Molecular Biology II: Vector Types and their Applications; Construction and Screening of cDNA and Genomic Libraries.

I. Learning objectives.

A. Compare and contrast different types of vectors and describe their uses.

B. Diagram the steps in cDNA and genomic library construction. Include representations of primers, templates, enzymes and products as appropriate.

C. List the uses of DNA libraries.

D. Explain how specific DNA clones can be obtained from DNA libraries.

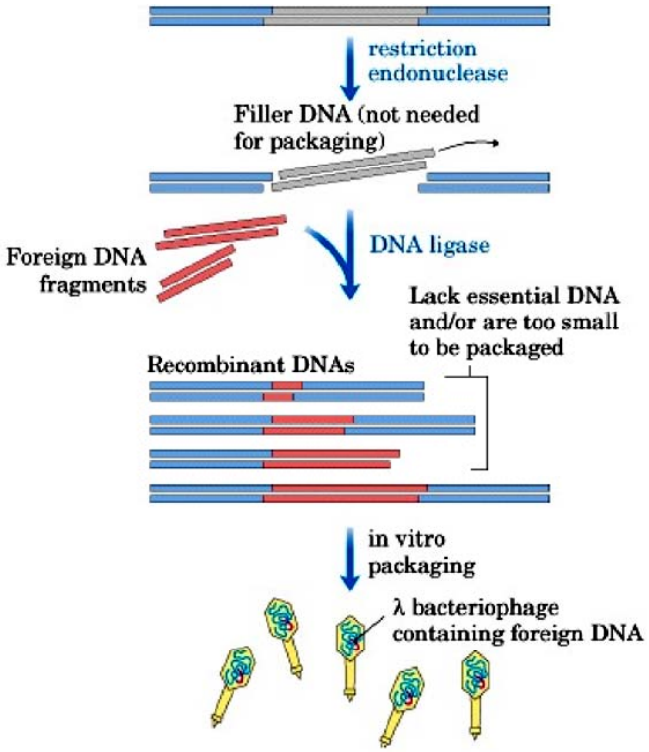
E. List the determinants of DNA hybridization.

F. Textbook reading for reference and clarification. — Lehninger, Chapter 29; Gelehrter, Chapter 5.

II. Types of vectors.

A. **plasmids** (accommodate DNA as large as 10 kb).

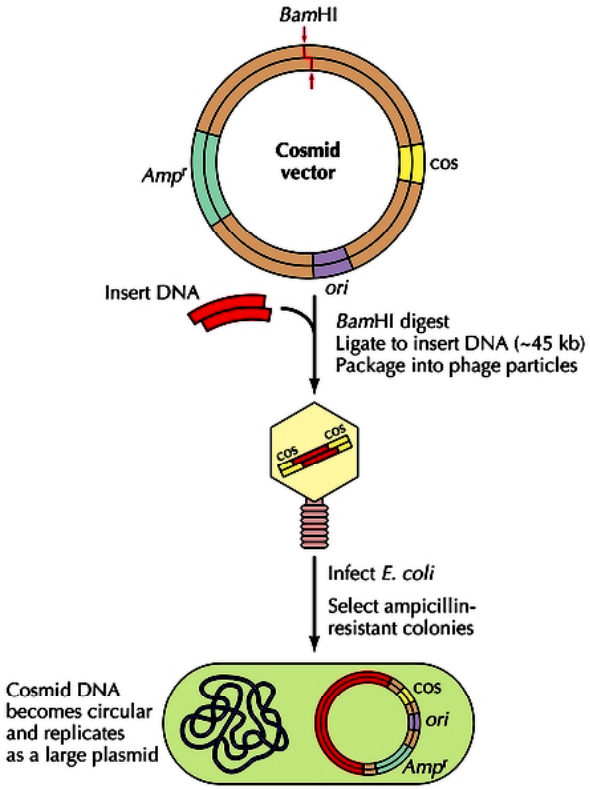
B. **bacteriophage** *(lambda)* – a double-stranded DNA virus of *E. coli* that accommodates DNA inserts in the range of 10-20 kb. *Recombinant phage* DNA molecules are packaged in vitro into viral particles using a preparation of purified phage structural coat proteins. Both the left and right arms of the phage and a minimum insert size are required for proper packaging. Thus, only recombinant phage with inserts are packaged. The packaged recombinant phages enter the host bacterium (*E. coli*) via a receptor in a process termed *infection*. The recombinant phages multiply in the host and eventually lyse the host cells. The phage particles released can infect other bacterial cells, which in turn lyse and release more phage. Cycles of infection and lysis with viral partical release result in a clear *plaque* on a lawn of *E. coli* embedded in an agar plate. Each plaque is a clone. Phage from a single plaque (clone) is used to infect large liquid cultures of *E. coli*. The phage particles released after lysis are the source of the purified cloned DNA. **Phage infection is more efficient than plasmid transformation**; even very small starting quantities of DNA fragments can be cloned using phage vectors. See figure above.

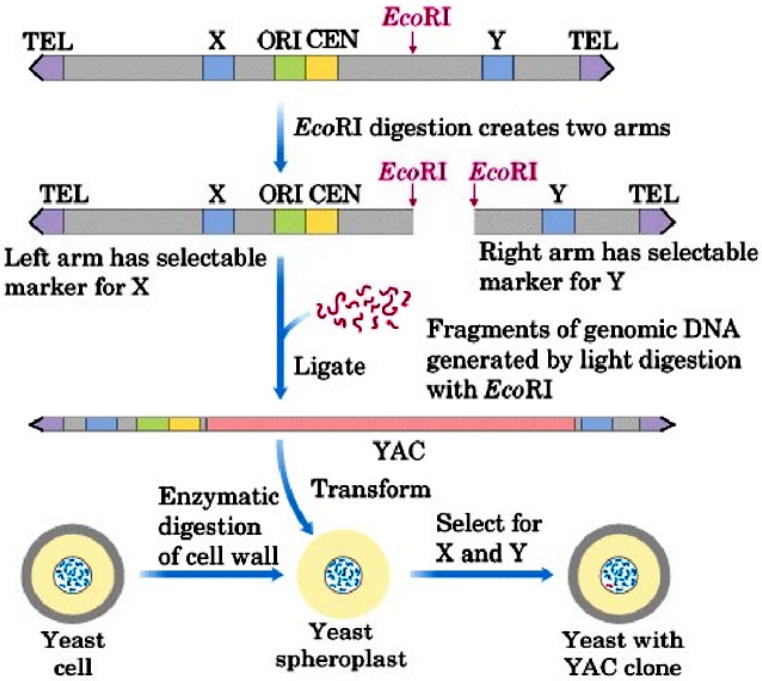


C. **Cosmids** - Plas<u>mid</u> vectors which contain the cohesive termini (<u>cos</u> sites) from the phage. Cosmids can accommodate ~45 kb of insert. See below left.

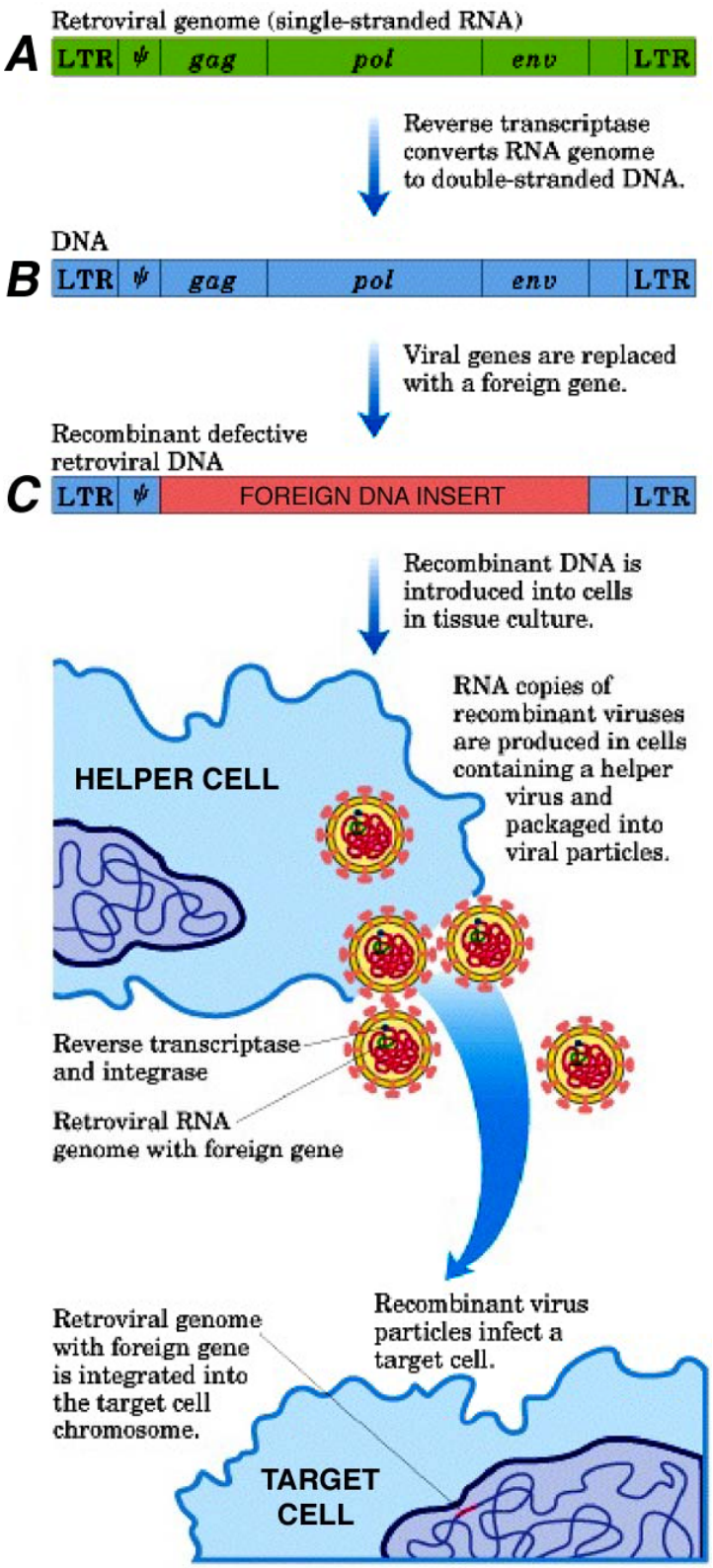
D. **BACs** – Plasmid that can accommodate an average insert size of 150 kb and can carry up to 400 kb. BACs are the preferred DNA clone resource for large genome sequencing programs. Like YACs below, they contain centromere-like regions that are required to partition them faithfully to daughter cells. See figure 1.10, Molecular Biology I.

E. **YACs** – Yeast Artificial Chromosomes – can accommodate very large DNA (200-1000 kb). Essential sequence elements: origin of replication, centromere, telomeres, selectable markers and unique restriction sites. See figure below, right.





F. **Retroviral vectors.** The retroviral genome is single-stranded RNA (***A***). To engineer retroviral vectors, the RNA is converted to double-stranded DNA using reverse transcriptase (***B***). To prepare the vector for insertion of a foreign DNA (i.e., the insert), the majority of the retrovirus genome is removed and replaced with the foreign DNA insert; the *gag, pol* and *env* genes of the retroviral genome, required for retroviral replication and viral particle assembly, are replaced with foreign DNA. Because these missing genes are essential for the assembly of virions, they must be must be provided by **helper cells** that package the recombinant retrovirus into viral particles. The viral particals are then used to transduce the **target cells** (mammalian cells). The viral particles also contain the viral reverse transcriptase and integrase enzymes (provided by the helper cells), which are introduced into the target cells. Thus, the resulting viral particle vectors introduce the recombinant viral RNA into target cells where the reverse transcriptase converts it to DNA, and the integrase integrates the DNA into the target cell genomic DNA. This integrated recombinant gene is transcribed into RNA for production of the encoded protein. However, the target cells do not produce infective particles because the recombinant virus lacks the gag, pol and env genes required for replication and packaging of the viral genome See figure.



III. Clinical correlations: Examples of viral vectors in gene therapy approaches.

A. *Ex vivo* retroviral infection

1. Goal is to enable normal blood cells to survive the toxicity of cancer chemotherapy. The patients first receive intensive chemotherapy, which lowers blood cell levels (toxic side-effect). The low blood cell levels are rescued with an autologous stem cell transplant. Patients then receive an additional course of chemotherapy, followed by a second transplant of stem cells that have been transduced with a *retroviral vector* containing the MDR-1 gene. The MDR-1 gene encodes a cytoplasmic membrane "multi-drug" pump. In the transduced cells, the drug pump allows the blood stem cells to survive and be "multi-drug" resistant. These cells increase 10-fold in the peripheral blood after subsequent cycles of chemotherapy. The expression of MDR-1 is causally linked to the increased survival of transduced cells during chemotherapy.
2. Goal is to treat SCID-X1, an X-linked member of SCID (severe combined immunodeficiency disease, "bubble boy disease"). SCID-X1 is characterized by an early block in T-cell and natural killer (NK) lymphocyte differentiation. The mutated gene causes malfunction of several cytokine receptors involved in signaling lymphocyte progenitors. The treatment involves harvesting stem cells from the bone marrow of five SCIP-X1 patients ranging in age from one month to nine months. The normal gene was inserted into the stem cells using the *retroviral vector*. The transduced stem cells were then transfused back into the patients. Encouraging therapeutic benefits were observed.

B. Infection *in vivo*.

Goal is to treat hemophilia by injection of *adenoassociated virus vector* directly into muscle to produce the missing clotting factor. The results: safe and successful gene transfer and expression of factor IX for the first three patients in a phase one/two trial. No antibodies to Factor IX have been observed to date. It might be expected that antibodies would be produced since these patients have never seen Factor IX before. Hemophilia is a good disease for gene therapy development because maintaining 1% of the normal level of the Factor IX is enough to prevent the major problems in the disease: joint damage and life-threatening bleeding complications. Also, the conventional treatment, intravenously administered concentrates of clotting factor, is associated with major problems: clotting factor has a short half-life of 8 to 12 hours which decreases throughout the patient's life; it is expensive and inconvenient and tissue damage continues; there is also risk of transmitting viral blood borne diseases through infusions such as AIDS and hepatitis A (the number one and two causes of death in hemophilia patients). Early clinical trials are underway.

IV. **Library construction.** *A DNA library is a collection of DNA clones*. The DNA clones are each composed of a vector and an insert. The clones in the library contain different inserts but all have the same vector.

A. **cDNA libraries, an overview** (cDNA = double stranded DNA copies of mRNA). A cDNA library is a collection of DNA copies of mRNA. [mRNA contain the nucleotide sequence that encodes protein as well as additional sequences 5' and 3' to the coding sequence that are known as UTRs (untranslated regions), which act to regulate mRNA stability and translation efficiency. Introns are not present in mRNA as they get spliced out in the nucleus before transport to the cytoplasm.]

B. **Examples of types of cDNA libraries**.

1. Tissue cDNA libraries contain copies of each of the mRNA species in a given tissue (e.g., substantia nigra, bone marrow, angiogenic regions of blood vessels, etc.) and can be used to identify, sequence and characterize the "expressed" DNA in a given tissue type.
2. Subtractive cDNA libraries – contain sequences present in one cell type but not another, or one condition but not another.
   a) mRNA present in one developmental stage of a tissue but not another
   b) mRNA present immediately after agent orange exposure but not before.
   c) mRNA present in old tissue but not young tissue
   d) mRNA present in diseased but not normal tissue

C. **Construction of cDNA libraries.** See figure below.

1. **Polymerization of cDNA.**

   a) *Polymerize 1st strand*.

      (1) Primers:

         (a) *oligo dT* is used to prime polyA.

         (b) Alternatively, *specific primers* can make specific mRNA subpopulations.

      (2) Enzyme: reverse transcriptase.

      (3) Template: mRNA.

      (4) Product: a mRNA/DNA duplex.

   b) *Hydrolysis of the mRNA* from the mRNA-DNA hybrid with *alkali*. This step prepares the 1st cDNA strand to be used as the template for synthesis of the 2nd DNA strand.

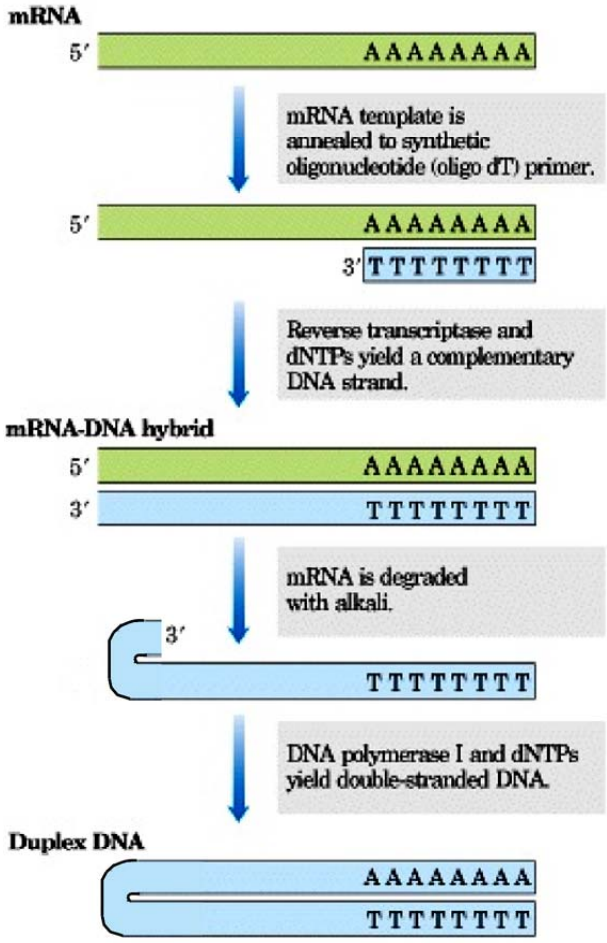
   c) *Polymerize 2nd strand*.

      (1) Primer: The 3' end of the 1st cDNA strand is the primer for synthesis of the sencond cDNA strand. This 3' end of the 1st strand loops back to use itself as the template for 2nd strand synthesis.

      (2) Enzyme: DNA polymerase.

      (3) Template: 1st strand cDNA.

      (4) Product: double stranded cDNA with a hairpin loop at one end. The hairpin loop is cleaved with a single stranded nuclease.



2. **Ligation of cDNA into a cloning vector (e.g., lambda).**

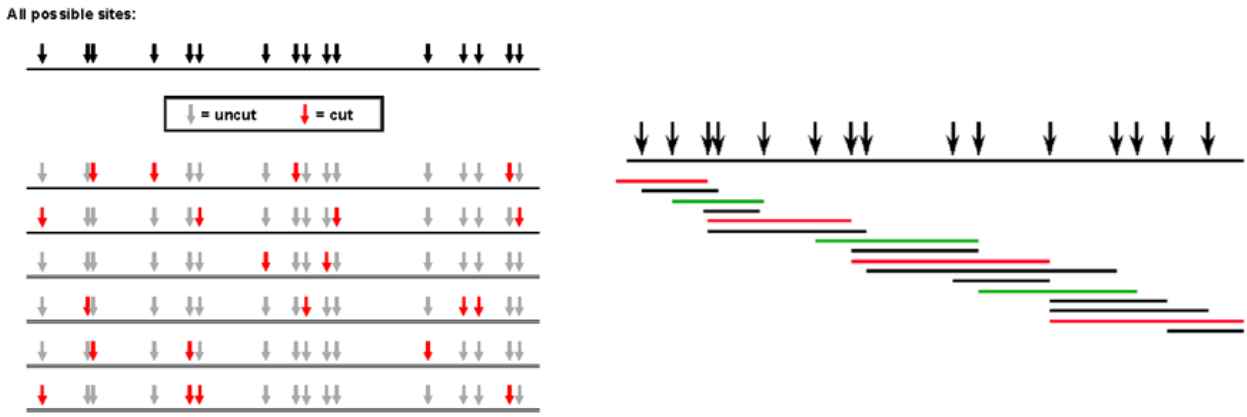
3. **Propagation of recombinant library.**

D. **Uses of cDNA libraries.**

1. **Sequencing.** Each of the clones in the library can be sequenced. If it is a cDNA library from a tissue, sequencing each clone in the library can be used to determine the genes that are turned on in that tissue.

2. **Obtain a specific DNA.** A specific cDNA can be obtained, for example, in order to engineer it for use in making large amounts of a therapeutic protein.

E. **Construction of genomic DNA libraries**. Overview. Genomic DNA is purified from cells. The particular cell type used doesn't matter since every cell from an individual has the same genomic DNA. The genomic DNA is digested with restriction endonucleases and then directly ligated into a vector. Large DNA pieces (20-200 kb) are inserted into the vectors. [A single gene can span as much as 200 kb.] Genomic DNA libraries are used for genome sequencing projects. All the individual clones of the library are sequenced. Since the clones overlap, they can be aligned to determine their order along chromosomes. *Bioinformatic* analyses can then identify the genes and predict the cDNA.

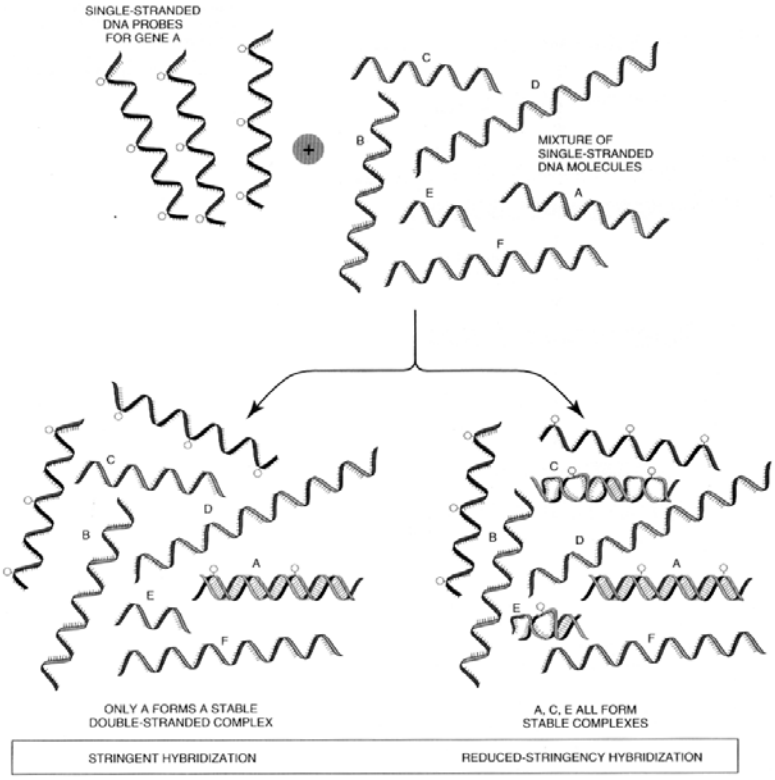


1. **Purify DNA**. To make a genomic library of a mammalian species, □white blood cells are often used since they can be readily obtained.
2. **Digest DNA with restriction enzymes**. A partial digestion is used: i.e., too little enzyme is used intentionally so that only a fraction of sites are cut and the resulting *restriction fragments* are a random population of overlapping DNA fragments. □□ See figure above. The size range of fragments appropriate for the vector are isolated.
3. **Ligate the DNA pieces into the vector**.
4. **Transform or infect the library into the host organism**. The resulting library of recombinants can be propagated in *E coli* indefinitely (the library can be used again and again).

V. **Nucleic acid hybridization** is the basis of nucleic acid sequence determination methods. Synthetic oligonucleotides (18 - 25 nucleotides long) are used as *hybridization probes*. The conditions for the hybridization are chosen such that exact hybridization is indicative of the nucleotide sequence.

A. **Overview**: The DNA double-helix can be denatured upon heating or extreme pH conditions. In methods of nucleic acid analysis, high temperature is typically used to denature the DNA. The temperature is then lowered to the exact point where only exactly complementary strands can *reanneal* to form stable duplexes. The basis of reannealing is the base-pairing language (A base-pairing with T, G base-pairing with C). In methods of nucleic acid analysis, a high concentration of oligonucleotide probe(s) (typically 18 – 25 nucleotides in length) is present, and when the temperature is lowered, the probe is much more likely to collide and *hybridize* with the target DNA than the target DNA

strands are to find each other and reanneal. The precise nucleotide sequence of a target DNA can be inferred from the hybridization of an oligonucleotide probe with its exact complementary sequence on the target DNA.



B. Above the melting temperature (T_m), the thermal energy is too high for probes to hybridize even with their exact complement. At temperatures roughly 5˚C below the T_m, probes hybridize rather selectively with their exact complements. At lower temperatures, probes can hybridize with their exact complements and with DNA that are less than exactly complementary (see figure). The smaller the probe, the lower the melting temperature. The lower the GC content of the probe, the lower the melting temperature.

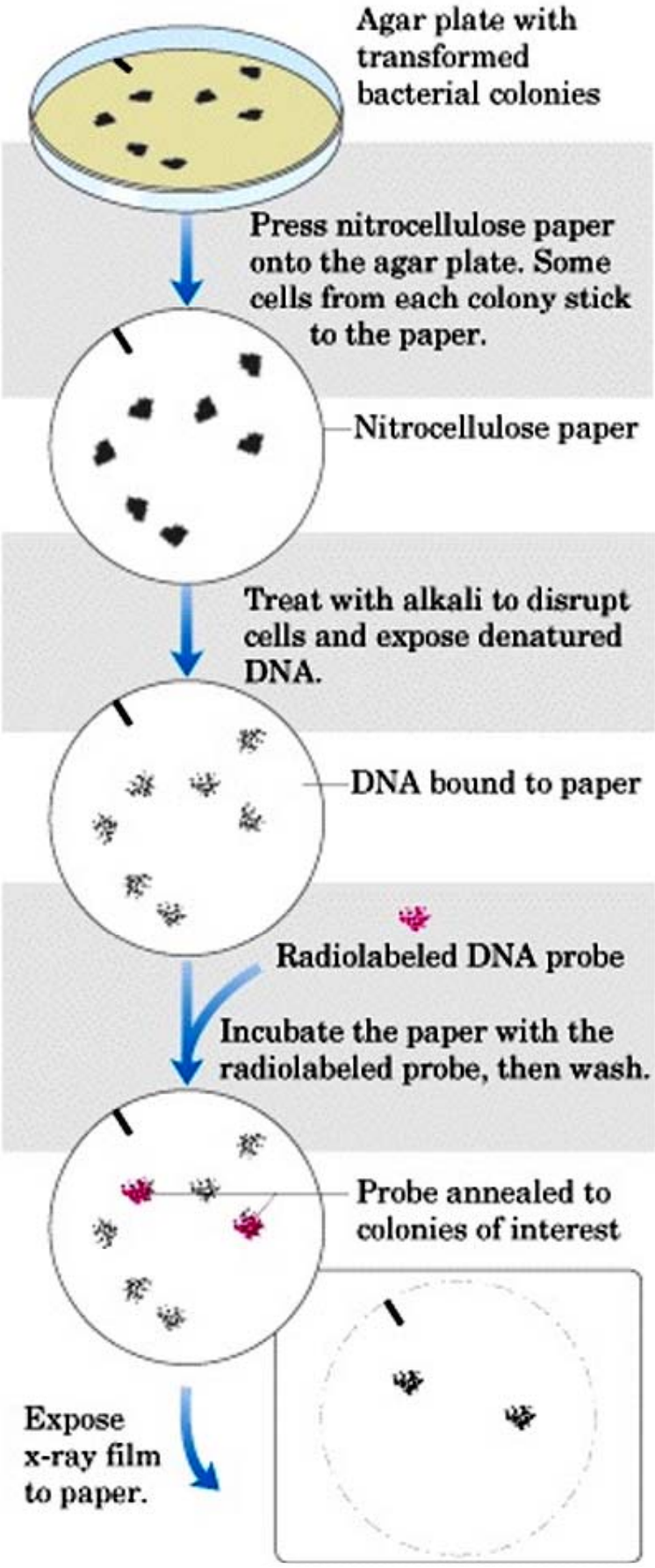
C. The ability of probes to hybridize with their exact complement vs. sequences similar to the complement is dependent on the **temperature**, **pH** and **ionic strength** conditions of the hybridization. These conditions are referred to as the *stringency* of hybridization.

VI. DNA Library screening.

A. **Introduction.** A DNA library is a mixture of DNA clones. A DNA library is typically *screened* using DNA hybridization probes to identify and isolate the DNA clones of interest. The method of library screening using radio-labeled probes is depicted in the figure below. The library is spread out onto the surface of agar plates and the individual clones are permitted to grow into bacterial colonies (for plasmid vectors) or to make plaques (for phage vectors). The colonies or plaques are then transferred to a membrane. The membrane is then *probed* with either a complimentary DNA probe or an antibody. The probes bind to and thereby identify the clones of interest.

B. **Screening with DNA probes.** If only a small part of the sequence of the DNA of interest is known, then this known sequence can be used as a hybridization probe to obtain the full length sequence. To make the probe, the DNA is *labeled* with a radionuclide or a fluorscent tag. The *labeled* probe is then hybridized to the colonies in order to identify the clones that contain the DNA of

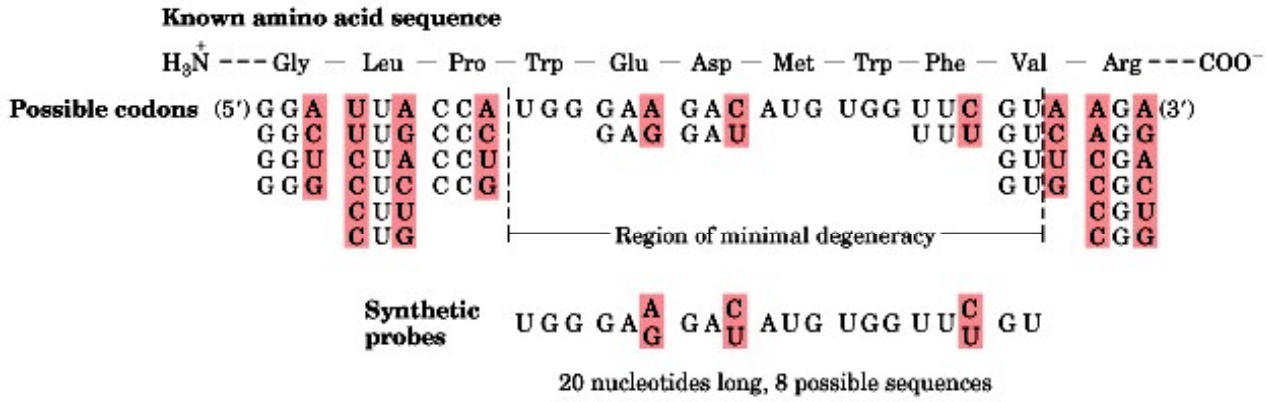
interest. The colonies are first broken open to expose their DNA to enable hybridization with *labeled probes.* The probes hybridze with and thereby can identify clones of interest from either cDNA or genomic DNA libraries.



1. **Types of hybridization probes**:

   a) *Synthetic single-stranded oligonucleotides with ^{32}P or fluorescent tags*. Short oligonucleotides can be ordered with 24 hrs notice at the current cost of about 50 cents per base (flourescent tags can be synthesized onto the oligo, or alternatively, enzymes can add radioactive or fluorescent labels to oligonucleotides).

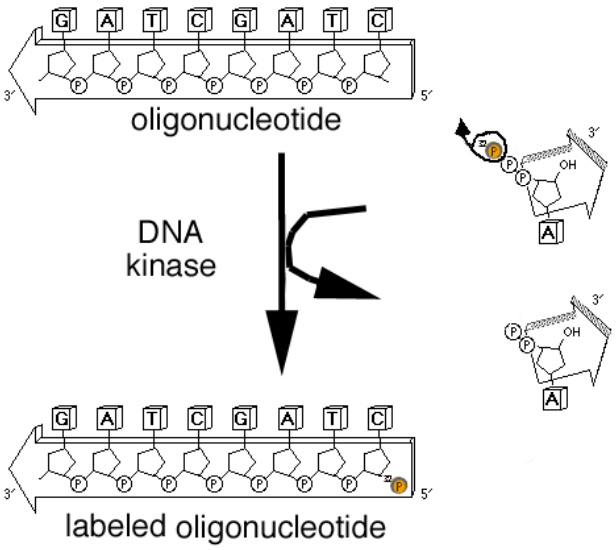
      (1) <u>An example of the use of an oligonucleotide hybridization probe</u>: If a short sequence of amino acids has been obtained from a protein, all the possible nucleotide sequences that can encode the amino acid sequence can be derived from codon tables (i.e., *reverse translation*). [Today this approach would be used if the protein is from a species whose genomic sequence is not known. If the genomic sequence is known, then the gene and mRNA encoding the protein can be obtained from a database search.] Due to the degeneracy of the genetic code, an amino acid sequence can be encoded by many different nucleotide sequences. Therefore, all the possible oligonucleotides that encode the peptide are synthesized and the combination of them is used as the probe. The strategy is to choose a short peptide sequence (~ 5 amino acids) that contains amino acids in which the codon is less degenerate (see figure below).



      Oligonucleotides can be labeled as shown in the figure on the next page, right.

      (2) **DNA probes made using a polymerase.** Cloned DNA can be used as a template for a DNA polymerase reaction in which a radioactive nucleotide (**^{32}P**) is incorporated into the polymerized DNA strand to create the probe. See figure, next page, left.

C. **Screening with antibody probes.** cDNA libraries can also be screened using antibodies if the library was engineered so that the protein encoded by each cDNA clone is synthesized in the host organism; i.e., the cDNA is inserted in the vector just downstream of a transcriptional control region (promoter) of the host species. An antibody that recognizes the protein is used to *screen* the library in a way that is analogous to DNA probes. A labeled *secondary antibody* that recognizes the *primary antibody* is used to detect the primary antibody.

One application of this method is to identify the cDNA encoding for the protein antigen targeted by an autoimmune antibody.

D. The sequencing of genomes of organisms of every type is providing unprecedented access to gene sequence information. Whereas the creation of one or more DNA libraries is an intermediate step in the sequencing of a genome, a gene or cDNA can readily be cloned *without the aid a library* when the genomic sequence is known using the polymerase chain reaction (PCR).