Reverse genetics - From protein or RNA to gene

Up until now, we've been following the classical genetic approach to creating an inventory of components, the path from mutation to gene. With the gene in hand, information from the DNA sequence can be obtained. In addition, having the gene enables a wide variety of other experimental approaches. Note that identification of a gene by finding a recessive mutation does not always mean that one even knows the phenotype expected from loss of function of the gene product. With the cloned DNA in hand, there are methods to construct knockout mutations (see the next chapter), or to see what happens if the gene is overexpressed or expressed in different times or in different tissues. The cloned gene can also be starting material for targeted or site-directed mutagenesis. The expression of the gene can be examined by creating fusions between regulatory sequences and reporter genes. The protein product, if there is one, can be overexpressed and purified in large amounts for studies ranging from structural studies to biochemical reconstitution.

Another path to the same ends is based on isolating genes by looking for sequences that encode a specific protein or RNA. In this case, one has a reason to suspect that a particular gene product is involved in the process you are interested in, and you want to test that idea by looking for the phenotypes associated with changing the level or activity of that gene product in the cell. The probable cause to implicate a particular protein can range from biochemical evidence that it has a particular activity, to the observation that the protein is present at the right time or place to be involved, to evidence for a particular function based on looking at a DNA sequence.

Protein to gene

Starting with a purified protein, there are two major approaches to isolating a clone corresponding to the gene. In the first class of approach, one looks for a detectable property of the protein when it is expressed in *E. coli*. Since *E. coli* does not have eukaryotic introns, screens based on protein expression involve screening cDNA libraries, where a population of cDNAs is cloned into a vector where an *E. coli* promoter drives expression of an mRNA encoding the desired polypeptide sequence in *E. coli*. The cDNA is cloned into a bacterial gene, such as a fragment of *lacZ*, so that the *lacZ* translation start assures that the sequence will also be translated.

The a cDNA library is expressed from a plasmid or λ vector so that each colony or plaque on a plate contains a different clone. A replica of the plate is made on a filter which is treated so that proteins stick to the filter. Detection of the expressed proteins can be accomplished in several ways. The most general is to use antibodies to the protein of interest. The filter is first blocked with nonspecific protein and then incubated with a primary antibody raised against the purified protein. Primary antibodies are usually prepared by immunizing rabbits or mice. Secondary antibodies can be purchased that are coupled to enzymes such as horseradish peroxidase or alkaline phosphatase. An enzyme-linked secondary antibody that recognizes species specific epitopes in rabbit or mouse immunoglobulins is used to treat the filters

Generating an antibody probe generally requires a purified protein to use as an antigen. Other kinds of probes can be used even if the desired protein has not been purified at all. For example, genetic analysis of a promoter region might identify a sequence that is involved in regulating a gene. That sequence is presumably recognized by some protein in some cell type. Binding to a short DNA fragment or synthetic oligonucleotide containing the sequence can often be detected in unfractionated nuclear extracts by methods such as gel mobility shifts or filter binding. When the nuclear extract is fractionated on a gel, blotted to a filter and then treated with conditions that can refold denatured proteins, one of the proteins from the gel can sometimes bind to a specific nucleic acid probe. When the probe is a DNA fragment, the experiment is called a Southwestern blot; with an RNA probe, it is a Northwestern blot. Northwestern and Southwestern blotting can also be used to probe filters from colonies or plaques generated by expression libraries.

Northwestern and Southwestern approaches rely on the ability of some fraction of the protein to renature into a form that can specifically bind the probe. In addition to potential problems with proteins failing to refold, multisubunit proteins will often lack the appropriate partners to reconstitute their normal activities.

The second class of approaches requires purified protein, but does not require an antibody or an activity. A partial amino acid sequence of the protein is obtained by conventional protein chemistry, and the possible DNA sequences that could encode the protein sequence are inferred from the genetic code. An degenerate oligonucleotide is synthesized such that some fraction will hybridize to the real gene. In the original method, the oligo is radioactively labelled and used as a probe. Since the development of PCR, it has been found that products can often be amplified between two degenerate oligonucleotides. The PCR product, being longer and more specific, makes a more efficient probe than an oligo alone.

The degenerate oligo approach requires protein sequence. N-terminal sequencing by Edman degradation is the easiest way to get enough amino acid sequence to design a probe. However, sometimes the sequence is unsuitable or the N-terminus is blocked. In these cases, internal peptide sequences are obtained. In practice, it seems that the degenerate oligo approach does not always work, and often several oligos from different parts of the protein have to be tried in order to find the right clone. Advances in protein sequencing are helping to make this approach easier.

Degenerate primers can even be synthesized to match a protein sequence motif common to a family of proteins, rather than as a match to a specific purified protein. For example, two-component regulators are a common family of regulatory factors in prokaryotes. These proteins are involved in regulating different cellular functions, but share sequences based on the activity of one component as a histidine protein kinase, and of the other component as a substrate. Degenerate PCR primers designed to match the consensus sequences of conserved motifs have been used to isolate genes encoding two-component regulators in a variety of bacteria.

RNA to gene

Genes that are expressed in a specific cell type are often useful in understanding the molecular basis for differentiation. It is assumed that the genes that define a cell's identity as belonging to a particular tissue or stage in the cell cycle are only transcribed in the cell type of interest. For example, hemoglobin is synthesized in red blood cells, and hemoglobin mRNA will be absent in other cell types.

There are two basic approaches to finding the genes that are expressed only under a certain kind of condition. In the first approach, you identify a probe that is specific for a particular RNA with the desired expression pattern. This probe is then used to find a specific clone in a library that represents as many genes as possible. The second approach involves making a library that is enriched for genes that are expressed in a certain cell type.

Finding a DNA probe that is specific for a specific cell type is very similar to finding a DNA polymorphism for map-based cloning. Methods analogous to RAPDs or AFLPs can be used with the cDNA as a source of material instead of genomic DNA. One can think of the cDNAs made from the RNA in different cell types as being polymorphic. Markers that are present in one cell type and not in another represent differences in the mRNA content instead of the genomic DNA. In other words, if a PCR product from RAPD or AFLP analysis is present in the cDNA made from mRNA found in one tissue, and it is absent in the cDNA made from mRNA from another tissue, that PCR product represents a gene whose mRNA is present in the first tissue and not in the second.

Applying high volume DNA marker methods to cDNA has the same basic advantages as applying them to mapping. Reactions can be run using different combinations of primers so that comparisons can be made with many different markers in parallel, and the reactions can be optimized for use on automated sequencers. When high volume marker methods are applied to the problem of finding genes whose expression changes in different tissues or under different environmental conditions, the experiment is called **differential display**.

As with map based cloning, the PCR product that is specific for a differentially expressed mRNA can be recovered from a gel and reamplified to use as a probe. The probe can then be used to find a cDNA or genomic DNA clone for the differentially expressed gene.

The second approach is to create what is called a **subtraction library**. Subtraction libraries were used to find genes that were expressed in T cells, but were not expressed in B cells, and to find male-specific genes. The description of how this is done that follows is based on the discussion of subtraction libraries in Current Protocols in Molecular Biology (aka the Red Book). References for the methods described can be found there (XX at least in this year's version of this handout). The idea behind subtraction libraries is to purify tissue-specific cDNAs away from undesired cDNAs.

One method for preparing a subtractive libraries is shown in Figure 7-1. Double stranded cDNA is synthesized from two cell types, A and B. The cDNA from cell type A is treated with a specific restriction enzyme that generates sticky ends allow it to be ligated into the cloning vector of choice. The cDNA from cell type B is also digested with the same enzyme, but is treated so that the ends of each fragment are blunt-ended.

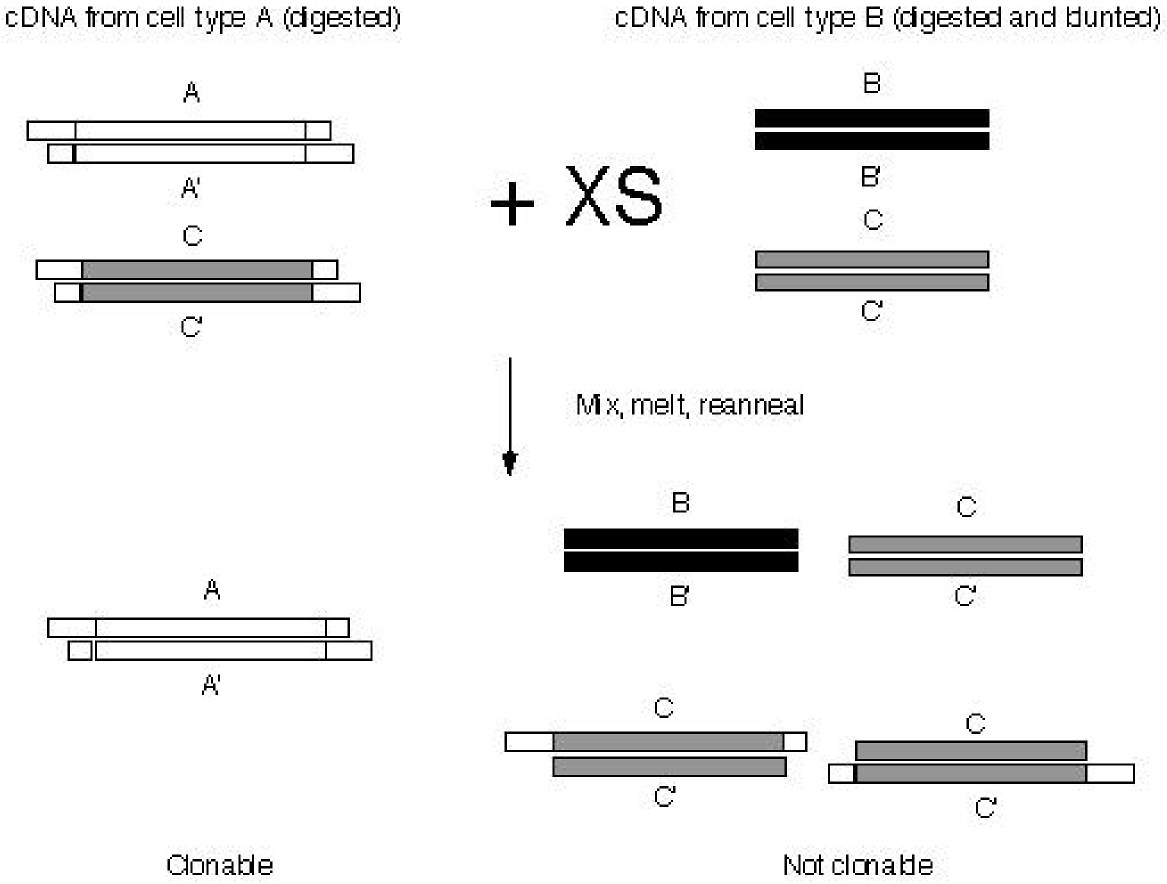
cDNA from cell type A (digested)
cDNA from cell type B (digested and blunted)
A
A'
C
C'
+ XS
B
B'
C
C'
Mix, melt, reanneal
B
B'
C
C'
A
A'
C
C'
C
C'
Clonable
Not clonable

Figure 7-1 Preparing inserts for subtractive cDNA libraries. A and A’ indicate the two strands from cDNAs that are specific to cell type A. B and B’ are specific to cell type B, and C and C’ represent shared sequences.

The two cDNA pools are mixed so that there is a molar excess of the cDNA from cell type B. The DNA is heated to denature it, and then cooled to allow the DNA molecules to rehybridize. The A-specific and B-specific cDNAs will reform double stranded DNA with molecules indistinguishable from their original partners. However, any cDNA that comes from an mRNA that is expressed in both cell types can form heteroduplexes with cDNA that originated from mRNA from the other cell type. Since the cDNA from cell type B is in excess, the shared cDNA molecules from cell type A will almost always rehybridize to a molecule from the B pool, which does not have the sticky ends from the restriction digest on its ends. These heteroduplex molecules are now unable to ligate into the cloning vector. The A-specific sequences will rehybridize so that the sticky ends are reformed, and will be able to ligate into the vector. This protocol can be modified by digesting the cDNA from cell type B with restriction enzymes that will cut the DNA into smaller fragments. This increases the molar concentration of potential partners for the common sequences without increasing the mass of B DNA needed.

Typically, this kind of procedure enriches for tissue-specific cDNAs. Clones corresponding to differentially expressed mRNAs usually comprise 20-50% of the library. The other clones will tend to be from mRNAs that are expressed at high levels in both cell types. Thus, any individual clone should be rechecked to make sure that it corresponds to something that is expressed preferentially in the expected cell type.

Note that this kind of subtractive library will allow isolation of cDNA fragments that correspond to differentially expressed genes. However, the cloned fragment can be used as a hybridization probe to isolate a clone from a cDNA library with more full length inserts, or from a genomic library.