and their putative functions • Efficient packaging system, easy for gene transfer into *E. coli*, more representative than plasmid libraries, subcloning and subsequent DNA manipulation process are less convenient than plasmid systems
Bacterial plasmids	Up to 10-15kb	• Relatively easy to transform *E. coli* cells although may not be efficient as the λ-phage system for large scale gene transfer • Less representative than λ-phage libraries, subcloning and subsequent DNA manipulation processes are more convenient than the λ-pahge systems.

Procedure in the Construction of cDNA Library:

The steps involved in the construction of a cDNA library are as follows:

1. Extraction of mRNA from the eukaryotic Cell:

Firstly, the mRNA is obtained and purified from the rest of the RNAs. Several methods exist for purifying RNA such as trizol extraction and column purification. Column purification is done by using oligomeric dT nucleotide coated resins where only the mRNA having the poly-A tail will bind.

The rest of the RNAs are eluted out. The mRNA is eluted by using eluting buffer and some heat to separate the mRNA strands from oligo-dT.

2. Construction of cDNA from the Extracted mRNA (Fig. 6.4):

There are different strategies for the construction of a cDNA. These are discussed as follows:

(a) The RNase Method:

The principle of this method is that a complementary DNA strand is synthesized using reverse transcriptase to make an RNA: DNA duplex. The RNA strand is then nicked and replaced by DNA. In this method the first step is to anneal a chemically synthesized oligo-dT primer to the 3′ polyA-tail of the RNA.

The primer is typically 10-15 residues long, and it primes (by providing a free 3′ end) the synthesis of the first DNA strand in the presence of reverse transcriptase and deoxyribonucleotides. This leaves an RNA: DNA duplex.

The next step is to replace the RNA strand with a DNA strand. This is done by using RNase H enzyme which removes the RNA from RNA: DNA duplex. The DNA strand thus left behind is then considered as the template and the second DNA strand is synthesized by the action of DNA polymerase II.

(b) The Self-Priming method:

This involved the use of an oligo-dT primer annealing at the polyadenylate tail of the mRNA to prime first DNA strand synthesis against the mRNA. This cDNA thus formed has the tendency to transiently fold back on itself, forming a hairpin loop. This results in the self-priming of the second strand.

After the synthesis of the second DNA strand, this loop must be cleaved with a single-strand-specific nuclease, e.g., SI nuclease, to allow insertion into the cloning vector. This method has a serious disadvantage. The cleavage with SI nuclease results in the loss of a certain amount of sequence at the 5′ end of the clone.

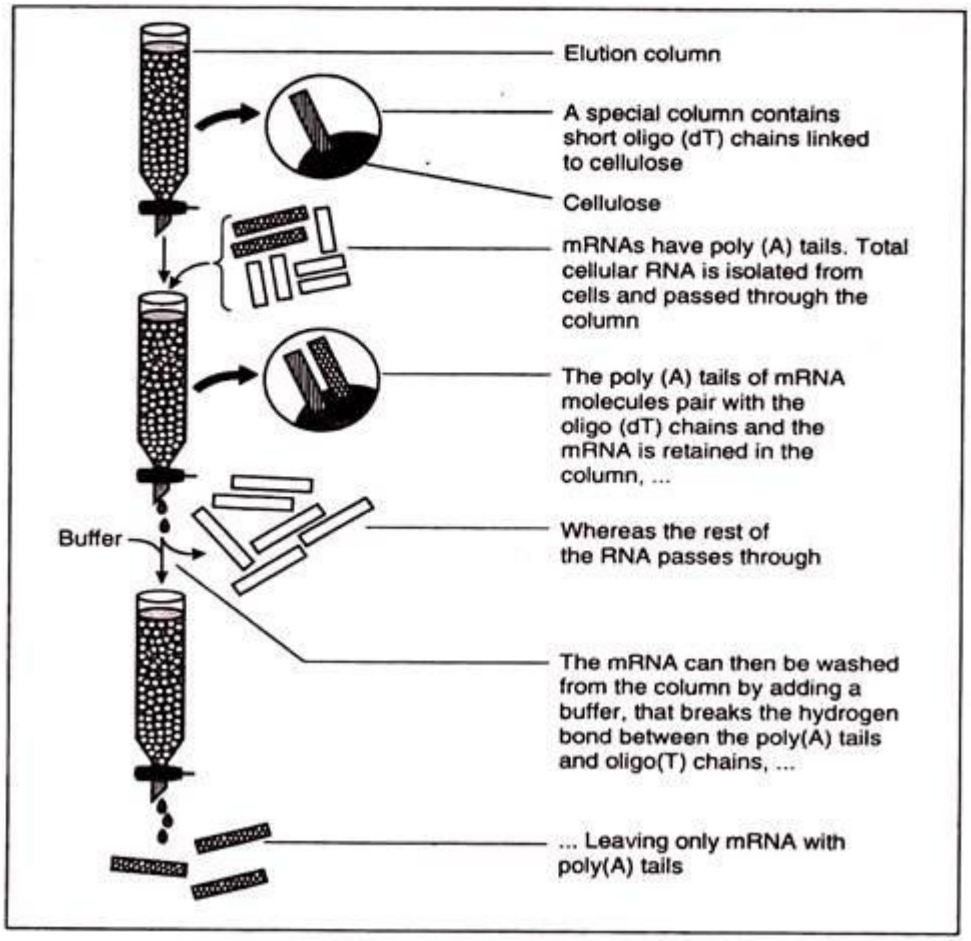

Elution column
A special column contains short oligo (dT) chains linked to cellulose
Cellulose
mRNAs have poly (A) tails. Total cellular RNA is isolated from cells and passed through the column
The poly (A) tails of mRNA molecules pair with the oligo (dT) chains and the mRNA is retained in the column, ...
Buffer
Whereas the rest of the RNA passes through
The mRNA can then be washed from the column by adding a buffer, that breaks the hydrogen bond between the poly(A) tails and oligo(T) chains, ...
... Leaving only mRNA with poly(A) tails

Fig. 6.4: Extraction of mRNA from the eukaryotic cell

(c) Land et al. Strategy:

After first-strand synthesis, which is primed with an oligo- dT primer as usual, the cDNA is tailed with a string of cytidine residues using the enzyme terminal transferase. This artificial oligo-dC tail is then used as an annealing site for a synthetic oligo-dG primer, allowing synthesis of the second strand.

(d) Homopolymer Tailing:

This approach uses the enzyme terminal transferase, which can polymerize nucleotides onto the 3′-hydroxyl of both DNA and RNA molecules. We carry out the synthesis of the first DNA strand essentially as before, to produce an RNA: DNA hybrid.

We then use terminal transferase and a single deoxyribonucleotide to add tails of that nucleotide to the 3′ ends of both RNA and DNA strands. The result of this is that the DNA strand now has a known sequence at its 3′ end Typically, dCTP or dATP are used.

A complementary oligomer (synthesized chemically) can now be annealed and used as a primer to direct second strand synthesis. This oligomer (and also the one used for first strand synthesis) may additionally incorporate a restriction site, to help in cloning the resulting double- stranded cDNA.

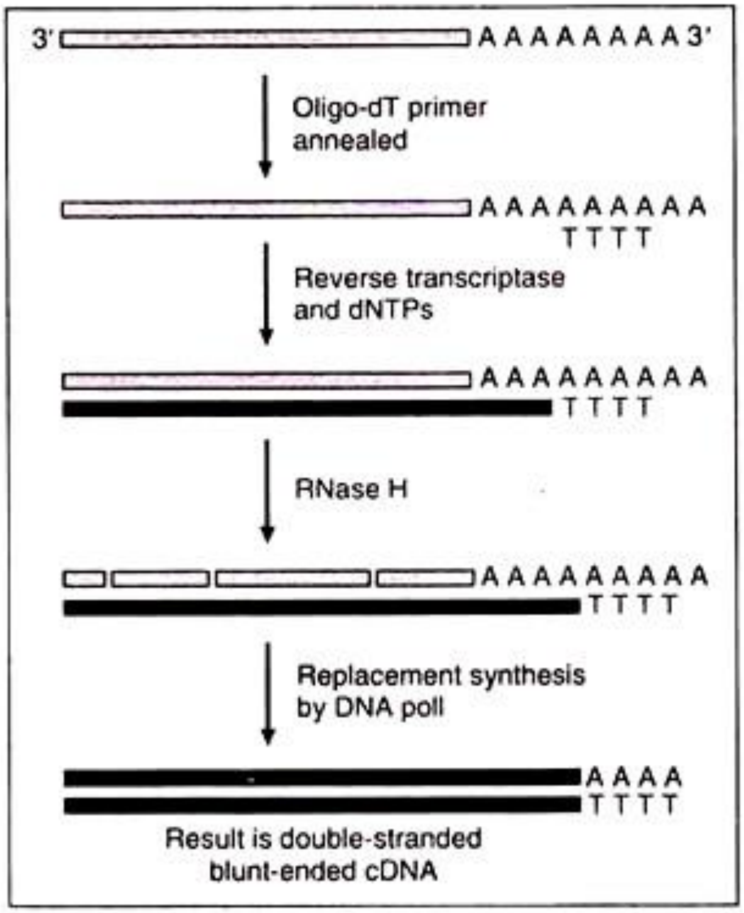

3'
AAAAAAAAA3'
Oligo-dT primer annealed
AAAAAAAAA
TTTT
Reverse transcriptase and dNTPs
AAAAAAAAA
TTTT
RNase H
AAAAAAAAA
TTTT
Replacement synthesis by DNA poll
AAAA
TTTT
Result is double-stranded blunt-ended cDNA

Fig. 6.5: The RNase H method of c-DNA synthesis

(e) Rapid Amplification of cDNA Ends (RACE):

It is sometimes the case that we wish to clone a particular cDNA for which we already have some sequence data, but with particular emphasis on the integrity of the 5′ or 3′ ends. RACE techniques (Rapid Amplification of cDNA Ends) are available for this. The RACE methods are divided into 3’RACE and 5’RACE, according to which end of the cDNA we are interested in.

(a) 3’RACE:

In this type of RACE, reverse transcriptase synthesis of a first DNA strand is carried out using a modified oligo-dT primer. This primer comprises a stretch of unique adaptor sequence followed by an oligo-dT stretch. The first strand synthesis is followed by a second strand synthesis using a primer internal to the coding sequence of interest.

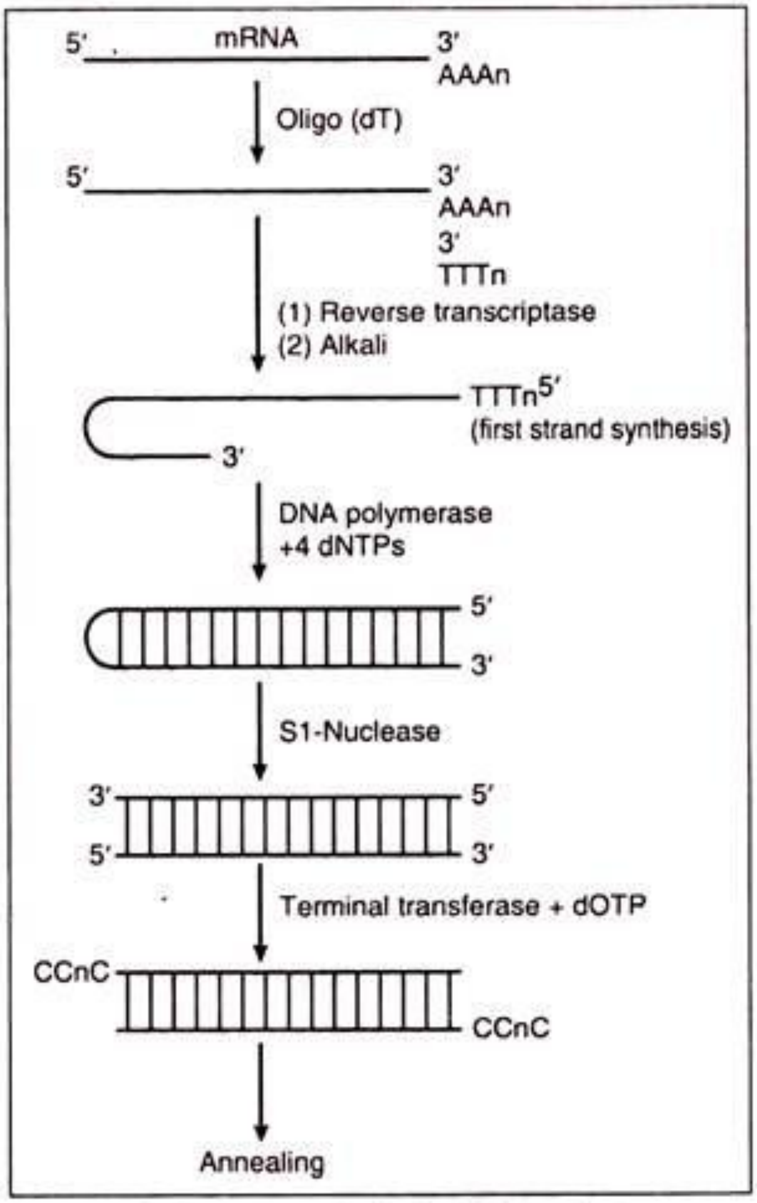


Fig. 6.6: Self-priming method of c-DNA synthesis

This is followed by PCR using

(i) The same internal primer and '

(ii) The adaptor sequence (i.e., omitting the oligo-dT). Although in theory it should be possible to use a simple oligo- dT primer throughout instead of the adaptor-oligo-dT and adaptor combination, the low melting temperature for an oligo-dT primer may interfere with the subsequent rounds of PCR.

(b) 5'RACE:

In this type of RACE first cDNA strand is synthesized with re- verse transcriptase and a primer from within the coding sequence. Unincorporated primer is removed and the cDNA strands are tailed with oligo-dA. A second cDNA strand is then synthesized with an adaptor-oligo-dT primer.

The resulting double-stranded molecules are then subject to PCR using

(i) A primer nested within the coding region and
(ii) The adaptor sequence. A nested primer is used in the final PCR to improve specificity. The adaptor sequence is used in the PCR because of the low melting temperature of a simple oligo-dT primer, as in 3'RACE above. A number of kits for RACE are commercially available.

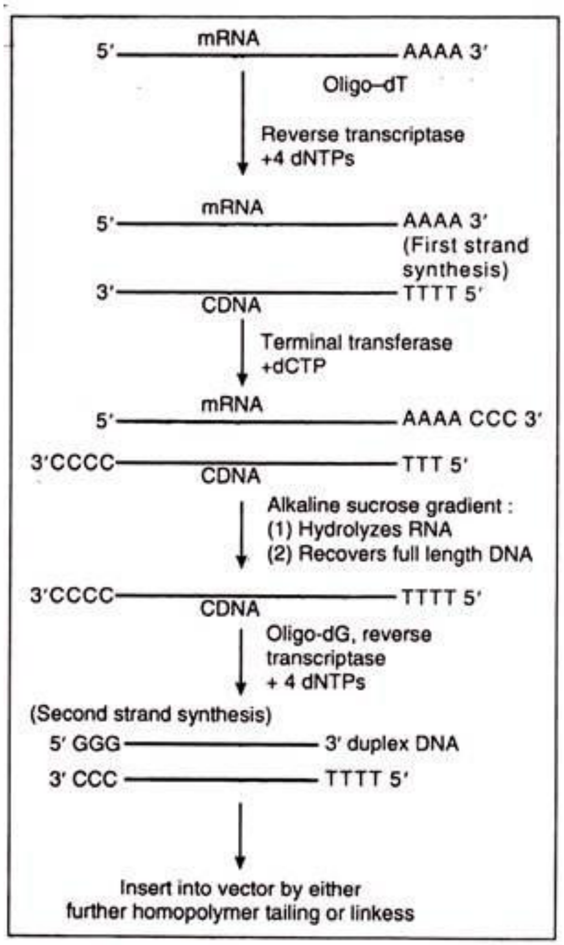


Fig. 6.7: Land et al. strategy

3. Cloning the c-DNA:

(a) Linkers:

The RNaseH and homopolymer tailing methods ultimately generate a collection of double-stranded, blunt-ended cDNA molecules. They must now be attached to the vector molecules. This could be done by blunt-ended ligation, or by the addition of linkers, digestion with the relevant enzyme and ligation into vector.

(b) Incorporation of Restriction Sites:

It is possible to adapt the homopolymer tailing method by using primers that are modified to incorporate restriction. In the diagram shown next page, the oligo-dT primer is modified to contain a restriction site (in the figure, a Sail site GTCGAC).

The 3′ end of the newly synthesized first cDNA strand is tailed with C’s. An oligo-dG primer, again preceded by a Sail site within a short double-stranded region of the oligonucleotide, is then used for second-strand synthesis.

Note that this method requires the use of an oligonucleotide containing a double-stranded region. Such oligonucleotides are made by synthesizing the two strands separately and then allowing them to anneal to one another.

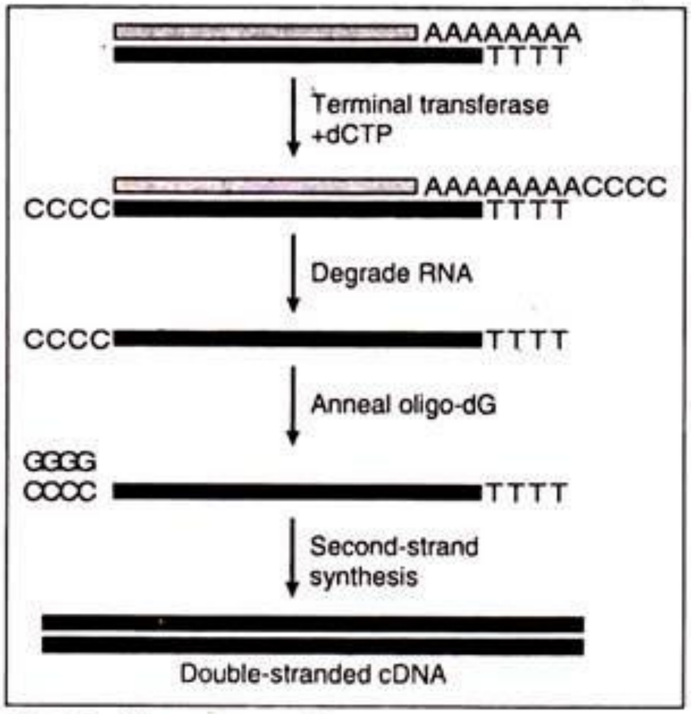


Fig. 6.8: Homopolymer tailing

(c) Homopolymer Tailing of cDNA:

Another option is to use terminal transferase again. Treatment of the blunt-ended double-stranded cDNA with terminal transferase and dCTP leads to the polymerization of several C residues (typically 20 or so) to the 3′ hydroxyl at each end.

Treatment of the vector with terminal transferase and dGTP leads to the incorporation of several G residues onto the ends of the vector. (Alternatively, dATP and dTTP can be used.) The vector and cDNA can now anneal, and the base-paired region is often so extensive that treatment with DNA ligase is unnecessary.

In fact, there may be gaps rather than nicks at the vector insert boundaries, but these are repaired by physiological processes once the recombinant molecules have been introduced into a host.

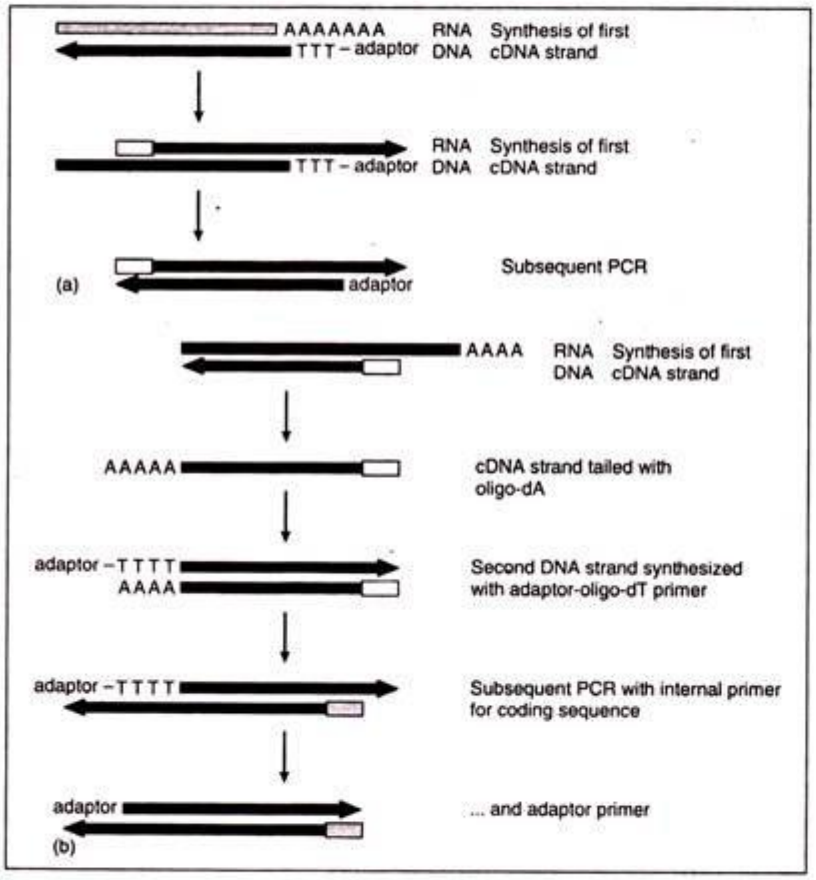

AAAAAAA
TTT – adaptor
RNA
DNA
Synthesis of first cDNA strand
RNA
DNA
Synthesis of first cDNA strand
TTT – adaptor
Subsequent PCR
(a)
adaptor
AAAA
RNA
DNA
Synthesis of first cDNA strand
AAAAA
cDNA strand tailed with oligo-dA
adaptor – TTTT
AAAA
Second DNA strand synthesized with adaptor-oligo-dT primer
adaptor – TTTT
Subsequent PCR with internal primer for coding sequence
adaptor
... and adaptor primer
(b)

Fig. 6.9: RACE. (a) 3'RACE. The first primer is the oligo-dT-adaptor molecule. The second primer (open box) is internal to the coding sequence of interest. This is used in conjunction with the adaptor primer (rather than the oligo-dT-adaptor primer) in subsequent PCR. (b) 5'RACE. Synthesis of the first cDNA strand uses a primer within the coding region (open box). The first cDNA strand is tailed with oligo-dA. A second DNA strand is synthesized with an adaptor-oligo-dT primer. This is followed by PCR with (i) a primer nested within the coding sequence (shaded box) and (ii) the adaptor.

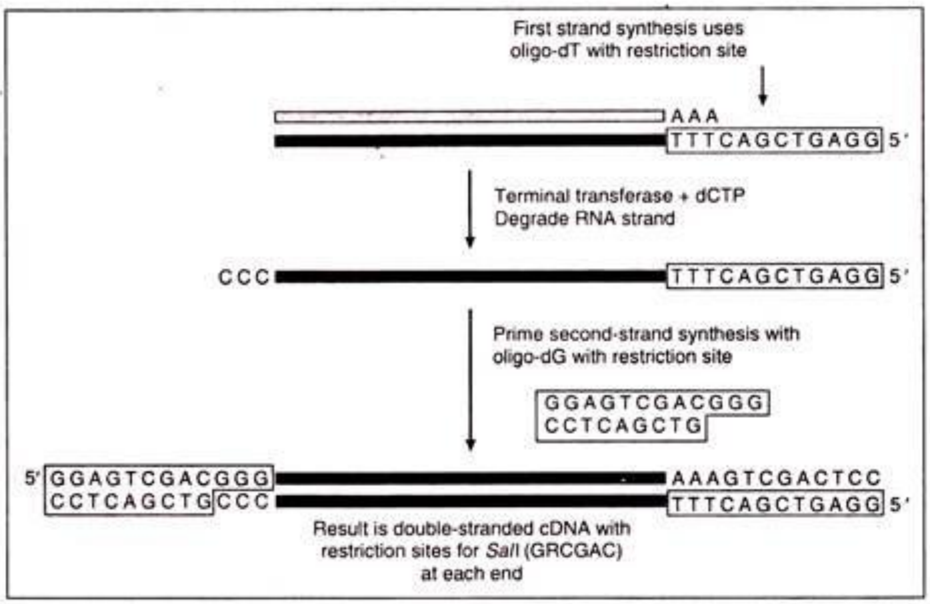

First strand synthesis uses oligo-dT with restriction site
AAA
TTTCAGCTGAGG 5'
Terminal transferase + dCTP
Degrade RNA strand
CCC
TTTCAGCTGAGG 5'
Prime second-strand synthesis with oligo-dG with restriction site
GGAGTCGACGGG
CCTCAGCTG
5' GGAGTCGACGGG
CCTCAGCTG CCC
AAAGTCGACTCC
TTTCAGCTGAGG 5'
Result is double-stranded cDNA with restriction sites for SalI (GRCGAC) at each end

Fig. 6.10: Modification of homopolymer tailing, incorporating restriction sites

Advantages of cDNA Library:

A cDNA library has two additional advantages. First, it is enriched with fragments from actively transcribed genes. Second, introns do not interrupt the cloned sequences; introns would pose a problem when the goal is to produce a eukaryotic protein in bacteria, because most bacteria have no means of removing the introns.

Disadvantages of cDNA Library:

The disadvantage of a cDNA library is that it contains only sequences that are present in mature mRNA. Introns and any other sequences that are altered after transcription are not present; sequences, such as promoters and enhancers that are not transcribed into RNA also are not present in a cDNA library.

It is also important to note that the cDNA library represents only those gene sequences expressed in the tissue from which the RNA was isolated. Furthermore, the frequency of a particular DNA sequence in a cDNA library depends on the abundance of the corresponding mRNA in the given tissue. In contrast, almost all genes are present at the same frequency in a genomic DNA library.

Applications of cDNA Library:

Following are the applications of cDNA libraries:

1. Discovery of novel genes.
2. Cloning of full-length cDNA molecules for in vitro study of gene function.
3. Study of the repertoire of mRNAs expressed in different cells or tissues.
4. Study of alternative splicing in different cells or tissues.

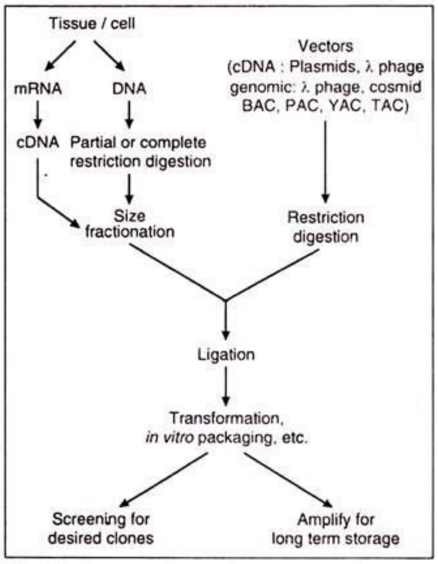

Tissue / cell
mRNA
DNA
cDNA
Partial or complete restriction digestion
Size fractionation
Vectors (cDNA : Plasmids, λ phage genomic: λ phage, cosmid BAC, PAC, YAC, TAC)
Restriction digestion
Ligation
Transformation, *in vitro* packaging, etc.
Screening for desired clones
Amplify for long term storage

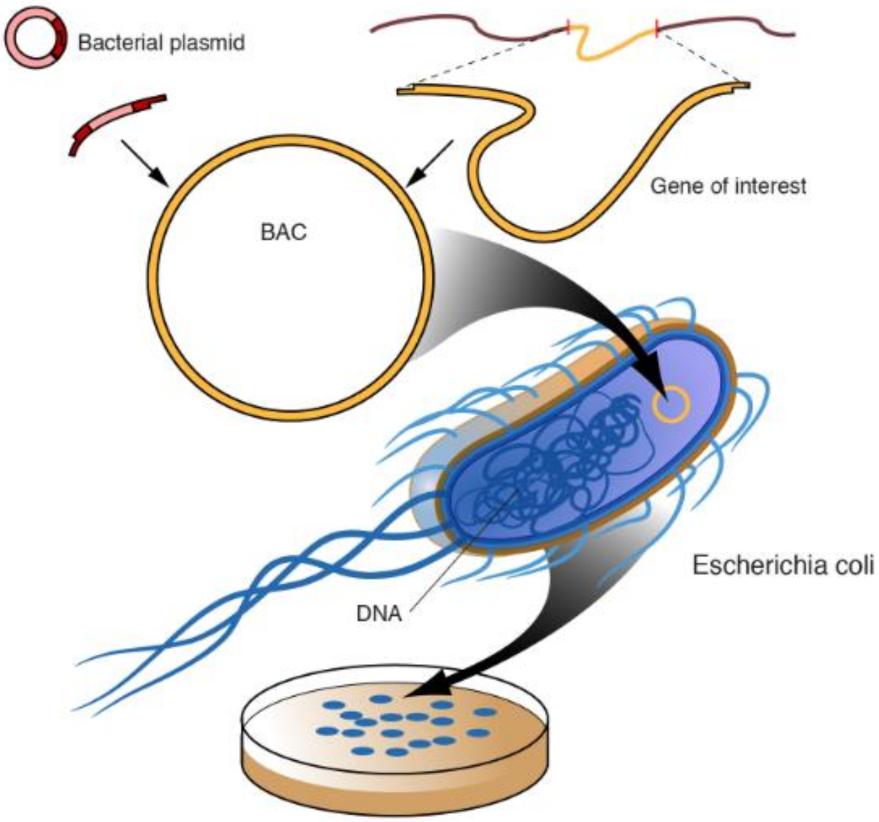
Fig. 6.11: Flow chart showing the construction of genomic and cDNA library

ARTIFICIAL CHROMOSOMES – BACS AND YACS

Bacterial Artificial Chromosome

A bacterial artificial chromosome (BAC) is an engineered DNA molecule used to clone DNA sequences in bacterial cells (for example, E. coli). BACs are often used in connection with DNA sequencing. Segments of an organism's DNA, ranging from 100,000 to about 300,000 base pairs, can be inserted into BACs. The BACs, with their inserted DNA, are then taken up by bacterial cells. As the bacterial cells grow and divide, they amplify the BAC DNA, which can then be isolated and used in sequencing DNA.

A large piece of DNA can be engineered in a fashion that allows it be propagated as a circular artificial chromosome in bacteria--so-called bacterial artificial chromosome, or BAC. Each BAC is a DNA clone containing roughly 100 to 300 thousand base pairs of cloned DNA. Because the BAC is much smaller than the endogenous bacterial chromosome, it is straightforward to purify the BAC DNA away from the rest of the bacteria cell's DNA, and thus have the cloned DNA in a purified form. This and other powerful features of BACs have made them extremely useful for mapping and sequencing mammalian genomes.

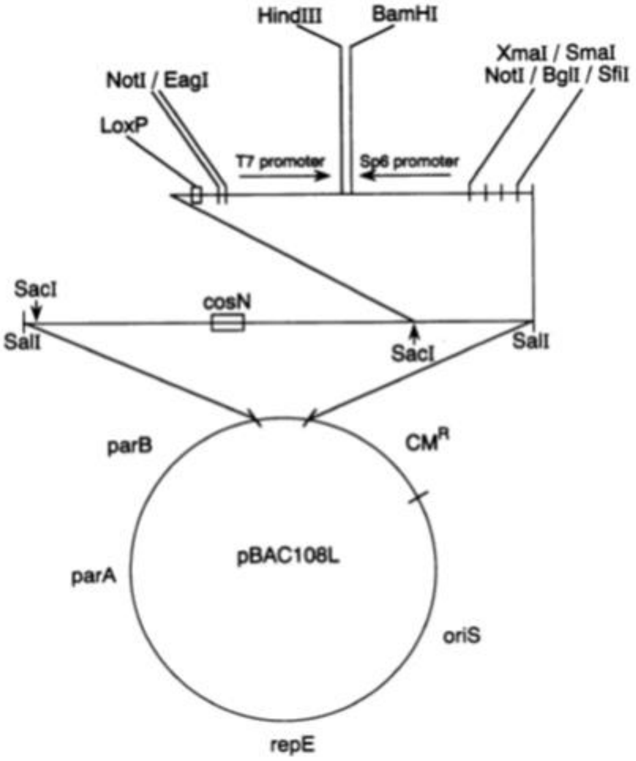


Bacterial artificial chromosomes (BACs)

- Bacterial artificial chromosomes (BACs) are simple plasmid which is designed to clone very large DNA fragments ranging in size from 75 to 300 kb.
- BACs basically have marker like sights such as antibiotic resistance genes and a very stable origin of replication (ori) that promotes the distribution of plasmid after bacterial cell division and maintaining the plasmid copy number to one or two per cell.
- BACs are basically used in sequencing the genome of organisms in genome projects (example: BACs were used in human genome project).
- Several hundred thousand base pair DNA fragments can be cloned using BACs.

Characteristic features of BAC vectors

- The original BAC vector, pBAC108L, is based on a mini-F plasmid, pMBO131 (Figure 1) which encodes genes essential for self-replication and regulates its copy number inside a cell. The unidirectional self-replicating genes are *oriS* and *repE* while *parA* and *parB* maintain copy number to one or two for each *E. coli* genome.
- Multiple cloning sites is present, flanked by "universal promoters" T7 and SP6, all flanked by GC-rich restriction enzyme sites for insert excision.



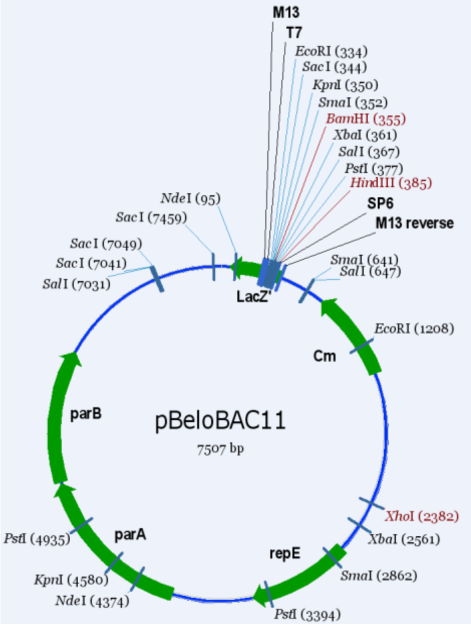
- Presence of *cosN* and *loxP* sites(cloned in by bacteriophage l terminase and P1 Cre recombinase, respectively) permits linearization of the plasmid for convenient restriction mapping.
- There is a chloramphenicol resistance gene for negative selection of non-transformed bacteria.
- Vector is 6900 bp in length and is capable of maintaining insert DNA in excess of 300 kilobases (kb).

Other BAC Vectors

- There have been many modifications done to increase the ease-of-use as well as for use in specific systems and situations.
- pBeloBAC11 2 and pBACe3.6 are modified BAC vectors based on pBAC108L and are commonly used as a basis for further modification.

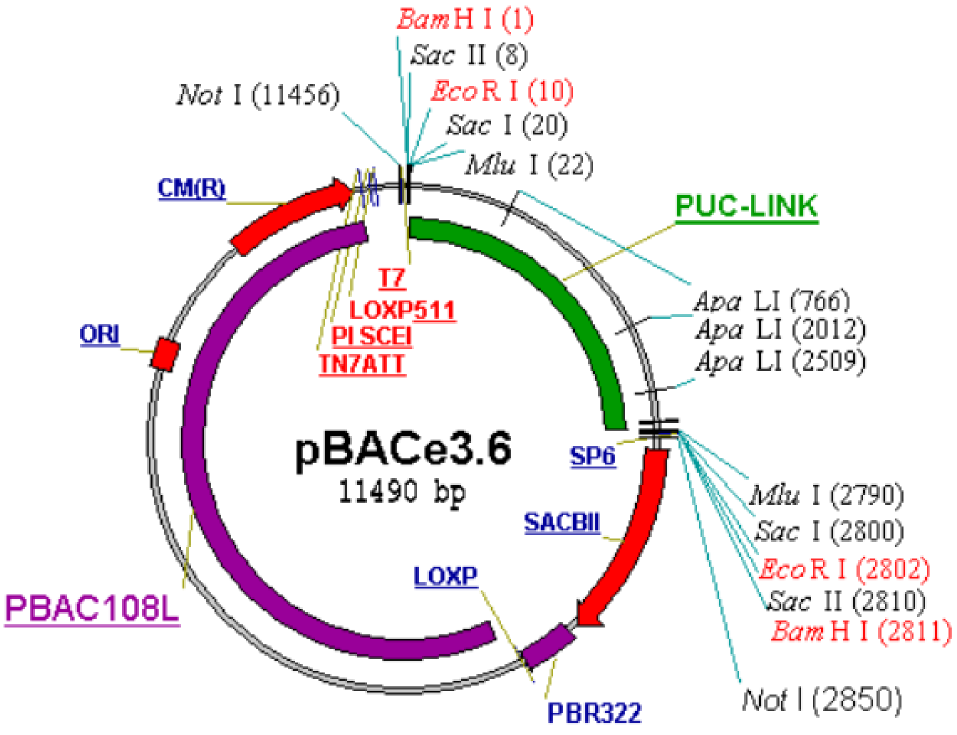
pBeloBAC11

- The primary characteristic of this vector is the addition of a *lacZ* gene into the multiple cloning sites 2 of pBAC108L.
- Plates supplemented with X-gal/IPTG, an intact lacZ gene encodes b-galactosidase which catalyses the supplemented substrate into a blue substance. Successful ligation of insert DNA into the vector inactivates *lacZ*, generating white colonies, indicating the presence of a successful vector-insert ligation.
- It is still a low-copy number plasmid due to presence of *parA* and *parB*.
- Size of vector is 7507 bp in length.



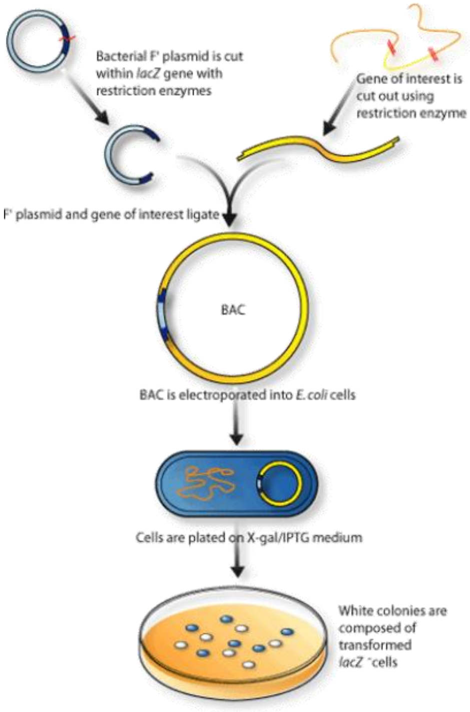
pBACe3.6

- This vector is based on pBAC108L but is more highly modified than pBeloBAC11.
- In order to overcome the issue of low plasmid copy numbers, the P1 replicon in F’ was deleted and a removable high copy number replicon originating from an inserted pUC19 was introduced.
- This vector contains 2.7 kb pUClink stuffer fragment which is flanked by two sets of six restriction sites within a *sacB* region.
- Levansucrase, a product of *sacB* gene, which converts sucrose (supplemented in the media) to levan, which is toxic to *E. coli* host cells. Hence, if the vector is re-ligated without an insert, the functional *sacB* produces levansucrase and the cells die before forming colonies. Successful ligation of an insert into the vector increases the distance from the promoter to the coding region of *sacB*, disrupting toxic gene expression in the presence of sucrose.



- In addition to these vectors, there are many specialized BAC vectors carrying a variety of different combinations of drug resistance genes. Besides, many different selection mechanisms and markers are available. Modifications of cloning sites (unique restriction endonuclease sites) are also common as per the addition of genes and promoters specific to different strains of bacteria.

Development of BAC vector


Bacterial F' plasmid is cut within *lacZ* gene with restriction enzymes

Gene of interest is cut out using restriction enzyme

F' plasmid and gene of interest ligate

BAC is electroporated into *E.coli* cells

Cells are plated on X-gal/IPTG medium

White colonies are composed of transformed *lacZ* ⁻cells

Advantages of BAC Vectors

1. The large size of BACs help to minimize site of integration effects, a phenomenon which has been defined as endogenous sequences (such as gene coding regions and distal regulatory elements) to be disrupted, and to produce potentially undesirable phenotypes in gene cloning technology.
2. Endogenous gene expression more accurately than other cloning systems.
3. The human genome BACs consist of the full gene structure (which play very important role in gene regulation). Therefore the human genome BACs will ensure full mRNA processing and splicing when genes are transcribed, and produce the full complement of protein isoforms once mRNAs are translated.
4. It can be transfected and expressed in mammalian cell lines even if transfection efficiency and copy numbers are low.

Disadvantages of BAC vectors

1. A construct containing a large genomic fragment is likely to contain non-related genes which may lead to indirect, non-specific gene expression and unanticipated changes in the cell phenotype.
2. Recombinant BAC constructs can be time-consuming and labor-intensive.
3. The large size BAC DNA constructs are more easily degraded and sheard during manipulation before transfection.

Applications of BAC vectors

BACs are useful for the construction of genomic libraries but their range of use is vast. It spans from basic science to economically rewarding industrial research, and fields as prosaic as animal husbandry.

In genomic analyses, it helps in determining phylogenetic lineage det between species.

Helps in study of horizontal gene transfer and since bacterial genes are usually clustered, the ability of BAC vectors to accommodate large inserts has allowed the study of entire bacterial pathways.

By isolating DNA directly from soil or from marine environments, the “metagenomes” of those organisms which are either uncultureable or are termed viable but uncultureable can be cloned into BAC vectors and indirectly studied.

In industrial research fields where BAC vectors are invaluable tools in cataloguing novel genomes is in the discovery of novel enzymes. Work has been done on identifying enzymes that are involved in biopolymer hydrolysis or even radioactive waste management.

BAC vectors have been instrumental in studying large double stranded DNA viruses both from an academic point of view and as a tool to develop improved vaccines.

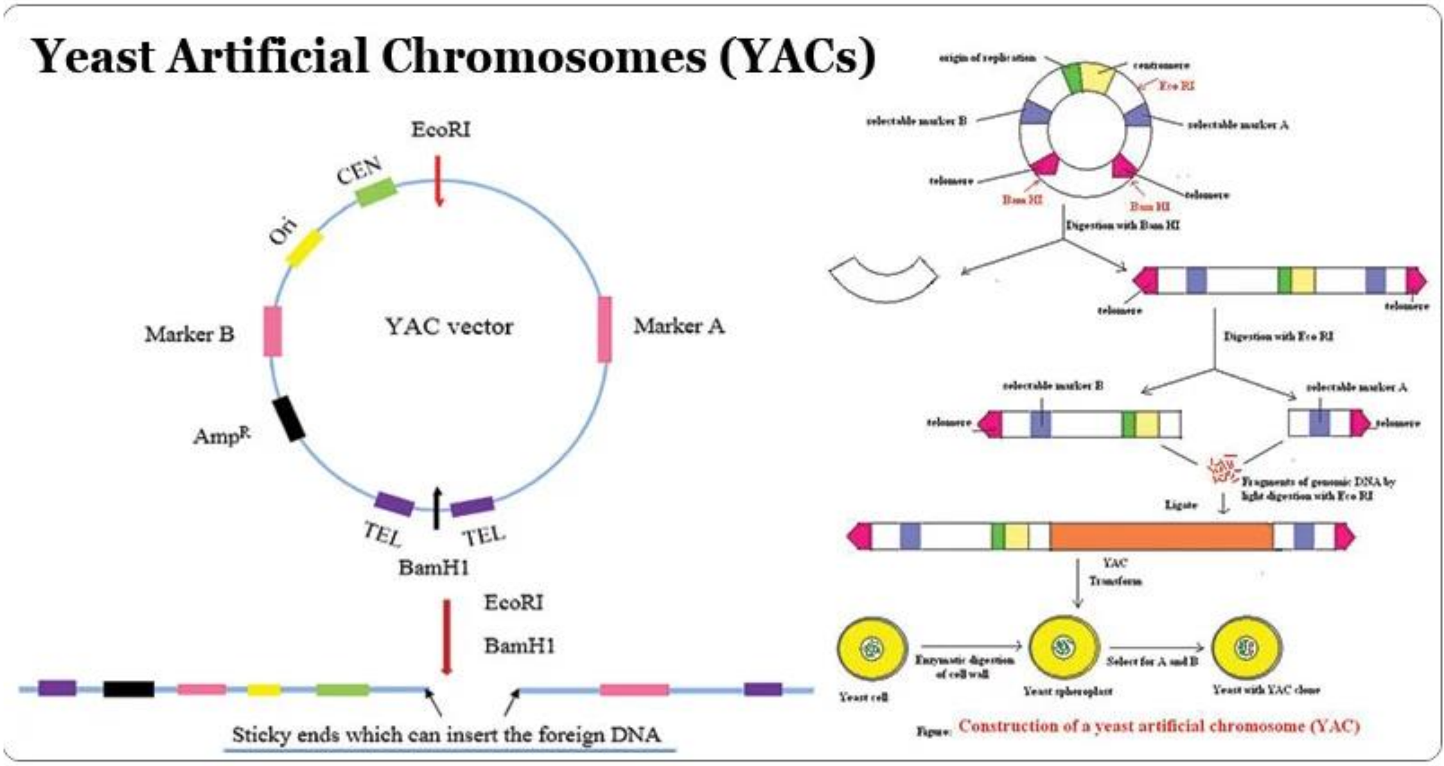
In genomic research, high throughput determination of gains and losses of genetic material using high resolution BAC arrays and comparative genomic hybridization (CGH) have been developed into the new tools for translational research in solid tumors and neurodegenerative disorders.

BAC technology is becoming the most upcoming method for genome sequencing. The technique uses an overlapping tailing part of large genomic fragments (150-200 kb) maintained within BACs. Every individual BAC is shotgun sequenced, where these large overlapping sequences of the BACs are assembled to produce the whole genome sequence.

BACs have also been used in mammalian genome mapping, genomic imprinting, vaccine development, gene therapy and studies of the evolutionary history and functional dynamics of sex chromosomes have recently been possible using BAC libraries.

Yeast Artificial Chromosomes (YACs)

- Yeast artificial chromosomes (YACs) are genetically engineered chromosomes derived from the DNA of the yeast.
- It is a human-engineered DNA molecule used to clone DNA sequences in yeast cells.
- They are the products of a recombinant DNA cloning methodology to isolate and propagate very large segments of DNA in a yeast host.
- By inserting large fragments of DNA, the inserted sequences can be cloned and physically mapped using a process called chromosome walking.
- The amount of DNA that can be cloned into a YAC is, on average, from 200 to 500 kb.
- However, as much as 1 Mb (mega, 106) can be cloned into a YAC.



Structure of Yeast Artificial Chromosomes

A yeast artificial chromosome cloning vector consists of two copies of a yeast telomeric sequence (telomeres are the sequences at the ends of chromosomes), a yeast centromere, a yeast ars (an autonomously replicating sequence where DNA replication begins), and appropriate selectable markers.

Working Principle of Yeast Artificial Chromosomes

The yeast artificial chromosome, which is often shortened to YAC, is an artificially constructed system that can undergo replication. The design of a YAC allows extremely large segments of genetic material to be inserted. Subsequent rounds of replication produce many copies of the inserted sequence, in a genetic procedure known as cloning.

- The principle is similar to that for plasmids or cosmids.
- The experimenter introduces some typical elements that are necessary for correct replication.
- In the case of YACs, the replication origins are the centromeres and telomeres of the yeast chromosomes, which must be inserted into the DNA being cloned.
- The constructs can be transformed in yeast Spheroplast and are then replicated there.
- In contrast to the vectors, YACs are not circular; they are made of linear DNA.

Process of Yeast Artificial Chromosomes

- YAC vector is initially propagated as circular plasmid inside bacterial host utilizing bacterial ori sequence.
- The circular plasmid is cut at a specific site using restriction enzymes to generate a linear chromosome with two telomere sites at terminals.
- The linear chromosome is again digested at a specific site with two arms with different selection marker.
- The genomic insert is then ligated into YAC vector using DNA ligase enzyme.
- The recombinant vectors are transformed into yeast cells and screened for the selection markers to obtain recombinant colonies.

Advantages of Yeast Artificial Chromosomes

- Yeast artificial chromosomes (YACs) provide the largest insert capacity of any cloning system.
- Yeast expression vectors, such as YACs, YIPs (yeast integrating plasmids), and YEPs (yeast episomal plasmids), have advantageous over bacterial artificial chromosomes (BACs). They can be used to express eukaryotic proteins that require post-translational modification.
- A major advantage of cloning in yeast, a eukaryote, is that many sequences that are unstable, underrepresented, or absent when cloned into prokaryotic systems, remain stable and intact in YAC clones.
- It is possible to reintroduce YACs intact into mammalian cells where the introduced mammalian genes are expressed and used to study the functions of genes in the context of flanking sequences.

Uses of Yeast Artificial Chromosomes

- Yeast artificial chromosomes (YACs) were originally constructed in order to study chromosome behavior in mitosis and meiosis without the complications of manipulating and destabilizing native chromosomes.
- YACs representing contiguous stretches of genomic DNA (YAC contigs) have provided a physical map framework for the human, mouse, and even Arabidopsis genomes.
- YACs are extremely popular for those trying to analyze entire genomes.

Limitations of Yeast Artificial Chromosomes

- A problem encountered in constructing and using YAC libraries is that they typically contain clones that are chimeric, i.e., contain DNA in a single clone from different locations in the genome.

- YAC clones frequently contain deletions, rearrangements, or noncontiguous pieces of the cloned DNA. As a result, each YAC clone must be carefully analyzed to be sure that no rearrangements of the DNA have occurred.
- The efficiency of cloning is low (about 1000 clones are obtained per microgram of vector and insert DNA).
- YACs have been found to be less stable than BACs.
- The yield of YAC DNA isolated from a yeast clone containing a YAC is quite low.
- The YAC DNA is only a few percents of the total DNA in the recombinant yeast cell. It is difficult to obtain even 1 μg of YAC DNA.
- The cloning of YACs is too complicated to be carried out by a lone researcher.

CHROMOSOMAL WALKING

Key Difference – Chromosome Walking vs Jumping

Chromosome walking and chromosome jumping are two technical tools used in molecular biology for locating genes on the chromosomes and physical mapping of the genomes. Chromosome walking is a technique used to clone a target gene in a genomic library by repeated isolation and cloning of adjacent clones of the genomic library. Chromosomal jumping is a special version of chromosomal walking which overcomes the breakpoints of chromosomal walking. Chromosomal walking can only sequence and map small lengths of chromosomes while chromosomal jumping enables sequencing of large parts of chromosomes. This is the key difference between Chromosomal walking and chromosomal jumping.

What is Chromosome Walking?

Chromosome walking is a tool which explores the unknown sequence regions of chromosomes by using overlapping restriction fragments. In chromosome walking, a part of a known gene is used as a probe and continued with characterizing the full length of the chromosome to be mapped or sequenced. This goes from the marker to the target length. In chromosome walking, the ends of each overlapping fragments are used for hybridization to identify the next sequence.

The probes are prepared from the end pieces of cloned DNA and they are subcloned. Then they are used to find the next overlapping fragment. All these overlapping sequences are used to construct the genetic map of the chromosome and locate the target genes. It is a method of analyzing long stretches of DNA by small overlapping fragments from the recontructed genomic library.

Chromosome Walking Technique – Steps

1. Isolation of a DNA fragment which contains the known gene or marker near target gene

2. Preparation of the restriction map of the selected fragment and subcloning the end region of the fragment to use as a probe

3. Hybridization of the probe with the next overlapping fragment

4. Preparation of the restriction map of the fragment 1 and subcloning of the end region of the fragment 1 to use as a probe for the identification of the next overlapping fragment.

5. Hybridization of the probe with the next overlapping fragment 2

6. Preparation of the restriction map of fragment 2 and subcloning of the end region of the fragment 2 to serve as a probe for the identification of the next overlapping fragment

Above steps should be continued till the target gene or up to 3' end of the total length of the sequence.

Chromosome walking is an important aspect of cytogenetic in finding SNPs of many organisms and analyzing the genetically transmitted diseases and finding mutations of relevant genes.

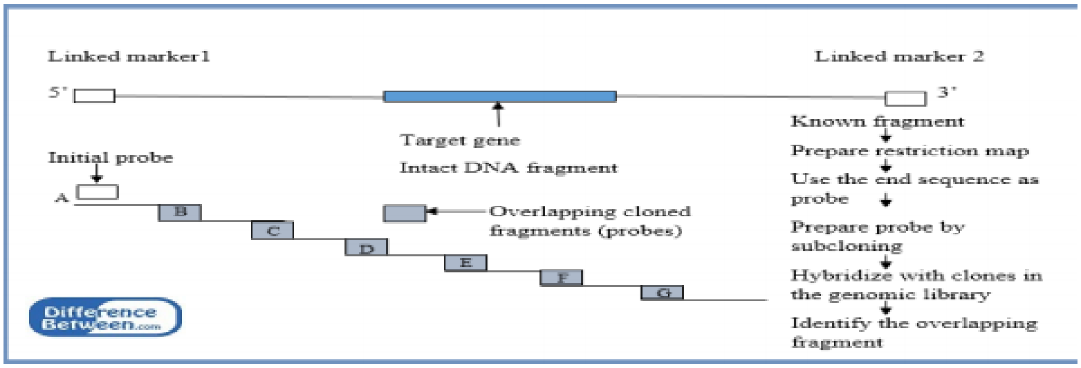

Linked marker1
Linked marker 2
5'
3'
Target gene
Intact DNA fragment
Initial probe
A
B
C
D
E
F
G
Overlapping cloned fragments (probes)
Known fragment
Prepare restriction map
Use the end sequence as probe
Prepare probe by subcloning
Hybridize with clones in the genomic library
Identify the overlapping fragment
DifferenceBetween.com

Figure 01: Chromosome Walking Technique

What is Chromosome Jumping?

- Chromosomal jumping is a technique used in molecular biology for physical mapping of genomes of the organisms.
- This technique was introduced to overcome a barrier of the chromosomal walking which arose upon finding the repetitive DNA regions during the cloning process.
- Therefore, chromosome jumping technique can be considered as a special version of chromosomal walking.
- It is a rapid method compared to chromosomal walking and enables bypassing of the repetitive DNA sequences which are not prone to be cloned during chromosomal walking.
- Chromosomal jumping narrows the gap between the target gene and the available known markers for genome mapping.
- Chromosome jumping tool starts with the cutting of a specific DNA with special restriction endonucleases and ligation of the fragments into circularized loops.

- Then a primer designed from a known sequence is used to sequence the circularized loops. This primer enables jumping and sequencing in an alternative manner.
- Hence, it can bypass the repetitive DNA sequences and rapidly walk through the chromosome for the search of the target gene.
- The discovery of the gene encodes for cystic fibrosis disease was done using the chromosomal jumping tool.
- Combined together, chromosomal jumping and walking can enhance the genome mapping process.

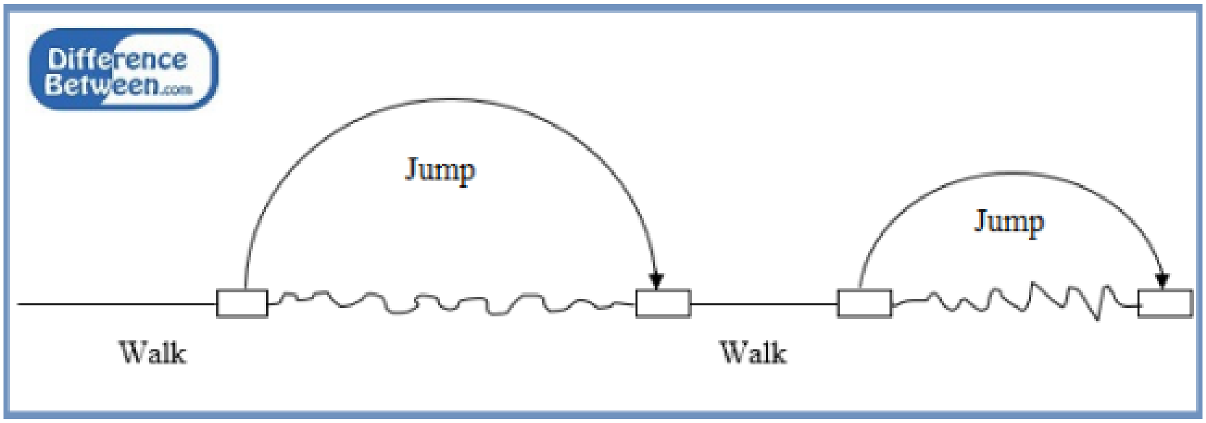


Figure 02: Chromosome Jumping

Summary – Chromosome Walking vs Jumping

Chromosomal walking is frequently applied when it is known that a particular gene is located near a previously cloned gene in a chromosome and it is possible to identify it with repeated isolation of adjacent genomic clones from the genomic library. However, when repetitive DNA regions are found during the chromosomal walking technique, the process cannot be continued. Hence, the technique breaks from that point. Chromosomal jumping is a molecular biological tool which overcomes this limitation for mapping genomes. It bypasses these repetitive DNA regions which are difficult to clone and helps in physical mapping of genomes. This is the main difference between chromosome walking and jumping.

SCREENING OF DNA LIBRARIES USING NUCLEIC ACID PROBES AND ANTISERA

SCREENING AND PRESERVATION OF DNA LIBRARIES

Introduction

Library screening is the process of identification of the clones carrying the gene of interest. Screening relies on a unique property of a clone in a library. The DNA libraries consist of a collection of probably many thousand clones in the form of either plaques or colonies on a plate. Screening of libraries can be done by following approaches based on-

- Detecting a particular DNA sequence and
- Gene expression.

Methods for screening based on detecting a DNA sequence

Screening by hybridization

- Nucleic acid hybridization is the most commonly used method of library screening first developed by Grunstein and Hogness in1975 to detect DNA sequences in transformed colonies using radioactive RNA probes.
- It relies on the fact that a single-stranded DNA molecule, used as a probe can hybridize to its complementary sequence and identify the specific sequences.
- This method is quick, can handle a very large number of clones and used in the identification of cDNA clones which are not full-length (and therefore cannot be expressed). The commonly used methods of hybridization are,

a) Colony hybridization

b) Plaque hybridization.

(a). Colony hybridization

Colony hybridization, also known as replica plating, allows the screening of colonies plated at high density using radioactive DNA probes. This method can be used to screen plasmid or cosmid based libraries

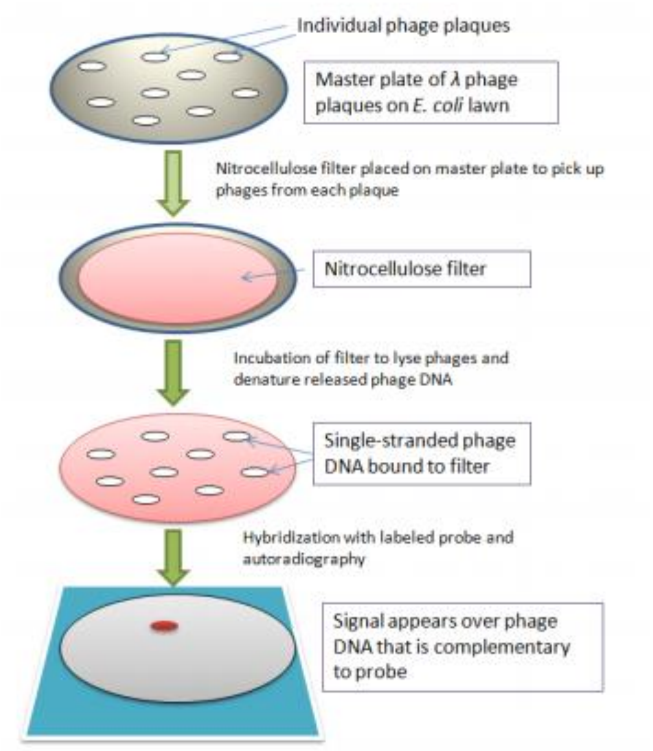
(b). Plaque hybridization

Plaque hybridization, also known as Plaque lift, was developed by Benton and Davis in 1977 and employs a filter lift method applied to phage plaques. This procedure is successfully applied to the isolation of recombinant phage by nucleic acid hybridization and probably is the most widely applied method of library screening. The method of screening library by plaque hybridization is described below-

- The nitrocellulose filter is applied to the upper surface of agar plates, making a direct contact between plaques and filter.
- The plaques contain phage particles, as well as a considerable amount of unpackaged recombinant DNA which bind to the filter.
- The DNA is denatured, fixed to the filter, hybridized with radioactive probes and assayed by autoradiography.

Advantages

- This method results in a ‘cleaner’ background and distinct signal (less background probe hybridization) for λ plaque screening due to less DNA transfer from the bacterial host to the nitrocellulose membrane while lifting plaques rather than bacterial colonies.
- Multiple screens can be performed from the same plate as plaques can be lifted several times.
- Screening can be performed at very high density by screening small plaques. High-density screening has the advantage that a large number of recombinant clones can be screened for the presence of sequences homologous to the probe in a single experiment.


Individual phage plaques
Master plate of λ phage plaques on E. coli lawn
Nitrocellulose filter placed on master plate to pick up phages from each plaque
Nitrocellulose filter
Incubation of filter to lyse phages and denature released phage DNA
Single-stranded phage DNA bound to filter
Hybridization with labeled probe and autoradiography
Signal appears over phage DNA that is complementary to probe

4-5.2.1(b). Schematic process for screening libraries by Plaque hybridization.

Probes used for hybridization

Cloned DNA fragments can be used as probes in hybridization reactions if a cDNA clone is available.

DNA or synthetic oligonucleotide probes can be used for identification of a clone from a genomic library instead of RNA probes, for example, to study the regulatory sequences which are not part of the cDNA clone.

A common method of labeling probes is the incorporation of a radioactive or other marker into the molecule. A number of alternative labeling methods are also available that involve an amplification process to detect the presence of small quantities of bound probe and avoid the use of radioactivity. These methods involve the incorporation of chemical labels such as digoxigenin or biotin into the probe which can be detected with a specific antibody or the ligand streptavidin, respectively.

Screening by PCR

PCR screening is employed for the identification of rare DNA sequences in complex mixtures of molecular clones by increasing the abundance of a particular sequence. It is possible to identify any clone by PCR only if there is available information about its sequence to design suitable primers.

Preparation of a library for screening by PCR can be done by following ways-

- The library can be plated as plaques or colonies on agar plates and individually inoculated into the wells of the multi-well plate. However it is a labor intensive process and can lead to bias in favor of larger colonies or plaques.
- The alternative method involves diluting the library. It involves plating out a small part of the original library (the packaging mix for a phage library, transformation for a plasmid library) and calculating the titer of the library. A larger sample is diluted to give a titer of 100 colonies per mL. Dispensing 100 µL into each well theoretically gives 10 clones in each well. These are then pooled and PCR reactions are carried out with gene-specific primers flanking a unique sequence in the target to identify the wells containing the clone of interest. This method is often used for screening commercially available libraries.

Screening methods based on gene expression

Immunological screening

This involves the use of antibodies that specifically recognize antigenic determinants on the polypeptide. It does not rely upon any particular function of the expressed foreign protein, but requires an antibody specific to the protein.

Earlier immuno-screening methods employed radio-labeled primary antibodies to detect antibody binding to the nitrocellulose sheet (Figure a). It is now superseded by antibody sandwiches resulting in highly amplified signals. The secondary antibody recognizes the constant region of the primary antibody and is, additionally, conjugated to an easily assayable enzyme (e.g. horseradish peroxidase or alkaline phosphatase) which can be assayed using colorimetric change or emission of light using X-ray film (Figure b).

- In this technique, the cells are grown as colonies on master plates and transferred to a solid matrix.

- These colonies are subjected to lysis releasing the proteins which bind to the matrix.

- These proteins are treated with a primary antibody which specifically binds to the protein (acts as antigen), encoded by the target DNA. The unbound antibodies are removed by washing.

- A secondary antibody is added which specifically binds to the primary antibody removing the unbound antibodies by washing.

- The secondary antibody carries an enzyme label (e.g., horse radishperoxidase or alkaline phosphatase) bound to it which converts colorless substrate to colored product. The colonies with positive results (i.e. colored spots) are identified and subcultured from the master plate.

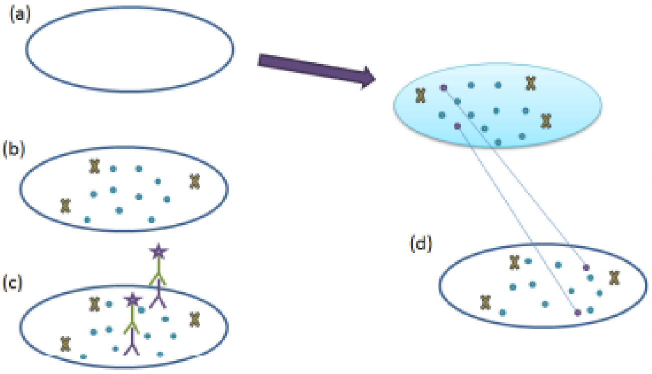


Figure a. Schematic process of immunological screening (a) a nitrocellulose disk is placed onto the surface of an agar plate containing the phage library. Both agar plate and disk are marked so as to realign them later. (b) When the nitrocellulose disk is lifted off again, proteins released from the bacteria by phage lysis bind to the disk. (c) These proteins bind to specific antibody. (d) Plaques formed by bacteriophage that express the protein bound to the antibody will be detected by emission of light. The positive clones can be identified by realignment. (Adapted from Lodge J. 2007.Gene cloning: principles and applications. Taylor & Francis Group)

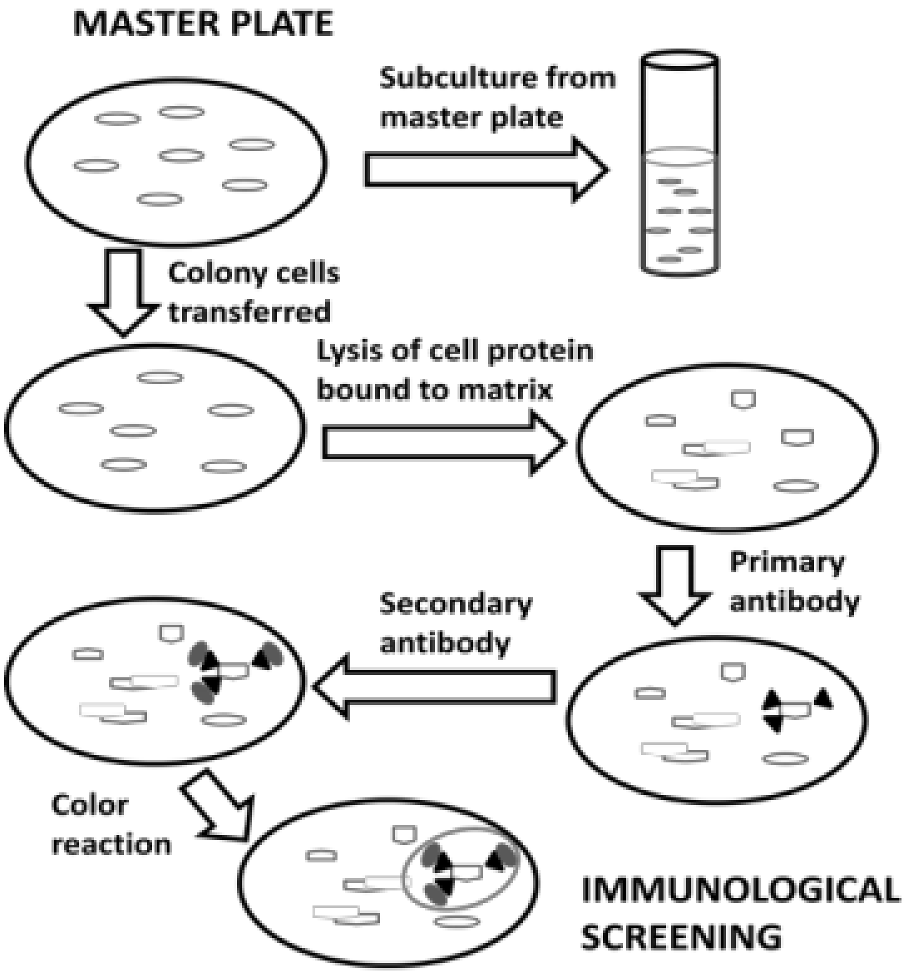


Figure 4-5.3.1(b). Schematic process of immunological screening using antibody sandwich.

The main difficulty with antibody-based screening is to raise a specific antibody for each protein to be detected by injecting a foreign protein or peptide into an animal. This is a lengthy and costly procedure and can only be carried out successfully with proteins produced in reasonably large amounts.

Screening by functional complementation

Functional complementation is the process of compensating a missing function in a mutant cell by a particular DNA sequence for restoring the wild-type phenotype. If the mutant cells are non-viable, the cells carrying the clone of interest can be positively selected and isolated. It is a very powerful method of expression cloning and also useful for identification of genes from an organism having same role as that of defective gene in another organism. The selection and identification of positive clones is based on either the gain of function or a visible change in phenotype. For example, the functional complementation in transgenic mice for the isolation of Shaker-2 gene applied by Probst et al in1988 shown in Figure 4-5.3.2.

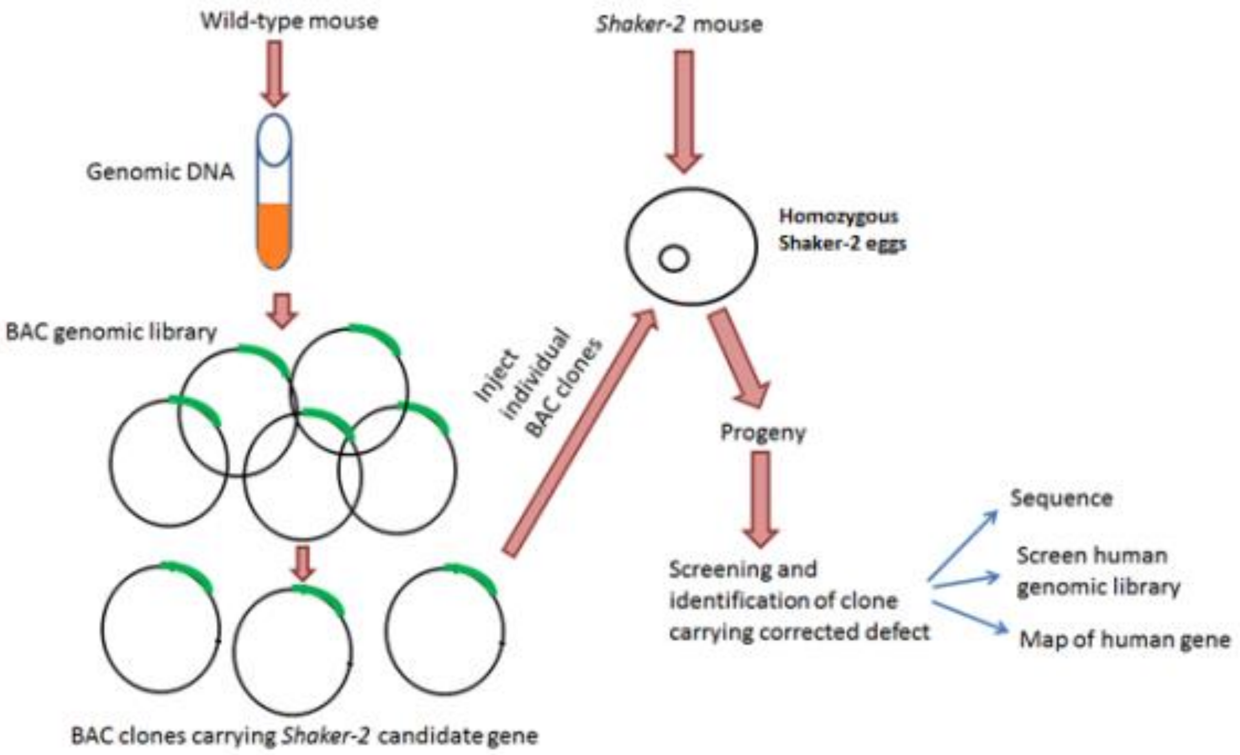

Wild-type mouse
Genomic DNA
BAC genomic library
BAC clones carrying *Shaker-2* candidate gene
Shaker-2 mouse
Homozygous Shaker-2 eggs
Inject individual BAC clones
Progeny
Screening and identification of clone carrying corrected defect
Sequence
Screen human genomic library
Map of human gene

Figure 4-5.3.2. Functional complementation in transgenic mice for isolation of *Shaker-2* gene.

(Adapted from Primrose SB, Twyman RM. 2006. Principles of gene manipulation and genomics. 7th ed. Blackwell Publishing.)

The Shaker-2 mutation is due to the defective gene associated with human deafness disorder. The BAC clone from the wild type mice are prepared and injected into the eggs of Shaker-2 mutants. The resulting mice are then screened for the presence of wild type phenotype. Thus the BAC clone carrying the functional Shaker-2 gene is identified which encodes a cytoskeletal myosin protein. This method can be used for screening human genomic libraries to identify equivalent human gene.

Drawbacks

- Presence of an assayable mutation within the host cell that can be compensated by the foreign gene expression which in most cases is not available. In addition, foreign genes may not fully compensate the mutations.

Applications

- This method can be used for the isolation of higher-eukaryotic genes (e.g. Drosophila topoisomerase II gene, a number of human RNA polymerase II transcription factors) from an organism.

- It can also be possible in transgenic animals and plants to clone a specific gene from its functional homologue.

Assessment:

Brief the following:

1. Construction of genomic DNA libraries.
2. Chromosome jumping.
3. Safety regulations in rDNA techniques.

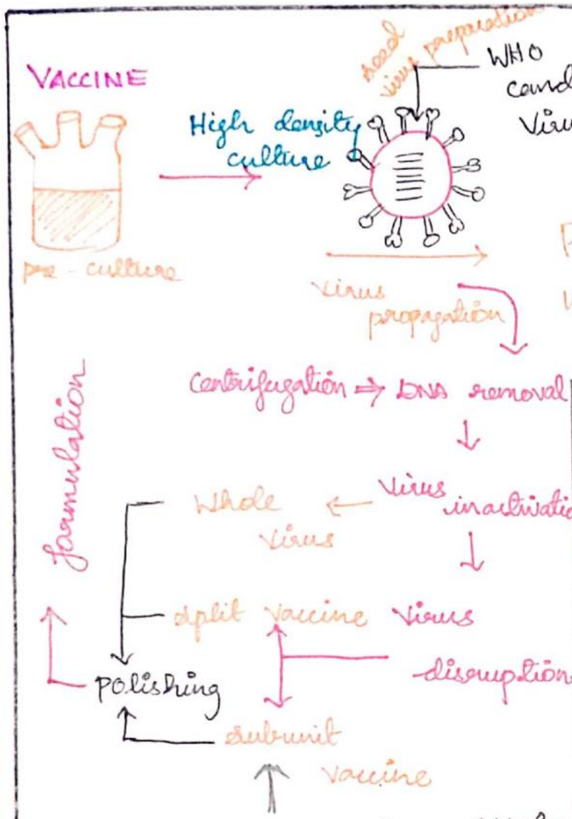
Detail the following:

4. Construction of genomic and complementary DNA libraries.
5. Site directed mutagenesis.

UNIT V

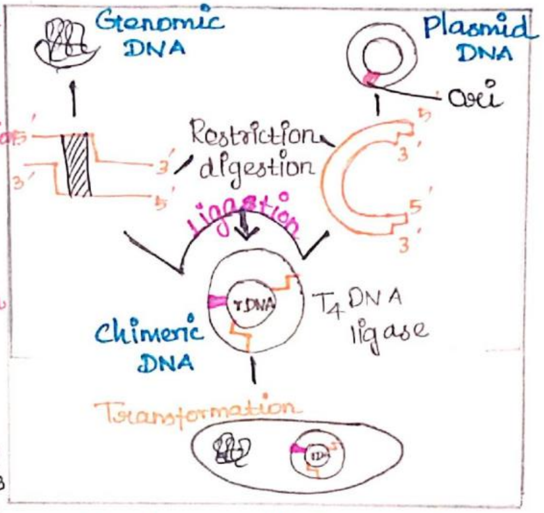
UNIT – V

VACCINE



GEMOS

Genetically modified Organism are Produced using scientific methods that include Recombinant DNA technology.



PRODUCTION OF HEALTH CARE PRODUCT FROM GEMOS.

INPUT → HEALTH CARE

Health care service provider

Health management and support worker.

PRODUCTION PROCESSES

Technology

Organizational structure

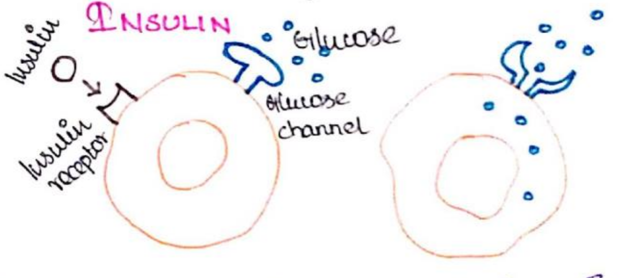
SERVICES

Types → Primary care, birth, delivery, HIV/AIDS treatment.

Quality

↓

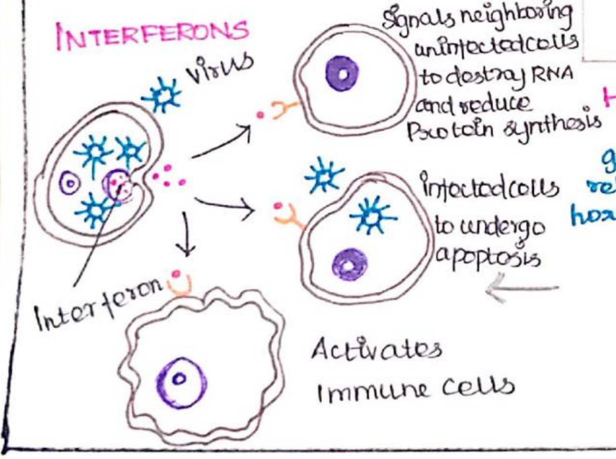
INSULIN



Langerhans Islets → α-helix → β-shirt

INSULIN ← β-galactosidase

INTERFERONS


Virus
Interferon
Signals neighboring uninfected cells to destroy RNA and reduce Protein synthesis
Infected cells to undergo apoptosis
Activates Immune cells

HUMAN GROWTH HORMONE

Hypothalamus

growth hormone releasing hormone (GHRH) + ↓ − growth hormone inhibiting hormone (GHIH)

Anterior pituitary

+ ↓ growth hormone (GH)

liver

+ ↓ Insulin-like growth factor 1

GENETICALLY ENGINEERED MICROORGANISMS

INTRODUCTION TO FUNCTIONAL GENOMICS

Functional genomics is the study of the function of genes contained within an organism's genome, or, put another way, an attempt to figure out what roles genes have in an organism. The earliest examples of functional genomics came in the form of "forward genetic" screens in model organisms such as bacteriophages, bacteria, budding yeast, fruit flies, and roundworms. Their genomes could be mutated at random (e.g., with chemical mutagens) to identify genetic loci (i.e., genes) required for developmental or physiological or molecular processes of interest (e.g., viral particle formation, cell growth, eye formation, reproductive cycle, etc.). These types of studies allowed researchers to infer by introducing a gene's function – an induced mutation in a particular gene causes a specific, measurable change. Importantly, this inference could be made because the screens and follow up experiments were performed in a "wild type" or reference organism. This allowed researchers to introduce only the single genetic change of interest into the reference genome. Since nothing else changes during the experiment, only the DNA mutation of interest could be responsible for the phenotypic change. This became a very powerful tool to assign functions to genes and, in effect, create functional genomic portraits of these model organisms.

However, these same techniques were difficult to carry out in mammals due to multiple experimental barriers, largely arising from the nature of the mammalian genome. First, the genome is diploid – at least two copies of each gene are present. Mutating or altering one gene copy still leaves the other one intact – foiling traditional mutagenesis approaches. Most model organisms allow for a "haploid" phase whereby mutations in single gene copies are "uncovered". Second, classic gene mutagenesis techniques are further hampered by the fact that mammalian genomes contain only <2% protein coding genes (compared to ~70% for yeast), making their use inefficient. Further, for human cells, there is no definable "wild type". This is because individual genomes naturally vary from one another by >10 million bases (usually occurring as single nucleotide polymorphisms). So no two humans or human cell lines are exactly the same (even monozygotic [identical] twins, which should be genetically identical, have small numbers of mutations that arise during development). So if we mutate or inhibit a gene in one human cell will it have the same effect in another isolate? (reviewed in Paddison and Hannon, Cancer Cell 2002)

Lack of expeditious gene manipulation techniques with which to perform functional genomic studies in mammals have hampered basic mammalian biological research and human disease research for decades. Fortunately, two powerful homology-based gene targeting technologies have come along that have helped overcome these barriers, revolutionizing functional genomics in mammals. These are RNAi and CRISPR-Cas9.

RNA interference or RNAi

RNAi emerged out of the pioneering work of Fire, Mello, and colleagues (1998) in the nematode Caenorhabditis elegans. Attempting to use antisense RNA to knock down gene expression, they found synergistic effects on gene silencing when antisense and sense RNA strands where delivered together as double-stranded RNA (dsRNA). While at first RNAi seemed a peculiarity of nematodes, the core machinery that underlies RNAi is conserved in virtually every experimental eukaryotic system and has been co-opted in most of them to trigger gene silencing. At least three core components of the RNAi pathway appear to be generally required for dsRNA-dependent silencing phenomena in higher eukaryotes: the Drosha, Dicer, and Argonaute (Ago) gene family members. Drosha and Dicer proteins sit atop the RNAi pathway in the first catalytic steps that convert various forms of dsRNA into smaller, guide dsRNAs of 21–25 nt. Ago proteins incorporate these small dsRNAs and use their sequence as a guide to identify and target homologous mRNAs for silencing. Some Ago proteins have nuclease activities that can cut or "slice" mRNA targets, triggering their destruction (reviewed in Paddison, 2008).

Uncovering and characterizing many of the components and biochemical determinants of RNAi in invertebrate systems has helped translate RNAi into a genetic tool in mammals via inhibiting mRNA translation. Today, we generally use two types of RNAi triggers to inhibit gene function in mammals: small interfering RNAs (siRNAs) and short hairpin RNAs (shRNAs). SiRNAs are generally chemically synthesized RNA duplexes that contain 21 nt of identity to a homologous mRNA target, 19 nt of dsRNA, and a 2-nt 3′ overhang. ShRNAs are RNA duplexes of 23–29 nt, contain a loop structure that joins both strands of the duplex, which are expressed from DNA-based plasmids or viral vectors. The Paddison Lab routinely performs siRNA and shRNA functional genomic screens in numerous cell types, including human and mouse stem and progenitor cells. As a graduate student Dr. Paddison helped design some of the first RNAi libraries targeting the human genome. (Paddison et al, Nature 2004)

CRISPR-Cas9

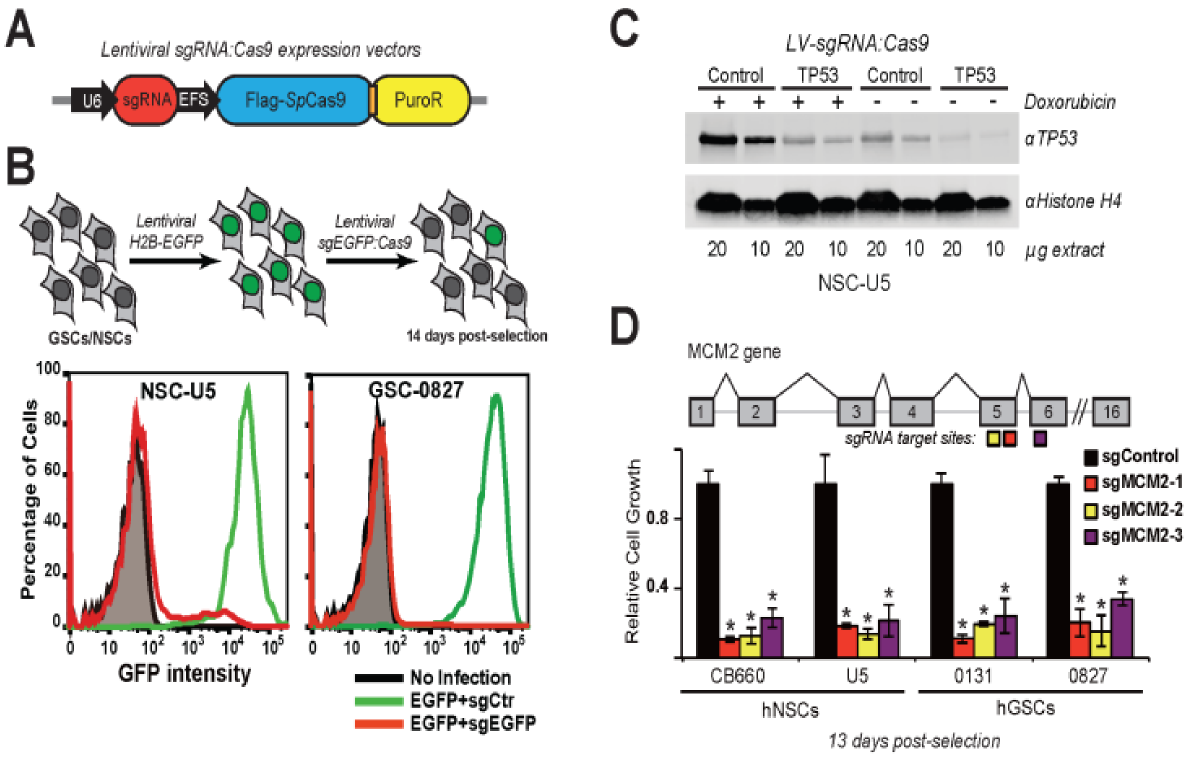
In bacteria, the CRISPR-Cas (Clustered, Regularly Interspaced, Short Palindromic Repeats (CRISPR)–CRISPR-associated (Cas)) pathway acts as an adaptive immune system, conferring resistance to genetic parasites and bacteriophage. Similar to the RNAi pathway in eukaryotes, the CRISPR-Cas system utilizes a single guide RNA (sgRNA) that is incorporated into a protein effector nuclease (e.g., Cas9) to target exogenous genomic sequences. Unlike RNAi, however, CRISPR-Cas systems are able to target and degrade DNA. This property has been harnessed for sgRNA-directed genome editing in prokaryotes and now eukaryotes. In specific, the type II CRISPR-Cas system from Streptococcus pyogenesto has been shown to elicit robust RNA-guided gene editing in multiple eukaryotic systems, including mammalian cells (Gasiunas et al., PNAS 2012; Jinek et al. Science 2012; Wiedenheft et al., Nature 2012; Mali et al., Science 2013; Cho et al., Nature Biotech. 2013).

In its simplest form for gene editing studies, the CRISPR-Cas system consists of two components, an sgRNA and a Cas protein (e.g., Cas9 from S. pyogenes). The sgRNA is a chimeric guide RNA composed of a ~20nt ‘protospacer’ sequence which is used for target recognition and a structural RNA required for complex sgRNA-Cas9 complex formation (i.e., tracrRNA). In addition, a DNA cleavage is "licensed" by an appropriate protospacer adjacent motif (PAM) at the 3' end of the protosequence in the targeted gene. For the type II S. pyogenes system, this sequence is "NGG", where N is any nucleotide. PAMs allow self versus non-self recognition in bacteria and without an appropriate PAM the Cas9-sgRNA complex fails to cut target DNA. The Cas9 gene from S. pyogenes has two catalytic nuclease domains (HNH and RuvC-like) that generate a blunt-ended, double-stranded break 3 bp upstream of the PAM. The HNH domain cleaves the strand of DNA complementary to the sgRNA, while its RuvC-like domain cleaves the non-complementary strand (reviewed by Jinek et al. Science 2012; Wiedenheft et al, Nature 2012; Mali et al., Nature Methods 2013).

The use of the two component CRISPR-Cas system in mammalian cells generally involves the expression of a codon-optimized Cas9 gene with a nuclear localization sequence and expression of a sgRNA from an RNA polymerase III promoter. When expressed together in a mammalian cell, Cas9 promotes gene editing, stimulated by triggering a DNA double-stranded break (DSB), which is repaired by the error-prone non-homologous end joining (NHEJ) or the higher fidelity homology-directed repair (HDR) pathway. While HDR can be error-free, it requires the presence of a homologous repair template, such as a sister chromatid or homology arms from an insertion construct, and appears to only be present in dividing cells, though efficiency can vary widely. By contrast, the NHEJ pathway repairs DSBs throughout the cell cycle in the absence of repair templates through DSB trimming, processing, and re-ligation. NHEJ leaves repair scars in the form of small insertion/deletion (indel) mutations. Thus, in the absence of a repair template the vast majority of Cas9 directed dsDNA cleavage events lead to indel formation at the target site, which, when occurring in an exon, cause frameshifts and premature stop codons in the target gene (reviewed by Jinek et al. Science 2012; Wiedenheft et al, Nature 2012; Mali et al., Nature Methods 2013).

My interest in functional genomics arose from working with model genetic organisms like bacteriophage T4 and budding yeast as an undergrad and also as a technician in Lee Hartwell's lab in the late 1990s. During my time in his lab, Lee Hartwell was fascinated by the notion of synthetic lethality. He thought this could be applied to cancer because cancer is a disease of genomic alterations. Synthetic lethality occurs when a cell or organism can tolerate the loss of gene A or gene B but not loss of gene A and gene B together. This usually means that gene A can compensate for gene B function and vice versa. Because cancer cells are riddled with genetic alterations, it is possible that some of these alterations will result in loss of these types of redundancy – in other words gene A goes missing and now gene B's loss cannot be tolerated. For cancer therapeutic targets, this represents a potentially ideal scenario because normal cells can live without gene B, while cancer cells cannot. This remains a long-term focus of the field of cancer therapeutic and also of my lab.

Validation of CRISPR-Cas9-based gene targeting in human GSCs and NSCs



(A) Cartoon of lentiviral construct used for sgRNA:Cas9 expression.

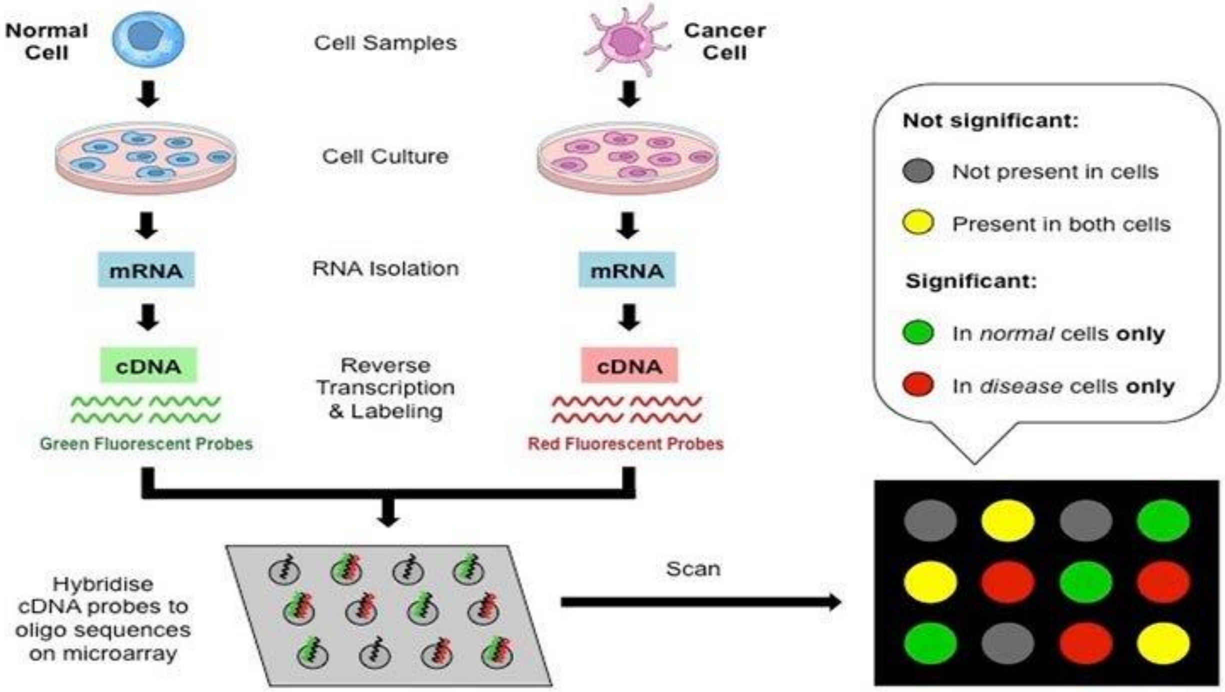
(B) sgEGFP:Cas9 was used to target stably expressed H2B-EGFP in GSCs and NSCs. Cells were first infected with LV-EGFP-H2B, and then infected with sgControl or sgEGFP at MOI<1, selected, outgrown for 14 days, and flow analyzed. At day 5 post-selection, for EGFP+sgEGFP NSC-CB660s, we noted 19.5% of cells still positive for GFP, while by D12, this number was reduced to <1%, suggesting that peak suppression probably occurs around D10 for a single, mono-allelic genomic target.

(C) Western blot confirmation of TP53 protein expression after targeting TP53 gene with sgRNA:Cas9 in NSC-U5s. Cells were outgrown for >21 days following selection. Doxorubicin treatment (0.75µg/ml for 6 hours) was used to stabilize TP53 in response to DNA damage.

(D) CRISPR-Cas9-based targeting of an essential gene, MCM2. Cells were infected with sgRNAs and seeded 3 days post-selection for a 10-day culture in triplicate. Cell viability was then measured using alamarBlue reagent. *p<0.01, student's t-test (unpaired, unequal variance).

MICROARRAYS

- DNA microarrays are solid supports, usually of glass or silicon, upon which DNA is attached in an organized pre-determined grid fashion.
- Each spot of DNA, called a probe, represents a single gene.
- DNA microarrays can analyze the expression of tens of thousands of genes simultaneously.
- There are several synonyms of DNA microarrays such as DNA chips, gene chips, DNA arrays, gene arrays, and biochips.



Principle of DNA Microarray Technique

- The principle of DNA microarrays lies on the hybridization between the nucleic acid strands.
- The property of complementary nucleic acid sequences is to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs.
- For this, samples are labeled using fluorescent dyes.
- At least two samples are hybridized to chip.
- Complementary nucleic acid sequences between the sample and the probe attached on the chip get paired via hydrogen bonds.
- The non-specific bonding sequences while remain unattached and washed out during the washing step of the process.
- Fluorescently labeled target sequences that bind to a probe sequence generate a signal.
- The signal depends on the hybridization conditions (ex: temperature), washing after hybridization etc while the total strength of the signal, depends upon the amount of target sample present.
- Using this technology the presence of one genomic or cDNA sequence in 1,00,000 or more sequences can be screened in a single hybridization.

Types of DNA Microarrays

There are 2 types of DNA Chips/Microarrays:

1. cDNA based microarray

2. Oligonucleotide based microarray

Spotted DNA arrays (“cDNA arrays”)

- Chips are prepared by using cDNA.
- Called cDNA chips or cDNA microarray or probe DNA.
- The cDNAs are amplified by using PCR.
- Then these immobilized on a solid support made up of nylon filtre of glass slide (1 x 3 inches). The probe DNA are loaded into a a spotting spin by capillary action.
- Small volume of this DNA preparation is spotted on solid surface making physical contact between these two.
- DNA is delivered mechanically or in a robotic manner.

Oligonucleotide arrays (Gene Chips)

- In oligonucleotide microarrays, short DNA oligonucleotides are spotted onto the array.
- Small number of 20-25mers/gene.
- The main feature of oligonucleotide microarray is that each gene is normally represented by more than one probe.
- Enabled by photolithography from the computer industry
- Off the shelf

Requirements of DNA Microarray Technique

There are certain requirements for designing a DNA microarray system, viz:

1. DNA Chip
2. Target sample (Fluorescently labelled)
3. Fluorescent dyes
4. Probes
5. Scanner

Steps Involved in cDNA based Microarray

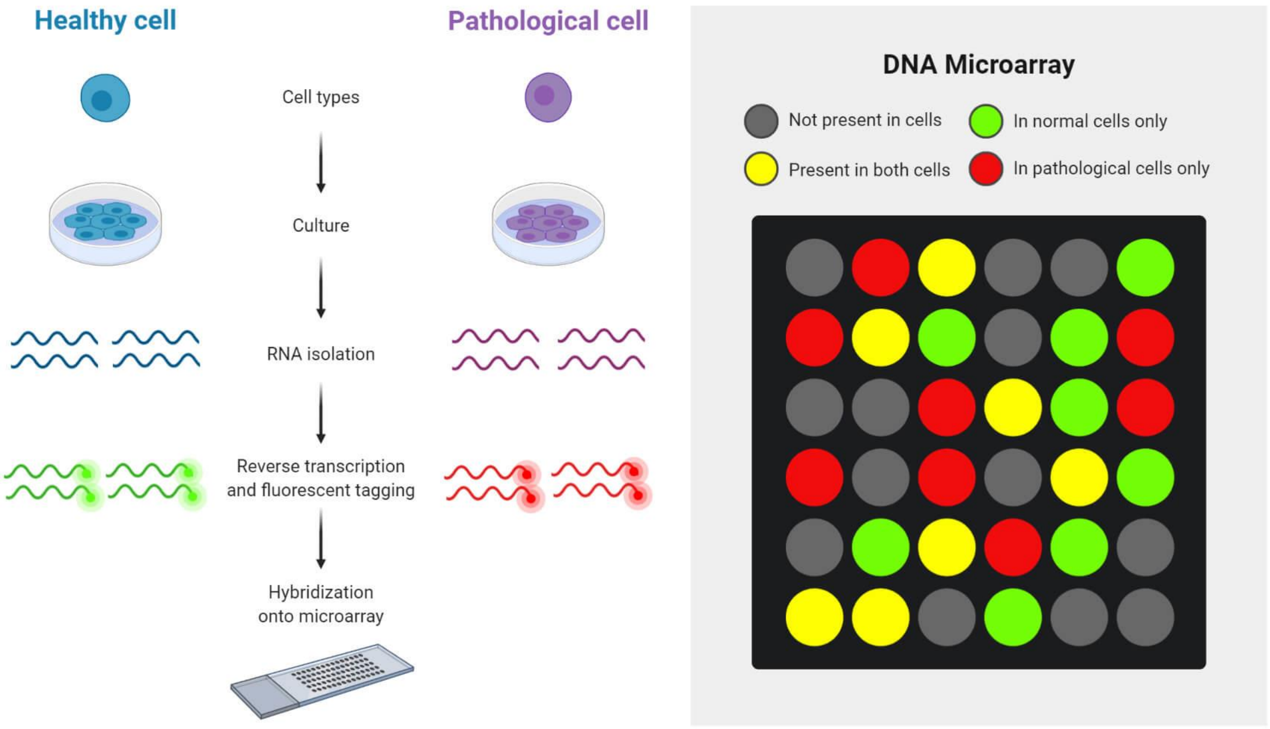


Image By Sagar Aryal, created using biorender.com

The reaction procedure of DNA microarray takes places in several steps:

1. **Collection of samples**

- The sample may be a cell/tissue of the organism that we wish to conduct the study on.
- Two types of samples are collected: healthy cells and infected cells, for comparison and to obtain the results.

2. **Isolation of mRNA**

- RNA is extracted from the sample using a column or solvent like phenol-chloroform.
- From the extracted RNA, mRNA is separated leaving behind rRNA and tRNA.
- As mRNA has a poly-A tail, column beads with poly-T-tails are used to bind mRNA.
- After the extraction, the column is rinsed with buffer to isolate mRNA from the beads.

3. **Creation of labeled cDNA**

- To create cDNA (complementary DNA strand), reverse transcription of the mRNA is done.
- Both the samples are then incorporated with different fluorescent dyes for producing fluorescent cDNA strands. This helps in distinguishing the sample category of the cDNAs.

4. **Hybridization**

- The labeled cDNAs from both the samples are placed in the DNA microarray so that each cDNA gets hybridized to its complementary strand; they are also thoroughly washed to remove unbounded sequences.

5. **Collection and analysis**

- The collection of data is done by using a microarray scanner.
- This scanner consists of a laser, a computer, and a camera. The laser excites fluorescence of the cDNA, generating signals.
- When the laser scans the array, the camera records the images produced.
- Then the computer stores the data and provides the results immediately. The data thus produced are then analyzed.
- The difference in the intensity of the colors for each spot determines the character of the gene in that particular spot.

Applications of DNA Microarray

- In humans, they can be used to determine how particular diseases affect the pattern of gene expression (the expression profile) in various tissues, or the identity (from the expression profile) of the infecting organism. Thus, in clinical medicine alone, DNA microarrays have huge potential for diagnosis.

Besides, it has applications in many fields such as:

- Discovery of drugs
- Diagnostics and genetic engineering
- Alternative splicing detection
- Proteomics
- Functional genomics
- DNA sequencing
- Gene expression profiling
- Toxicological research (Toxicogenomics)

Advantages of DNA Microarray

- Provides data for thousands of genes in real time.
- Single experiment generates many results easily.
- Fast and easy to obtain results.
- Promising for discovering cures to diseases and cancer.
- Different parts of DNA can be used to study gene expression.

Disadvantages of DNA Microarray

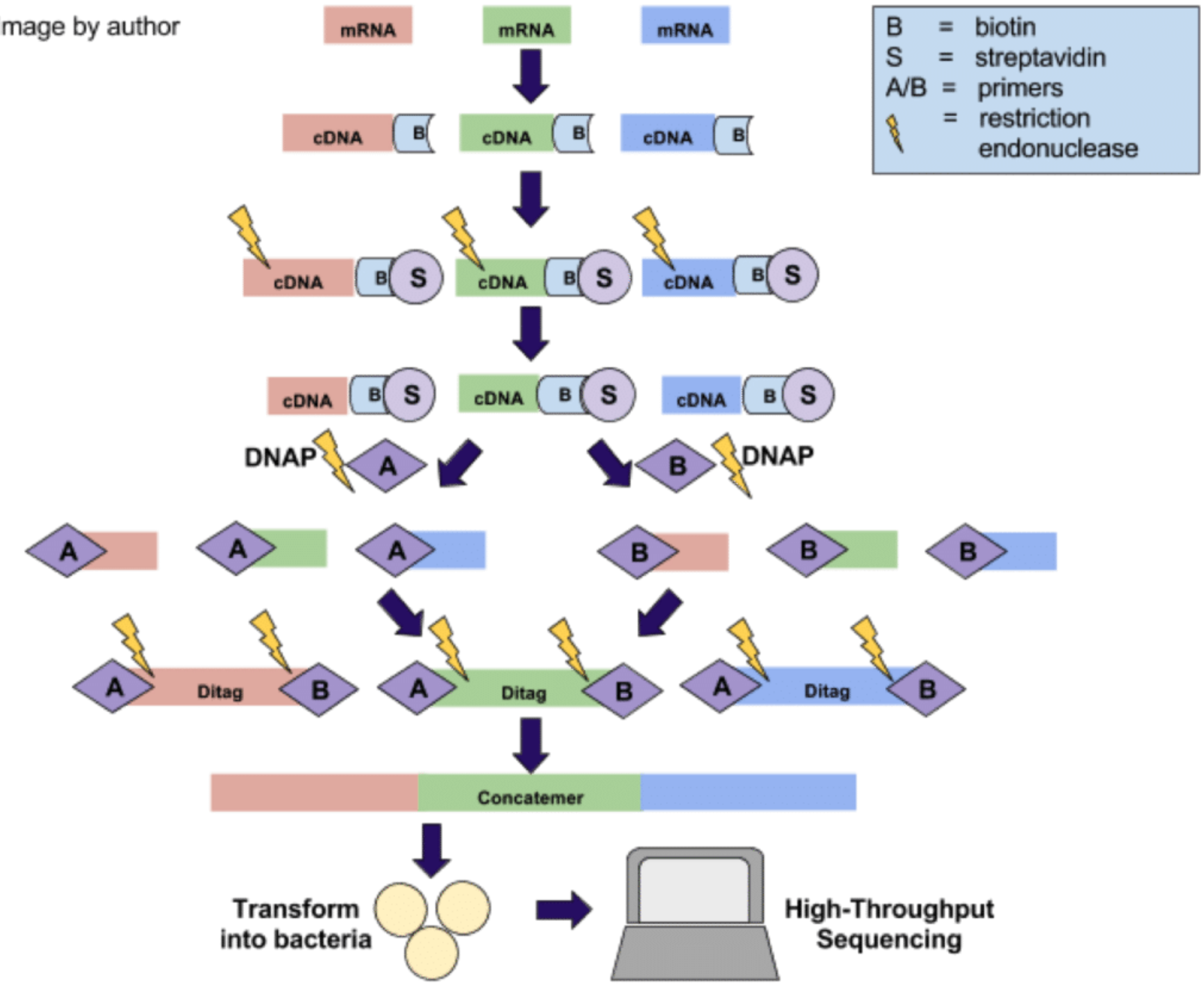
- Expensive to create.
- The production of too many results at a time requires long time for analysis, which is quite complex in nature.
- The DNA chips do not have very long shelf life.

SERIAL ANALYSIS OF GENE EXPRESSION (SAGE)

The SAGE™ technique is one of the more comprehensive methods available for detailed analysis of vast number of cellular transcripts. Initially developed to study differential gene expression in colon cancer samples, SAGE™ analysis compares quantitative expression data for multiple genes in a given specimen to create accurate relative expression profiles.

The ability to count many thousands of genes means that you can detect genes expressed at very low levels in a high-throughput manner. The three-step SAGE™ process allows for simultaneous analysis of sequences derived from various cell populations or tissues.

Image by author



- SAGE was first described and published by Velculescu *et al.* in 1995. At the time, techniques like RNA blotting and expressed sequence tagging were used to study gene expression. However techniques like these were slow and very limited. The speed of SAGE and the ability to study many genes, as small as 10-14bp was a huge step forward in genetics.

SAGE allows you to digitally analyze gene expression patterns. Not just of a few genes but for a cell's complete gene expression profile. SAGE starts with mRNA and end with neat graphs that allow you to compare the gene expression of normal, developmental and diseased cells. I bet you can imagine plenty of times this would be useful. So let's take a look how this technique works.

Working Principle of SAGE

Step 1: Isolate your mRNA and perform reverse transcription using reverse transcriptase and biotinylated primers to generate the corresponding cDNA. Using biotinylation will allow you to isolate your cDNA fragments later on in the process.

Step 2: Mix your cDNA with streptavidin beads. These beads will bind to the biotin-cDNA complex. (You might recognize streptavidin-biotin interaction from western blotting and immunohistological staining techniques – streptavidin-biotin is a very strong bond useful in lots of techniques.)

Step 3: Next cleave the cDNA using a restriction endonuclease enzyme, called an anchoring enzyme. If you remember your Biochemistry 101: restriction enzymes cut at specific points, called a restriction site. So the enzyme you choose will depend on where you would like it to cut. Chef's choice. And since each cDNA fragment is different, each one will be cut at a different place. Where depends on where the restriction enzymes corresponding site is located on the individual fragments. The result of this cleavage is that the beads are bound to cDNA fragments of various lengths with the same sequence at their exposed end.

Step 4: Cleaved cDNA that is no longer bound to the beads is now removed by rinsing. And the remaining bound cDNA is divided into two solutions.

Step 5: Next an oligonucleotide – either A or B – is added to each solution. You can see oligo A and B being added in **Image 1** after your sample is split into two samples. These A and B oligonucleotides have a few notable features: 1) An attachment site or "sticky ends" containing the anchoring enzyme cut site. These attachment sites when digested bind the cleaved cDNA. 2) A recognition site for another type of restriction enzyme called a tagging enzymes. 3) And a short primer sequence that can bind adaptor A or B (this will be used during the PCR step to follow). The adaptors A & B ligate to the cDNA.

Step 6: Now a tagging enzyme is used to cleaved the cDNA. This removes the cDNA from the beads to create a short "tag" of around 11 nucleotides (+4 nucleotides that correspond to the anchoring enzyme recognition site).

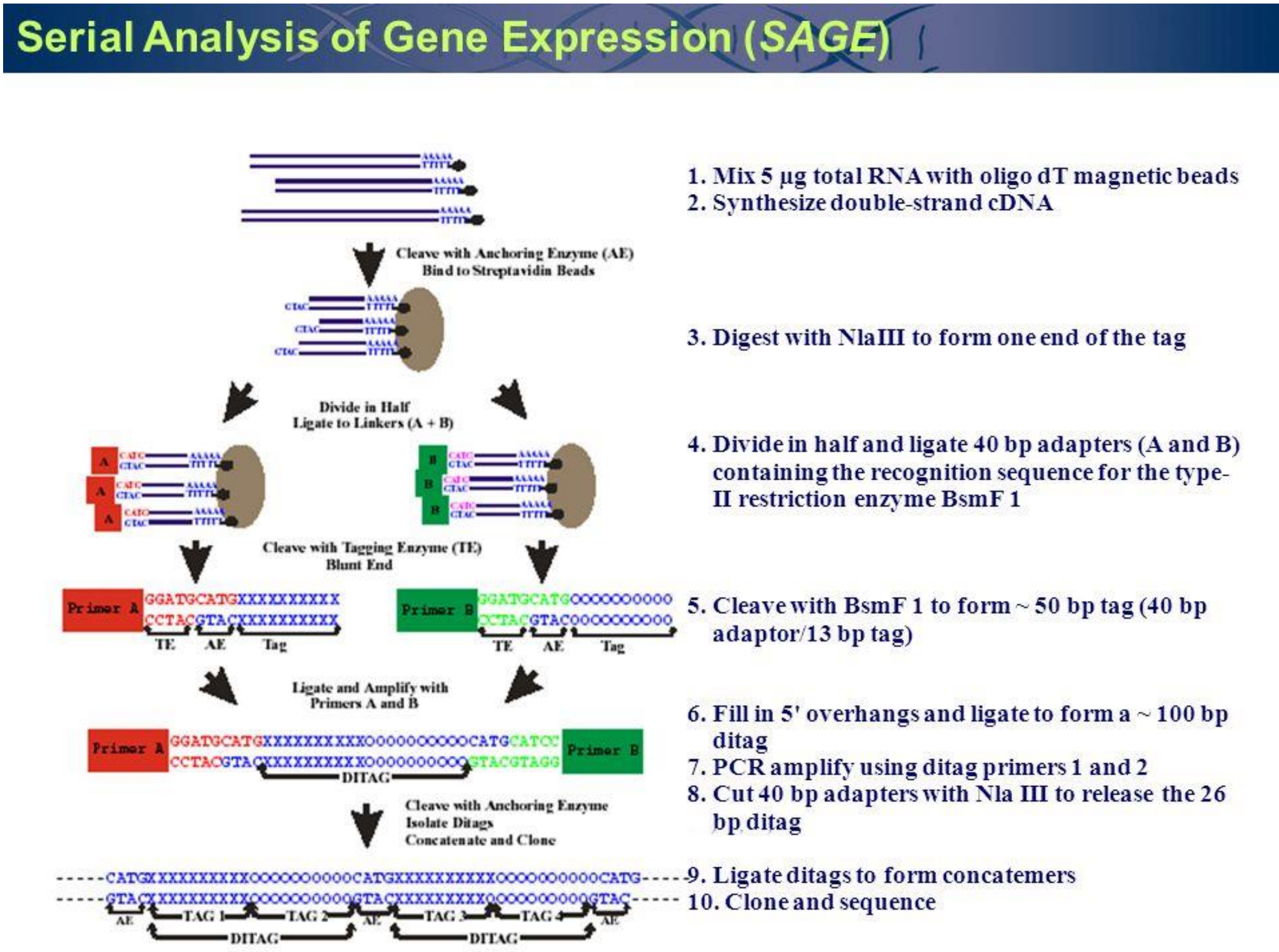
Step 7: These tags have sticky ends but are repaired using DNA polymerase (DNAP). This gives you blunt end fragments that are still bound to the adaptor primer-anchoring enzyme site-tagging enzyme site oligonucleotide.

Step 8: Now it is time to ligate the blunt-end tags together to generate ditags with A and B adaptor ends. This string is then amplified by PCR using A and B primers.

Step 9: The anchoring enzyme is then used to cleave the ditags to remove the A and B oligonucleotides and allow the ditags to form long chains cDNA, called cDNA concatemers, where each ditag is separated by an anchoring enzymes recognition site.

Step 10: Then transform your concatemers into bacteria and allow the bacteria to replicate to form high quantities of your concatemers.

Step 11: The final step (Yay! You made it.) is to isolate your concatemers using your favorite protocol. Then use high-throughput DNA sequencing to quantify each individual tag. And create a gene expression profile for your original sample of mRNA-containing cells.



Serial Analysis of Gene Expression (*SAGE*)

1. Mix 5 µg total RNA with oligo dT magnetic beads
2. Synthesize double-strand cDNA
3. Digest with NlaIII to form one end of the tag
4. Divide in half and ligate 40 bp adapters (A and B) containing the recognition sequence for the type-II restriction enzyme BsmF 1
5. Cleave with BsmF 1 to form ~ 50 bp tag (40 bp adaptor/13 bp tag)
6. Fill in 5' overhangs and ligate to form a ~ 100 bp ditag
7. PCR amplify using ditag primers 1 and 2
8. Cut 40 bp adapters with Nla III to release the 26 bp ditag
9. Ligate ditags to form concatemers
10. Clone and sequence

SUBTRACTIVE HYBRIDIZATION

- **Subtractive hybridization is** a technique for identifying and characterizing differences between two populations of nucleic acids. It detects differences between the RNA in different cells, tissues, organisms, or sexes under normal conditions, or during different growth phases, after various treatments (ie, hormone application, heat shock) or in diseased (or mutant) versus healthy (or wild-type) cells. Subtractive hybridization also detects DNA differences between different genomes or between cell types where deletions or certain types of genomic rearrangements have occurred. Subtractive hybridization techniques have identified many differentially expressed sequences from a wide variety of organisms.

- For example, recent studies using subtractive hybridization have identified mouse complementary DNA (cDNA) associated with retinoic acid-induced growth arrest, bacterial DNA differentiating clinical from nonclinical populations of Staphylococcus aureus, and plant senescence-associated genes from Brassica napus.

- Subtractive hybridization requires two populations of nucleic acids; the tester (or tracer) contains the target nucleic acid (the DNA or RNA differences that one wants to identify), and the driver lacks the target sequences. The two populations are hybridized with a driver to tester ratio of at least 10:1.

- Because of the large excess of driver molecules, tester sequences are more likely to form driver-tester hybrids than double-stranded tester. Only the sequences in common between the tester and the driver hybridize, however, leaving the remaining tester sequences either single-stranded or forming tester-tester pairs.
- The driver-tester, double-stranded driver and any single-stranded driver molecules are subsequently removed (the "subtractive" step), leaving only tester molecules enriched for sequences not found in the driver. Usually multiple rounds of subtractive hybridization are necessary to identify truly tester-specific nucleic acid sequences.

There are five basic steps to subtractive hybridization:

(1) Choosing material for isolating tester and driver nucleic acids
(2) Producing tester and driver;
(3) Hybridizing;
(4) Removing driver-tester hybrids and excess driver (subtraction); and
(5) Isolating of the complete sequence of the remaining target nucleic acid. Variations are possible at each step, and the materials used and methods chosen depend on the desired results.

When choosing appropriate sources for driver and tester, it must be kept in mind that the less complex the source of tester and driver and the more sequences they have in common, the easier it is to isolate specific target sequence differences. For example, it is easier to identify RNA differences between cell types than it is to identify differences between tissues because fewer genes are expressed in single cells.

1. Preparation of Driver and Tester

In principle, both tester and driver samples can be either DNA or RNA, but it is often most practical for the tester to be DNA (because the tester is present in a low concentration, and DNA is more stable than RNA), and for the driver to be RNA (after hybridization, excess driver RNA can be eliminated enzymatically or by alkali degradation). In the basic subtractive hybridization protocol, RNA from the tester source is reverse transcribed into complementary DNA (cDNA) and hybridized to poly A+ driver RNA. The tester-driver hybrids are removed, excess fresh driver is added, and the hybridization is repeated once. The remaining "target" cDNA is either cloned or used to make a probe. This basic procedure is useful if the starting material is not very complex and is easy to isolate.

If little starting tissue is available or if the starting material is complex, multiple rounds of hybridization-subtraction are needed, and it is necessary to use a library- or a PCR-based technique. Tester and driver are prepared from cDNA libraries as phagemids or as library inserts amplified by PCR or in vitro transcription. Alternatively, cDNA from tester and driver sources is ligated to different primers, amplified by PCR, and hybridized. The steps are repeated as needed.

2. Hybridization

When single-stranded nucleic acids are hybridized to each other, more abundant sequences anneal more rapidly because they encounter each other more frequently (see CQt curve). During subtractive hybridization, the hybridization step is driven by the excess driver sequences, so tester sequences that have complementary sequences in the driver population rapidly form driver-tester hybrids, whereas sequences unique to the tester population remain single-stranded or form tester-tester pairs more slowly. Rare sequences from either population take longer to pair up than abundant sequences. The ratio of driver to tester, the overall concentration of driver, the temperature, and the length of hybridization should be chosen based on the complexity of the driver and tester, the abundance class of the target nucleic acids, and the length of the driver and tester sequences used.

2.1. Subtraction

The purpose of the subtraction step is to remove driver-tester hybrids formed during the hybridization step, leaving behind tester enriched for the target sequences. Many different methods are used for subtraction, depending on the nature of the driver and the tester. A few possibilities are mentioned.

Hydroxyapatite chromatography is used to bind double-stranded driver and driver-tester hybrids, leaving single-stranded nucleic acids behind. This is a good choice if the driver is RNA because single-stranded RNA can be removed chemically or enzymatically, leaving only single-stranded cDNA tester after the subtraction.

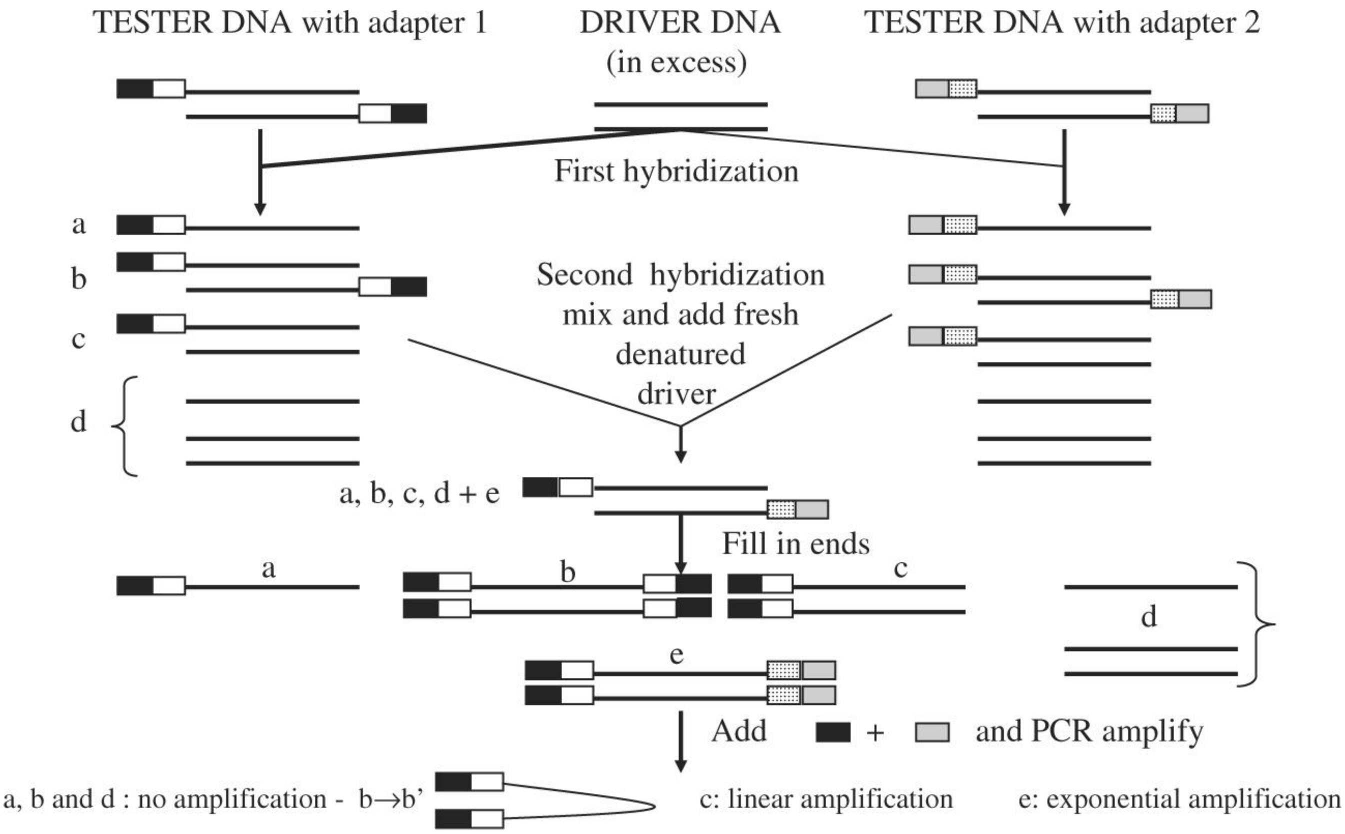
If the tester is a single-stranded phagemid library and the driver is first-strand cDNA, after hybridization the double-stranded driver-tester hybrids can be digested with a frequent-cutting restriction enzyme, and the hybridization mixture used to infect bacteria. Only the single-stranded tester phagemids infect, and they can thus be isolated.

A common procedure is to use biotin-streptavidin binding to separate nucleic acids. Streptavidin binds to biotinylated driver sequences, and phenol extraction is used to remove the streptavidin protein and the bound driver and driver-tester hybrids. Streptavidin can also be attached to beads or to a column and used to remove excess driver and driver-tester hybrids.

The effectiveness of the subtraction is monitored by using radiolabeled tester and determining whether the levels of single-stranded tester decrease after subtraction. Alternatively, enrichment for target sequences is monitored. If there are known genes common to the driver and tester and one or more specific to the tester, it can be determined, after each round of hybridization and subtraction, whether the tester-specific gene is becoming more abundant compared with the common genes.

2.2. Isolation of Target Sequences

After one or more hybridization and subtraction steps, the resulting tester nucleic acids should be greatly enriched for target sequences. However, it is still possible that rare sequences common to both the driver and the tester remain, and in many cases the sequences isolated are only partial gene sequences. The remaining tester sequences are isolated and analyzed in a variety of ways. Tester can be made into an enriched library and probed with driver and tester sequences to look for tester-specific clones, or the tester is labeled and used to probe tester and driver libraries and to isolate full-length clones. It is necessary to further analyze isolated tester sequences by Northern blotting, in situ hybridization or PCR methods to determine whether the sequences are truly tester-specific.



2.3. Alternatives to Standard Subtractive Hybridization Techniques

2.3.1. Positive Selection

An important alternative to subtractive hybridization is positive selection. Hybridization of tester and driver are still carried out but, rather than removing unwanted driver-tester and driver sequences by subtraction during step 4, double-stranded tester sequences are positively selected for selective cloning or selective amplification. Again, various methods are employed to carry out positive selection. A simple method is to digest tester with a restriction enzyme producing cohesive ends, while using sonication to shear the driver DNA randomly. After hybridization, DNA Ligase and vector DNA are added. Only double-stranded tester is cloned into the vector, and then it can be used to transform bacteria.

2.3.2. Representational Difference Analysis (RDA)

RDA is a positive selection technique employing PCR. RDA was originally used to identify differences between complex genomes, such as those caused by chromosomal rearrangements or losses due to cancer, infections with pathogens, and polymorphisms between individuals (4), and it was later adapted to analyze differences in gene expression (5). In both cases, tester and driver are ligated to adapters, amplified by PCR, the original adapters are removed, and new adapters (T2) are ligated only to the tester. After hybridization, only tester-tester DNA is amplified by using primers specific for the T2 adapter. The amplified tester is used again in further rounds of hybridization.

2.3.3. Suppression Subtractive Hybridization (6)

In this positive selection technique, both driver and tester are digested with a frequent-cutting restriction enzyme to give blunt ends. Tester is divided into two samples, which are ligated to different adapters, P1 and P2, and then hybridized to excess driver. Then the two tester populations are mixed, and additional driver is added. Hybrids formed between members of the two subtracted tester populations are selectively amplified by PCR using primers specific to P1 and P2. Molecules that have either P1 or P2 adapters at both ends form "panhandles" as the adapters hybridize to each other. and these molecules are not amplified by PCR (this results in the "suppression").

2.3.4. Differential Display (7)

Differential display is a PCR-based technique that uses random amplification of cDNAs in different populations to identify differences between the populations. For each reaction, one primer is 5′-T^NN, where NN are any two specific nucleotides. This primer binds to a subset of cDNAs containing the two nucleotides complementary to NN immediately adjacent to the poly A tail. The second primer is an arbitrary 10-mer. PCR using these two primers amplifies the same subset of expressed messenger RNA in each sample. Differences in amplification products between different samples are visualized by running the products on a sequencing gel. Band differences on the gel are cut out, and the DNA is eluted and cloned for further analysis. Differential display has the advantage that, for each primer set, a large number of different populations are comparable side by side on one sequencing gel. However, one disadvantage is that a large number of primers and reactions are needed to survey all of the expressed genes in a population.

2.3.5. Serial Analysis of Gene Expression (SAGE) (8)

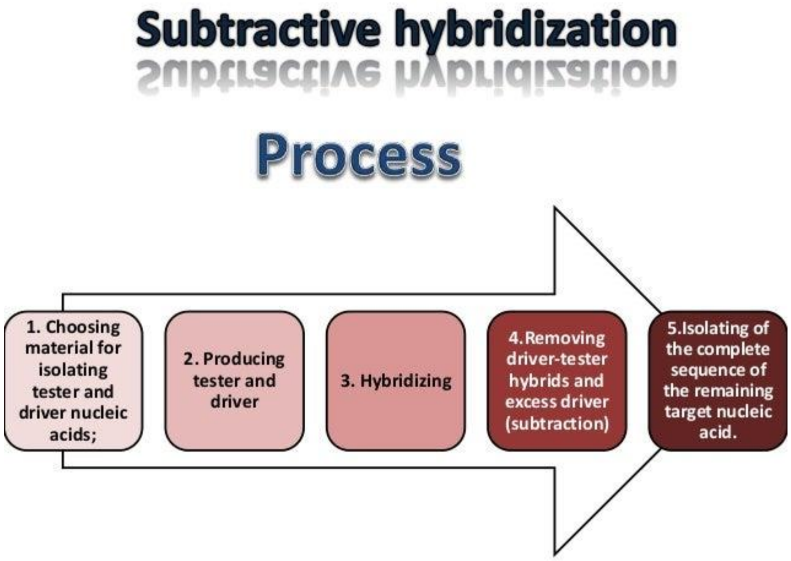
SAGE is based on the fact that nine bp of sequence located at its 3′-end is all the sequence information needed to identify a gene unambiguously. The first step in SAGE involves generating a 9-bp cDNA tag for each of the mRNAs in a population. Then many unrelated tags are concatenated, the concatenated tags are cloned, and random clones are sequenced. These sequences give a spectrum of the genes expressed in the tissue and indicate their relative abundance. Many genes can be analyzed at once, because only nine bp of each gene are sequenced and many tags are sequenced in a single reaction. To be useful for most purposes, however, full-length genes corresponding to the tags must be subsequently identified and isolated.

2.3.6. Microarrays

cDNAs representing either known or unknown genes can be spotted onto glass and probed with different sources of fluorescently-labeled mRNA (9). Alternatively, 20-mer oligonucleotides (oligos) can be synthesized in situ in high-density arrays (10). Oligo sequences are derived from known gene sequences, and a number of oligos are prepared for each gene, so that there are many internal controls. Arrays can be synthesized that contain hundreds of thousands of oligos and can thus simultaneously monitor differences in the expression of tens of thousands of different genes corresponding to the cDNAs. A number of different probes can be analyzed at once because differently colored fluorescent labels are available. As the full repertoire of expressed sequences becomes available for different organisms, this technique will be extremely useful in monitoring genomewide changes in gene expression in different tissues, developmental states, and mutant backgrounds.

3. Conclusion

Many different methods have been used to analyze DNA differences between genomes and differences in gene expression (RNA differences). The use of subtractive hybridization has resulted in the isolation of many useful tissue-specific markers and interesting genes, but it is quite labor-intensive. Various positive selection techniques are more rapid, but they may not result in identifying all differentially expressed genes. Both subtractive hybridization and positive selection only compare two nucleic acid populations at once. The final products are used to make a subtracted probe or to produce a subtracted library, which then are used to identify and isolate full-length target sequences. If full-length cDNA libraries are used, it is not necessary to include this final step. Techniques, such as differential display, SAGE, and the analysis of microarrays, allow comparing a number of different mRNA populations, but differentially expressed genes are not specifically selected. Target genes are identified as a band on a gel (differential display), a very short sequence (SAGE), or a cDNA sequence (microarrays), so that a further step is needed to identify and/or clone full-length sequences. All of the techniques described are used with success, so the choice of technique for detecting DNA and RNA differences depends on the materials available to the investigator and the end results desired.



Subtractive hybridization

Process

1. Choosing material for isolating tester and driver nucleic acids;

2. Producing tester and driver

3. Hybridizing

4.Removing driver-tester hybrids and excess driver (subtraction)

5.Isolating of the complete sequence of the remaining target nucleic acid.

DIFFERENCE IN GEL ELECTROPHORESIS (DIGE)

Any sample from which proteins can be isolated can be analyzed by DIGE. This includes all biological and clinical samples can be analyzed, such as blood, serum, plasma, tissue, cultured cells, culture media, tumor biopsies, and laser micro dissected samples. All animal, plant, nematode, microbial, insect and fish samples can also be analyzed.

What type of buffer is suitable/best for 2D-DIGE experiments?

Basically, any buffer system suitable for the upstream process can be accommodated. However, it may be necessary to include a precipitation and buffer exchange step. If samples are submitted in a different buffer then a buffer exchange step will be required prior to the 2D-DIGE process. This step may lead to unavoidable losses and changes in the protein content.

The 2D-DIGE process is very sensitive and the use of inappropriate buffer systems could affect the labeling efficiency, reproducibility and accuracy of the experiment. The best buffer for 2D-DIGE is the buffer system that is specifically prepared for 2D-DIGE. Based on more than 8 years experience we recommend the ToPI-DIGE (Total Protein Isolation Kit for 2D Difference Gel Electrophoresis) specifically formulated and validated for 2D-DIGE.

What are the steps for the 2D-DIGE process at ITSI-Biosciences?

– Proteins are isolated from samples using ToPI-DIGE kit.
– If client sent samples already partially processed, then samples are precipitated to remove interfering substances (if necessary) and the samples are re-suspended in 2D-DIGE compatible buffer.
– Protein concentration is determined using the ToPA Bradford protein assay.
– If necessary, protein samples are qualified using the Agilent BioAnalyzer.
– 50µg of each protein is labeled with 200pmole Cy3 or Cy5 dye respectively, for the minimal labeling approach. If sample concentration is limiting, 30µg protein can be analyzed using the saturated labeling approach. Cy2 is used to label equal amounts of protein from all samples and used as the universal internal standard/control.
– The labeled samples are loaded on a 24cm IEF strip, pH 3-10. IEF strips with different (narrow and wide) pH ranges can be also used.
– The strips are rehydrated in the presence of the samples for 12 hrs at 30 volts.
– The rehydrated IEF strips are focused for a total of 65,000 volt hours.
– The focused strips are loaded onto a 24cm x 20cm, 12.5% SDS-PAGE gel, and run for 4 hours. The gel strength can be varied to select for low or high molecular weight proteins.
– The gels are scanned on a DIGE-enabled Typhoon Digital Imager at 3 different wavelengths.
– The images are analyzed using DeCyder analysis software (GE Healthcare). Depending on the experimental design, the statistical analysis is performed with the Difference In gel Analysis (DIA) module or the Biological Variation Analysis (BVA) module of DeCyder. The data obtained will be in p-values and ANOVA.
– Spots representing proteins-of-interest picked and tryptically digested with spot handling robots and identified by LC/MS/MS.

What is the maximum number of samples that can be run on a 2D-DIGE gel?

We can run up to three samples on one 2D-DIGE. Two samples will be labeled with Cy3 and Cy5 and the third sample labeled with Cy2 is the universal internal control that allows gel-to-gel comparison.

What is the amount of protein required to run a 2D-DIGE gel?

The minimum amount of an individual total protein sample recommended for a 2D-DIGE experiment is 30µg. However, we recommend sending at least 100µg of each sample. If protein amount is limiting then the “saturated” labeling approach is used. If protein concentration is not limiting then the “minimal” labeling approach is used. We have experience analyzing less than

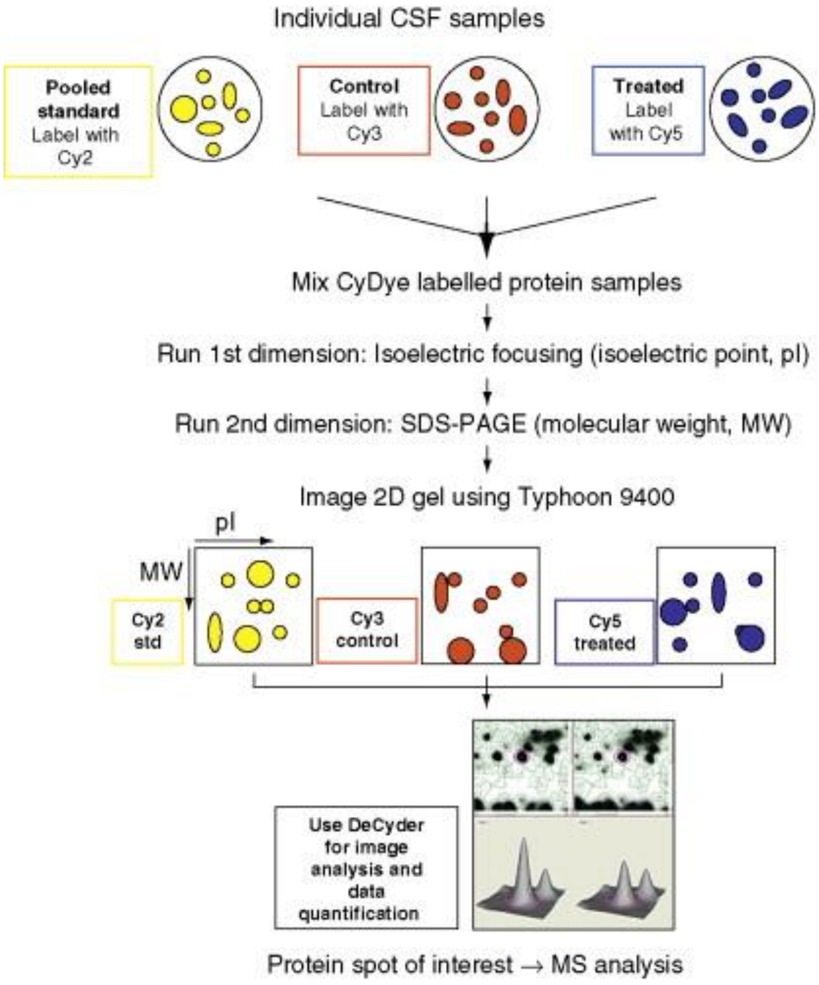
30µg (as low as 5µg has been analyzed) but then only the high abundant proteins will be detected and downstream identification of a candidate spot by mass spectrometry becomes challenging due to the small amount of peptides that will be available for sequencing.

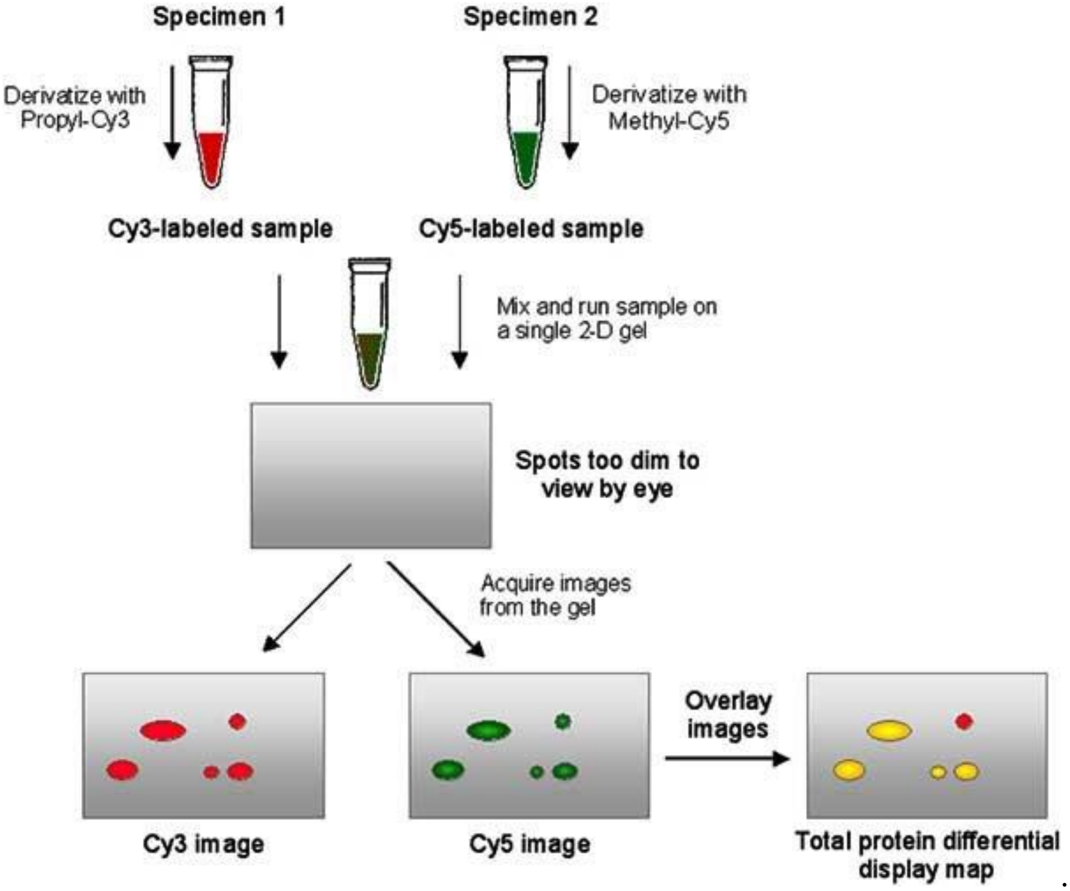
What is the difference between 2D-DIGE and Classical 2D(2D-PAGE)?

In the DIGE (Difference in-gel electrophoresis) process, protein samples are directly labeled with fluorescent dyes (typically Cy2, Cy3 and Cy5) before being separated by 2D electrophoresis. Since samples can be labeled with different fluorescent dyes they can be multiplexed on one gel, reducing gel to gel variations. In a 2D-PAGE only one sample per gel can be analyzed. Also to visualize a 2D- gel the gel must be stained, increasing the background and reducing sensitivity of detection, whereas in a 2D-DIGE gel the labeled proteins can be detected without staining the gel allowing low abundant proteins to be detected. With 2D-DIGE a few number of gels can be used for comparative proteomics.

What is the turnaround time for 2D-DIGE analysis?

The turnaround time for the 2D-DIGE aspect is 3 – 5 days for 10 samples or less. The turnaround time for an entire project that includes spot picking, in gel digestion and identification of the candidate proteins of interest by mass spectrometry is 2 – 3 weeks for most experiments.


Individual CSF samples
Pooled standard Label with Cy2
Control Label with Cy3
Treated Label with Cy5
Mix CyDye labelled protein samples
Run 1st dimension: Isoelectric focusing (isoelectric point, pI)
Run 2nd dimension: SDS-PAGE (molecular weight, MW)
Image 2D gel using Typhoon 9400
pI
MW
Cy2 std
Cy3 control
Cy5 treated
Use DeCyder for image analysis and data quantification
Protein spot of interest → MS analysis


Specimen 1
Specimen 2
Derivatize with Propyl-Cy3
Derivatize with Methyl-Cy5
Cy3-labeled sample
Cy5-labeled sample
Mix and run sample on a single 2-D gel
Spots too dim to view by eye
Acquire images from the gel
Overlay images
Cy3 image
Cy5 image
Total protein differential display map

TOGA VIRUS

Structure of Togaviruses:

The virus is enveloped and forms spherical particle of 65-70 nm diameter. It includes icosahedral capsid, which is made up of 240 monomers (Fig. 17.26). It has a triangulation number of (74). Envelope consists of 80 trimer spikes; dimension of each spike is 3xE1/E2 heterodimers. The spikes consist of glycoprotein that acts as attachment proteins to receptor of host cell membrane.

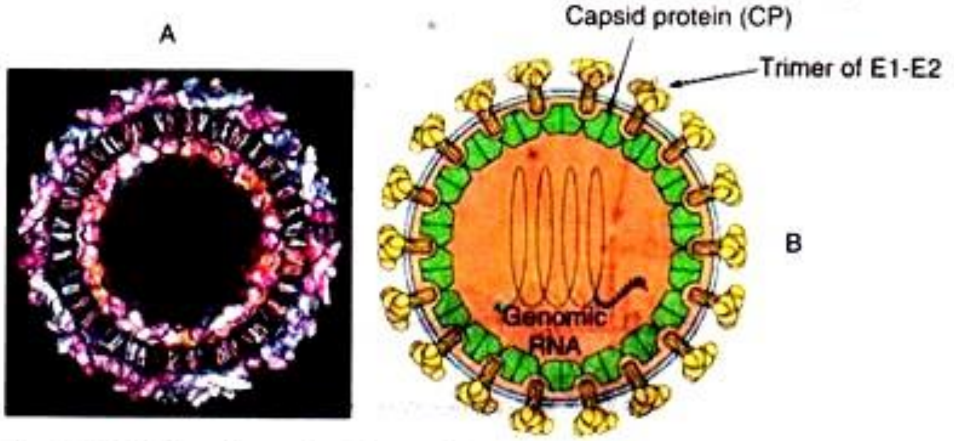


Fig. 17.26: Structure of a virion of Togavirus; A- a computer generated model; B-enveloped virus shows fibres attached to capsid and genomic RNA.

Capsid consists of a mono-partite, single-stranded, (+) sense, non-segmented RNA of about 11.7 kb long (consisting of about 10,000-12,000 nucleotides). It account for 4-8% total weight of particle. The 5′-terminus carries a methylated nucleotide cap (5’cap) and the 3′-terminus has a polyadenylated tail (3′ poly-A) or a genome-linked protein (VPg); therefore, its genome resembles to cellular mRNA, (Fig. 17.27).

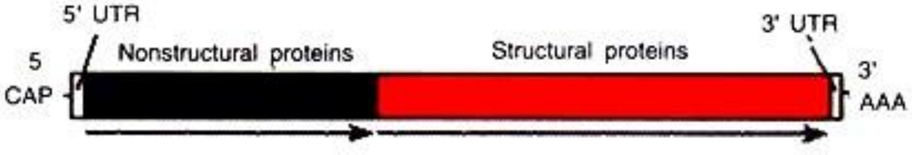


Fig. 17.27: Togavirus having genome of about 11.7 Kb.

Replication of Togaviruses:

Virus attaches to host receptors though E glycoprotein. Thereafter, fusion of virus membrane with the vesicle membrane occurs. Consequently, virus is endocytosed into vesicles in the host cell. RNA genome is released into the cytoplasm. After entry into the cell, gene expression and replication takes place within the cytoplasm (Fig. 17.28).

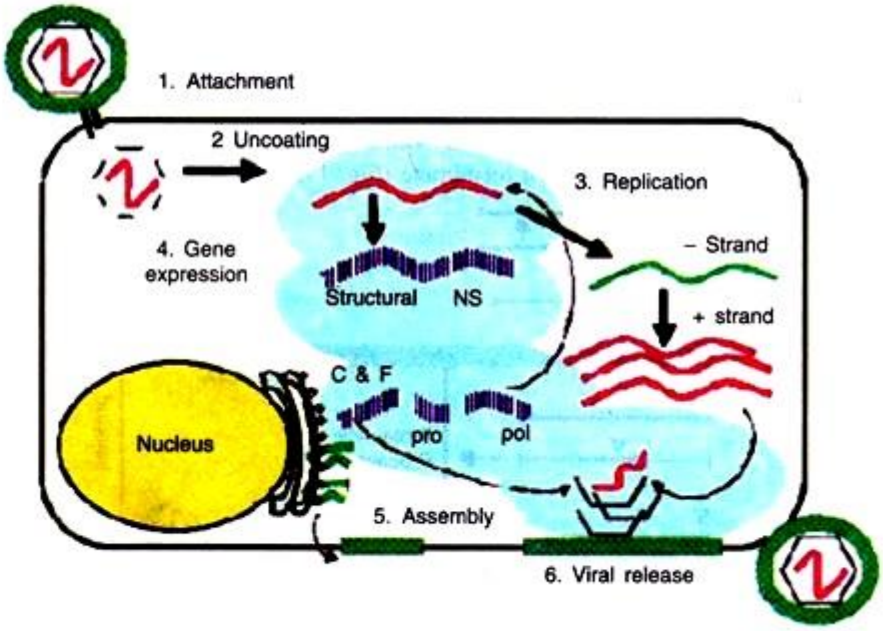


Fig. 17.28 : Multiplication cycle of *Togavirus* inside the host cell (diagrammatic).

The virion RNA is infectious and serves as both genome and viral messenger RNA (Fig. 17.29). The whole genome is translated in a non-structural (NS) polyprotein which is processed by host and viral proteases. Structural polyprotein is expressed through a sub-genomic mRNA (Fig. 17.29).

Characteristically, there are two rounds of translation: (+) sense genomic RNA (’49S’ = 11.7kb) acts directly as mRNA and is partially translated (5′ end) to produce NS proteins. These proteins are responsible for replication, forming a complementary (-) sense strand (cRNA) as the template for further (+) sense strand synthesis.

Consequently, two species of (+) RNA are synthesized, full length genomic RNA and sub-genomic mRNA (’26S’ = 4.1kb). Translation of the newly synthesized sub-genomic RNA results in the production of structural proteins from 3′ end of genome (Fig. 17.29).

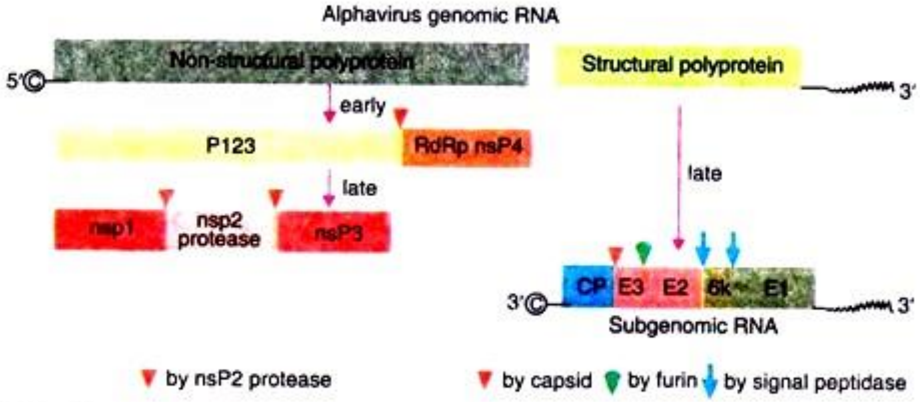


Fig. 17.29 : The genomic RNA *of Alphavirus* which expresses non-structural (NS) and structural proteins; Replication occurs in sub-genomic RNA; C-coat protein; E-envelope protein.

Replication takes place in cytoplasm of the host which occurs at the surface of endoplasmic reticulum. By using the genomic RNA as a template, a negative-sense single-stranded complementary RNA (cRNA) is synthesized. Both new genomic RNA as well as sub-genomic RNA are synthesized by using the negative-sense RNA as a template.

Sub-genomic RNA is translated in structural proteins. Virus assembly occurs at the endoplasmic reticulum. The virion buds at the endoplasmic reticulum is transported to the Golgi apparatus. Assembly occurs al the cell surface, and the envelope is acquired as the virus buds from the cell. Release and maturation almost simultaneous and then bud from the cell membrane (Fig. 17.30).

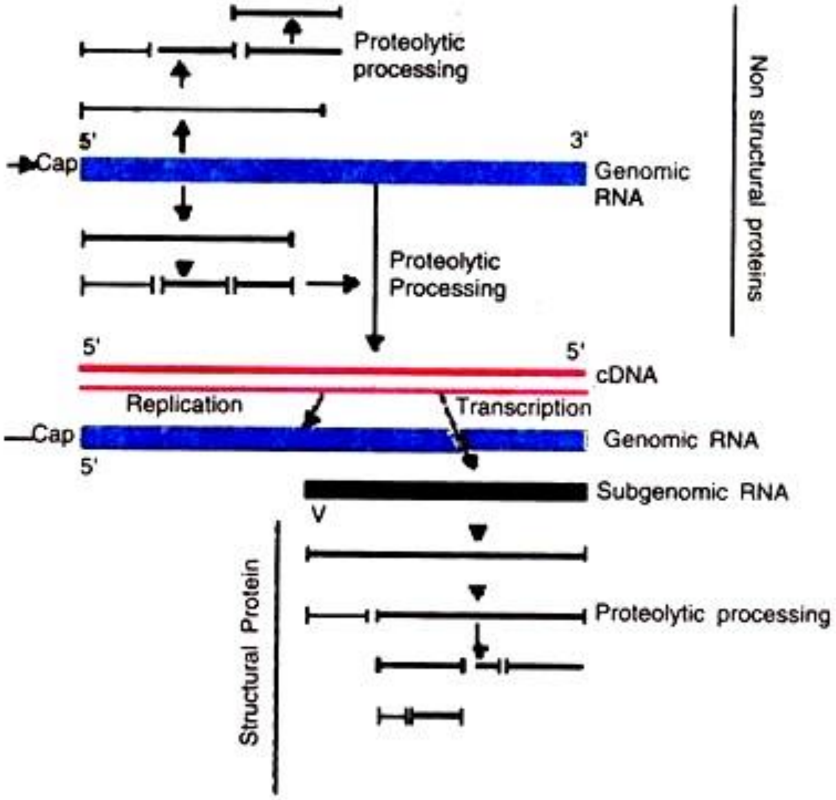
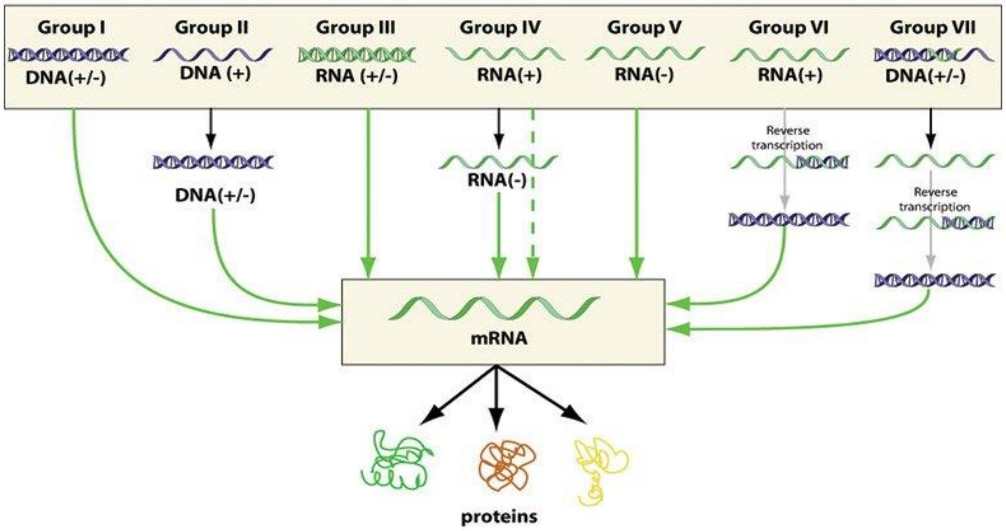


Fig. 17.30 : Protein synthesis (both non-structural and structural), its proteolytic processing and RNA replication (diagrammatic).



YEAST TWO HYBRID SYSTEM

The yeast-2-hybrid system is a simple scientific technique used to screen a library of proteins for potential interactions

- Firstly, a transcription factor is broken into two parts – a DNA-binding domain (BD) and a catalytic activation domain (AD)
- The DNA-binding domain is fused to a protein of interest called the bait (e.g. an enzyme)
- The activation domain is fused to a number of potential binding partners – called the prey (e.g. different ligands)

- If the bait and prey interact, the two parts of the transcription factor are reconstituted and activate transcription of a gene
- If the bait and prey do not interact, the two parts of the transcription factor remain separate and transcription doesn't occur

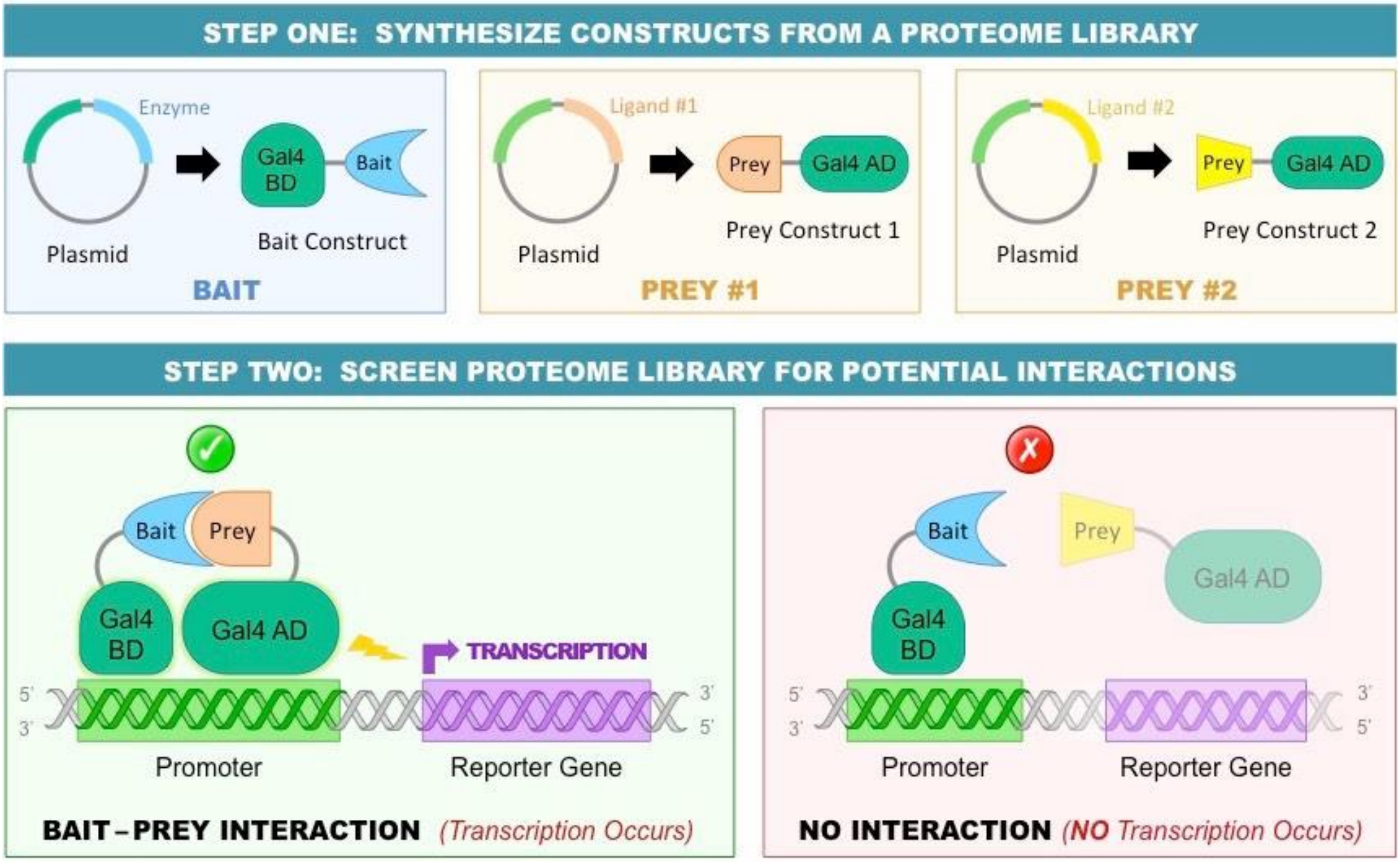
The yeast-2-hybrid system detects protein-protein interactions according to the activation of a reporter gene

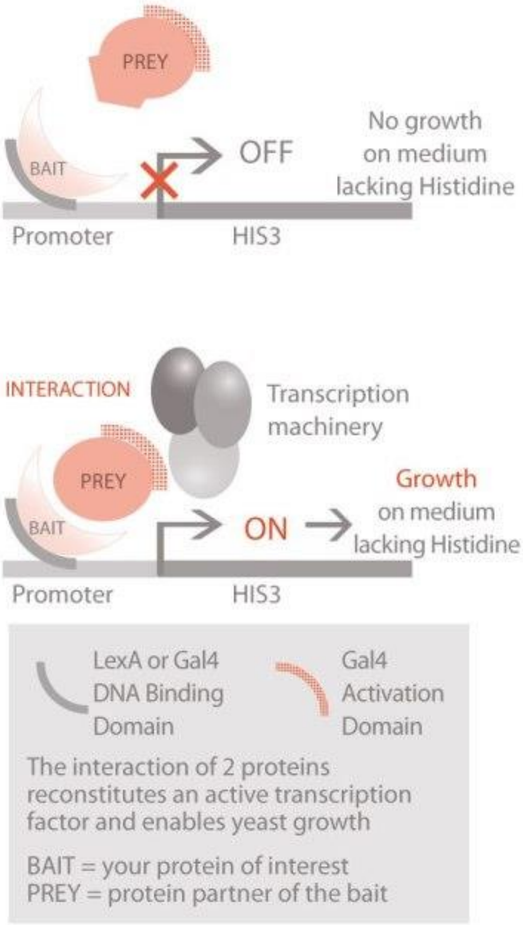
- The reporter gene may encode for the production of a protein that causes a visible colour change (e.g. ß-galactosidase)
- Alternatively, the reporter gene may encode for the production of an essential amino acid that is required for the yeast to grow on a deficient media (hence yeast growth would indicate successful interaction between bait and prey)

Yeast-2-hybrid screens are a simple technique and hence have a relatively high rate of false positives (partial interactions)

- Consequently, the yeast-2-hybrid system is typically only used as an initial test to identify possible protein interactions
- The development and gradual improvements of the **yeast two-hybrid system (Y2H)** since the early 90's revolutionized the way protein interactions could be detected1.
- **Yeast two-hybrid** is based on the reconstitution of a functional transcription factor (TF) when two proteins or polypeptides of interest interact. This takes place in genetically modified yeast strains, in which the transcription of a reporter gene leads to a specific phenotype, usually growth on a selective medium or change in the color of the yeast colonies. The most popular reporter genes are HIS3 to select yeast on a medium lacking histidine, and LacZ to screen yeast in a colorimetric assay.

Overview of the Yeast-2-Hybrid System





The interaction of 2 proteins reconstitutes an active transcription factor and enables yeast growth

BAIT = your protein of interest
PREY = protein partner of the bait

COMPARATIVE GENOMICS

Comparative genomics is a field of biology where the genome of different species is compared to each other to understand evolutionary and molecular differences between species. The development of low cost, next-generation sequencing has enabled the analysis of a plethora of related genomes using comparative genomics.

Genome sequencing and genome comparison

Genetic information is encoded by four nucleosides: adenine, cytosine, guanine, and thymine. Determining the order of these nucleosides in linear DNA forms the basis of sequencing. Along with the human genome, the genomes of several model organisms has now been sequenced - including chimpanzees, mice, fruit flies, puffer fish, roundworms, baker's yeast, and bacteria. In total, the genomes of more than 1000 prokaryotic organisms and 1300 species have been sequenced to date.

The first step in comparative genomics is to compare general features such as: genome size, number of genes and chromosome number. For example, *Arabidopsis* (a plant) has a smaller genome compared to *Drosophila*, the fruit fly which has twice as many genes. Interestingly, the genome size of *Arabidopsis* is similar to humans, suggesting that genome size is not an indicator of complexity or evolutionary status.

DNA sequencing and synteny

Synteny is a method by which genes are arranged in similar blocks across species to identify similar and dissimilar regions. The extent of similarity and dissimilarity may vary across the chromosomes. For example, chromosome 20 of humans corresponds almost completely to the second chromosome of the mouse.

Similarly, the seventeenth chromosome of humans corresponds with chromosome 11 of the mouse. Thus, analysis can show how chromosomal changes that have happened in mouse and human chromosomes since they diverged from a common ancestor almost 75–80 million years ago.

Homologous DNA analysis

Another method used in comparative genomics is homology analysis, where homologous chromosomes of different species are aligned. For example, in one study, the gene for enzyme pyruvate kinase in humans was aligned with the homologous enzyme sequence form dog, mouse, chicken, and zebrafish (among others), and subsequently, the regions of high sequence similarity were plotted.

Such as analysis showed high similarity in the enzyme sequences of human and macaque (a primate), whereas chicken and zebrafish showed similarity only in the coding regions. Such analysis can be used to find which genomic features have been preserved during the course of evolution and, conversely, which features have diversified.

Phylogenetic distance

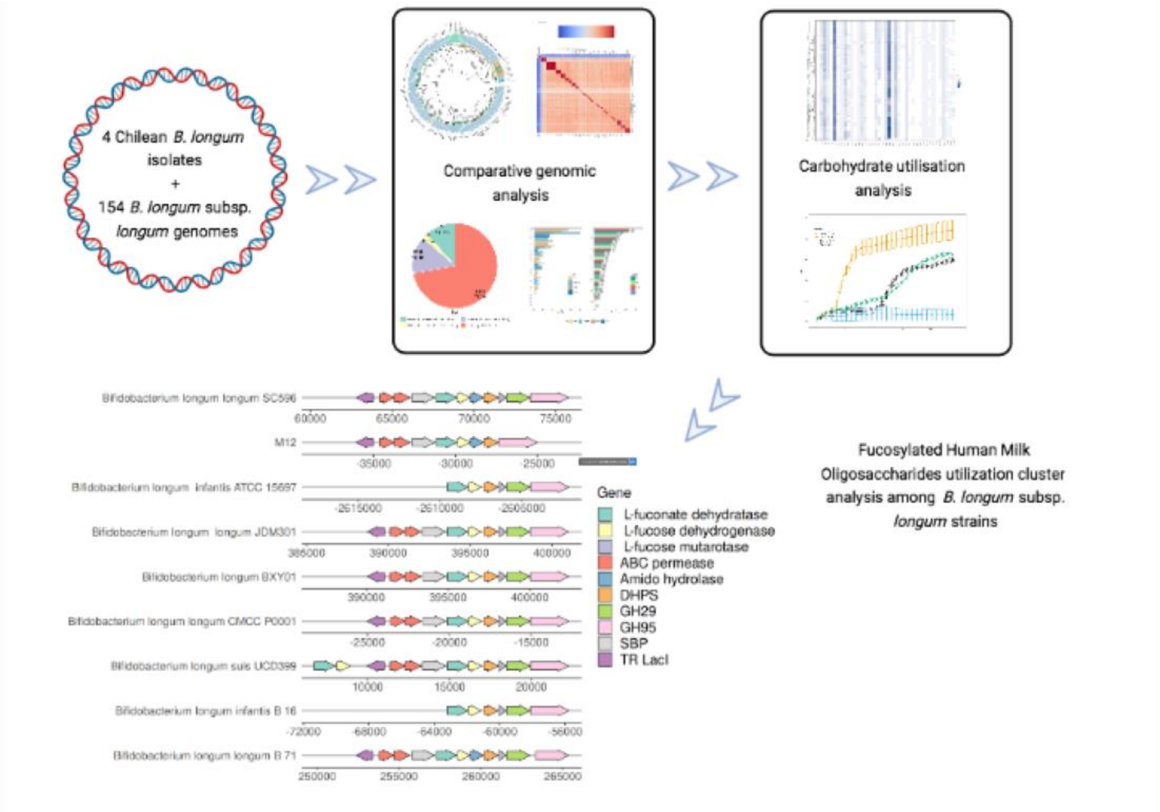
Phylogenetic distance is a non-parametric feature used to measure the degree of separation between two organisms. This parameter is based on the number of sequence changes that have accumulated over a period of years or generations. This distance is inversely proportional to the sequence similarity between the organisms – *i.e.* less the sequence similarity, more is the phylogenetic distance between them.

Over longer phylogenetic distances (such as one billion years) since the organisms separated, only general inferences can be gathered. However, for closer phylogenetic distances, such as 50–200 million years since separation, functional and non-functional DNA may be discriminated, which can subsequently lead to the identification of coding regions, non-coding RNAs, regulatory regions, etc.

For phylogenetic distances less than 5 million years, sequence differences can be used to infer smaller and subtle differences in shape and form. Therefore, comparative genomic differences can provide a lot of powerful information.

Advantages of comparative genomics

Comparative genomes have led to interesting insights, such as that human genome and fruit fly share 60% of their genes. Also, almost two-thirds of the cancer genes have homologous genes in the fruit fly. Such results have a profound impact on human health research. Apart from health, it also has implications in various fields like agriculture, biotechnology, zoology, conservation biology, etc.


4 Chilean *B. longum* isolates
+
154 *B. longum* subsp. *longum* genomes
Comparative genomic analysis
Carbohydrate utilisation analysis
Fucosylated Human Milk Oligosaccharides utilization cluster analysis among *B. longum* subsp. *longum* strains
Bifidobacterium longum longum SC596
60000
65000
70000
75000
M12
-35000
-30000
-25000
Bifidobacterium longum infantis ATCC 15697
-2615000
-2610000
-2605000
Bifidobacterium longum longum JDM301
385000
390000
395000
400000
Bifidobacterium longum BXY01
390000
395000
400000
Bifidobacterium longum longum CMCC P0001
-25000
-20000
-15000
Bifidobacterium longum suis UCD399
10000
15000
20000
Bifidobacterium longum infantis B 16
-72000
-68000
-64000
-60000
-56000
Bifidobacterium longum longum B 71
250000
255000
260000
265000
Gene
L-fuconate dehydratase
L-fucose dehydrogenase
L-fucose mutarotase
ABC permease
Amido hydrolase
DHPS
GH29
GH95
SBP
TR LacI

PROTEOGENOMICS

Proteogenomics is a branch of biology that uses a mix of proteomics, genomics, and transcriptomics to help in peptide discovery and identification. Proteogenomics is a technique for discovering novel peptides by comparing MS/MS spectra to a protein database built from genomic and transcriptome data.

Proteogenomics is a term that refers to research that uses proteomic data, which is frequently acquired from mass spectrometry, to enhance gene annotations. Genomics is the study of complete organisms' genetic code, whereas transcriptomics is the study of RNA sequencing and transcripts.

To discover and analyse the activities of proteins, proteomics uses tandem mass spectrometry and liquid chromatography. Proteomics is a technique for identifying all of the proteins expressed by an organism, which is referred to as its proteome. The problem with proteomics is that it assumes that current gene models are right and that proper protein sequences can be obtained using a reference protein sequence database; however, this is not always the case, as certain peptides are not detected in the database. Mutations can also result in the emergence of new protein sequences.

Proteomic, genomic, and transcriptomic data can be used to address these difficulties. Proteogenomics, which combines proteomics and genomics, was established as a separate science in 2004.

1. Single-cell proteogenomics is a term that has recently been used to describe the profiling of surface proteins and mRNA transcripts from single cells using techniques such as CITE-Seq, despite the fact that the aim of these investigations are not linked to peptide discovery.

Since the year 2019, these techniques have been referred to as multimodal omics or multi-comics.

Methodology : The fundamental idea behind the proteogenomic technique is to find peptides by comparing MS/MS data to protein databases with predicted protein sequences. The protein database is created in a number of methods using genomic and transcriptomic data.

Applications:

Proteogenomics can be used in a variety of ways. Gene annotation enhancement in multiple species is one application. Gene annotation entails the identification of genes and their activities.

In prokaryotic species, proteogenomics has proven particularly effective in the identification and refinement of gene annotations. Various microorganisms, such as *Escherichia* coli, *Mycobacterium*, and numerous species of *Shewanella* bacteria, have had their genomic annotation investigated using the proteogenomic technique.

Proteogenomic investigations can reveal the existence of planned frameshifts, N-terminal methionine excision, signal peptides, proteolysis, and other posttranslational changes, in addition to enhancing gene annotations. Proteogenomics has medical benefits, particularly in oncology research.

Methylation, translocation, and somatic mutations are all examples of genetic alterations that cause cancer.

According to research, understanding the molecular changes that contribute to cancer require both genomic and proteomic knowledge. This has been assisted by proteogenomics, which has identified protein sequences that may have functional functions in cancer.

A recent example of this was a research involving colon cancer that led to the discovery of prospective cancer therapeutic targets. Proteogenomics has also led to customised cancer targeting immunotherapies, in which antibody epitopes for cancer antigens are predicted using proteogenomics, allowing for the development of drugs that target the patient's individual tumour.

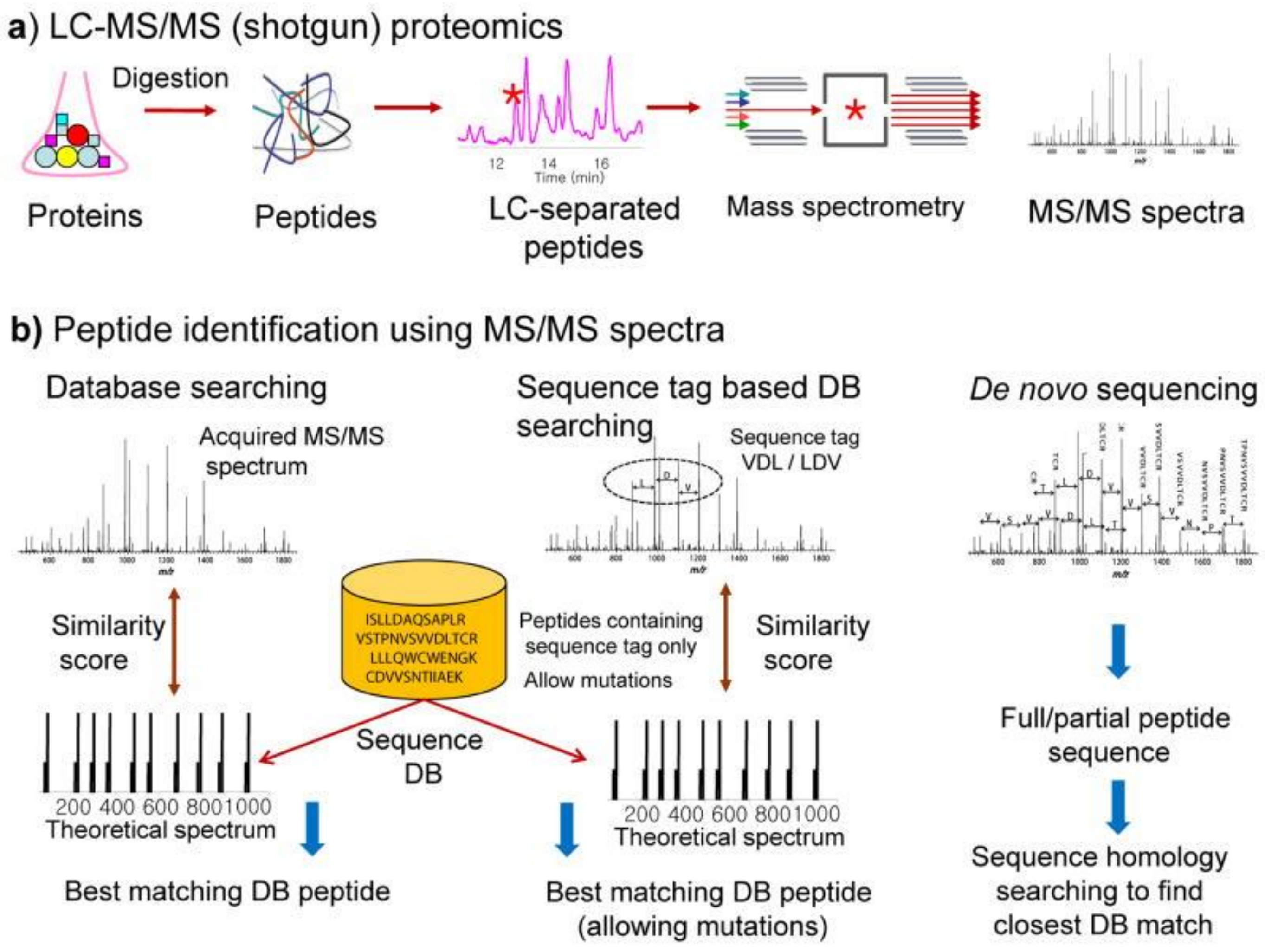
Proteogenomics may give insight into cancer diagnosis in addition to treatment. Proteogenomics was used to discover somatic mutations in colon and rectal cancer investigations.

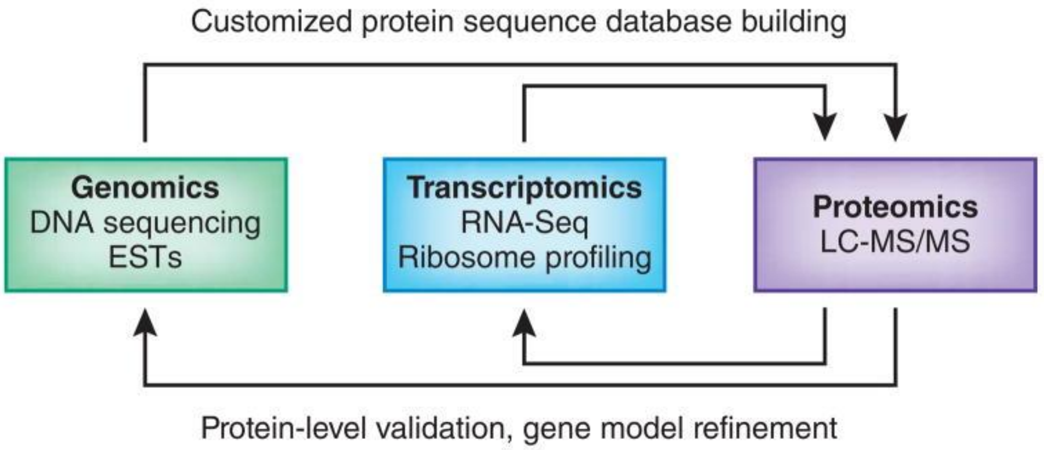
The detection of somatic mutations in patients might aid in the diagnosis of cancer.

Proteogenomics may provide peptide identification techniques without the drawbacks of proteomics, such as inadequate or erroneous protein databases; nonetheless, the proteogenomic approach comes with its own set of obstacles.

The sheer amount of protein databases created is one of the most difficult aspects of proteogenomics.

According to statistics, a big protein database is more likely to result in inaccurate protein database data matching to MS/MS data, which can obstruct the discovery of novel peptides.




Customized protein sequence database building
Genomics
DNA sequencing
ESTs
Transcriptomics
RNA-Seq
Ribosome profiling
Proteomics
LC-MS/MS
Protein-level validation, gene model refinement

The concept of proteogenomics

In a proteogenomics approach, genomics (DNA sequencing, expressed sequence tags (ESTs) and transcriptomics (RNA-Seq, ribosome profiling) data is used to generate customized protein sequence databases to help interpret proteomics (LC-MS/MS) data. In turn, the proteomics data provides protein-level validation of the gene expression data, as well as helping to refine gene models. The enhanced gene models can help improve protein sequence databases for traditional proteomics analysis.

WEB RESOURCES FOR GENOMICS

GENOMICS

Many of the tools that one needs for the analysis of genomes can be found in the DNA Sequence Analysis section. Here we have unique tools for genomic analysis which do not fit easily in that section.

1. DNA sequencing
2. Sequencing errors
3. Genome annotation
4. Correcting genome annotations
5. Specialized annotation - general (inteins, plasmids, typing, vaccine candidates)
6. Two-component and other regulatory proteins
7. Orthologous genes/proteins
8. Specialized annotation - antibiotic resistance
9. Specialized annotation - CRISPR
10. Specialized annotation - virulence determinants
11. Specialized annotation - Genomic Islands
12. Genome comparisons and synteny
13. Phylogeny (AAI and ANI)
14. Genome visualization
15. Synthetic genes
16. Metagenomics
17. Meta sites
18. Naming your bacteriophage

DNA sequencing:

- DNA Sequence Quality - Phred - provides base calling, chromatogram display and high quality sequence region evaluation and presentation for up to five sequences simultaneously.

- Sequence assembly - you don't need your own contig assembly program when you can use:

- EGassember - aligns and merges sequence fragments resulting from shotgun sequencing or gene transcripts (EST) fragments in order to reconstruct the original segment or gene (Reference: A. Masoudi-Nejad et al. 2006. Nucl. Acids Res. 34: W459-462).

- CGE Assembler 1.2 - assembles Illumina, 454, SOLid and Ion Torrent data (Reference: Larsen MV, et al. J. Clin. Micobiol. 2012. 50(4): 1355-1361).
- CGE SPAdes 3.9 - assembles Illumina and Ion Torrent data (Reference: S. Nurk et al. Research in Computational Molecular Biology: pp 158-170).

- CAP3 (*PBIL, France*), (Reference: Huang,X. & Madan A. 1999. Genome Res. 9: 868-877), and here.
- CAP EST Assembler *(Istituto FIRC di Oncologia Molecolare, Italy)* - Maximum sequence length for each sequence is 30 kb - Maximum number of sequences 10 kb

- MicroScope web site (hosted at Genoscope), provides an environment for expert annotation and comparative genomics. Genome project: Annotation and comparative analyses of finished or draft genome sequences. For pre-annotated sequences, they only integrate annotations from NCBI RefSeq complete genome section. Metagenome project: Annotation and comparative analyses of assembled metagenomic sequences. Currently, they are able to integrate datasets below 20 Mb of contigs per bin.

- NanoPipe - was developed in consideration of the specifics of the MinION sequencing technologies, providing accordingly adjusted alignment parameters. The range of the target species/sequences for the alignment is not limited, and the descriptive usage page of NanoPipe helps a user to succeed with NanoPipe analysis. The results contain alignment statistics, consensus sequence, polymorphisms data, and visualization of the alignment. (Reference: Shabardina V et al. (2019) Gigascience 8(2). pii: giy169).

- COV2HTML: a visualization and analysis tool of bacterial next generation sequencing (NGS) data for postgenomics life scientists - allows performing both coverage visualization and analysis of NGS alignments performed on prokaryotic organisms (bacteria and phages).

It combines two processes: a tool that converts the huge NGS mapping or coverage files into light specific coverage files containing information on genetic elements; and a visualization interface allowing a real-time analysis of data with optional integration of statistical results. (Reference: Monot M. et al. 2014. OMICS 18(3): 184-95).

- DCA Divide-and-Conquer Multiple Sequence Alignment *(Universitat Bielefeld, Germany)* - is a program for producing fast, high quality simultaneous multiple sequence alignments of amino acid, RNA, or DNA sequences. (Reference: Brinkmann, G. et al. Mathematical Programming 79: 71-97, 1997).

- PhageTerm - is a fast and user-friendly software package which can be used to determine bacteriophage termini and packaging mode from randomly fragmented NGS data. It is part of the Galaxy package, and can be found in the "NGS: Mapping" directory. Ideal is you want an automated answer. (Reference: Garneau JR, et al. 2017. Sci Rep. 7(1):8292).

- QUAST - a quality assessment tool for evaluating and comparing genome assemblies. This tool improves on leading assembly comparison software with new ideas and quality metrics. QUAST can evaluate assemblies both with a reference genome, as well as without a reference. QUAST produces many reports, summary tables and plots to help scientists in their research and in their publications. (Reference: A. Gurevich et al. 2013. Bioinformatics, 29(8): 1072–1075). N.B. This server is as of April 2020, but there are hopes that it will be back online (see here for software downloads).

- Sequencing errors: - if your DNA sequence doesn't match the expected protein sequence you can check for errors at GeneWise (EMBL-EBI) which compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors. Other programs include:

- FrameD (Reference: T. Schliex et al. 2003. Nucl. Acids Res. 31: 3738-3741)
- AMIGene - annotation of microbial genes (Reference: Bocs S et al. (2003) *Nucleic Acids Res.* 13(31): 3723-3726).
- path :: protein back-translation and alignment - addresses the problem of finding distant protein homologies where the divergence is the result of frameshift mutations and substitutions. Given two input protein sequences, the method implicitly aligns all the possible pairs of DNA sequences that encode them, by manipulating memory-efficient graph representations of the complete set of putative DNA sequences for each protein. (Reference: Gîrdea M et al. 2010. Algorithms for Molecular Biology 5:)

- In-silico.com *(Dr. Joseba Bikandi & co-workers, Faculty of Pharmacy, in the University of the Basque Country)* - allows *in silico* experiments including theoretical PCR amplification, AFLP-PCR , restriction analysis and pulsed field gel electrophoresis [PFGE] with bacterial & archael genomes found in the public database.

- **Genome annotation:**

- NCBI Prokaryotic Genomes Automatic Annotation Pipeline. This will completely annotate your bacterial genome and provide you with a Sequin submission file. N.B. an NCBI Phage Automatic Annotation Pipeline is in developement.

- RAST (Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylogenetic tree. Requires registration. (Reference: Aziz, RK et al. 2008. BMC Genomics 9:75.).

- BASys Bacterial Annotation Tool - this incredible tool supports automated, in-depth annotation of bacterial genomic sequences. It accepts raw DNA sequence data and an optional list of gene identification information (Glimmer) and provides extensive textual annotation and hyperlinked image output. BASys uses >30 programs to determine 60 annotation subfields for each gene, including gene/protein name, GO function, COG function, possible paralogues and orthologues, molecular weight, isoelectric point, operon structure, subcellular localization, signal peptides, transmembrane regions, secondary structure, 3D structure, reactions and pathways. (Reference: G.H. Van Domselaar et al. 2005. Nucl. Acids Res. 33(Web Server issue): W455-W459).

- MicroScope - (CEA, Institut de Génomique - Genoscope, France) is a microbial genome annotation & analysis platform which provides access to a wide range of tools including COG analysis, comparative genomics ... (Reference: Vallenet D et al. (2017) Nucleic Acids Res. 45(D1): D517-D528). Requires registration.

- MAKER Web Annotation Service (MWAS) is an easily configurable web-accesible genome annotation pipeline. It's purpose is to allow research groups with small to intermediate amounts of eukaryotic and prokaryotic genome sequence (i.e. BAC clones, small whole genomes, preliminary sequencing data, etc.) to independently annotate and analyse their data and produce output that can be loaded into a genome database. (Reference: Holt, C. & Yandell, M. 2011. BMC Bioinformatics 12:491).

- MITOS - a pipeline is designed to provide consistent and high quality de novo annotation of Metazoan mitochondrial genomes sequences. We show that the results of MITOS match RefSeq and MitoZoa in terms of annotation coverage and quality. At the same time we avoid biases, inconsistencies of nomenclature, and typos originating from manual curation strategies. (Reference: M. Bernt et al. 2013. Molecular Phylogenetics & Evolution 69:313-319).

- GenSAS - Genome Sequence Annotation Server - provides a one-stop website with a single graphical interface for running multiple structural and functional annotation tools, enabling visualization and manual curation of genome sequences. Users can upload sequences into their account and run gene prediction programs, protein homology searches, map ESTs, identify repeats, ORFs and SSRs with custom parameter settings. Each analysis is displayed on separate tracks of the graphical interface with custom editabe tracks to select final annotation of features and create gff3 files for upload to genome browsers such as GBrowse. Additional programs can be easily added using this Drupal based software.

- Viral Genome ORF Reader (VIGOR) - supports high throughput feature prediction and annotation. VIGOR employs an extrinsic strategy and boasts sensitivity and specificity greater than 98% for the RNA viral genomes we tested. Genome-specific features identified by VIGOR include frameshifts, ribosomal slippage, RNA editing, stop codon read-through, overlapping genes, embedded genes, and mature peptide cleavage sites. Genotyping capability for influenza and rotavirus is built into the program. (Reference: S. Wang et al. 2011. BMC Bioinformatics 2010, 11:451)

- FLAN (FLu ANnotation) is an NCBI web server for genome annotation of influenza virus is a tool for user-provided influenza A virus or influenza B virus sequences. It can validate and predict protein sequences encoded by an input flu sequence. (Reference: Y. Bao et al. 2007. Nucleic Acids Res. Web Server issue) 35: W280-W284.)

- CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission Tool) - allows accurate chloroplast genome annotation, the generation of circular maps, the provision of useful analysis results of the annotated genome, the creation of files that can be submitted to GenBank directly. (Reference: C. Liu et al. 2012. BMC Genomics 13: 715)

- Genome Annotation Transfer Utility (GATU) annotates a genome based on a very closely related reference genome. The proteins/mature peptides of the reference genome are BLASTed against the genome to be annotated in order to find the genes/mature peptides in the genome to be annotated (Reference: T. Tcherepanov et al. 2006. BMC Genomics 7:150.)

- BioGPS *(The Scripps Research Institute, USA)* - is a one-stop gene annotation portal that emphasizes user-customizability and community-extensibility It is a customizable gene annotation portal and a complete resource for learning about gene and protein function.

- BAGEL (*Groningen Biomolecular Sciences and Biotechnology Institute, Haren, the Netherlands*) - will determine from an existing or non submitted GenBank file the presence of bacteriocins based on a database containing information of known bacteriocins and adjacent genes involved in bacteriocin activity. An alternative site for bacteriocins is BACTIBASE which is a data repository of bacteriocin natural antimicrobial peptides. See.LABioicin if you are interested in the topic of Lactic Acid Bacteria (LAB) and its bacteriocins.

- MICheck (MIcrobial genome Checker) - enables rapid verification of sets of annotated genes and frameshifts in previously published bacterial genomes, or genomes for which the user has a *.gbk file. This tool can be seen as a preliminary step before the functional re-annotation step to check quickly for missing or wrongly annotated genes. It worked nicely with phage genomes from 43-135kb. (Reference: S. Cruveiller et al. 2005. Nucl. Acids Res. 33: W471- W479).

- WebGeSTer - Genome Scanner for Terminators - my favourite terminator search program is finally web enabled. Please note that if you want to analyze data from a *.gbk file you need to use their conversion program "GenBank2GeSTer" first. A complete description of each terminator including a diagram is produced by this program. This site linked to an extensive database of transcriptional terminators in bacterial genome (WebGeSTer DB) (Reference: Mitra A. et al. 2011. Nucl. Acids Res. 39(Database issue):D129-35).

- RibEx: Riboswitch Explorer - scans <40kb DNA for potential genes (which are linked to BLASTP) and several hundred regulatory elements, including riboswitches. If you click on the "search for attenuators" it finds terminators and antiterminators. It presents the capculated genes and perits BLAST analysis at NCBI (Reference: C. Abreu-Goodger & E. Merino. 2005. Nucl. Acids Res. 33: W690-W692).

- tRNAs: tRNAscan-SE - is incredibly sensitive & also provides secondary structure diagrams of the tRNA molecules (Reference: Schattner, P. et al. 2005. Nucleic Acids Res. 33: W686-689). Alternatively use ARAGORN (Reference: Laslett, D. & Canback. 2004. Nucleic Acids Research 32:11-16).
  Test sequences.

- LTR_Finder - is an efficient program for finding full-length LTR retrotranspsons in genome sequences. The size of input file is now limited to 50MB (Reference: Z. Xu & H. Wang. 2007. Nucl. Acids Res.35(Web Server issue): W265-W268).
- RTAnalyzer - finds retrotransposons and detects L1 retrotransposition signatures (Reference: J-F. Lucier et al. 2007. Nucl. Acids Res. 35(Web Server issue):W269-W274

- MG-RAST (Metagenome Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating metagenome samples. It provides annotation of sequence fragments, their phylogenetic classification and an initial metabolic reconstruction. The service also provides means for comparing phylogenetic classifications and metabolic reconstructions of metagenomes (Reference: F. Meyer et al. 2008. BMC Bioinformatics 9: 386).

The following four programs can be used to prediction phage proteins:

- PVPred (Reference: Ding H et al (2014) Mol Biosyst 10(8): 2229-2235).
- PHPred (Reference: Ding H (2016) Computers Biol Med 71: 156–161).
- PVP-SVM (Reference: Manavalan B et al. (2018) Front Microbiol 9: 476).
- PVPred-SCM (Reference: Charoenkwan P et al. (2020) Cells 9(2) pii: E353.

Chromosome replication origin:

- Ori-Finder and Ori-Finder 2 - are useful platforms for the identification and analysis of replication origins (oriCs) in the bacterial and archaeal genomes, respectively. (Reference: Luo H et al. (2019) Brief Bioinform 20(4): 1114-1124). Please note that these tools have been used to create DoriC - a database of replication origins in prokaryotic genomes including chromosomes and plasmids. (Reference: Luo H & Gao F (2019) Nucleic Acids Res. 47(D1): D74-D77).

Correcting genome annotations:

- One of the problems with GenBank is that scientists do not update their submission data nor correct errors. In part this is due to laziness; but is also due to the fact that GenBank is, in most cases, unwilling to accept a new version of the Sequin file. Tbl2asn is a command-line program that automates the creation of sequence records for submission to GenBank but, from my perspective, it is not easy to use. The only online program is GenBank 2 Sequin which generates not only a Sequin file (*.sqn), but also a five-column "Annotation Table" (*.tbl). This together with the fasta-formatted DNA sequence can be submitted to GenBank by Email (gb-admin@ncbi.nlm.nih.gov). In its absence I recommend the perl script gbf2tbl.pl available for downloading here.

Specialized annotation - general

- PlasmidFinder 1.3 - identifies plasmids in total or partial sequenced isolates of bacteria. The method uses BLAST for identification of replicons of plasmids belonging to the major incompatibility (Inc) groups of *Enterobacteriaceae*. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms. See also pMLST (Reference: Carattoli A et al. 2014. Antimicrob. Agents Chemother. 58: 3895-903)

- PHACTS can be used to quickly classify the lifestyle of a phage (temperate or lytic). All that is needed is the proteome of the phage to be classified and PHACTS will predict the lifestyle of that phage and return a confidence value for that prediction. (Reference: K. McNair et al. 2012. Bioinformatics 28: 614-618).

- SpeciesFinder 1.0 *(Danish Technical University)* - predicts the species of bacteria from pre-assembled, complete or partial genomes, and short sequence reads. The prediction is based on the 16S rRNA gene.

- CSI Phylogeny 1.1 (Call SNPs & Infer Phylogeny) - calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs. (Reference: Kaas, R.S. et al. PLoS ONE 2014; 9: e104984.)

- KmerFinder 2.0 – predicts the species of bacteria from pre-assembled, complete or partial genomes, and short sequence reads. The prediction is based on the number of co-occurring k-mers (substrings of k nucleotides in DNA sequence data, in this case 16-mers) between the genomes of reference bacteria in a database and the genome provided by the user. (Reference: Hasman H et al. 2013. J Clin Microbiol. 52:139-146)

- VIOLIN: Vaccine Investigation and Online Information Network - allows easy curation, comparison and analysis of vaccine-related research data across various human pathogens VIOLIN is expected to become a centralized source of vaccine information and to provide investigators in basic and clinical sciences with curated data and bioinformatics tools for vaccine research and development. VBLAST: Customized BLAST Search for Vaccine Research allows various search strategies against against 77 genomes of 34 pathogens. (Reference: He, Y. et al. 2014. Nucleic Acids Res. 42 (Database issue):D1124-32).

- MLST 1.8 (MultiLocus Sequence Typing) - currently only works with assembled genomes and contigs (Reference: Larsen MV et al. 2012. J. Clin. Micobiol. 50: 1355-1361).

- ECFfinder - extracytoplasmic function (ECF) sigma factors - the largest group of alternative sigma factors - represent the third fundamental mechanism of bacterial signal transduction, with about six such regulators on average per bacterial genome. Together with their cognate anti-sigma factors, they represent a highly modular design that primarily facilitates transmembrane signal transduction. (Reference: Staron A et al. (2009) Mol Microbiol 74(3): 557-581).

- BacWGSTdb - is designed for monitoring the emergence and outbreak of important bacterial pathogens. In detail, it serves two particular purposes: Typing & Tracking. The former refers to an integrated genotyping at both the traditional multi-locus sequence typing (MLST) and whole-genome sequencing typing (WGST) level. The latter refers to source tracking (i.e., finding highly similar isolates) according to the typing result and isolates information stored in BacWGSTdb. (Reference: Z. Ruan 7 Y. Feng, Nucleic Acids Research. 2016; 44(D1): D682-D687).

- SISTR: *Salmonella* In Silico Typing Resource - *(Public Health Agency of Canada, Laboratory for Foodborne Zoonoses)* is a bioinformatics resource for rapidly interpreting in silico data for multiple Salmonella subtyping methods from draft bacterial genome assemblies. In addition to performing serovar prediction by genoserotyping, this resource integrates sequence-based typing analyses for: Multi-Locus Sequence Typing (MLST), ribosomal MLST (rMLST), and core genome MLST (cgMLST). Google Chrome is recommended; Firefox is also supported but the SVG visualizations within this app may not be as responsive. Internet Explorer is unsupported.

- FSFinder2 (Frameshift Signal Finder) - Programmed ribosomal frameshifting is involved in the expression of certain genes from a wide range of organisms such as virus, bacteria and eukaryotes including human. In programmed frameshifting, the ribosome switches to an alternative frame at a specific site in response to a special signal in a messanger RNA. Programmed frameshift plays role in viral particle morphogenesis, autogenous control, and alternative enzymatic activities. The common frameshift is a -1 frameshift, in which the ribosome shifts a single nucleotide in the upstream direction. The major elements of -1 frameshifting consist of a slippery site, where the ribosome changes reading frames, and a stimulatory RNA structure such as pseudoknot or stem-loop located a few nucleotides downstream. +1 frameshifts are much less common than -1 frameshifting but are observed in diverse organisms.

● InBase, The Intein Database and Registry - Protein splicing is defined as the excision of an intervening protein sequence (the INTEIN) from a protein precursor and the concomitant ligation of the flanking protein fragments (the EXTEINS) to form a mature extein host protein and the free intein (Perler 1994). Protein splicing results in a native peptide bond between the ligated exteins. This is a database site which permits BLAST analysis. (Reference: Perler, F.B. 2002. Nucleic Acids Res. 30: 383-384).

● Two-component and other regulatory proteins:

● P2RP (Predicted Prokaryotic Regulatory Proteins) - users can input amino acid or genomic DNA sequences, and predicted proteins therein are scanned for the possession of DNA-binding domains and/or two-component system domains. RPs identified in this manner are categorised into families, unambiguously annotated. (Reference: Barakat M, et al. 2013. BMC Genomics 14:269).

● P2CS (Prokaryotic 2-Component Systems) is a comprehensive resource for the analysis of Prokaryotic Two-Component Systems (TCSs). TCSs are comprised of a receptor histidine kinase (HK) and a partner response regulator (RR) and control important prokaryotic behaviors. It can be searched using BLASTP. (Reference: P. Ortet et al. 2015. Nucl. Acids Res. 43 (D1): D536-D541).

● Orthologous genes/proteins

COG analysis - Clusters of Orthologous Groups - COG protein database was generated by comparing predicted and known proteins in all completely sequenced microbial genomes to infer sets of orthologs. Each COG consists of a group of proteins found to be orthologous across at least three lineages and likely corresponds to an ancient conserved domain (CloVR). Sites which offer this analysis include:

● WebMGA (Reference: S. Wu et al. 2011. BMC Genomics 12:444), RAST (Reference: Aziz RK et al. 2008. BMC Genomics 9:75), and BASys (Bacterial Annotation System; Reference: Van Domselaar GH et al. 2005. Nucleic Acids Res. 33(Web Server issue):W455-459.) and JGI IMG (Integrated Microbial Genomes; Reference: Markowitz VM et al. 2014. Nucl. Acids Res. 42: D560-D567.)

Other sites:

- EggNOG - A database of orthologous groups and functional annotation that derives Nonsupervised Orthologous Groups (NOGs) from complete genomes, and then applies a comprehensive characterization and analysis pipeline to the resulting gene families. (Reference: Powell S et al. 2014. Nucleic Acids Res. 42 (D1): D231-D239

- OrthoMCL - is another algorithm for grouping proteins into ortholog groups based on their sequence similarity. The process usually takes between 6 and 72 hours.(Reference: Fischer S et al. 2011. Curr Protoc Bioinformatics; Chapter 6:Unit 6.12.1-19).

- KAAS (KEGG Automatic Annotation Server) provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database. The result contains KO (KEGG Orthology) assignments and automatically generated KEGG pathways. (Reference: Moriya Y et al. 2007. Nucleic Acids Res. 35(Web Server issue):W182-185).

- **Specialized annotation - antibiotic resistance.**

- ResFinder (Acquired antimicrobial resistance gene finder) - uses BLAST for identification of acquired antimicrobial resistance genes in whole-genome data. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms. Tested with 1411 different resistance genes with 100% identity. (Reference: Zankari E et al. 2012. J Antimicrob Chemother. 67:2640-2644)

- ARG-ANNOT (Antibiotic Resistance Gene-ANNOTation) is a new tool that was created to detect existing and putative new antibiotic resistance (AR) genes in bacterial genomes. ARG-ANNOT uses a local blast program in Bio-Edit software that allows the user to analyze sequences without web interface (Reference: Gupta, S.K. et al. 2014. Antimicrob Agents Chemother. 58: 212–220).

- CARD (The Comprehensive Antibiotic Resistance Database) - a rigorously curated collection of known resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO) and AMR gene detection models (Reference: Jia, B. et al. 2017. Nucleic Acids Research, 45: D566-573).

- MEGARes - is a hand-curated antimicrobial resistance database and annotation structure that provides a foundation for the development of high throughput acyclical classifiers and hierarchical statistical analysis of big data (Reference: Lakin, S.N.. et al. 2017. Nucleic Acids Research, 45: D574-D580).

- BacMet (Antibacterial Biocide & Metal Resistance Genes Database) - a database of biocide and metal resistance genes with highly reliable content. In BacMet version 1.1, the experimentally confirmed database contains 704 resistance genes, whereas the predicted database contains 40,556 resistance genes (Reference: Pal, C. et al. 2014. Nucleic Acids Research, 42: D737-743).

- Specialized annotation - CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats):

- CRISPRfinder - enables the easy detection of CRISPRs in locally-produced data and consultation of CRISPRs present in the database. It also gives information on the presence of CRISPR-associated (cas) genes when they have been annotated as such. . (Reference: I. Grissa et al. 2007. Nucl. Acids Res. 35 (Web Server issue): W52-W57).

- CRISPRmap -provides a quick and detailed insight into repeat conservation and diversity of both bacterial and archaeal systems. It comprises the largest dataset of CRISPRs to date and enables comprehensive independent clustering analyses to determine conserved sequence families, potential structure motifs for endoribonucleases, and evolutionary relationships. (Reference: S.J. Lange et al. 2013. Nucleic Acids Research, 41: 8034-8044).

- CRISPI : a CRISPR Interactive database - includes a complete repertory of associated CRISPR-associated genes (CAS). A user-friendly web interface with many graphical tools and functions allows users to extract results, find CRISPR in personal sequences or calculate sequence similarity with spacers.(Reference: Rousseau C et al. 2009. Bioinformatics. 25: 3317–3318).

- CRISPRTarget - that predicts the most likely targets of CRISPR RNAs. This can be used to discover targets in newly sequenced genomic or metagenomic data. (Reference: Biswas A et al. 2013. RNA Biol. 10:817-827).

- CRISPy-web - is an easy to use web tool based on CRISPy to design sgRNAs for any user-provided microbial genome. CRISPy-web allows researchers to interactively select a region of their genome of interest to scan for possible sgRNAs. After checks for potential off-target matches, the resulting sgRNA sequences are displayed graphically and can be exported to text files. (Reference: K. Blin et al. 2016. Synthetic and Systems Biotechnology 1(2): 118-121).

- Specialized annotation - virulence determinants: This is of particular interest to those working on bacteriophages for therapy

- VirulenceFinder *(Danish Technical University)* – identification of virulence genes. The method uses BLAST for identification of known virulence genes in *Escherichia coli*. The method is being extended to also include virulence genes for *Enterococcus* and *Staphylococcus aureus*. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms.

- ClanTox: a classifier of short animal toxins - predicts whether each sequence is toxin-like and provides a ranked list of positively predicted candidates according to statistical confidence. For each protein, additional information is presented including the presence of a signal peptide, the number of cysteine residues and the associated functional annotations. (Reference: G. Naamati et al. 2009. Nucleic Acids Res. 37(Web Server issue): W363–W368).

- t3db the Toxin and Toxin Target Database - combines detailed toxin data with comprehensive toxin target information. The database currently houses 3,053 toxins which are linked to 1,670 corresponding toxin target records. Each toxin record (ToxCard) contains over 50 data fields and holds information such as chemical properties and descriptors, toxicity values, molecular and cellular interactions, and medical information. (Reference: Lim E et al. 2010. Nucleic Acids Res. 38(Database issue): D781-786).

- TAfinder 2.0 - is a web-based tool to identify Type II toxin-antitoxin loci in bacterial genome (Reference: Xie Y et al. (2018) Nucleic Acids Res. 46(D1): D749-D753).

- DBETH Database of Bacterial ExoToxins for Humans is a database of sequences, structures, interaction networks and analytical results for 229 exotoxins, from 26 different human pathogenic bacterial genus. All toxins are classified into 24 different Toxin classes. The aim of DBETH is to provide a comprehensive database for human pathogenic bacterial exotoxins. (Reference: Chakraborty A et al. 2012. Nucleic Acids Res. 40(Database issue): D615-620).

- VFDB - is an integrated and comprehensive database of virulence factors for bacterial pathogens (also including Chlamydia and Mycoplasma). (Reference: L.H. Chen et al. 2012. Nucleic Acids Res. 40(Database issue): D641-D645).

- PAIDB (Pathogenicity Island Database) - Pathogenicity islands (PAIs) and resistance islands (REIs) are key to the evolution of pathogens and appear to play complimentary roles in the process of bacterial infection. While PAIs promote disease development, REIs give a fitness advantage to the host against multiple antimicrobial agents. An anncillary program, PAI Finder, identifies PAI-like regions or REI-like regions in a multi-sequence query. (Reference: S.H Yoon et al. 2015. Nucl. Acids Res. 43 (D1): D624-D630).

- IslandViewer - includes a new interactive genome visualization tool, IslandPlot, and expanded virulence factor, antimicrobial resistance gene, and pathogen-associated gene annotations, as well as homologs of these genes in closely related genomes. Notably, incomplete genomes are accepted as input in IslandViewer 3, though they strongly urge users to use complete genomes whenever possible. (Reference: B.K. Dhillon et al. 2015. Nucl. Acids Res. 43 (W1): W104-W108).

- Gypsy Database - an open editable database about the evolutionary relationship of viruses, mobile genetic elements (MGEs; Ty3/Gypsy, Retroviridae, Ty1/Copia and Bel/Pao LTR retroelements and the *Caulimoviridae* pararetroviruses of plants) and other genomic repeats. Equipped for BLAST and HMM searches. (Reference: Llorens, C et al. 2011. Nucl. Acids Res. 39(suppl 1): D70-D74).

- PanDaTox (Pan Genomic Database for Genomic Elements Toxic to Bacteria) - is a database of genes and intergenic regions that are unclonable in E. coli, to aid n the discovery of new antibiotics and biotechnologically beneficial functional genes. It is also designed to improve the efficiency of metabolic engineering. BLAST Search feature included. (Reference: Mitai G & Sorek R. 2012. Bioengineered, 3: 218-221.)

- PathogenFinder (predicts pathogenic potential) – Based on complete genomes from 513 bacteria annotated as human non-pathogens and 372 bacteria annotated as human pathogens, a database of protein families, which are either mainly associated with non-pathogens or with pathogens have been created. This database is then used for predicting the pathogenic potential of bacteria. As input, the method can use both pre-assembled, complete or partial genomes, and short sequence reads from four different sequencing platforms. (Reference: Cosentino S et al. 2013. PLoS ONE 8: e77302)

- VirulentPred - is a SVM based method to predict bacterial virulent proteins sequences, which can be used to screen virulent proteins in proteomes. Together with experimentally verified virulent proteins, several putative, non annotated and hypothetical protein sequences have been predicted to be high scoring virulent proteins by the prediction method. (Reference: Garg A & Gupta G. 2008. BMC Bioinformatics 9: 62).

- The Type III secretion system (T3SS) is an essential mechanism for host-pathogen interaction in the infection process. The proteins secreted through the T3SSmachinery of many Gram-negative bacteria are known as T3SS effectors (T3SEs). These can either be localized subcellularly in the host, or be part of the needle tip of the T3SS that interacts directly with the host membrane to bring other effectors into the target cell. T3SEdb represents such an effort to assemble a comprehensive database of all experimentally determined and putative T3SEs into a web-accessible site. BLAST search is available. (Reference: Tay DM et al. 2010. BMC Bioinformatics. 11 Suppl 7:S4).

- Effective (*University of Vienna, Austria & Technical University of Munich, Germany*) - Bacterial protein secretion is the key virulence mechanism of symbiotic and pathogenic bacteria.Thereby effector proteins are transported from the bacterial cytosol into the extracellular medium or directly into the eukaryotic host cell. The Effective portal provides precalculated predictions on bacterial effectors in all publicly available pathogenic and symbiontic genomes as well as the possibility for the user to predict effectors in own protein sequence data.

- SIEVE Server is a public web tool for prediction of type III secreted effectors. The SIEVE Server scores potential secreted effectors from genomes of bacterial pathogens with type III secretion systems using a model learned from known secreted proteins. The SIEVE Server requires only protein sequences of proteins to be screened and returns a conservative probability that each input protein is a type III secreted effector. (Reference: McDermott JE et al. 2011. Infect Immun. 79:23-32).

- T3SE - Type III secretion system effector prediction (Reference: Löwer M, & Schneider G. 2009. PLoS One. 4:e5917. Erratum in: PLoS One. 2009;4(7).

Specialized annotation - Genomic Islands:

- Phage_Finder - was created to identify prophage regions in completed bacterial genomes. Using a test dataset of 42 bacterial genomes whose prophages have been manually identified, *Phage_Finder* found 91% of the regions, resulting in 7% false positive and 9% false negative prophages. A search of 302 complete bacterial genomes predicted 403 putative prophage regions, accounting for 2.7% of the total bacterial DNA. Analysis of the 285 putative attachment sites revealed tRNAs are targets for integration slightly more frequently (33%) than intergenic (31%) or intragenic (28%) regions, while tmRNAs were targeted in 8% of the regions. (Reference: D.E. Fouts. 2006. Nucleic Acids Res. 34: 5839–5851).

- Prophinder - is the tool used for detecting prophages in bacterial genomes. Select a GenBank formatted file.

- PHAST (PHAge Search Tool) - is designed to rapidly and accurately identify, annotate and graphically display prophage sequences within bacterial genomes or plasmids. It accepts either raw DNA sequence data or partially annotated GenBank formatted data and rapidly performs a number of database comparisons as well as phage "cornerstone" feature identification steps to locate, annotate and display prophage sequences and prophage features. Relative to other prophage identification tools, PHAST is up to 40 times faster and up to 15% more sensitive. It is also able to process and annotate both raw DNA sequence data and Genbank files, provide richly annotated tables on prophage features and prophage "quality" and distinguish between intact and incomplete prophage. PHAST also generates downloadable, high quality, interactive graphics that display all identified prophage components in both circular and linear genomic views.Furthermore, tests indicate that PHAST is as accurate or slightly more accurate than all available phage finding tools, with sensitivity of 85.4% and positive predictive value of 94.2%. (Reference: Zhou, Y. et al. 2011. Nucl. Acids Res. 39(suppl 2): W347-W352).

- PHASTER PHAge Search Tool Enhanced Release - is a significant upgrade to PHAST for the rapid identification and annotation of prophage sequences within bacterial genomes and plasmids. Numerous software improvements and significant hardware enhancements have now made PHASTER faster, more efficient, more visually appealing and much more user friendly. In particular, PHASTER is now 4.3X faster than PHAST. (Reference: D. Arndt et al. Nucleic Acids Res. 2016; 44(W1):W16-21).

- Prophage Hunter - provides a one-stop web service to extract prophage genomes from bacterial genomes, evaluate the activity of the prophages, identify phylogenetically related phages, and annotate the function of phage proteins. (Reference: Song W et al. (2019) Nucleic Acids Res 47(W1): W74–W80).

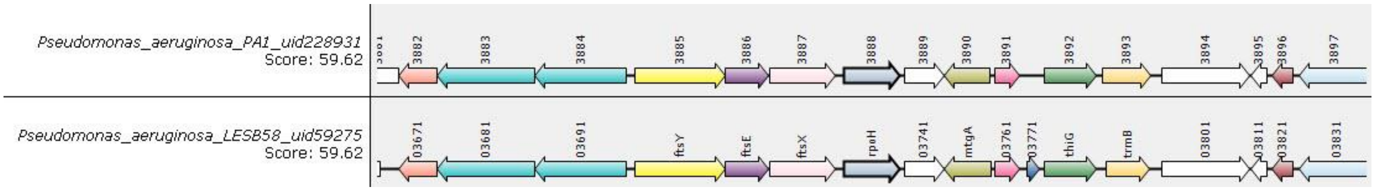
- IslandViewer - integrates two sequence composition GI prediction methods SIGI-HMM and IslandPath-DIMOB, and a single comparative GI prediction method IslandPick (Reference: Langille et al. 2008. BMC Bioinformatics 9: 329).

- PAIDB (PAthogenicity Island DataBase) has made an effort to collect known PAIs and to detect the potential PAI regions in the prokaryotic complete genomes. Pathogenicity islands (PAIs) are distinct genetic elements of pathogens encoding various virulence factors. (Reference: Yoon SH et al. 2007. Nucleic Acids Res. 35 (Database Issue): D395-D400).

- MTGIpick can identify genomic islands from a single genome, without annotated information of genomes or prior knowledge from other datasets. In simulations with alien fragments from artificial and real genomes, MTGIpick reported robust results across different experiments (Reference: Dai Q et al. (2018) Brief Bioinform 19(3): 361-373).

- **Genome comparisons and synteny:**

- SyntTax - is a web server linking synteny to prokaryotic taxonomy. SyntTax incorporates a full hierarchical taxonomic tree allowing intuitive access to all completely sequenced prokaryotes (Archaea and Bacteria). Single or multiple organisms can be chosen on the basis of their lineage by selecting the corresponding rank nodes in the tree. This is my favourite among the synteny programs (Reference: Oberto J. 2013. BMC Bioinformatics. 14:4). The results below were generated using the heat-shock sigma factor (RpoH) from *Salmonella Typhimurium* against the *Pseudomonadales*.



- Cinteny Server for Synteny Identification and Analysis of Genome Rearrangement *(A. U. Sinha & J. Meller, University of Cincinnati, USA)* - this server can be used for finding regions syntenic across multiple genomes and measuring the extent of genome rearrangement using reversal distance as a measure. You may create a project and upload your own data or work with pre-loaded prokaryote or eukaryote data.

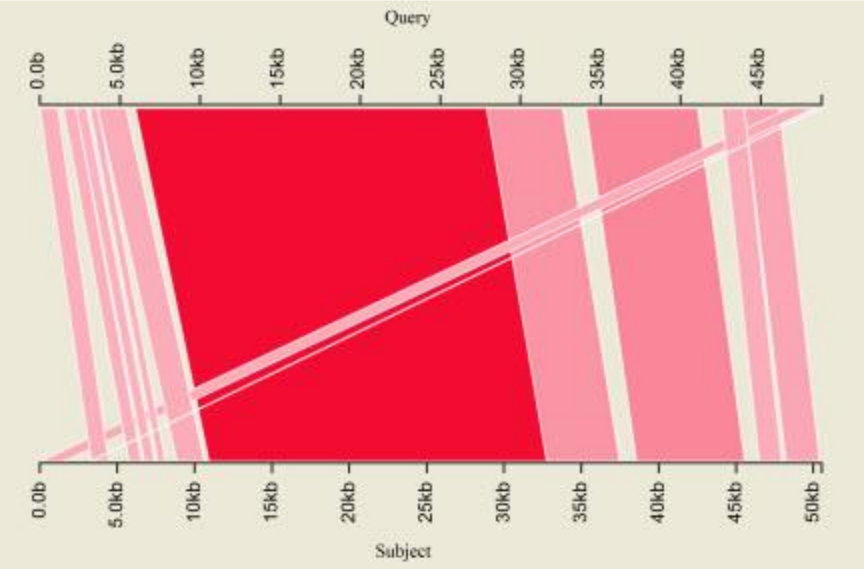
- SimpleSynteny - provides a pipeline for evaluating the synteny of a preselected set of gene targets across multiple organismal genomes. An emphasis has been placed on ease-of-use, and users are only required to submit FASTA files for their genomes and genes of interest. SimpleSynteny then guides the user through an iterative process of exploring and customizing genomes individually before combining them into a final high-resolution figure. (Reference: Veltri D et al. 2016. Nucleic Acids Res. 44(Web Server issue): W41–W45).

- Synteny Portal - eukaryotic genome users can easily (i) construct synteny blocks among multiple species by using prebuilt alignments in the UCSC genome browser database, (ii) visualize and download syntenic relationships as high-quality images, (iii) browse synteny blocks with genetic information and (iv) download the details of synteny blocks to be used as input for downstream synteny-based analyses, all in an intuitive and easy-to-use web-based interface. (Reference: Lee J et al. 2016. Nucleic Acids Res 44(W1): W35–W40).

- AutoGRAPH is an integrated web server for multi-species comparative genomic analysis. It is designed for constructing and visualizing synteny maps between two or three species, determination and display of macrosynteny and microsynteny relationships among species, and for highlighting evolutionary breakpoints. The web server constructs synteny maps by pairwise comparison of marker/anchor orders between a reference chromosome and one or two tested genome(s). It permits users to visualize and characterize several features: Conserved segments (CS), Conserved Segments Ordered (CSO) and breakpoints. (Reference: Derrien T et al. 2007. Bioinformatics 23:498-499).

- Sibelia *(University of California San Diego, USA)* - is a tool for finding synteny blocks in multiple closely related microbial genomes using iterative de Bruijn graphs. Unlike most other tools, Sibelia can find synteny blocks that are repeated within genomes as well as blocks shared by multiple genomes. It represents synteny blocks in a hierarchy structure with multiple layers, each of which representing a different granularity level.

- Kablammo helps you create interactive visualizations of BLAST results from your web browser. Find your most interesting alignments, list detailed parametersfor each, and export a publication-ready vector image. Incredibly easy to use - here are the results for a BLASTN comparison to *Escherichia* phages T1 (query) and ADB-2. (Reference: Wintersinger JA et al. Bioinformatics 31:1305-1306).

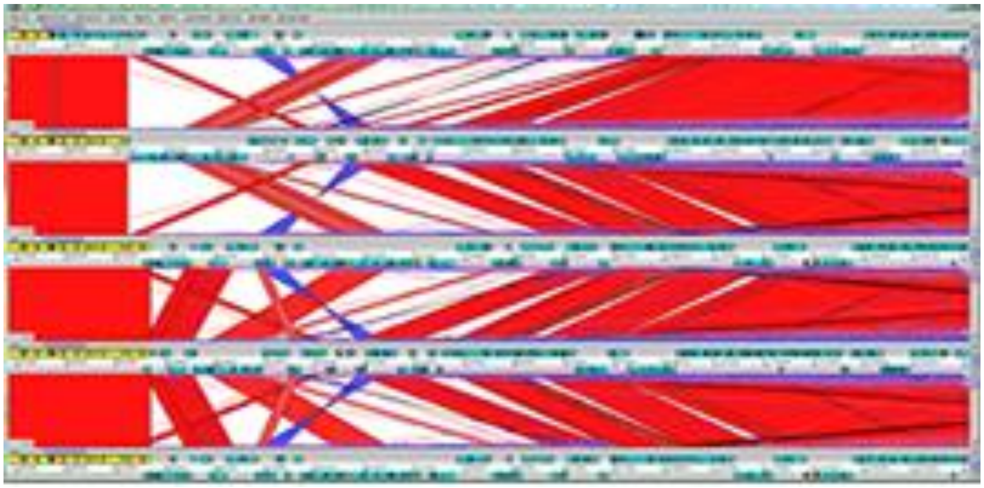


- M1CR0B1AL1Z3R - is a 'one-stop shop' for conducting microbial genomics data analyses via a simple graphical user interface. Some of the features implemented in M1CR0B1AL1Z3R are: (i) extracting putative open reading frames and comparative genomics analysis of gene content; (ii) extracting orthologous sets and analyzing their size distribution; (iii) analyzing gene presence-absence patterns; (iv) reconstructing a phylogenetic tree based on the extracted orthologous set; (v) inferring GC-content variation among lineages. M1CR0B1AL1Z3R facilitates the mining and analysis of dozens of bacterial genomes using advanced techniques. (Reference: Avram O et al. (2019) Nucleic Acids Res. 47(W1): W88-W92).

- GeneOrder 4.0 *(D. Seto, Bioinformatics & Computational Biology, George Mason Univ., U.S.A.)* is designed to can be used to compare the gene order between two bacterial genomes (Reference: Mahadevan P. & Seto D. 2010. BMC Research Notes 3:41).
- CoreGenes *(D. Seto & P. Mahadevan, Bioinformatics & Computational Biology, George Mason Univ., U.S.A)* - tallies the total number of genes in common between the two genomes being compared; displays the percent value of genes in common with a specific genome; determines the unique genes contained in a pair of proteomes. CoreGenes 3.5 is the batch CoreGenes server. I have extensively used this set of resources in the classification of bacterial viruses.

- If you have a a gbk file for a phage which has not yet been deposited in GenBank you can use these instructions to convert your data into CoreGenes format for use here.

- WebACT - this is the web version of ACT (Artemis Comparison Tool) a DNA sequence comparison viewer based on Artemis (Reference: T.J. Carver et al. Bioinformatics 21: 3422 - 3423). Visit the database page of EMBL-EBI and select EMBL and "Standard Query Form" to determine the EMBL accession number for the sequence you are interested in.



- Panseq *(Chad Laing, Public Health Agency of Canada)* - a group of tools for the analysis of the 'pan genome' of a group of genomic sequences. The pan-genome of a bacterial species consists of a core genome and an accessory gene pool, the latter of which allows subpopulations of the organism to adapt to specific environments. These include Novel Region Finder, which will find sequences that are unique to a strain or group of strains with respect to another strain or group of strains. Pan-genome Analysis identifies the pan-genome among your sequences; and, finds SNPs in the core genome and determine the distribution of accessory genomic regions.Loci Selector identifies loci that offer the best discrimination among your dataset. (Reference: Laing, C. et al. 2010. BMC Bioinformatics. 11: 461).

- PARIGA - enables users to perform all-against-all BLAST searches on two sets of sequences selected by the user. Moreover, since it stores the two BLAST output in a python-serialized-objects database, results can be filtered according to several parameters in real-time fashion, without re-running the process and avoiding additional programming efforts. (Reference: Orsini M. et al. 2013. PLoS One 8(5):e62224).

- EDGAR (Efficient Database framework for comparative Genome Analyses using BLAST score Ratios) - EDGAR is designed to automatically perform genome comparisons in a high throughput approach and can be used for core genome, pan genome and singleton analysis, and Venn diagram construction. (Reference: Blom J. et al. 2009. BMC Bioinformatics 10: 154).

- OrthoVenn - is a web server for genome wide comparison and annotation of orthologous clusters across multiple species. It provides coverage of vertebrates, metazoa, protists, fungi, plants and bacteria for the comparison of orthologous clusters and also supports uploading of customized protein sequences from user-defined species. An interactive Venn diagram, summary counts, and functional summaries of the disjunction and intersection of clusters shared between species are displayed as part of the OrthoVenn result. OrthoVenn also includes in-depth views of the clusters using various sequence analysis tools. Furthermore, it identifies orthologous clusters of single copy genes and allows for a customized search of clusters of specific genes through key words or BLAST. (Reference: Y. Yang et al. 2015. Nucl. Acids Res. 43 (W1): W78-W84). Also found here.

- BEACON is a software tool that compares annotations of a particular genome from different Annotation Methods (AMs). It uses GenBank format as input and derives Extended Annotation

(EA) along side listing original annotations from individual AMs. (Reference: Kalkatawi M, BMC Genomics. 2015;16(1): 1-8).

● **Phylogeny (AAI and ANI)**

● ANI (Average Nucleotide Identity) calculator - estimates the average nucleotide identity using both best hits (one-way ANI) and reciprocal best hits (two-way ANI) between two genomic datasets. Typically, the ANI values between genomes of the same species are above 95% (e.g., Escherichia coli). Values below 75% are not to be trusted, and AAI should be used instead. This tool supports both complete and draft genomes (multi-fasta). (Reference: Goris J et al. 2007. Int J Syst Evol Microbiol. 57(Pt 1): 81-91).

● Average Nucleotide Identity (ANI) calculator - their ANI Calculator uses the OrthoANIu algorithm, an improved iteration of the original OrthoANI algorithm, which uses USEARCH instead of BLAST (Reference: Yoon, S. H. et al. (2017). Antonie van Leeuwenhoek. 110:1281–1286).

● VIRIDIC (Virus Intergenomic Distance Calculator; *C. Moraru, Institute for Chemistry and Biology of the Marine Environment, Germany*) - the first level of bacteriophage classification by ICTV involves computing the overall DNA sequence identity between two viruses. This new tool computes pairwise intergenomic distances/similarities amongst phage genomes. To run it, upload a single fasta file with all phage genomes of interest, create a project and press run. Save the project ID that will be displayed when the project is created. You will need it to access the data if the calculations take a long time.

● GGDC (Genome-To-Genome Distance Calculator) - provides methods for inferring whole-genome distances which are well able to mimic DNA-DNA hybridization (DDH). Values calculated with GGDC yield a somewhat better correlation with wet-lab DDH values than alternative approaches such as "ANI". These distance functions can also cope with heavily reduced genomes and repetitive sequence regions. Some of them are also very robust against missing fractions of genomic information (due to incomplete genome sequencing). Thus, this web service can be used for genome-based species delineation. (Reference: Meier-Kolthoff JP et al. 2013. BMC Bioinformatics 14: 60).

● POGO-DB - Based on computationally intensive whole-genome BLASTs, POGO-DB provides several metrics on pairwise genome: (a) Average Amino Acid Identity of all bi-directional best blast hits that covered at least 70% of the sequence and had 30% sequence identity; (b) Genomic Fluidity that estimates the similarity in gene content between two genomes; (c) Number of orthologs shared between two genomes (as defined by two criteria); (d) Pairwise identity of the most similar 16S rRNA genes; (e) Pairwise identity of 73 additional globally-conserved marker genes (which were determined by us to exist in at least 90% of all the genomes). (Reference: Lan Y et al. 2014. Nucl. Acids Res. 42 (D1): D625-D632).

● VICTOR (Virus Classification and Tree Building Online Resource; *Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen Gmb*H). This web service compares bacterial and archaeal viruses ("phages") using their genome or proteome sequences. The results include phylogenomic trees inferred using the Genome-BLAST Distance Phylogeny method (GBDP), with branch support, as well as suggestions for the classification at the species, genus and family level. (The service can be applied to other kinds of viruses, too, but has not yet been tested in this respect.) Upload your FASTA files, GenBank files and/or GenBank accession IDs. (Reference: JP Meier-Kolthoff & M Göker. 2017. Bioinformatics 33(21): 3396–3404).

- VIRFAM is dedicated to the recognition of head-neck-tail modules and of recombinase genes in phage genomes. You can use this server to search for remote homologs of specific protein families within protein sequences of bacteriophages. Input: protein sequences you're your phage; output includesd a phylogenetic tree with the placement of your virus. (Reference: Lopes A et al. Nucleic Acids Res. (2010) 38(12): 3952-62).

- Seeker - is a deep-learning tool for reference-free identification of phage sequences. Seeker allows rapid detection of phages in sequence datasets and clean differentiation of phage sequences from bacterial ones, even for phages with little sequence similarity to established phage families. We comprehensively validate Seeker ability to identify unknown phages and employ Seeker to detect unknown phages, some of which are highly divergent from known phage families. (Reference: Auslander N et al. (2020) doi.org/10.1101/2020.04.04.025783)

- VipTree - generates a "proteomic tree" of viral genome sequences based on genome-wide sequence similarities computed by tBLASTx. The original proteomic tree concept (i.e., "the Phage Proteomic Tree") was developed by Rohwer and Edwards, 2002. A proteomic tree is a dendrogram that reveals global genomic similarity relationships between tens, hundreds, and thousands of viruses. It has been shown that viral groups identified in a proteomic tree well correspond to established viral taxonomies. (Reference: Nishimura Y et al. (2017) Bioinformatics 33: 2379–2380).

- MiGA (Microbial Genomes Atlas) - a webserver that allows the classification of an unknown query genomic sequence, complete or partial, against all taxonomically classified taxa with available genome sequences, as well as comparisons to other related genomes including uncultivated ones, based on the genome-aggregate Average Nucleotide and Amino Acid Identity (ANI/AAI) concepts. (Reference: Rodriguez-R et al (2018) Nucleic Acids Research 46(W1): W282-W288).

- **Genome visualization:**

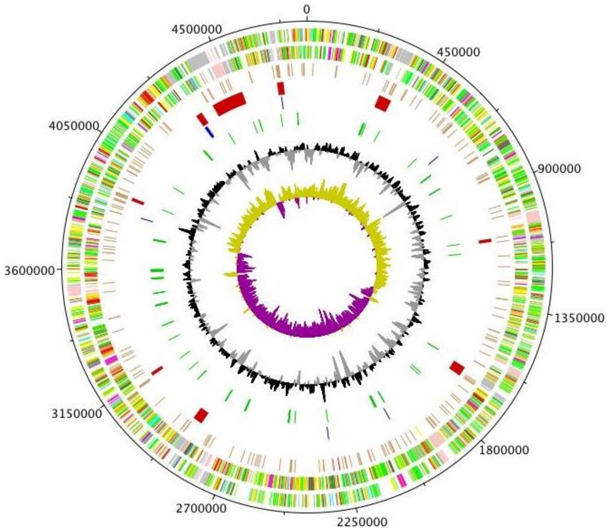
- CGView Server - is a comparative genomics tool for circular genomes that allows sequence feature information to be visualized in the context of sequence analysis results. A genome sequence is supplied to the program in FASTA, GenBank, EMBL or raw format. Up to three comparison sequences (or sequence sets) in FASTA format can also be submitted. The CGView Server uses BLAST to compare the genome sequence to the comparison sequences, and then converts the results and any available feature information (from the GenBank, EMBL or optional GFF file) or analysis information (from an optional GFF file) into a high-quality graphical map showing the entire genome sequence, or a zoomed view of a region of interest. Several options are available for specifying how the BLAST comparisons are conducted, and for controlling how results are displayed.(Reference: Grant JR & Stothard P. 2008. Nucleic Acids Res. 36(Web Server issue): W181-184)

- Jena Prokaryotic Genome Viewer (JPGV) - from a GenBank flatfile (*.gbk) generates linear or circular plots; including if desired GC content, GC skew, purine excess and keto excess can be displayed. Also allows BLAST analysis against related genomes. Requires free registration.

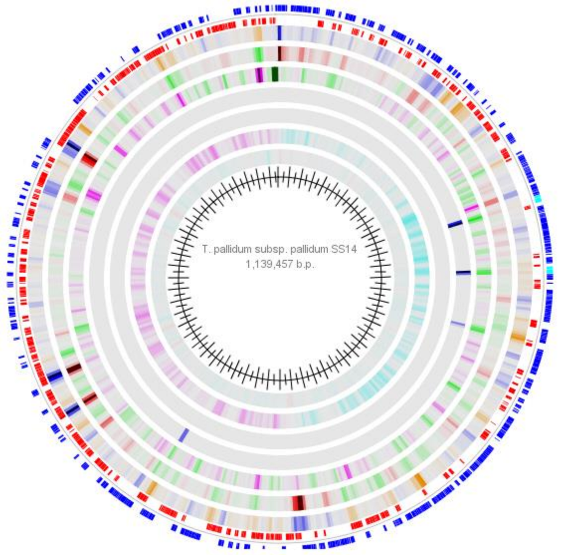
- GenomeVx - makes editable, publication-quality, maps of mitochondrial and chloroplast genomes and of large plasmids. These maps show the location of genes and chromosomal features as well as a position scale. The program takes as input either raw feature positions or GenBank records. In the latter case, features are automatically extracted and colored, an example of which is given. Output is in the Adobe Portable Document Format (PDF) and can be edited by programs such as Adobe Illustrator.(Reference: G. Conant & K. Woolfe. 2008. Bioinformatics 24:861-862).

- myGenomeBrowser - is a web-based environment that provides biologists with a way to build, query and share their genome browsers. This tool, that builds on JBrowse, is designed to give users more autonomy while simplifying and minimizing intervention from system administrators. They have extended genome browser basic features to allow users to query, analyze and share their data. (Reference: S. Carrere & J. Gouzy. Bioinformatics (2017) 33 (8): 1255-1257).

- DNAPlotter - is an interactive Java application for generating circular and linear representations of genomes. Making use of the Artemis libraries to provide a user-friendly method of loading in sequence files (EMBL, GenBank, GFF) as well as data from relational databases, it filters features of interest to display on separate user-definable tracks. It can be used to produce publication quality images for papers or web pages.(Reference: Carver, T. et al. 2008. Bioinformatics 25:119-120)



- GeneWiz (*Center for Biological Sequence Analysis, Danish Technical University*) produces linear or circular genome altases such as the one below. They have ready name ones for most bacteria, but by uploading custom data in GenBank format (.gbk) one can make one's own diagram showing the genetic and physical properties of your genome.



- OrganellarGenomeDRAW - is a suite of software tools that enable users to create high-quality visualrepresentations of both circular and linear annotated genome sequences provided as GenBank files oraccession numbers. Although all types of DNA sequences are accepted as input, the software has beenspecifically optimized to properly depict features of organellar genomes. A recent extension facilitates theplotting of quantitative gene expression data, such as transcript or protein abundance data, directly ontothe genome map (Reference: Lohse M, et al. 2013. Nucleic Acids Res. 41(Web Server issue):W575-81).

- PlasmaDNA - Starting with a primary DNA sequence, PlasmaDNA looks for restriction sites, open reading frames, primer annealing sequences, and various common domains. The databases are easily expandable by the user to fit his most common cloning needs. PlasmaDNA can manage and graphically represent multiple sequences at the same time, and keeps in memory the overhangs at the end of the sequences if any. This means that it is possible to virtually digest fragments, to add the digestion products to the project, and to ligate together fragments with compatible ends to generate the new sequences. Excellent package for plasmids. (Reference: Angers-Loustau A et al. 2007. BMC Mol Biol. 2007; 8:77).

- GSDraw (Gene Structure Draw Server) is a web server for gene family to draw gene structure schematic diagrams. Users can submit genomic, CDS and transcript sequences. GSDraw uses this information to obtain the gene structure, protien motif and phylogenetics tree, then draw diagram for it. (Reference: Wang Y, et al. 2013. Nucleic Acids Res. 41(Database issue):D1159-66).

- GECA is a user-friendly tool for representing gene exon/intron organization and highlighting changes in gene structure among members of a gene family. It relies on protein alignment, completed with the identification of common introns in the corresponding genes using CIWOG. GECA produces a main graphical representation showing the resulting aligned set of gene structures, where exons are to scale. The important and original feature of GECA is that it combines these gene structures with a symbolic display highlighting sequence similarity between subsequent genes. It is worth noting that this combination of gene structure with the indications of similarities between related genes allows rapid identification of possible events of gain or loss of introns, or points to erroneous structural annotations. The output image is generated in a portable network graphics format which can be used for scientific publications.

(Reference: Fawal N, et al. 2012. Bioinformatics; 28:1398-9).

● **Synthetic genes:**

● GeneDesign - is an excellent resource for designing synthetic genes. It includes tools for codon optimization and removal of restriction sites (Reference: Richarson, S.M. et al. 2006. Genome Research 16:550-556)

● **Metagenomics:**

● Orphelia - Orphelia is a metagenomic ORF finding tool for the prediction of protein coding genes in short, environmental DNA sequences with unknown phylogenetic origin. Orphelia is based on a two-stage machine learning approach that was recently introduced by our group. After the initial extraction of ORFs, linear discriminants are used to extract features from those ORFs. Subsequently, an artificial neural network combines the features and computes a gene probability for each ORF in a fragment. A greedy strategy computes a likely combination of high scoring ORFs with an overlap constraint. (Reference: K.J. Hoff et al. 2009. Nucl. Acids Res. 37(Web Server issue:W101-W105).

● WebMGA is a customizable web server for fast metagenomic analysis which includes over 20 commonly used tools for analyses such as ORF calling, sequence clustering, quality control of raw reads, removal of sequencing artifacts and contaminations, taxonomic analysis, functional annotation etc. All the tools behind WebMGA were implemented to run in parallel on our local computer cluster. (Reference: Wu S, et al. 2011. BMC Genomics. 12:444).

● MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data. The server primarily provides upload, quality control, automated annotation and analysis for prokaryotic metagenomic shotgun samples. (Reference: Wilke A, et al. 2016. Nucleic Acids Res. 44(D1):D590-4).

● MetaBin Comprehensive Taxonomic Assignment of Metagenomic Sequences (Laboratory for Integrated Bioinformatics, RIKEN, Japan) web server and standalone program allow faster and more accurate taxonomic assignment of single and paired-end sequence reads of varying lengths (≥45 bp) obtained from both Sanger and next-generation sequencing platforms. Has a tutorial.

● AmphoraNet - uses 31 bacterial and 104 archaeal protein coding marker genes for metagenomic and genomic phylotyping. Most of these are single copy genes, therefore AmphoraNet is suitable for estimating the taxonomic composition of bacterial and archaeal communities from metagenomic shotgun sequencing data. (Reference: Kerepesi C, et al. 2014. Gene. 533:538-40).

● METAGENassist - allows users to take bacterial census data from different environment sites or different biological hosts, and perform comprehensive multivariate statistical analyses on the data. These multivariate analyses can be done using either taxonomic or automatically generated phenotypic labels and visualized using a variety of high quality graphical tools. The bacterial census data can be derived from 16S rRNA data, NextGen shotgun sequencing or even classical microbial culturing techniques. Includes a tutorial. (Reference: Arndt D, et al. 2012. Nucleic Acids Res. 40(Web Server issue):W88-95).

- Real Time Metagenomics *(Dr. Robert Edwards, San Diego State University, USA)* - is the next revolution in metagenome annotation: Real time data processing and analysis. You can finally annotate a metagenome in real time, with no waiting. You can upload your own data for analysis. They accept either fasta or fastq files, and you can provide zip or gzip compressed data.

- EBI Metagenomics *(EMBL-EBI)* - is an automated pipeline for the analysis and archiving of metagenomic data that aims to provide insights into the phylogenetic diversity as well as the functional and metabolic potential of a sample. You can freely browse all the public data in the repository. The service identifies rRNA sequences, using rRNASelector, and performs taxonomic analysis upon 16S rRNAs using Qiime. The remaining reads are submitted for functional analysis of predicted protein coding sequences using the InterPro sequence analysis resource. InterPro uses diagnostic models to classify sequences into families and to predict the presence of functionally important domains and sites. By utilising this resource, the service offers a powerful and sophisticated alternative to BLAST-based functional metagenomic analyses. Data submitted to the EBI Metagenomics service is automatically archived in the European Nucleotide Archive (ENA). Accession numbers are supplied for sequence data.

- Kaiju - is a fast and sensitive taxonomic classification for metagenomics which takes nucleotide sequences in compressed FASTA or FASTQ format. Reads are directly assigned to taxa using the NCBI taxonomy and a reference database of protein sequences from bacterial, archaeal and viral genomes. By default, Kaiju uses either the available complete genomes from NCBI RefSeq or the microbial subset of the non-redundant protein database nr used by NCBI BLAST.Kaiju translates reads into amino acid sequences, which are then searched in the database using a modified backward search on a memory-efficient implementation of the Burrows-Wheeler transform, which finds maximum exact matches (MEMs), optionally allowing mismatches in the protein alignment. (Reference: Menzel P et al. 2016. (Nat. Commun. 7:11257)

- PhyloPythiaS - is a fast and accurate sequence composition-based classifier that utilizes the hierarchical relationships between clades. Taxonomic assignments with the web server can be made with a generic model, or with sample-specific models that users can specify and create. Several interactive visualization modes and multiple download formats allow quick and convenient analysis and downstream processing of taxonomic assignments. (Reference: Patil KR, et al. 2012. PLoS One. 7:e38581).

- Virtual Metagenome - A web server to reconstruct metagenomes from 16S rRNA sequences. a novel method for the rapid and efficient reconstruction of a virtual metagenome in environmental microbial communities without using large-scale genomic sequencing. We demonstrate this approach using 16S rRNA gene sequences obtained from denaturing gradient gel electrophoresis analysis, mapped to fully sequenced genomes, to reconstruct virtual metagenome-like organizations. (Reference: Okuda S, et al. 2012. Nat Commun. 3:1203.)

- MetaPhlAn2 (version 2.0.0) - is a computational tool for profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes and Viruses) from metagenomic shotgun sequencing data with species level resolution. It is also able to identify specific strains and to track strains across samples for all species. It allows for unambiguous taxonomic assignments, accurate estimation of organismal relative abundance, and species-level resolution for bacteria, archaea, eukaryotes and viruses. (Reference: Segata N, et al. 2012. Nature Methods 8: 811–814).

- CoMet-Universe — a web-server for comparative analysis of metagenomes based on protein domain signatures. Starting with an upload of your DNA sequences the CoMet pipeline performs all necessary steps for a comprehensive metagenome analysis including gene prediction, protein domain detection using Pfam 27, metabolic profiling based on KEGG pathways and taxon abundance estimation across all domains of life and viruses. (Reference: Aßhauer KP et al. Int J Mol Sci. 2014; 15(7):12364-78).
- 16S Classifier - is a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. On real metagenomic datasets, it showed up to 99.7% accuracy at the phylum level and up to 99.0% accuracy at the genus level. (Reference: N. Chaudhary et al. 2015. PLoS One 10(2): e0116106). It can also be accessed here

- **Meta sites:**

- DNAATLAS *(DNA2.0 Inc., U.S.A.)* - A place for all your sequences. Easily import all your constructs including Genbank, Gene Designer, Excel, Word, and nearly any text-based format. DNA Atlas immediately parses your upload files and infers whether each sequence is a feature, construct, primer, DNA or amino acid. Upload features and primers to see them annotated in your sequences. Instantly view constructs annotated with our curated list of over 1000 features, or add your own. Use the BLAST-based sequence search to quickly align and compare your sequences.Keep track of your sequences, features, and primers. Categorize them using tags - from freezer locations to characterization data. (requires registration).

- SuperPhy *(Chad Laing & Vic Gannon, Public Health Agency of Canada)* is an online tool for the predictive genomics of *Escherichia coli.* The platform integrates the analyses tools and genome sequence data for all publicly available *E. coli* genomes and facilitates the upload of new genome sequences from users under public or private settings. SuperPhy provides real-time analyses of thousands of genome sequences based on strain metadata, including geospatial and phylogenetic context.

- Naming your bacteriophage: This is of prime importance for members of the bacterial virus community to name their newly isolated phages appropriately. A good place to start is "How to Name and Classify Your Phage: An Informal Guide." (Reference: Adriaenssens E & Brister JR. 2017. Viruses 9(4). pii: E70) to which I will add the following points (a) please check that the name you propose has not been used already; and, (b) Do not name your phage Enterobacteria phage ø1234 or Enterobacteria phage 2017/ABC_567 since these names are incompatable with the creation of new species and genera taxa by the International Committee on Taxonomy of Viruses (ICTV). To find if your proposed name is unique consult:

- Phage Name Check *(Stephen T. Abedon, Ohio State University, USA)* - to see whether 'your' phage name is currently found on Google Scholar, Google Books, PubMed, or even Bacteriophage Names 2000.

- CPT Phage Name Search *(Center for Phage Technology at Texas A&M University)*

APPLICATIONS OF GENOME ANALYSIS AND GENOMICS.

The clinical applications of genomic technologies

The clinical applications of genomic technologies are vast and offer opportunities to improve healthcare across the breadth of medical specialities.

1. Gene discovery and diagnosis of rare monogenic disorders Genomic technologies can be used by clinicians from all specialities to diagnose their patients who have high-risk genetic errors causing disease. Researchers are using these techniques to identify new genes which cause genetic disease at an astonishing rate – over 4000 diseases now have a known single genetic cause, compared to around 50 in 1990.

2. Identification and diagnosis of genetic factors contributing to common disease Genomic technologies are increasingly being used to understand the contribution of both rare and common genetic factors to the development of common diseases, such as high blood pressure, diabetes and cancer.

3. Pharmacogenetics and targeted therapy Genetic information may be used to predict whether a person will respond to a particular drug, how well they will respond to that drug and whether they are likely to get any side effects from the use of a specific drug. This allows their treating team to make individualised decisions about the right drug treatment. In some cases, such as cancer, we can identify the genetic drivers of disease and then give drugs which specifically target that pathway. This is known as targeted therapy.

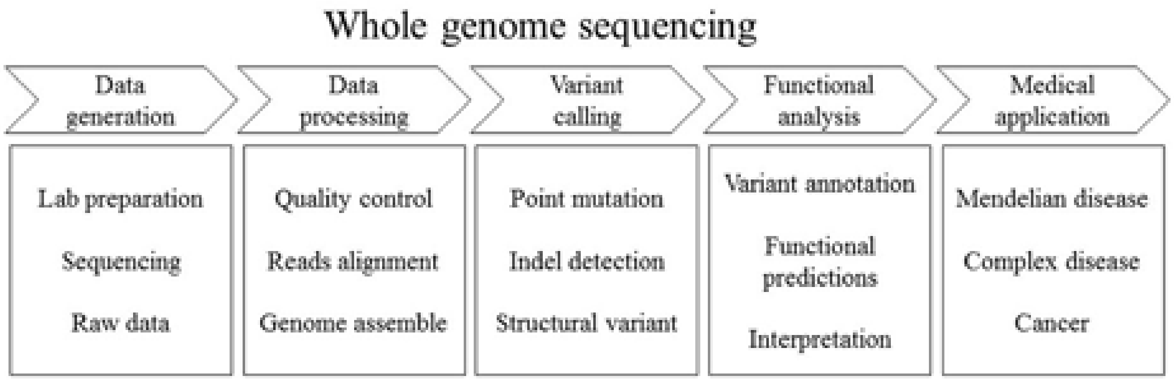
4. Prenatal diagnosis and testing Genetic diseases are often devastating and may cause significant disability and even death in childhood. Prenatal diagnosis of genetic diseases allows parents to make decisions about whether to continue with the pregnancy or to allow early diagnosis and possible treatment in utero or at birth. Whilst previous approaches to prenatal diagnosis could put the pregnancy at risk, new methods using genomic technology can look directly at the DNA of the fetus from a maternal blood test, without increasing the risk of miscarriage – this is known as non-invasive prenatal testing. The use of NGS and array technology in prenatal samples is also on the increase to improve diagnostic yields in a pregnancy.

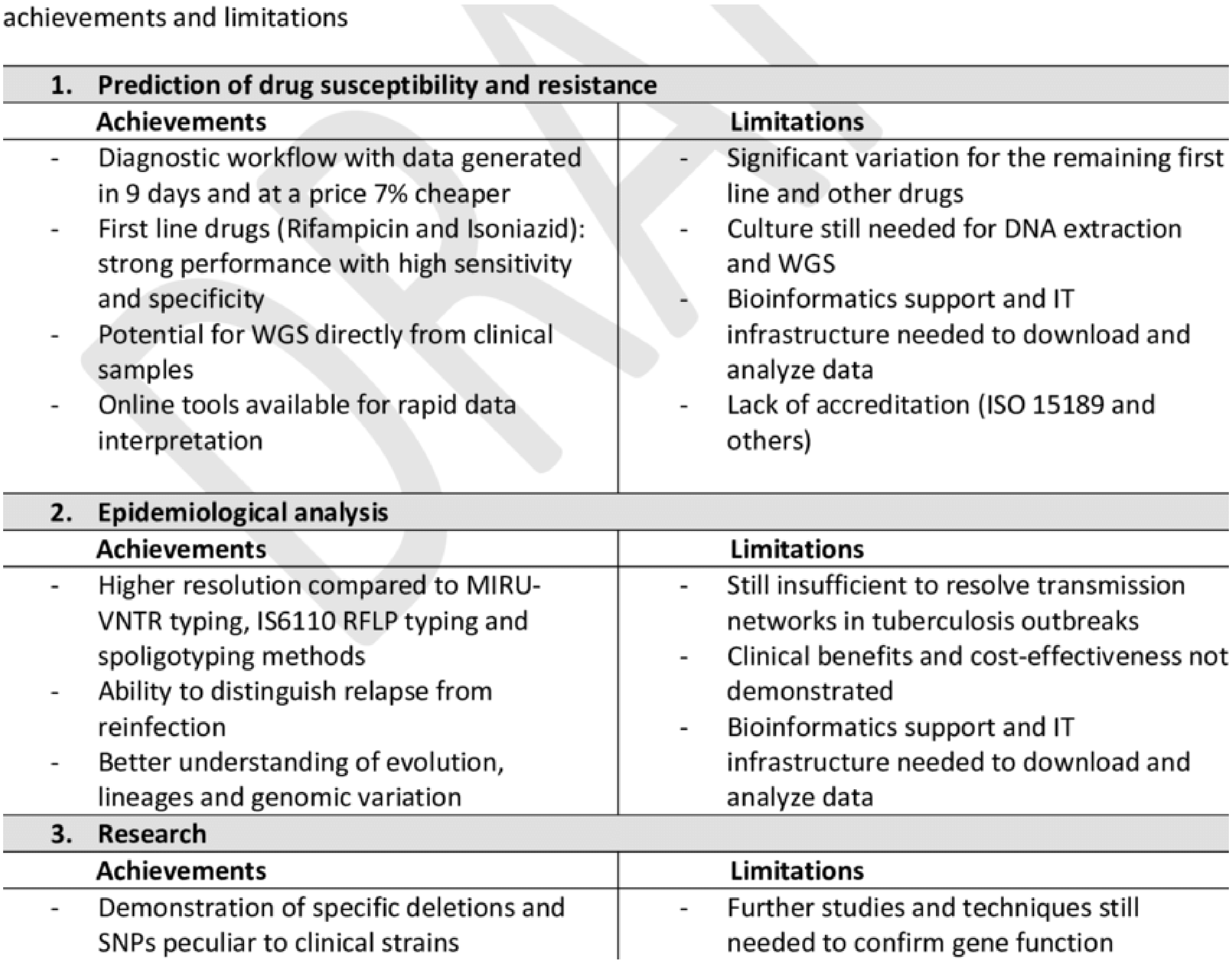
5. Infectious diseases Sequencing the genomes of microorganisms which cause human infection can identify the exact organism causing symptoms, help to trace the cause of infectious outbreaks, and give information as to which antibiotics are most likely to be effective in treatment.

6. Personalised medicine As the exact DNA sequence of the genome of each human is unique to them, we will all have unique disease susceptibilities and treatment responses. Personalised medicine describes the use of our genetic information to tailor health care intervention to our own individual need.

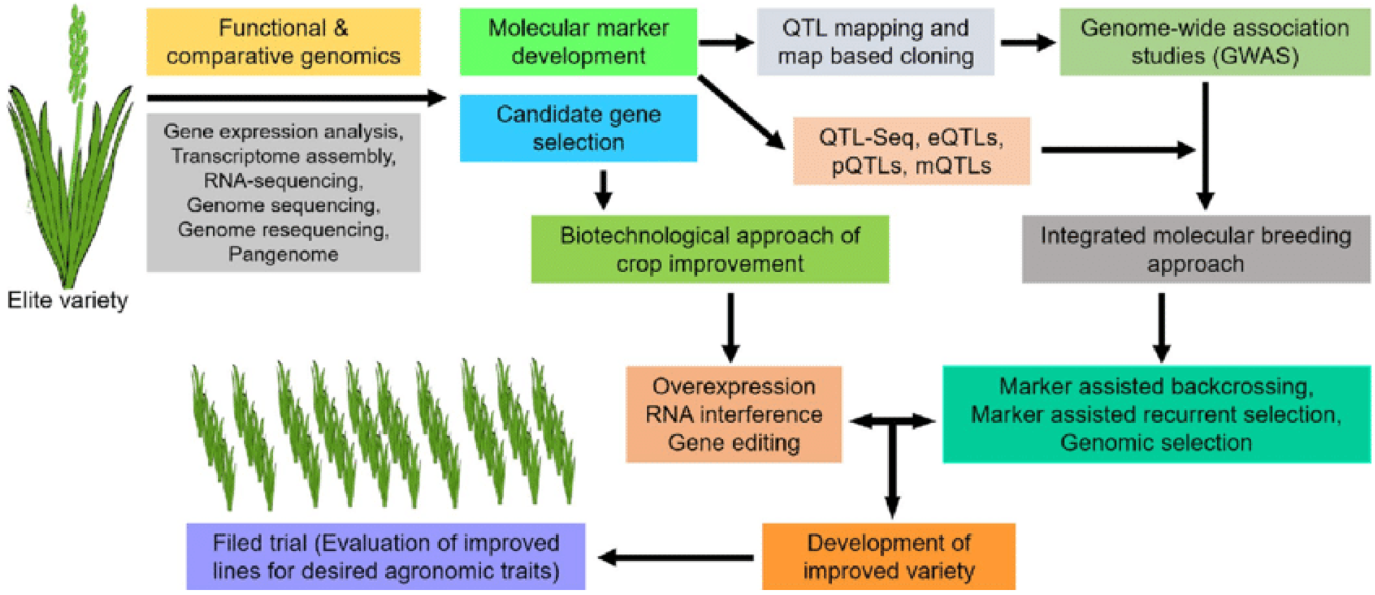
7. Gene therapy Gene therapy involves the administration of DNA or RNA, in order to correct a genetic abnormality, or modify the expression of genes.

8. Genome editing Genome editing uses molecular techniques to modify the genome – genome editing can add in, cut out, or replace sections of the DNA sequence.




achievements and limitations

1.	**Prediction of drug susceptibility and resistance**	
	Achievements	**Limitations**
	- Diagnostic workflow with data generated in 9 days and at a price 7% cheaper - First line drugs (Rifampicin and Isoniazid): strong performance with high sensitivity and specificity - Potential for WGS directly from clinical samples - Online tools available for rapid data interpretation	- Significant variation for the remaining first line and other drugs - Culture still needed for DNA extraction and WGS - Bioinformatics support and IT infrastructure needed to download and analyze data - Lack of accreditation (ISO 15189 and others)
2.	**Epidemiological analysis**	
	Achievements	**Limitations**
	- Higher resolution compared to MIRU-VNTR typing, IS6110 RFLP typing and spoligotyping methods - Ability to distinguish relapse from reinfection - Better understanding of evolution, lineages and genomic variation	- Still insufficient to resolve transmission networks in tuberculosis outbreaks - Clinical benefits and cost-effectiveness not demonstrated - Bioinformatics support and IT infrastructure needed to download and analyze data
3.	**Research**	
	Achievements	**Limitations**
	- Demonstration of specific deletions and SNPs peculiar to clinical strains	- Further studies and techniques still needed to confirm gene function


Functional & comparative genomics
Molecular marker development
QTL mapping and map based cloning
Genome-wide association studies (GWAS)
Gene expression analysis, Transcriptome assembly, RNA-sequencing, Genome sequencing, Genome resequencing, Pangenome
Candidate gene selection
QTL-Seq, eQTLs, pQTLs, mQTLs
Elite variety
Biotechnological approach of crop improvement
Integrated molecular breeding approach
Overexpression RNA interference Gene editing
Marker assisted backcrossing, Marker assisted recurrent selection, Genomic selection
Filed trial (Evaluation of improved lines for desired agronomic traits)
Development of improved variety

References

1. http://library.umac.mo/ebooks/b28050393.pdf
2. http://www.aun.edu.eg/molecular_biology/Proceeding_Dec2011/DNA%20sequencing.pdf
3. http://www.bio.miami.edu/dana/dox/restrictionenzymes.html
4. http://www.dspmuranchi.ac.in/pdf/Blog/SCREENING%20OF%20DNA%20LIBRARIES.pdf
5. http://www.indiastudychannel.com/resources/155090-The-principles-techniques-application-DNA-fingerprinting.aspx
6. http://www.pathologyoutlines.com/topic/molecularpathdnaseqmaxam.html
7. http://www.yourarticlelibrary.com/dna/dna-fingerprinting-principles-and-techniques-of-dna-fingerprinting/12211.
8. https://24hoursofbiology.com/human-genome-project/
9. https://ab.inf.uni-tuebingen.de/teaching/ws09/bioinformatics-i/10-sequencing.pdf
10. https://bio.libretexts.org/Bookshelves/Biochemistry/Supplemental_Modules_(Biochemistry)/1%3A_DNA/1.4%3A_DNA_Modifying_Enzymes
11. https://en.wikipedia.org/wiki/Restriction_enzyme
12. https://getrevising.co.uk/grids/the_human_genome_project_2
13. https://ghr.nlm.nih.gov/primer/hgp/accomplishments
14. https://international.neb.com/products/restriction-endonucleases/restriction-endonucleases
15. https://microbenotes.com/dna-microarray/

16. https://nptel.ac.in/courses/102103013/module1/lec4/11.html

17. https://nptel.ac.in/courses/102103017/pdf/lecture%2038.pdf

18. https://web.wpi.edu/Pubs/E-project/Available/E-project-011306-130417/unrestricted/IQP.pdf

19. https://www.biologydiscussion.com/viruses/animal-viruses/togaviruses-structure-and-replication-microbiology/65689

20. https://www.biologyexams4u.com/2014/05/dna-fingerprinting-procedure.html#.W4QNuPkzbIU

21. https://www.britannica.com/event/Human-Genome-Project

22. https://www.britannica.com/science/restriction-enzyme

23. https://www.britannica.com/science/yeast-artificial-chromosome

24. https://www.caister.com/genomeanalysis

25. https://www.differencebetween.com/difference-between-chromosomewalking-and-vs-jumping/

26. https://www.encyclopedia.com/science-and-technology/biology-and-genetics/cell-biology/yeast-artificial-chromosome-yac

27. https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/

28. https://www.nature.com/scitable/topicpage/dna-sequencing-technologies-key-to-the-human-828

29. https://www.ncbi.nlm.nih.gov/pubmed/7859160

30. https://www.ncbi.nlm.nih.gov/pubmed/9291964

31. https://www.sciencedirect.com/topics/neuroscience/yeast-artificial-chromosome

32. https://www.slideshare.net/gurya87/yeast-artificial-chromosomes-yacs-44970900

Assessment:

Brief the following:

1. Insulin.
2. Human growth hormone.
3. Vaccine.

Detail the following:

4. Cultivation of GEMOs.
5. Genetically engineered microorganisms.