

## Common Terms in Molecular Biology

**3' end/5' end:** A nucleic acid strand is inherently directional, and the "5 prime end" has a free hydroxyl (or phosphate) on a 5' carbon and the "3 prime end" has a free hydroxyl (or phosphate) on a 3' carbon (carbon atoms in the sugar ring are numbered from 1' to 5'). That's simple enough for an RNA strand or for single-stranded (ss) DNA. However, for double-stranded (ds) DNA it's not so obvious - each strand has a 5' end and a 3' end, and the 5' end of one strand is paired with the 3' end of the other strand (it is "antiparallel"). One would talk about the 5' end of ds DNA only if there was some reason to emphasize one strand over the other - for example if one strand is the sense strand of a gene. In that case, the orientation of the sense strand establishes the direction.

**3' flanking region:** A region of DNA which is NOT copied into the mature mRNA, but which is present adjacent to 3' end of the gene. It was originally thought that the 3' flanking DNA was not transcribed at all, but it was discovered to be transcribed into RNA, but quickly removed during processing of the primary transcript to form the mature mRNA. The 3' flanking region often contains sequences which affect the formation of the 3' end of the message. It may also contain enhancers or other sites to which proteins may bind.

**3' untranslated region:** A region of the DNA which IS transcribed into mRNA and becomes the 3' end of the message, but which does not contain protein coding sequence. Everything between the stop codon and the poly(A) tail is considered to be 3' untranslated. The 3' untranslated region may affect the translation efficiency of the mRNA or the stability of the mRNA. It also has sequences which are required for the addition of the poly(A) tail to the message (including one known as the "hexanucleotide", AAUAAA).

**5' flanking region:** A region of DNA which is NOT transcribed into RNA, but rather is adjacent to 5' end of the gene. The 5'-flanking region contains the promoter, and may also contain enhancers or other protein binding sites.

**5' untranslated region:** A region of a gene which IS transcribed into mRNA, becoming the 5' end of the message, but which does not contain protein coding sequence. The 5'-untranslated region is the portion of the DNA starting from the cap site and extending to the base just before the ATG translation initiation codon. While not itself translated, this region may have sequences which alter the translation efficiency of the mRNA, or which affect the stability of the mRNA.

**Ablation experiment:** An experiment designed to produce an animal deficient in one or a few cell types, in order to study cell lineage or cell function. The idea is to make a transgenic mouse with a toxin gene (often diphtheria toxin) under control of a specialized promoter which activates only in the target cell type. When embryo development progresses to the point where it starts to form the target tissue, the toxin gene is activated, and that specific tissue dies. Other tissues are unaffected.

**Acrylamide gels:** A polymer gel used for electrophoresis of DNA or protein to measure their sizes (in daltons for proteins, or in base pairs for DNA). See "Gel Electrophoresis". Acrylamide gels are especially useful for high resolution separations of DNA in the range of tens to hundreds of nucleotides in length.

**Agarose gels:** A polysaccharide gel used to measure the size of nucleic acids (in bases or base pairs). See "Gel Electrophoresis". This is the gel of choice for DNA or RNA in the range of thousands of bases in length, or even up to 1 megabase if you are using pulsed field gel electrophoresis.

**Amp resistance:** See "Antibiotic resistance".

**Anneal:** Generally synonymous with "hybridize".

**Antibiotic resistance:** Plasmids generally contain genes which confer on the host bacterium the ability to survive a given antibiotic. If the plasmid pBR322 is present in a host, that host will not be killed by (moderate levels of) ampicillin or tetracycline. By using plasmids containing antibiotic resistance genes, the researcher can kill off all the bacteria which have not taken up his plasmid, thus ensuring that the plasmid will be propagated as the surviving cells divide.

**Anti-sense strand:** See discussion under "Sense strand".

**AP-1 site:** The binding site on DNA at which the transcription "factor" AP-1 binds, thereby altering the rate of transcription for the adjacent gene. AP-1 is actually a complex between c-fos protein and c-jun protein, or sometimes is just c-jun dimers. The AP-1 site consensus sequence is (C/G)TGACT(C/A)A. Also known as the TPA-response element (TRE). [TPA is a phorbol ester, tetradecanoyl phorbol acetate, which is a chemical tumor promoter]

**ATG or AUG:** The codon for methionine; the translation initiation codon. Usually, protein translation can only start at a methionine codon (although this codon may be found elsewhere within the protein sequence as well). In eukaryotic DNA, the sequence is ATG; in RNA it is AUG. Usually, the first AUG in the mRNA is the point at which translation starts, and an open reading frame follows - i.e. the nucleotides taken three at a time will code for the amino acids of the protein, and a stop codon will be found only when the protein coding region is complete.

**BAC:** Bacterial Artificial Chromosome — a cloning vector capable of carrying between 100 and 300 kilobases of target sequence. They are propagated as a mini-chromosome in a bacterial host. The size of the typical BAC is ideal for use as an intermediate in large-scale genome sequencing projects. Entire genomes can be cloned into BAC libraries, and entire BAC clones can be shotgun-sequenced fairly rapidly.

**Band shift assay:** see Gel shift assay.

**Bacteriophage lambda:** A virus which infects *E. coli*, and which is often used in molecular genetics experiments as a vector, or cloning vehicle. Recombinant phages can be made in which certain non-essential  $\lambda$  DNA is removed and replaced with the DNA of interest. The phage can accommodate a DNA "insert" of about 15-20 kb. Replication of that virus will thus replicate the investigator's DNA. One would use phage  $\lambda$  rather than a plasmid if the desired piece of DNA is rather large.

**Binding site:** A place on cellular DNA to which a protein (such as a transcription factor) can bind. Typically, binding sites might be found in the vicinity of genes, and would be involved in activating transcription of that gene (promoter elements), in enhancing the transcription of that gene (enhancer elements), or in reducing the transcription of that gene (silencers). NOTE that whether the protein in fact performs these functions may depend on some condition, such as the presence of a hormone, or the tissue in which the gene is being examined. Binding sites could also be involved in the regulation of chromosome structure or of DNA replication.

**Blotting:** A technique for detecting one RNA within a mixture of RNAs (a Northern blot) or one type of DNA within a mixture of DNAs (a Southern blot). A blot can prove whether that one species of RNA or DNA is present, how much is there, and its approximate size. Basically, blotting involves gel electrophoresis, transfer to a blotting membrane (typically nitrocellulose or activated nylon), and incubating with a radioactive probe. Exposing the membrane to X-ray film produces darkening at a spot correlating with the position of the DNA or RNA of interest. The darker the spot, the more nucleic acid was present there. (see figure, below)

**BP:** Abbreviation for base pair(s). Double stranded DNA is usually measured in bp rather than nucleotides (nt).

**Cap:** All eukaryotes have at the 5' end of their messages a structure called a "cap", consisting of a 7-methylguanosine in 5'-5' triphosphate linkage with the first nucleotide of the mRNA. It is added post-transcriptionally, and is not encoded in the DNA.

**Cap site:** Two usages: In eukaryotes, the cap site is the position in the gene at which transcription starts, and really should be called the "transcription initiation site". The first nucleotide is transcribed from this site to start the nascent RNA chain. That nucleotide becomes the 5' end of the chain, and thus the nucleotide to which the cap structure is attached (see "Cap"). In bacteria, the CAP site (note the capital letters) is a site on the DNA to which a protein factor (the Catabolite Activated Protein) binds.

**CAT assay:** An enzyme assay. CAT stands for chloramphenicol acetyl transferase, a bacterial enzyme which inactivates chloramphenicol by acetylating it. CAT assays are often performed to test the function of a promoter. The gene coding for CAT is linked onto a promoter (transcription control region) from another gene, and the construct is "transfected" into cultured cells. The amount of CAT enzyme produced is taken to indicate the transcriptional activity of the promoter (relative to other promoters which must be tested in parallel). It is easier to perform a CAT assay than it is to do a Northern blot, so CAT assays were a common method for testing the effects of sequence changes on promoter function. Largely supplanted by the reporter gene luciferase.

**CCAAT box:** (CAT box, CAAT box, other variants) A sequence found in the 5' flanking region of certain genes which is necessary for efficient expression. A transcription factor (CCAAT-binding protein, CBP) binds to this site.

**cDNA clone:** "complementary DNA"; a piece of DNA copied from an mRNA. The term "clone" indicates that this cDNA has been spliced into a plasmid or other vector in order to propagate it. A cDNA clone may contain DNA copies of such typical mRNA regions as coding sequence, 5'-untranslated region, 3' untranslated region or poly(A) tail. No introns will be present, nor any promoter sequences (or other 5' or 3' flanking regions). A "full-length" cDNA clone is one which contains all of the mRNA sequence from nucleotide #1 through to the poly(A) tail.

**ChIP:** See Chromatin Immunoprecipitation (below).

**Chromatin Immunoprecipitation:** This is a method for isolating and characterizing the specific pieces of DNA out of an entire genome, to which is bound a protein of interest. The protein of interest could for example be a transcription factor, or a specific modified histone, or any other DNA binding protein. This procedure requires an antibody to that protein of interest.

One isolates chromosomal material with all the proteins still bound to the genomic DNA. After fragmenting the DNA, you use the antibody to immunoprecipitate all chunks that contain your protein of interest. Isolate the DNA from those chunks, and you can characterize the specific DNA sites to which your protein was bound.

There are two common ways to characterize the DNA so isolated: ChIP-chip (or "ChIP-on-chip") or ChIP-seq.

- **ChIP-chip:** In this variant, the DNA isolated from a ChIP experiment is characterized by labeling it with a fluorescent dye, then hybridizing it to a DNA array (an oligonucleotide array, for example). Array spots that "light up" are taken as evidence that their specific sequence is present in your ChIP product. Unfortunately, designing these arrays requires that you have some idea what to expect in your ChIP isolates.
- **ChIP-seq:** A newer variant for characterizing ChIP results, one can simply sequence everything that immunoprecipitated with the antibody. It requires no fore-knowledge of the expected products, as would ChIP-chip.

**Chromosome walking:** A technique for cloning everything in the genome around a known piece of DNA (the starting probe). You screen a genomic library for all clones hybridizing with the probe, and then figure out which one extends furthest into the surrounding DNA. The most distal piece of this most distal clone is then used as a probe, so that ever more distal regions can be cloned. This has been used to move as much as 200 kb away from a given starting point (an immense undertaking). Typically used to "walk" from a starting point towards some nearby gene in order to clone that gene. Also used to obtain the remainder of a gene when you have isolated a part of it.

**Clone (verb):** To "clone" something is to produce copies of it. To clone a piece of DNA, one would insert it into some type of vector (say, a plasmid) and put the resultant construct into a host (usually a bacterium) so that the plasmid and insert replicate with the host. An individual bacterium is isolated and grown and the plasmid containing the "cloned" DNA is re-isolated from the bacteria, at which point there will be many millions of copies of the DNA - essentially an unlimited supply. Actually, an investigator wishing to clone some gene or cDNA rarely has that DNA in a purified form, so practically speaking, to "clone" something involves screening a cDNA or genomic library for the desired clone. See also "Probe" for a description of how one might start a cloning project, and "Screening" for how the probe is used.

One can also clone more complex organisms, with considerable difficulty. The much-publicized Scottish research that resulted in the sheep 'Dolly' exemplifies this approach.

**Clone (noun):** The term "clone" can refer either to a bacterium carrying a cloned DNA, or to the cloned DNA itself. If you receive a clone from a collaborator, you should first figure out if they send you DNA or bacteria. If it is DNA, your first job is to introduce it ("transform" it) into bacteria [see "Transformation (with respect to bacteria)"]. Occasionally, someone might send just the "insert", rather than the whole plasmid. "Your assignment, Jim, if you decide to accept it", is to splice that DNA into a convenient vector, and only then can you transform it into bacteria.

**Coding sequence:** The portion of a gene or an mRNA which actually codes for a protein. Introns are not coding sequences; nor are the 5' or 3' untranslated regions (or the flanking regions, for that matter - they are not even transcribed into mRNA). The coding sequence in a cDNA or mature mRNA includes everything from the AUG (or ATG) initiation codon through to the stop codon, inclusive.

**Coding strand:** an ambiguous term intended to refer to one specific strand in a double-stranded gene. See "Sense strand".

**Codon:** In an mRNA, a codon is a sequence of three nucleotides which codes for the incorporation of a specific amino acid into the growing protein. The sequence of codons in the mRNA unambiguously defines the primary structure of the final protein. Of course, the codons in the mRNA were also present in the genomic DNA, but the sequence may be interrupted by introns.

**Consensus sequence:** A 'nominal' sequence inferred from multiple, imperfect examples. Multiple lanes of shotgun sequence can be merged to show a consensus sequence. The optimal sequence of nucleotides recognized by some factor. A DNA binding site for a protein may vary substantially, but one can infer the consensus sequence for the binding site by comparing numerous examples. For example, the (fictitious) transcription factor ZQ1 usually binds to the sequences AAAGTT, AAGGTT or AAGATT. The consensus sequence for that factor is said to be AARRTT (where R is any purine, i.e. A or G). ZQ1 may also be able to weakly bind to ACAGTT (which differs by one base from the consensus).

**Contig:** Several uses, all nouns. The term comes from a shortening of the word 'contiguous'. A 'contig' may refer to a map showing placement of a set of clones that completely, contiguously cover some segment of DNA in which you are interested. Also called the 'minimal tiling path'. More often, the term 'contig' is used to refer to the final product of a shotgun sequencing project. When individual lanes of sequence information are merged to infer the sequence of the larger DNA piece, the product consensus sequence is called a 'contig'.

**Cosmid:** A type of vector used for cloning 35-45 kb of DNA. These are plasmids carrying a phage  $\lambda$  cos site (which allows packaging into  $\lambda$  capsids), an origin of replication and an antibiotic resistance gene. A plasmid of 40 kb is very difficult to put into bacteria, but can replicate once there. Cosmids, however, have a cos site, and thus can be packaged into  $\lambda$  phage heads (a reaction which can be performed *in vitro*) to allow efficient introduction into bacteria (you'll have to look up the cos site elsewhere).

**DNase:** Deoxyribonuclease, a class of enzymes which digest DNA. The most common is DNase I, an endonuclease which digests both single and double-stranded DNA.

**Dot blot:** A technique for measuring the amount of one specific DNA or RNA in a complex mixture. The samples are spotted onto a hybridization membrane (such as nitrocellulose or activated nylon, etc.), fixed and hybridized with a radioactive probe. The extent of labeling (as determined by autoradiography and densitometry) is proportional to the concentration of the target molecule in the sample. Standards provide a means of calibrating the results.

**Downstream:** See "Upstream/Downstream".

***E. coli:*** A common Gram-negative bacterium useful for cloning experiments. Present in human intestinal tract. Hundreds of strains of *E. coli* exist. One strain, K-12, has been completely sequenced.

**Electrophoresis:** See "Gel electrophoresis".

**Endonuclease:** An enzyme which digests nucleic acids starting in the middle of the strand (as opposed to an exonuclease, which must start at an end). Examples include the restriction enzymes, DNase I and RNase A.

**Enhancer:** An enhancer is a nucleotide sequence to which transcription factor(s) bind, and which increases the transcription of a gene. It is NOT part of a promoter; the basic difference being that an enhancer can be moved around anywhere in the general vicinity of the gene (within several thousand nucleotides on either side or even within an intron), and it will still function. It can even be clipped out and spliced back in backwards, and will still operate. A promoter, on the other hand, is position- and orientation-dependent. Some enhancers are "conditional" - in other words, they enhance transcription only under certain conditions, for example in the presence of a hormone.

**ERE:** Estrogen Response Element. A binding site in a promoter to which the activated estrogen receptor can bind. The estrogen receptor is essentially a transcription factor which is activated only in the presence of estrogens. The activated receptor will bind to an ERE, and transcription of the adjacent gene will be altered. See also "Response element".

**Evolutionary Footprinting:** One can infer which portions of a gene are important by comparing the sequence of that gene with its cognates from other species. A plot showing the regions of high conservation will presumably reflect the regions that are functional in all the test species. In theory, the more species involved in the comparison, the more stringent the result can be (i.e. the more the conserved regions will reflect truly important sequences). Care must be taken, however, to use species in which the function of the gene has not diverged excessively, or the outcome will be uninformative.

**Exon:** Those portions of a genomic DNA sequence which WILL be represented in the final, mature mRNA. The term "exon" can also be used for the equivalent segments in the final RNA. Exons may include coding sequences, the 5' untranslated region or the 3' untranslated region.

**Exonuclease:** An enzyme which digests nucleic acids starting at one end. An example is Exonuclease III, which digests only double-stranded DNA starting from the 3' end.

**Expression:** To "express" a gene is to cause it to function. A gene which encodes a protein will, when expressed, be transcribed and translated to produce that protein. A gene which encodes an RNA rather than a protein (for example, a rRNA gene) will produce that RNA when expressed.

**Expression clone:** This is a clone (plasmid in a bacteria, or maybe a  $\lambda$  phage in bacteria) which is designed to produce a protein from the DNA insert. Mammalian genes do not function in bacteria, so to get bacterial expression from your mammalian cDNA, you would place its coding region (i.e. no introns) immediately adjacent to bacterial transcription/translation control sequences. That artificial construct (the "expression clone") will produce a pseudo-mammalian protein if put back into bacteria. Often, that protein can be recognized by antibodies raised against the authentic mammalian protein, and vice versa.

**Footprinting:** A technique by which one identifies a protein binding site on cellular DNA. The presence of a bound protein prevents DNase from "nicking" that region, which can be detected by an appropriately designed gel.

**Gel electrophoresis:** A method to analyze the size of DNA (or RNA) fragments. In the presence of an electric field, larger fragments of DNA move through a gel slower than smaller ones. If a sample contains fragments at four different discrete sizes, those four size classes will, when subjected to electrophoresis, all migrate in groups, producing four migrating "bands". Usually, these are visualized by soaking the gel in a dye (ethidium bromide) which makes the DNA fluoresce under UV light.

**Gel shift assay:** (aka gel mobility shift assay (GMSA), band shift assay (BSA), electrophoretic mobility shift assay (EMSA)) A method by which one can determine whether a particular protein preparation contains factors which bind to a particular DNA fragment. When a radiolabeled DNA fragment is run on a gel, it shows a characteristic mobility. If it is first incubated with a cellular extract of proteins (or with purified protein), any protein-DNA complexes will migrate slower than the naked DNA - a shifted band.

**Gene:** A unit of DNA which performs one function. Usually, this is equated with the production of one RNA or one protein. A gene contains coding regions, introns, untranslated regions and control regions.

**Genome:** The total DNA contained in each cell of an organism. Mammalian genomic DNA (including that of humans) contains  $6 \times 10^9$  base pairs of DNA per diploid cell. There are somewhere in the order of a hundred thousand genes, including coding regions, 5' and 3' untranslated regions, introns, 5' and 3' flanking DNA. Also present in the genome are structural segments such as telomeric and centromeric DNAs and replication origins, and intergenic DNA.

**Genomic blot:** A type of Southern blot specifically used to analyze a mixture of DNA fragments derived from total genomic DNA. Because genomic DNA is very complicated, when it has been digested with restriction enzymes, it produces a complex set of fragments ranging from tens of bp to tens of thousands of bp. However, any specific gene will be reproducibly found on only one or a few specific fragments. A million identical cells will produce a

million identical restriction fragments for any given gene, so probing a genomic Southern with a gene-specific probe will produce a pattern of perhaps one or just a few bands.

**Genomic clone:** A piece of DNA taken from the genome of a cell or animal, and spliced into a bacteriophage or other cloning vector. A genomic clone may contain coding regions, exons, introns, 5' flanking regions, 5' untranslated regions, 3' flanking regions, 3' untranslated regions, or it may contain none of these...it may only contain intergenic DNA (usually not a desired outcome of a cloning experiment!).

**Genotype:** Two uses: one is a verb, the other a noun. To 'genotype' (verb) is to example polymorphisms (e.g. RFLPs, microsatellites, SNPs) present in a sample of DNA. You might be looking for linkage between a microsatellite marker and an unknown disease gene. With such information, you can infer the chromosomal location of the unknown gene, and can sometimes identify the gene. As a noun, a 'genotype' is the result of a genotyping experiment, be it a SNP or microsat or whatever.

**GRE: Glucocorticoid Response Element:** A binding site in a promoter to which the activated glucocorticoid receptor can bind. The glucocorticoid receptor is essentially a transcription factor which is activated only in the presence of glucocorticoids. The activated receptor will bind to a GRE, and transcription of the adjacent gene will be altered. See also "Response element".

**Helix-loop-helix:** A protein structural motif characteristic of certain DNA-binding proteins.

**hnRNA:** Heterogeneous nuclear RNA; refers collectively to the variety of RNAs found in the nucleus, including primary transcripts, partially processed RNAs and snRNA. The term hnRNA is often used just for the unprocessed primary transcripts, however.

**Host strain (bacterial):** The bacterium used to harbor a plasmid. Typical host strains include HB101 (general purpose *E. coli* strain), DH5 $\alpha$  (ditto), JM101 and JM109 (suitable for growing M13 phages), XL1-Blue (general-purpose, good for blue/white lacZ screening). Note that the host strain is available in a form with no plasmids (hence you can put one of your own into it), or it may have plasmids present (especially if you put them there). Hundreds, perhaps thousands, of host strains are available.

**Hybridization:** The reaction by which the pairing of complementary strands of nucleic acid occurs. DNA is usually double-stranded, and when the strands are separated they will re-hybridize under the appropriate conditions. Hybrids can form between DNA-DNA, DNA-RNA or RNA-RNA. They can form between a short strand and a long strand containing a region complementary to the short one. Imperfect hybrids can also form, but the more imperfect they are, the less stable they will be (and the less likely to form). To "anneal" two strands is the same as to "hybridize" them.

**Insert:** In a complete plasmid clone, there are two types of DNA - the "vector" sequences and the "insert". The vector sequences are those regions necessary for propagation, antibiotic resistance, and all those mundane functions necessary for useful cloning. In contrast, however, the insert is the piece of DNA in which you are really interested.

**Intergenic:** Between two genes; e.g. intergenic DNA is the DNA found between two genes. The term is often used to mean non-functional DNA (or at least DNA with no known importance to the two genes flanking it). Alternatively, one might speak of the "intergenic distance" between two genes as the number of base pairs from the polyA site of the first gene to the cap site of the second. This usage might therefore include the promoter region of the second gene.

**Intron:** Introns are portions of genomic DNA which ARE transcribed (and thus present in the primary transcript) but which are later spliced out. They thus are not present in the mature mRNA. Note that although the 3' flanking region is often transcribed, it is removed by endonucleolytic cleavage and not by splicing. It is not an intron.

**KB:** abbreviation for kilobase, one thousand bases.

**Kinase:** A kinase is in general an enzyme that catalyzes the transfer of a phosphate group from ATP to something else. In molecular biology, it has acquired the more specific verbal usage for the transfer onto DNA of a radiolabeled phosphate group. This would be done in order to use the resultant "hot" DNA as a probe.

**Knock-out experiment:** A technique for deleting, mutating or otherwise inactivating a gene in a mouse. This laborious method involves transfecting a crippled gene into cultured embryonic stem cells, searching through the thousands of resulting clones for one in which the crippled gene exactly replaced the normal one (by homologous recombination), and inserting that cell back into a mouse blastocyst. The resulting mouse will be chimaeric but, if you are lucky (and if you've gotten this far, you obviously are), its germ cells will carry the deleted gene. A few rounds of careful breeding can then produce progeny in which both copies of the gene are inactivated.

**Lambda:** see Bacteriophage Lambda.

**Leucine zipper:** A motif found in certain proteins in which Leu residues are evenly spaced through an  $\alpha$ -helical region, such that they would end up on the same face of the helix. Dimers can form between two such proteins. The Leu zipper is important in the function of transcription factors such as Fos and Jun and related proteins.

**Library:** A library might be either a genomic library, or a cDNA library. In either case, the library is just a tube carrying a mixture of thousands of different clones - bacteria or  $\lambda$  phages. Each clone carries an "insert" - the cloned DNA.

A cDNA library is usually just a mixture of bacteria, where each bacteria carries a different plasmid. Inserted into the plasmids (one per plasmid) are thousands of different pieces of cDNA (each typ. 500-5000 bp) copied from some source of mRNA, for example, total liver mRNA. The basic idea is that if you have a large enough number of different liver-derived cDNAs carried in those bacteria, there is a 99% probability that a cDNA copy of any given liver mRNA exists somewhere in the tube. The real trick is to find the one you want out of that mess - a process called screening (see "Screening").

A genomic library is similar in concept to a cDNA library, but differs in three major ways - 1) the library carries pieces of genomic DNA (and so contains introns and flanking regions, as well as coding and untranslated); 2) you need bacteriophage  $\lambda$  or cosmids, rather than plasmids, because... 3) the inserts are usually 5-15 kb long (in a  $\lambda$  library) or 20-40 kb (in a cosmid library). Therefore, a genomic library is most commonly a tube containing a mixture of  $\lambda$  phages. Enough different phages must be present in the library so that any given piece of DNA from the source genome has a 99% probability of being present.

**Ligase:** An enzyme, T4 DNA ligase, which can link pieces of DNA together. The pieces must have compatible ends (both of them blunt, or else mutually compatible sticky ends), and the ligation reaction requires ATP.

**Ligation:** The process of splicing two pieces of DNA together. In practice, a pool of DNA fragments are treated with ligase (see "Ligase") in the presence of ATP, and all possible splicing products are produced, including circularized forms and end-to-end ligation of 2, 3 or more pieces. Usually, only some of these products are useful, and the investigator must have some way of selecting the desirable ones.

**Linker:** A small piece of synthetic double-stranded DNA which contains something useful, such as a restriction site. A linker might be ligated onto the end of another piece of DNA to provide a desired restriction site.

**Marker:** Two typical usages:

Molecular weight size marker: a piece of DNA of known size, or a mixture of pieces with known size, used on electrophoresis gels to determine the size of unknown DNA's by comparison.

Genetic marker: A known site on the chromosome. It might for example be the site of a locus with some recognizable phenotype, or it may be the site of a polymorphism that can be experimentally discerned. See 'Microsatellite', 'SNP', 'Genotyping'.

**Message:** see mRNA.

**Microsatellite:** A microsatellite is a simple sequence repeat (SSR). It might be a homopolymer ('...TTTTTTT...'), a dinucleotide repeat ('...CACACACACACACA...'), trinucleotide repeat ('...AGTAGTAGTAGTAGT...') etc. Due to polymerase slip (a.k.a. polymerase chatter), during DNA replication there is a slight chance these repeat sequences may become altered; copies of the repeat unit can be created or removed. Consequently, the exact number of repeat units may differ between unrelated individuals. Considering all the known microsatellite

markers, no two individuals are identical. This is the basis for forensic DNA identification and for testing of familial relationships (e.g. paternity testing).

**mRNA:** "messenger RNA" or sometimes just "message"; an RNA which contains sequences coding for a protein. The term mRNA is used only for a mature transcript with polyA tail and with all introns removed, rather than the primary transcript in the nucleus. As such, an mRNA will have a 5' untranslated region, a coding region, a 3' untranslated region and (almost always) a poly(A) tail. Typically about 2% of the total cellular RNA is mRNA.

**M13:** A bacteriophage which infects certain strains of *E. coli*. The salient feature of this phage is that it packages only a single strand of DNA into its capsid. If the investigator has inserted some heterologous DNA into the M13 genome, copious quantities of single-stranded DNA can subsequently be isolated from the phage capsids. M13 is often used to generate templates for DNA sequencing.

**Nick translation:** A method for incorporating radioactive isotopes (typically  $^{32}\text{P}$ ) into a piece of DNA. The DNA is randomly nicked by DNase I, and then starting from those nicks DNA polymerase I digests and then replaces a stretch of DNA. Radiolabeled precursor nucleotide triphosphates can thus be incorporated.

**Non-coding strand:** Anti-sense strand. See "Sense strand" for a discussion of sense strand vs. anti-sense strand.

**Northern blot:** A technique for analyzing mixtures of RNA, whereby the presence and rough size of one particular type of RNA (usually an mRNA) can be ascertained. See "Blotting" for more information. After Dr. E. M. Southern invented the Southern blot, it was adapted to RNA and named the "Northern" blot.

**NT:** Abbreviation for nucleotide; i.e. the monomeric unit from which DNA or RNA are built. One can express the size of a nucleic acid strand in terms of the number of nucleotides in its chain; hence 'nt' can be a measure of chain length.

**Nuclear run-on:** A method used to estimate the relative rate of transcription of a given gene, as opposed to the steady-state level of the mRNA transcript (which is influenced not just by transcription rates, but by the stability of the RNA). This technique is based on the assumption that a highly-transcribed gene should have more molecules of RNA polymerase bound to it than will the same gene in a less-active state. If properly prepared, isolated nuclei will continue to transcribe genes and incorporate  $^{32}\text{P}$  into RNA, but only in those transcripts that were in progress at the time the nuclei were isolated. Once the polymerase molecules complete the transcript they have in progress, they should not be able to re-initiate transcription. If that is true, then the amount of radiolabel incorporated into a specific type of mRNA is theoretically proportional to the number of RNA polymerase complexes present on that gene at the time of isolation. A very difficult technique, rarely applied appropriately from what I understand.

**Nuclease:** An enzyme which degrades nucleic acids. A nuclease can be DNA-specific (a DNase), RNA-specific (RNase) or non-specific. It may act only on single stranded nucleic acids, or only on double-stranded nucleic acids, or it may be non-specific with respect to strandedness. A nuclease may degrade only from an end (an exonuclease), or may be able to start in the middle of a strand (an endonuclease). To further complicate matters, many enzymes have multiple functions; for example, Bal31 has a 3'-exonuclease activity on double-stranded DNA, and an endonuclease activity specific for single-stranded DNA or RNA.

**Nuclease protection assay:** See "RNase protection assay".

**Oncogene:** A gene in a tumor virus or in cancerous cells which, when transferred into other cells, can cause transformation (note that only certain cells are susceptible to transformation by any one oncogene). Functional oncogenes are not present in normal cells. A normal cell has many "proto-oncogenes" which serve normal functions, and which under the right circumstances can be activated to become oncogenes. The prefix "v-" indicates that a gene is derived from a virus, and is generally an oncogene (like *v-src*, *v-ras*, *v-myb*, etc). See also "Transformation (with respect to cultured cells)".

**Open reading frame:** Any region of DNA or RNA where a protein could be encoded. In other words, there must be a string of nucleotides (possibly starting with a Met codon) in which one of the three reading frames has no stop codons. See "Reading frame" for a simple example.

**Origin of replication:** Nucleotide sequences present in a plasmid which are necessary for that plasmid to replicate in the bacterial host. (Abbr. "ori")

**pBR322:** A common plasmid. Along with the obligatory origin of replication, this plasmid has genes which make the *E. coli* host resistant to ampicillin and tetracycline. It also has several restriction sites (BamHI, PstI, EcoRI, HindIII etc.) into which DNA fragments could be spliced in order to clone them.

**PCR:** see Polymerase Chain Reaction.

**Phagemid:** A type of plasmid which carries within its sequence a bacteriophage replication origin. When the host bacterium is infected with "helper" phage, the phagemid is replicated along with the phage DNA and packaged into phage capsids.

**Plasmid:** A circular piece of DNA present in bacteria or isolated from bacteria. *Escherichia coli*, the usual bacteria in molecular genetics experiments, has a large circular genome, but it will also replicate smaller circular DNAs as long as they have an "origin of replication". Plasmids may also have other DNA inserted by the investigator. A bacterium carrying a plasmid and replicating a million-fold will produce a million identical copies of that plasmid. Common plasmids are pBR322, pGEM, pUC18.

**PolyA tail:** After an mRNA is transcribed from a gene, the cell adds a stretch of A residues (typically 50-200) to its 3' end. It is thought that the presence of this "polyA tail" increases the stability of the mRNA (possibly by protecting it from nucleases). Note that not all mRNAs have a polyA tail; the histone mRNAs in particular do not.

**Polymerase:** An enzyme which links individual nucleotides together into a long strand, using another strand as a template. There are two general types of polymerase — DNA polymerases (which synthesize DNA) and RNA polymerase (which makes RNA). Within these two classes, there are numerous sub-types of polymerase, depending on what type of nucleic acid can function as template and what type of nucleic acid is formed. A DNA-dependant DNA polymerase will copy one DNA strand starting from a primer, and the product will be the complementary DNA strand. A DNA-dependant RNA polymerase will use DNA as a template to synthesize an RNA strand.

**Polymerase chain reaction:** A technique for replicating a specific piece of DNA *in-vitro*, even in the presence of excess non-specific DNA. Primers are added (which initiate the copying of each strand) along with nucleotides and Taq polymerase. By cycling the temperature, the target DNA is repetitively denatured and copied. A single copy of the target DNA, even if mixed in with other undesirable DNA, can be amplified to obtain billions of replicates. PCR can be used to amplify RNA sequences if they are first converted to DNA via reverse transcriptase. This two-phase procedure is known as 'RT-PCR'.

Polymerase Chain Reaction (PCR) is the basis for a number of extremely important methods in molecular biology. It can be used to detect and measure vanishingly small amounts of DNA and to create customized pieces of DNA. It has been applied to clinical diagnosis and therapy, to forensics and to vast numbers of research applications. It would be difficult to overstate the importance of PCR to science.

**Post-transcriptional regulation:** Any process occurring after transcription which affects the amount of protein a gene produces. Includes RNA processing efficiency, RNA stability, translation efficiency, protein stability. For example, the rapid degradation of an mRNA will reduce the amount of protein arising from it. Increasing the rate at which an mRNA is translated will increase the amount of protein product.

**Post-translational processing:** The reactions which alter a protein's covalent structure, such as phosphorylation, glycosylation or proteolytic cleavage.

**Post-translational regulation:** Any process which affects the amount of protein produced from a gene, and which occurs AFTER translation in the grand scheme of genetic expression. Actually, this is often just a buzzword for regulation of the stability of the protein. The more stable a protein is, the more it will accumulate.

**PRE: Progesterone Response Element:** A binding site in a promoter to which the activated progesterone receptor can bind. The progesterone receptor is essentially a transcription factor which is activated only in the presence of progesterone. The activated receptor will bind to a PRE, and transcription of the adjacent gene will be altered. See also "Response element".

**Primary transcript:** When a gene is transcribed in the nucleus, the initial product is the primary transcript, an RNA containing copies of all exons and introns. This primary transcript is then processed by the cell to remove the introns, to cleave off unwanted 3' sequence, and to polyadenylate the 5' end. The mature message thus formed is then exported to the cytoplasm for translation.

**Primer:** A small oligonucleotide (anywhere from 6 to 50 nt long) used to prime DNA synthesis. The DNA polymerases are only able to extend a pre-existing strand along a template; they are not able to take a naked single strand and produce a complementary copy of it *de-novo*. A primer which sticks to the template is therefore used to initiate the replication. Primers are necessary for DNA sequencing and PCR.

**Primer extension:** This is a method used to figure out how far upstream from a fixed site the start of an mRNA is. For example, perhaps you have isolated a cDNA clone, but you don't think that the clone has all of the 5' untranslated region. To find out how much is missing, you would first sequence the part you have, and figure out which strand is coding strand (usually the coding strand will have a large open reading frame). Next, you ask the DNA Synthesis Facility to make an oligonucleotide complementary to the 5'-most region of the coding strand (and thus complementary to the mRNA). This "primer" is hybridized to mRNA (say, a mixture of mRNA containing the one in which you are interested), and reverse transcriptase is added to copy the mRNA from the primer out to the 5' end. The size of the resulting DNA fragment shows how far away from the 5' end your primer is.

**Probe:** A fragment of DNA or RNA which is labeled in some way (often incorporating <sup>32</sup>P or <sup>35</sup>S), and which is used to hybridize with the nucleic acid in which you are interested. For example, if you want to quantitate the levels of alpha subunit mRNA in a preparation of pituitary RNA, you might make a radiolabeled RNA *in-vitro* which is complementary to the mRNA, and then use it to probe a Northern blot of the pit RNA. A probe can be radiolabeled, or tagged with another functional group such as biotin. A probe can be cloned DNA, or might be a synthetic DNA strand. As an example of the latter, perhaps you have isolated a protein for which you wish to obtain a cDNA or genomic clone. You might (pay to) microsequence a portion of the protein, deduce the nucleic acid sequence, (pay to) synthesize an oligonucleotide carrying that sequence, radiolabel it and use it as a probe to screen a cDNA library or genomic library. A better way is to call up someone who already has the clone.

**Processing:** The reactions occurring in the nucleus which convert the primary RNA transcript to a mature mRNA. Processing reactions include capping, splicing and polyadenylation. The term can also refer to the processing of the protein product, including proteolytic cleavages, glycosylation, etc.

**Promoter:** The first few hundred nucleotides of DNA "upstream" (on the 5' side) of a gene, which control the transcription of that gene. The promoter is part of the 5' flanking DNA, i.e. it is not transcribed into RNA, but without the promoter, the gene is not functional. Note that the definition is a bit hazy as far as the size of the region encompassed, but the "promoter" of a gene starts with the nucleotide immediately upstream from the cap site, and includes binding sites for one or more transcription factors which can not work if moved farther away from the gene.

**Proto-oncogene:** A gene present in a normal cell which carries out a normal cellular function, but which can become an oncogene under certain circumstances. The prefix "c-" indicates a cellular gene, and is generally used for proto-oncogenes (examples: *c-myb*, *c-myc*, *c-fos*, *c-jun*, etc).

**Pulsed field gel electrophoresis:** (PFGE) A gel technique which allows size-separation of very large fragments of DNA, in the range of hundreds of kb to thousands of kb. As in other gel electrophoresis techniques, populations of molecules migrate through the gel at a speed related to their size, producing discrete bands. In normal electrophoresis, DNA fragments greater than a certain size limit all migrate at the same rate through the gel. In PFGE, the electrophoretic voltage is applied alternately along two perpendicular axes, which forces even the larger DNA fragments to separate by size.

**Random primed synthesis:** If you have a DNA clone and you want to produce radioactive copies of it, one way is to denature it (separate the strands), then hybridize to that template a mixture of all possible 6-mer oligonucleotides. Those oligos will act as primers for the synthesis of labeled strands by DNA polymerase (in the presence of radiolabeled precursors).

**Reading frame:** When mRNA is translated by the cell, the nucleotides are read three at a time. By starting at different positions, the groupings of three that are produced can be entirely different.

If we shift the grouping again, we will just get the first reading frame again. The reading frame that is actually used is determined by the first methionine codon (the initiation codon). Once that first AUG is recognized, the pattern of triplet groupings follows unambiguously.

**Repetitive DNA:** A surprising portion of any genome consists not of genes or structural elements, but of frequently repeated simple sequences. These may be short repeats just a few nt long, like CACACA etc. They can also range up to a few hundred nt long. Examples of the latter include Alu repeats, LINEs, SINEs. The function of these elements is often unknown. In shorter repeats like di- and tri-nucleotide repeats, the number of repeating units can occasionally change during evolution and descent. They are thus useful markers for familial relationships and have been used in paternity testing, forensic science and in the identification of human remains.

**Response element:** By definition, a "response element" is a portion of a gene which must be present in order for that gene to respond to some hormone or other stimulus. Response elements are binding sites for transcription factors. Certain transcription factors are activated by stimuli such as hormones or heat shock. A gene may respond to the presence of that hormone because the gene has in its promoter region a binding site for hormone-activated transcription factor. Example: the glucocorticoid response element (GRE).

**Restriction:** To "restrict" DNA means to cut it with a restriction enzyme. See "Restriction Enzyme".

**Restriction enzyme:** A class of enzymes ("restriction endonucleases") generally isolated from bacteria, which are able to recognize and cut specific sequences ("restriction sites") in DNA. For example, the restriction enzyme BamHI locates and cuts any occurrence of:



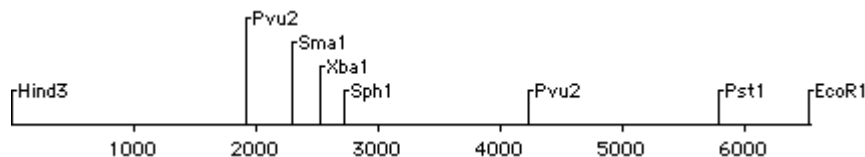
Note that both strands contain the sequence GGATCC, but in antiparallel orientation. The recognition site is thus said to be palindromic, which is typical of restriction sites. Every copy of a plasmid is identical in sequence, so if BamHI cuts a particular circular plasmid at three sites producing three "restriction fragments", then a million copies of that plasmid will produce those same restriction fragments a million times over. There are more than six hundred known restriction enzymes.

Bacteria produce restriction enzymes for protection against invasion by foreign DNA such as phages. The bacteria's own DNA is modified in such a way as to prevent it from being clipped.

**Restriction fragment:** The piece of DNA released after restriction digestion of plasmids or genomic DNA. See "Restriction enzyme". One can digest a plasmid and isolate one particular restriction fragment (actually a set of identical fragments). The term also describes the fragments detected on a genomic blot which carry the gene of interest.

**Restriction fragment length polymorphism:** See "RFLP".

**Restriction map:** A "cartoon" depiction of the locations within a stretch of known DNA where restriction enzymes will cut.



The map usually indicates the approximate length of the entire piece (scale on the bottom), as well as the position within the piece at which designated enzymes will cut. This map happens to be of a plasmid, and the two ends are joined together with about 25 nt between the EcoRI and HindIII sites.

**Restriction site:** See Restriction enzyme.

**Reverse transcriptase:** An enzyme which will make a DNA copy of an RNA template - a DNA-dependant RNA polymerase. RT is used to make cDNA; one begins by isolating polyadenylated mRNA, providing oligo-dT as a primer, and adding nucleotide triphosphates and RT to copy the RNA into cDNA.

**RFLP:** Restriction fragment length polymorphism; the acronym is pronounced "riflip". Although two individuals of the same species have almost identical genomes, they will always differ at a few nucleotides. Some of these differences will produce new restriction sites (or remove them), and thus the banding pattern seen on a genomic Southern will thus be affected. For any given probe (or gene), it is often possible to test different restriction enzymes until you find one which gives a pattern difference between two individuals - a RFLP. The less related the individuals, the more divergent their DNA sequences are and the more likely you are to find a RFLP.

**Ribonuclease:** see "RNAse".

**Riboprobe:** A strand of RNA synthesized *in-vitro* (usually radiolabeled) and used as a probe for hybridization reactions. An RNA probe can be synthesized at very high specific activity, is single stranded (and therefore will not self anneal), and can be used for very sensitive detection of DNA or RNA.

**Ribosome:** A cellular particle which is involved in the translation of mRNAs to make proteins. Ribosomes are a complex consisting of ribosomal RNAs (rRNA) and several proteins.

**RNAi:** 'RNA interference' (a.k.a. 'RNA silencing') is the mechanism by which small double-stranded RNAs can interfere with expression of any mRNA having a similar sequence. Those small RNAs are known as 'siRNA', for short interfering RNAs. The mode of action for siRNA appears to be via dissociation of its strands, hybridization to the target RNA, extension of those fragments by an RNA-dependent RNA polymerase, then fragmentation of the target. Importantly, the remnants of the target molecule appears to then act as an siRNA itself; thus the effect of a small amount of starting siRNA is effectively amplified and can have long-lasting effects on the recipient cell.

The RNAi effect has been exploited in numerous research programs to deplete the call of specific messages, thus examining the role of those messages by their absence.

**RNAse:** Ribonuclease; an enzyme which degrades RNA. It is ubiquitous in living organisms and is exceptionally stable. The prevention of RNAse activity is the primary problem in handling RNA.

**RNAse protection assay:** This is a sensitive method to determine (1) the amount of a specific mRNA present in a complex mixture of mRNA and/or (2) the sizes of exons which comprise the mRNA of interest. A radioactive DNA or RNA probe (in excess) is allowed to hybridize with a sample of mRNA (for example, total mRNA isolated from tissue), after which the mixture is digested with single-strand specific nuclease. Only the probe which is hybridized to the specific mRNA will escape the nuclease treatment, and can be detected on a gel. The amount of radioactivity which was protected from nuclease is proportional to the amount of mRNA to which it hybridized. If the probe included both intron and exons, only the exons will be protected from nuclease and their sizes can be ascertained on the gel.

**rRNA:** "ribosomal RNA"; any of several RNAs which become part of the ribosome, and thus are involved in translating mRNA and synthesizing proteins. They are the most abundant RNA in the cell (on a mass basis).

**RT-PCR:** See 'Polymerase Chain Reaction'.

**Run-off:** see Nuclear run-on.

**Run-on:** see Nuclear run-on.

**S1 end mapping:** A technique to determine where the end of an RNA transcript lies with respect to its template DNA (the gene). Can't be described in a short paragraph. See "RNAse Protection assay" for a closely related technique.

**S1 nuclease:** An enzyme which digests only single-stranded nucleic acids.

**Screening:** To screen a library (see "Library") is to select and isolate individual clones out of the mixture of clones. For example, if you needed a cDNA clone of the pituitary glycoprotein hormone alpha subunit, you would need to make (or buy) a pituitary cDNA library, then screen that library in order to detect and isolate those few bacteria carrying alpha subunit cDNA.



**Slot blot:** Similar to a dot blot, but the analyte is put onto the membrane using a slot-shaped template. The template produces a consistently shaped spot, thus decreasing errors and improving the accuracy of the analysis. See Dot blot.

**snRNA:** Small nuclear RNA; forms complexes with proteins to form snRNPs; involved in RNA splicing, polyadenylation reactions, other unknown functions (probably).

**snRNP:** "snerps", Small Nuclear RiboNucleoProtein particles, which are complexes between small nuclear RNAs and proteins, and which are involved in RNA splicing and polyadenylation reactions.

**SNP:** Single Nucleotide Polymorphism (SNP) - a position in a genomic DNA sequence that varies from one individual to another. It is thought that the primary source of genetic difference between any two humans is due to the presence of single nucleotide polymorphisms in their DNA. Furthermore, these SNPs can be extremely useful in genetic mapping (see 'Genetic Mapping') to follow inheritance of specific segments of DNA in a lineage. SNP-typing is the process of determining the exact nucleotide at positions known to be polymorphic.

**Solution hybridization:** A method closely related to RNase protection (see "RNase protection assay"). Solution hybridization is designed to measure the levels of a specific mRNA species in a complex population of RNA. An excess of radioactive probe is allowed to hybridize to the RNA, then single-strand specific nuclease is used to destroy the remaining unhybridized probe and RNA. The "protected" probe is separated from the degraded fragments, and the amount of radioactivity in it is proportional to the amount of mRNA in the sample which was capable of hybridization. This can be a very sensitive detection method.

**Southern blot:** A technique for analyzing mixtures of DNA, whereby the presence and rough size of one particular fragment of DNA can be ascertained. See "Blotting". Named for its inventor, Dr E. M. Southern.

**SSR:** Simple Sequence Repeat. See 'Microsatellite'.

**Stable transfection:** A form of transfection experiment designed to produce permanent lines of cultured cells with a new gene inserted into their genome. Usually this is done by linking the desired gene with a "selectable" gene, i.e. a gene which confers resistance to a toxin (like G418, aka Geneticin). Upon putting the toxin into the culture medium, only those cells which incorporate the resistance gene will survive, and essentially all of those will also have incorporated the experimenter's gene.

**Sticky ends:** After digestion of a DNA with certain restriction enzymes, the ends left have one strand overhanging the other to form a short (typically 4 nt) single-stranded segment. This overhang will easily re-attach to other ends like it, and are thus known as "sticky ends". For example, the enzyme BamHI recognizes the sequence GGATCC, and clips after the first G in each strand:



The overhangs thus produced can still hybridize ("anneal") with each other, even if they came from different parent DNA molecules, and the enzyme ligase will then covalently link the strands. Sticky ends therefore facilitate the ligation of diverse segments of DNA, and allow the formation of novel DNA constructs.

**Stringency:** A term used to describe the conditions of hybridization. By varying the conditions (especially salt concentration and temperature) a given probe sequence may be allowed to hybridize only with its exact complement (high stringency), or with any somewhat related sequences (relaxed or low stringency). Increasing the temperature or decreasing the salt concentration will tend to increase the selectivity of a hybridization reaction, and thus will raise the stringency.

**Sub-cloning:** If you have a cloned piece of DNA (say, inserted into a plasmid) and you need unlimited copies of only a part of it, you might "sub-clone" it. This involves starting with several million copies of the original plasmid,

cutting with restriction enzymes, and purifying the desired fragment out of the mixture. That fragment can then be inserted into a new plasmid for replication. It has now been subcloned.

**Taq polymerase:** A DNA polymerase isolated from the bacterium *Thermophilis aquaticus* and which is very stable to high temperatures. It is used in PCR procedures and high temperature sequencing.

**TATA box:** A sequence found in the promoter (part of the 5' flanking region) of many genes. Deletion of this site (the binding site of transcription factor TFIIID) causes a marked reduction in transcription, and gives rise to heterogeneous transcription initiation sites.

**Tet resistance:** See "Antibiotic resistance".

**Tissue-specific expression:** Gene function which is restricted to a particular tissue or cell type. For example, the glycoprotein hormone alpha subunit is produced only in certain cell types of the anterior pituitary and placenta, not in lungs or skin; thus expression of the glycoprotein hormone alpha-chain gene is said to be tissue-specific. Tissue specific expression is usually the result of an enhancer which is activated only in the proper cell type.

**Tm:** The melting point for a double-stranded nucleic acid. Technically, this is defined as the temperature at which 50% of the strands are in double-stranded form and 50% are single-stranded, i.e. midway in the melting curve. A primer has a specific Tm because it is assumed that it will find an opposite strand of appropriate character.

**Transcription factor:** A protein which is involved in the transcription of genes. These usually bind to DNA as part of their function (but not necessarily). A transcription factor may be general (i.e. acting on many or all genes in all tissues), or tissue-specific (i.e. present only in a particular cell type, and activating the genes restricted to that cell type). Its activity may be constitutive, or may depend on the presence of some stimulus; for example, the glucocorticoid receptor is a transcription factor which is active only when glucocorticoids are present.

**Transcription:** The process of copying DNA to produce an RNA transcript. This is the first step in the expression of any gene. The resulting RNA, if it codes for a protein, will be spliced, polyadenylated, transported to the cytoplasm, and by the process of translation will produce the desired protein molecule.

**Transfection:** A method by which experimental DNA may be put into a cultured mammalian cell. Such experiments are usually performed using cloned DNA containing coding sequences and control regions (promoters, etc) in order to test whether the DNA will be expressed. Since the cloned DNA may have been extensively modified (for example, protein binding sites on the promoter may have been altered or removed), this procedure is often used to test whether a particular modification affects the function of a gene.

**Transformation (with respect to bacteria):** The process by which a bacteria acquires a plasmid and becomes antibiotic resistant. This term most commonly refers to a bench procedure performed by the investigator which introduces experimental plasmids into bacteria.

**Transformation (with respect to cultured cells):** A change in cell morphology and behavior which is generally related to carcinogenesis. Transformed cells tend to exhibit characteristics known collectively as the "transformed phenotype" (rounded cell bodies, reduced attachment dependence, increased growth rate, loss of contact inhibition, etc). There are different "degrees" of transformation, and cells may exhibit only a subset of these characteristics. Not well understood, the process of transformation is the subject of intense research.

**Transgenic mouse:** A mouse which carries experimentally introduced DNA. The procedure by which one makes a transgenic mouse involves the injection of DNA into a fertilized embryo at the pro-nuclear stage. The DNA is generally cloned, and may be experimentally altered. It will become incorporated into the genome of the embryo. That embryo is implanted into a foster mother, who gives birth to an animal carrying the new gene. Various experiments are then carried out to test the functionality of the inserted DNA.

**Transient transfection:** When DNA is transfected into cultured cells, it is able to stay in those cells for about 2-3 days, but then will be lost (unless steps are taken to ensure that it is retained - see Stable transfection). During those 2-3 days, the DNA is functional, and any functional genes it contains will be expressed. Investigators take advantage of this transient expression period to test gene function.

**Translation:** The process of decoding a strand of mRNA, thereby producing a protein based on the code. This process requires ribosomes (which are composed of rRNA along with various proteins) to perform the synthesis, and tRNA to bring in the amino acids. Sometimes, however, people speak of "translating" the DNA or RNA when they are merely reading the nucleotide sequence and predicting from it the sequence of the encoded protein. This might be more accurately termed "conceptual translation".

**Tumor suppressor:** A gene that inhibits progression towards neoplastic transformation. The best-known examples of tumor suppressors are the proteins **p53** and **Rb**.

**tRNA:** "transfer RNA"; one of a class of rather small RNAs used by the cell to carry amino acids to the enzyme complex (the ribosome) which builds proteins, using an mRNA as a guide. Fairly abundant.

**Upstream activator sequence:** A binding site for transcription factors, generally part of a promoter region. A UAS may be found upstream of the TATA sequence (if there is one), and its function is (like an enhancer) to increase transcription. Unlike an enhancer, it can not be positioned just anywhere or in any orientation.

**Upstream/Downstream:** In an RNA, anything towards the 5' end of a reference point is "upstream" of that point. This orientation reflects the direction of both the synthesis of mRNA, and its translation - from the 5' end to the 3' end. In DNA, the situation is a bit more complicated. In the vicinity of a gene (or in a cDNA), the DNA has two strands, but one strand is virtually a duplicate of the RNA, so its 5' and 3' ends determine upstream and downstream, respectively. NOTE that in genomic DNA, two adjacent genes may be on different strands and thus oriented in opposite directions. Upstream or downstream is only used on conjunction with a given gene.

**Vector:** The DNA "vehicle" used to carry experimental DNA and to clone it. The vector provides all sequences essential for replicating the test DNA. Typical vectors include plasmids, cosmids, phages and YACs.

**Western blot:** A technique for analyzing mixtures of proteins to show the presence, size and abundance of one particular type of protein. Similar to Southern or Northern blotting (see "Blotting"), except that (1) a protein mixture is electrophoresed in an acrylamide gel, and (2) the "probe" is an antibody which recognizes the protein of interest, followed by a radioactive secondary probe (such as <sup>125</sup>I-protein A).

**YAC:** Yeast artificial chromosome. This is a method for cloning very large fragments of DNA. Genomic DNA in fragments of 200-500 kb are linked to sequences which allow them to propagate in yeast as a mini-chromosome (including telomeres, a centromere and an ARS - an autonomous replication sequence). This technique is used to clone large genes and intergenic regions, and for chromosome walking.

**Zinc finger:** A protein structural motif common in DNA binding proteins. Four Cys residues are found for each "finger" and one finger can bind a molecule of zinc. A typical configuration is:  
CysXxxXxxCys--(intervening 12 or so aa's)--CysXxxXxxCys.