I. Libraries

A **library** is a collection of clones representing many different DNA fragments. The fragments can be from all the DNA in an organism, all the DNA from a one chromosome, all the genes expressed in a particular stage, etc.

In a **genomic library**, each vector molecule carries a fragment of DNA from the genome.
1) Isolate total genomic DNA
2) Digest it into pieces of appropriate size (*e.g.*, 40 kb for a cosmid vector). Usually this involves **partial digestion** – DNA is incubated with restriction enzyme for a short time so that only 10% of the sites are cut. This allows for different clones to have overlapping fragments.
3) Select fragments of the appropriate size, if necessary.
4) Ligate mixture of fragments into desired vector.

Number of clones required for coverage of the human genome (3 x 10^9 bp):

vector	insert size	No. clones
λ	15 kb	200,000
cosmid	40 kb	75,000
BAC	300 kb	10,000

There is never a completely random library with full coverage of a genome. Generally, one should use 4 or 5-fold the number of clones for 1X coverage to have a 95% chance of containing any given sequence (*e.g.*, a million lambda clones for the human genome).

II. cDNA synthesis

cDNA (**complementary DNA**) = DNA copy of mRNA

reverse transcriptase is an RNA-dependent DNA polymerase. RNA is the template. If we make a synthetic primer of T nucleotides (oligo-dT or poly-dT), it will anneal to the 3’ poly-A tract of mRNAs. This synthesis is called **reverse transcription**. Reverse transcription is not always complete – sometimes synthesis stops before the 5' end of the mRNA is reached; therefore, many cDNAs are not **full-length**!

The product of reverse transcription is an RNA/DNA hybrid. There are various methods to remove the RNA strand and use the DNA strand as a template for another synthesis with DNA polymerase, giving a double-stranded DNA molecule.

There are various techniques to clone cDNAs into plasmid (or other) vectors **directionally**, so that we know which end in the vector is from the 5’ end of the gene.

cDNAs are extremely useful:

The entire protein-coding sequence can be determined, even if the gene has many exons spread through a large genomic region.

Because cDNAs come from mRNAs, they tell us that a gene was transcribed in the tissue from which the cDNA was obtained. One can construct a **cDNA library** using all the mRNA from a particular source (a certain tissue, a cancer cell line, embryos, etc).

III. ESTs

Adams *et al*. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252:1651-1656:

> Automated partial DNA sequencing was conducted on more than 600 randomly selected human brain complementary DNA (cDNA) clones to generate expressed sequence tags (ESTs). ESTs have applications in the discovery of new human genes, mapping of the human genome, and identification of coding regions in genomic sequences. Of the sequences generated, 337 represent new genes, including 48 with significant similarity to genes from other organisms, such as a yeast RNA polymerase II subunit; Drosophila kinesin, Notch, and Enhancer of split; and a murine tyrosine kinase receptor. This fast approach to cDNA characterization will facilitate the tagging of most human genes in a few years at a fraction of the cost of complete genomic sequencing, provide new genetic markers, and serve as a resource in diverse research fields.

expressed sequence tag: expressed sequence, because it comes from mRNA so we know it's transcribed; tag, because we only sequence the ends of the cDNA (200-600 bp), which is enough information to uniquely identify the gene.

See http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html for information on the current EST collection.