Gene Library

Ma Hongbao

Brookdale University Hospital and Medical Center, Brooklyn, New York 11212, USA
ma8080@gmail.com

Abstract: A gene library is a collection of gene clones that represents the genetic material of an organism. There are different types of gene libraries, including cDNA libraries, genomic libraries and randomized mutant libraries. The applications of these libraries depend on the source of the original DNA fragments. Generally in a gene library each DNA fragment is uniquely inserted into a cloning vector and the pool of recombinant DNA molecules is then transferred into a population of bacteria or yeast such that each organism contains on average one construct. As the population of organisms is grown in culture, the DNA molecules contained within them are copied and propagated.
[Ma Hongbao. **Gene Library.** *N Y Sci J* 2016;9(10):19-23]. ISSN 1554-0200 (print); ISSN 2375-723X (online). http://www.sciencepub.net/newyork. 4. doi:10.7537/marsnys091016.04.

Keywords: gene; library; clone; organism; DNA; molecule; copy

A gene library is a collection of gene clones that represents the genetic material of an organism. In molecular biology, a library is a collection of DNA fragments stored and propagated in a population of micro-organisms through the process of molecular cloning. There are different types of gene libraries, including cDNA libraries, genomic libraries and randomized mutant libraries. The applications of these libraries depend on the source of the original DNA fragments. Generally in a gene library each DNA fragment is uniquely inserted into a cloning vector and the pool of recombinant DNA molecules is then transferred into a population of bacteria or yeast such that each organism contains on average one construct. As the population of organisms is grown in culture, the DNA molecules contained within them are copied and propagated.

A genomic library is a set of clones that contains the entire genome of a given organism. The number of clones that constitute a genomic library depends on: (1) The size of the genome in question; (2) The insert size tolerated by the particular cloning vector system. In the most case, the genomic library is same for the tissues from different source of an organism as each cell of the body contains identical DNA. Applications of genomic libraries include: (1) Determining the complete genome sequence of a given organism; (2) Serving as a source of genomic sequence for generation of transgenic animals through genetic engineering; (3) Study of the function of regulatory sequences *in vitro;* (4) Study of genetic mutations in tissues.

A cDNA library is a collection of cDNA converted by mRNA from a particular source by the enzyme reverse transcriptase. It contains the genes that are actively transcribed in that particular source under the physiological, developmental, or environmental conditions that existed when the mRNA is purified. cDNA libraries are useful in reverse genetics, but they only represent a very small (less than 1%) portion of the overall genome in a given organism. The applications of cDNA libraries include: (1) Discovery of novel genes; (2) Cloning of full-length cDNA molecules for *in vitro* study of gene function; (3) Study of the functions of mRNAs expressed in different cells or tissues; (4) Study of alternative splicing in different cells or tissues.

A randomized mutant library is created by de novo synthesis of a gene. During synthesis, alternative nucleotides or codons are incorporated into the DNA sequence at specific positions. This results in a mixture of double stranded DNA molecules which represent variants of the original gene. These variants can then be ligated into an expression vector, individual clones can be created, and the encoded protein variants can be expressed. The expressed proteins can then be screened for variants which exhibit favorable properties.

1. Genomic library

A genomic library is the collection of the total genomic DNA from a single organism. The DNA is stored in a population of identical vectors, each containing a different insert of DNA. In order to construct a genomic library, the organism's DNA is extracted from cells and then digested with a restriction enzyme to cut the DNA into fragments of a specific size. The fragments are then inserted into the vector using DNA ligase. Next, the vector DNA is taken up by a host organism, such as E coli or yeast, with each cell containing only one vector molecule. The host cell carrying the vector is easy to amplify and retrieve for the specific clones from the library for analysis.

Libraries from organisms with larger genomes require vectors having larger inserts and fewer vector

molecules are needed to make the library. It also could choose a vector considering the ideal insert size to find a desired number of clones necessary for full genome coverage. Genomic libraries can be used for sequencing applications.

The steps for creating a genomic library from a large genome can be done as the following: (1) Extract and purify DNA; (2) Digest the DNA with a restriction enzyme. This creates fragments that are similar in size, each containing one or more genes; (3) Insert the fragments of DNA into vectors that are cut with the same restriction enzyme. DNA ligase can be used to seal the DNA fragments into the vector. This creates a large pool of recombinant molecules; (4) These recombinant molecules are taken up by a host bacterium by transformation, creating a DNA library.

After a genomic library is constructed with a viral vector, the titer of the library can be determined. A titer is a way of expressing concentration. A specific example is a viral titer, which is the lowest concentration of virus that still infects cells. Calculating the titer lets us know how many infectious viral particles are successfully created in the library. To do this, dilutions of the library are used to transfect cultures of E. coli of known concentrations. The cultures are then plated on agar plates and incubated overnight. The number of viral plaques is counted and can be used to calculate the total number of infectious viral particles in the library.

A similar method can be used to titer genomic libraries made with non-viral vectors, such as plasmids and BACs. A test ligation of the library can be used to transform E. coli. The transformation is then spread on agar plates and incubated overnight. The titer of the transformation is determined by counting the number of colonies present on the plates. These vectors generally have a selectable marker allowing the differentiation of clones containing an insert from those that do not.

In order to isolate clones that contain regions needed from a library, the library should be screened first. One method of screening is hybridization. Each transformed host cell of a library will contain only one vector with one insert of DNA. The whole library can be plated onto a filter over media. The filter and colonies are prepared for hybridization and then labeled with a probe. The target DNA can be identified by detection such as autoradiography. Another method of screening is with polymerase chain reaction (PCR). Some libraries are stored as pools of clones and screening by PCR is an efficient way to identify pools containing specific clones.

A marker gene is a gene to determine if a needed DNA sequence has been successfully inserted into an organism's DNA. There are two types of marker genes: a selectable marker and a marker for screening. A selectable marker protects the organism from a selective agent that would normally kill it or prevent its growth. In a transformation reaction, depending on the transformation efficiency, only one in a several million to billion cells may take up DNA. A selective agent is used to kill all cells that do not contain the foreign DNA, leaving only the desired cells. A screenable marker makes cells containing the gene different. There are 3 types of screening: (1) Green fluorescent protein (GFP) makes cells glow green under UV light, and it is commonly used to measure gene expression; (2) GUS assay is a method for detecting a single cell by staining it blue without using any complicated equipment; (3) Blue white screen is used in both bacteria and eukaryotic cells.

GFP is a protein composed of 238 amino acid residues (26.9 kDa) that exhibits bright green fluorescence when exposed to light in the blue to ultraviolet range. Although many other marine organisms have similar green fluorescent proteins, GFP traditionally refers to the protein first isolated from the jellyfish *Aequorea victoria*. The GFP from *A. victoria* has a major excitation peak at a wavelength of 395 nm and a minor one at 475 nm. Its emission peak is at 509 nm, which is in the lower green portion of the visible spectrum. The fluorescence quantum yield of GFP is 0.79. The GFP from the sea pansy *Renilla reniformis* has a single major excitation peak at 498 nm. GFP gene can be used as a reporter of expression in research, as the biosensor. GFP can be introduced into animals or other species through transgenic techniques, and maintained in their genome and that of their offspring.

A plasmid is a double stranded circular DNA molecule commonly used for molecular cloning. Plasmids are generally 2 to 4 kb in length and are capable of carrying inserts up to 15kb. Plasmids contain an origin of replication allowing them to replicate inside a bacterium independently of the host chromosome. Plasmids commonly carry a gene for antibiotic resistance that allows for the selection of bacterial cells containing the plasmid. Many plasmids also carry a reporter gene.

Phage λ is a double-stranded DNA virus that infects *E. coli*. The λ chromosome is 48.5kb and can carry inserts up to 25kb. These inserts replace non-essential viral sequences in the λ chromosome, while the genes required for formation of viral particles and infection remain intact. The insert DNA is replicated with the viral DNA and are packaged into viral particles. These particles are very efficient at infection and multiplication leading to a higher production of the recombinant λ chromosomes. As the smaller insert size, libraries made with λ phage may require many clones for full genome coverage.

Cosmid vectors are plasmids that contain a small

region of bacteriophage λ DNA (cos sequence) that allows the cosmid to be packaged into bacteriophage λ particles. These particles are introduced into the host cell by transduction. Once inside the host, the cosmids circularize with the aid of the host's DNA ligase and then function as plasmids. Cosmids are capable of carrying inserts up to 45kb in size.

Bacteriophage P1 vectors can hold inserts 70 – 100kb in size. They begin as linear DNA molecules packaged into bacteriophage P1 particles. These particles are injected into an E. coli strain expressing Cre recombinase. The linear P1 vector becomes circularized by recombination between two loxP sites in the vector. P1 vectors generally contain a gene for antibiotic resistance and a positive selection marker to distinguish clones containing an insert from those that do not. P1 vectors also contain a P1 plasmid replicon, which ensures only one copy of the vector is present in a cell. However, there is a second P1 replicon that is controlled by an inducible promoter that promoter allows the amplification of more than 1 copy of the vector per cell prior to DNA extraction. P1 artificial chromosomes have features of both P1 vectors and Bacterial Artificial Chromosomes. Bacterial artificial chromosomes are circular DNA molecules, usually about 7kb in length that are capable of holding inserts up to 300kb in size. Yeast artificial chromosomes are linear DNA molecules containing the necessary features of an authentic yeast chromosome, including telomeres, a centromere, and an origin of replication. Large inserts of DNA up to 2000kb can be ligated into the middle of the yeast artificial chromosome so that there is an arm of the YAC on either side of the insert.

Vector selection requires one to ensure the library made is representative of the entire genome. Any insert of the genome derived from a restriction enzyme should have an equal chance of being in the library compared to any other insert. Furthermore, recombinant molecules should contain large enough inserts ensuring the library size is able to be handled conveniently. This is particularly determined by the amount of clones needed to have in a library. The amount of clones to get a sampling of all the genes is determined by the size of the organism's genome as well as the average insert size. Increasing the insert size by choice of vector will be good for fewer clones needed to represent a genome.

After a library is created, the genome of an organism can be sequenced to elucidate how genes affect an organism or to compare similar organisms at the genome-level. Genes can be isolated through genomic libraries and used on human cell lines or animal models to further research.

2. cDNA library

A cDNA library is a combination of cloned cDNA fragments inserted into a collection of host cells, which together constitute some portion of the transcriptome of the organism. cDNA is produced from fully transcribed mRNA found in the nucleus and therefore contains only the expressed genes of an organism. In eukaryotic cells the mature mRNA is already spliced so that the cDNA produced lacks introns and can be readily expressed in a bacterial cell. While information in cDNA libraries is a powerful and useful tool since gene products are easily identified, the libraries lack information about enhancers, introns, and other regulatory elements in a genomic DNA library.

cDNA is created from a mature mRNA from a eukaryotic cell with the use of reverse transcriptase. In eukaryotes, a poly-(A) tail distinguishes mRNA from tRNA and rRNA and can therefore be used as a primer site for reverse transcription. This has the problem that not all transcripts, such as those for the histone, encode a poly-A tail.

The mRNA can be extracted and purified from several methods, such as Trizol extraction and column purification. Trizol method uses the RNA extract reagent to precipitate whole RNA. Column purification is done by using oligomeric dT nucleotide coated resins where only the mRNA having the poly-A tail will bind and the rest of RNAs are eluted out. The mRNA is eluted by using eluting buffer and some heat to separate the mRNA strands from oligo-dT.

Once mRNA is purified, *oligo-dT* is tagged as a complementary primer which binds to the poly-A tail providing a free 3'-OH end that can be extended by reverse transcriptase to create the complementary DNA strand. Now, the mRNA is removed by using an RNase enzyme leaving a single stranded cDNA (sscDNA). This sscDNA is converted into a double stranded DNA with the help of DNA polymerase. However, for DNA polymerase to synthesize a complementary strand a free 3'-OH end is needed. This is provided by the sscDNA itself by generating a *hairpin loop* at the 3' end by coiling on itself. The polymerase extends the 3'-OH end and later the loop at 3' end is opened by the scissoring action of S_1 *nuclease*. Restriction endonucleases and DNA ligase are then used to clone the sequences into bacterial plasmids. The cloned bacteria are then selected, commonly through the use of antibiotic selection. Once selected, stocks of the bacteria are created which can later be grown and sequenced to compile the cDNA library.

cDNA libraries are commonly used when reproducing eukaryotic genomes, as the amount of information is reduced to remove the large numbers of non-coding regions from the library. cDNA libraries are used to express eukaryotic genes in prokaryotes.

Prokaryotes do not have introns in their DNA and therefore do not possess any enzymes that can cut it out during transcription process. cDNA does not have introns and therefore can be expressed in prokaryotic cells. cDNA libraries are most useful in reverse genetics where the additional genomic information is of less use, and it is useful for subsequently isolating the gene that codes for that mRNA.

cDNA molecules can be cloned by using restriction site linkers. Linkers are short, double stranded pieces of DNA about 8 to 12 nucleotide pairs long that include a restriction endonuclease cleavage site, e.g. BamHI. Both the cDNA and the linker have blunt ends which can be ligated together using a high concentration of T4 DNA ligase. Then sticky ends are produced in the cDNA molecule by cleaving the cDNA ends with the appropriate endonuclease. A cloning vector is then also cleaved with the appropriate endonuclease. Following sticky end ligation of the insert into the vector the resulting recombinant DNA molecule is transferred into *E. coli* host cell for cloning.

cDNA library lacks the non-coding and regulatory elements found in genomic DNA. Genomic DNA libraries provide more detailed information about the organism, but are more resource-intensive to generate and maintain.

3. Randomized mutant libraries

A randomized mutant library is created by de novo synthesis of a gene. During synthesis, alternative nucleotides or codons are incorporated into the DNA sequence at specific positions. This results in a mixture of double stranded DNA molecules which represent variants of the original gene. These variants can then be ligated into an expression vector, individual clones can be created, and the encoded protein variants can be expressed.

The expressed proteins can then be screened for variants which exhibit favorable properties. Typically the properties that are to be improved by screening a randomized mutant library are the binding affinity of antibodies or other protein-protein interactions, the activity of enzymes, or the stability of a protein. Multiple cycles of creating gene variants and screening the expression products are typically involved in directed evolution experiments.

The nucleotide sequences of interest are preserved as inserts to a plasmid or the genome of a bacteriophage that has been used to infect bacterial cells.

cDNA is double-stranded DNA synthesized from a single stranded RNA template in a reaction catalyzed by the enzyme reverse transcriptase.

cDNA is most often synthesized from mature mRNA using the enzyme reverse transcriptase. This enzyme, which naturally occurs in retroviruses, operates on a single strand of mRNA, generating its complementary DNA based on the pairing of RNA base pairs (A, U, G and C) to their DNA complements (T, A, C and G, respectively). cDNA is often used in gene cloning or as gene probes or in the creation of a cDNA library.

A cloning vector is a small piece of DNA, taken from a virus, a plasmid or a cell of a higher organism, which can be stably maintained in an organism, and into which a foreign DNA fragment can be inserted for cloning purposes. The vector therefore contains features that allow for the convenient insertion or removal of DNA fragment in or out of the vector, for example by treating the vector and the foreign DNA with a restriction enzyme that creates the same overhang, then ligating the fragments together. After a DNA fragment has been cloned into a cloning vector, it may be further subcloned into another vector designed for more specific use.

All commonly used cloning vectors in molecular biology have key features necessary for their function, such as a suitable cloning site and selectable marker. Others may have additional features specific to their use. For reason of ease and convenience, cloning is often performed using *E. coli*. Thus, the cloning vectors used often have elements necessary for their propagation and maintenance in *E. coli*, such as a functional origin of replication. The ColE1 origin of replication is found in many plasmids. Some vectors also include elements that allow them to be maintained in another organism in addition to *E. coli*, and these vectors are called shuttle vector. ColE1 replication begins at the origin 555bp upstream from this point, RNA polymerase-initiated transcription of RNAII which acts as a pre-primer and begins the synthesis of the leader strand. The transcript folds into a secondary structure which stabilizes the interaction between the nascent RNA and the origin's DNA. This hybrid is attacked by RNase H, which cleaves the RNA strand, exposing a 3' hydroxyl group. This allows the extension of the leading strand by DNA Polymerase I. Lagging strand synthesis is later initiated by a primase encoded by the host cell.

Molecular cloning is a set of experimental methods in molecular biology that are used to assemble recombinant DNA molecules and to direct their replication within host organisms. The use of the word *cloning* refers to the fact that the method involves the replication of one molecule to produce a population of cells with identical DNA molecules. Molecular cloning generally uses DNA sequences from two different organisms: the species that is the source of the DNA to be cloned, and the species that will serve as the living host for replication of the recombinant DNA. Molecular cloning methods are

central to many contemporary areas of modern biology and medicine.

Note:

Most of the contents in this article are collected from the Internet and other opened literature sites, which are offered to the readers as a convenient reference as the aim.

References

1. Baidu. http://www.baidu.com. 2016.
2. Google. http://www.google.com. 2016.
3. https://en.wikipedia.org/wiki/Gene_library. 2016.
4. https://en.wikipedia.org/wiki/Genomic_library. 2016.
5. https://en.wikipedia.org/wiki/Library_(biology). 2016.
6. https://en.wikipedia.org/wiki/Marker_gene#Marker_for_screening. 2016.
7. https://en.wikipedia.org/wiki/Titer. 2016.
8. Ma H, *Chen G*. Stem cell. The Journal of American Science 2005;1(2):90-92.
9. Ma H, Cherng S. *Eternal Life and Stem Cell.* Nature and Science. 2007;5(1):81-96.
10. Ma H, Cherng S. Nature of Life. Life Science Journal 2005;2(1):7-15.
11. Ma H. The Nature of Time and Space. Nature and science 2003;1(1):1-11. Nature and science 2007;5(1):81-96.
12. National Center for Biotechnology Information, U.S. National Library of Medicine. http://www.ncbi.nlm.nih.gov/pubmed. 2016.
13. Wikipedia. The free encyclopedia. http://en.wikipedia.org. 2016.

10/3/2016