

Lecture 6 - DNA Libraries and Mutagenesis (AMG text pp. 59-78)
September 6, 2001

Overview of DNA Libraries

The two basic types of DNA libraries; Genomic libraries and cDNA libraries.

The most common strategy for library construction is to generate a representative library that contains a true representation of all the DNA sequences present in the starting material. Many types of DNA libraries are available commercially, but you still need to know how to find your gene of interest by library screening.

The total number of primary recombinants N that need to be screened to achieve a 99% probability P that a clone will be found in a library, depends on the average insert size (use 20 kb for a lambda phage genomic library), and the size of the genome being screened (the human genome is 3 billion base pairs), using the term f as the fraction of the genome in each recombinant. This relationship can be represented by:

$$\begin{aligned} N &= \ln(1-P)/\ln(1-f) \\ N &= \ln(1-0.99)/\ln(1-[20 \times 10^3/3 \times 10^9]) \\ N &= -4.6/\ln(1-[6.6 \times 10^{-6}]) \\ N &= -4.6/-6.6 \times 10^{-6} = 6.9 \times 10^5 \end{aligned}$$

It is common to screen between 500,000 and 1 million recombinants to be relatively sure that a 20 kb segment of the genome will be found.

What characteristic of most genomes makes this calculation less than accurate?

What could you do if you failed to isolate a desired fragment from a given library?

Why would this equation not work for calculating the number of recombinants needed to screen a cDNA library?

A genomic DNA library is constructed using overlapping DNA fragments that have been generated using a mechanical shearing strategy, or partial digestion with a frequently cutting enzymes such as *Sau3A*, to create multiple segment endpoints within the population of fragments.

The term "library walking" refers to the reiterative process of using terminal sequences of captured clones to rescreen the library in an attempt to isolate a new set of overlapping clones. This will only work if the library is constructed in such a way that random fragments are generated and are stably propagated in the library.

A contiguous map called a "contig" can be generated using the minimum number of overlapping clones that contain extended areas of sequence complementarity. A low resolution map can be made using landmarks such as rare cutting restriction enzymes or repetitive elements. High throughput DNA sequencing is performed to confirm the overlaps.

A phage library is screened by physically separating 500,000 plaques across 20-25 large petri plates. The exact location of each plaque on the plates needs to be carefully monitored so that a positive signal on an X-ray film can be lined up with the master plate. This can be done a number of ways and is analogous to spreading out 500,000 jigsaw puzzle pieces on a large table and hunting for one specific piece (Waldo's red striped hat!). Automated procedures are being developed to replace library plating which are based on affinity methods or arrayed recombinants. These services are available commercially but are expensive.

Screening a Lambda library using conventional laboratory methods.

The first step in this low tech approach is to spread all the phage out by "plating the library" using large plates and an accurate count of phage titer:

The second step is to "lift the phage" off the top agar using nitrocellulose filters that are carefully laid down on the master plate, registered with a needle puncture, and then carefully removed to capture a small fraction of the phage in the plaque. About 20% of the phage particles in the plaque adhere non-covalently to the nitrocellulose, leaving behind ~80% of the viable phage for propagation later.

What is the purpose of laying down a second filter onto the same plate to create a duplicate lift?

How do you prevent the second filter from misaligning with the master plate?

The third step is to denature the DNA and covalently link it to the nitrocellulose by UV irradiation. The filters are then processed by hybridization to the labeled probe. The developed film is re-aligned with the master plates and agar plugs are removed from the area of the plate corresponding to the positive signal. This last step is dependent on accurate registration marks and well preserved master plates.

The final step in library screening is to purify the candidate phage by limiting dilution. It is at this step of plaque purification where positive signals are confirmed and false-positives phage are discarded. The purified λ phage in the final stage are called "plaque pure", meaning that 100% of the phage in the tertiary plating hybridize to the probe.

Why is it necessary to further purify the phage in this primary pick?

What are two explanations for "faint" signals on the film that repeatedly test positive, but do not get any darker on repeated purifications?

What would be the meaning of intense hybridization signals from >100 plaques when screening a genomic library with an uncharacterized probe? When screening a cDNA library?

Restriction Enzyme Mapping

One of the most difficult tasks in this process is preliminary characterization of purified phage. Restriction enzyme mapping, cross-hybridization and DNA sequencing are approaches that are often used to get a first glance at isolated inserts.

Restriction enzyme mapping can be combined with subcloning to permit high throughput DNA sequencing and bioinformatic analyses.

Once a cloned insert has been sequenced, the next step is to determine function. This could mean investigating protein coding sequence or transcriptional regulatory regions. Either way, the DNA will need to be manipulated by deletion or site-directed mutagenesis to assess function.

In Vitro Mutagenesis

Deletion mutagenesis using restriction enzymes and exonucleases is one way to create a set of DNA fragments for functional analysis. Based on the restriction map of a plasmid DNA insert, it is possible to design a strategy to remove nucleotides from the termini, or within the cloned segment, using intramolecular ligation reactions. BglII and BamHI digestion result in compatible termini (GATC). Exonuclease III (Exo III) degrades one strand of double-stranded DNA from the 3' end, but only if the DNA termini contains either a 5' overhang or a blunt end, but not an extended single-strand 3' overhang.

Linker Scanning mutagenesis was first described by Steve McKnight and is a systematic approach to base pair substitutions across a segment of DNA. It has most often been used to map transcriptional regulatory regions. In this strategy, a collection of 5' and 3' deletants are created and the termini are ligated to an oligonucleotide linker. Based on the DNA sequence of individual deletants, paired combinations are chosen and used to create a new DNA fragment in which the linker sequence precisely replaces 8 bp of the original sequence without altering the spacing of surrounding nucleotides.

Why would linker scanning NOT be a good strategy for investigating protein structure and function?

A key strategy in oligonucleotide-directed mutagenesis methods is to increase the relative number of mutated plasmids in the pooled population. The dUTP incorporation strategy shown below relies on the degradation and repair of template DNA has been "marked" with deoxyuridine prior to the in vitro DNA synthesis step. Thomas Kunkel developed the dUTP incorporation strategy by taking advantage of an E. coli strain that contains defects in genes responsible for preventing dUTP incorporation into DNA.

Oligonucleotide-directed mutagenesis is an efficient method to generate specific nucleotide alterations in cloned DNA. dUTP incorporation into M13 DNA is done by using a phagemid vector to produce single strand DNA in an E. coli strain defective in

the two enzymes uracil-N-glycosylase (ung) and dUTPase (dut). The mutagenic oligo is annealed to uracil-containing template DNA (U), and in vitro DNA synthesis is performed to generate double-stranded DNA.

Following transformation of ung⁺, dut⁺ E. coli, the template strand is degraded and the surviving double strand phagemid plasmid is isolated and sequenced. In this example, the mutant oligo was used to engineer a novel EcoRI site into the DNA insert.

Alanine Scanning Mutagenesis

The term “alanine scanning” refers to the systematic substitution of alanine codons at selected positions within a protein coding region using site directed mutagenesis. Alanine scanning mutagenesis is used to systematically substitute specific amino acids with alanine by altering the codon sequence using standard site-directed in vitro mutagenesis procedures.

What is the advantage of using site directed mutagenesis rather than deletion analysis to study protein structure and function?

In what type of gene analysis would be best suited for an initial study using terminal deletion analysis, followed by high resolution site directed mutagenesis (hint: what part of a gene is required for transcriptional regulation)?

Why is alanine a good amino acid to use for replacement in a protein coding sequence?

What other amino acid would be another good choice based on its chemical structure?

What amino acids would be good choices if you are trying to study the effects of charge on protein function?

What amino acids might you want to change if you are studying the effect of phosphorylation on protein function?