DNA Libraries

In molecular biology, a library is a mixture of DNA molecules. In most cases, the DNA molecules are short to medium sized DNA fragments. A library contains a large variety of different DNA molecules (in theory, **all** of the relevant sequences) from a common source.

In most cases, the DNA fragments are added to a common carrier DNA such as a plasmid or bacteriophage genome. The most commonly used vectors for libraries are variants of bacteriophage with specific properties designed to assist in screening of the library.

Libraries allow the isolation of small subsets of the total DNA.

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

cDNA libraries are produced only from higher eukaryotes. cDNA is "complementary DNA"; it is produced by reverse transcribing mRNA. Because mRNA has the introns removed, cDNA, like its parent mRNA, consists of continuous coding sequences. Lower eukaryotes (such as yeast) do not contain introns, and preparation of cDNA is generally unnecessary for these organisms. The vast majority of higher eukaryotic mRNA molecules have poly(A) tails, and therefore can be purified from cells using an oligo-T resin.

A cDNA library includes only the sequences that are actively expressed in the source material, and messages that are present in low quantities may not be found in the library. In addition, long messages may be incomplete; if the reverse transcriptase does not fully copy the mRNA molecule, the 5´ end of the coding region may not be present in the library.

cDNA libraries are required for cloning a gene and expressing a protein from a higher eukaryote in a foreign organism. For lower eukaryotes and prokaryotes, cDNA libraries are unnecessary (and are difficult to produce due to lack of poly(A) tails on the mRNA of these organisms). The lack of introns in these organisms means that contiguous coding sequences are present in the genomic DNA.

Why make libraries?

Genomic libraries are comprised of whole genomic DNA separated into smaller, more manageable pieces. By using a screening method, it is possible to extract one relatively small segment that probably contains the DNA of interest.

cDNA libraries reflect the expressed genes within a cell. They therefore allow the specific isolation of the coding regions. Note that these are used for multicellular organisms and therefore not all cDNAs expressed in an organism will be present in any given library.

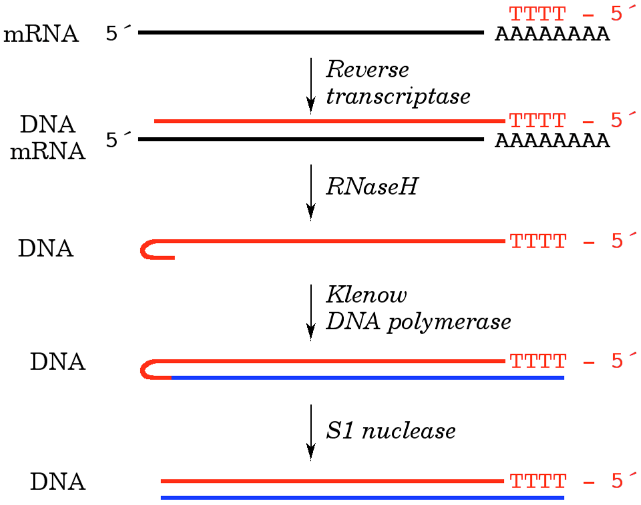
Making a prokaryotic library

Prokaryotic organisms lack introns, and do not attach poly(A) tails to their mRNA. As a result, in general, only genomic libraries are made from prokaryotic organisms. The process involves cleaving the genomic DNA either enzymatically or by using shearing forces. The DNA fragments are then attached to something (either a plasmid, or more frequently, a bacteriophage) that can be replicated.

Making a eukaryotic library

Making a genomic library from a eukaryotic organism is similar to the process used for generating prokaryotic libraries. The only potential problem is the larger size of the eukaryotic genome, which means that in many cases, larger fragments are necessary to allow the use of a reasonable number of clones for the entire library. Genomic libraries for higher eukaryotes are used for genome sequencing projects, and to allow analysis of transcriptional control elements and intron sequences.

In higher eukaryotes, the presence of introns means that the protein coding regions are discontinuous. Generating a library that contains uninterrupted coding regions, requires the preparation of cDNA. The general procedure requires the purification of the mRNA (the poly(A) tail of mRNA molecules will bind to a poly(T) resin, which simplifies the mRNA isolation).



The first strand of the cDNA is synthesized using reverse transcriptase. The RNA template must then be degraded using RNaseH, which specifically digests the RNA portion of an RNA-DNA duplex. Klenow or a similar DNA polymerase is then used to synthesize the second strand.

Both the reverse transcriptase and Klenow polymerases, like all DNA polymerases, require a primer. In the case of the reverse transcriptase, the primer is a short poly-T sequence that binds to the poly(A) tail. The synthesis of the second strand takes advantage of the hairpin loop of DNA that frequently forms at the 3´ end of single stranded nucleic acids. Depending on the size and location of the hairpin loop, some sequences present at the 5´ end of the mRNA may not be present in the final cDNA.

Prior to ligation, the loop at the end of the DNA molecule must be cleaved using S1 nuclease. The double stranded DNA molecules can then be ligated into the host molecule for the library.

Propagating plasmid DNA

Sometimes the library DNA molecules are ligated to plasmids. While bacteriophages and viruses are specialized for the process of injecting DNA into cells, plasmid DNA transfer requires other methods. Plasmids are somewhat more difficult to insert into cells, but are frequently used because manipulating phage DNA is often more difficult than manipulating plasmid DNA.

Replication of a plasmid requires the insertion of the plasmid DNA into the bacteria being used for the replication process. In general, bacteria are reluctant to take up DNA from their environment (more precisely, they generally will not do so without degrading the DNA first). In order to improve the probability that the cells will actually internalize the plasmid DNA, the cells should first be "competent" to absorb the plasmid DNA. **Competent cells** are cells treated with a salt solution that increases their probability of DNA uptake, although the mechanism by which these salt solutions function is not understood.

Transformation of bacteria is a low efficiency process. In most cases, only about 1 in 10^6 to at best 1 in 1000 bacteria actually internalize plasmid DNA. Of the DNA molecules added, only about 1 in 100,000 to at best 1 in 10 plasmid molecules are taken into cells. By comparison, addition of a single (functional) phage particle to a culture is usually enough to allow infection of a cell, and essentially all of the cells present will become infected if enough phage is present.

Electroporation is a modified transformation technique, in which high voltage is applied to a cell suspension for a short time (usually milliseconds). This results in DNA uptake. Electroporation is a higher efficiency process than normal transformation, and is also used for eukaryotic cells.

The mechanism by which competent cells and electroporated cells take up DNA is poorly understood.

Conjugation

An alternative method for transferring DNA from one cell to another uses a cellular process called conjugation. Conjugation is a bacterial method for exchanging genetic information. Most plasmids used in molecular biology, and many *E. coli* laboratory strains are not capable of this process. Conjugation is currently rarely used as a method for transferring experimental DNA constructs, because there are easier, more controllable methods.

Conjugation usually uses a large circular DNA molecule (the F-plasmid or episome, a 150-300 kb molecule present in very low copy number).

Bacterial artificial chromosomes (BACs) are very large plasmids based on the F episome. These are present at one to two copies per cell, but can allow replication of more than 100 kb of DNA. BACs are used to propagate large DNA fragments, and have been very important in the genome sequencing efforts.

Bacteriophage □

A bacteriophage is a virus that infects bacteria. One bacteriophage called bacteriophage □ has been intensively studied, and has some properties useful for molecular biology.

As with all viruses, bacteriophage □ is specialized for inserting its genetic material into a host organism. For the wild-type bacteriophage □, two possible processes occur after an *E. coli* cell becomes infected. One process is called the **lytic cycle**. In this case, phage production occurs at a high rate until the cell is irreversibly damaged and lyses. The lytic process produces ~100 phage particles per cell killed.

The second process forms a **lysogen**. In some cells, the bacteriophage □ DNA is incorporated into host genome, and remains present without reproducing autonomously. The modified cell is called a □ lysogen. At some later time when the progeny cells are stressed, viral protein synthesis occurs, and the doomed cells begins actively producing phage particles in a resumption of the lytic process.

About 20 kb of the 50 kb □ genome is involved in lysogen cycle. This DNA can be deleted, resulting in a phage that can only perform lytic processes. In most molecular biological experiments, the lysogen stage is undesirable, and therefore this DNA is superfluous.

The bacteriophage □ particle can hold ~50 kb of DNA (more than this will not physically fit into the particle). The bacteriophage □ particle must hold at least 40 kb in order to be functional. Because it is possible to purify empty (*i.e.* lacking DNA) particles, and because these can be combined with DNA *in vitro* to produce functional phage particles, bacteriophage □ is very useful for library propagation.

In libraries based on bacteriophage □, 10 to 20 kb of library DNA is added to the phage DNA sequences required for the lytic cycle (~30 kb), and this DNA is then inserted into empty phage particle to produce functional phage.

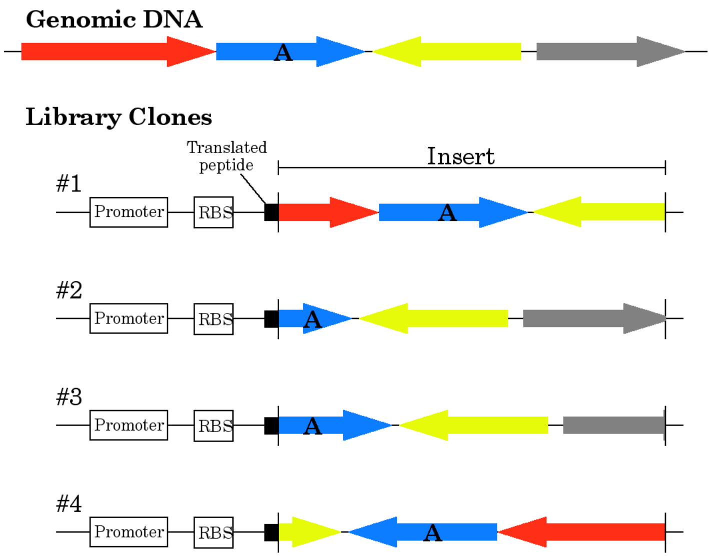
Unlike transformation with a plasmid, infection of a cell with a phage is a very high efficiency process. Libraries in bacteriophage □ are therefore readily replicated with very simple techniques. Specialized variants of bacteriophage □ have been developed that simplify the process of screening the library.

Screening libraries

In order to find a sequence of interest from within a library, it is necessary to know something about the DNA you are attempting to find.
1) You may know something about the protein
2) You may know part of the sequence
3) You may know the entire sequence
4) You may know almost nothing about the protein or the sequence.

Protein-based methods

Expression libraries are one type of library, designed for finding protein-coding sequences based on previously available information about the protein. In these libraries, the foreign DNA is inserted after a promoter sequence.



In the drawing at right, the genomic DNA was fragmented and inserted into a library vector of some type (such as a variant of bacteriophage). Some of the clones containing Gene A (either the entire gene, or portions therefore) are also shown. Clones 1 and 4 contain the entire Gene A, in opposite orientations. Clones 2 and 3 contain portions of the gene, arranged so that the vector promoter will drive transcription and the vector ribosome-binding site (RBS) and vector peptide will drive translation. If the Gene A fragment fused to the vector peptide in the correct reading frame, it should be translated. The cells will therefore sometimes transcribe and translate the protein coded by the library DNA. In clone 1, a Gene A transcript might not be translated, because the coding sequence is not fused to the vector peptide. In clone 4, the coding sequence is in the wrong orientation, and the Gene A peptide product will not be translated from any transcript driven by the vector promoter. Thus, the **DNA sequence must be in the correct orientation, and must be in the correct reading frame** for the translation. (Note: although other methods have been used for protein expression from library DNA, this method, where the library gene is fused to a peptide coding region from the vector is the most commonly used.

Screening the library for the expressed protein requires some method for specifically finding the protein. Two methods are frequently used, depending on the protein of interest:

1) Using an antibody raised against the protein
2) Using an assay monitors the cells for the activity of the expressed protein.

Antibody-based screening

In some cases, the protein of interest has been purified using standard biochemical methods, and the purified protein used to raise antibodies specific for the protein. However, the coding sequence for the protein is not known. In this case, an investigator might decide to screen an expression library for the presence of the protein using the antibody; once a clone containing the gene for the protein was isolated, it could be sequenced. (This method was common a few years ago; because many proteins have had their genes identified and sequenced, it is currently a much less commonly used technique.)

The general procedure is to spread out cells containing the library on culture plates, such that each individual clone is reasonably well separated from the others. (The plate shown below is more sparsely populated than would normally be done. A typical library contains ~10^6 different clones; screening even a significant proportion of the clones requires a minimum of several hundred clones per plate.) A filter membrane[2] comparable in size to the plate is then placed on the plate, and lifted off. Some of the cells from each colony stick to the membrane, forming a "colony lift".

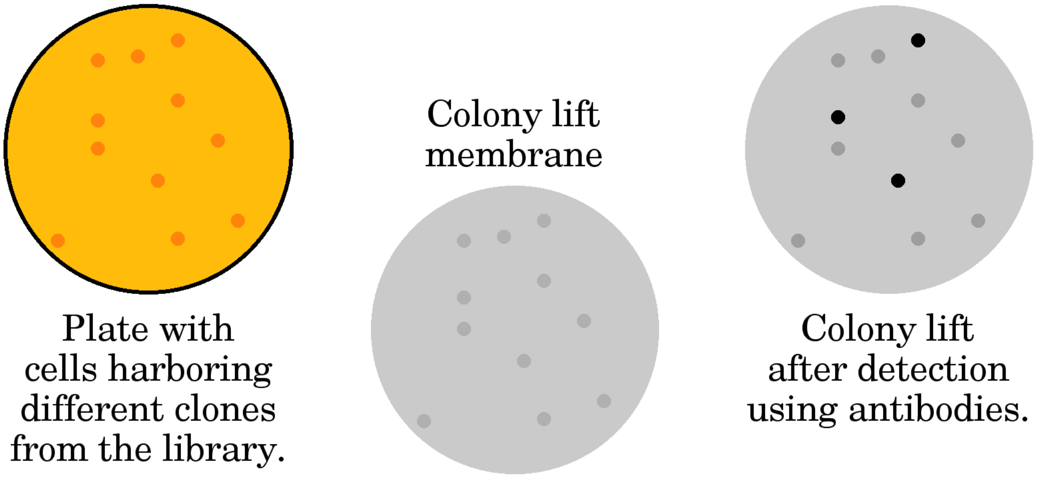


Plate with cells harboring different clones from the library.

Colony lift membrane

Colony lift after detection using antibodies.

The colony lift filter is then exposed to reagents that lyse the cells (in order to release the expressed protein; for bacteriophage libraries, this is often unnecessary, because the phage will lyse the cells without intervention). The next step is to allow antibodies to bind to the proteins, followed by washing off the excess antibodies. Finally, some method is used to detect the bound antibodies. We will cover some methods that could be used later when we discuss ELISA assays and Western blotting.

Activity-based methods

In principle, clones from the expression library that contain full-length, or close to full-length, open reading frames will result in the production of functional proteins. For proteins with readily detectable activities, assays for the activity allow the isolation of clones containing the protein coding sequence. This method has more stringent requirements for the library construct, because peptides that will bind antibodies are frequently too truncated to exhibit activity. In addition, with relatively few exceptions, testing for activity requires more effort than screening using antibodies.

DNA-based methods:

An alternative to looking for expressed protein is to test for the DNA of interest directly. The advantage of DNA-based screening methods is that the reading frame and orientation of the gene sequence within the library vector are irrelevant. On the other hand, these methods require more knowledge about the DNA of interest than the protein-based methods.

[2] The filter used for protein binding is usually made from either nitrocellulose or polyvinyl difluoride (PVDF), because proteins bind tightly to these membranes. For DNA-based methods, the filter is usually nitrocellulose, which also binds DNA.

Hybridization screening

If only a partial sequence of the protein or the protein-coding region is available, a method called hybridization screening is frequently used to find the DNA of interest. In hybridization screening, a nitrocellulose filter is placed on top of the culture plate containing a subset of the library. Some of the phage particles will stick to the filter. The filter is treated with reagents that denature the phage particles and separate the strands of the DNA. The single stranded DNA molecules are then covalently cross-linked to the filter, and the filter is probed with a DNA molecule corresponding to the sequence of interest. The probing process involves adding DNA that is labeled in some way (most commonly, the DNA is made radioactive by attaching a radioactive isotope of phosphorus to its 5´ end using polynucleotide kinase). The probe DNA will be allowed to form regions of double stranded DNA with the DNA on the filter (this is called "hybridization"). The excess DNA can then be washed off, and the bound DNA detected by autoradiography (placing the filter on a piece of film, developing the film, and looking for dark spots). The general appearance of the process is similar to that used for antibody based screening, although the specific reagents are different.

Brute-force methods

If very little information is known, it is sometimes still possible to find a protein coding sequence using the brute force method of sequencing all of the DNA molecules in the library. This sounds inefficient, but advances in sequencing technology have meant that this method is used more and more frequently.

Brute force sequencing has been used to find aberrant genes of poorly understood or unknown function, based on linkage analysis. For example, a few years ago, the gene for Huntington's chorea was known to be present on chromosome 4 based on genetic studies. An extensive group of scientists sequenced regions on chromosome 4 looking for possible genes until they found the gene. This process was rather expensive; the genome sequencing efforts are actually less expensive than this type of brute force search.

Genome sequencing is the ultimate brute force method. Efficiencies of scale, however, and improved techniques have also made genome sequencing the ultimate method for finding new genes.

Polymerase chain reaction

If the entire sequence is known, or at minimum portions of the sequence at each end of the sequence are known, the sequence of interest can be isolated by PCR. If only a portion of the sequence is known, PCR amplification may still be possible. One possible technique requires knowing the N-terminal sequence of the protein. This sequence information can then be used to predict the DNA sequence that codes for the protein. (The degenerate nature of the genetic code means that multiple DNA sequences could code for the same amino acid sequence, which limits the success of these methods.)

For PCR, two primers are necessary. In this case, the first primer would be designed to bind the 5´ end of the coding sequence based on the N-terminal sequence information. Designing the 3´ primer is more difficult. One option is to design a primer that should bind to the bacteriophage □ sequence; a bacteriophage □ particle that contains the DNA of interest may allow use of these two primers to amplify the DNA. (If the 3´ primer binds too far downstream, or is located upstream of the 5´ end, this method will not work.)