RECOMBINANT DNA TECHNOLOGY AND BIOTECHNOLOGY

Cloning Strategies and Screening of Recombinant DNA Clones

Santosh Dhillon
Department of Biotechnology & Molecular Biology
A.K. Chhabra
Department of Plant Breeding
and
Pushpa Kharb
Department of Biotechnology & Molecular Biology
CCS Haryana Agricultural University
Hisar

CONTENTS
Gene cloning
Definition
Purpose
Approaches for gene cloning
Cell-based approach
Polymerase chain reaction (PCR)
Cell based cloning vs PCR
Methods for isolation of gene/DNA fragment of interest
Genomic libraries
Complementary (cDNA) libraries
Chemical synthesis
Transposon tagging
Map based cloning
PCR
Screening of transformants for presence of desired gene/gene product
Direct selection for the desired gene
Identification of the clone from a gene library
Some important techniques used in characterization of genes/gene products
Blotting and hybridization techniques
DNA fingerprinting
Restriction fragment length polymorphism (RFLP)
PCR based techniques
DNA chips and microarrays

Keywords
Gene cloning; Recombinant DNA technology tools; Cell base gene cloning; Polymerase chain reaction (PCR); Gene isolation; DNA fragment isolation; Genomic libraries; Complimentary DNA libraries; DNA fingerprinting; DNA chips; DNA microarrays; Gene characterization.
Gene cloning

Definition
A clone is a group of genetically identical cells derived from a single cell. In literal sense, gene cloning means to make many exact copies of the gene. A gene is a stretch of DNA which codes for a function e.g. a gene for rRNA or a protein such as insulin. It includes the coding sequence and its regulatory elements i.e. promoter, operator etc. Gene cloning can be achieved by two different approaches: (1) cell based and (2) polymerase chain reaction (PCR).

Purpose
Gene cloning has three main purposes or objectives:

i) To obtain a large number of copies of a desired gene or segment of DNA

In order to study genes in the laboratory, it is necessary to have sufficient quantity of DNA for analysis such as determination of its sequence i.e. the determination of the precise order of all the base pairs in the gene. Sequence determination allows us to study structure, function and regulation of the gene.

ii) To obtain protein product of the gene in large quantities

Cloned genes also make it easier to study the proteins they encode. Cloning a gene which encodes a desired protein may allow that gene to be over-expressed so that the protein can be produced in bulk. Many important pharmaceutical products are produced from cloned genes e.g. insulin, interferon, clotting factors, human growth hormone, cytokines (cell growth stimulants) and several anticancer drugs. Similarly, gene cloning is also used to produce enzymes used for recombinant DNA research such as restriction endonucleases, DNA ligase, DNA polymerases etc. and many enzymes of industrial importance such as proteases, amylases, lipases etc. The proteins or enzymes can also be genetically engineered to alter their properties.

iii) To obtain genetically modified organisms

Cloning allows engineering animals, plants and microbes. The ability to create genetically modified animals and plants (transgenics) has led to their use for research and for therapeutic and commercial purposes. The technology may lead to the development of new therapies for the treatment of diseases (gene therapy).

Two approaches for gene cloning

Cell-based approach

In the cell-based approach, a vector is required to carry the DNA fragment of interest into the host cell. A host cell is a cell in which the vector carrying the gene of interest can multiply. The DNA fragment or gene of interest is isolated from an organism such as a plant or an animal and ligated to a vector molecule to make a recombinant vector / DNA. Vector molecule is also known as cloning vehicle. Usually bacterial plasmids and viruses are used as vectors for cloning genes in bacteria. The recombinant vector carrying the DNA fragment is then introduced into the host. E. coli is the most commonly used host organism though other bacteria, plant, yeast or animal cells can also be used. As the host cell multiplies, the recombinant vector carrying the
desired DNA fragment or gene also multiplies. This process is also known as Recombinant DNA Technology (RDT).

**Tools of recombinant DNA technology**

Following basic tools are required for RDT:
1. Restriction endonucleases and other enzymes
2. Vector
3. Host Cell

**Restriction endonucleases and other enzymes**

*What are restriction endonucleases?*

Restriction endonucleases (REs) also called restriction enzymes were identified independently by two groups, Arber and Linn (1968) and Meselson and Yuan (1968). Later, their importance was realized in 1970 when Hamilton Smith discovered type II restriction endonucleases. The *in vivo* function of restriction enzymes is to protect bacteria from invasion by viruses. There are three types of restriction endonucleases namely, type I, type II and type III. Out of these only types II or class II enzymes are important for gene cloning. These are the enzymes that recognize specific sequences in double stranded DNA and cleave within these sequences i.e. they are very specific in their action.

The nomenclature of RE was given by Smith and Nathans (1973). According to this nomenclature, each enzyme is given four letter name derived from the name of the bacterium from which it was isolated e.g. *EcoR*I

Most of the type II restriction enzymes recognize 4 to 6 base pair long palindromic (have two folded rotational symmetry) sequences and cleave within or near to these sequences.

\[
E \ Escherichia \quad (\text{genus}) \\
co \ coli \quad (\text{species}) \\
R \ RY13 \quad (\text{strain}) \\
I \quad \text{First identified order in bacterium}
\]

Majority of the type II enzymes make cleavage in such a way that the fragments produced have protruding ends or cohesive tails that are complementary to each other. For example, the enzyme *EcoR*I recognizes a 6 bp long sequence GAATTTC and cuts at a place shown by arrow so that fragments produced have 5’ overhangs or tails as shown in fig. These ends can be easily joined together with the help of another enzyme called DNA ligase. Any two fragments, regardless of their origin (animal, plant, fungal, bacterial) can be joined *in vitro* to form recombinant molecules, if they are obtained by cutting with the same restriction endonuclease. This is in fact the basis of recombinant DNA technology or gene cloning.

There are more than 1000 type II restriction enzymes known so far. Some of the commonly used restriction enzymes along with their recognition sequences and cleavage sites are shown in Table 1. Depending upon their recognition sequences, they can be tetra-, penta- or hexa- cutters and may produce 5’ or 3’ overhangs or blunt ends.
Table 1: Recognition sequences and cleavage patterns of some restriction enzymes

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Source</th>
<th>Recognition Sequence</th>
<th>Cleavage pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>EcoRI</td>
<td>Escherichia coli</td>
<td>5'GAATTC 3'CTTAAG</td>
<td>5'---G AATTC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CTTAA G---5'</td>
</tr>
<tr>
<td>BamHI</td>
<td>Bacillus amyloliquefaciens</td>
<td>5'GGATCC 3'CCTAGG</td>
<td>5'---G GATCC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CCTAG G---5'</td>
</tr>
<tr>
<td>HindIII</td>
<td>Haemophilus influenzae</td>
<td>5'AAGCTT 3'TTCGAA</td>
<td>5'---A AGCTT---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---TTCGA A---5'</td>
</tr>
<tr>
<td>TaqI</td>
<td>Thermus aquaticus</td>
<td>5'TCGA 3'AGCT</td>
<td>5'---T CGA---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---AGC T---5'</td>
</tr>
<tr>
<td>NotI</td>
<td>Nocardia otitidis</td>
<td>5'GCGGCCGC 3'CGCCGGCG</td>
<td>5'---GC GGCCGC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CGCCGG CG---5'</td>
</tr>
<tr>
<td>HinfII</td>
<td>Haemophilus influenzae</td>
<td>5'GANTC 3'CTNAG</td>
<td>5'---G ANTC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CTNA G---5'</td>
</tr>
<tr>
<td>Sau3A</td>
<td>Staphylococcus aureus</td>
<td>5'GATC 3'CTAG</td>
<td>5'--- G GATC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CTAG G---3'</td>
</tr>
<tr>
<td>HaeIII</td>
<td>Haemophilus egytius</td>
<td>5'GGCC 3'CCGG</td>
<td>5'---GG CC---3'</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3'---CC GG---5'</td>
</tr>
</tbody>
</table>

Other Enzymes used in RDT:

i) DNA ligase is used for joining DNA molecules.
ii) Alkaline phosphatase is used for dephosphorylation of the vector i.e. removal of 5’ phosphate to avoid recircularization of the cut vector.
iii) S1 nuclease is used for cutting single stranded nucleic acids.
iv) Terminal transferase is used for adding homopolymer tails.
v) Reverse transcriptase is used for cDNA synthesis.
vi) DNA polymerases are used for DNA replication.
vii) RNase H is used for removal of RNA from RNA/DNA hybrid.

Vectors

Vector is an agent that can carry a DNA fragment into a host cell in which it is capable of replication. If it is used only for reproducing the DNA fragment, it is called a cloning vector. If it is used for expression of foreign gene, it is called an expression vector. Properties of a good vector:

(1) It should be autonomously replicating i.e. it should have ori region.
(2) It should contain at least one selectable marker e.g. gene for antibiotic resistance.
(3) It should have unique restriction enzyme site (only one site for one RE) for different REs to insert foreign DNA.
(4) It should be preferably small in size for easy handling.
(5) It should have relaxed control of replication so that multiple copies can be obtained.

Vectors are of different types depending on the host. These are as follows:
1. Bacterial vectors
2. Yeast vectors
3. Plant vectors
4. Animal vectors

**Bacterial vectors**

*E. coli* is the most commonly used bacterium for gene cloning though other bacteria such as *Bacillus* are also used. Vectors for cloning in these bacteria are described below:

**Vectors for cloning in E.coli**

A number of vectors are used for cloning in *E. coli*. These are categorized as plasmids, phages, cosmids, phagemids and bacterial artificial chromosomes.

i) **Plasmid vectors**

Plasmids are autonomously replicating circular, double stranded DNA molecules found in bacteria. They have their own origin of replication (*ori* region), and can replicate independently of the host chromosome. The size of plasmids ranges from a few kb to 200 kb. Plasmid vectors are often used for cloning DNA segments of small size (upto 10 kilobases). Some of the commonly used plasmid vectors are described below:

**pBR322**

The first plasmid vector that has been constructed artificially is pBR322. It is named after the scientists Bolivar and Rodriguiz who constructed it in 1977. It is 4362bp in size. It has an origin of replication derived from a colicin-resistance plasmid (ColE1). This origin allows a fairly high copy number, about 100 copies of the plasmid per cell. Plasmid pBR322 carries two selectable markers viz. genes for resistance to ampicillin (Ap<sup>r</sup>) and tetracycline (Tc<sup>r</sup>). Several unique RE sites are present within these genes for insertion of foreign DNA (Fig 1). When a foreign DNA segment is inserted in any of these genes, the antibiotic resistance by that particular gene is lost. This is called insertional inactivation. For instance, insertion of a restriction fragment in the SalI site of the Tc<sup>r</sup> gene inactivates that gene. One can still select for Ap<sup>r</sup> colonies, and then screen to see which ones have lost Tc<sup>r</sup>.

**pUC**

A series of small plasmids (about 2.7 kb) have been developed at the University of California and hence the name pUC e.g. pUC7, 8, 18 and 19 etc. (Fig 2). These are high copy number plasmids that carry an ampicillin resistance gene and an origin of replication, both from pBR322. They also have a multiple cloning site (MCS) – a sequence of DNA that carries unique sites for many REs. The MCS contains a portion of lacZ gene that codes for the enzyme β-galactosidase. When such plasmids are introduced into *E. coli*, the colonies are blue on plates containing X-gal (substrate for β- galactosidase) and IPTG (isopropyl thiogalactoside, an inducer). When a foreign
DNA is introduced in MCS, the β-galactosidase activity is lost. Thus cells containing recombinant plasmids form white (not blue) colonies.

![Plasmid Vector](image)

### ii) Phage vectors

Bacteriophages or phages are viruses that specifically infect bacteria. The phage particle attaches to the outer surface of bacterium and injects its DNA into the cell. The phage DNA is then replicated inside the host and its genes are expressed to make phage capsid proteins and new phage particles are assembled and released from the bacterium.

Phage vectors can accommodate more DNA (upto 25 kb) than plasmids and are often used for preparation of genomic libraries. They also have higher transformation efficiency as compared to plasmids.

Two bacteriophages namely, Lambda (λ) and M13 have been commonly used for construction of vectors for cloning in *E. coli*.

**Lambda (λ) phage vectors**

Lambda is a temperate bacteriophage with a genome size of 48.5 kb. Its entire DNA sequence is known. The lambda genome is a linear, double-stranded molecule with single-stranded, complementary ends. These ends can hybridize with each other (and do so when the DNA is within an infected cell) and are thus termed cohesive (cos) sites. The phage can have two modes of life cycles i.e. lytic and lysogenic. During lytic cycle, it replicates independently in the host cell and produces a large number of phage particles which are released by lysis of the host. Alternatively, it can take up lysogenic growth, meaning that it integrates its DNA into the bacterial chromosome and multiplies along with it.
Two types of vectors have been constructed from lambda phage. Theses are insertional and replacement vectors (Fig 3). Insertional vectors have one unique restriction site for a particular restriction enzyme and can accommodate 6-7 kb DNA. Examples of insertional vectors are λgt10, λgt11 and λZAP II. On the other hand, replacement vectors have two cleavage sites for a restriction enzyme and can accommodate up to 25 kb DNA. When vector is cut with a restriction endonuclease, a stuffer fragment is removed and replaced with a foreign DNA. Some examples of replacement vectors are EMBL3, EMBL3A, EMBL4, λDASH, λFIX, GEM11 and GEM12.

Bacteriophage lambda can be reconstituted in a test tube by simply mixing phage DNA with a mixture of phage proteins, an infective viral particle with the DNA inside the phage head can be produced. This process is called in vitro packaging. There is a strict size requirement for the piece of DNA that goes into the phage head. That is, it should not be more than 52 kb and less than 38 kb. This feature allows only the recombinants to be packaged inside the phage head. In addition, some lambda phage vectors have a stuffer fragment that carries the β-galactosidase gene. When it is removed or when foreign DNA is cloned within the gene, β-galactosidase activity may be abolished. The accompanying loss of activity may be used to select recombinant clones.

**M13 Phage vectors**

M13 is a filamentous bacteriophage of *E. coli* and contains a single stranded circular DNA of 7.2 kb. A series of vectors (M13 mp series) have been developed from this phage. These vectors have a polylinker with unique restriction enzyme sites in lac Z gene that complements host (e.g. JM 103 or JM 104). Screening of recombinants is done based on formation of blue/white plaques. M13 vectors are used for obtaining sufficient quantity of DNA for sequencing by Sanger's dideoxy chain termination method.

**iii) Phagemids**

Phagemids are hybrid vectors derived from plasmids and phages e.g. pBluescript. They contain origin of replication from M13 or F1 phage and remaining features of the vector from plasmids. Infection of transformed bacteria (containing the phagemid) with a helper virus (e.g. derived from M13) will cause the M13 origin to be activated, and progeny viruses carrying single-stranded copies of the phagemid can be obtained.
M13 Phage vectors
M13 is a filamentous bacteriophage of E. coli and contains a single stranded circular DNA of 7.2 kb. A series of vectors (M13 mp series) have been developed from this phage. These vectors have a poly linker with unique restriction enzyme sites in lac Z gene that complements host (e.g. JM 103 or JM 104). Screening of recombinants is done based on formation of blue/white plaques. M13 vectors are used for obtaining sufficient quantity of DNA for sequencing by Sanger's dideoxy chain termination method.

ii) Phagemids
Phagemids are hybrid vectors derived from plasmids and phages e.g. pBluescript. They contain origin of replication from M13 or F1 phage and remaining features of the vector from plasmids. Infection of transformed bacteria (containing the phagemid) with a helper virus (e.g. derived from M13) will cause the M13 origin to be activated, and progeny viruses carrying single-stranded copies of the phagemid can be obtained.

iv) Cosmids
The cosmid vector is a combination of the plasmid and bacteriophage lambda. It is small (5-7 kb) circular DNA containing an origin for DNA replication (ori), selectable markers and restriction sites from plasmid plus a sequence from lambda needed for packaging the DNA (cos site). Cosmids may be used to clone large DNA molecules of up to 45 kb. They also have high transformation efficiency. Some examples of cosm id vectors include pJB, PWE and SuperCos series (Fig 4).

v) Bacterial artificial chromosomes (BAC)
BACs are based on bacterial mini-F plasmids, which are small pieces of episomal bacterial DNA that give the bacteria the ability to initiate conjugation with adjacent bacteria. They have a
cloning capacity of 75-300 kb. These have a lower copy number (like F) but are stable and relatively easy to work with. BACs have become one of the most frequently used vectors for preparation of genomic libraries.

**Vectors for cloning in Bacillus**

Most strains of *Bacillus* contain extra chromosomal DNA molecules called plasmids. However, plasmids present in *Bacillus* are cryptic which are devoid of any selectable markers. Ehrlich (1977) observed that plasmids of *Staphylococcus aureus* were able to replicate in *Bacillus subtilis*. So most of the plasmid vectors used for cloning in *B. subtilis* are derived from *S. aureus* plasmids. As none of the natural *S. aureus* plasmids carries more than one selectable marker, so improved vectors have been constructed by gene manipulation, e.g. PHV11. It has been derived from pC194 and carries TcR gene of pT127. Because of difficulties in direct cloning in *B. subtilis* hybrid plasmids called shuttle vectors (vectors that can multiply in two different hosts) have been constructed so that they can be used for cloning in *E. coli* as well as *B. subtilis*. *E. coli* replicon required for development of shuttle vectors has been obtained from pBR322. Additional antibiotic resistance genes or selectable markers derived from other *S. aureus* plasmids, chromosomal DNA of other strains of *Bacilli* or from pBR322 have also been incorporated. Some examples of shuttle vectors that can multiply both in *Bacillus* and *E. coli* are pHV 33, pLB5 and pPL603.

**Vectors for cloning in yeast**

The discovery of a 2μm plasmid in most strains of *Saccharomyces cerevisiae* led to the development of cloning vectors in yeast. The 2μm plasmid is 6 kb in size. It is present in 50-100 copies per cell. A number of shuttle vectors based on 2μm plasmid and bacterial plasmids have been constructed which can replicate either in *E. coli* or yeast. Yeast plasmid vectors are of four types, yeast episomal plasmids (YEps), yeast integrative plasmids (YIps) yeast replicative plasmids (YRps) and yeast centromeric plasmids (Ycps). In addition to plasmid vectors, yeast artificial chromosomes (YACs) are also used as vectors for cloning large pieces of DNA. A brief description of these vectors is given below:
i) Yeast episomal plasmids (YEps)
These are derived from 2µm plasmid. Some YEps contain the entire 2µm plasmid; others include just the 2µm origin of replication. An example of latter type is YEp13 (Fig 5). It is a shuttle vector and can be replicated both in *E.coli* and yeast. It contains 2µm origin of replication, yeast gene *leu2* as selectable marker and entire sequence of pBR322. The *leu2* gene codes for an enzyme involved in biosynthesis of amino acid leucine. YEps may replicate autonomously or integrate in one of the yeast chromosomes by homologous recombination. They have high transformation frequency of 10,000 to 100,000 transformants/ µg DNA.

![A SHUTTLE VECTOR](image)

**Fig. 5**

ii) Yeast integrative plasmids (YIps)
These are basically bacterial plasmids carrying a yeast gene. YIp5 is an example of yeast integrative plasmid. It has *ura3* gene inserted in pBR322. The gene *ura3* codes for an enzyme involved in biosynthesis of pyrimidine nucleotides and acts as selectable marker. The plasmid cannot replicate autonomously as it lacks 2µm origin of replication and survives by integrating in yeast chromosomal DNA. They have very low transformation frequency, less than 100 transformants/ µg DNA.

iii) Yeast replicative plasmids (YRps)
They carry a part of chromosomal DNA with an origin of replication and one or two selectable markers and are capable of independent replication. They have transformation frequency between 1000 and 10,000 transformants/ µg DNA.

iv) Yeast centromeric plasmids (YCps)
These are shuttle vectors that behave as small chromosomes and replicate only once during each cell division. They contain i) origin of replication called ARS sequence , ii) CEN sequence (for
proper segregation of chromosomes) and iii) a selectable marker such as \textit{leu2} from yeast and sequences from bacterial plasmid having \textit{ori} region and selectable marker (Ap'). They are stably maintained at one copy per cell.

\textit{v) Yeast Artificial Chromosomes (YACs)}

YACs are artificial chromosomes that replicate in yeast cells. Main features of these vectors are:

1. Autonomous replicating sequence (ARS) necessary for the replication in yeast cells (Fig 6).
2. Telomeres (TEL), which are ends of chromosomes involved in the replication and stability of chromosomes.
3. A yeast centromere (CEN), required for proper segregation of chromosomes
4. Selectable markers that allow the easy isolation of yeast cells that have taken up the artificial chromosome.
5. Unique RE sites.

YACs are capable of carrying a large DNA fragment (up to 3000 kb), but their transformation efficiency is very low.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{yeast_artificial_chromosome_vector.png}
\caption{Yeast Artificial Chromosome Vector}
\end{figure}

\textbf{Host Cell}

A good host should have the following properties:

1. It should be easy to transform.
2. It should not hinder replication of recombinant vector.
3. It should not have restriction and methylase activities.
4. It should be deficient in recombination function so that the introduced recombinant vector is not altered.
5. It should be easy to grow.
6. The recombinant vector should be easily retrieved from the transformed host.
A number of hosts have been used in RDT e.g. bacteria, yeast, plant and animal cells. *E.coli*, *B.subtilis* and yeast as hosts are described below:

**a. E. coli as a host**
*E.coli* is a rod shaped Gram negative bacterium. It is the host of choice for cloning experiments due to several reasons. It has a short doubling time (20 min.) and its genetics is well understood. It can be easily transformed and a number of plasmids are known that can multiply in *E.coli*. The *E.coli* strain K12 is the most commonly used; it has several sub strains such as HB 101, JM 103, C 600, DH5α etc. For most cloning applications, DH5α host cells are used. These cells are compatible with lacZ blue/white selection procedures, are easily transformed, and good quality plasmid DNA can be recovered from transformants.

**b. Bacillus subtilis as a host**
*Bacillus subtilis* is a Gram positive bacterium and is an important host organism for cloning genes though *E.coli* is the most commonly used host. *B. subtilis* has several advantages over *E.coli*. It is generally regarded as safe and has been used in industry for a long time for fermentation on large scale. In addition, it synthesizes a number of important enzymes such as amylases and proteases which are secreted into the growth medium. These and other proteins can thus be obtained directly from culture fluids and do not require isolation and purification from bacteria. This feature is very important for expression of heterologous proteins which are often degraded inside bacteria.

**c. Yeast as a host**
The yeast, *S. cerevisiae* is an important eukaryotic model organism for molecular genetic studies because its entire genome has been sequenced, and is used as a reference to human and other higher eukaryotic genes. It has been used in brewing and bread making and is considered as a safe organism for production of proteins to be used in medicine or food. It is easily genetically manipulated and is a popular organism to clone and express eukaryotic DNA.

**Steps of cell based gene cloning**
Cell based cloning process or recombinant DNA technology basically involves the following steps:-

a. Preparation of vector and insert DNA.
b. Construction of recombinant DNA (rDNA).
c. Transformation.
d. Selection of transformants.

The basic steps of a gene cloning experiment are diagrammatically shown in (Fig 7 At each of the above steps a suitable strategy is employed depending upon the objectives of the cloning experiment, types of vectors, host organism and available information about the gene (s) to be cloned (Fig 15). For example, a gene may be isolated either from a cDNA (complementary DNA) library/ genomic library or it may be chemically synthesized if the gene product is known.

**a. Preparation of vector and insert DNA**
This is usually done by cleaving both vector and foreign DNA with a suitable RE to generate complementary/compatible ends.
b. Construction of recombinant DNA (rDNA)

The joining of foreign DNA to vector is done with the help of enzyme DNA ligase. This step is also known as ligation. T4 phage encoded DNA ligase is most commonly used because of its high efficiency of ligation as well as ability to join even the blunt ends. This DNA ligase is obtained from *E.coli* that has been infected with T4 phage. The enzyme joins 5’ P and 3’ OH ends of DNA molecules and requires ATP for this reaction.
Strategy for efficient ligation of vector and insert DNA
To achieve good efficiency of ligation of foreign DNA into a vector and avoid high background of non-recombinants, following strategies are used:
1. Remove 5' phosphate groups from the cut vector by alkaline phosphatase to prevent self-ligation.
2. Sometimes the foreign DNA may get inserted in inverse direction when both the vector and foreign DNA are cut with one RE. To avoid this problem, ends of the vector are generated with two different REs. The ends of the insert are also obtained in the same way so that the foreign DNA is ligated in only one direction (Fig 8).
3. If vector and insert DNA are cut with an enzyme which leaves blunt DNA ends, the background of non-recombinant plasmids can be high. The problem can be solved by using high concentrations of both DNAs (vector and insert) and of the DNA ligase enzyme. Another way is to convert blunt ended molecules into those with cohesive ends. This can be achieved by adding homopolymer tails or addition of linkers/adapters.

Homopolymer tailing
A homopolymer is a polymer in which all the subunits are same e.g. a DNA strand made up of many deoxy- guanosine residues i.e. poly (dG). Homopolymer tailing is done with the help of enzyme terminal deoxynucleotidyl transferase which adds nucleotides to 3’ OH termini of a double stranded DNA molecule. The enzyme has the property to add same type of nucleotides
when the reaction is carried out in the presence of only one type of deoxynucleoside triphosphate resulting in formation of homopolymer tails.

Complementary homopolymer tails are added to the vector and the DNA insert e.g. if poly (G) tail is added to vector then poly C is added to the insert (Fig.9).

**Linkers and adaptors**

Linkers are short pieces of double stranded DNA, of known nucleotide sequence that are blunt ended but have R.E. site in them (Fig.10). They are chemically synthesized. Linkers are attached to the blunt ended DNA inserts with the help of enzyme DNA ligase. These are then cut with a suitable R.E. to produce cohesive ends.

Adaptors, like linkers are short synthetic oligonucleotides that have one blunt end and one sticky end. The blunt end of the adaptor is joined to the DNA insert to produce DNA molecules with sticky ends. These sticky ends have a known sequence which is complementary to the ends generated by cutting the vector with a R.E. so that these can be joined to the vector. Also two different adaptors may be joined to the two ends of DNA insert for its ligation in correct orientation.

c. Transformation

It is the process by which plasmids (or other DNA) are introduced into a host cell. The transformation of different host cells is described below:

i) Transformation of *E.coli*

The bacterial cells are made competent by incubation in the presence of divalent cations (usually Ca$^{2+}$) and a brief heat shock (42°C) is given which induces the *E. coli* cells to take up the foreign DNA. The efficiency of transformation is calculated as the number of transformants/μg of input
DNA. Alternatively transformation can also be done by electroporation. Here the foreign DNA and bacterial cells are mixed together and a brief pulse of high voltage is applied. This increases permeability of the bacterial cells and allows uptake of foreign DNA.

A recombinant DNA constructed with phage vector can be introduced in *E.coli* either by transfection or *in vitro* packaging.

**Transfection**
This process is same as transformation except that purified phage DNA i.e. recombinant phage vector is used in place of plasmid for introduction into the host.

**In vitro packaging**
Recombinant DNA molecules constructed with lambda phage vectors are mixed in a test tube with *in vitro* packaging extracts that contain phage head and tail structures. Assembly of the phage particles occurs automatically in the test tube and mature assembled phages are allowed to infect *E.coli* cells. This method is highly efficient for introduction of foreign DNA.

**ii) Transformation of B.subtilis**
For transformation of *B.subtilis*, protoplasts are obtained by treatment of bacterial cells with a lysozyme. These are mixed with recombinant plasmids in the presence of polyethylene glycol.
which allows uptake of DNA. High transformation frequency of more than $10^7$ transformants per microgram of DNA are generally obtained.

**iii) Transformation of yeast**

To introduce DNA into yeast, cell walls are enzymatically digested to produce spheroplasts which can take up DNA following treatment with CaCl$_2$. Cell walls then regenerate in specific media.

d. **Selection of transformants**

A host cell having foreign DNA introduced in it is called a transformant. At the end of transformation experiment, we get bacterial cells that may contain non-recombinant vector, desired recombinant vector or undesired recombinant vector or may not contain any vector i.e. non-transformants. To identify the clone containing desired piece of DNA from among several others, screening is carried out in two steps:

**Selection of clones carrying recombinants**

The selection of recombinants is generally done on the basis of marker genes present in the vector. There are two types of marker genes, selectable marker and a reporter gene or scorable marker.

**Selectable markers**

A selectable marker gene codes for a function which enables only those cells which possess it to survive under suitable conditions. For example, genes conferring resistance to antibiotics like ampicillin, tetracycline and kanamycin are good selectable markers. When a population of bacterial cells is plated on an ampicillin containing medium, only those cells that have ampicillin resistance genes survive and form colonies.

**Reporter genes**

A reporter gene produces a protein product whose activity can be assayed and permits either an easy selection or quick identification of cells in which it is present. Therefore it is also called as scorable marker. Among the more commonly used reporter genes are gus (codes for β-glucuronidase which produces blue colour in the presence of suitable substrate), lux (luciferase, produces phosphorescence, gfp (green fluorescence protein, fluoresces on irradiation with U.V.). Some examples making use of selectable and or scorable markers to identify recombinant clones are listed below:

i) **Insertional inactivation of antibiotic resistance gene**

This can be explained with the help of pBR322 which has two selectable markers i.e. Ap$'$ and Tc$'$. Both these genes have unique cloning sites in them. Insertion of foreign DNA in any of the sites causes inactivation of that gene. The recombinants thus become susceptible to one of the antibiotics while non-recombinants are resistant to both.

ii) **Insertional inactivation of lacZ gene**

Some vectors contain a gene or sometimes only part of a gene, which complements a function missing in their host cells, e.g. lacZ gene (encodes fragment of β-galactosidase) in the pUC vectors, M13 and some λ phage vectors which complements defective lacZ gene (encodes part of
β- galactosidase) in *E. coli* host strains. Insertion of foreign gene in the vector causes inactivation of *lacZ* gene. The recombinants are identified by formation of white colonies /plaques while non-recombinants form blue colonies /plaques (Fig.11).

**iii) Insertional inactivation of λ cl gene**

Some phage vectors contain unique restriction sites in *cl* gene e.g. λgt10. Insertional inactivation of *cl* gene causes change in plaque morphology. Normal plaques appear turbid while recombinants with inactive *cl* give rise to clear plaques.

**Selection of recombinant clones by insertional inactivation of *lacZ* gene**

Blue / white screening of transformants on nutrient agar plate containing ampicilin, X-GAL and IPTG

- = Non recombinant clone
- = Recombinant

**iv) Selection on the basis of λ genome size**

The λ packaging system can only insert DNA molecules between 38 and 52 kb size into the phage head. Anything less than 38 kb is not packaged. In some vectors e.g. λ replacement vectors, the size of the vector is less than 38 kb. The length of DNA insert can be so adjusted as to allow the packaging of only the recombinant DNA.

**v) Selection based on Spi (sensitive to phage infection) phenotype**

Phage λ cannot infect *E.coli* cells that have P2 phage integrated in their genome and is said to be Spi⁺. Insertion of foreign DNA causes change in phenotype from Spi⁺ to Spi⁻. The recombinants can infect *E.coli* having P2 phage while non-recombinants cannot.
vi) Selection based on growth on minimal medium
In yeast, an auxotrophic mutant that has non-functional \textit{leu2} gene is used as a host. Such a mutant is able to survive only if leucine is supplied in the growth medium. However, transformants are able to grow on a minimal medium (contains no added leucine) due to presence of \textit{leu2} gene in the vector (e.g. YEps).

Expression of cloned genes
For expression of cloned genes, special type of vectors called \textit{expression vectors} are used, which are designed so that a foreign gene can be placed under the control of the transcriptional and translational machinery of the host cell. Under appropriate conditions, the foreign gene can be transcribed, and the resulting messenger RNA translated, in the host cell to produce the protein encoded by the gene. Expression of cloned genes in suitable host can be used in a variety of ways. For example, they may be used for production of proteins of pharmaceutical and industrial importance in large quantities. Biochemical function of the protein can also be studied. Since the gene is cloned, it is relatively simple to mutate it, and obtain mutant protein, to see how changes in the gene can affect the function of the protein it encodes. Some other uses for such proteins are in raising antibodies, and in structural studies.

Expression of cloned genes in \textit{E. coli}
Gene expression in bacteria, as in all cells, involves the following steps:
1. Transcription: the production of messenger RNA from DNA template by RNA polymerase.
2. Translation: synthesis of the proteins using messenger RNA as template.

The three most important features required for expression of genes in \textit{E.coli} are as follows:
1. Promoter sequence: required for binding of RNA polymerase to initiate transcription.
2. Termination sequence: site which marks the point at the end of the gene where transcription stops.
3. The ribosome binding site (RBS): a purine-rich sequence, about 10 nucleotides upstream of the initiation (AUG) codon. This sequence is also called \textit{Shine-Dalgarno sequence}. It contains a sequence complementary to the 3’ end of 16S ribosomal RNA, which is part of the 30S ribosomal subunit.

The genes of eukaryotic organisms have different control regions (promoter and other regulatory elements) than bacteria. A eukaryotic gene is inactive in \textit{E.coli} simply because the bacterium does not recognize its expression signals. This problem is solved by inserting the foreign gene into the vector in such a way that it is placed under the control of \textit{E.coli} expression signals. Most \textit{E.coli} expression vectors utilize either the promoter of the \textit{lac} or of \textit{trp} operons from \textit{E.coli} or \textit{\lambda} P\textsubscript{L} promoter responsible for transcription of \textit{\lambda} DNA. The T7 promoter which is specific for RNA polymerase coded by T7 bacteriophage is also used in some vectors. It allows high level expression of genes as T7 RNA polymerase is much more active than the \textit{E.coli} RNA polymerase.

There are two main strategies for construction of expression vectors:
(i) Vectors that can synthesize pure proteins exclusively encoded by inserted gene.
(ii) Vectors which allow the synthesis of fusion protein (hybrid protein).
i) Vectors for synthesizing pure proteins

Such vectors are constructed by linking a suitable prokaryotic promoter, bacterial Shine-Dalgarno sequence and the start codon in such a way that when a foreign gene is cloned, these lie in front of the gene. This will allow transcription and translation of the foreign gene and gene product obtained is a pure protein. If the expression vector carries only the promoter and relies on the translational signals present in the foreign DNA, it is referred to as transcriptional fusion vector. This strategy of gene expression is also called transcriptional fusion (Fig. 12).

![Fig. 12]

ii) Vectors for synthesis of fusion proteins

The foreign gene is inserted into the coding region of a vector gene in such a way that the product of the gene expression is obtained as a hybrid protein consisting of the short peptide (e.g. part of β-galactosidase) coded by vector gene fused to the amino-terminus of foreign protein. This strategy is known as translational fusion (Fig. 13). It is important to note that the insert must be in frame with the start codon. Some examples of translational fusion vectors are pUC 18 and λ gt 11. The fusion protein is then cleaved to release the foreign protein. The fusion system has following advantages:

1. The presence of bacterial peptide at the start of fusion protein provides stability to the protein and prevents its degradation in the host. In contrast, the foreign proteins that lack a bacterial peptide are often destroyed by the host.

2. The bacterial peptide may constitute a signal peptide responsible for export of proteins outside the cell. This allows export of the recombinant protein either into the culture medium or into the periplasmic space between the inner and outer cell membranes. It simplifies the problem of purification of recombinant protein from the bacterial culture.

After the fusion protein is obtained, the bacterial peptide is removed by treating the protein with a suitable chemical or enzyme that cleaves the polypeptide chain at or near the junction. For example, if a methionine residue is present at the junction, the fusion protein can be cleaved with
cyanogen bromide, which cuts polypeptides specifically at methionine residues. But care must be taken that recognition sequences for the cleavage agent do not occur within the recombinant protein.

Fig. 13

iii) General problems with the production of recombinant proteins in E.coli

In spite of development of expression vectors, there are some difficulties in production of proteins from foreign genes cloned in E.coli. These problems can be either due to the sequence of the foreign gene or due to the limitations of E.coli as a host for recombinant protein synthesis.

a) Problems due to sequence of the foreign gene

There are three reasons which might prevent expression of foreign genes in E.coli:

(i) The foreign gene might contain introns. As E.coli genes do not contain introns, there is no splicing mechanism for removal of introns in E.coli. Therefore, for expression of eukaryotic genes in bacteria, cDNA (a complementary copy of mRNA) which is intron free is used.

(ii) The foreign gene might contain sequences that act as transcription termination signals in E.coli and may result in premature termination of mRNA synthesis. Site directed mutagenesis could be used to modify these sequences without altering the coding region.

(iii) Poor translation of foreign gene due to codon bias. Codon bias refers to the preference for certain codons over others from among multiple codons specifying the same amino acid. Although all organisms use the same genetic code, each organism exhibits a different pattern of codon preferences. In vitro mutagenesis could be used to modify and replace unfavorable codons with those that are preferred by E. coli. Alternatively if the size of the gene is small, it can be synthesized chemically.
b) Problems due to host cell
Most proteins undergo post-translational modifications after their synthesis. Some problems that are encountered during processing or folding of proteins are given below:

(i) Lack of protein processing: Processing of proteins is often required for their biological activity. These events involve modification of amino acids like methylation, phosphorylation, glycosylation etc. and are different in bacteria and eukaryotes. For example, glycosylation mechanisms which involve addition of sugar residues in proteins are absent in bacteria but present in eukaryotes. Such proteins that need to be glycosylated cannot be expressed in E.coli.

(ii) Incorrect folding: E.coli may not be able to fold the recombinant protein correctly and as it cannot synthesize disulfide bonds present in most eukaryotic proteins. An incorrectly folded protein often forms insoluble inclusion bodies (structures in the cell that contain aggregates of insoluble protein). This makes it difficult to recover the protein easily. This problem can be solved by using E.coli hosts that overproduce chaperon proteins (required for proper folding of proteins).

(iii) Degradation of recombinant proteins: Some foreign proteins may get degraded by the action of host proteases. In such cases, a strain of E. coli deficient in proteases (lon minus strains) may be used as host. The problem of stabilizing an expressed protein is also sometimes solved by expressing it as a fusion protein. Other ways to get around this problem include adding a signal sequence to the protein, causing it to be secreted into the periplasmic space. This can also, in some cases, make it easier to purify the protein.

Expression of proteins in B. subtilis
The transcription and translation machinery of B.subtilis differs from that of E.coli and some additional sequences are required. Promoters in Bacillus contain the -35 and -10 regions found in E.coli. In addition, they also contain an essential TGTG motif at -16 position. Ribosomes of Bacillus recognize only homologous mRNA (from same source). This selectivity is due to lack of counterpart of the E.coli ribosomal protein S1. The additional sequence requirements for efficient transcription and translation in B. subtilis explain why many E.coli genes are not expressed in Bacillus. Specialized vectors with control regions from Bacillus are thus used for expression of heterologous proteins in Bacillus.

Expression of heterologous proteins in yeast
Although E. coli is still the first choice for the production of heterologous proteins, the yeast S. cerevisiae has some attractive features. Proteins produced in yeast, unlike those produced in E. coli, are free from endotoxins. Yeast being eukaryotic, has several post translational processing mechanisms which are absent in bacteria. Such posttranslational modifications include particle assembly, amino terminal acetylation, myristylation and proteolytic processing. In addition, heterologous proteins secreted from specially engineered host strains are easily harvested from yeast culture media. The importance of yeast for production of protein products by recombinant DNA methods is illustrated by the fact that the first approved human vaccine, hepatitis B core antigen, and the first food product, rennin, were produced in yeast.

There are numerous varieties of expression vectors currently available for producing heterologous proteins in yeast, and these are derivatives of the Ylp, YEp and YCp plasmids. Cloned genes are placed under the control of the gal promoter which is induced by galactose. Other useful promoters are pho5, regulated by phosphate level in growth medium and cup1
induced by copper. Most yeast expression vectors also carry a termination sequence from *S. cerevisiae* gene because animal termination signals do not work efficiently in yeast.

**Polymerase chain reaction (PCR)**

Polymerase chain reaction also called *in vitro* cloning technique is an alternate method for producing many copies of a specific DNA fragment. It was developed by Karry Mullis in 1985. In this method, DNA fragment is amplified many fold *in vitro* with the help of enzyme DNA polymerase. Thus it differs from cell based cloning where DNA is inserted into a vector and amplified inside the host whereas PCR amplifies DNA in a test-tube.

*What does one need to amplify DNA using PCR?*

For PCR all the usual ingredients needed for DNA replication are required: A template (the DNA containing the target sequence that is to be copied), primers (to initiate the synthesis of the new DNA strands), thermostable DNA polymerase e.g. Taq DNA polymerase (to carry out the synthesis), dNTPs (nucleotide precursors) and Mg $^{++}$ ions required for enzyme activity.

The primers are synthetic single stranded DNA molecules whose sequence matches a region flanking the target DNA to be amplified.

*How does PCR work?*

PCR works by repeated cycles, each cycle consists of three steps i) denaturation i.e. DNA strand separation which is done at 94°C for 3-5 min., ii) annealing or binding of primers at 50-60°C for 1-2 min. and iii) extension of the primed strands at 72 °C for 2 min. This cycle is repeated 25-30 times. After completion of one cycle, two molecules of target DNA are formed from a single DNA molecule. This number goes on increasing exponentially ($2^n$, where n = no. of cycles) with each completed cycle e.g. after 30 cycles, millions of copies of target DNA are formed (Fig 14).

**Cell based cloning vs PCR**

PCR approach is used when the primers that anneal to either side of the gene of interest are available. Many copies of that gene can be obtained in a few hours by PCR, whereas it takes about a week to obtain the same by cell based cloning. However, there is a limit to the length of DNA sequence that can be copied by PCR. Five kilobases (kb) can be copied fairly easily with PCR, but this is shorter than the lengths of many genes, especially those of humans and other vertebrates. So cell based cloning must be used for long genes.

**Methods for isolation of gene/DNA fragment of interest**

Gene/DNA fragment of interest can be isolated by any of the following methods:

1. From genomic libraries
2. From cDNA libraries
3. Chemical synthesis
4. Transposon tagging
5. Map based cloning
6. PCR
The Polymerase Chain Reaction (PCR) is an important tool for many applications. For example, it can be used to amplify a sample of DNA when there isn't enough to analyze. The method uses specifically designed primers that are complementary to the sequence to be amplified. The primers provide a starting point for the extension of the DNA by a DNA polymerase (usually Taq or Pfu polymerase). Amplification is carried out in cycles. First, the DNA sample is heated up to separate the double strands. The sample is cooled slowly, allowing the primers to bind. Then, the sample is incubated at 72°C so that the DNA Polymerase can extend the primers, creating a long complementary strand of DNA. In this way, one double strand of DNA becomes 2, 2 become 4, 4 become 8 and so on.
**Genomic libraries**

*What is genome?*

Genome represents the total DNA present in the cell.

*What is a genomic library?*

A genomic library consists of cloned DNA fragments representing the entire genome of an organism or we can say that it is a collection of recombinant clones representing total genomic DNA of an organism. Genomic libraries are commonly used for isolating genes when information about product of the gene is not known or if we want to study the regulatory sequences of the gene.

---

**Figure 15**

**CLONING STRATEGIES**

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Cell Based Cloning</th>
<th>PCR based <em>in vitro</em> cloning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source of DNA/Gene</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic DNA (gene product</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not known or to study</td>
<td></td>
<td></td>
</tr>
<tr>
<td>regulatory sequences)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cDNA (Tissue specific</td>
<td></td>
<td></td>
</tr>
<tr>
<td>expression/gene product</td>
<td></td>
<td></td>
</tr>
<tr>
<td>known)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical synthesis (protein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sequence known)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Others: Transposon tagging</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Map-based cloning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic DNA or mRNA (gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>product known)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DNA Fragment /Gene**

- R.E. Digestion
- Mechanical Shearing
- cDNA
- chemically synthesized gene
- Blunt Ends
- Blunt /cohesive Ends
- Desired gene
- Random primers
- Several Amplified Fragments
- Modification of ends
- Cohesive ends
- Blunt or protruding ends
- Homopolymer tailing
- Linkers/ adaptors

**Ligation or joining to vector**

- Recombinant Vector

**Introduction into host**

- Transformation (Recombinant plasmid, BAC, YAC)
- Transfection (Recombinant phage, cosmid vector, *in vitro* packaging)
- Transformation

**Screening of recombinants**

- Hybridization with probes (DNA / RNA)
- Immunochemical (antibodies as probes)
- Direct selection
- PCR
**How is genomic library prepared?**

For construction of a genomic library, total DNA of the cell is isolated. Since DNA in all cells of an organism is same, genomic libraries can be prepared from any cell of an individual. Genes isolated from genomic libraries contain both coding (exons) as well as non-coding (introns) sequences. The DNA is then cut into small pieces with a restriction enzyme. Usually partial digestion with a tetra cutter such as Sau3A or Mbo I is done to obtain overlapping fragments of about 20 kb size DNA. Each piece of DNA is then joined to a vector or cloning vehicle. Vectors derived from λ phage are commonly used for preparation of genomic libraries. The recombinant DNA is then introduced into bacterial cells such as *E. coli*. The bacteria are plated on agar plates containing suitable medium on which they grow and form colonies or plaques (when vector is a phage). Each colony/plaque represents a recombinant clone carrying a different piece of genomic DNA. Various steps used in preparation of genomic library are shown in Fig 16. A collection of all these clones is called a genomic library. To isolate genes, genomic library is screened with the help of probes. Genomic libraries are useful to study detailed structures of genes, to identify regulatory regions, i.e. DNA sequences needed for correct expression of the gene.

**How many clones of genomic library should be screened?**

This depends on the size of the DNA fragments cloned and genome size of the organism whose genomic library is constructed. The number of clones to be screened is given by the following formula:

\[
N = \frac{\ln (1-p)}{\ln (1- f/n)}
\]

Where \(N\) represents the number of clones to be screened, \(p\) is probability of finding a clone, \(\ln\) represents natural log, \(f\) is the fragment size of cloned DNA and \(n\) is the size of the genome. As is clear from the Table 2, larger is the size of the genome i.e. \(n\), more will be the number of clones \((N)\) to be screened and larger is the size of DNA which is cloned \((f)\), less number of clones are required to be screened.

**Complementary (cDNA) libraries**

cDNA is a copy of double stranded DNA complementary to messenger RNA (mRNA) of the cell. This approach is used when we have information about expression of the gene. For gene isolation using this approach usually a cDNA library is prepared.

**What is a cDNA library?**

It is a collection of recombinant clones representing total mRNA of a cell or tissue at particular stage of development.

The advantage of cDNA library is that it contains only the coding region of a genome. It is the method of choice for cloning eukaryotic genes in bacteria as they lack splicing mechanism.
CONSTRUCTION OF GENOMIC LIBRARY

GENOMIC DNA

PARTIAL DIGESTION WITH Sav3A

λ PHAGE REPLACEMENT VECTOR

BamHI

PHAGE REPLACEMENT VECTOR CUT WITH BamHI

STUFFER FRAGMENT

LEFT ARM

RIGHT ARM

LIGATION

RECOMBINANT VECTORS

In vitro packaging

TRANSFECTION IN E. coli

GENOMIC LIBRARY

Fig. 16
Table 2: Genome sizes of some organisms and their estimated library sizes

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome size</th>
<th>Vector type</th>
<th>Insert size</th>
<th>Library size*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>4x10^6 bp</td>
<td>plasmid</td>
<td>4 kb</td>
<td>4.6x10^3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>phage</td>
<td>17 kb</td>
<td>1.5x10^4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cosmid</td>
<td>40 kb</td>
<td>458</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BAC</td>
<td>300 kb</td>
<td>59</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>1.5x 10^7 bp</td>
<td>phage</td>
<td>17 kb</td>
<td>4.6x10^4</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>1.65x10^8bp</td>
<td>phage</td>
<td>17 kb</td>
<td>4.5x10^4</td>
</tr>
<tr>
<td>Human</td>
<td>3x10^9 bp</td>
<td>plasmid</td>
<td>4 kb</td>
<td>3.5x10^6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>phage</td>
<td>17 kb</td>
<td>8.1x10^5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cosmid</td>
<td>40 kb</td>
<td>3.5x10^5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BAC</td>
<td>300 kb</td>
<td>4.6x10^4</td>
</tr>
</tbody>
</table>

*Number of clones to be screened for finding a gene with a probability of 99% (p=099)

Steps for preparation of cDNA library

The various steps for preparation of cDNA library are as follows:

1. **Isolation of mRNA**: The total mRNA is isolated from the cell/tissue of interest (where the desired gene is expressed). Because eukaryotic mRNAs contain a poly A tail at the 3’ end, they can easily be separated from other types of RNA (ribosomal and transfer RNA) molecules present in the cell.

2. **Preparation of single stranded (ss) cDNA**: The enzyme reverse transcriptase is used to synthesize a DNA strand complementary to each mRNA molecule (Fig. 17). A cDNA-RNA hybrid is formed from which RNA is removed with the help of RNase H or alkali.

3. **Preparation of double stranded (ds) cDNA**: The single stranded cDNA has the property to form a transient hair-pin loop at 3’end which is used to prime the synthesis of second strand of cDNA by DNA polymerase I. After the single-stranded DNA molecules are converted into double-stranded DNA, the hair-pin loop is removed with the help of S1 nuclease (acts only on single strands).

4. **Modification of ends**: Ends of double stranded cDNA are suitably modified by homopolymer tailing or addition of linkers and adapters to allow its insertion into a vector.

5. **Transformation of the host cells**: The recombinant vectors containing cDNAs corresponding to mRNAs are introduced into suitable host cells. This collection of transformed host cells represents the cDNA library from that particular cell/tissue.

The cDNA library is then screened for identification of the clone carrying the desired gene. The number of clones to be screened is given by the following formula:

\[ N = \frac{\ln (1-p)}{\ln (1-1/n)} \]

Where N represents the number of clones to be screened, p is probability of finding a clone, ln represents natural log and n is the number of mRNA molecules in the cell.
Strategies for cDNA cloning

The above described method for cDNA synthesis results in synthesis of somewhat shorter lengths of cDNAs due to removal of hair-pin loop or incomplete copying of mRNAs. So strategies are used to obtain full length cDNA molecules. One such strategy is to use a specially designed $E.coli$ vector to which mRNA is attached before copying and subsequent steps are followed while it is attached to vector. Similarly, use of S1 nuclease is avoided by adding poly G or poly C or poly T tail to 3’ end of single stranded cDNA. An oligonucleotide having sequence complementary to the tail is then used as a primer to synthesize double stranded cDNA.

A number of strategies are used for cloning cDNA depending upon information about expression of genes i.e. whether mRNA is present in abundance or as low copy number or is a rare species. If purified mRNA is available for a gene, cDNA is directly prepared. However, for isolation of
cDNA clones in the moderate and low abundance classes it is usually necessary to construct a cDNA library. Some commonly used vectors for cDNA cloning are \( \lambda \) insertion vectors namely \( \lambda \)gt10, \( \lambda \)gt11 and ZAP II.

cDNA libraries are tissue specific and are prepared using mRNA isolated from tissues where its gene is expressed. For example, for isolation of gene for insulin (which is synthesized in pancreas), mRNA for construction of cDNA library is isolated from pancreatic cells and not from liver or other cells. Similarly, for isolation of genes for gliadin, a seed protein in wheat, cDNA library from developing wheat grains is used.

**Chemical synthesis**

This method is used when we know the amino acid sequence of the protein i.e. product of the gene and the size of the gene is small. On the basis of codons for various amino acids, it is possible to deduce the nucleotide sequence of the gene from the sequence of amino acids in the protein. Once the base sequence of the gene is deduced, the polynucleotide of the same base sequence can be synthesized chemically. However, degeneracy of genetic code may present some problems and has to be looked into. There are three methods for chemical synthesis of genes; i) Phosphodiester method ii) Phoshotriester method and iii) Phosphite triester method. Among these phosphite triester method is the most efficient and is used for automatic synthesis of oligonucleotides in gene machines. Some examples of chemically synthesized genes include somatostatin, interferon and \textit{cry} gene.

**Transposon tagging**

Transposons are movable genetic elements found in prokaryotes as well as eukaryotes e.g. Tn5 in bacteria, Ac/Ds in maize. Transposon tagging is used for cloning of genes whose functions are not known. Here transposons are used for induction of mutations by insertional inactivation. The mutants are screened for changed phenotype. The mutant phenotype and the position of the mutation in the genetic map indicate that the correct gene has been hit. Using the DNA sequence of the transposon, DNA clones that contain the target gene can be identified, consequently the gene is cloned.

Transposon tagging requires that the organism must have well-characterized transposon system such as maize (corn). It is not feasible in many species including humans.

**Map based gene cloning**

Map-based gene cloning also called positional cloning is the isolation of a gene based solely on its position on a genetic map. The strategy is used when no information is available about the gene or its product. However, it requires the availability of molecular maps. Gene is located on a particular chromosome through molecular marker linkage analysis. Using the linked molecular marker (such as RFLP, RAPD, AFLP, SSRs etc) the gene is isolated through chromosome walking (Fig. 18).
PCR
Reverse transcription followed by polymerase chain reaction (RT-PCR) can be used for amplification of RNA sequences in cDNA form. Using gene specific primers, it is possible to clone specific cDNA molecules from total cellular RNA without need to purify mRNA. Under certain situations such as when starting DNA material is very small, for example, single cells, it is difficult to prepare genomic libraries by cell based cloning. In these cases, PCR is the only available alternative for gene isolation if specific primers are available. However, besides amplifying specific fragments, PCR can also be used to generate random genomic libraries with the help of random primers. One limitation of PCR is that the enzyme Taq polymerase can amplify only small fragments (1-2 kb). In recent years this has been overcome to some extent and it has been possible to amplify DNA fragments up to 22 kb by method called long PCR.

Screening of transformants for presence of desired gene/gene product
A number of methods have been devised for screening of desired transformants. These include the following:-
1. Direct selection for the desired gene
2. Identification of the clone from a gene library

Direct selection for the desired gene
In this method, the cloning experiment is designed in such a way that only the desired recombinant clones are obtained. Selection here occurs at the plating out stage. An example of direct selection is cloning of genes that specify antibiotic resistance such as kanamycin, tetracycline or ampicillin resistance. For example, let us consider selection of recombinants that contain gene for kanamycin resistance. The transformants are plated on agar medium containing
kanamycin. Only the cells that contain the cloned kanamycin resistance gene are able to survive and form colonies. Hence the technique is called direct selection.

Another situation where direct selection of recombinant clones is done is the marker rescue technique (Fig. 19). It makes use of auxotrophic mutant strains that have nutritional defects as the hosts for transformation. For example, suppose an *E.coli* strain is available which has a mutation in a gene encoding an enzyme involved in the biosynthesis of amino acid leucine. Such a mutant strain will only grow in a medium supplemented with leucine. By cloning DNA from a normal strain i.e. one that can synthesize its own leucine, in the mutant strain and selecting those transformants that can grow in absence of leucine it is possible to isolate the gene of interest.

Marker rescue is applicable for most genes that encode for biosynthetic enzymes as clones of these genes can be studied on minimal medium as described for leucine. Auxotrophic strains of yeast and filamentous fungi are also available and marker rescue has been used to select genes cloned in these organisms. However, the technique has two limitations: i) a mutant strain must be available for the gene in question ii) a medium on which only the wild type can survive is needed.
**Identification of the clones from a gene library**

A library has to be screened in order to find a clone. The identification of a specific clone from a DNA library can be carried out by exploiting either the sequence of the clone or the structure/function of its expressed product. The strategy for screening depends upon information about the gene of interest, the availability of probe and the cloning method used.

One of the key elements required to identify a gene during screening is a probe. A probe is a piece of DNA or RNA that contains a portion of the sequence complementary to the desired gene for which we are searching. It is used to detect specific nucleic acid sequences by hybridization (based on complementarity). The probe can be labeled radioactively (with P$^{32}$) or non-radioactively (biotin, digoxigenin and fluorescent dyes etc). Probes can be chemically synthesized based on the amino acid sequence of the protein coded by the gene. Probes can be homologous or heterologous.

**Homologous probe** - a probe that is exactly complementary to the nucleic acid sequence for which we are searching; e.g. a human cDNA used for searching a human genomic library.

**Heterologous probe** - a probe that is similar to, but not exactly complementary to the nucleic acid sequence for which we are searching; e.g., a mouse cDNA probe used to search a human genomic library.

There are several methods for screening DNA libraries. Some of the commonly used methods are described below:

1. Methods based on nucleic acid hybridization
2. Immunochemical methods
3. Screening DNA libraries using PCR

**Methods based on nucleic acid hybridization**

*a* Colony/plaque hybridization

This method was given by Grunstein and Hogness (1975). It is used for screening both genomic as well as cDNA libraries and is the most common method of library screening. The procedure has following steps:

1. The recombinant bacterial colonies or phage plaques to be screened are transferred from the culture plate on to a nitrocellulose filter paper by replica plating (Fig. 20).
2. The filter with colony replicas is treated with NaOH to lyse the cells/phages and to denature DNA.
3. The filter is then baked at 80°C for two hours in vacuum oven to fix the DNA.
4. The filter is allowed to hybridize with a labeled probe.
5. The filter is washed to remove the unbound excess probe, dried and then subjected to autoradiography if the probe is radioactively labeled. The washing conditions are kept stringent (high temperature and low salt concentration) when the probe is homologous and non-stringent (lower temperature and high salt concentration) in case of heterologous probe.
**Immunochemical methods**

This method is used to screen libraries for presence of desired gene product (protein), e.g. libraries constructed with expression vectors such as λgt11. In a similar method to the colony/plaque hybridization method, proteins expressed from the library are bound to membranes or filters rather than the DNA. The filters/membranes are then incubated with a specific labeled antibody (¹²⁵I) as probe to detect the desired protein.

There are many different expression screening methods that can be used to isolate a particular gene, and are designed according to what is known about the protein function.

Besides the above methods, when specific probes are not available, many indirect approaches may be used for identification. Two of such procedures that use *in vitro* translation system and identification of resulting polypeptide(s) are: Hybrid arrested translation (HART) and Hybrid released translation (HRT).

**Screening DNA libraries using PCR**

PCR can be used as alternative to hybridization for screening of genomic and cDNA libraries. It is possible to identify any clone by PCR but only if there is sufficient information about its sequence to make suitable primers. To isolate a specific clone, PCR is carried out with gene
specific primers that flank a unique sequence in the target. Pools of clones are maintained in multi-well plates. Each well is screened by PCR and positive wells are identified. The clones in each positive well are then diluted into a series in a secondary set of plates and screened again. The process is repeated until wells having homogenous clones corresponding to the gene of interest have been identified.

**Some important techniques used in characterization of genes/gene products**

A number of techniques are used for analyzing the genes/gene products. Some of these techniques are blotting and hybridization, DNA finger printing and Micro-array analysis. These are briefly discussed below.

**Blotting and hybridization techniques**

These techniques are used for detecting a specific gene sequence or its expression, number of copies of a gene, relatedness among organisms etc. There are three types of blotting techniques: Southern, Northern and Western blotting.

**Southern blotting**

Southern blotting is a technique for transferring DNA molecules from agarose gel to a solid support such as nitrocellulose paper or nylon membrane. Blotting is required to carry out hybridization with a probe so that specific DNA fragment in a complex mixture of DNA fragments can be detected. The technique was invented in 1975 by Edward Southern and is named after its inventor. Following are the steps of this technique:

1. The DNA to be analyzed is digested with a restriction enzyme and then separated by agarose gel electrophoresis.
2. The DNA fragments in the gel are denatured with alkali. Denaturation of DNA fragments prior to blotting is essential so that the probe can hybridize.
3. Denatured DNA is then transferred to a nitrocellulose filter or nylon membrane by blotting. A buffer saturated Whatman No.1 filter paper is placed on top of a support such as glass plate and two edges of the filter are dipped in buffer solution. Alternatively sponge dipped in buffer can be used as support. The gel is laid on top of the filter paper placed on the support (Fig. 21).
4. A sheet of nitrocellulose membrane cut to the size of the gel is placed on top of the gel. A stack of dry rough filter papers of the size of the gel are then placed on top of the nitrocellulose filter. A weight of about 0.5 kg is then placed on top of this. The DNA molecules move upward by capillary action of buffer and on coming in contact with the nitrocellulose filter they get bound to it.
5. The membrane is then heated at 80 °C for about 2h in a vacuum oven or exposed briefly (3-5 minutes) to UV radiations for firm binding of DNA to the nitrocellulose filter.

This nitrocellulose filter can be used for hybridization with a labeled probe. This is known as Southern hybridization. The location of the DNA fragment that hybridizes with the probe can be detected by autoradiography (when probe is radioactively labeled).
Northern blotting

When RNA molecules are transferred from gel to nitrocellulose membrane, it is known as Northern blotting. The technique is called "Northern" simply because it is similar to "Southern", not because it was invented by a person named "Northern".

Total RNA is extracted and separated by gel electrophoresis. Prior to blotting, the gel is treated with formaldehyde to ensure linear conformation of RNA molecules (no alkali treatment is given as RNA is degraded by alkali). These molecules are then transferred to a nitrocellulose filter in the same way as in Southern Blotting. This membrane can be used for detection of specific RNA molecules by hybridization with a labeled probe.

Western blotting

The transfer of proteins from gel to the nitrocellulose filter is called Western blotting and is used to detect a particular protein in a mixture of proteins. The technique is also called "immunoblotting". The proteins are separated by polyacrylamide gel electrophoresis and transferred to a nitrocellulose filter by electro blotting. The filter is then probed with a specific labeled antibody (\(^{125}\)I). The antibody will bind to a specific protein and is detected by autoradiography. If radiolabel is not used, bound antibody may be detected by a second antibody linked to an enzyme such as horse radish peroxidase. The enzyme catalyses a reaction that produces a colored product and is detected.

Dot and slot blots

In dot blots, DNA samples to be analysed are directly applied as a dot or spot adjacent to each other on a nitrocellulose membrane. The DNA is then denatured and firmly bound to membrane as single strands. The membrane is then hybridized with radioactively labeled DNA or RNA probe and signal is detected by autoradiography.
Slot blotting is similar to dot blotting except that the DNA samples are applied on the membrane in the form of slots with the help of slot blotting apparatus.

Dot and slot blots are used to study homology between organisms. More is homology; more intense will be the signal and vice-versa.

**DNA fingerprinting**

A DNA fingerprint is a genetic photograph or blueprint of an organism, whether a plant, microbe, animal or human. Variations in DNA sequence occur due to mutations, translocations, DNA rearrangements etc. These variations in sequence of DNA of two or more closely related individuals of a species can be detected by a number of techniques and are the basis of DNA fingerprinting. DNA fingerprinting has a number of applications such as identification of varieties of plants and protection of breeder’s rights; paternity determination in case of disputes; in forensics for identification of suspects of crime; diagnosis of genetic defects etc.

There are two broad categories of DNA fingerprinting techniques- a) hybridization based techniques such as restriction fragment length polymorphism (RFLP) and b) PCR based techniques. Besides fingerprinting, all these techniques have a number of other applications such as construction of molecular maps; tagging and map-based cloning of genes; marker-assisted selection; studying phylogenetic/ evolutionary relationship among organisms etc. A brief introduction to some of these techniques is given below:

*Restriction fragment length polymorphism (RFLP)*

RFLP technique was first developed by Botstein and co-workers in 1980. Total DNA from cells/tissues or whole organism is isolated and digested with a restriction enzyme. The DNA fragments produced are then separated by agarose gel electrophoresis. Due to inherited mutational changes in DNA, nucleotide base sequence differs between the individuals which results in a loss or gain of restriction enzyme sites in DNA. Hence, digestion of total DNA of an organism with a restriction enzyme will generate a unique restriction profile for an individual. Restriction enzyme digested DNA from two closely related individuals will appear as different sized fragments on agarose gel electrophoresis. This difference in length of fragments obtained when DNA of two or more individuals is digested with a restriction enzyme is called restriction fragment length polymorphism (Fig.22). In case when genome size is large e.g. plant or human DNA, detection of RFLPs involves additional steps such as blotting and hybridization with labeled probes.

The technique of DNA fingerprinting was first developed in 1985 by Alec Jeffreys and his colleagues using RFLPs. They used hypervariable tandem repeats of short sequence (11-60 bp long) called variable number tandem repeats (VNTR) or minisatellites as probes. These sequences are distributed throughout the eukaryotic genomes and reveal polymorphism. When restriction enzyme digestion cuts DNA flanking the VNTRs (the flanking sequences are conserved among species), the lengths of the resultant fragments will be variable depending upon the number of repeats at a given locus. Many different VNTR loci have been identified and are extremely useful for DNA fingerprint analysis in humans, plants as well as animals.
PCR based techniques

These include a number of techniques such as randomly amplified polymorphic DNA (RAPD), amplification length polymorphism (AFLP), inter simple sequence repeats (ISSR) and simple sequence repeats (SSRs) etc. These techniques have a number of advantages over RFLP in that i) a very small quantity (ng) of DNA is required whereas RFLP requires µg quantities of DNA ii) are less time consuming as no blotting and hybridization steps are involved, iii) are less cumbersome and iv) can be used to amplify even partially degraded DNA samples.

i) Randomly amplified polymorphic DNA (RAPD)

RAPD technique was given by Williams et al (1980) and makes use of single short synthetic oligonucleotide (8-10 bases long) primer of arbitrary sequence to amplify DNA fragments from the target DNA. The amplified fragments are then separated by agarose gel electrophoresis. Polymorphism is generated based on presence or absence of complementary sequence in the genome of different individuals and is scored as presence or absence of amplified band (Fig. 23). The technique is simple, rapid, highly polymorphic and has successfully been applied to distinguish varieties, genotypes and species of various plants as well as animals.

ii) Amplification fragment length polymorphism (AFLP)

AFLP technique given by Vos et al. (1995) is a combination of RFLP and RAPD techniques and has proven both powerful as well as reliable in fingerprinting. Genomic DNA from various individuals is isolated, cut with a restriction enzyme and adaptors are attached to fragments obtained by digestion with restriction enzyme. This is followed by PCR amplification using primers complementary to adaptors but with 1 or 2 arbitrary bases. The PCR products are
separated by denaturing polyacrylamide gel electrophoresis (PAGE) and detected by silver staining. About 60 to 80 bands per sample are usually obtained (Fig. 24).

**iii) Simple sequence repeats (SSRs) or microsatellites**

Simple sequence repeats or microsatellites are DNA sequences that consist of two to five nucleotide core units which are randomly repeated such as (AT)n, (CTT)n, and (ATGT)n. The
regions flanking the microsatellites are generally conserved among genotypes of the same species. PCR primers to the flanking regions are used to amplify the SSR containing DNA fragments which are then separated by PAGE. Polymorphism is created when PCR products from different individuals vary in length as a result of variation in the number of repeat units in the SSR (Fig. 25).

DNA chips and microarrays

DNA chips or DNA microarrays are arrays of genomic or cDNA sequences immobilized on glass slide or silicon wafers. They allow the rapid and simultaneous screening of thousands of genes for their expression. Besides study of gene expression pattern of an organism, DNA chips can be used for detection of genetic diseases such as cancer and cystic fibrosis; detection of single nucleotide polymorphisms (SNPs) etc. DNA chips are of different types –

(i) **oligonucleotide based chips** which contain high density of short oligonucleotides of known sequence, generally 20-25 nucleotides long and

(ii) **cDNA based chips**, which contain a high density of cDNA samples and are used for studying gene expression pattern. The DNA chips can be prepared by two different ways:

- DNA segments from known genes a few dozen to hundreds of nucleotides long are amplified by PCR and placed on a solid surface, using robotic devices that accurately deposit nanoliter quantities of DNA solution. Many thousands of such spots are deposited in a predesigned array on a surface area of just a few square centimeters.

- By photolithography. Here DNA is directly synthesized on the solid surface and makes use of nucleotide precursors that are activated by light, joining one nucleotide to the next in a photoreaction (as opposed to the chemical synthesis). Once the chip is constructed, it can be probed with mRNAs (or cDNAs from a particular cell type or cell culture to identify the genes being expressed in those
cells (Fig. 26). The analysis based on microarrays are very rapid and highly sensitive. In a single assay, identity of genes expressed in a specific tissue at a given time can be determined. In addition to DNA molecules, proteins have also been immobilized to form protein chips.

(iii) **Protein Chips** contain an array of antibodies immobilized as individual spots on a solid surface. A sample of proteins is added, and if the protein that binds any of the antibodies is present in the sample, it can be detected by a solid-state form of the ELISA assay.

Suggested readings