

Module 2 – Construction and Screening of DNA Libraries

Resources

General overview

http://www.rvc.ac.uk/Extranet/DNA_1/0_intro.htm

Vectors

<http://www.web-books.com/MoBio/Free/Ch9A4.htm>

Plasmid vectors

NEB description of M13 derived

<http://www.neb.com/nebecomm/products/productN4019.asp>

Overview of cloning into plasmids

<http://www.accessexcellence.org/RC/VL/GG/plasmid.html>

Bacterial artificial chromosomes

<http://www.msstate.edu/research/mgel/newbac.htm>

http://www.msstate.edu/research/mgel/pubs_pdf/pete00.pdf

<http://www.nal.usda.gov/pgdic/Probe/v5n2/chromlib.html>

Yeast artificial chromosomes

<http://www.accessexcellence.org/RC/VL/GG/YAC.html>

General Background

A DNA library is a collection of fragments ligated into a cloning vector and stored individually in a host cell. Libraries are important resources for molecular cloning and genome sequencing projects. They can be formed from genomic DNA or messenger RNA (cDNA), insert size can vary over several orders of magnitude (100's of bp to Mb), and they can be maintained in yeast or, more commonly, *E. coli*.

Construction of a library follows a series of steps that you have all ready been exposed to in the first Module of 604.03; cutting DNA, ligation into a vector, and transformation into a host. In constructing the library, preparation of the insert is the first step. If the goal of the library is to represent the entire genome, DNA is extracted from somatic tissue. If the library is to represent only the coding region of the genome that is expressed in a particular tissue, organ, or stage of life, mRNA is extracted and enzymatically copied into DNA (cDNA). The foundation of many “genomics” projects is high-throughput sequencing of cDNA libraries to generate “Expressed Sequence Tags”. Expressed Sequence Tags are partial, single-pass sequences from either end of a cDNA clone. The EST strategy was developed to allow rapid identification of expressed genes found in different tissues and conditions by sequence analysis. Genomic DNA libraries provide the basis of sequencing projects that aim to sequence the complete genome, not just the expressed sequences. Complete genome sequencing projects follow

different strategies. Genome sequencing strategies include “shotgun” approaches that involve the sequencing of many small insert genomic DNA libraries developed in plasmid vectors. Minimum tiling strategies involve the development of overlapping contiguous sequence (contigs) in large insert bacterial artificial chromosomes (BAC) or yeast artificial chromosome vectors (YAC). Examples follow.

Vector Choice

Cloning vectors are extra-chromosomal DNA which can replicate in the host. Vectors have physical properties that facilitate purification separately from the chromosomal DNA of the host. Vector choice for library construction will vary by the use and goal of the library. The most important criteria is the desired size of the insert.

Vector systems and approximate insert sizes for libraries

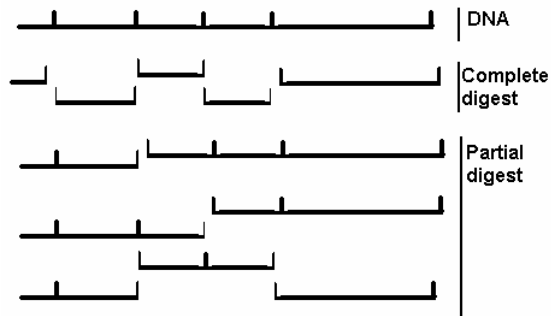
| Vector | | Insert Size | Use | Host |
|---------------------------------|-----|-----------------|-------------|----------------------|
| plasmid | | 100 bp to 5 Kb | DNA or cDNA | <i>E. coli</i> |
| phage lambda | λ | 5 Kb to 20 Kb | DNA or cDNA | <i>E. coli</i> |
| cosmid | | 20 Kb to 50 Kb | DNA | <i>E. coli</i> |
| bacterial artificial chromosome | BAC | 20 Kb to 150 Kb | DNA | <i>E. coli</i> |
| yeast artificial chromosome | YAC | 50 Kb to 1 Mb | DNA | <i>S. cerevisiae</i> |

Preparation of DNA

Genomic DNA isolation is tailored to the desired insert size. All DNA isolation procedures involve steps to disrupt tissue (mechanically or chemically) and steps to separate DNA from contaminating proteins and carbohydrates. For large-insert libraries (BAC and YAC), DNA isolation is often done with tissue, protoplasts, cells or nuclei imbedded in agarose to protect the DNA from sheering (see supplemental protocol).

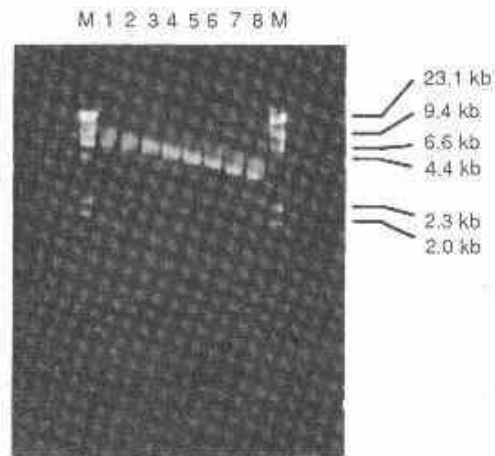
Size selection follows DNA isolation and is an important step to insure that genomic libraries consist largely of inserts that are of the desired size range. DNA is “cut” to size by controlled sheering or enzymatic digestion. The restriction endonuclease *Sau3a* (or *Mbo I*) is often used for controlled partial digestion of DNA. *Sau3a* is used because the recognition site, 5'-GATC-3', and digestion produce overhanging sticky ends that are compatible with *BamHI* sites found in many cloning “polylinkers”. Following partial digestion, selection of a specific size range can be facilitated by centrifugation over sucrose gradients and collection of an appropriate fraction, or isolation of a specific size range from agarose gels.

Partial digestion



Partial digestion with *Sau3a* (sites indicated by vertical hatch marks) is used to create a series of overlapping and random fragments of DNA with an overhanging sticky end 5'-GATC-3'.

Size Selection



Fractions from a sucrose gradient containing *Sau3a* partially digested DNA are separated on an agarose gel. Fractions 1 and 2 contain DNA between 5 and 10 Kb in size.

Cloning inserts into vectors

Ligation of size selected insert DNA into a vector is followed by transformation into the host. You have gained experience with these steps in Module 1. Methodology for phage (virus-based) vectors differs slightly from that used for plasmid-based vectors. Methods used for phage vectors are described in supplemental protocol xxx.

Library Quality and Size

The quality of a library is gauged by the average insert size and the number of clones. The goal of most genomic DNA libraries is to have a sufficient number of cloned fragments to cover the genome 5 to 10 times. Genome coverage several fold is necessary to insure a high probability that any single clone or fragment of DNA is represented at least once. The equation $N = \ln(1-P)/\ln(1-f)$ can be used to estimate the number of clones in a library that must be screened to identify a specific fragment or gene in the library.

P is a chosen probability of desired success (usually 0.95 – 0.999)

f is the expected frequency of an event. For libraries, f is the fraction of the genome represented by the average size clone. For a tomato phage library with an average insert size of 15 Kb, f is 15 Kb/ 950,000 Kb (the genome size of tomato).

Table of number of clones (N) to screen for probabilities of success ranging between 95% and 99.9% of identifying a single positive clone.

| Probability | Insert size | N |
|-------------|-------------|---------|
| 0.95 | 2 Kb | 1348078 |
| 0.99 | | 2072324 |
| 0.999 | | 3108486 |
| 0.95 | 15 Kb | 179742 |
| 0.99 | | 276308 |
| 0.999 | | 414462 |
| 0.95 | 20 Kb | 134807 |
| 0.99 | | 207230 |
| 0.999 | | 310846 |
| 0.95 | 100 Kb | 26960 |
| 0.99 | | 41444 |
| 0.999 | | 62166 |

Again, using tomato as an example, the importance of fold coverage can be illustrated. Tomato has an estimated genome size of 950 MB. We therefore need to screen 207,230 clones of average insert size 20 Kb to have a 99% probability of identifying a specific clone of interest. It is important to note that screening 207,230 represents a 4.6 fold coverage of tomato genome. In order to increase our probability of success to 99.9% we need to screen the equivalent of 6.9 genomes. A decrease in average insert size to 19 Kb would require that we screen ~10,000-20,000 more clones.

Screening libraries for specific clones

It is increasingly common for libraries to be developed as a “community resource”. Typically such resource libraries are stored and distributed by genome centers and/or companies. The availability of these resources to the public has been facilitated by information science (in the case of cDNA information) and automation for the replication, distribution, and screening of large-insert genomic DNA libraries. This fundamental change in “how biology is done” has occurred over the last 5-10 years. Examples of the genome resources for plants follow. Similar, and even more expansive resources are available for animals and microbes.

Resources for cDNA libraries

Information from cDNA libraries is now typically “retrieved” from databases. As introduced in the “Genomics and Database Mining” exercise between Module 1 and Module 2. Rather than screening a cDNA library for a specific clone, our first option is to perform a database search using either the National Center for Biotechnology Information < <http://www.ncbi.nlm.nih.gov/> > database or species specific indices of genes < <http://www.tigr.org/tdb/tgi/> >.

Partial list of cDNA libraries for Tomato

< http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=tomato >

| Cat# | #ESTs | Library Name # |
|--------|-------|--|
| 8LB | 163 | tomato root subtraction cDNA library |
| CAR | 28 | tomato thick-juice fruit subtracted by cDNAs of thin-juice fruit |
| CAS | 23 | tomato thin-juice fruit subtracted by cDNAs of thick-juice fruit |
| CN1 | 1521 | normalized cDNA library from ripening tomato pericarp |
| T1005 | 1049 | tomato shoot |
| T10227 | 5425 | wild tomato pollen |
| T10284 | 5204 | tomato crown gall |
| T10304 | 9124 | tomato shoot/meristem |
| T10393 | 7010 | tomato flower library from a mixture of developmental stages |
| T1045 | 9950 | tomato ovary |
| T1048 | 591 | tomato seed |
| T10600 | 8026 | tomato suspension culture, untreated |
| T1079 | 5317 | tomato pseudomonas susceptible |
| T1080 | 5195 | tomato pseudomonas resistant |
| T1207 | 14160 | tomato callus tissue |
| T1297 | 9539 | tomato mixed elicitor |
| T1356 | 5354 | tomato mature green fruit |
| T1391 | 3895 | tomato red ripe fruit T1437 |

Note the diversity of tissues used to generate these libraries and the large number of “Expressed Sequence Tags (ESTs)” generated for several libraries. ESTs are single-pass sequences generated using high-throughput approaches. This sequence data is then archived in public databases.

Resources for Bacterial Artificial Chromosome (BAC) libraries

Table: A partial list of the 87 plant, animal, and microbial BAC libraries available from the Clemson University Genome Center.

| Library | Common Name | Genus | Species |
|---------|--------------|--------------|------------|
| BAC | Soybean | Glycine | tomentella |
| BAC | Soybean | Glycine | max |
| BAC | Soybean | Glycine | max |
| BAC | Cotton | Gossypium | hirsutum |
| BAC | Cotton | Gossypium | hirsutum |
| BAC | Barley | Hordeum | vulgare |
| BAC | Tulip poplar | Liriodendron | tulipifera |
| BAC | Tomato | Lycopersicon | esculentum |
| BAC | Tomato | Lycopersicum | esculentum |

<http://www.genome.arizona.edu/orders/>

<https://www.genome.clemson.edu/cgi-bin/orders?page=serviceHome&service=bacrc>