Principles of DNA cloning

John S Miles, C Roland Wolf

The nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), are common to all living organisms. It is the DNA within a person's genes that holds the unique blueprint necessary for that person's growth, differentiation, development, and reproduction. Consideration of the so called "central dogma" of molecular biology—that is, the transmission of information from DNA to RNA to protein (fig 1)—shows why this is so. DNA sequences (genes) encode

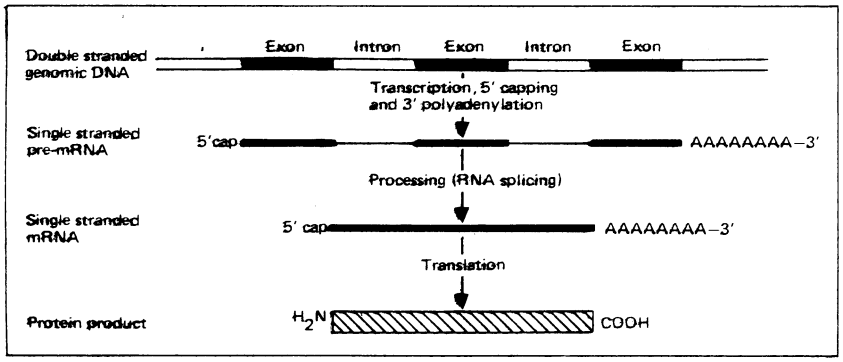

FIG 1—*Genes contained within the DNA are transcribed into premessenger RNA and this is processed to give mature messenger RNA, which is then translated into protein products that determine the functions expressed within each cell*

proteins and therefore ultimately determine the essential functions of each cell as well as the specific functions that are characteristic of each cell type. Changes in function due to mutations carried in the germline DNA represent the molecular basis for inherited diseases such as sickle cell disease, phenylketonuria, and Duchenne type muscular dystrophy[1-3] whereas somatic cell mutations can cause changes in cellular function which may lead to certain types of cancer.[4]

The ability to isolate and clone specific DNA molecules has brought about a revolution in medical genetics, cell and molecular biology, and biochemistry in the past 15 years. These cloned DNA molecules can be used as probes to detect mutations and polymorphisms associated with disease[5] and unequivocally to determine family pedigrees by Southern blotting and hybridisation methods.[6,7] They are also useful in studying tissue specific genes and temporal expression of selected genes by northern blotting[7] and in identifying factors that regulate gene expression. Sequencing DNA can disclose the primary structure of the protein that it encodes and has given new insights into the relation, function, and evolution of numerous genes and proteins. Similarly, the introduction and expression of cloned genes in suitable host cells or animals (transgenic animals) has helped to establish the function of protein products that may otherwise be unsuitable for classical biochemical studies because of their low abundance or instability.

DNA cloning has touched, or has the potential for touching, almost every aspect of biology and medicine. This article aims at outlining the principles of DNA cloning, and more detailed accounts of the principles and practice of DNA cloning can be found elsewhere.[8,9]

Imperial Cancer Research Fund Laboratory of Molecular Pharmacology and Drug Metabolism, Department of Biochemistry, University of Edinburgh, Edinburgh EH8 9XD
John S Miles, PHD, *research fellow*
C Roland Wolf, PHD, *head*

This series has been edited by Drs P C Hayes and J D Hayes, Royal Infirmary, Edinburgh.

Br Med J 1989;299:1019-22

What is DNA cloning?

Cloning is the method whereby individual DNA molecules can be faithfully copied in vivo to produce numerous identical molecules or clones. The DNA molecule containing the sequence of interest can be isolated from a mixture and propagated in a suitable host cell. This yields a limitless supply of this molecule and enables the gene or genes contained within it or any protein product to be studied. The universality of DNA as the genetic material (except in retroviruses, which contain RNA) means that DNA from one organism can be copied accurately and maintained in another.

GENERAL STRATEGY

A general strategy for DNA cloning is outlined in figure 2. Linear DNA molecules from a suitable source are mixed with vector DNA and joined together (ligated) with a specific enzyme (DNA ligase). This procedure generates a library of recombinant DNA molecules. Recombinant simply means that DNA from one source has been joined to DNA from another source. The vector contains DNA sequences that enable it and the accompanying DNA to be copied (replicated) after entering a suitable host cell. The most common host cell is a laboratory (non-pathogenic) strain of the gut bacterium *Escherichia coli*, although bakers' yeast (*Saccharomyces cerevisiae*) and other micro-organisms are also used. The host cell provides the enzymatic machinery necessary for generating copies of the recombinant DNA molecule. Vectors are usually either plasmids (circular DNA molecules that can replicate independently of the bacterial chromosome and subsequently can be present in many copies in each cell) or bacteriophages, which are essentially bacterial viruses.

Host cells, each containing a single representative from the library of recombinants, are grown and

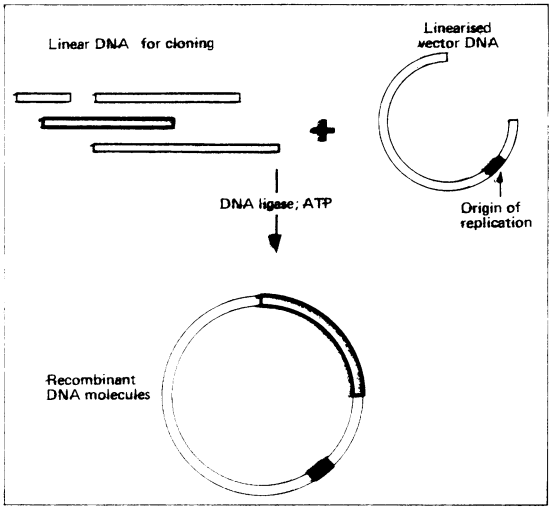

FIG 2—*General strategy for DNA cloning (with plasmid vector). Linear fragments of DNA for cloning (usually produced by digestion with restriction enzymes) are joined to vector DNA molecules in vitro with the enzyme DNA ligase. As the vector is a plasmid the resulting recombinant DNA molecules are circular. Plasmids contain an origin of replication which allows them to be copied by the host cell*

multiplied on agar plates. Plasmid vectors containing a gene that encodes a protein that confers antibiotic resistance are used so that *E coli* host cells are grown as colonies on a selective medium (a medium containing antibiotic). On the other hand, bacteriophage infection produces clearings (plaques) in a lawn of bacterial host cells as the virus reproduces, lyses the cell, and infects surrounding cells.

Each bacterial colony or bacteriophage plaque contains many thousands of copies of a single recombinant DNA molecule. When the bacterial cell that contains the cloned DNA of interest has been selected by one or more of the procedures described in more detail below, it can be grown, and large quantities (amounts measurable in μg or mg) can be isolated fairly easily. This DNA can be used directly for experiments. More commonly it is transferred (subcloned) into other vectors that have been designed for a special purpose, such as DNA sequencing or production of DNA probes.

RESTRICTION ENZYMES

Probably the most important tools for DNA cloning are restriction enzymes, which recognise specific DNA sequences and cut the DNA within or near the recognition site. For example, the enzyme *Eco*RI, isolated from a strain of *E coli*, recognises the sequence

```
      ↓
5' - GAATTC - 3'
3' - CTTAAG - 5'
          ↑
```

and cuts the molecule as indicated by the arrows. A large series of such enzymes is available, and they recognise well over 100 different target sites. DNA cut by many of these enzymes has short, overhanging, single stranded DNA ends called cohesive termini or "sticky ends." Thus DNA cut with *Eco*RI has the following sticky ends:

```
5' - G        and AATTC - 3'
3' - CTTAA          G - 5'.
```

These two ends can be joined back together again in vitro by ligation because of complementarity: the ability of the single stranded overhangs to base pair.

DNA from any source which contains *Eco*RI recognition sites can be cut and joined in vitro to another DNA molecule from another source which has been cut by *Eco*RI. Restriction enzymes are normally used at some stage to generate the linear DNA and linearised vector DNA that are joined to form the recombinant molecules as described in figure 2.

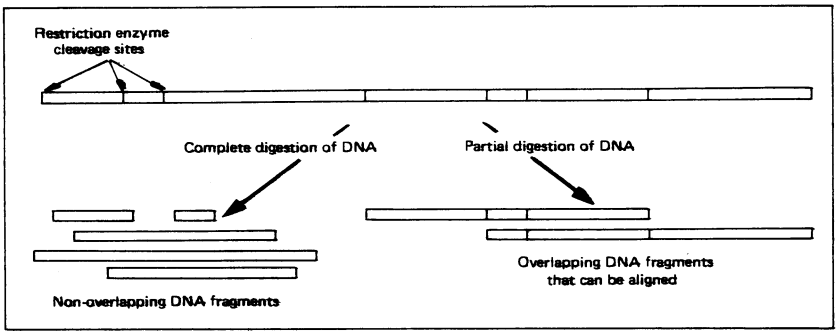

FIG 3—*Complete digestion of DNA with a restriction enzyme produces non-overlapping DNA fragments, and all information as to the order of the fragments is lost, whereas partial digestion yields overlapping fragments of DNA to be cloned in the library; these can be ordered later into a physical map*

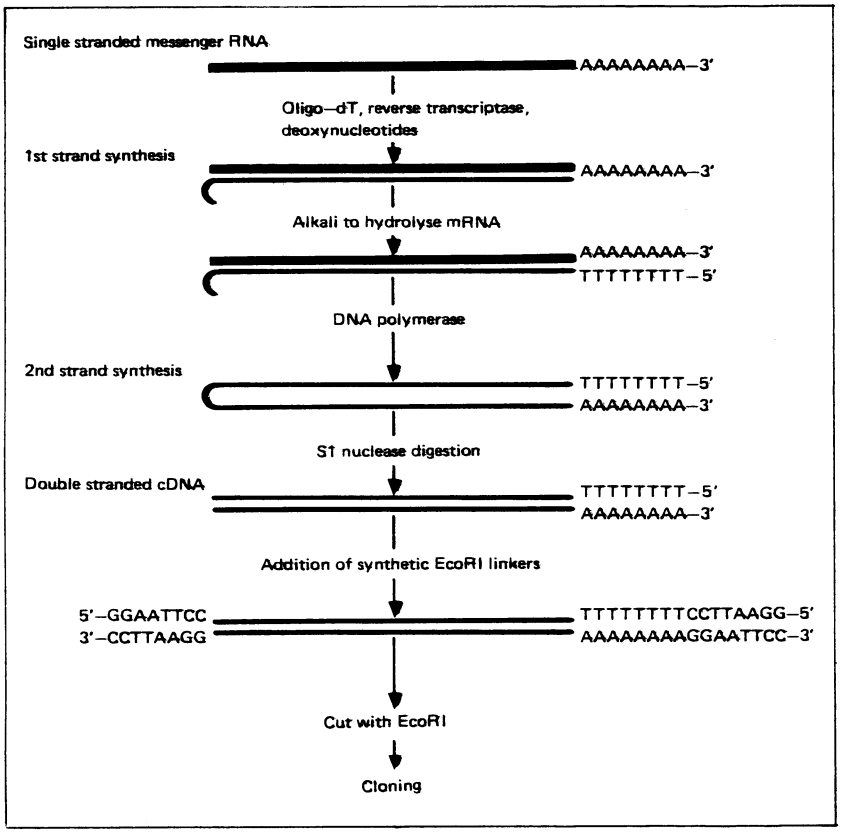

FIG 4—*Conversion of single stranded mature messenger RNA into double stranded complementary DNA in vitro with purified enzymes to catalyse the reactions*

Types of recombinant DNA libraries

The complexity of human DNA is such that there are about $3{\cdot}3 \times 10^9$ base pairs per haploid genome. The capacity of bacterial cloning vectors in common use is up to 45×10^3 base pairs (45 kilobase (kb) pairs). This means that for DNA fragments of around 20 kb (generated by restriction enzyme digests of human DNA) $7{\cdot}6 \times 10^5$ recombinant molecules are required to give a 99% chance of having cloned a particular DNA sequence.[8] Artificial chromosomes are currently being developed for yeast to increase the amount of DNA that can be cloned and consequently to reduce the number of recombinants required to generate a representative library.

When total genomic DNA is used to produce a collection of recombinant molecules, this is called a genomic library (of genomic clones). Because, in general, each somatic cell from a person contains the same complement of DNA, genomic libraries are usually made from a readily available source such as leucocyte DNA.

Only a small proportion of the total amount of DNA codes for proteins, partly because of intervening sequences (introns) interspersed within the coding sequences (exons) of genes (fig 1). After gene transcription and RNA processing, however, all of the information required to code for proteins resides in the messenger RNA. To make this information accessible DNA copies of the messenger RNA can be made in vitro and cloned. A library of these copied molecules of complementary DNA reflects the variety and number of messenger RNAs present in the starting population of the cells under study. The source of messenger RNA used to produce a complementary DNA library is important because not all genes are expressed (and hence not all messenger RNAs are present) in all tissues or cells, and not all genes are expressed all of the time within the same tissue or cell type. Therefore complementary DNA libraries are usually made from a specific tissue or cell type, chosen because the protein encoded by the desired messenger RNA is known to be present.

BUILDING A LIBRARY OF GENOMIC DNA

To construct a library of genomic clones it is necessary to cut the genomic DNA with restriction enzymes into small enough pieces for cloning. If the DNA is cut at every possible restriction site before

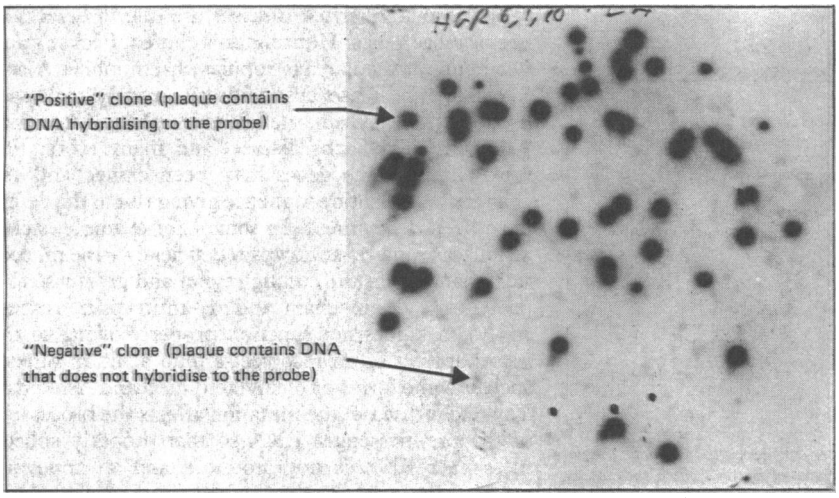

FIG 5—*Autoradiograph of a membrane support bearing recombinant DNA from bacteriophage plaques which has been hybridised to a DNA probe. Dark spots are "positive" clones (containing the target DNA sequence); light spots are "negative" clones (containing unrelated DNA sequences)*

cloning all information as to the order of the DNA pieces is lost (fig 3). The usual procedure to circumvent this problem is to cut the DNA partially (randomly but not at every site) with an enzyme that has a large number of recognition sites within the DNA. For example, the enzyme *Sau*3A recognises the tetranucleotide sequence

GATC
CTAG

and so on average will cut the DNA every 256 base pairs. The partially cut DNA is fractionated by size with agarose gel electrophoresis,[10] and fragments of a suitable size are isolated and cloned by the strategy shown in figure 2. With this procedure the genomic library will be made up of clones that contain DNA that overlaps with other clones within the library. This is important for ordering the isolated clones and generating a physical restriction map of the DNA of interest (fig 3).

BUILDING A LIBRARY OF COMPLEMENTARY DNA

The construction of a library of complementary DNA requires generating a double stranded complementary DNA molecule from single stranded messenger RNA before the cloning step. A general procedure for generating complementary DNA molecules is outlined in figure 4. All, or almost all, eukaryotic messenger RNAs have a tail of adenosine molecules (poly-A tail) at their 3′ end. An oligo-dT primer (a short synthetic single stranded DNA molecule made up solely of thymine residues) is base paired to this tail and is used to initiate synthesis of the first DNA strand. An enzyme, reverse transcriptase, isolated from cells infected with a retrovirus, can synthesise a single strand DNA copy of the RNA using deoxynucleotides. Alkaline hydrolysis removes the RNA from these hybrid DNA-RNA duplex molecules. The 5′ end of the first DNA strand can initiate synthesis of its own second strand, and the enzyme DNA polymerase carries out this reaction. S1-nuclease digests the small loop of DNA present after this procedure and leaves a blunt ended complementary DNA molecule with no sticky ends. Lastly, short synthetic "linker" DNA molecules containing restriction enzyme sites are ligated to the complementary DNA. These sites are then cut by the appropriate enzyme, and the DNA is ready for cloning by ligating into a vector cut with the same restriction enzyme.

Synthesis of complementary DNAs by this procedure does not always ensure the generation of a full size complementary copy of the messenger RNA. This is because the reverse transcriptase enzyme does not always get as far as the 5′ end of the messenger RNA before dropping off. Several more complex procedures have been devised to maximise the chance of generating full length complementary DNAs that better reflect the original messenger RNA.[8]

SELECTING THE REQUIRED RECOMBINANT DNA CLONE

A library of recombinant DNA molecules may contain many tens of thousands of distinct clones, but only a few of these clones may contain the sequence of interest. All sequences of genomic DNA should be represented equally within the library. In complementary DNA libraries there will be more clones corresponding to abundant messenger RNAs and fewer clones for rare messenger RNAs. Either way, powerful identification procedures are required to isolate the desired clone.

SCREENING WITH DNA-DNA HYBRIDISATION

One of the most widespread methods for selecting a clone from both genomic DNA and complementary DNA libraries uses DNA-DNA hybridisation (fig 5). Bacterial colonies or bacteriophage plaques containing the recombinant DNA molecules are grown on agar or bacterial lawns respectively, and DNA is transferred from these clones to a membrane (nitrocellulose or nylon filter) to provide a replica. The DNA is denatured to form single strands on the membrane, and then a radioactively labelled probe of single stranded DNA, specifically designed to identify the clone of interest, is hybridised to the DNA on the membrane. The base pairs of the probe will match only those of the relevant clone so that its location can be deduced by autoradiography of the membrane to detect the hybridisation signal. Alignment of the membrane replica with the original agar plate allows the plaque or colony containing the DNA of interest to be identified. The clone can be isolated, purified, and grown to provide an unlimited supply of DNA.

Synthetic single stranded DNA molecules (oligonucleotides) of known sequence provide a good way of selecting a clone encoding a protein for which there is some amino acid sequence available. This is because if the amino acid sequence is known it is possible to predict the DNA sequences that will code for it. As most amino acids are encoded by more than one base triplet (codon) a mixture of oligonucleotides is often used (figure 6).

An alternative approach using DNA probes is to go between species to find equivalent genes. Equivalent genes between, for example, rat and man commonly have more than 60% identical nucleotide sequences, which can, under suitable conditions, permit base pairing between DNAs. Hence rat genes can be used as hybridisation probes for isolating human genes.

ANTIBODY SCREENING

A selection strategy that is suitable for use only with complementary DNA libraries is that of antibody

Met	Glu	Lys	Trp	Ser	Gly	Protein sequence
ATG	GAA	AAA	TGG	TCT	GGT	Possible nucleotide sequences
	G	G		C	C	
				A	A	
				G	G	
				AGT		
				AGC		
1	2	2	1	6	4	=96 Combinations

FIG 6—*Amino acid sequence information for designing oligonucleotide probes. One base triplet (codon) encodes the amino acids methionine (Met) and tryptophan (Trp), two encode glutamate (Glu) and lysine (Lys), four encode glycine (Gly), and six encode serine (Ser); thus 96 possible oligonucleotides could encode this sequence of six amino acids*

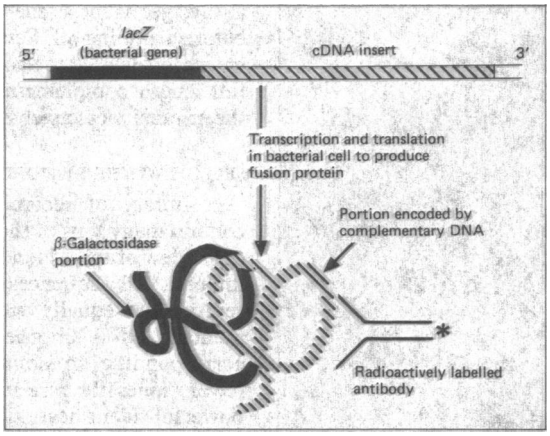

FIG 7—*Recombinant DNA libraries can be made in which complementary DNA is fused to the lacZ gene of E coli, which encodes the enzyme β-galactosidase. The hybrid genes can be transcribed and translated in E coli to form a fusion protein partly of β-galactosidase and partly of protein encoded by complementary DNA. The fusion proteins are immobilised on to membrane supports and radioactively labelled antibody (against the protein for which the complementary DNA is required) is used as a probe*

screening. The principle of antibody screening is outlined in figure 7. It relies on the complementary DNA library being constructed in a special vector that generates a hybrid between the bacterial gene and the complementary DNA, which is expressed and results in the production of a fusion protein, part of which is encoded by the bacterial gene and part by the complementary DNA. Fusion proteins generated by the bacterial colonies are immobilised on to a membrane in much the same way as for DNA, and the membranes are probed with radioactive antibodies directed towards the protein of interest. Clones containing complementary DNAs encoding the protein will produce a fusion protein that will be able to crossreact with the specific antibody and thus give a signal on autoradiography.

COMPLEMENTATION

Complementation as a means of selection requires the expression of a complementary DNA sequence to give a functional (protein) product. This is because the host cell chosen is defective in the particular function of interest—that is, it contains a mutation in the gene encoding the function. Thus growth of the host cell relies on a functional copy of a gene being introduced (as a recombinant DNA molecule) to compensate for the defective gene and hence restore its function. Complementation has been used successfully across species, such that human genes (complementary DNAs) encoding vital control functions of the cell cycle have been able to restore defects in mutant yeast host cells.[11]

OTHER METHODS OF SELECTION

Several other methods of selection have been developed which rely on a series of other biochemical and genetic criteria. Most DNA sequences can be cloned and isolated given enough ingenuity, time, and luck.

Conclusion

Cloned DNA sequences provide the tools by which many biological processes can be studied, and our understanding of cell and molecular biology has expanded enormously by their use. Expression of cloned genes has allowed many different pharmacoactive proteins—for example, insulin, growth hormone, and factor VIII— to be produced on a scale and to a purity not seen before. Recombinant DNA technology will become more important pharmaceutically as new therapeutic protein and peptide molecules are made available.

Numerous inherited diseases are due to defects in one or more genes: Huntington's chorea, Becker's and Duchenne muscular dystrophies, haemophilia A and B, sickle cell disease, cystic fibrosis, α and β thalassaemias, α_1-antitrypsin deficiency, phenylketonuria, Ashkenazi Tay-Sachs disease, and many more. For some of these the genes have been cloned and the changes in nucleotide sequence giving rise to the defect have been determined. In some—for example, sickle cell disease and α_1-antitrypsin deficiency—the nucleotide change is in the coding region and gives rise to a protein (β globin chain and α_1-antitrypsin, respectively) that does not function properly owing to the incorporation of an incorrect amino acid. In others, such as some forms of phenylketonuria and Ashkenazi Tay-Sachs disease, the mutation affects the processing of the pre-messenger RNA so that properly spliced messenger RNA is not produced and an abnormal protein product, or none, is generated. Deletions of different sized parts of the dystrophin gene give rise to incorrect protein products and result in Duchenne and Becker's muscular dystrophies.

The use of cloned DNA probes in medical genetics permits prenatal and presymptomatic diagnosis of many inherited diseases. Diagnosis is most accurate when the nucleotide sequence changes causing the defect are known. Recent technical advances have led to the development of a method, the polymerase chain reaction,[12] that allows very small amounts of specific target DNA (such as genomic DNA isolated from blood samples and blood stains, mouth washes, or semen samples) to be amplified millions of times in vitro. Amplification of, for example, the β globin gene followed by DNA sequencing or restriction enzyme analysis would determine whether a person's genes carry the mutation for sickle cell disease,[13] and the method can be applied to many other diseases. Changes in the structure and sequence of DNA within somatic cells that occur during oncogenesis (the activation, through mutation, of certain oncogenes) can be studied much more readily with this powerful new procedure. Because DNA can be amplified from minute quantities of starting material, it may also be possible to screen embryos produced by in vitro fertilisation before implantation, using DNA from a single cell.

Cloning of the gene responsible for an inherited disease such as cystic fibrosis,[14] in which little is known about the underlying biochemical processes, will provide new approaches to the treatment of the disease. It may even be possible one day to repair defective genes by introducing cloned DNA molecules into affected people.

1 Kan YW, Dozy AM. Antenatal diagnosis of sickle cell anaemia by DNA analysis of amniotic fluid cells. *Lancet* 1978;ii:910-2.
2 Woo SL, Lidsky AS, Guttler F, Chandra T, Robson JH. Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. *Nature* 1983;**306**:234-8.
3 Monaco A, Neve R, Colletti-Feener C, Bertelson C, Kurnit D, Kunkel L. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* 1986;**323**:646-50.
4 Solomon E, Voss R, Hall V, *et al*. Chromosome 5 allele loss in human colorectal carcinoma. *Nature* 1987;**328**:616-9.
5 Solomon E, Bodmer WF. Evolution of sickle variant gene. *Lancet* 1979;i:923.
6 Jeffreys AJ. Highly variable minisatellites and DNA fingerprints. *Biochem Soc Trans* 1987;**15**:309-17.
7 Hayes PC, Wolf CR, Hayes JD. Blotting techniques for the study of DNA, RNA, and proteins. *Br Med J* 1989;**299**:965-8.
8 Glover DM, ed. *DNA cloning*. Vol I. *A practical approach*. Oxford: IRL Press, 1985.
9 Maniatis T, Fritsch EF, Sambrook J. *Molecular cloning, a laboratory manual*. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1982.
10 Hayes JD, Stockman PK. Electrophoresis of proteins and nucleic acids. II. Techniques and applications. *Br Med J* 1989;**299**:907-10.
11 Lee MG, Nurse P. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. *Nature* 1987;**327**:31-5.
12 Saiki RK, Gelfand DH, Stoffel S, *et al*. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988;**239**:487-91.
13 Saiki RK, Scharf S, Faloona F, *et al*. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 1985;**230**:1350-4.
14 Estivill X, Farrall M, Scambler PJ, *et al*. A candidate for the cystic fibrosis locus isolated by selection for methylation-free islands. *Nature* 1987; **326**:840-5.