Genomic Libraries and cDNA Libraries

DNA (Gene) Libraries:

A DNA library is a set of cloned fragments that collectively represent the genes of a particular organism. Particular genes can be isolated from DNA libraries, much as books can be obtained from conventional libraries.

The secret is knowing where and how to look. There are two general types of gene library: a genomic library, which consists of the total chromosomal DNA of an organism; and a cDNA library, which represents the mRNA from a cell or tissue at a specific point of time.

The choice of the particular type of gene library depends on a number of factors, the most important being the final application of any DNA fragment derived from the library. If the ultimate aim understands the control of protein production for a particular gene or its architecture, then genomic libraries must be used.

However, if the goal is the production of new or modified proteins, or the determination of tissue-specific expression of timing patterns, cDNA libraries are more appropriate. The main consideration in the construction of genomic or cDNA libraries is, therefore, the nucleic acid starting material. Since the genome of an organism is fixed, chromosomal DNA may be isolated from almost any cell type in order to prepare genomic DNA.

In contrast, however, cDNA libraries represent only mRNA being produced from a specific cell type at a particular time in the cell's development. Thus, it is important to consider carefully the cell or tissue type from which the mRNA is to be deriver in the construction of cDNA libraries.

There are a variety of cloning vectors available, many based on naturally occurring molecules such as bacterial plasmids or bacteria-infecting viruses. The choice of vector also depends on whether a genomic library or cDNA library is constructed.

Constructing Gene Libraries:

Digesting Genomic DNA Molecules:

After genomic DNA has been isolated and purified, it is digested with restriction endonucleases. These enzymes are the key to molecular cloning because of the specificity they have for particular DNA sequences. It is important to note that every copy of a given DNA

molecule from a specific organism will give the same set of fragments when digested with a particular enzyme.

DNA from different organisms will, in general, give different sets of fragments when treated with the same enzyme. By digesting complex genomic DNA from an organism it is possible to reproducibly divide its genome into a large number of small fragments, each approximately the size of a single gene. Some enzymes cut straight across the DNA to give flush or blunt ends.

Other restriction enzymes make staggered single-strand cuts, producing short single-stranded projections at each end of the digested DNA. These ends are not only identical but complementary and will base-pair with each other; they are, therefore, known as cohesive or sticky ends. In addition, the 5′-end projection of the DNA always retains the phosphate groups.

Over 500 restriction enzymes, recognizing more than 200 different sites, have been characterized. The choice of which enzyme to use depends on a number of factors. For example, the recognition sequence of 6 bp will occur, on average, every 4096 (4^6) bases, assuming a random sequence of each of the four bases.
This means that digesting genomic DNA with EcoRI, which recognizes the sequence 5′-GAATTC-3′, will produce fragments each of which is, on average, just over 4 kb. Enzymes with 8 bp recognition sequences produce much longer fragments. Therefore, very large genomes, such as human DNA, are usually digested with enzymes that produce long DNA fragments. This makes subsequent steps more manageable, since a smaller number of those fragments need to be cloned and subsequently analyzed.

Ligating DNA Molecules:
The DNA products resulting from restriction digestion to form sticky ends may be joined to any other DNA fragments treated with the same restriction enzyme. Thus, when the two sets of fragments are mixed; base-pairing between sticky ends will result in the annealing of fragments that were derived from different starting DNA. There will, of course, also be pairing of fragments derived from the same starting DNA molecules, termed re-annealing.

All these pairing are transient, owing to the weakness of hydrogen bonding between the few bases in the sticky ends, but they can be stabilized by use of an enzyme, DNA ligase, in a process termed ligation. This enzyme, usually isolated from bacteriophage T4 and called T4 DNA ligase, forms a covalent bond between the 5′- phosphate at the end of one strand and the 3′-hydroxyl of the adjacent strand.

The reaction which is ATP dependent is often carried out at 10°C to lower the kinetic energy of the molecules, and so reduce the chances of base-paired sticky ends parting before they have been stabilized by ligation. However, long reaction times are needed to compensate for the low activity of DNA ligase in the cold. It is also possible to join blunt ends of DNA molecules, although the efficiency of this reaction is much lower than in sticky-ended ligations.

Since ligation reconstructs the site of cleavage, recombinant molecules produced by ligation of sticky ends can be cleaved again at the 'joins', using the same restriction enzyme that was used to generate the fragments initially. In order to propagate digested DNA from an organism it is necessary to join or ligate that DNA with a specialized DNA carrier molecule termed a vector.

Each DNA fragment is inserted by ligation into vector DNA molecule, which allows the whole recombinant DNA to then be replicated indefinitely within microbial cells. In this way a DNA fragment can be cloned to provide sufficient material for further detailed analysis or for further manipulations. Thus, all of the DNA extracted from an organism and digested with a restriction enzyme will result in a collection of clones. This collection of clones is known as a gene library.

Genomic Libraries:

Any particular gene constitutes only a small part of an organism's genome. For example, if the organism is a mammal whose entire genome encompasses some 106 kbp and the gene is 10 kbp, then the gene represents only 0.001% of the total nuclear DNA. It is impractical to attempt to recover such rare sequences directly from isolated nuclear DNA because of the overwhelming amount of extraneous DNA sequences.

Instead, a genomic library is prepared by isolating total DNA from the organism, digesting it into fragments of suitable size, and cloning the fragments into an appropriate vector. This approach is called shotgun cloning because the strategy has no way of targeting a particular gene but instead seeks to clone all the genes of the organism at one time.

The intent is that at least one recombinant clone will contain at least part of the gene of interest. This can be achieved by partial restriction digestion with an enzyme that recognizes tetra nucleotide sequences. Complete digestion with such an enzyme would produce a large number of very short fragments, but, if the enzyme is allowed to cleave only a few of its potential restriction sites before the reaction is stopped, each DNA molecule will be cut into relatively large fragments.

Average fragment size will depend on the relative concentrations of DNA and restriction enzyme and, in particular, on the conditions and durations of incubation. It is also possible to produce fragments of DNA by physical shearing, although the ends of the fragments may need to be repaired to make them flush ended. This is achieved by using a modified DNA polymerase termed Klenow polymerase.

This is prepared by cleavage of DNA polymerase with subtilizing, giving a large enzyme fragment which has no 5'→3′ exonuclease activity, but which still acts as 5'→3′ polymerase. Using the appropriate dNTPs, this will fill in any recessed 3′ ends on the sheared DNA. The mixture of DNA fragments is then ligated with a vector, and subsequently cloned.

If enough clones are produced there will be a very high chance that any particular DNA fragment, such as a gene, will be present in at least one of the clones. To keep the number of clones to a manageable size, fragments about 10 kb in length are needed for prokaryotic libraries, but the length must be increased to about 40 kb for mammalism libraries.

Genomic libraries have been prepared from hundreds of different species. Many clones must be created to be confident that the genomic library contains the gene of interest. The probability, P, that some number of clones, N, contains a particular fragment representing a fraction, f, of the genome is

$P = 1 – (1 -f)^N$.
Thus, N = In (1 – P)/ In (1 -f).

For example, if the library consists of 10 kbp fragments of the E. coli genome (4640 kbp total), over 2000 individual clones must be screened to have a 99% probability (P = 0.99) of finding a particular fragment. Since/ =10/4640 = 0.0022 and P – 0.99, N = 2093. For a 99% probability of finding a particular sequence within the 3×10^6 kbp human genome, N would equal almost 1.4 million if the cloned fragments averaged 10 kbp in size. The need for cloning vectors capable of carrying very large DNA inserts becomes obvious from these numbers.

Combinatorial Libraries:

Specific recognition and binding of other molecules is a defining characteristic of any protein or nucleic acid. Often, target ligands of a particular protein are unknown, or, in other instances, a unique ligand for a known protein may be sought in the hope of blocking the activity of the protein or otherwise perturbing its function.

Combinatorial libraries are the products of emerging strategies to facilitate the identification and characterization of possible ligands for a protein. These strategies are also applicable to the study of nucleic acids. Unlike genomic libraries, combinatorial libraries consist of synthetic oligomers. Arrays of synthetic oligonucleotides printed as tiny dots on miniature solid supports are known as DNA chips.

Specifically, combinatorial libraries contain very large numbers of chemically synthesized molecules (such as peptides or oligonucleotides) with randomized sequences or structures. Such libraries are designed and constructed with the hope that one molecule among a vast number will be recognized as a ligand by the protein (or nucleic acid) of interest.

If so, perhaps that molecule will be useful in a pharmaceutical application, for instance as a drug to treat a disease involving the protein to which it binds. An example of this strategy is the preparation of a synthetic combinatorial library of hexapeptides. The maximum number of sequence combinations for hexapeptides is 20^6 or 64,000,000.

One approach to simplify preparation and screening possibilities for such a library is to specify the first two amino acids in the hexapeptide while the next four are randomly chosen. In this approach, 400 libraries (20^2) are synthesized, each of which is unique in terms of the amino acids at positions 1 and 2 but random at the other four positions (as in AAXXXX, ACXXXX, ADXXXX, etc.) so that each of the 400 libraries contains 20^4 or 1,60,000 different sequence combinations.

Screening these libraries with the protein of interest reveals which of the 400 libraries contains a ligand with high affinity. This library is then systematically expanded by specifying the first 3 amino acids (knowing from the chosen 1-of-400 libraries which amino acids are best as the first 2); only 20 synthetic libraries (each containing 20^3 or 8000 hexapeptides) are made here (one for each third-position possibility, the remaining three positions being randomized).

Selection for ligand binding, again with the protein of interest, reveals the best of these 20, and this particular library is then varied systematically at the fourth position, creating 20 more libraries (each containing 20^2 or 400 hexapeptides). This cycle of synthesis, screening, and selection is repeated until all six positions in the hexapeptide are optimized to create the best ligand for the protein.

A variation on this basic strategy using synthetic oligonucleotides rather than peptides identified a unique 15-mer (sequence GGTTGGTGTGGTTGG) with high affinity (K_D = 2.7 nM) toward thrombin, a serine protease in the blood coagulation pathway. Thrombin is a major target for the pharmacological prevention of clot formation in coronary thrombosis.

Screening Libraries:

A common method of screening plasmid-based genomic libraries is to carry out a colony hybridization experiment. The protocol is similar for phage-based libraries except that bacteriophage plaques, not bacterial colonies, are screened. In a typical experiment, host bacteria containing either a plasmid based or bacteriophage-based library are plated out on a petri dish and allowed to grow overnight to form colonies (or in the case of phage libraries, plaques) (Fig 4.10).

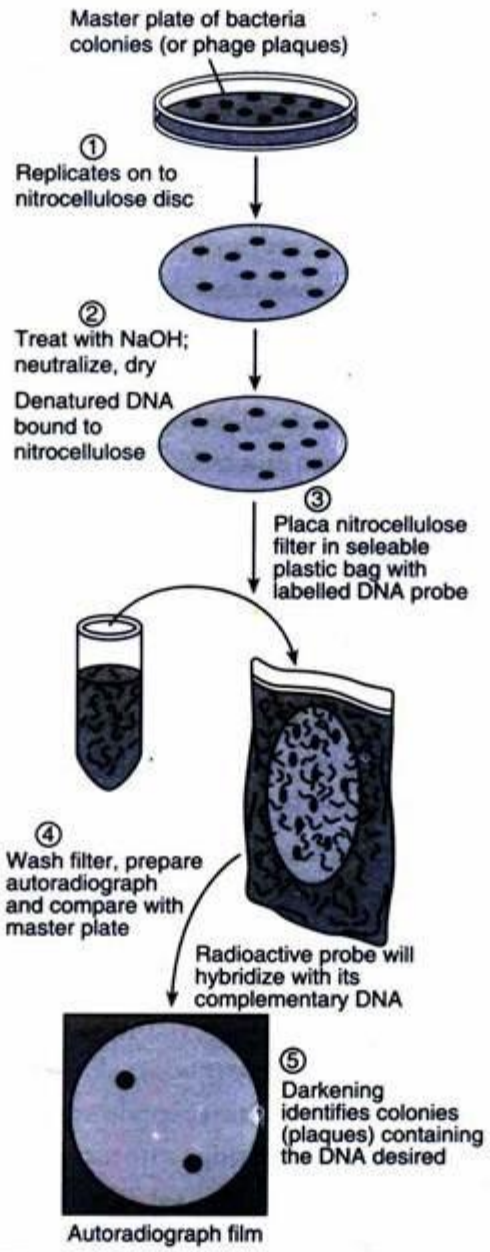

Fig. 4.10: Screening a genomic library by colony hybridization (or plaque hybridization). Host bacteria transformed with a plasmid-based genomic library or infected with a bacteriophage-based genomic library are plated on a petri plate and incubated overnight to allow bacterial colonies (or phage plaques) to form. A replica of the bacterial colonies (or plaques) is then obtained by overlaying the plate with a nitrocellulose disc. (1) Nitrocellulose strongly binds nucleic acids; single-stranded nucleic acids are bound more tightly than double-stranded nucleic acids. (Nylon membranes with similar nucleic acid – and protein-binding properties are also used.) Once the nitrocellulose disc has taken up an impression of the bacterial colonies (or plaques), it is removed and the petri plate is set aside and saved. The disc is treated with 2 M NaOH, neutralized, and dried. (2) NaOH both lyse any bacteria (or phage particles) and dissociates the DNA strands. When the disc is dried, the DNA strands become immobilized on the filter. The dried disc is placed in a sealable plastic bag, and a solution containing heat-denatured (single-stranded), labelled probe is added. (3) The bag is incubated to allow annealing of the probe DNA to any target DNA sequences that might be present on the nitrocellulose. The filter is then washed, dried, and placed on a piece of X-ray film to obtain an autoradiogram. (4) The position of any spots on the X-ray film reveals where the labelled probe has hybridized with target DNA. (5) The location of these spots can be used to recover the genomic clone from the bacteria (or plaques) on the original petri plate.

A replica of the bacterial colonies (or plaques) is then obtained by overlaying the plate with a nitrocellulose disc. The disc is removed, treated with alkali to dissociate bound DNA duplexes into single-stranded DNA, dried, and placed in a sealed bag with labelled probe. If the probe DNA is duplex DNA, it must be denatured by heating at 70°C.

The probe and target DNA complementary sequences must be in a single stranded form if they are to hybridize with one another. Any DNA sequences complementary to probe DNA will be revealed by autoradiography of the nitrocellulose disc. Bacterial colonies (phage plaques) containing clones bearing target DNA are identified on the film and can be recovered from the master plate.

Probes for Southern Hybridization:

Clearly, specific probes are essential reagents if the goal is to identify a particular gene against a background of innumerable DNA sequences. Usually, the probes that are used to screen libraries are nucleotide sequences that are complementary to some part of the target gene. To make useful probes requires some information about the gene's nucleotide sequence.

Sometimes such information is available. Alternatively, if the amino acid sequence of the protein encoded by the gene is known, it is possible to work backward through the genetic code to the DNA sequence (Fig. 4.11). Because the genetic code is degenerate (that is, several codons may specify the same amino acid), probes designed by this approach are usually degenerate oligonucleotides about 17 to 50 residues long (such oligonucleotides are so-called 17- to 50-mers).

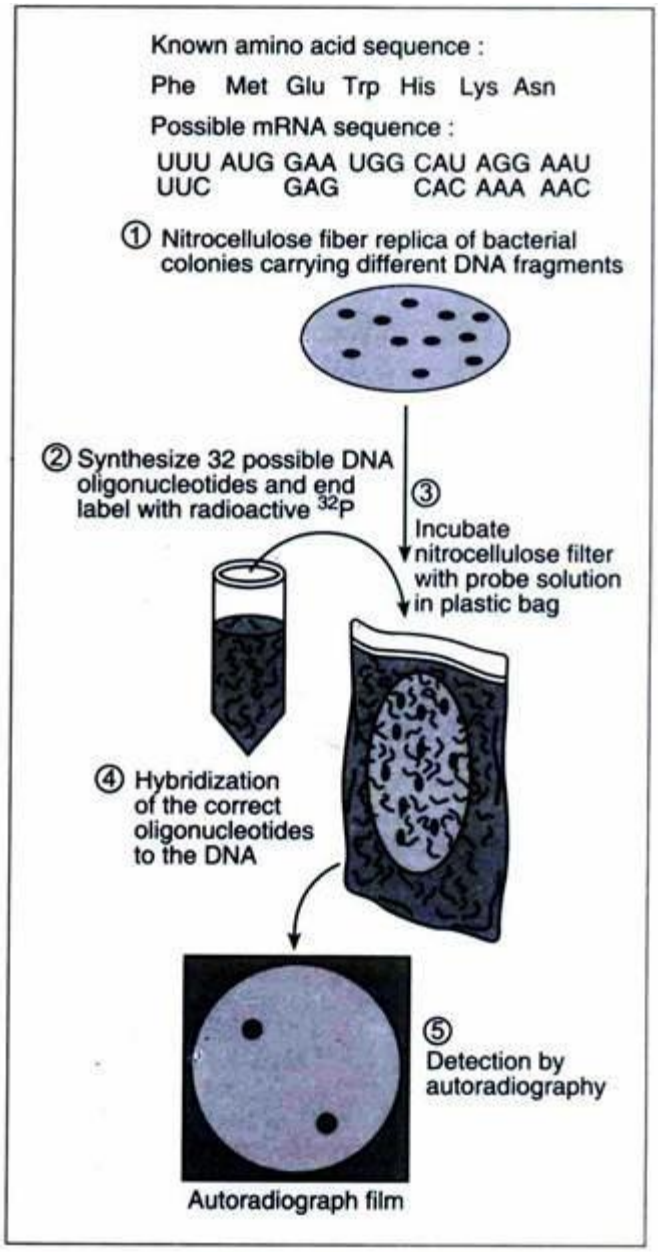

Fig. 4.11: sequence. A radioactively labelled set of DNA (degenerate) oligonucleotides representing all possible mRNA coding sequences is synthesized. (In this case, there are 25 or 32.) The complete mixture is used to probe the genomic library by colony hybridization.

The oligonucleotides are synthesized so that different bases are incorporated at sites where degeneracies occur in the codons. The final preparation thus consists of a mixture of equal-length oligonucleotides whose sequences vary to accommodate the degeneracies. Presumably, one oligonucleotide sequence in the mixture will hybridize with the target gene. These

oligonucleotide probes are at least 17-mers because shorter degenerate oligonucleotides might hybridize with sequences unrelated to the target sequence.

A piece of DNA from the corresponding gene in a related organism can also be used as a probe in screening a library for a particular gene. Such probes are termed heterologous probes because they are not derived from the homologous (same) organism. Problems arise if a complete eukaryotic gene is the cloning target; eukaryotic genes can be tens or even hundreds of kilo-base pairs in size.

Genes of this size are fragmented in most cloning procedures. Thus, the DNA identified by the probe may represent a clone that carries only part of the desired gene. However, most cloning strategies are based on a partial digestion of the genomic DNA, a technique that generates an overlapping set of genomic fragments.

This being so, DNA segments from the ends of the identified clone can now be used to probe the library for clones carrying DNA sequences that flanked the original isolate in the genome. Repeating this process ultimately yields the complete gene among a subset of overlapping clones.

cDNA Libraries:

cDNAs are DNA molecules copied from mRNA templates. cDNA libraries are constructed by synthesizing cDNA from purified cellular mRNA. These libraries present an alternative strategy for gene isolation, especially eukaryotic genes. Because most eukaryotic mRNAs carry 3′-poly(A) tails, mRNA can be selectively isolated from preparations of total cellular RNA by oligo(dT)-cellulose chromatography (Fig. 4.12). DNA copies of the purified mRNAs are synthesized by first annealing short oligo (dT) chains to the poly(A) tails.

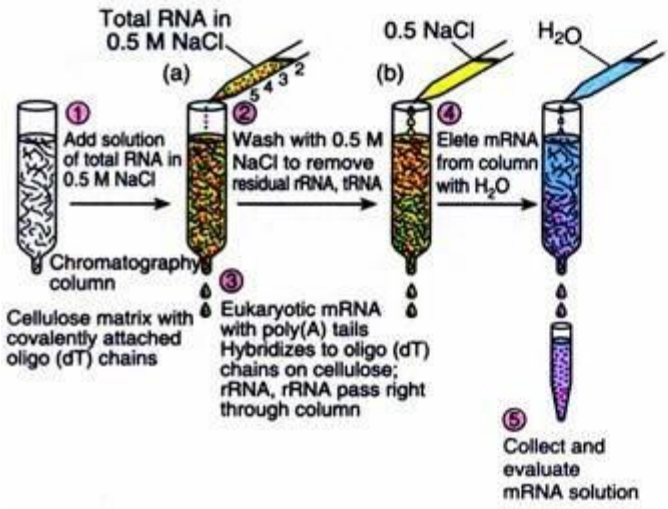

Fig. 4.12: Isolation of eukaryotic mRNA via oligo(dT)-cellulose chromatography. (a) In the presence of 0.5 M NaCl, the poly(A) tails of eukaryotic mRNA anneal with short oligo(dT) chains covalently attached to an insoluble chromatographic matrix such as cellulose. Other RNAs, such as rRNA (green), pass right through the chromatography column. (b) The column is washed with more 0.5 M NaCl to remove residual contaminants. (c) Then the poly(A) mRNA is recovered by washing the column with water because the base pairs formed between the poly(A) tails of the mRNA and the oligo(dT) chains are unstable in solutions of low ionic strength.

These oligo(dT) chains serve as primers for reverse transcriptase-driven synthesis of DNA (Fig. 4.13). (Random oligonucleotides can also be used as primers, with the advantages being less dependency on poly(A) tracts and increased likelihood of creating clones representing the 5′-ends of mRNAs.) Reverse transcriptase is an enzyme that synthesizes a DNA strand, copying RNA as the template. DNA polymerase is then used to copy the DNA strand and form a double-stranded (duplex DNA) molecule.

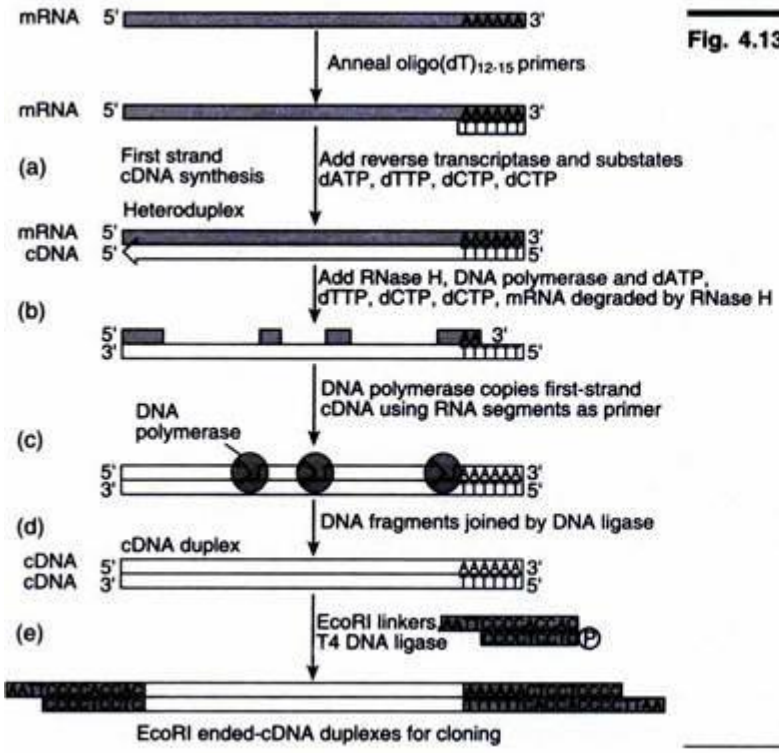

Fig. 4.13: Reverse transcriptase–driven synthesis of cDNA from oligo(dT) primers annealed to the poly(A) tails of purified eukaryotic mRNA. (a) Oligo(dT) chains serve as primers for synthesis of a DNA copy of the mRNA by reverse transcriptase. Following completion of first-strand cDNA synthesis by reverse transcriptase, RNase H and DNA polymerase are added (b). RNase H specifically digests RNA strands in DNA:RNA hybrid duplexes. DNA polymerase copies the first-strand cDNA, using as primers the residual RNA segments after RNase H has created nicks and gaps (c). DNA polymerase has a 5′→3′ exonuclease activity that removes the residual RNA as it fills in with DNA. The nicks remaining in the second-strand DNA are sealed by DNA ligase (d), yielding duplex cDNA. *Eco*RI adapters with 5′-overhangs are then ligated onto the cDNA duplexes (e) using phage T4 DNA ligase to create *Eco*RI ended cDNA for insertion into a cloning vector.

Ligation of blunt-ended DNA fragments is not as efficient as ligation of sticky ends; therefore, with cDNA molecules additional procedures are undertaken before ligation with cloning

vectors. One approach is to add cDNA small, double stranded molecules with one internal site for a restriction endonuclease; these are termed nucleic acid linkers. Numerous linkers are commercially available with internal restriction for many of the most commonly used restriction enzymes.

Linkers are blunt end ligated to cDNA but since they are added much in excess of the cDNA, the ligation process is reasonably successful. Subsequently the linkers are digested with the appropriate restriction enzyme, which provides the sticky ends for efficient ligation to a vector digested with the same enzyme. This process may be made easier by the addition of adaptors rather than linkers, which are identical except that the sticky ends are performed and so there is no need of restriction digestion following ligation.

Therefore, lastly Linkers are added to the DNA duplexes rendered from the mRNA templates, and the cDNA is cloned into a suitable vector. Once a cDNA derived from a particular gene has been identified, the cDNA becomes an effective probe for screening genomic libraries for isolation of the gene itself.

Because different cell types in eukaryotic organisms express selected subsets of genes, RNA preparations from cells or tissues in which genes of interest are selectively transcribed are enriched for the desired mRNAs. cDNA libraries prepared from such mRNA are representative of the pattern and extent of gene expression that uniquely define particular kinds of differentiated cells.

cDNA libraries of many normal and diseased human cell types are commercially available, including cDNA libraries of many tumour cells. Comparison of normal and abnormal cDNA libraries, in conjunction with two dimensional gel electrophoretic analysis of the proteins produced in normal and abnormal cells is a promising new strategy in clinical medicine to understand disease mechanisms.