CHAPTER 6
Cloning strategies

Introduction

In the early chapters of this book, we discussed DNA cutting and joining techniques, and introduced the different types of vectors that are available for cloning DNA molecules. Any cell-based cloning procedure has four essential parts: (i) a method for generating the DNA fragment for cloning; (ii) a reaction that inserts that fragment into the chosen cloning vector; (iii) a means for introducing that recombinant vector into a host cell wherein it is replicated; and (iv) a method for selecting recipient cells that have acquired the recombinant (Fig. 6.1). To simplify the description of such procedures, the assumption is made that we know exactly what we are cloning. This is indeed the case with simple *subcloning* strategies, where a defined restriction fragment is isolated from one cloning vector and inserted into another. However, we also need to consider what happens in cases where the source of donor DNA is very complex. We may wish, for example, to isolate a single gene from the human

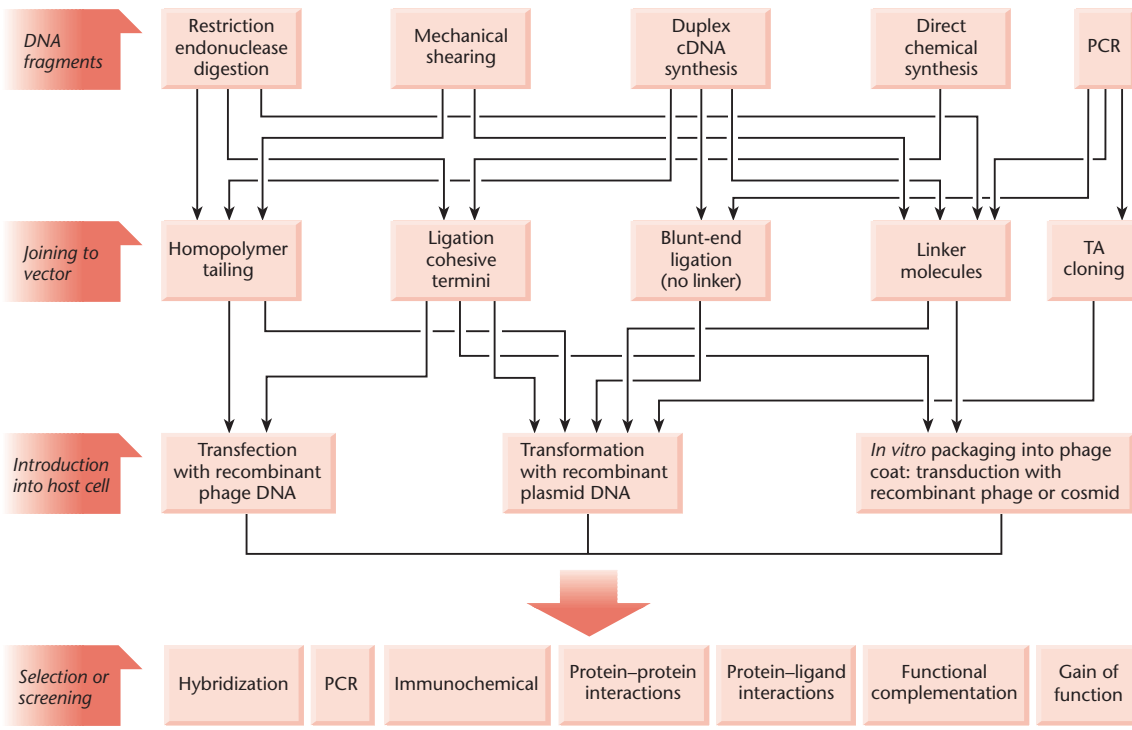


Fig. 6.1 Generalized overview of cloning strategies, with favoured routes shown by arrows. Note that in cell-based cloning strategies, DNA fragments are initially generated and cloned in a non-specific manner, so that screening for the desired clone is carried out at the end of the process. Conversely, when specific DNA fragments are obtained by PCR or direct chemical synthesis, there is no need for subsequent screening.

genome, in which case the target sequence could be diluted over a millionfold by unwanted genomic DNA. We need to find some way of rapidly sifting through, or *screening*, large numbers of unwanted sequences to identify our particular target.

There are two major strategies for isolating sequences from complex sources such as genomic DNA. The first, a cell-based cloning strategy, is to divide the source DNA into manageable fragments and *clone everything*. Such a collection of clones, representative of the entire starting population, is known as a *library*. We must then *screen the library* to identify our clone of interest, using a procedure that discriminates between the desired clone and all the others. A number of such procedures are discussed later in the chapter. The second strategy is to selectively amplify the target sequence directly from the source DNA using the polymerase chain reaction (PCR), and then clone this individual fragment. Each strategy has its advantages and disadvantages. Note that, in the library approach, screening is carried out after the entire source DNA population has been cloned indiscriminately. Conversely, in the PCR approach, the screening step is built into the first stage of the procedure, when the fragments are generated, so that only selected fragments are actually cloned. In this chapter we consider principles for the construction and screening of genomic and complementary DNA (cDNA) libraries, and compare the library-based route of gene isolation to equivalent PCR-based techniques.

Cloning genomic DNA

Genomic DNA libraries

Producing representative genomic libraries in λ cloning vectors

Following the example above, let us suppose that we wish to clone a single-copy gene from the human genome. How might this be achieved? We could simply digest total human DNA with a restriction endonuclease, such as *Eco*RI, insert the fragments into a suitable phage-λ vector and then attempt to isolate the desired clone. How many recombinants would we have to screen in order to isolate the right one? Assuming *Eco*RI gives, on average, fragments of about 4 kb, and given that the size of the human haploid genome is 2.8×10^6 kb, we can see that over 7×10^5 independent recombinants must be prepared and screened in order to have a reasonable chance of including the desired sequence. In other words we have to obtain a very large number of recombinants, which together contain a complete collection of all of the DNA sequences in the entire human genome, a human *genomic library*. The sizes of some other genomes are listed in Table 6.1.

There are two problems with the above approach. First, the gene may be cut internally one or more times by *Eco*RI so that it is not obtained as a single fragment. This is likely if the gene is large. Also, it may be desirable to obtain extensive regions flanking

Table 6.1 Genome sizes of selected organisms.

Organism	Genome size (kb) (haploid where appropriate)
Escherichia coli	4.0×10^3
Yeast (*Saccharomyces cerevisiae*)	1.35×10^4
Arabidopsis thaliana (higher plant)	1.25×10^5
Tobacco	1.6×10^6
Wheat	5.9×10^6
Zea mays	1.5×10^7
Drosophila melanogaster	1.8×10^5
Mouse	2.3×10^6
Human	2.8×10^6
Xenopus laevis	3.0×10^6

the gene or whole gene clusters. Fragments averaging about 4 kb are likely to be inconveniently short. Alternatively, the gene may be contained on an *Eco*RI fragment that is larger than the vector can accept. In this case the appropriate gene would not be cloned at all.

These problems can be overcome by cloning *random* DNA fragments of a large size (for λ replacement vectors, approximately 20 kb). Since the DNA is randomly fragmented, there will be no systematic exclusion of any sequence. Furthermore, clones will overlap one another, allowing the sequence of very large genes to be assembled and giving an opportunity to 'walk' from one clone to an adjacent one (p. 107). Because of the larger size of each cloned DNA fragment, fewer clones are required for a complete or nearly complete library. How many clones are required? Let n be the size of the genome relative to a single cloned fragment. Thus, for the human genome (2.8×10^6 kb) and an average cloned fragment size of 20 kb, $n = 1.4 \times 10^5$. The number of independent recombinants required in the library must be greater than n, because sampling variation will lead to the inclusion of some sequences several times and the exclusion of other sequences in a library of just n recombinants. Clarke and Carbon (1976) have derived a formula that relates the probability (P) of including any DNA sequence in a random library of N independent recombinants:

$$N = \frac{\ln(1-P)}{\ln\left(1-\frac{1}{n}\right)}$$

Therefore, to achieve a 95% probability ($P = 0.95$) of including any particular sequence in a random human genomic DNA library of 20 kb fragment size:

$$N = \frac{\ln(1-0.95)}{\ln\left(1-\frac{1}{1.4\times10^5}\right)} = 4.2\times10^5$$

Notice that a considerably higher number of recombinants is required to achieve a 99% probability, for here $N = 6.5 \times 10^5$.

How can appropriately sized random fragments be produced? Various methods are available. Random breakage by mechanical shearing is appropriate because the average fragment size can be controlled,

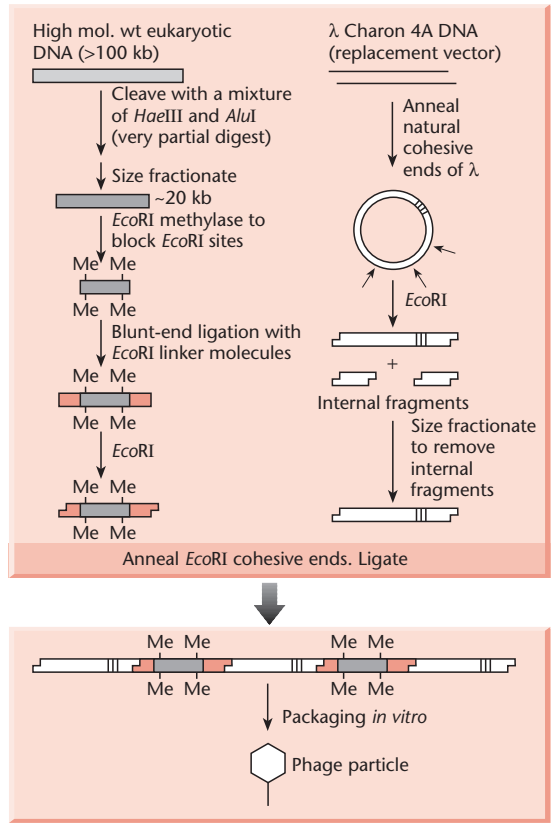


Fig. 6.2 Maniatis' strategy for producing a representative gene library.

but insertion of the resulting fragments into vectors requires additional steps. The more commonly used procedure involves restriction endonucleases. In the strategy devised by Maniatis *et al.* (1978) (Fig. 6.2), the target DNA is digested with a mixture of *two* restriction enzymes. These enzymes have tetranucleotide recognition sites, which therefore occur frequently in the target DNA and in a complete double-digest would produce fragments averaging less than 1 kb. However, only a partial restriction digest is carried out, and therefore the majority of the fragments are large (in the range 10–30 kb). Given that the chances of cutting at each of the available restriction sites are more or less equivalent, such a reaction effectively produces a random set of overlapping fragments. These can be size-fractionated, e.g. by gel electrophoresis, so as to give a random population of fragments of about

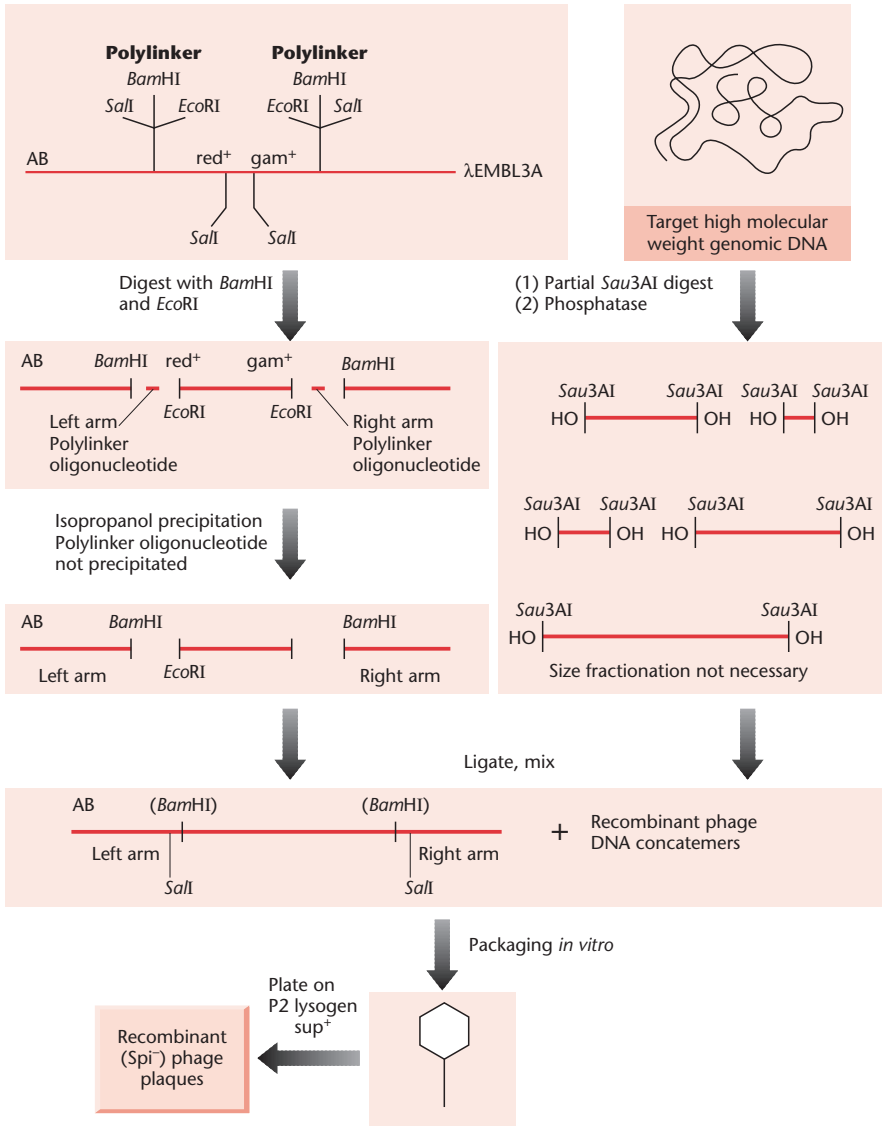


Fig. 6.3 Creation of a genomic DNA library using the phage-λ vector EMBL3A. High-molecular-weight genomic DNA is partially digested with *Sau*3AI. The fragments are treated with phosphatase to remove their 5′ phosphate groups. The vector is digested with *Bam*HI and *Eco*RI, which cut within the polylinker sites. The tiny *Bam*HI/*Eco*RI polylinker fragments are discarded in the isopropanol precipitation, or alternatively the vector arms may be purified by preparative agarose gel electrophoresis. The vector arms are then ligated with the partially digested genomic DNA. The phosphatase treatment prevents the genomic DNA fragments from ligating together. Non-recombinant vector cannot re-form because the small polylinker fragments have been discarded. The only packageable molecules are recombinant phages. These are obtained as plaques on a P2 lysogen of *sup*$^+$ *E. coli*. The Spi$^-$ selection ensures recovery of phage lacking *red* and *gam* genes. A *sup*$^+$ host is necessary because, in this example, the vector carries amber mutations in genes *A* and *B*. These mutations increase biological containment, and can be applied to selection procedures, such as recombinational selection, or tagging DNA with a *sup*$^+$ gene. Ultimately, the foreign DNA can be excised from the vector by virtue of the *Sal*I sites in the polylinker. (*Note*: Rogers *et al.* (1988) have shown that the EMBL3 polylinker sequence is not exactly as originally described. It contains an extra sequence with a previously unreported *Pst*I site. This does not affect most applications as a vector.)

20 kb, which are suitable for insertion into a λ replacement vector. Packaging *in vitro* (p. 58) ensures that an appropriately large number of independent recombinants can be recovered, which will give an almost completely representative library.

Development of λ replacement vectors for genomic library construction

In the Maniatis strategy, the use of two different restriction endonucleases with completely unrelated recognition sites, *Hae*III and *Alu*I, assists in obtaining fragmentation that is nearly random. These enzymes both produce blunt ends, and the cloning strategy requires linkers (see Fig. 6.2). Therefore, in the early days of vector development, a large number of different vectors became available with alternative restriction sites and genetic markers suitable for varied cloning strategies. A good example of this diversity is the Charon series, which included both insertion- and replacement-type vectors (Blatttner *et al.* 1977, Williams & Blattner 1979).

A convenient simplification can be achieved by using a *single* restriction endonuclease that cuts frequently, such as *Sau*3AI. This will create a partial digest that is slightly less random than that achieved with a pair of enzymes. However, it has the great advantage that the *Sau*3AI fragments can be readily inserted into λ replacement vectors, such as λEMBL3 (Frischauf *et al.* 1983), which have been digested with *Bam*HI (Fig. 6.3). This is because *Sau*3AI and *Bam*HI create the same cohesive ends (see p. 32). Due to the convenience and efficiency of this strategy, the λEMBL series of vectors have been very widely used for genomic library construction (p. 58). Note that λEMBL vectors also carry the *red* and *gam* genes on the stuffer fragment and a *chi* site on one of the vector arms, allowing convenient positive selection on the basis of the Spi phenotype (see p. 58). Most λ vectors currently used for genomic library construction are positively selected on this basis, including λ2001 (Karn *et al.* 1984), λDASH and λFIX (Sorge 1988). λDASH and λFIX (and recently improved versions, λDASHII and λFIXII) are particularly versatile because the multiple cloning sites flanking the stuffer fragment contain opposed promoters for the T3 and T7 RNA polymerases. If the recombinant vector is digested with a restriction endonuclease that cuts frequently, only short fragments of insert DNA are left attached to these promoters. This allows RNA probes to be generated corresponding to the *ends* of any genomic insert. These are ideal for probing the library to identify overlapping clones as part of a chromosome walk (p. 107) and have the great advantage that they can be made conveniently, directly from the vector, without recourse to subcloning. Vector maps of λDASH and λFIX are shown in Fig. 6.4. λFIX is similar to λDASH, except that it incorporates additional *Xho*I sites flanking the stuffer fragment. Digestion of the vector with *Xho*I followed by partial filling of the sticky ends prevents vector religation. However, partially filled *Sau*3AI sticky ends are compatible with the partially filled *Xho*I ends, although not with each other. This strategy prevents the ligation of vector arms without genomic DNA, and also prevents the insertion of multiple fragments.

Genomic libraries in high-capacity vectors

In place of phage-λ derivatives, a number of higher-capacity cloning vectors such as cosmids, bacterial artificial chromosomes (BACs), P1-derived artificial chromosomes (PACs) and yeast artificial chromosomes (YACs) are available for the construction of genomic libraries. The advantage of such vectors is that the average insert size is much larger than for λ replacement vectors. Thus, the number of recombinants that need to be screened to identify a particular gene of interest is correspondingly lower, large genes are more likely to be contained within a single clone and fewer steps are needed for a chromosome walk (p. 107). Generally, strategies similar to the Maniatis method discussed above are used to construct such libraries, except that the partial restriction digest conditions are optimized for larger fragment sizes, and size fractionation must be preformed by specialized electrophoresis methods that can separate fragments over 30 kb in length. Pulsed-field gel electrophoresis (PFGE) and field-inversion gel electrophoresis (FIGE) have been devised for this purpose (p. 10). High-molecular-weight donor DNA fragments can also be prepared using restriction enzymes that cut very rarely.

Cosmids may be favoured over λ vectors because they accept inserts of up to 45 kb. However, since

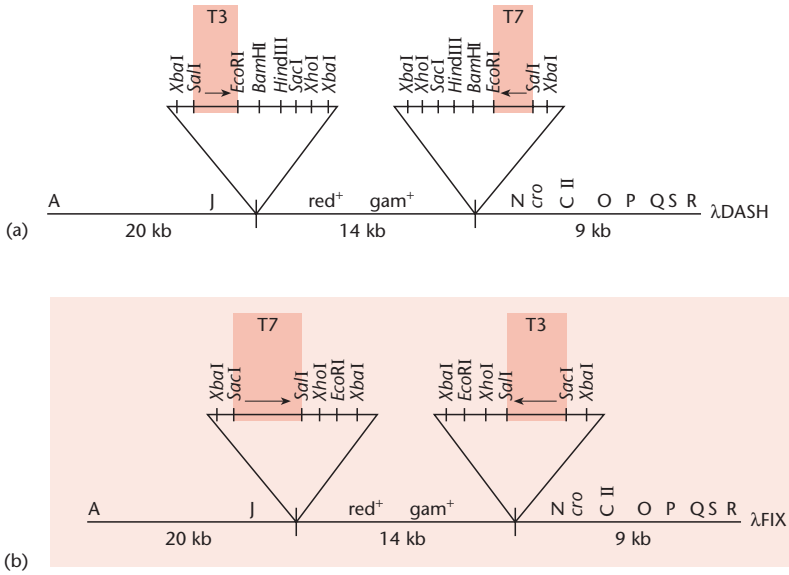


Fig. 6.4 The replacement vectors λDASH and λFIX. Promoters specific for the bacteriophage T3 and T7 RNA polymerases are located adjacent to the cloning sites, allowing RNA probes to be generated that correspond to each end of the insert.

these vectors are maintained as high-copy-number plasmids in *Escherichia coli*, they have a tendency to be unstable, undergoing deletions that favour increased replication. Most workers also find that plaques give less background hybridization than colonies, so screening libraries of phage-λ recombinants by plaque hybridization gives cleaner results than screening cosmid libraries by colony hybridization (see p. 104). It may also be desired to retain and store an *amplified genomic library*. With phage, the initial recombinant DNA population is packaged and plated out, and can be screened at this stage. Alternatively, the plates containing the recombinant plaques can be washed to give an amplified library of recombinant phage. The amplified library can then be stored almost indefinitely, because phage have a long shelf-life. The amplification is so great that samples of the amplified library can be plated out and screened with different probes on hundreds of occasions. With bacterial colonies containing cosmids, it is also possible to store an amplified library, but bacterial populations cannot be stored as readily as phage populations, and there is often an unacceptable loss of viability (Hanahan & Meselson 1980). A word of caution is necessary, however, when considering the use of any amplified library. This is due to the possibility of *distortion*. Not all recombinants in a population will propagate equally well, e.g. variations in target DNA size or sequence may affect replication of a recombinant phage, plasmid or cosmid. Therefore, when a library is put through an amplification step, particular recombinants may be increased in frequency, decreased in frequency or lost altogether. The development of modern vectors and cloning strategies has simplified library construction to the point where many workers now prefer to create a new library for each screening, rather than risk using a previously amplified one. Furthermore, pre-made libraries are available from many commercial sources and the same companies often offer custom library services. These libraries are often of high quality and such services are becoming increasingly popular.

The highest-capacity vectors – BACs, PACs and YACs – would seem to be ideal for library construction because of the very large insert sizes. However, such libraries are generally more difficult to prepare, and the larger inserts can be less than straightforward to work with. Unless the genomic target sequence is known to be very large and needs to be isolated as a single clone, λ replacement vectors or cosmids remain the most appropriate choice for many experiments. The main application of BAC, PAC and YAC libraries is for genome mapping, sequencing and the assembly of clone contigs.

Subgenomic libraries

Genomic libraries have been prepared from single human chromosomes separated by flow cytometry (e.g. Davies *et al*. 1981). Even greater enrichment for particular regions of the genome is possible using the technically demanding technique of *chromosome microdissection*. In *Drosophila*, it has been possible to physically excise a region of the salivary gland chromosome by micromanipulation, and then digest the DNA and clone it in phage-λ vectors, all within a microdrop under oil (Scalenghe *et al.* 1981). Similarly, this has been achieved with specific bands of human chromosomes using either extremely fine needles or a finely focused laser beam (Ludecke *et al*. 1989, 1990). Regardless of the species, this is a laborious and difficult technique and is prone to contamination with inappropriate DNA fragments. It has been rendered obsolete with the advent of high-capacity vectors, such as YACs.

PCR as an alternative to genomic DNA cloning

The PCR is a robust technique for amplifying specific DNA sequences from complex sources. In principle, therefore, PCR with specific primers could be used to isolate genes directly from genomic DNA, obviating the need for the production of genomic libraries. However, a serious limitation is that standard PCR conditions are suitable only for the amplification of short products. The maximum product size that can be obtained is about 5 kb, although the typical size is more likely to be 1–2 kb. This reflects the poor processivity of PCR enzymes such as *Taq* polymerase, and their lack of proofreading activity. Both of these deficiencies increase the likelihood of the enzyme detaching from the template, especially if the template is long. The extreme reaction conditions required for the PCR are also thought to cause damage to bases and generate nicks in DNA strands, which increases the probability of premature termination on long templates.

Long PCR

Modifications to reaction conditions can improve polymerase processivity by lowering the reaction temperature and increasing the pH, thereby protecting the template from damage (Foord & Rose 1994). The use of such conditions in combination with two DNA polymerases, one of which is a proofreading enzyme, has been shown to dramatically improve the performance of PCR using long templates (Barnes 1994, Cheng *et al*. 1994a). Essentially, the improvements come about because the proofreading enzyme removes mismatched bases that are often incorporated into growing DNA strands by enzymes such as *Taq* polymerase. Under normal conditions, *Taq* polymerase would stall at these obstructions and, lacking the intrinsic proofreading activity to correct them, the reaction would most likely be aborted.

Using such polymerase mixtures, it has been possible to amplify DNA fragments of up to 22 kb directly from human genomic DNA, almost the entire 16.6 kb human mitochondrial genome and the complete or near-complete genomes of several viruses, including 42 kb of the 45 kb phage-λ genome (Cheng *et al*. 1994a,b). Several commercial companies now provide cocktails of enzymes suitable for long PCR, e.g. TaqPlus Long PCR system, marketed by Stratagene, which is essentially a mixture of *Taq* polymerase and the thermostable proofreading enzyme *Pfu* polymerase. The technique has been applied to the structural analysis of human genes (e.g. Ruzzo *et al*. 1998, Bochmann *et al*. 1999) and viral genomes, including human immunodeficiency virus (HIV) (Dittmar *et al*. 1997). Long PCR has particular diagnostic value for the analysis of human triplet-repeat disorders, such as Friedreich's ataxia (Lamont *et al*. 1997). However, while long PCR is useful for the isolation of genes where sequence information is already available, it is unlikely to replace the use of genomic libraries, since the latter represent a permanent, full-genome resource that can be shared by numerous laboratories. Indeed, genomic libraries may be used in preference to total genomic DNA as the starting-point for gene isolation by long PCR.

Fragment libraries

Traditional genomic libraries cannot be prepared from small amounts of starting material, e.g. single cells, or from problematical sources, such as fixed tissue. In these cases, PCR is the only available strategy

for gene isolation. However, as well as being useful for the isolation of specific fragments, the PCR can be used to generate libraries, i.e. by amplifying a representative collection of random genomic fragments. This can be achieved using either random primers followed by size selection for suitable PCR products or a strategy in which genomic DNA is digested with restriction enzymes and then linkers are ligated to the ends of the DNA fragments, providing annealing sites for one specific type of primer (e.g. Zhang *et al.* 1992, Cheung & Nelson 1996). These techniques are powerful because they allow genomic fragment libraries to be prepared from material that could not yield DNA of suitable quality or quantity for conventional library construction, but competition among the templates generally does not allow the production of a truly representative library.

cDNA cloning

Properties of cDNA

cDNA is prepared by reverse-transcribing cellular RNA. Cloned eukaryotic cDNAs have their own special uses, which derive from the fact that they lack introns and other non-coding sequences present in the corresponding genomic DNA. Introns are rare in bacteria but occur in most genes of higher eukaryotes. They can be situated within the coding sequence itself, where they then interrupt the collinear relationship of the gene and its encoded polypeptide, or they may occur in the 5′ or 3′ untranslated regions. In any event, they are copied by RNA polymerase when the gene is transcribed. The primary transcript goes through a series of processing events in the nucleus before appearing in the cytoplasm as mature mRNA. These events include the removal of intron sequences by a process called *splicing.* In mammals, some genes contain numerous large introns that represent the vast majority of the sequence. For example, the human dystrophin gene contains 79 introns, representing over 99% of the sequence. The gene is nearly 2.5 Mb in length and yet the corresponding cDNA is only just over 11 kb. Thus, one advantage of cDNA cloning is that in many cases the size of the cDNA clone is significantly lower than that of the corresponding genomic clone. Since removal of eukaryotic intron transcripts by splicing does not occur in bacteria, eukaryotic cDNA clones find application where bacterial expression of the foreign DNA is necessary, either as a prerequisite for detecting the clone (see p. 109) or because expression of the polypeptide product is the primary objective. Also, where the sequence of the genomic DNA is available, the position of intron/exon boundaries can be assigned by comparison with the cDNA sequence.

cDNA libraries

Under some circumstances, it may be possible to prepare cDNA directly from a purified mRNA species. Much more commonly, a *cDNA library* is prepared by reverse-transcribing a population of mRNAs and then screened for particular clones. An important concept is that the cDNA library is representative of the RNA population from which it was derived. Thus, whereas genomic libraries are essentially the same, regardless of the cell type or developmental stage from which the DNA was isolated, the contents of cDNA libraries will vary widely according to these parameters. A given cDNA library will also be enriched for abundant mRNAs but may contain only a few clones representing rare mRNAs. Furthermore, where a gene is differentially spliced, a cDNA library will contain different clones representing alternative splice variants.

Table 6.2 shows the abundances of different classes of mRNAs in two representative tissues. Generally, mRNAs can be described as abundant, moderately abundant or rare. Notice that, in the chicken oviduct, one mRNA type is superabundant. This encodes ovalbumin, the major egg-white protein. Therefore, the starting population is naturally so enriched in ovalbumin mRNA that isolating the ovalbumin cDNA can be achieved without the use of a library. An appropriate strategy for obtaining such abundant cDNAs is to clone them directly in an M13 vector, such as M13 mp8. A set of clones can then be sequenced immediately and identified on the basis of the polypeptide that each encodes. A successful demonstration of this strategy was reported by Putney *et al.* (1983), who determined DNA sequences of 178 randomly chosen muscle cDNA clones. Based on the amino acid sequences available for 19 abundant muscle-specific proteins, they were able to

Table 6.2 Abundance classes of typical mRNA populations.

Source	Number of different mRNAs	Abundance (molecules/cell)
Mouse liver cytoplasmic poly(A)$^+$	9	12 000
	700	300
	11 500	15
Chick oviduct polysomal poly(A)$^+$	1	100 000
	7	4 000
	12 500	5

References: mouse (Young *et al*. 1976); chick oviduct (Axel *et al*. 1976).

identify clones corresponding to 13 of these 19 proteins, including several protein variants.

For the isolation of cDNA clones in the moderate- and low-abundance classes, it is usually necessary to construct a cDNA library. Once again, the high efficiency obtained by packaging *in vitro* makes phage-λ vectors attractive for obtaining large numbers of cDNA clones. Phage-λ insertion vectors are particularly well suited for cDNA cloning and some of the most widely used vectors are discussed in Box 6.1.

Typically, 10^5–10^6 clones are sufficient for the isolation of low-abundance mRNAs from most cell types, i.e. those present at 15 molecules per cell or above. However, some mRNAs are even less abundant than this, and may be further diluted if they are expressed in only a few specific cells in a particular tissue. Under these circumstances, it may be worth enriching the mRNA preparation prior to library construction, e.g. by size fractionation, and testing the fractions for the presence of the desired molecule. One way in which

Box 6.1 Phage-λ vectors for cDNA cloning and expression

λgt10 and λgt11

Most early cDNA libraries were constructed using plasmid vectors, and were difficult to store and maintain for long periods. They were largely replaced by phage-λ libraries, which can be stored indefinitely and can also be prepared to much higher titres. λgt10 and λgt11 were the standard vectors for cDNA cloning until about 1990. Both λgt10 and λgt11 are insertion vectors, and they can accept approximately 7.6 kb and 7.2 kb of foreign DNA, respectively. In each case, the foreign DNA is introduced at a unique *Eco*RI cloning site. λgt10 is used to make libraries that are screened by hybridization. The *Eco*RI site interrupts the phage *cI* gene, allowing selection on the basis of plaque morphology. λgt11 contains an *E. coli lacZ* gene driven by the *lac* promoter. If inserted in the correct orientation and reading frame, cDNA sequences cloned in this vector can be expressed as β-galactosidase fusion proteins, and can be detected by immunological screening or screening with other ligands (see p. 109). λgt11 libraries can also be screened by hybridization, although λgt10 is more appropriate for this screening strategy because higher titres are possible.

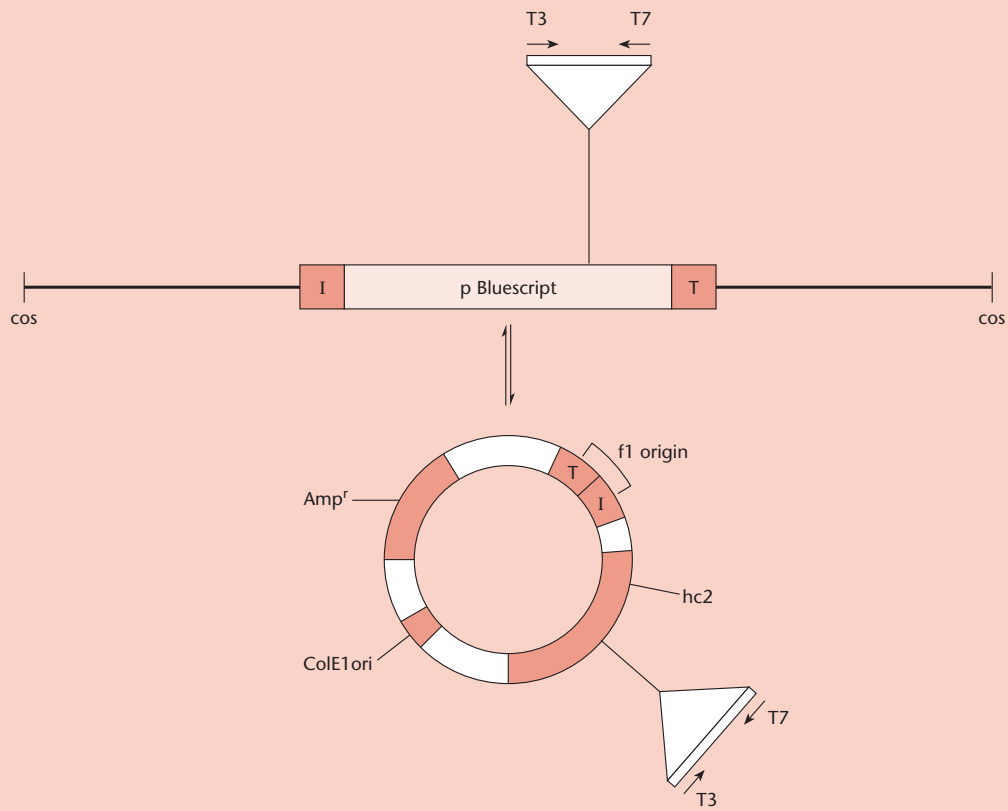
λZAP series

While phage-λ vectors generate better libraries, they cannot be manipulated *in vitro* with the convenience of plasmid vectors. Therefore, phage clones have to

continued

Box 6.1 *continued*

be laboriously subcloned back into plasmids for further analysis. This limitation of conventional phage-λ vectors has been addressed by the development of hybrids, sometimes called *phasmids*, which possess the most attractive features of both bacteriophage λ and plasmids (see Chapter 5). The most popular current vectors for cDNA cloning are undoubtedly those of the λZAP series marketed by Stratagene (Short *et al*. 1988). A map of the original λZAP vector is shown below. The advantageous features of this vector are: (i) the high capacity – up to 10 kb of foreign DNA can be cloned, which is large enough to encompass most cDNAs; (ii) the presence of a polylinker with six unique restriction sites, which increases cloning versatility and also allows directional cloning; and (iii) the availability of T3 and T7 RNA polymerase sites flanking the polylinker, allowing sense and antisense RNA to be prepared from the insert. Most importantly, all these features are included within a plasmid vector called pBluescript, which is itself inserted into the phage genome. Thus the cDNA clone can be recovered from the phage and propagated as a high-copy-number plasmid without any subcloning, simply by coinfecting the bacteria with a helper f1 phage that nicks the λZAP vector at the flanks of the plasmid and facilitates excision. Another member of this series, λZAP Express, also includes the human cytomegalovirus promoter and SV40 terminator, so that fusion proteins can be expressed in mammalian cells as well as bacteria. Thus, cDNA libraries can be cloned in the phage vector in *E. coli*, rescued as plasmids and then transfected into mammalian cells for expression cloning.



this can be achieved is to inject mRNA fractions into *Xenopus* oocytes (p. 215) and test them for production of the corresponding protein (Melton 1987). See also the discussion of subtraction cloning on p. 115.

Preparation of cDNA for library construction

The cDNA synthesis reaction

The synthesis of double-stranded cDNA suitable for insertion into a cloning vector involves three major steps: (i) first-strand DNA synthesis on the mRNA template, carried out with a reverse transcriptase; (ii) removal of the RNA template; and (iii) second-strand DNA synthesis using the first DNA strand as a template, carried out with a DNA-dependent DNA polymerase, such as *E. coli* DNA polymerase I. All DNA polymerases, whether they use RNA or DNA as the template, require a primer to initiate strand synthesis.

Development of cDNA cloning strategies

The first reports of cDNA cloning were published in the mid-1970s and were all based on the homopolymer tailing technique, which is described briefly in Chapter 3. Of several alternative methods, the one that became the most popular was that of Maniatis *et al.* (1976). This involved the use of an oligo-dT primer annealing at the polyadenylate tail of the mRNA to prime first-strand cDNA synthesis, and took advantage of the fact that the first cDNA strand has the tendency to transiently fold back on itself, forming a hairpin loop, resulting in self-priming of the second strand (Efstratiadis *et al.* 1976). After the synthesis of the second DNA strand, this loop must be cleaved with a single-strand-specific nuclease, e.g. S1 nuclease, to allow insertion into the cloning vector (Fig. 6.5).

A serious disadvantage of the hairpin method is that cleavage with S1 nuclease results in the loss of a certain amount of sequence at the 5′ end of the clone. This strategy has therefore been superseded by other methods in which the second strand is primed in a separate reaction. One of the simplest strategies is shown in Fig. 6.6 (Land *et al.* 1981). After first-strand synthesis, which is primed with an oligo-dT primer as usual, the cDNA is tailed with a string of cytidine residues using the enzyme terminal transferase. This artificial oligo-dC tail is then used as an annealing site for a synthetic oligo-dG primer, allowing synthesis of the second strand. Using this method, Land *et al.* (1981) were able to isolate a full-length cDNA corresponding to the chicken lysozyme gene. However, the efficiency can be lower for other cDNAs (e.g. Cooke *et al.* 1980).

For cDNA expression libraries, it is advantageous if the cDNA can be inserted into the vector in the correct orientation. With the self-priming method, this can be achieved by adding a synthetic linker to the double-stranded cDNA molecule before the hairpin loop is cleaved (e.g. Kurtz & Nicodemus 1981; Fig. 6.7a). Where the second strand is primed separately, direction cloning can be achieved using an oligo-dT primer containing a linker sequence (e.g. Coleclough & Erlitz 1985; Fig. 6.7b). An alternative is to use primers for cDNA synthesis that are already linked to a plasmid (Fig. 6.7c). This strategy was devised by Okayama and Berg (1982) and has two further notable characteristics. First, full-length cDNAs are *preferentially obtained* because an RNA–DNA hybrid molecule, the result of first-strand synthesis, is the substrate for a terminal transferase reaction. A cDNA that does not extend to the end of the mRNA will present a shielded 3-hydroxyl group, which is a poor substrate for tailing. Secondly, the second-strand synthesis step is primed by nicking the RNA at multiple sites with RNase H. Second-strand synthesis therefore occurs by a nick-translation type of reaction, which is highly efficient. Simpler cDNA cloning strategies incorporating replacement synthesis of the second strand are widely used (e.g. Gubler & Hoffman 1983, Lapeyre & Amalric 1985). The Gubler–Hoffman reaction, as it is commonly known, is show in Fig. 6.8.

Full-length cDNA cloning

Limitations of conventional cloning strategies

Conventional approaches to the production of cDNA libraries have two major drawbacks. First, where oligo-dT primers are used to initiate first-strand synthesis, there is generally a 3′-end bias (preferential recovery of clones representing the 3′ end of cDNA

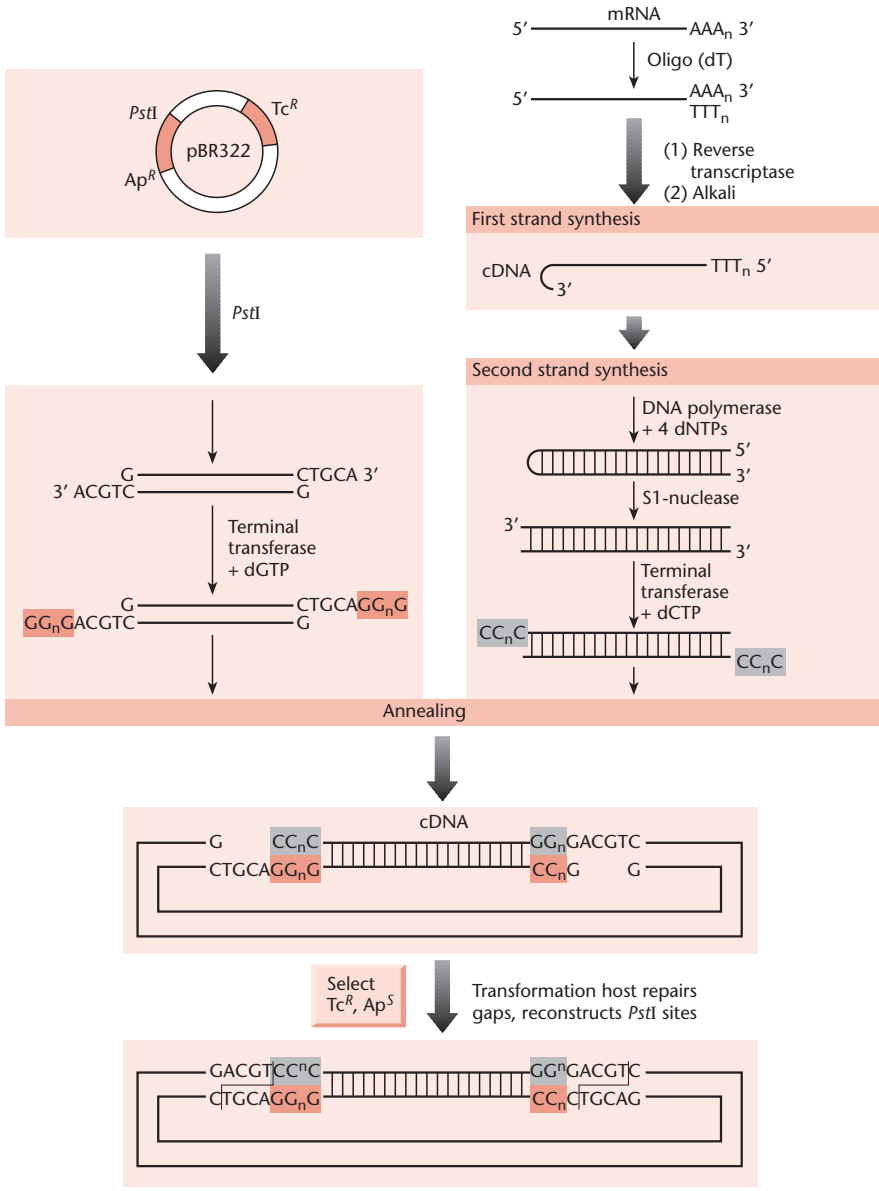


Fig. 6.5 An early cDNA cloning strategy, involving hairpin-primed second-strand DNA synthesis and homopolymer tailing to insert the cDNA into the vector.

sequences) in the resulting library. This can be addressed through the use of random oligonucleotide primers, usually hexamers, for both first- and second-strand cDNA synthesis. However, while this eliminates 3′-end bias in library construction, the resulting clones are much smaller, such that full-length cDNAs must be assembled from several shorter fragments. Secondly, as the size of a cDNA increases, it becomes progressively more difficult to isolate full-length clones. This is partly due to deficiencies in the reverse-transcriptase enzymes used for first-strand cDNA synthesis. The enzymes

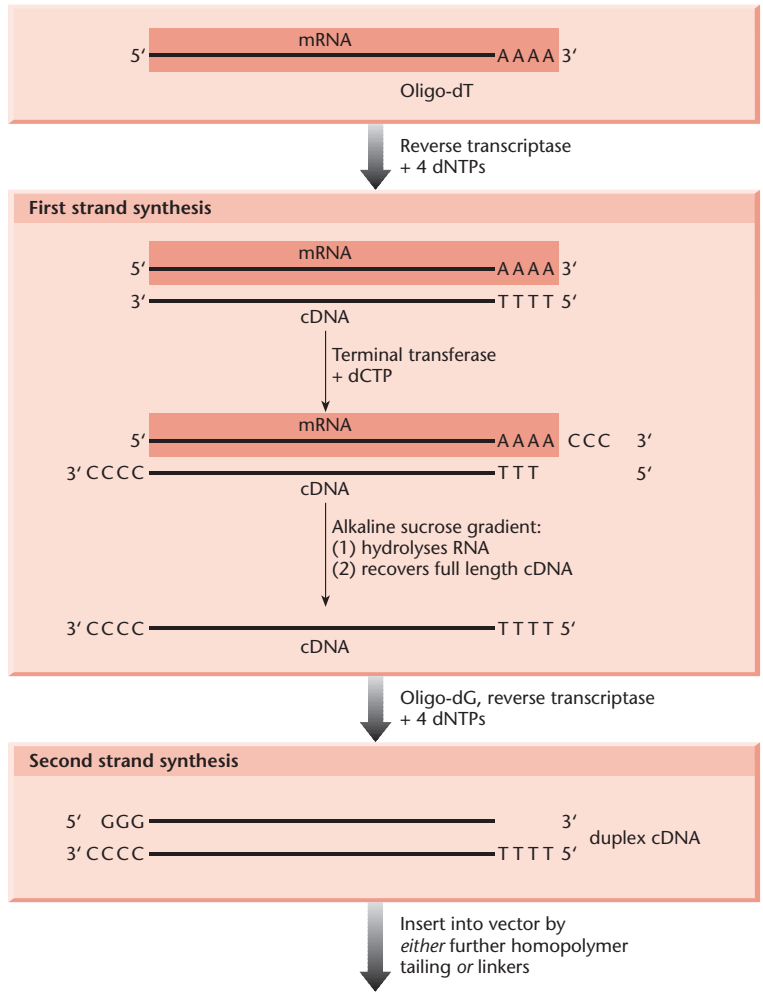
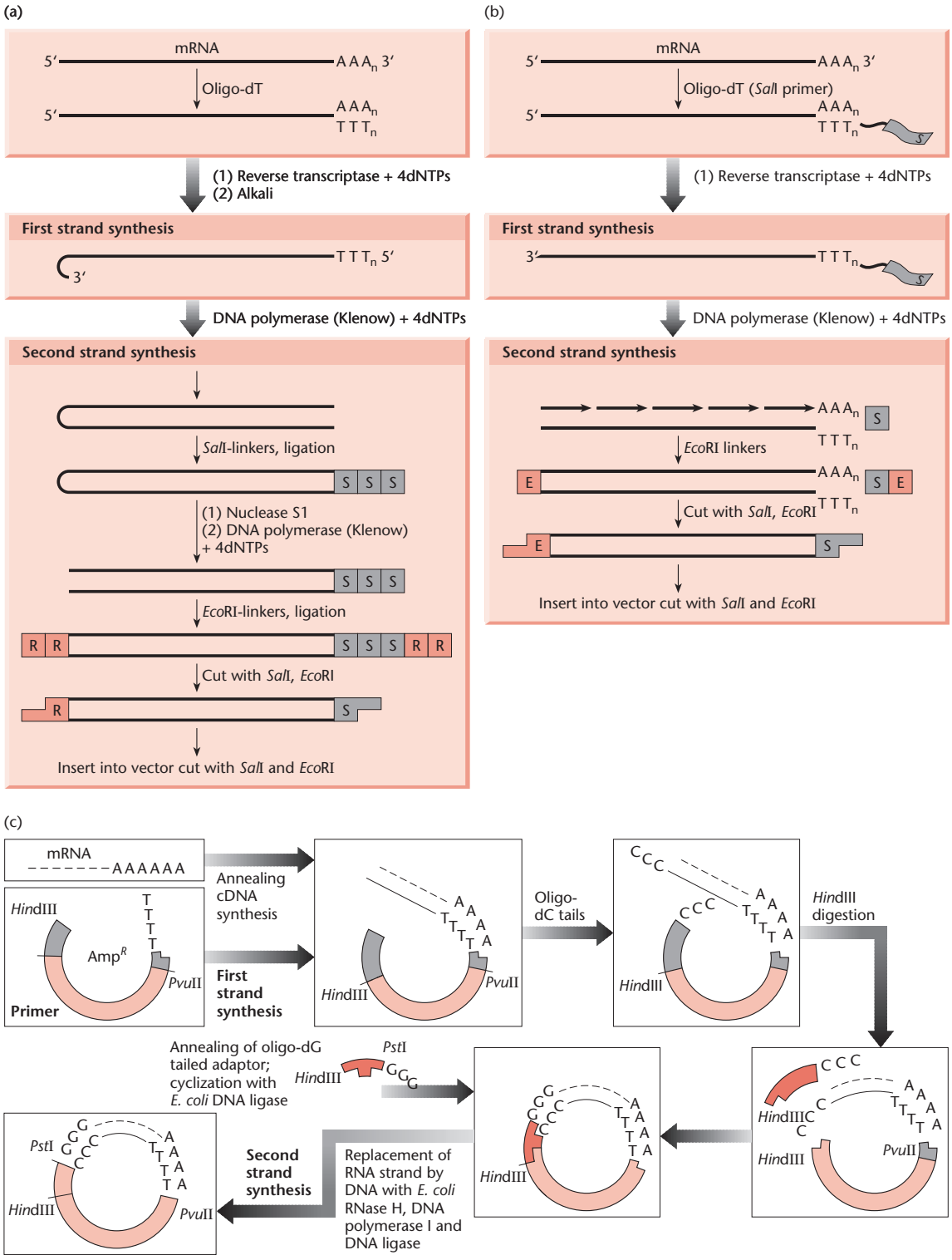


Fig. 6.6 Improved method for cDNA cloning. The first strand is tailed with oligo(dC) allowing the second strand to be initiated using an oligo(dG) primer.

are usually purified from avian myeloblastosis virus (AMV) or produced from a cloned Moloney murine leukaemia virus (MMLV) gene in *E. coli.* Native enzymes have poor processivity and intrinsic RNase activity, which leads to degradation of the RNA template (Champoux 1995). Several companies produce engineered murine reverse transcriptases that lack RNase H activity, and these are more efficient in the production of full-length cDNAs (Gerard & D'Allesio 1993). An example is the enzyme SuperScript II, marketed by Life Technologies (Kotewicz *et al.* 1988). This enzyme can also carry out reverse transcription at temperatures of up to 50°C. The native enzymes function optimally at 37°C and therefore tend to stall at sequences that are rich in secondary structure, as often found in 5′ and 3′ untranslated regions.

Selection of 5′ mRNA ends

Despite improvements in reverse transcriptases, the generation of full-length clones corresponding to large mRNAs remains a problem. This has been addressed by the development of cDNA cloning strategies involving the selection of mRNAs with intact 5′ ends. Nearly all eukaryotic mRNAs have a 5′ end cap, a specialized, methylated guanidine residue that is inverted with respect to the rest of the strand and is recognized by the ribosome prior to the initiation of protein synthesis. Using a combination



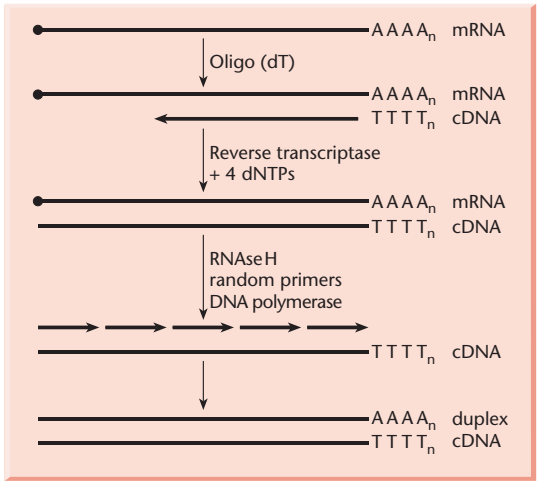


Fig. 6.8 The Gubler–Hoffman method, a simple and general method for non-directional cDNA cloning. First-strand synthesis is primed using an oligo(dT) primer. When the first strand is complete, the RNA is removed with RNase H and the second strand is random-primed and synthesized with DNA polymerase I. T4 DNA polymerase is used to ensure that the molecule is blunt-ended prior to insertion into the vector.

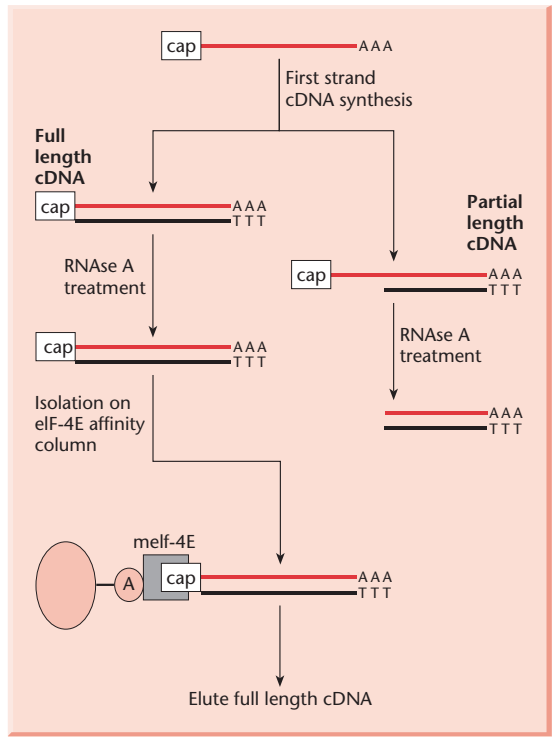


Fig. 6.9 The CAPture method of full-length cDNA cloning, using the eukaryotic initiation factor eIF-4E to select mRNAs with caps protected from RNase digestion by a complementary DNA strand.

of cap selection and nuclease treatment, it is possible to select for full-length first-strand cDNAs and thus generate libraries highly enriched in full-length clones.

An example is the method described by Edery *et al.* (1995) (Fig. 6.9). In this strategy, first-strand cDNA synthesis is initiated as usual, using an oligo-dT primer. Following the synthesis reaction, the hybrid molecules are treated with RNase A, which only digests single-stranded RNA. DNA–RNA hybrids therefore remain intact. If the first-strand cDNA is full-length, it reaches all the way to the 5′ cap of the mRNA, which is therefore protected from cleavage by RNase A. However, part-length cDNAs will leave a stretch of unprotected single-stranded RNA between the end of the double-stranded region and the cap, which is digested away with the enzyme. In the next stage of the procedure, the eukaryotic translational initiation factor eIF-4E is used to isolate full-length molecules by affinity capture. Incomplete cDNAs and cDNAs synthesized on broken templates will lack the cap and will not be retained. A similar method based on the biotinylation of mRNA has also been reported (Caminci *et al.* 1996). Both methods, however, also co-purify cDNAs resulting from the mispriming of first-strand synthesis, which can account for up to 10% of the clones in a library. An alternative method, *oligo-capping*, addresses this

Fig. 6.7 (*opposite*) Methods for directional cDNA cloning. (a) An early strategy in which the formation of a loop is exploited to place a specific linker (in this example, for *Sal*I) at the open end of the duplex cDNA. Following this ligation, the loop is cleaved and trimmed with S1 nuclease and *Eco*R1 linkers are added to both ends. Cleavage with *Eco*R1 and *Sal*I generates a restriction fragment that can be unidirectionally inserted into a vector cleaved with the same enzymes. (b) A similar strategy, but second-strand cDNA synthesis is random-primed. The oligo(dT) primer carries an extension forming a *Sal*I site. During second-strand synthesis, this forms a double-stranded *Sal*I linker. The addition of further *Eco*RI linkers to both ends allows the cDNA to be unidirectionally cloned, as above. (c) The strategy of Okayama and Berg (1982), where the mRNA is linked unidirectionally to the plasmid cloning vector prior to cDNA synthesis, by virtue of a cDNA tail.

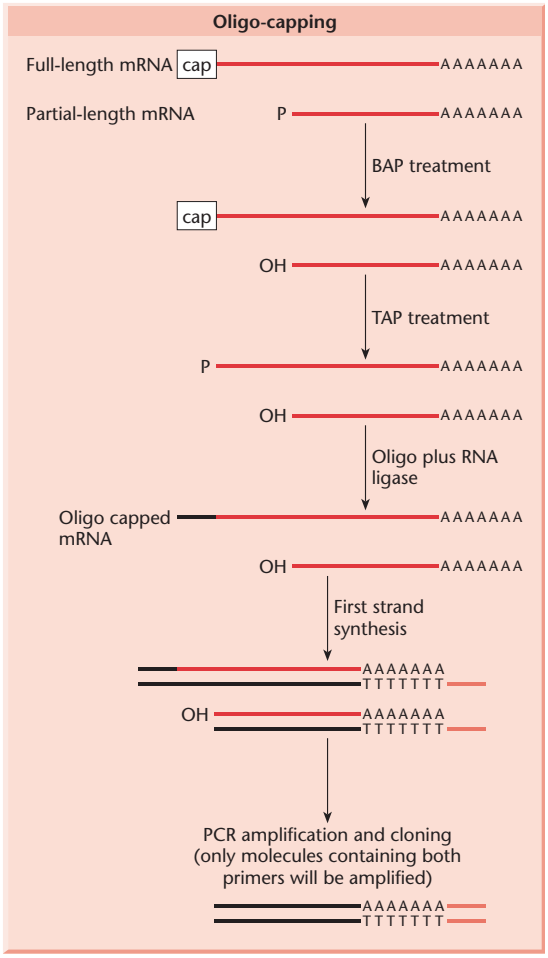


Fig. 6.10 Oligo-capping, the addition of specific oligonucleotide primers to full-length RNAs by sequential treatment with alkaline phosphatase and acid pyrophosphatase. Once the oligo cap has annealed to the 5′ end of the mRNA, it can serve as a primer binding site for PCR amplification.

problem by performing selection at the RNA stage (Maruyama & Sugano 1994, Suzuki *et al.* 1997, 2000; Fig. 6.10). The basis of the method is that RNA is sequentially treated with the enzymes alkaline phosphatase and acid pyrophosphatase. The first enzyme removes phosphate groups from the 5′ ends of uncapped RNA molecules, but does not affect full-length molecules with a 5′ cap. The second treatment removes the cap from full-length RNAs, leaving a 5′-terminal residue with a phosphate group. Full-length molecules can be ligated to a specific oligonucleotide, while broken and degraded molecules cannot. The result is an oligo-capped population of full-length mRNAs. This selected population is then reverse-transcribed using an oligo-dT primer. Second-strand synthesis and cloning is then carried out by PCR using the oligo-dT primer and a primer annealing to the oligonucleotide cap. Only full-length cDNAs annealing to both primers will be amplified, thus eliminating broken or degraded RNAs, incomplete first cDNA strands (which lack a 5′ primer annealing site) and misprimed cDNAs (which lack a 3′ primer annealing site).

PCR as an alternative to cDNA cloning

Reverse transcription followed by the PCR (RT-PCR) leads to the amplification of RNA sequences in cDNA form. No modification to the basic PCR strategy (p. 19) is required, except that the template for PCR amplification is generated in the same reaction tube in a prior reverse-transcription reaction (see Kawasaki 1990, Dieffenbach & Dvesler 1995). Using gene-specific primers, RT-PCR is a sensitive means for detecting, quantifying and cloning specific cDNA molecules. Reverse transcription is carried out using a specific 3′ primer that generates the first cDNA strand, and then PCR amplification is initiated following the addition of a 5′ primer to the reaction mix. The sensitivity is such that total RNA can be used as the starting material, rather than the poly(A)$^+$ RNA which is used for conventional cDNA cloning. Total RNA also contains ribosomal RNA (rRNA) and transfer RNA (tRNA), which can be present in a great excess to mRNA.

Due to the speed with which RT-PCR can be carried out, it is an attractive approach for obtaining a specific cDNA sequence for cloning. In contrast, screening a cDNA library is laborious, even presuming that a suitable cDNA library is already available and does not have to be constructed for the purpose. Quite apart from the labour involved, a cDNA library may not yield a cDNA clone with a full-length coding region, because, as described above, generating a full-length cDNA clone may be technically challenging, particularly with respect to long mRNAs. Furthermore, the sought-after cDNA may be very rare even in specialized libraries. Does this mean

that cDNA libraries have been superseded? Despite the advantages of RT-PCR, there are still reasons for constructing cDNA libraries. The first reflects the availability of starting material and the permanence of the library. A sought-after mRNA may occur in a source that is not readily available, perhaps a small number of cells in a particular human tissue. A good-quality cDNA library has only to be constructed once from this tissue to give a virtually infinite resource for future use. The specialized library is permanently available for screening. Indeed, the library may be used as a source from which a specific cDNA can be obtained by PCR amplification. The second reason concerns screening strategies. The PCR-based approaches are dependent upon specific primers. However, with cDNA libraries, screening strategies are possible that are based upon expression, e.g. immunochemical screening, rather than nucleic acid hybridization (see below).

As discussed above for genomic libraries, PCR can be used to provide the DNA for library construction when the source is unsuitable for conventional approaches, e.g. a very small amount of starting material or fixed tissue. Instead of gene-specific primers, universal primers can be used that lead to the amplification of all mRNAs, which can then be subcloned into suitable vectors. A disadvantage of PCR-based strategies for cDNA library construction is that the DNA polymerases used for PCR are more error-prone than those used conventionally for second-strand synthesis, so the library may contain a large number of mutations. There is also likely to be a certain amount of distortion due to competition among templates, and a bias towards shorter cDNAs.

A potential problem with RT-PCR is false results resulting from the amplification of contaminating genomic sequences in the RNA preparation. Even trace amounts of genomic DNA may be amplified. In the study of eukaryotic mRNAs, it is therefore desirable to choose primers that anneal in different exons, such that the products expected from the amplification of cDNA and genomic DNA would be different sizes or, if the intron is suitably large, so that genomic DNA would not be amplified at all. Where this is not possible (e.g. when bacterial RNA is used as the template), the RNA can be treated with DNase prior to amplification to destroy any contaminating DNA.

Rapid amplification of cDNA ends (RACE)

Another way to address the problem of incomplete cDNA sequences in libraries is to use a PCR-based technique for the *rapid amplification of cDNA ends* (RACE) (Frohman *et al.* 1988). Both 5′ RACE and 3′ RACE protocols are available, although 3′ RACE is usually only required if cDNAs have been generated using random primers. In each case, only limited knowledge of the mRNA sequence is required. A single stretch of sequence within the mRNA is sufficient, so an incomplete clone from a cDNA library is a good starting-point. From this sequence, specific primers are chosen which face *outwards* and which produce overlapping cDNA fragments. In the two RACE protocols, extension of the cDNAs from the ends of the transcript to the specific primers is accomplished by using primers that hybridize either at the natural 3′ poly(A) tail of the mRNA, or at a synthetic poly(dA) tract added to the 5′ end of the first-strand cDNA (Fig. 6.11). Finally, after amplification, the overlapping RACE products can be combined if desired, to produce an intact full-length cDNA.

Although simple in principle, RACE suffers from the same limitations that affect conventional cDNA cloning procedures. In 5′ RACE, for example, the reverse transcriptase may not, in many cases, reach the authentic 5′ end of the mRNA, but all first-strand cDNAs, whether full length or truncated, are tailed in the subsequent reaction, leading to the amplification of a population of variable-length products. Furthermore, as might be anticipated, since only a single *specific* primer is used in each of the RACE protocols, the specificity of amplification may not be very high. This is especially problematical where the specific primer is degenerate. In order to overcome this problem, a modification of the RACE method has been devised which is based on using nested primers to increase specificity (Frohman & Martin 1989). Strategies for improving the specificity of RACE have been reviewed (Schaefer 1995, Chen 1996).

Screening strategies

The identification of a specific clone from a DNA library can be carried out by exploiting either the sequence of the clone or the structure/function of its

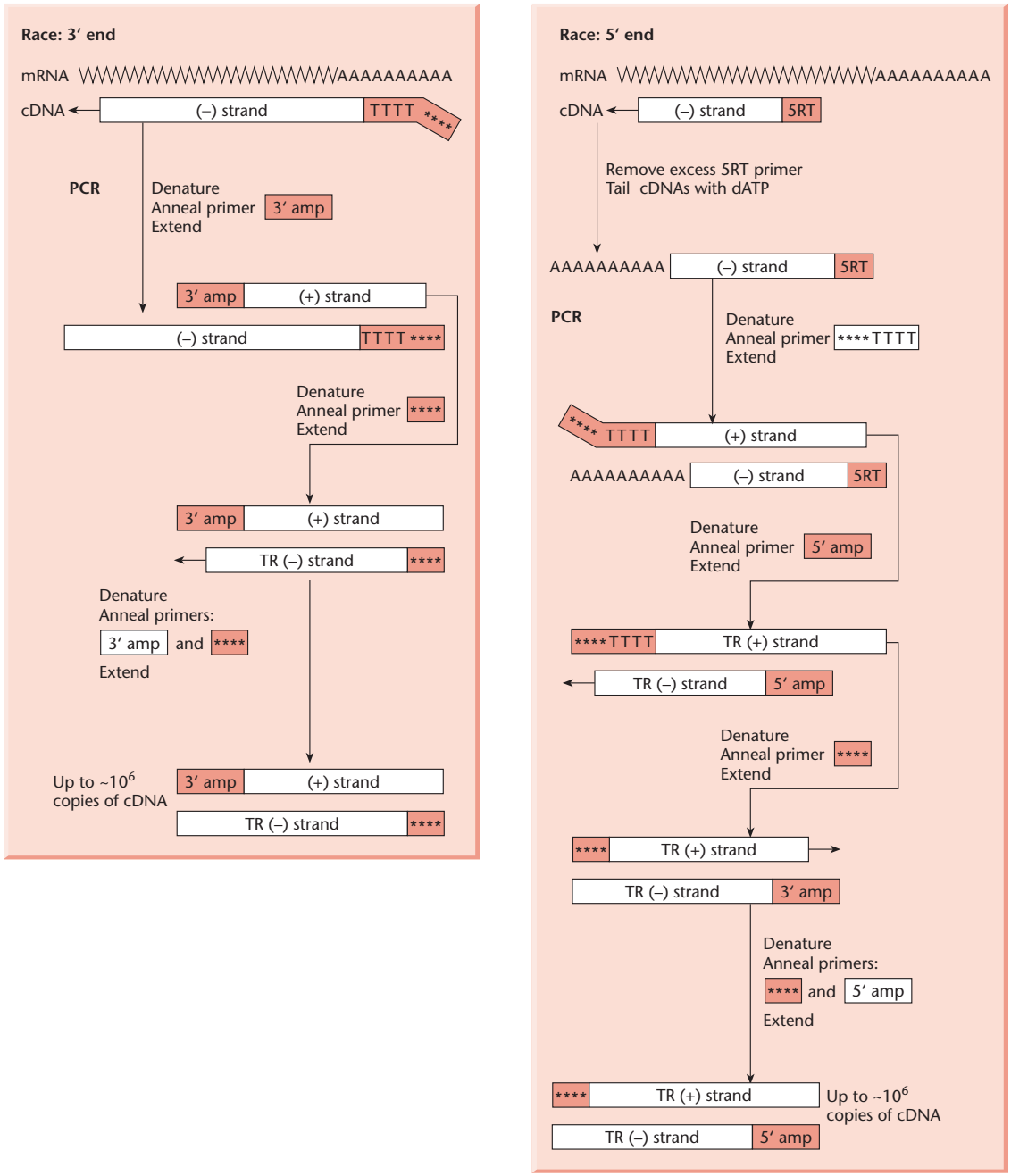


Fig. 6.11 Rapid amplification of cDNA ends (RACE) (Frohman *et al.* 1988). *3′ Protocol*. The mRNA is reverse transcribed using an oligo(dT_{17}) primer which has a 17 nucleotide extension at its 5′-end. This extension, the anchor sequence, is designed to contain restriction sites for subsequent cloning. Amplification is performed using the anchor 17-mer (which has a T_m higher than oligo(dT_{17})) and a primer specific for the sought-after cDNA. *5′ Protocol*. The mRNA is reverse transcribed from a specific primer. The resultant cDNA is then extended by terminal transferase to create a poly(dA) tail at the 3′-end of the cDNA. Amplification is performed with the oligo(dT_{17})/anchor system as used for the 3′ protocol, and the specific primer. Open boxes represent DNA strands being synthesized; coloured boxes represent DNA from a previous step. The diagram is simplified to show only how the *new* product from a previous step is used. Molecules designated TR, truncated, are shorter than full-length (+) or (–) strands.

Box 6.2 Expressed sequence tags (ESTs) for high-throughput genome research

The gold standard in the analysis of individual genes is a full-length cDNA clone that has been independently sequenced several times to ensure accuracy. Such clones are desirable for accurate archiving and for the detailed mapping of genomic transcription units, i.e. to determine the transcriptional start and stop sites, and all intron/exon boundaries. However, as discussed in the main text, such clones can be difficult and expensive to obtain. Technology has not yet advanced to the stage where full-length cDNAs can be produced and sequenced in a high-throughput manner.

Fortunately, full-length cDNA clones are not required for many types of analysis. Even short cDNA sequence fragments can be used to unambiguously identify specific genes and therefore map them on to physical gene maps or provide information about their expression patterns. The development of high-throughput sequencing technology has allowed thousands of clones to be picked randomly from cDNA libraries and subjected to single-pass sequencing to generate 200–300 bp cDNA signatures called *expressed sequence tags* (ESTs) (Wilcox *et al*. 1991, Okubo *et al*. 1992). Although short and somewhat inaccurate, very large numbers of sequences can be collected rapidly and inexpensively and deposited into public databases that can be searched using the Internet. The vast majority of database sequences are now ESTs rather than full cDNA or genomic clones. ESTs have been used for gene discovery, as physical markers on genomic maps and for the identification of genes in genomic clones (e.g. Adams *et al*. 1991, 1992, Banfi *et al*. 1996). Over two million ESTs from numerous species are currently searchable using the major public EST database, dbEST (Boguski *et al*. 1993). The development of EST informatics has been reviewed (Boguski 1995, Gerhold & Caskey 1996, Hartl 1996, Okubo & Matsubara 1997). The advent of ESTs prompted a wide public debate on the concept of patenting genes. The National Institutes of Health attempted to patent more than 1000 of the first EST sequences to be generated, but the patent application was rejected, predominantly on the grounds that ESTs are incomplete sequences and lack precise functional applications (for discussion, see Roberts 1992).

As well as their use for mapping, ESTs are useful for expression analysis. PCR primers designed around ESTs have been used to generate large numbers of target sequences for cDNA microarrays (see Box 6.5), and the partial cDNA fragments used for techniques such as differential-display PCR are also essentially ESTs (p. 116). The ultimate EST approach to expression analysis is *serial analysis of gene expression* (SAGE). In this technique, the size of the sequence tag is only 9–10 bp (the minimum that is sufficient to uniquely identify specific transcripts) and multiple tags are ligated into a large concatemer allowing expressed genes to be 'read' by cloning and sequencing the tags serially arranged in each clone (for details see Velculescu *et al*. 1995).

expressed product. The former applies to any type of library, genomic or cDNA, and can involve either nucleic acid hybridization or the PCR. In each case, the design of the probe or primers can be used to home in on one specific clone or a group of structurally related clones. Note that PCR screening can also be used to isolate DNA sequences from uncloned genomic DNA and cDNA. Screening the product of a clone applies only to expression libraries, i.e. libraries where the DNA fragment is expressed to yield a protein. In this case, the clone can be identified because its product is recognized by an antibody or a ligand of some nature, or because the biological activity of the protein is preserved and can be assayed in an appropriate test system.

Sequence-dependent screening

Screening by hybridization

Nucleic acid hybridization is the most commonly used method of library screening because it is rapid,

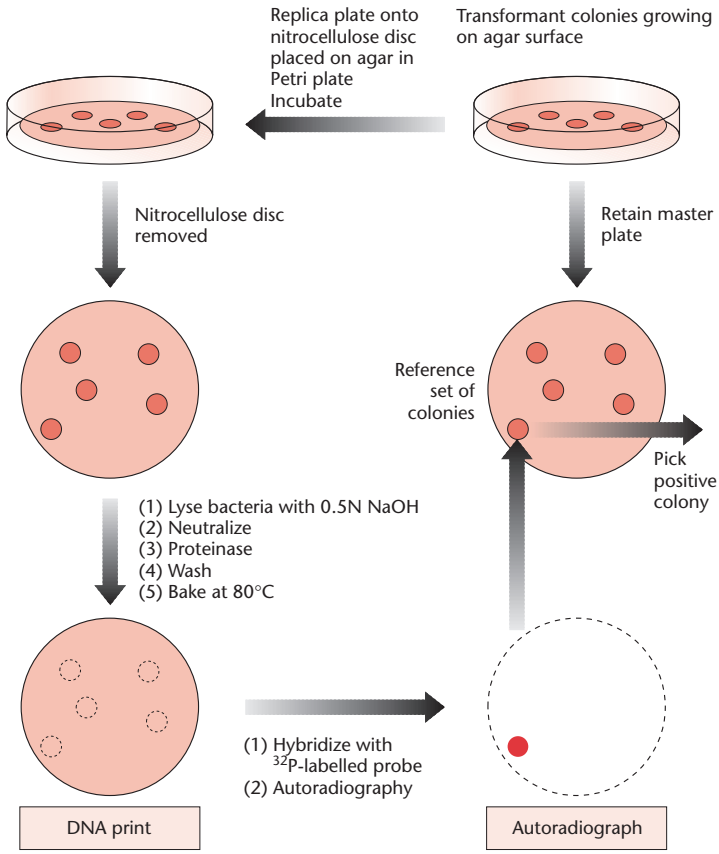


Fig. 6.12 Grunstein–Hogness method for detection of recombinant clones by colony hybridization.

it can be applied to very large numbers of clones and, in the case of cDNA libraries, can be used to identify clones that are not full-length (and therefore cannot be expressed).

Grunstein and Hogness (1975) developed a screening procedure to detect DNA sequences in transformed colonies by hybridization *in situ* with radioactive RNA probes. Their procedure can rapidly determine which colony among thousands contains the target sequence. A modification of the method allows screening of colonies plated at a very high density (Hanahan & Meselson 1980). The colonies to be screened are first replica-plated on to a nitrocellulose filter disc that has been placed on the surface of an agar plate prior to inoculation (Fig. 6.12). A reference set of these colonies on the master plate is retained. The filter bearing the colonies is removed and treated with alkali so that the bacterial colonies are lysed and the DNA they contain is denatured. The filter is then treated with proteinase K to remove protein and leave denatured DNA bound to the nitrocellulose, for which it has a high affinity, in the form of a 'DNA print' of the colonies. The DNA is fixed firmly by baking the filter at 80°C. The defining, labelled RNA is hybridized to this DNA and the result of this hybridization is monitored by autoradiography. A colony whose DNA print gives a positive autoradiographic result can then be picked from the reference plate.

Variations of this procedure can be applied to phage plaques (Jones & Murray 1975, Kramer *et al.* 1976). Benton and Davis (1977) devised a method called *plaque lift*, in which the nitrocellulose filter is applied to the upper surface of agar plates, making direct contact between plaques and filter. The plaques contain phage particles, as well as a considerable amount of unpackaged recombinant DNA. Both phage and unpackaged DNA bind to the filter and

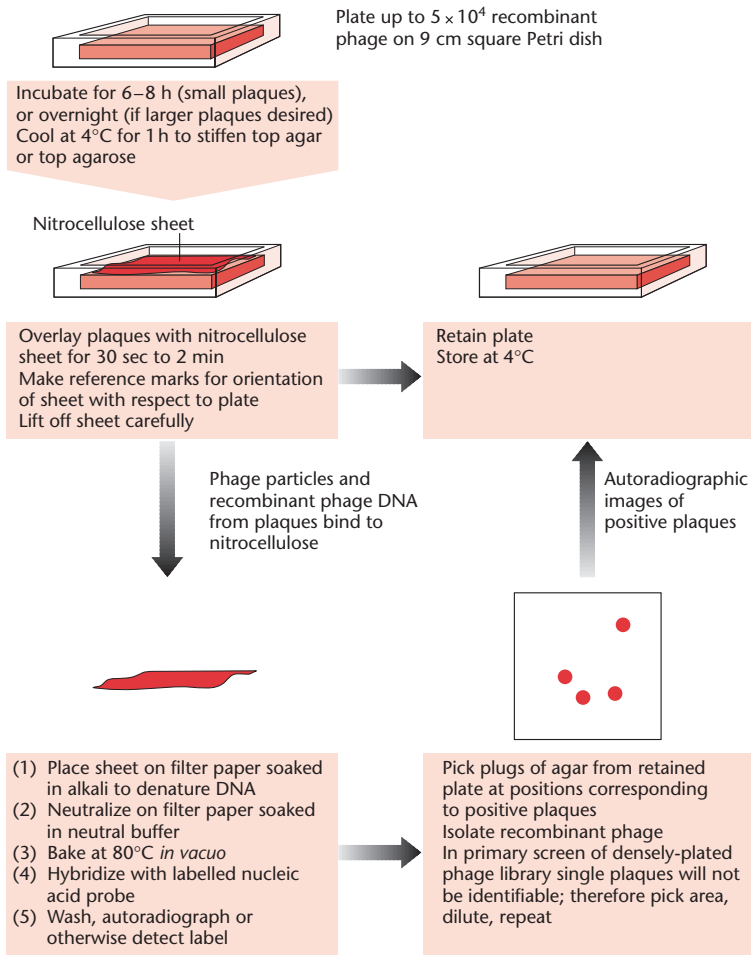


Fig. 6.13 Benton and Davis' plaque-lift procedure.

can be denatured, fixed and hybridized. This method has the advantage that several identical DNA prints can easily be made from a single-phage plate: this allows the screening to be performed in duplicate, and hence with increased reliability, and also allows a single set of recombinants to be screened with two or more probes. The Benton and Davis (1977) procedure is probably the most widely applied method of library screening, successfully applied in thousands of laboratories to the isolation of recombinant phage by nucleic acid hybridization (Fig. 6.13). More recently, however, library presentation and screening have become increasingly automated. Box 6.3 considers the advantages of gridded reference libraries.

In place of RNA probes, DNA or synthetic oligonucleotide probes can be used. A number of alternative labelling methods are also available that avoid the use of radioactivity. These methods involve the incorporation of chemical labels into the probe, such as digoxigenin or biotin, which can be detected with a specific antibody or the ligand streptavidin, respectively.

Probe design

A great advantage of hybridization for library screening is that it is extremely versatile. Conditions can be used in which hybridization is very stringent, so that only sequences identical to the probe are identified. This is necessary, for example, to identify genomic clones corresponding to a specific cDNA or to identify overlapping clones in a chromosome walk (see below). Alternatively, less stringent conditions can be used to identify both identical and related sequences. This is appropriate where a probe from one species

Box 6.3 Gridded (arrayed) hybridization reference libraries

Traditionally, library screening by hybridization involves taking a plaque lift or colony blot, which generates a replica of the distribution of clones on an agar plate. However, an alternative is to individually pick clones and arrange them on the membrane in the form of a regular grid. Once a laborious process, gridding or arraying has been considerably simplified through the use of robotics. Machines can be programmed to pick clones from microtitre dishes and spot them onto membranes at a high density; then the membrane can be hybridized with a radioactive probe as normal. Using traditional libraries, positive clones are detected by autoradiography and the X-ray film must be aligned with the original plates in order to identify the corresponding plaques. With gridded libraries, however, positive signals can be used to obtain sets of coordinates, which then identify the corresponding clone from the original microtitre dishes. Since identical sets of membranes can be easily prepared, duplicates can be distributed to other laboratories for screening. These laboratories can then determine the coordinates of their positive signals and order the corresponding clone from the source laboratory. Thus, one library can serve a number of different users and all data can be centralized (Zehetner & Lehrach 1994). Gridded libraries, while convenient for screening and data sharing, are more expensive to prepare than traditional libraries. Therefore, they are often prepared for high-value libraries with wide applications, such as genomic libraries cloned in high-capacity P1, BAC or YAC vectors (Bentley *et al*. 1992) and also for valuable cDNA libraries (Lennon & Lehrach 1991). It is possible to plate libraries at a density of one clone per well, although for PACs and BACs it is more common to pool clones in a hierarchical manner, so that individual clones may be identified by successive rounds of screening on smaller subpools (e.g. Shepherd *et al*. 1994, Shepherd & Smoller 1994).

is being used to isolate a homologous clone from another species (e.g. see Old *et al*. 1982). Probes corresponding to a conserved functional domain of a gene may also cross-hybridize with several different clones in the same species at lower stringency, and this can be exploited to identify members of a gene family. The identification of the vertebrate *Hox* genes provides an example in which cross-species hybridization was used to identify a family of related clones (Levine *et al*. 1984). In this case a DNA sequence was identified that was conserved between the *Drosophila* developmental genes *fushi tarazu* and *Antennapedia*. When this sequence, the homoeobox, was used to screen a Southern blot of DNA from other species, including frogs and humans, several hybridizing bands were revealed. This led to the isolation of a number of clones from vertebrate cDNA libraries representing the large family of *Hox* genes that play a central role in animal development.

Hybridization thus has the potential to isolate any sequence from any library *if a probe is available*. If a suitable DNA or RNA probe cannot be obtained from an existing cloned DNA, an alternative strategy is to make an oligonucleotide probe by chemical synthesis. This requires some knowledge of the amino acid sequence of the protein encoded by the target clone. However, since the genetic code is degenerate (i.e. most amino acids are specified by more than one codon), degeneracy must be incorporated into probe design so that a mixture of probes is made, at least one variant of which will specifically match the target clone. Amino acid sequences known to include methionine and tryptophan are particularly valuable because these amino acids are each specified by a single codon, hence reducing the degeneracy of the resulting probe. Thus, for example, the oligopeptide His-Phe-Pro-Phe-Met may be identified and chosen to provide a probe sequence, in which case 32 different oligonucleotides would be required:

$$5' \quad CA\begin{matrix}T\\C\end{matrix}TT\begin{matrix}T\\C\end{matrix}CCCTT\begin{matrix}T\\C\end{matrix}ATG \quad 3'$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad A$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad G$$

These 32 different sequences do not have to be synthesized individually because it is possible to perform

a mixed addition reaction for each polymerization step. This mixture is then end-labelled with a single isotopic or alternatively labelled nucleotide, using an exchange reaction. This mixed-probe method was originally devised by Wallace and co-workers (Suggs *et al.* 1981). To cover all codon possibilities, degeneracies of 64-fold (Orkin *et al.* 1983) or even 256-fold (Bell *et al.* 1984) have been employed successfully. What length of oligonucleotide is required for reliable hybridization? Even though 11-mers can be adequate for Southern blot hybridization (Singer-Sam *et al.* 1983), longer probes are necessary for good colony and plaque hybridization. Mixed probes of 14 nucleotides have been successful, although 16-mers are typical (Singer-Sam *et al.* 1983).

An alternative strategy is to use a single longer probe of 40–60 nucleotides. Here the uncertainty at each codon is largely ignored and instead increased probe length confers specificity. Such probes are usually designed to incorporate the most commonly used codons in the target species, and they may include the non-standard base inosine at positions of high uncertainty because this can pair with all four conventional bases. Such probes are sometimes termed *guessmers*. Hybridization is carried out under low stringency to allow for the presence of mismatches. This strategy is examined theoretically by Lathe (1985) and has been applied to sequences coding for human coagulation factor VIII (Toole *et al.* 1984, Wood *et al.* 1984) and the human insulin receptor (Ullrich *et al.* 1985).

Chromosome walking

Earlier in this chapter, we discussed the advantages of making genomic libraries from random DNA fragments. One of these advantages is that the resulting fragments overlap, which allows genes to be cloned by *chromosome walking*. The principle of chromosome walking is that overlapping clones will hybridize to each other, allowing them to be assembled into a contiguous sequence. This can be used to isolate genes whose function is unknown but whose genetic location is known, a technique known as *positional cloning*.

To begin a chromosome walk, it is necessary to have in hand a genomic clone that is known to lie very close to the suspected location of the target gene. In humans, for example, this could be a restriction fragment length polymorphism that has been genetically mapped to the same region. This clone is then used to screen a genomic library by hybridization, which should reveal any overlapping clones. These overlapping clones are then isolated, labelled and used in a second round of screening to identify further overlapping clones, and the process is repeated to build up a contiguous map. If the same library is used for each round of screening, previously identified clones can be distinguished from new ones, so that walking back and forth along the same section of DNA is prevented. Furthermore, modern vectors, such as λDASH and λFIX, allow probes to be generated from the end-points of a given genomic clone by *in vitro* transcription (see Fig. 6.4), which makes it possible to walk specifically in one direction. In *Drosophila*, the progress of a walk can also be monitored by using such probes for *in situ* hybridization against polytene chromosomes. Monitoring is necessary due to the dangers posed by repetitive DNA. Certain DNA sequences are highly repetitive and are dispersed throughout the genome. Hybridization with such a sequence could disrupt the orderly progress of a walk, in the worst cases causing a 'warp' to another chromosome. The probe used for stepping from one genomic clone to the next must be a unique sequence clone, or a subclone that has been shown to contain only a unique sequence.

Chromosome walking is simple in principle, but technically demanding. For large distances, it is advisable to use libraries based on high-capacity vectors, such as BACs and YACs, to reduce the number of steps involved. Before such libraries were available, some ingenious strategies were used to reduce the number of steps needed in a walk. In one of the first applications of this technology, Hogness and his co-workers (Bender *et al.* 1983) cloned DNA from the *Ace* and *rosy* loci and the homoeotic *Bithorax* gene complex in *Drosophila*. The number of steps was minimized by exploiting the numerous strains carrying well-characterized inversions and translocations of specific chromosome regions. A different strategy, called *chromosome jumping*, has been used for human DNA (Collins & Weissman 1984, Poustka & Lehrach 1986). This involves the circularization of very large genomic fragments

Box 6.4 A landmark publication. Identification of the cystic fibrosis gene by chromosome walking and jumping

Cystic fibrosis (CF) is a relatively common severe autosomal recessive disorder. Until the CF gene was cloned, there was little definite information about the primary genetic defect. The cloning of the CF gene was a breakthrough for studying the biochemistry of the disorder (abnormal chloride-channel function), for providing probes for prenatal diagnosis and for potential treatment by somatic gene therapy or other means. The publication is especially notable for the generality of the cloning strategy. In the absence of any direct functional information about the CF gene, the chromosomal location of the gene was used as the basis of the cloning strategy. Starting from markers identified by linkage analysis as being close to the CF locus on chromosome 7, a total of about 500 kb was encompassed by a combination of chromosome walking and jumping. Jumping was found to be very important to overcome problems caused by 'unclonable' regions which halted the sequential walks, and in one case achieved a distance of 100 kb (Collins *et al*. 1987). In this work, large numbers of clones were involved, obtained from several different phage and cosmid genomic libraries. Among these libraries, one was prepared using the Maniatis strategy using the λCharon 4A vector, and several were prepared using the λDASH and λFIX vectors (Fig. 6.4) after partial digestion of human genomic DNA with *Sau*3AI. Cloned regions were aligned with a map of the genome in the CF region, obtained by long-range restriction mapping using rare-cutting enzymes, such as *Not*I, in combination with pulsed-field gel electrophoresis (p. 10). The actual CF gene was detected in this cloned region by a number of criteria, such as the identification of open reading frames, the detection of cDNAs hybridizing to the genomic clones, the detection of cross-hybridizing sequences in other species and the presence of CpG islands, which are known to be associated with the 5′ ends of many genes in mammals.

From: Rommens *et al.* (1989) *Science* 245: 1059–65.

generated by digestion with endonucleases, such as *Not*I, which cut at very rare target sites. This is followed by subcloning of the region covering the closure of the fragment, thus bringing together sequences that were located a considerable distance apart. In this way a *jumping library* is constructed, which can be used for long-distance chromosome walks (Collins *et al*. 1987, Richards *et al*. 1988). The application of chromosome walking and jumping to the cloning of the human cystic fibrosis gene is discussed in Box 6.4.

Screening by PCR

The PCR is widely used to isolate specific DNA sequences from uncloned genomic DNA or cDNA, but it also a useful technique for library screening (Takumi, 1997). As a screening method, PCR has the same versatility as hybridization, and the same limitations. It is possible to identify any clone by PCR but only if there is sufficient information about its sequence to make suitable primers.*

To isolate a specific clone, PCR is carried out with gene-specific primers that flank a unique sequence in the target. A typical strategy for library screening by PCR is demonstrated by Takumi and Lodish (1994). Instead of plating the library out on agar, as would be necessary for screening by hybridization, pools of clones are maintained in multiwell plates. Each well is screened by PCR and positive wells are identified. The clones in each positive well are then

* Note that, in certain situations, clever experimental design can allow the PCR to be used to isolate specific but unknown DNA sequences. One example of this is 5′ RACE, which is discussed on p. 101. Another is inverse PCR (p. 267), which can be used to isolate unknown flanking DNA surrounding the insertion site of an integrating vector. In each case, primers are designed to bind to known sequences that are joined to the DNA fragment of interest, e.g. synthetic homopolymer tails, linkers or parts of the cloning vector.

diluted into a series in a secondary set of plates and screened again. The process is repeated until wells carrying homogeneous clones corresponding to the gene of interest have been identified.

There are also several applications where the use of *degenerate primers* is favourable. A degenerate primer is a mixture of primers, all of similar sequence but with variations at one or more positions. This is analogous to the use of degenerate oligonucleotides as hybridization probes, and the primers are synthesized in the same way. A common circumstance requiring the use of degenerate primers is when the primer sequences have to be deduced from amino acid sequences (Lee *et al.* 1988). Degenerate primers may also be employed to search for novel members of a known family of genes (Wilks 1989) or to search for homologous genes between species (Nunberg *et al.* 1989). As with oligonucleotide probes, the selection of amino acids with low codon degeneracy is desirable. However, a 128-fold degeneracy in each primer can be successful in amplifying a single-copy target from the human genome (Girgis *et al.* 1988). Under such circumstances, the concentration of any individual primer sequence is very low, so mismatching between primer and template must occur under the annealing conditions chosen. Since mismatching of the 3′-terminal nucleotide of the primer may prevent efficient extension, degeneracy at this position is to be avoided.

Screening expression libraries (expression cloning)

If a DNA library is established using expression vectors (see Chapter 5), each individual clone can be expressed to yield a polypeptide. While all libraries can be screened by hybridization or PCR, as discussed above, expression libraries are useful because they allow a range of alternative techniques to be employed, each of which exploits some structural or functional property of the gene product. This can be important in cases where the DNA sequence of the target clone is completely unknown and there is no strategy available to design a suitable probe or set of primers.

For higher eukaryotes, all expression libraries are cDNA libraries, since these lack introns and the clones are in most cases of a reasonable size. Generally, a random primer method is used for cDNA synthesis, so there is a greater representation of 5′ sequences. As discussed above, such libraries are representative of their source, so certain cDNAs are abundant and others rare. However, it should be noted that bacterial expression libraries and many yeast expression libraries are usually genomic, since there are few introns in bacteria and some yeasts and very little intergenic DNA. Efficient expression libraries can be generated by cloning randomly sheared genomic DNA or partially digested DNA, and therefore all genes are represented at the same frequency (Young *et al.* 1985). A potential problem with such libraries is that clones corresponding to a specific gene may carry termination sequences from the gene lying immediately upstream, which can prevent efficient expression. For this reason, conditions are imposed so that the size of the fragments for cloning is smaller than that of the target gene, and enough recombinants are generated for there to be a reasonable chance that each gene fragment will be cloned in all six possible reading frames (three in each orientation). Considerations for cloning DNA in *E. coli* expression vectors are discussed further in Chapter 8.

Immunological screening

Immunological screening involves the use of antibodies that specifically recognize antigenic determinants on the polypeptide synthesized by a target clone. This is one of the most versatile expression-cloning strategies, because it can be applied to any protein for which an antibody is available. Unlike the screening strategies discussed below, there is also no need for that protein to be functional. The molecular target for recognition is generally an *epitope*, a short sequence of amino acids that folds into a particular three-dimensional conformation on the surface of the protein. Epitopes can fold independently of the rest of the protein and therefore often form even when the polypeptide chain is incomplete or when expressed as a fusion with another protein. Importantly, many epitopes can form under denaturing conditions, when the overall conformation of the protein is abnormal.

The first immunological screening techniques were developed in the late 1970s, when expression libraries were generally constructed using plasmid

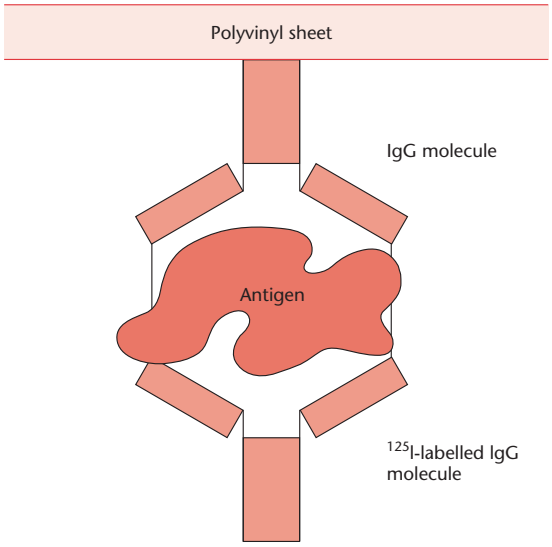


Fig. 6.14 Antigen–antibody complex formation in the immunochemical detection method of Broome and Gilbert. (See text for details.)

vectors. The method of Broome and Gilbert (1978) was widely used at the time. This method exploited the fact that antibodies adsorb very strongly to certain types of plastic, such as polyvinyl, and that IgG antibodies can be readily labelled with ^{125}I by iodination *in vitro*. As usual, transformed cells were plated out on Petri dishes and allowed to form colonies. In order to release the antigen from positive clones, the colonies were lysed, e.g. using chloroform vapour or by spraying with an aerosol of virulent phage (a replica plate is required because this procedure kills the bacteria). A sheet of polyvinyl that had been coated with the appropriate antibody was then applied to the surface of the plate, allowing antigen–antibody complexes to form. The sheet was then removed and exposed to ^{125}I-labelled IgG specific to a *different* determinant on the surface of the antigen (i.e. a determinant not involved in the initial binding of the antigen to the antibody-coated sheet (Fig. 6.14)). The sheet was then washed and exposed to X-ray film. The clones identified by this procedure could then be isolated from the replica plate. Note that this 'sandwich' technique is applicable only where two antibodies recognizing different determinants of the same protein are available. However, if the protein is expressed as a fusion, antibodies that bind to each component of the fusion can be used, efficiently selecting for recombinant molecules.

While plasmid libraries have been useful for expression screening (Helfman *et al.* 1983, Helfman & Hughes 1987), it is much more convenient to use bacteriophage-λ insertion vectors, because these have a higher capacity and the efficiency of *in vitro* packaging allows large numbers of recombinants to be prepared and screened. Immunological screening with phage-λ cDNA libraries was introduced by Young and Davies (1983) using the expression vector λgt11, which generates fusion proteins with β-galactosidase under the control of the *lac* promoter (see Box 6.1 for a discussion of λgt11 and similar fusion vectors, such as λZAP). In the original technique, screening was carried out using colonies of induced lysogenic bacteria, which required the production of replica plates, as above. A simplification of the method is possible by directly screening plaques of recombinant phage. In this procedure (Fig. 6.15), the library is plated out at moderately high density (up to 5×10^4 plaques/9 cm^2 plate), with *E. coli* strain Y1090 as the host. This *E. coli* strain overproduces the *lac* repressor and ensures that no expression of cloned sequences (which may be deleterious to the host) takes place until the inducer isopropyl-β-D-thiogalactoside (IPTG) is presented to the infected cells. Y1090 is also deficient in the *lon* protease, hence increasing the stability of recombinant fusion proteins. Fusion proteins expressed in plaques are absorbed on to a nitrocellulose membrane overlay, and this membrane is processed for antibody screening. When a positive signal is identified on the membrane, the positive plaque can be picked from the original agar plate (a replica is not necessary) and the recombinant phage can be isolated.

The original detection method using iodinated antibodies has been superseded by more convenient methods using non-isotopic labels, which are also more sensitive and have a lower background of non-specific signal. Generally, these involve the use of unlabelled primary antibodies directed against the polypeptide of interest, which are in turn recognized by secondary antibodies carrying an enzymatic label. As well as eliminating the need for isotopes, such methods also incorporate an amplification step, since two or more secondary antibodies bind to the primary antibody. Typically, the secondary antibody

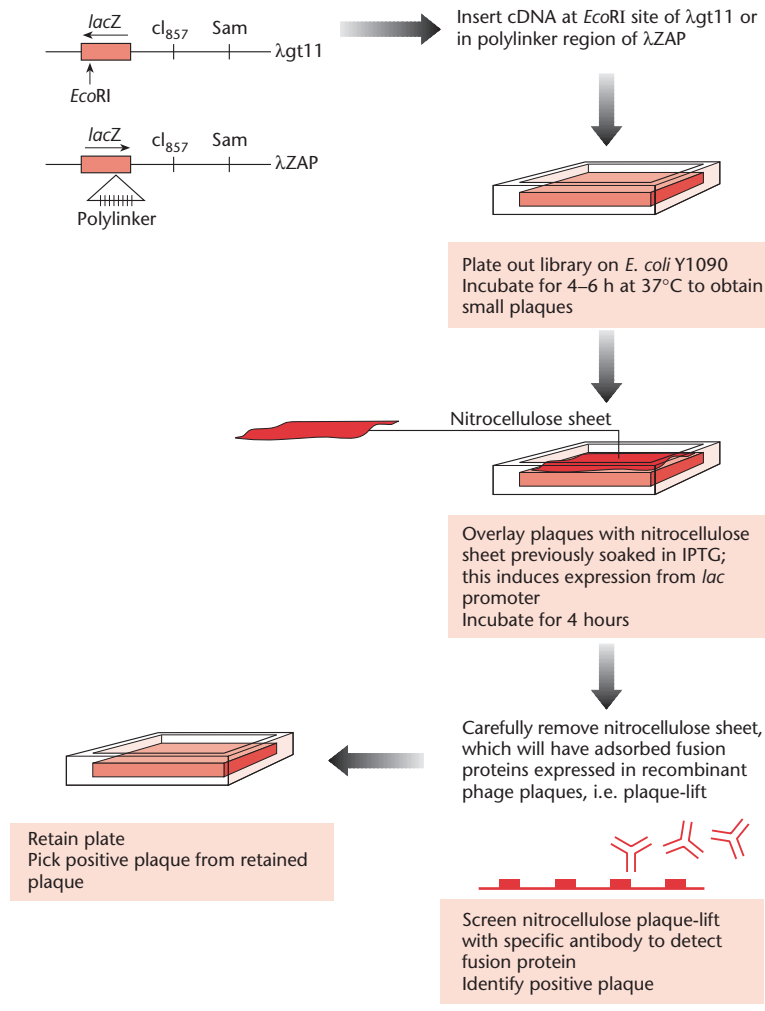


Fig. 6.15 Immunochemical screening of λgt11 or λZAP recombinant plaques.

recognizes the species-specific constant region of the primary antibody and is conjugated to either horseradish peroxidase (De Wet *et al.* 1984) or alkaline phosphatase (Mierendorf *et al.* 1987), each of which can in turn be detected using a simple colorimetric assay carried out directly on the nitrocellulose filter. Polyclonal antibodies, which recognize many different epitopes, provide a very sensitive probe for immunological screening, although they may also cross-react to proteins in the expression host. Monoclonal antibodies and cloned antibody fragments can also be used, although the sensitivity of such reagents is reduced because only a single epitope is recognized.

South-western and north-western screening

We have seen how fusion proteins expressed in plaques produced by recombinant λgt11 or λZAP vectors may be detected by immunochemical screening. A closely related approach has been used for the screening and isolation of clones expressing sequence-specific DNA-binding proteins. As above, a plaque lift is carried out to transfer a print of the library on to nitrocellulose membranes. However, the screening is carried out, without using an antibody, by incubating the membranes with a radiolabelled *double-stranded* DNA oligonucleotide probe, containing the recognition sequence for the target DNA-binding

protein. This technique is called *south-western* screening, because it combines the principles of Southern and western blots. It has been particularly successful in the isolation of clones expressing cDNA sequences corresponding to certain mammalian transcription factors (Singh *et al.* 1988, Staudt *et al.* 1988, Vinson *et al.* 1988, Katagiri *et al.* 1989, Williams *et al.* 1991, Xiao *et al.* 1991). A limitation of this technique is that, since individual plaques contain only single cDNA clones, transcription factors that function only in the form of heterodimers or as part of a multimeric complex do not recognize the DNA probe and the corresponding cDNAs cannot be isolated. Clearly the procedure can also be successful only in cases where the transcription factor remains functional when expressed as a fusion polypeptide. It is also clear that the affinity of the polypeptide for the specific DNA sequence must be high, and this has led to the preferential isolation of certain types of transcription factor (reviewed by Singh 1993). More recently, a similar technique has been used to isolate sequence-specific RNA-binding proteins, in this case using a single-stranded RNA probe. By analogy to the above, this is termed *north-western screening* and has been successful in a number of cases (e.g. see Qian & Wilusz 1993; reviewed by Bagga & Wilusz 1999). Both south-western and north-western screening are most efficient when the oligonucleotide contains the binding sequence in multimeric form. This may mean that several fusion polypeptides on the filter bind to each probe, hence greatly increasing the average dissociation time. To minimize non-specific binding, a large excess of unlabelled double-stranded DNA (or single-stranded RNA) is mixed with the specific probe. However, it is usually necessary to confirm the specificity of binding in a second round of screening, using the specific oligonucleotide probe and one or more alternative probes containing a similar sequences that are not expected to be recognized.

Screening with alternative ligands

As well as DNA and RNA, a whole range of alternative 'ligands' can be used to identify polypeptides that specifically bind certain molecules (for example, as an alternative to south-western screening). Such techniques are not widely used because they generally have a low sensitivity and their success depends on the preservation of the appropriate interacting domain of the protein when exposed on the surface of a nitrocellulose filter. Furthermore, as discussed in Chapter 9, the yeast two-hybrid system and its derivatives now provide versatile assay formats for many specific types of protein–protein interaction, with the advantage that such interactions are tested in living cells, so the proteins involved are more likely to retain their functional interacting domains.

Functional cloning

Finally, we consider screening methods that depend on the full biological activity of the protein. This is often termed *functional cloning*. In contrast to positional cloning, described above, functional cloning is possible in complete ignorance of the whereabouts of the gene in the genome and requires no prior knowledge of the nucleotide sequence of the clone or the amino acid sequence of its product. As long as the expressed protein is functional and that function can be exploited to screen an expression library, the corresponding clone can be identified.

Screening by functional complementation

Functional complementation is the process by which a particular DNA sequence compensates for a missing function in a mutant cell, and thus restores the wild-type phenotype. This can be a very powerful method of expression cloning, because, if the mutant cells are non-viable under particular growth conditions, cells carrying the clone of interest can be positively selected, allowing the corresponding clones to be isolated.

Ratzkin and Carbon (1977) provide an early example of how certain eukaryotic genes can be cloned on the basis of their ability to complement auxotrophic mutations in *E. coli*. These investigators inserted fragments of yeast DNA, obtained by mechanical shearing, into the plasmid ColEl, using a homopolymer-tailing procedure. They transformed *E. coli his*B mutants, which are unable to synthesize histidine, with the recombinant plasmids and plated the bacteria on minimal medium. In this way, they selected for complementation of the mutation and isolated clones carrying an expressed yeast *his* gene. If the function of the gene is highly conserved, it is

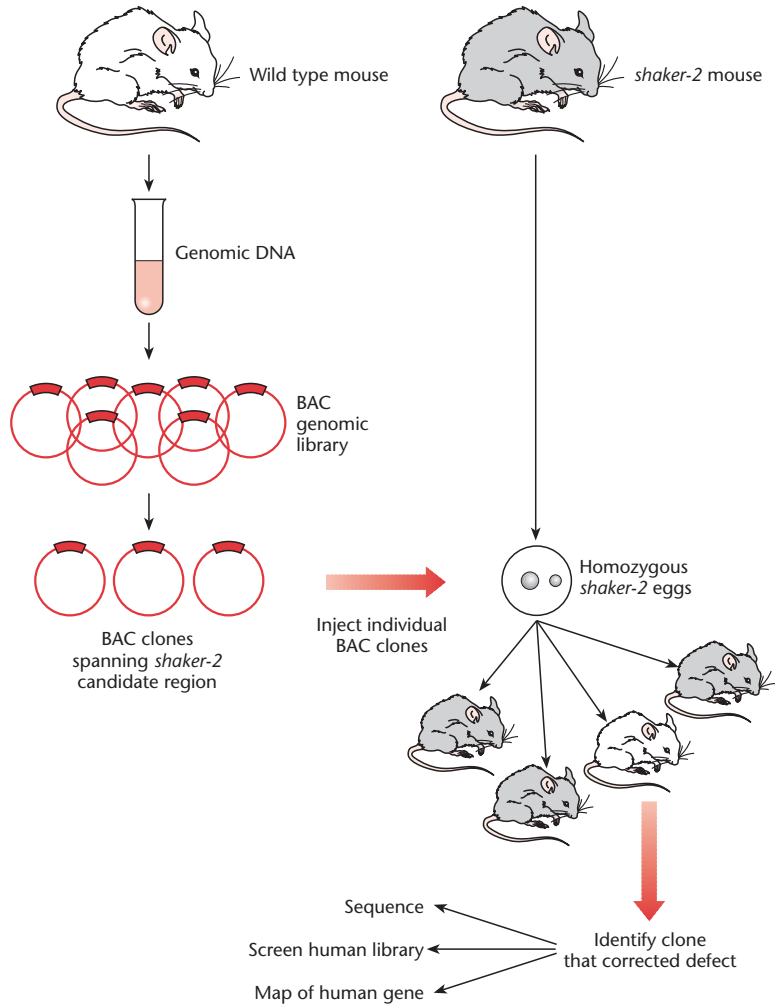


Fig. 6.16 Functional complementation in transgenic mice to isolate the *Shaker-2* gene. Homozygous *shaker-2* fertilized mouse eggs were injected with BAC clones derived from the *Shaker-2* candidate region of a wild-type mouse. Progeny were screened for restoration of the wild-type phenotype, thus identifying the BAC clone corresponding to the *Shaker-2* gene. This clone is then sequenced and used to isolate and map the corresponding human disease gene *DFNB3*.

quite possible to carry out functional cloning of, for example, mammalian proteins in bacteria and yeast. Thus, complementation in yeast has been used to isolate cDNAs for a number of mammalian metabolic enzymes (e.g. Botstein & Fink 1988) and certain highly conserved transcription factors (e.g. Becker *et al.* 1991), as well as regulators of meiosis in plants (Hirayama *et al.* 1997). This approach can also be used in mammalian cells, as demonstrated by Strathdee *et al.* (1992), who succeeded in isolating the *FACC* gene, corresponding to complementation group C of Fanconi's anaemia. Generally a pool system is employed, where cells are transfected with a complex mix of up to 100 000 clones. Pools that successfully complement the mutant phenotype are then subdivided for a further round of transfection, and the procedure repeated until the individual cDNA responsible is isolated.

Functional complementation is also possible in transgenic animals and plants. In this way, Probst *et al.* (1998) were able to clone the mouse deafness-associated gene *Shaker-2*, and from there its human homologue, *DFNB3* (Fig. 6.16). The *shaker-2* mutation had previously been mapped to a region of the mouse genome that is syntenic to the region involved in a human deafness disorder. BAC clones corresponding to this region were therefore prepared from wild-type mice and microinjected into the eggs of *shaker-2* mutants. The resulting transgenic mice were screened for restoration of a normal hearing

phenotype, allowing a BAC clone corresponding to the functional *Shaker-2* gene to be identified. The gene was shown to encode a cytoskeletal myosin protein. This was then used to screen a human genomic library, resulting in the identification of the equivalent human gene. Note that no sequence information was required for this screening procedure, and without the functional assay there would have been no way to identify either the mouse or human gene except through a laborious chromosome walk from a linked marker. The recent development of high-capacity transformation vectors for plants (p. 236) has allowed similar methods to be used to identify plant genes (e.g. Sawa *et al.* 1999, Kubo & Kakimoto 2001).

Screening by 'gain of function'

Complementation analysis can be used only if an appropriate mutant expression host is available. In many cases, however, the function of the target gene is too specialized for such a technique to work in a bacterial or yeast expression host and, even in a higher eukaryotic system, loss of function in the host may be fully or partially compensated by one or more other genes. As an alternative, it may be possible to identify clones on the basis that they confer a gain of function on the host cell. In some cases, this gain of function is a selectable phenotype that allows cells containing the corresponding clone to be positively selected. For example, in an early example of the expression of a mammalian gene in *E. coli*, Chang *et al.* (1978) constructed a population of recombinant plasmids containing cDNA derived from unfractionated mouse mRNA. This population of mRNA molecules was expected to contain the transcript for dihydrofolate reductase (DHFR). Mouse DHFR is much less sensitive to inhibition by the drug trimethoprim than *E. coli* DHFR, so growing transformants in medium containing the drug allowed selection for those cells containing the mouse *Dhfr* cDNA.

In other cases, the phenotype conferred by the clone of interest is not selectable, but can be detected because it causes a visible change in phenotype. In mammalian cells, for example, clones corresponding to cellular oncogenes have been identified on the basis of their ability to stimulate the proliferation of quiescent mouse 3T3 fibroblast cells either in culture or when transplanted into 'nude mice' (e.g. Brady *et al.* 1985). Many different specific assays have also been developed for the functional cloning of cDNAs encoding particular types of gene product. For example, *Xenopus* melanophores have been used for the functional cloning of G-protein-coupled receptors. Melanophores are dark cells containing many pigment organelles, called melanosomes. A useful characteristic of these organelles is that they disperse when adenyl cyclase or phospholipase C are active and aggregate when these enzymes are inhibited. Therefore, the expression of cDNAs encoding G-protein-coupled receptors and many types of receptor tyrosine kinases leads to redistribution of pigmentation within the cell, which can be used as an assay for the identification of receptor cDNAs (reviewed be Lerner 1994).

Difference cloning

Difference cloning refers to a range of techniques used to isolate sequences that are represented in one source of DNA but not another. Normally this means differentially expressed cDNAs, representing genes that are active in one tissue but inactive in another, but the technique can also be applied to genomic DNA to identify genes corresponding to deletion mutants. There are a number of cell-based differential cloning methods and also a range of PCR techniques. Each method follows one of two principles: either the differences between two sources are displayed, allowing differentially expressed clones to be visually identified, or the differences are exploited to generate a collection of clones that are enriched for differentially expressed sequences. The analysis of differential gene expression has taken on new importance recently with the advent of high-throughput techniques allowing the monitoring of many and, in some cases, all genes simultaneously.

Difference cloning with DNA libraries

Displaying differences – differential screening

An early approach to difference cloning was *differential screening*, a simple variation on normal hybridization-based library screening protocols that is useful for

the identification of differentially expressed cDNAs that are also moderately abundant (e.g. Dworkin & Dawid 1980). Let us consider, for example, the isolation of cDNAs derived from mRNAs which are abundant in the gastrula embryo of the frog *Xenopus* but which are absent, or present at low abundance, in the egg. A cDNA library is prepared from gastrula mRNA. Replica filters carrying identical sets of recombinant clones are then prepared. One of these filters is then probed with ^{32}P-labelled mRNA (or cDNA) from gastrula embryos and one with ^{32}P-labelled mRNA (or cDNA) from the egg. Some colonies will give a positive signal with both probes; these represent cDNAs derived from mRNA types that are abundant at both stages of development. Some colonies will not give a positive signal with either probe; these correspond to mRNA types present at undetectably low abundance in both tissues. This is a feature of using *complex probes*, which are derived from mRNA populations rather than single molecules: only abundant or moderately abundant sequences in the probe carry a significant proportion of the label and are effective in hybridization. Importantly, some colonies give a positive signal with the gastrula probe, but not with the egg probe. These can be visually identified and should correspond to differentially expressed sequences.

A recent resurgence in the popularity of differential screening has come about through the development of DNA microarrays (Schena *et al*. 1995). In this technique, cDNA clones are transferred to a miniature solid support in a dense grid pattern and screened simultaneously with complex probes from two sources, which are labelled with different fluorophores. Clones that are expressed in both tissues will fluoresce in a colour that represents a mixture of fluorophores, while differentially expressed clones will fluoresce in a colour closer to the pure signal of one or other of the probes. A similar technique involves the use of DNA chips containing densely arrayed oligonucleotides. These methods are compared in Box 6.5.

Enrichment for differences – subtractive cloning

An alternative to differential screening is to generate a library that is enriched in differentially expressed clones by removing sequences that are common to two sources. This is called a *subtracted cDNA library* and should greatly assist the isolation of rare cDNAs. If we use the same example as above, the aim of the experiment would be to generate a library enriched for cDNAs derived from gastrula-specific mRNAs. This would be achieved by hybridizing first-strand cDNAs prepared from gastrula mRNA with a large excess of mRNA from *Xenopus* oocytes. If this driver population is labelled in some way, allowing it to be removed from the mixed population, only gastrula-specific cDNAs would remain behind. A suitable labelling method would be to add biotin to all the oocyte mRNA, allowing oocyte/gastrula RNA/cDNA hybrids as well as excess oocyte mRNA to be subtracted by binding to streptavidin, for which biotin has great affinity (Duguid *et al*. 1988, Rubinstein *et al*. 1990). Highly enriched libraries can be prepared by several rounds of extraction with driver mRNA, resulting in highly enriched subtracted libraries (reviewed by Sagerstrom *et al*. 1997).

An example of subtractive cDNA cloning and differential screening is provided by Nedivi *et al*. (1993). These investigators were interested in the isolation of rat cDNAs that are induced in a particular region of the brain (the dentate gyrus (DG)) known to be involved in learning and memory. The inducing stimulus was kainate, a glutamate analogue that induces seizures and memory-related synaptic changes. Poly(A)$^+$ RNA was extracted from the DG of kainate-treated animals and used for first-strand cDNA synthesis. Ubiquitous sequences present in the activated DG cDNA preparation were hybridized with an excess of poly(A)$^+$ RNA from total uninduced rat brain. This RNA had previously been biotinylated (using a photobiotinylation procedure) and so hybrids and excess RNA could be removed using a streptavidin extraction method (Sive & St John 1988). The unhybridized cDNA was then converted into double-stranded form by conventional methods and used to construct the subtracted cDNA library in λZAP. This subtracted library was differentially screened using radiolabelled cDNA from activated and non-activated DG as the differential probes. A large number of activated DG clones were isolated, of which 52 were partially sequenced. One-third of these clones corresponded to known genes; the remainder were new.

Box 6.5 Differential screening with DNA chips

cDNA microarrays

Miniaturization and automation have facilitated the development of DNA microarrays, in which DNA sequences are displayed on the surface of a small 'chip' of either nylon or glass. In the initial description of this technology (Schena *et al*. 1995), up to 10 000 cDNA clones, each in the region of several hundred nucleotides in length, could be arrayed on a single microscope slide. The cDNAs were either obtained from an existing library or generated *de novo* by PCR. In each case, the machine transfers a small amount of liquid from a standard 96-well microtitre plate on to a poly-L-lysine-coated microscope slide, and the DNA is fixed in position by UV irradiation. Arrays are used predominantly for the multiplex analysis of gene expression profiles. Total RNA is used to prepare fluorescently labelled cDNA probes and signals are detected using a laser. Each hybridization experiment generates a large amount of data. Comparisons of expression profiles generated using probes from different sources can identify genes that are differentially expressed in various cell types, at different developmental stages or in response to induction (reviewed by Schena *et al*. 1998, Xiang & Chen 2000). There have been many successes with this relatively new technology, including the identification of genes involved in the development of the nervous system (Wen *et al*. 1998) and genes involved in inflammatory disease (Heller *et al*. 1997). Arrays have been constructed including every gene in the genome of *E. coli* (Tao *et al*. 1999) and most genes of the yeast *Saccharomyces cerevisiae* (De Risi *et al*. 1997). This has allowed comprehensive parallel analysis of the expression of all genes simultaneously in a variety of experimental assays (Cho *et al*. 1998, Chu *et al*. 1998, Spellman *et al*. 1998, Jelinsky & Samson 1999). It is likely that complete genome arrays will be available for higher eukaryotes, including humans, within the next few years, offering an unprecedented ability to capture functional snapshots of the genome in action.

Oligonucleotide chips

An alternative to spotting presynthesized cDNAs or ESTs on to slides is to synthesize oligonucleotides *in situ* on silicon or glass wafers, using similar processes to those used in the manufacture of semiconductors (Lockhart *et al*. 1996, Shalon *et al*. 1996). Using current techniques, up to 1 000 000 oligonucleotides can by synthesized in tightly packed regular arrays on chips approximately 1 cm^2 (Lipshutz *et al*. 1999). Unlike cDNA arrays, a hybridizing probe sequence is recognized not by a single cognate cDNA, but by a combination of short oligonucleotides, from which its sequence can be deduced (reviewed by Southern 1996a,b). Chips are more versatile than arrays, because they can be used not only for expression analysis but also for DNA sequencing (resequencing) (Chee *et al*. 1996) and the analysis of differences between genomes at the level of single nucleotide polymorphisms (Hacia 1999).

Difference cloning by PCR

Displaying differences – differential-display PCR and arbitrarily primed PCR

As expected, PCR-based methods for difference cloning are more sensitive and rapid than library-based methods, and can be applied to small amounts of starting material. Two similar methods have been described which use pairs of short arbitrary primers to amplify pools of partial cDNA sequences. If the same primer combinations are used to amplify cDNAs from two different tissues, the products can be fractionated side by side on a sequencing gel, and differences in the pattern of bands generated, the *mRNA fingerprint*, therefore reveal differentially expressed genes (Fig. 6.17). Essentially, the distinction between the two techniques concerns the

Box 6.6 A landmark publication. Subtraction cloning of the human Duchenne muscular dystrophy (*DMD*) gene

While most subtractive cloning experiments involve cDNAs, this publication reports one of the few successful attempts to isolate a gene using a subtracted *genomic* library. The study began with the identification of a young boy, known as 'BB', who suffered from four X-linked disorders, including DMD. Cytogenetic analysis showed that the boy had a chromosome deletion in the region Xp21, which was known to be the DMD locus. Kunkel's group then devised a subtraction-cloning procedure to isolate the DNA sequences that were deleted in BB. Genomic DNA was isolated from BB and randomly sheared, generating fragments with blunt ends and non-specific overhangs. DNA was also isolated from an aneuploid cell line with four (normal) X chromosomes. This DNA was digested with the restriction enzyme *Mbo*I, generating sticky ends suitable for cloning. The *Mbo*I fragments were mixed with a large excess of the randomly sheared DNA from BB, and the mixture was denatured and then persuaded to reanneal extensively, using phenol enhancement. The principle behind the strategy was that, since the randomly fragmented DNA was present in a vast excess, most of the DNA from the cell line would be sequestered into hybrid DNA molecules that would be unclonable. However, those sequences present among the *Mbo*I fragments but absent from BB's DNA due to the deletion would only be able to reanneal to complementary strands from the cell line. Such strands would have intact *Mbo*I sticky ends and could therefore be ligated into an appropriate cloning vector. Using this strategy, Kunkel and colleagues generated a genomic library that was highly enriched for fragments corresponding to the deletion in BB. Subclones from the library were tested by hybridization against normal DNA and DNA from BB to confirm that they mapped to the deletion. To confirm that the genuine DMD gene had been isolated, the positive subclones were then tested against DNA from many other patients with DMD, revealing similar deletions in 6.5% of cases.

From Kunkel (1986) *Nature* 322: 73–77.

primer used for first-strand cDNA synthesis. In the differential-display PCR technique (Liang & Pardee 1992), the antisense primer is an oligo-dT primer with a specific two-base extension, which thus binds at the 3′ end of the mRNA. Conversely, in the arbitrarily primed PCR method (Welsh *et al.* 1992), the antisense primer is arbitrary and can in principle anneal anywhere in the message. In each case, an arbitrary sense primer is used, allowing the amplification of partial cDNAs from pools of several hundred mRNA molecules. Following electrophoresis, differentially expressed cDNAs can be excised from the gel and characterized further, usually to confirm its differential expression.

Despite the fact that these display techniques are problematical and appear to generate a large number of false-positive results, there have been remarkable successes. In the original report by Liang and Pardee (1992), the technique was used to study differences between tumour cells and normal cells, resulting in the identification of a number of genes associated with the onset of cancer (Liang *et al.* 1992). Further cancer-related gene products have been discovered by other groups using differential display (Sager *et al.* 1993, Okamato & Beach 1994). The technique has also been used successfully to identify developmentally regulated genes (e.g. Adati *et al.* 1995) and genes that are induced by hormone treatment (Nitsche *et al.* 1996). An advantage of display techniques over subtracted libraries is that changes can be detected in related mRNAs representing the same gene family. In subtractive-cloning procedures, such differences are often overlooked because the excess of driver DNA can eliminate such sequences (see review by McClelland *et al.* 1995).

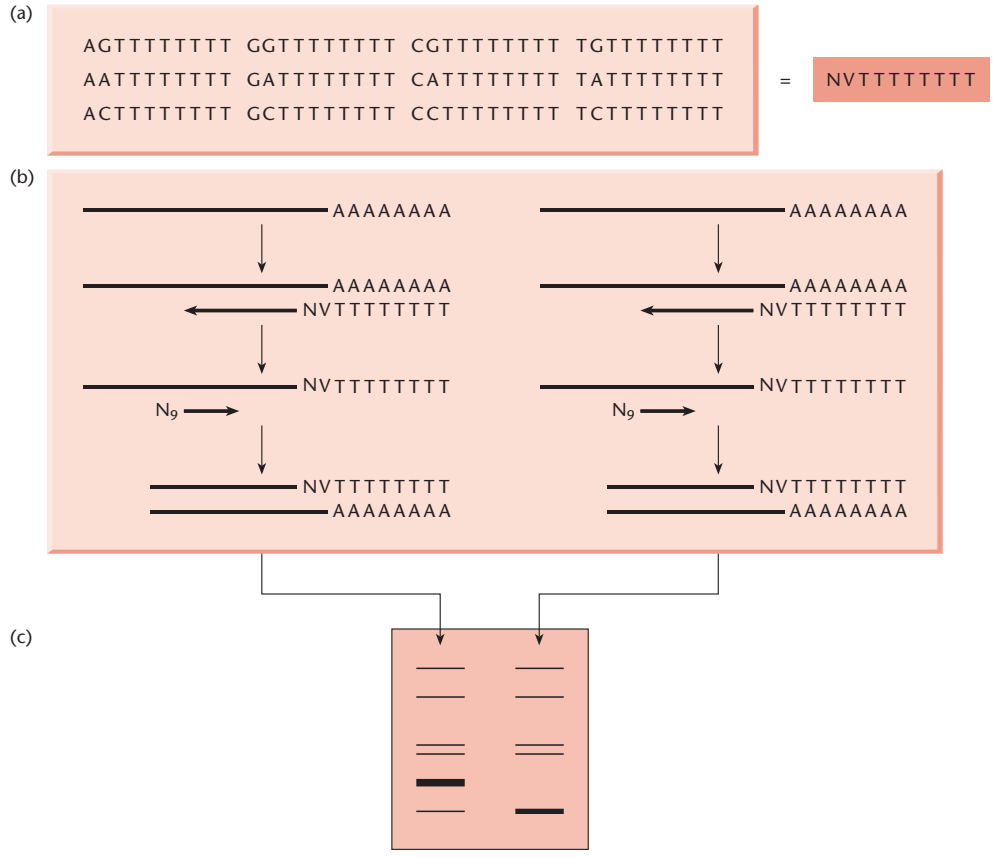


Fig. 6.17 Summary of the differential mRNA display technique, after Liang and Pardee (1992). (a) A set of 12 oligo(dT) primers is synthesized, each with a different two-base extension; the generic designation for this primer set is NVTTTTTTTTT, where N is any nucleotide and V is any nucleotide except T. (b) Messenger RNA from two sources is then converted into cDNA using these primers, generating 12 non-overlapping pools of first strand cDNA molecules for each source, The PCR is then carried out using the appropriate oligo(dT) primer and a set of arbitrary 9-mers (N_9), which may anneal anywhere within the cDNA sequence. This facilitates the amplification of pools of cDNA fragments, essentially the same as expressed sequence tages (ESTs). (c) Pools of PCR products, derived from alternative mRNA sources but amplified with the same pair of primers, are then compared side by side on a sequencing gel. Bands present in one lane but absent from the other are likely to represent differentially-expressed genes. The corresponding bands can be excised from the sequencing gel and the PCR products subcloned, allowing sequence annotation and expression analysis, e.g. by northern blot or *in situ* hybridization, to confirm differential expression.

Enrichment for differences – representational difference analysis

Representational difference analysis is a PCR subtraction technique, i.e. common sequences between two sources are eliminated prior to amplification. The method was developed for the comparative analysis of genomes (Lisitsyn *et al.* 1993) but has been modified for cloning differentially expressed genes (Hubank & Schatz 1994). Essentially, the technique involves the same principle as subtraction hybridization, in that a large excess of a DNA from one source, the driver, is used to make common sequences in the other source, the tester, unclonable (in this case unamplifiable). The general scheme is shown in Fig. 6.18. cDNA is prepared from two sources, digested with restriction enzymes and amplified. The amplified products from one source

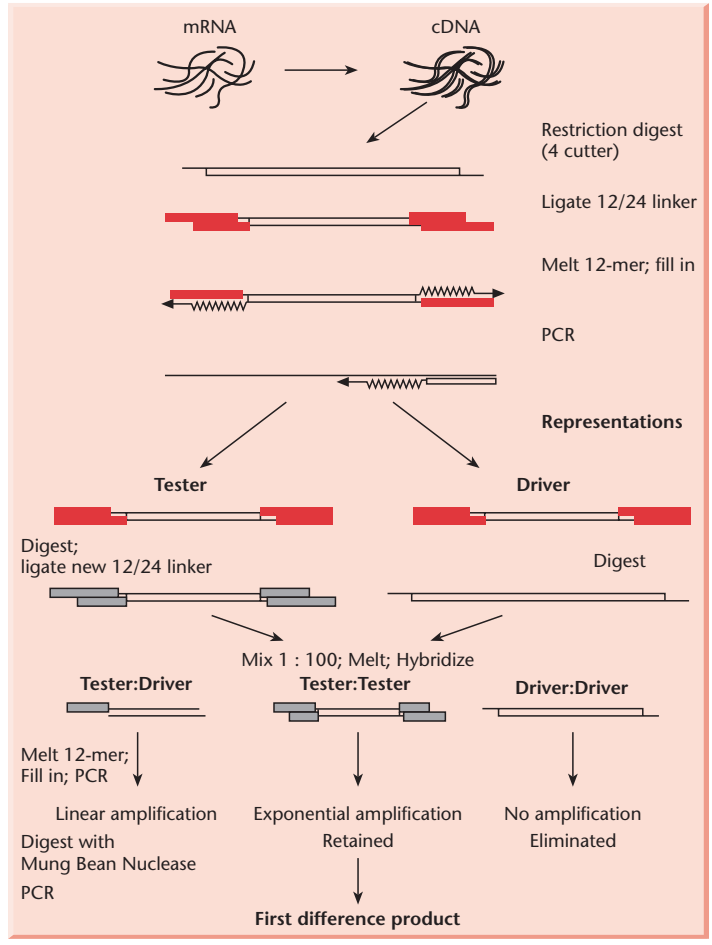


Fig. 6.18 Basic strategy for cDNA-based representational difference analysis. See text for details.

are then annealed to specific linkers that provide annealing sites for a unique pair of PCR primers. These linkers are not added to the driver cDNA. A large excess of driver cDNA is then added to the tester cDNA and the populations are mixed. Driver/driver fragments possess no linkers and cannot be amplified, while driver/tester fragments possess only one primer annealing site and will only be amplified in a linear fashion. However, cDNAs that are present only in the tester will possess linkers on both strands and will be amplified exponentially and can therefore be isolated and cloned.